Inferring the timing and strength of natural selection and gene migration in the evolution of chicken from ancient DNA data

Wenyang Lyu[a], Xiaoyang Dai[b,1], Mark Beaumont[b], Feng Yu[a,*], Zhangyi He[c,2,*]

[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*
[c]*MRC Toxicology Unit, University of Cambridge, Cambridge CB2 1QR, United Kingdom*

Abstract

With the rapid growth of the number of sequenced ancient genomes, there has been increasing interest in using this new information to study past and present adaptation. Such an additional temporal component has the promise of providing improved power for the estimation of natural selection. Over the last decade, statistical approaches for detection and quantification of natural selection from ancient DNA (aDNA) data have been developed. However, most of the existing methods do not allow us to estimate the timing of natural selection along with its strength, which is key to understanding the evolution and persistence of organismal diversity. Additionally, most methods ignore the fact that natural populations are almost always structured. This can result in overestimation of the effect of natural selection. To address these issues, we propose a novel Bayesian framework for the inference of natural selection and gene migration from aDNA data with Markov chain Monte Carlo techniques, co-estimating both timing and strength of natural selection and gene migration. Such an advance enables us to infer drivers of natural selection and gene migration by correlating genetic evolution with potential causes such as the changes in the ecological context in which an organism has evolved. The performance of our procedure is evaluated through extensive simulations, with its utility shown with an application to ancient chicken samples.

Keywords: Natural selection, Gene migration, Continent-island model, Wright-Fisher diffusion, Hidden Markov model, Blockwise particle marginal Metropolis-Hastings

*Corresponding author.
Email addresses: `feng.yu@bristol.ac.uk` (Feng Yu), `z.he@imperial.ac.uk` (Zhangyi He)
[1]Present address: The Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, United Kingdom
[2]Present address: Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, United Kingdom

1. Introduction

With modern advances in ancient DNA (aDNA) techniques, there has been a rapid increase in the availability of time serial samples of segregating alleles across one or more related populations. The temporal aspect of such samples reflects the combined evolutionary forces acting within and among populations such as genetic drift, natural selection and gene migration, which can contribute to our understanding of how these evolutionary forces are responsible for the patterns observed in contemporaneous samples. One of the most powerful applications of such genetic time series is to study the action of natural selection since the expected changes in allele frequencies over time are closely related to the strength of natural selection. For example, Ludwig et al. (2009) sequenced 87 ancient horse samples ranging from a pre- to a post-domestication period for eight loci encoding horse coat colour and found that natural selection strongly acted on the two loci encoding the Agouti signalling peptide (*ASIP*) and the melanocortin 1 receptor (*MC1R*), respectively. A detailed review of the most recent findings in aDNA can be found in Leonardi et al. (2017).

Over the past fifteen years, there has been a growing literature on the statistical inference of natural selection from time series data of allele frequencies, especially in aDNA (see Malaspinas, 2016; Dehasque et al., 2020, for excellent reviews), which benefits from the fact that aDNA data can provide direct information regarding the temporal changes in allele frequencies. Typically, estimating natural selection from time series genetic data is built on the hidden Markov model (HMM) framework proposed by Bollback et al. (2008), where the allele frequency trajectory of the underlying population through time was modelled as a latent variable following the Wright-Fisher model introduced by Fisher (1922) and Wright (1931), and the allele frequency of the sample drawn from the underlying population at each sampling time point was treated as a noisy observation of the underlying population allele frequency. In their likelihood computation, the Wright-Fisher model was approximated through its standard diffusion limit, called the Wright-Fisher diffusion, which was then discretised for numerical integration through a finite difference scheme. Their approach was applied by Ludwig et al. (2009) to analyse aDNA data associated with horse coat colouration.

A number of statistical approaches have been developed based on the HMM framework of Bollback et al. (2008) for the inference of natural selection from time series allele frequency data.

Malaspinas et al. (2012) proposed a likelihood-based approach that allowed joint estimation of the selection coefficient, the population size and the allele age, where the Wright-Fisher diffusion was approximated with a one-step Markov process in their likelihood computation. Steinrücken et al. (2014) introduced a likelihood-based procedure to estimate the selection coefficient with a spectral representation of the Wright-Fisher diffusion devised by Song & Steinrücken (2012), which allowed for general diploid models of natural selection, *e.g.*, natural selection could act in an over- or underdominant fashion. Ferrer-Admetlla et al. (2016) extended the HMM framework of Bollback et al. (2008) by approximating the Wright-Fisher diffusion with a coarse-grained Markov model that preserves the long-term behaviour of the Wright-Fisher diffusion, which can additionally estimate the mutation rate. Schraiber et al. (2016) presented a Bayesian procedure for the joint inference of natural selection and allele age, where they applied a path augmentation approach to carry out their posterior computation through Markov chain Monte Carlo (MCMC). Such an MCMC-based procedure can be easily extended to handle more complex demographic scenarios such as non-constant population sizes and general diploid models of natural selection. He et al. (2020c) provided an alternative likelihood-based method for co-estimating the selection coefficient and the allele age, where they conditioned the Wright-Fisher diffusion to survive until the time of the most recent sample under the HMM framework of Bollback et al. (2008). This conditioning takes advantage of the fact that aDNA samples are often collected from only those loci that are segregating in the most recent sample. Their likelihood computation was carried out backwards in time, which greatly reduced the computational cost. He et al. (2020b) extended the HMM framework of Bollback et al. (2008) to estimate the selection coefficients from time series data of allele frequencies across linked loci with the particle marginal Metropolis-Hastings (PMMH) of Andrieu et al. (2010). Their procedure enables us to explicitly model the genetic recombination effect and the local linkage information, which greatly improves the inference of natural selection, especially for tightly-linked loci.

To our knowledge, natural populations are almost always structured, which affects the relative effect of natural selection and genetic drift on the changes in allele frequencies over time. This can lead to overestimation of the selection coefficient (Mathieson et al., 2015). However, as far as we know, all existing methods based on the Wright-Fisher model for the inference of natural selection from time series data of allele frequencies lack the ability to account for the

confounding effect of gene migration, with the exception of Mathieson & McVean (2013), which could model population structure in their inference of natural selection. Mathieson & McVean (2013) is an extension of Bollback et al. (2008) for the inference of metapopulations, which enables joint estimation of the selection coefficient and the migration rate from time series allele frequency data. However, their approach could become computationally cumbersome for large population sizes and evolutionary timescales since their likelihood computation was carried out with the Wright-Fisher model.

More recently, Loog et al. (2017) developed a Bayesian statistical framework for quantifying the timing and strength of natural selection from genetic time series while explicitly accounting for gene migration from external sources. Loog et al. (2017) also allowed joint estimation of the allele frequency trajectory of the underlying population through time, which was important for understanding the drivers of natural selection. However, the population size in their work was assumed to be large enough to ignore genetic drift, which greatly simplifies the computation of their likelihood but highly limits the application of their method to aDNA data.

In this work, we develop a novel MCMC-based method for the Bayesian inference of natural selection and gene migration to re-analyse the temporally-spaced ancient chicken samples from Loog et al. (2017). Our approach is built upon the HMM framework of Bollback et al. (2008), but unlike most existing methods, it enables the joint estimation of the timing and strength of natural selection and gene migration. Such an advance allows us to infer the drivers of natural selection and gene migration by correlating genetic evolution with ecological and cultural shifts. Our key innovation is to propose a multi-allele Wright-Fisher diffusion for a single locus evolving under natural selection and gene migration, including the timing of natural selection and gene migration. This diffusion process characterises the allele frequency trajectories of the underlying population through time, where the alleles that migrate from external sources are no longer treated as the same as those that originate in the underlying population. Such a setup allows us to take full advantage of known quantities such as the proportion of the modern European chicken that have Asian origin as a direct result of gene migration. Our Bayesian inference of natural selection and gene migration is carried out by the PMMH of Andrieu et al. (2010) with blockwise sampling, which permits a joint update of the underlying population allele frequency trajectories. We partition the population genetic quantities of interest into two disjoint blocks,

composed of selection-related and migration-related parameters, respectively, and iteratively update one block at a time through the PMMH.

We evaluate the performance of our procedure through extensive simulations, showing that our method allows for efficient and accurate inference of natural selection and gene migration from time series data of allele frequencies, even if the samples are sparsely distributed with small uneven sizes. We then apply our procedure to re-analyse the time serial samples of segregating alleles from ancient chicken reported in Loog et al. (2017), further demonstrating that positive selection acted on the locus encoding for the thyroid-stimulating hormone receptor (*TSHR*), and the timing of the onset of positive selection coincided with a historically attested cultural shift in food preferences and husbandry practices in Medieval Europe.

2. Materials and Methods

In this section, we introduce the multi-allele Wright-Fisher diffusion for a single locus evolving under natural selection and gene migration. We present our Bayesian procedure for the joint inference of natural selection and gene migration from time series allele frequency data, *e.g.*, how to set up the HMM framework and how to compute the posterior probability distribution for the population genetic quantities of interest with MCMC techniques.

2.1. Wright-Fisher diffusion

We consider a population of randomly mating diploid individuals at a single locus $\mathcal{A}$ with discrete and nonoverlapping generations. Suppose that at locus $\mathcal{A}$ there are two possible allele types, labelled $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively. We attach the symbol $\mathcal{A}_1$ to the mutant allele, which is assumed to arise only once in the population and be positively selected once the evolution starts to act through natural selection. We attach the symbol $\mathcal{A}_2$ to the ancestral allele, which is assumed to originally exist in the population.

According to Loog et al. (2017), we characterise the population structure with the continent-island model (see, *e.g.*, Hamilton, 2011, for an introduction). More specifically, the population is subdivided into two demes, the continental population and the island population. To distinguish between the alleles found on the island but emigrated from the continent or were originally on the island, the mutant and ancestral alleles that originated on the island are labelled $\mathcal{A}_1^i$ and $\mathcal{A}_2^i$, respectively, and the mutant and ancestral alleles that were results of emigration from the

continent are labelled $\mathcal{A}_1^c$ and $\mathcal{A}_2^c$, respectively. We assume that the continent population is large enough such that the gene migration between the continent and the island does not influence the genetic composition of the continent population. Our interests in this work primarily focus on the island population dynamics under natural selection and gene migration. In what follows, the population refers to the island population unless noted otherwise.

2.1.1. Wright-Fisher model with selection and migration

We consider a population of N diploid individuals, with the population size N finite and constant over time. To investigate the population dynamics on the island under natural selection and gene migration, we need to specify the life cycle of the island population, which starts with zygotes that natural selection acts on. Suppose that natural selection takes the form of viability selection. The relative viabilities of all possible genotypes at locus $\mathcal{A}$ are illustrated in Table 1, where $s \in [0, 1]$ is the selection coefficient, and $h \in [0, 1]$ is the dominance parameter. After natural selection, a fraction m of the adults on the continent migrate into the population of mating adults on the island, which results in the change of the genetic composition of the island population, *i.e.*, fraction m of the adults on the island are immigrants from the continent, and fraction $1-m$ of adults were originally already on the island. Finally, there is the Wright-Fisher reproduction introduced by Fisher (1922) and Wright (1931), which corresponds to randomly sampling $2N$ gametes with replacement from an effectively infinite gamete pool to form new zygotes in the next generation through random union of gametes, consequently completing the life cycle.

	$\mathcal{A}_1^i$	$\mathcal{A}_2^i$	$\mathcal{A}_1^c$	$\mathcal{A}_2^c$
$\mathcal{A}_1^i$	1	$1-hs$	1	$1-hs$
$\mathcal{A}_2^i$	$1-hs$	$1-s$	$1-hs$	$1-s$
$\mathcal{A}_1^c$	1	$1-hs$	1	$1-hs$
$\mathcal{A}_2^c$	$1-hs$	$1-s$	$1-hs$	$1-s$

Table 1: Relative viabilities of all possible genotypes at locus $\mathcal{A}$ when we distinguish between the alleles that originate on the island and the alleles that emigrate from the continent.

Let $\boldsymbol{X}^{(N)}(k) = (X_1^{(N)}(k), X_2^{(N)}(k), X_3^{(N)}(k), X_4^{(N)}(k))$ denote frequencies of the $\mathcal{A}_1^i$, $\mathcal{A}_2^i$, $\mathcal{A}_1^c$ and $\mathcal{A}_2^c$ alleles in N zygotes of generation $k \in \mathbb{N}$ on the island. As discussed in Loog et al. (2017), the population dynamics might not be influenced by natural selection and gene migration in the early stage, thereby necessitating two additional population genetic quantities, the starting

times of natural selection and gene migration on the island, denoted by k_s and k_m, respectively. We then rewrite the selection coefficient and the migration rate as

$$s(k) = \begin{cases} 0, & \text{if } k < k_s \\ s, & \text{otherwise} \end{cases} \qquad \text{and} \qquad m(k) = \begin{cases} 0, & \text{if } k < k_m \\ m, & \text{otherwise,} \end{cases}$$

where the selection coefficient s and the migration rate m are both fixed from the timing of the onset of natural selection and gene migration up to present. Conditional on the allele frequencies of the population at generation k, we have

$$\boldsymbol{X}^{(N)}(k+1) \mid \boldsymbol{X}^{(N)}(k) = \boldsymbol{x} \sim \frac{1}{2N}\,\text{Multinomial}(2N, \boldsymbol{p}), \tag{1}$$

where

$$\begin{aligned}
p_1 &= (1-m(k))x_1 \frac{1 - hs(k)(x_2+x_4)}{1 - 2hs(k)(x_1+x_3)(x_2+x_4) - s(k)(x_2+x_4)^2} \\
p_2 &= (1-m(k))x_2 \frac{1 - hs(k)(x_1+x_3) - s(k)(x_2+x_4)}{1 - 2hs(k)(x_1+x_3)(x_2+x_4) - s(k)(x_2+x_4)^2} \\
p_3 &= (1-m(k))x_3 \frac{1 - hs(k)(x_2+x_4)}{1 - 2hs(k)(x_1+x_3)(x_2+x_4) - s(k)(x_2+x_4)^2} + m(k)x_c \\
p_4 &= (1-m(k))x_4 \frac{1 - hs(k)(x_1+x_3) - s(k)(x_2+x_4)}{1 - 2hs(k)(x_1+x_3)(x_2+x_4) - s(k)(x_2+x_4)^2} + m(k)(1-x_c)
\end{aligned} \tag{2}$$

are the allele frequencies of an effectively infinite population on the island after random mating, natural selection and gene migration from generation k to $k+1$. In Eq. (2), x_c is the frequency of the $\mathcal{A}_1^c$ allele in the continent population, which is fixed over time.

We define the multi-allele Wright-Fisher model with selection and migration to be the random process $\boldsymbol{X}^{(N)} = \{\boldsymbol{X}^{(N)}(k), k \in \mathbb{N}\}$ evolving in the state space

$$\Omega_{\boldsymbol{X}^{(N)}} = \left\{ \boldsymbol{x} \in \left\{0, \frac{1}{2N}, \ldots, 1\right\}^4 : \sum_{i=1}^{4} x_i = 1 \right\}$$

with multinomial sampling probabilities described in Eq. (2). The transition probabilities of the allele frequencies from one generation to the next depend only on the current generation, which implies that the Wright-Fisher model $\boldsymbol{X}^{(N)}$ is a Markov process.

2.1.2. Wright-Fisher diffusion with selection and migration

Due to the interplay of stochastic and deterministic forces, the Wright-Fisher model with selection and migration presents analytical challenges beyond the Wright-Fisher model for neutral populations. The analysis of the Wright-Fisher model is now greatly facilitated by its diffusion approximation, which can be traced back to Kimura (1964) and has already been successfully applied in population genetic studies (see Tataru et al., 2017, and references therein).

We assume that the selection coefficient and the migration rate are of order $1/(2N)$ once the population dynamics starts to be influenced by natural selection and gene migration, *i.e.*, $s = \mathcal{O}(1/(2N))$ and $m = \mathcal{O}(1/(2N))$. We run time at rate $2N$, *i.e.*, $t = k/(2N)$, and scale the selection coefficient and the migration rate as

$$\alpha(t) = \begin{cases} 0, & \text{if } t < t_s \\ 2Ns, & \text{otherwise} \end{cases} \qquad \text{and} \qquad \beta(t) = \begin{cases} 0, & \text{if } t < t_m \\ 4Nm, & \text{otherwise,} \end{cases}$$

where t_s and t_m denote the starting times of natural selection and gene migration on the island measured in the unites of $2N$ generations. As the population size N increases toward infinity, the Wright-Fisher model $\boldsymbol{X}^{(N)}$ converges to a diffusion process, denoted by $\boldsymbol{X} = \{\boldsymbol{X}(t), t \geq t_0\}$, evolving in the state space (*i.e.*, a three-simplex)

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

and satisfying the stochastic differential equation (SDE) of the form

$$d\boldsymbol{X}(t) = \boldsymbol{\mu}(\boldsymbol{X}(t), t)dt + \boldsymbol{\sigma}(\boldsymbol{X}(t), t)d\boldsymbol{W}(t), \qquad t \geq t_0 \tag{3}$$

with initial condition $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$. In Eq. (3), $\boldsymbol{\mu}(\boldsymbol{x}, t)$ is the drift term

$$\begin{aligned}
\mu_1(\boldsymbol{x}, t) &= \alpha(t)x_1(x_2 + x_4)\left[(x_1 + x_3)h + (x_2 + x_4)(1 - h)\right] - \frac{1}{2}\beta(t)x_1 \\
\mu_2(\boldsymbol{x}, t) &= -\alpha(t)x_2(x_1 + x_3)\left[(x_1 + x_3)h + (x_2 + x_4)(1 - h)\right] - \frac{1}{2}\beta(t)x_2 \\
\mu_3(\boldsymbol{x}, t) &= \alpha(t)x_3(x_2 + x_4)\left[(x_1 + x_3)h + (x_2 + x_4)(1 - h)\right] - \frac{1}{2}\beta(t)(x_3 - x_c) \\
\mu_4(\boldsymbol{x}, t) &= -\alpha(t)x_4(x_1 + x_3)\left[(x_1 + x_3)h + (x_2 + x_4)(1 - h)\right] - \frac{1}{2}\beta(t)(x_4 + x_c - 1),
\end{aligned} \tag{4}$$

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

$$\boldsymbol{\sigma}(\boldsymbol{x}, t) = \begin{pmatrix} \sqrt{x_1 x_2} & \sqrt{x_1 x_3} & \sqrt{x_1 x_4} & 0 & 0 & 0 \\ -\sqrt{x_2 x_1} & 0 & 0 & \sqrt{x_2 x_3} & \sqrt{x_2 x_4} & 0 \\ 0 & -\sqrt{x_3 x_1} & 0 & -\sqrt{x_3 x_2} & 0 & \sqrt{x_3 x_4} \\ 0 & 0 & -\sqrt{x_4 x_1} & 0 & -\sqrt{x_4 x_2} & -\sqrt{x_4 x_3} \end{pmatrix}, \tag{5}$$

and $\boldsymbol{W}(t)$ is a six-dimensional standard Brownian motion. Notice that the explicit formula for the diffusion term $\boldsymbol{\sigma}(\boldsymbol{x}, t)$ in Eq. (5) is obtained by following He et al. (2020a). The proof of the convergence follows in the similar manner to that employed for the neutral two-locus case in Durrett (2008, p. 323). We refer to the stochastic process $\boldsymbol{X} = \{\boldsymbol{X}(t), t \geq t_0\}$ that satisfies the SDE in Eq. (3) as the multi-allele Wright-Fisher diffusion with selection and migration.

2.2. Bayesian inference of natural selection and gene migration

Suppose that the available data are sampled from the underlying island population at time points $t_1 < t_2 < \ldots < t_K$, which are measured in units of $2N$ generations to be consistent with the Wright-Fisher diffusion timescale. At the sampling time point t_k, there are c_k mutant alleles (*i.e.*, the $\mathcal{A}_1^i$ and $\mathcal{A}_1^c$ alleles) and d_k continent alleles (*i.e.*, the $\mathcal{A}_1^c$ and $\mathcal{A}_2^c$ alleles) observed in the sample of n_k chromosomes drawn from the underlying island population. Note that in real data, the continent allele count of the sample may not be available at each sampling time point, *e.g.*, the proportion of European chicken that have Asian origin is only available in the modern sample (Loog et al., 2017). The population genetic quantities of interest are the scaled selection coefficient $\alpha = 2Ns$, the dominance parameter h, the selection time t_s, the scaled migration rate $\beta = 4Nm$, and the migration time t_m, as well as the allele frequency trajectories of the underlying island population. For simplicity, in the sequel we let $\boldsymbol{\vartheta}_s = (\alpha, h, t_s)$ be the selection-related parameters and $\boldsymbol{\vartheta}_m = (\beta, t_m)$ be the migration-related parameters, respectively.

2.2.1. Hidden Markov model

We employ a HMM framework similar to the one proposed in Bollback et al. (2008). We assume that the underlying population evolves according to the Wright-Fisher diffusion in Eq. (3) and the observations are modelled through independent sampling from the underlying population at each given time point. Unlike Loog et al. (2017), we co-estimate the timing and strength

of natural selection and gene migration, including the allele frequency trajectories of the underlying population. Our Wright-Fisher diffusion can trace the changes over time in the frequency of the allele in the island population that results from emigrants from the continent population, which allows us to make the most of other available information such as the proportion of the modern European chicken with Asian ancestry in the most recent sample reported in Loog et al. (2017). This provides valuable information regarding the timing and strength of gene migration.

We let $\boldsymbol{x}_{1:K} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K)$ denote the allele frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$. Under our HMM framework, the joint posterior probability distribution for the population genetic quantities of interest and the allele frequency trajectories of the underlying population is

$$p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K}) \propto p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m), \tag{6}$$

where $p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the prior probability distribution for the population genetic quantities of interest and can be taken to be a uniform prior over the parameter space if their prior knowledge is poor, $p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the probability distribution for the allele frequency trajectories of the underlying population at the sampling time points $\boldsymbol{t}_{1:K}$, and $p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the probability of the observations at the sampling time points $\boldsymbol{t}_{1:K}$ conditional on the underlying population allele frequency trajectories.

With the Markov property of the Wright-Fisher diffusion, we have

$$p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) = p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) \prod_{k=1}^{K-1} p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m), \tag{7}$$

where $p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the prior probability distribution for the allele frequencies of the underlying population at the initial sampling time point, commonly taken to be non-informative (*e.g.*, flat over the entire state space if the prior knowledge is poor), and $p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the transition probability density function of the Wright-Fisher diffusion between two consecutive sampling time points for $k = 1, 2, \ldots, K-1$, which satisfies the Kolmogorov backward equation (or its adjoint) resulting from the Wright-Fisher diffusion in Eq. (3). Unless otherwise specified, in this work we take the prior $p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ to be a uniform distribution over the state space $\Omega_{\boldsymbol{X}}$, known as the flat Dirichlet distribution, if gene migration starts before the first sampling

time point, *i.e.*, $t_m \leq t_1$; otherwise, the prior $p(\boldsymbol{x}_1 \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is set to be a uniform distribution over the state space $\Omega_{\boldsymbol{X}}$ restricted to the line satisfying the condition that $x_3 = x_4 = 0$, *i.e.*, there is no continent allele in the island population.

Given the allele frequency trajectories of the underlying population, the observations at each sampling time point are independent. Therefore, we have

$$p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) = \prod_{k=1}^{K} p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m), \tag{8}$$

where $p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is the probability for the observations at the k-th sampling time point given its relevant allele frequencies of the underlying population for $k = 1, 2, \ldots, K$. To calculate the emission probability $p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$, we introduce $\boldsymbol{z}_k = (z_{1,k}, z_{2,k}, z_{3,k}, z_{4,k})$ to denote the counts of the $\mathcal{A}_1^i$, $\mathcal{A}_2^i$, $\mathcal{A}_1^c$ and $\mathcal{A}_2^c$ alleles in the sample at the k-th sampling time point if the continent allele count of the sample is available at the k-th sampling time point. We then have

$$p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) = \sum_{\boldsymbol{z}_k \in \Omega_{\boldsymbol{Z}_k}} \frac{n_k!}{\prod_{i=1}^{4} z_{i,k}!} \prod_{i=1}^{4} x_{i,k}^{z_{i,k}} \mathbb{1}_{\{z_{1,k}+z_{3,k}=c_k, z_{3,k}+z_{4,k}=d_k\}}(\boldsymbol{z}_k),$$

where

$$\Omega_{\boldsymbol{Z}_k} = \left\{ \boldsymbol{z}_k \in \mathbb{N}^4 : \sum_{i=1}^{4} z_{i,k} = n_k \right\}$$

and $\mathbb{1}_A$ is the indicator function that equals to 1 if condition A holds and 0 otherwise. If the continent allele count of the sample at the k-th sampling time point is not available, the emission probability $p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ can be reduced to

$$p(c_k, d_k \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) = \frac{n_k!}{c_k!(n_k - c_k)!} (x_{1,k} + x_{3,k})^{c_k} (x_{2,k} + x_{4,k})^{n_k - c_k}.$$

2.2.2. Particle marginal Metropolis-Hastings

The most challenging part in the computation of the posterior $p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K})$ is obtaining the transition probability density function $p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k; \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ for $k = 1, 2, \ldots, K-1$. Although it can be achieved by numerically solving the Kolmogorov backward equation (or its adjoint) associated with the Wright-Fisher diffusion in Eq. (3) like Bollback et al. (2008) and He et al. (2020c), typically using a finite difference scheme, this requires a fine discretisation of the

state space $\Omega_{\boldsymbol{X}}$ to guarantee numerically stable computation of the solution. Moreover, how fine the discretisation needs to be strongly depends on the underlying population genetic quantities that we aim to estimate (He et al., 2020a). We thus resort to the PMMH algorithm developed by Andrieu et al. (2010) in this work that only involves simulating the Wright-Fisher SDE in Eq. (3), which also permits a joint update of the population genetic parameters of interests and the allele frequency trajectories of the underlying population. This is similar to what is done in He et al. (2020b). Full details of the PMMH algorithm can be found in Andrieu et al. (2010). Fearnhead & Künsch (2018) provided a detailed review of Monte Carlo methods for parameter estimation in the HMM based on the particle filter.

In our PMMH-based method, the likelihood $p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ is estimated through the bootstrap particle filter introduced by Gordon et al. (1993). In the bootstrap particle filter, the particles are generated by simulating the Wright-Fisher diffusion in Eq. (3) with the Euler-Maruyama scheme, and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}$ are sampled once from the final set of particles with their corresponding weights. We compute the product of average weights of the set of particles at the sampling time points $\boldsymbol{t}_{1:K}$, which yields an unbiased estimate of the likelihood $p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$. To our knowledge, it is challenging to find a proper proposal for the PMMH with full dimensional updates like in He et al. (2020b), which may also cause a low acceptance ratio. Since the strength of natural selection and gene migration can be strongly correlated with their timing, here we resort to a blockwise updating scheme to avoid the small acceptance ratio of the PMMH with full dimensional updates. We first partition the population genetic quantities of interest into two disjoint blocks, composed of the selection-related parameters $\boldsymbol{\vartheta}_s$ and the migration-related parameters $\boldsymbol{\vartheta}_m$, respectively. We then iteratively update one block at a time through the PMMH.

More specifically, we first generate a set of the initial candidates of the parameters $(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ from the prior $p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$. We then run a bootstrap particle filter with the proposed parameters $(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ to obtain an initial candidate of the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}$ and a bootstrap particle filter's estimate of the marginal likelihood

$$p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) = \int_{\Omega_{\boldsymbol{X}}^K} p(\boldsymbol{x}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m) p(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{x}_{1:K}, \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)\, d\boldsymbol{x}_{1:K}.$$

Repeat the following steps until a sufficient number of the samples of the parameters $(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$

and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}$ have been obtained:

Step 1: Update the selection-related parameters $\boldsymbol{\vartheta}_s$.

Step 1a: Draw a sample of new candidates of the selection-related parameters $\boldsymbol{\vartheta}_s^\star$ from the proposal $q_s(\cdot \mid \boldsymbol{\vartheta}_s)$.

Step 1b: Run a bootstrap particle filter with the parameters $(\boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)$ to yield the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^*$ and the marginal likelihood estimate $\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)$.

Step 1c: Accept the parameters $\boldsymbol{\vartheta}_s^\star$ and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^*$ with probability equal to the Metropolis-Hastings ratio

$$A = \frac{p(\boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)}{p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)} \frac{\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)}{\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)} \frac{q_s(\boldsymbol{\vartheta}_s \mid \boldsymbol{\vartheta}_s^\star)}{q_s(\boldsymbol{\vartheta}_s^\star \mid \boldsymbol{\vartheta}_s)}.$$

If the new candidates are rejected, put the parameters $\boldsymbol{\vartheta}_s^\star = \boldsymbol{\vartheta}_s$ and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^* = \boldsymbol{x}_{1:K}$.

Step 2: Update the migration-related parameters $\boldsymbol{\vartheta}_m$.

Step 2a: Draw a sample of new candidates of the migration-related parameters $\boldsymbol{\vartheta}_m^\star$ from the proposal $q_m(\cdot \mid \boldsymbol{\vartheta}_m)$.

Step 2b: Run a bootstrap particle filter with the parameters $(\boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m^\star)$ to yield the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^\star$ and the marginal likelihood estimate $\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m^\star)$.

Step 2c: Accept the parameters $\boldsymbol{\vartheta}_m^\star$ and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^\star$ with probability equal to the Metropolis-Hastings ratio

$$A = \frac{p(\boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m^\star)}{p(\boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)} \frac{\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m^\star)}{\hat{p}(\boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K} \mid \boldsymbol{\vartheta}_s^\star, \boldsymbol{\vartheta}_m)} \frac{q_m(\boldsymbol{\vartheta}_m \mid \boldsymbol{\vartheta}_m^\star)}{q_m(\boldsymbol{\vartheta}_m^\star \mid \boldsymbol{\vartheta}_m)},$$

If the new candidates are rejected, put the parameters $\boldsymbol{\vartheta}_m^\star = \boldsymbol{\vartheta}_m$ and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}^\star = \boldsymbol{x}_{1:K}^*$.

In this work we use random walk proposals for both selection- and migration-related parameters in our blockwise PMMH algorithm unless otherwise specified.

Once enough samples of the parameters $(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ and the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}$ have been yielded, we can compute the posterior $p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m \mid \boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K})$

from the samples of the parameters $(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m)$ using nonparametric density estimation techniques (see Izenman, 1991, for a detailed review) and achieve the maximum a posteriori probability (MAP) estimates for the population genetic quantities of interest. Our estimates for the underlying population allele frequency trajectories $\boldsymbol{x}_{1:K}$ are the posterior mean of the stored samples of the underlying population allele frequency trajectories. Our method can be readily extended to the analysis of multiple (independent) loci. Given that the migration-related parameters are shared by all loci, in each iteration of our procedure we only need to replicate Step 1 once to update selection-related parameters specified for each locus and then update migration-related parameters with shared by all loci in Step 2, where the likelihood is replaced by the product of the likelihoods for each locus.

In our method, it is desirable to generate a large number of particles in the bootstrap particle filter to achieve an accurate marginal likelihood estimate in the PMMH, which however leads to greater computational cost since each iteration of the PMMH requires a run of the bootstrap particle filter (see Andrieu & Roberts, 2009, for a discussion of the relationship between the accuracy of the marginal likelihood estimate and the efficiency of the pseudo-marginal method). Balancing the particle number and the MCMC iteration number to obtain good performance at a reasonable computational cost was investigated by Pitt et al. (2012) and Doucet et al. (2015). Nevertheless, the PMMH targets the posterior $p(\boldsymbol{\vartheta}_s, \boldsymbol{\vartheta}_m, \boldsymbol{x}_{1:K} \mid \boldsymbol{c}_{1:K}, \boldsymbol{d}_{1:K})$ for any number of particles. Moreover, in each run of the bootstrap particle filter, we simulate the particles from the Wright-Fisher diffusion using the Euler-Maruyama scheme between each two consecutive sampling time points. It is desirable to choose L, the number of subintervals that we divide each generation in the Euler-Maruyama scheme, large enough to obtain an accurate approximation of the Wright-Fisher diffusion. This also causes higher computational cost. Stramer & Bognar (2011) suggested choosing L to be $L^\star$ so that the marginal likelihood estimates are approximately the same for any $L > L^\star$.

3. Results

In this section, we first evaluate the performance of our approach using simulated datasets with various population genetic parameter values. We then apply it to re-analyse the time series allele frequency data from ancient chicken in Loog et al. (2017) for *TSHR*, which is hypothesised to have undergone strong and recent natural selection in domestic chicken.

3.1. Robustness and performance

To assess the performance of our method, we run forward-in-time simulations of the multi-allele Wright-Fisher model with selection and migration described in Section 2.1.1 and examine the bias and the root mean square error (RMSE) of our estimates obtained from these replicate simulations. We choose the selection coefficient $s \in \{0.003, 0.006, 0.009\}$, and adopt the selection time $k_s = 180$ and the dominance parameter $h = 0.5$ (*i.e.*, the heterozygous fitness is the arithmetic average of the homozygous fitness, called genic selection). We set the migration rate $m = 0.005$ and vary the migration time $k_m \in \{90, 360\}$. The starting times of natural selection and gene migration, k_s and k_m, respectively, are measured in generations. Moreover, we choose the population size $N \in \{5000, 50000, 500000\}$. In principle, the conclusions we arrive at in this section hold for other values of the population genetic parameters in similar ranges.

For each of the 18 possible combinations of the selection coefficient, the migration time and the population size, we perform 300 replicated runs. For each run, we take the starting allele frequencies of the underlying island population at generation 0 (*i.e.*, the first sampling time point) to be $\boldsymbol{x}_1 = (0.4, 0.6, 0, 0)$ and the mutant allele frequency of the underlying continent population to be $x_c = 0.9$. These values are similar to the allele frequencies of ancient chicken samples reported in Loog et al. (2017). We simulate a total of 500 generations according to the Wright-Fisher model in Eq. (1) and generate a multinomial sample of 100 chromosomes from the underlying island population every 50 generations from generation 0, 11 sampling time points in total. Given that in real data only mutant allele counts and continent allele counts are available, in our simulation studies we generate the mutant allele count of the sample by summing the first and third components of the simulated sample allele counts, and the continent allele count by summing the third and fourth components at each sampling time point. Moreover, considering that the continent allele count of the sample may not be available at each sampling time point in real data (*e.g.*, Loog et al., 2017), in our simulation studies we assume that continent allele counts are unavailable in the sample at first three and seven sampling time points, respectively, for each run (as shown in simulated dataset A and B, respectively, in Figure 1).

In our procedure, we choose a uniform prior over the interval $[-1, 1]$ for the selection coefficient s and a uniform prior over the interval $[0, 1]$ for the migration rate m. We assume that the starting times of natural selection and gene migration k_s and k_m are uniformly distributed

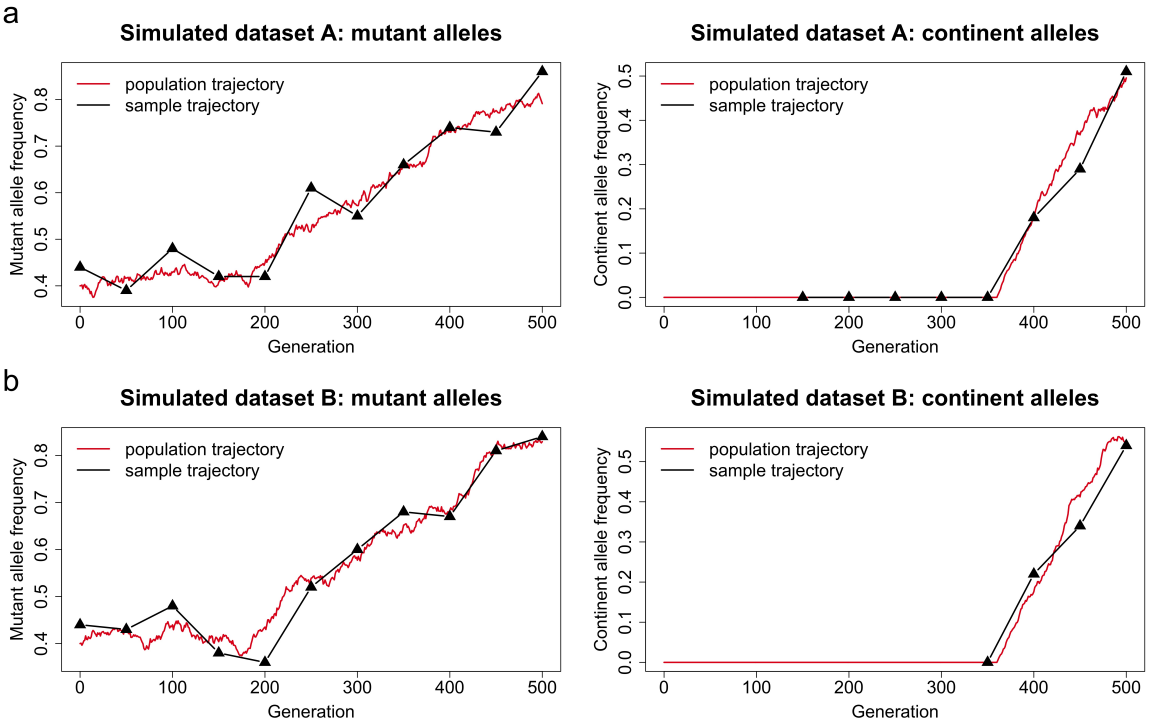

Figure 1: Representative examples of the datasets simulated using the Wright-Fisher model with selection and migration. We take the selection coefficient and time to be $s = 0.006$ and $k_s = 180$ and the migration rate and time to be $m = 0.005$ and $k_m = 360$, respectively. We set the dominance parameter $h = 0.5$ and the population size $N = 5000$. We adopt the starting allele frequencies of the underlying island population $\boldsymbol{x}_1 = (0.4, 0.6, 0, 0)$ and the mutant allele frequency of the underlying continent population $x_c = 0.9$. We sample 100 chromosomes at every 50 generations from generation 0 to 500. (a) simulated dataset A: continent allele counts are not available at the first three sampling time points. (b) simulated dataset B: continent allele counts of the sample are not available at the first seven sampling time points.

over the set of all possible time points, *i.e.*, $k_s, k_m \in \{n \in \mathbb{Z} : n \leq 500\}$. In the Euler-Maruyama scheme we divide each generation into five subintervals and run 10000 iterations of the blockwise PMMH with 1000 particles. We discard the first half of the iterations as the burn-in period, and then thin the remaining PMMH output by selecting every fifth value. See Figures 2 and 3 for the posteriors of the timing and strength of natural selection and gene migration produced with our method from the simulated datasets illustrated in Figure 1, including our estimates for the mutant and continent allele frequency trajectories of the underlying island population. Evidently our method is capable of identifying the signature of natural selection and gene migration and accurately estimate their timing and strength in these two examples. The mutant and continent allele frequency trajectories of the underlying island population are well matched with our estimates, *i.e.*, the mutant and continent allele frequency trajectories of the underlying island population fluctuate slightly around our estimates and are completely covered by our 95% highest posterior density (HPD) intervals.

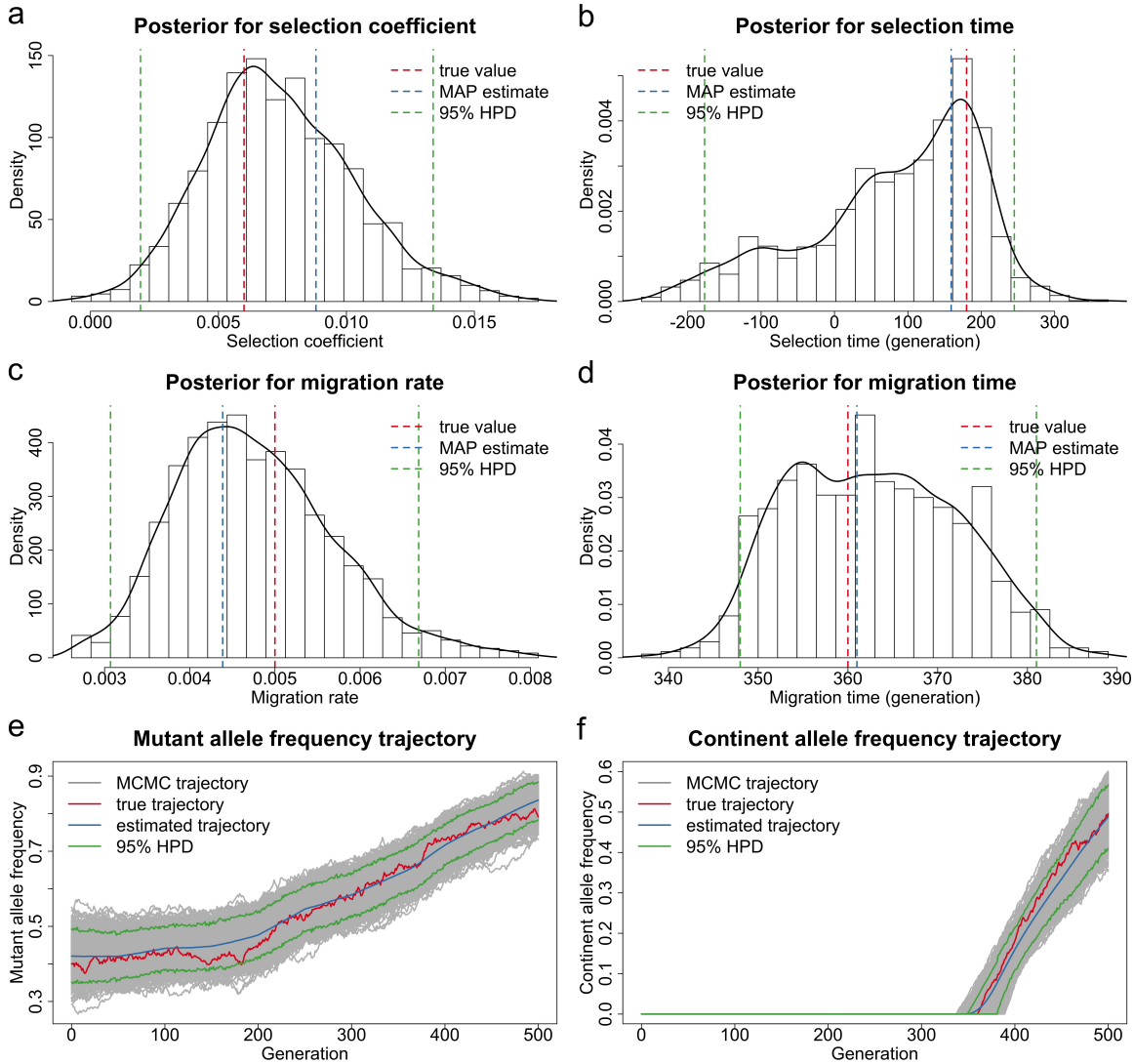

Figure 2: Bayesian estimates for the dataset shown in Figure 1a simulated for the case of the continent allele counts unavailable at the first three sampling time points. Posteriors for (a) the selection coefficient (b) the selection time (c) the migration rate and (d) the migration time. The MAP estimate is for the joint posterior, and may not correspond to the mode of the marginals. Estimated underlying trajectories of (e) the mutant allele frequency and (f) the continent allele frequency of the island population.

In Figure 4, we present the boxplots of our estimates for the scenario where continent allele counts are not available at the first three sampling time points. These boxplots illustrate the relative bias of (a) the selection coefficient estimates, (b) the selection time estimates, (c) the migration rate estimates and (d) the migration time estimates across 18 different combinations of the selection coefficient, the migration time and the population size. The tips of the whiskers represent the 2.5%-quantile and the 97.5%-quantile, and the boxes denote the first and third quartiles with the median in the middle. We also summarise the bias and the RMSE of these estimates in Supplemental Material, Tables S1 and S2.

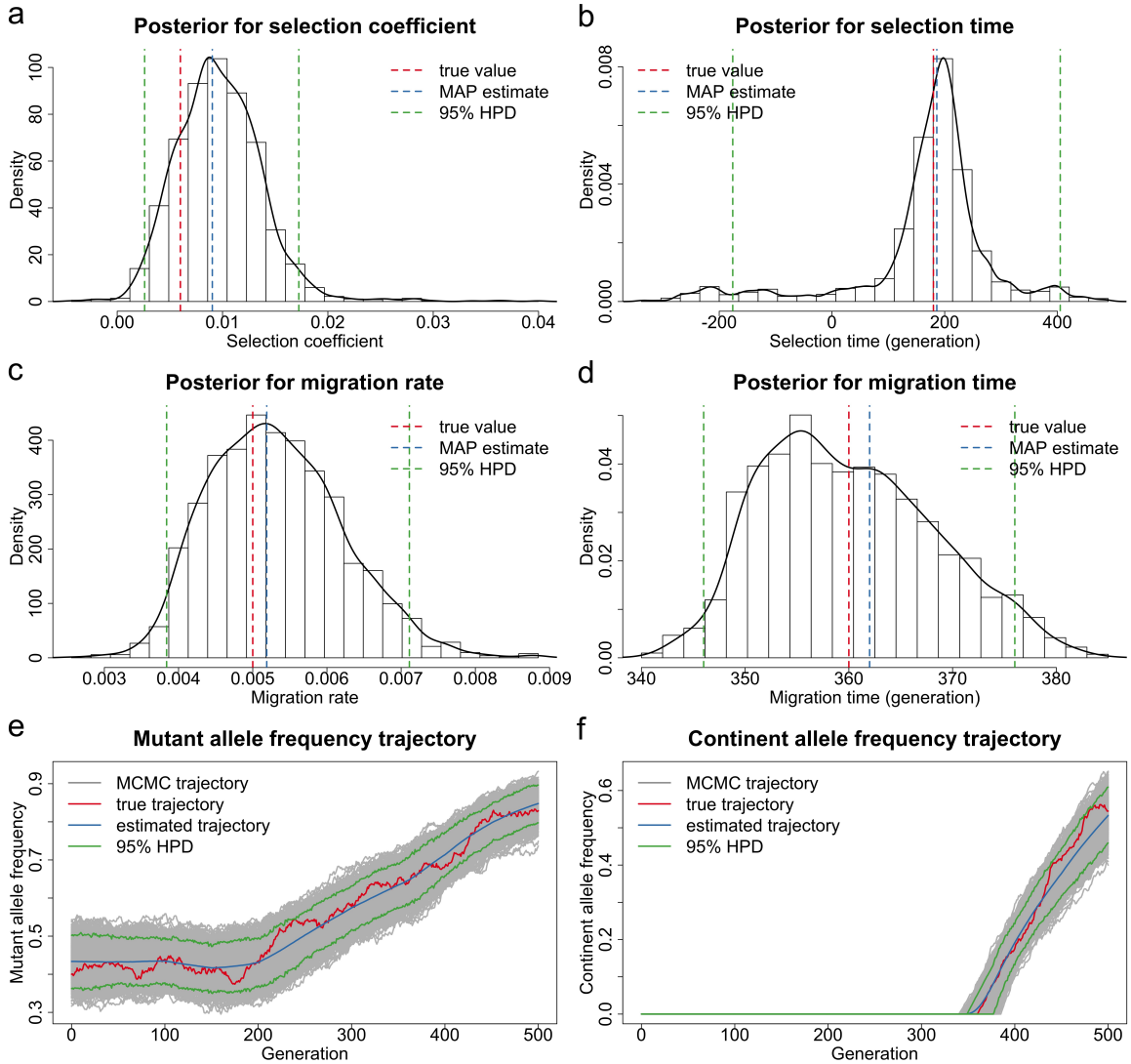

Figure 3: Bayesian estimates for the dataset shown in Figure 1b simulated for the case of continent allele counts unavailable at the first seven sampling time points. Posteriors for (a) the selection coefficient (b) the selection time (c) the migration rate and (d) the migration time. The MAP estimate is for the joint posterior, and may not correspond to the mode of the marginals. Estimated underlying trajectories of (e) the mutant allele frequency and (f) the continent allele frequency of the island population.

As shown in Figure 4, our estimates for the selection coefficient and time are approximately median-unbiased across 18 different parameter combinations, but the migration rate and time are both slightly overestimated, *i.e.*, our estimates show a small positive bias. An increase in the population size results in better overall performance of our estimates, *i.e.*, smaller bias with smaller variance. In particular, the average proportion of the replicates where the signature of natural selection can be identified (*i.e.*, the 95% HPD interval does not contain the value of 0) increases from 17.17% to 59.33% and then to 80.67% as the population size grows. Such an improvement in the performance of our estimation is to be expected as all else being equal large

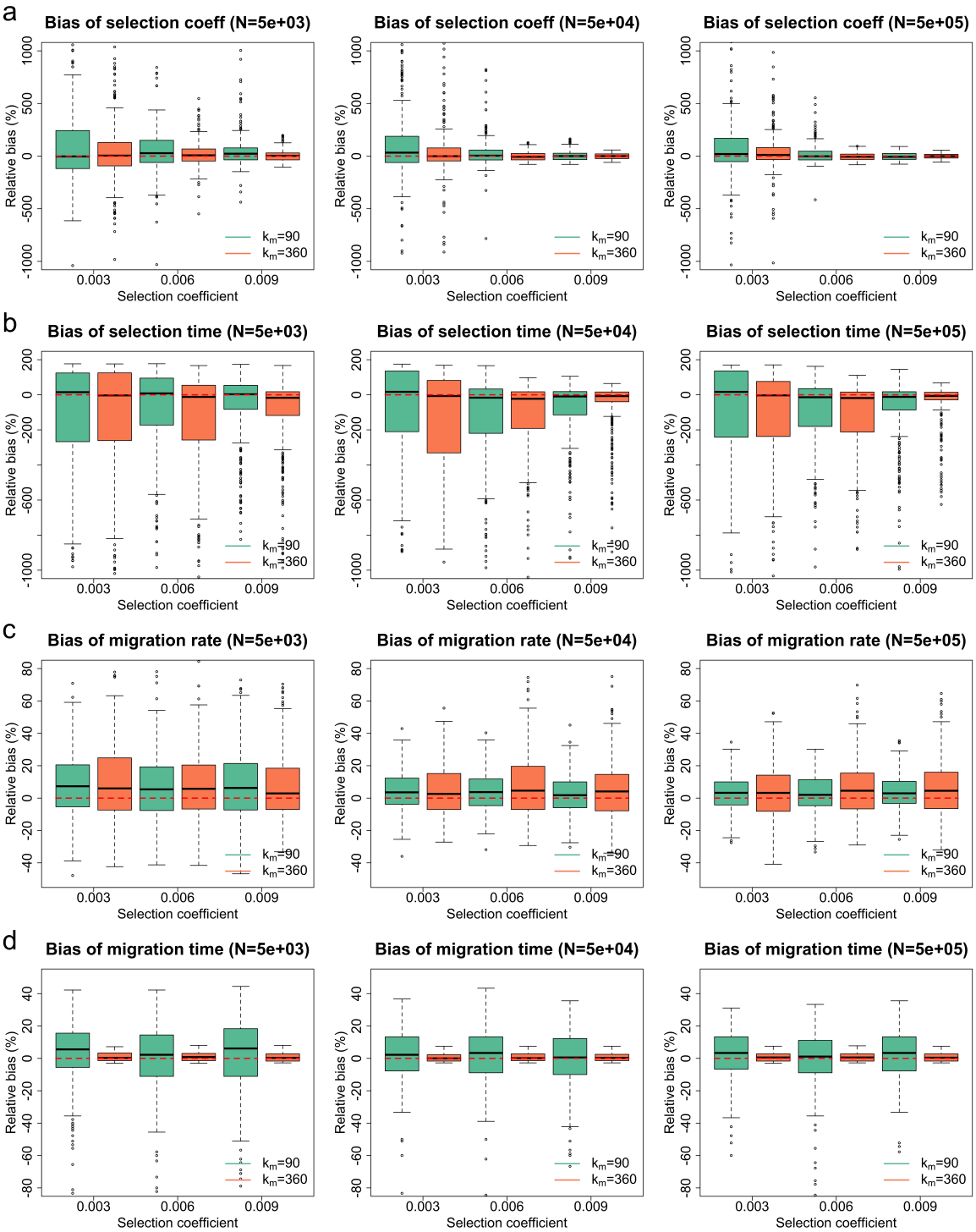

Figure 4: Empirical distributions of the estimates for 300 datasets simulated for the case of continent allele counts unavailable at the first three sampling time points. Aquamarine boxplots represent the estimates produced for the case of natural selection starting after gene migration, and coral boxplots represent the estimates produced for the case of natural selection starting before gene migration. Boxplots of the relative bias of (a) the selection coefficient estimates (b) the selection time estimates (c) the migration rate estimates and (d) the migration time estimates. To aid visual comparison, we have picked the y axes here so that boxes are of a relatively large size. This causes some outliers to lie outside the plots. Boxplots containing all outliers can be found in Supplemental Material, Figure S1.

population sizes reduce the magnitude of the stochastic effect on the changes in allele frequencies due to genetic drift, which degrades evidence of natural selection and gene migration.

Compared to the case of natural selection starting after gene migration (i.e. $k_m = 90$), our estimates for the case of natural selection starting before gene migration (i.e. $k_m = 360$) reveal smaller bias and variances in both the selection coefficient and time, with the average proportion of the replicates that the signature of natural selection can be identified increasing by 15.96%. This might be due to the fact that if natural selection starts before gene migration we have a period of time when the mutant allele frequency trajectory of the underlying population is only under natural selection. In contrast, our estimates perform better for the migration rate when natural selection starts after gene migration, but the performance of the migration time deteriorates somewhat unexpectedly. This may be caused by our parameter setting where the starting time of gene migration is within the period of availability of continent allele counts for $k_m = 360$, but not for $k_m = 90$.

In addition, we see from Figure 4 that the bias and variance of our estimates for the selection coefficient and time are largely reduced as the selection coefficient increases, especially in terms of outliers. The average proportion of the replicates where the signature of natural selection can be identified increases from 27.56% to 63.11% and then to 66.50% as the selection coefficient increases, with 97.17% for the case of large population size ($N = 500000$) and selection coefficient ($s = 0.009$). For weak natural selection, the underlying trajectory of allele frequencies is extremely stochastic so that it is difficult to disentangle the effects of genetic drift and natural selection (Schraiber et al., 2013). An increase in the strength of natural selection leads in more pronounced changes through time in allele frequencies, making the signature of natural selection more identifiable. In contrast, an increase in the selection coefficient shows little effect on our estimates of the migration rate and time.

In Figure 5, we present the boxplots of our estimates for the case where continent allele counts are not available at the first seven sampling time points, with their bias and RMSE summarised in Supplemental Material, Tables S3 and S4. They reveal similar behaviour in estimation bias and variance, although the estimates for the migration-related parameters show significantly larger bias and variances, probably due to the increased length of time when continent allele counts are unavailable. This, however, has little effect on our estimation of the selection-related

parameters, with similar average proportions of the replicates where the signature of natural selection can be identified (52.39% *vs.* 52.17%).

In conclusion, our approach can produce reasonably accurate joint estimates of the timing and strength of natural selection and gene migration from time series data of allele frequencies across different parameter combinations. Our estimates for the selection coefficient and time are approximately median-unbiased, with smaller variances as the population size or the selection coefficient (or both) increases. Our estimates for the migration rate and time both show little positive bias. Its performance improves with an increase in population size or the number of the sampling time points when continent allele counts are available (or both).

3.2. Application to ancient chicken samples

We re-analyse aDNA data of 452 European chicken genotyped at the *TSHR* locus (position 43250347 on chromosome 5) from earlier studies of Flink et al. (2014) and Loog et al. (2017). The time from which the data come ranges from approximately 2200 years ago to the present. The raw sample information and genotyping results can be found in Loog et al. (2017). The data shown in in Table 2 come from grouping the raw chicken samples from Loog et al. (2017) by their sampling time points. The derived *TSHR* allele has been associated with reduced aggression to conspecifics and faster onset of egg laying (Belyaev, 1979; Rubin et al., 2010; Karlsson et al., 2015, 2016). This mutant allele was hypothesised to have undergone strong and recent selection in domestic chicken (Rubin et al., 2010; Karlsson et al., 2015) from the time period when changes in Medieval dietary preferences and husbandry practices across northwestern Europe occurred (Loog et al., 2017).

To avoid overestimating the role of natural selection in allele frequency changes, we need to account for recent gene migration in domestic chicken from Asia to Europe in our analyses. More specifically, in our approach the European chicken population is represented as the island population while the Asian chicken population is represented as the continent population with a constant mutant allele frequency of $x_c = 0.99$. This means that the derived *TSHR* allele is nearly fixed in the Asian chicken population. In order to minimise the possibility of overestimating natural selection, we fix the frequency of the derived *TSHR* allele in the Asian chicken population at the time of the gene migration to be a conservative estimate chosen in Loog et al. (2017). Gene migration from Asia in domestic chicken, beginning around 250 years ago and continuing

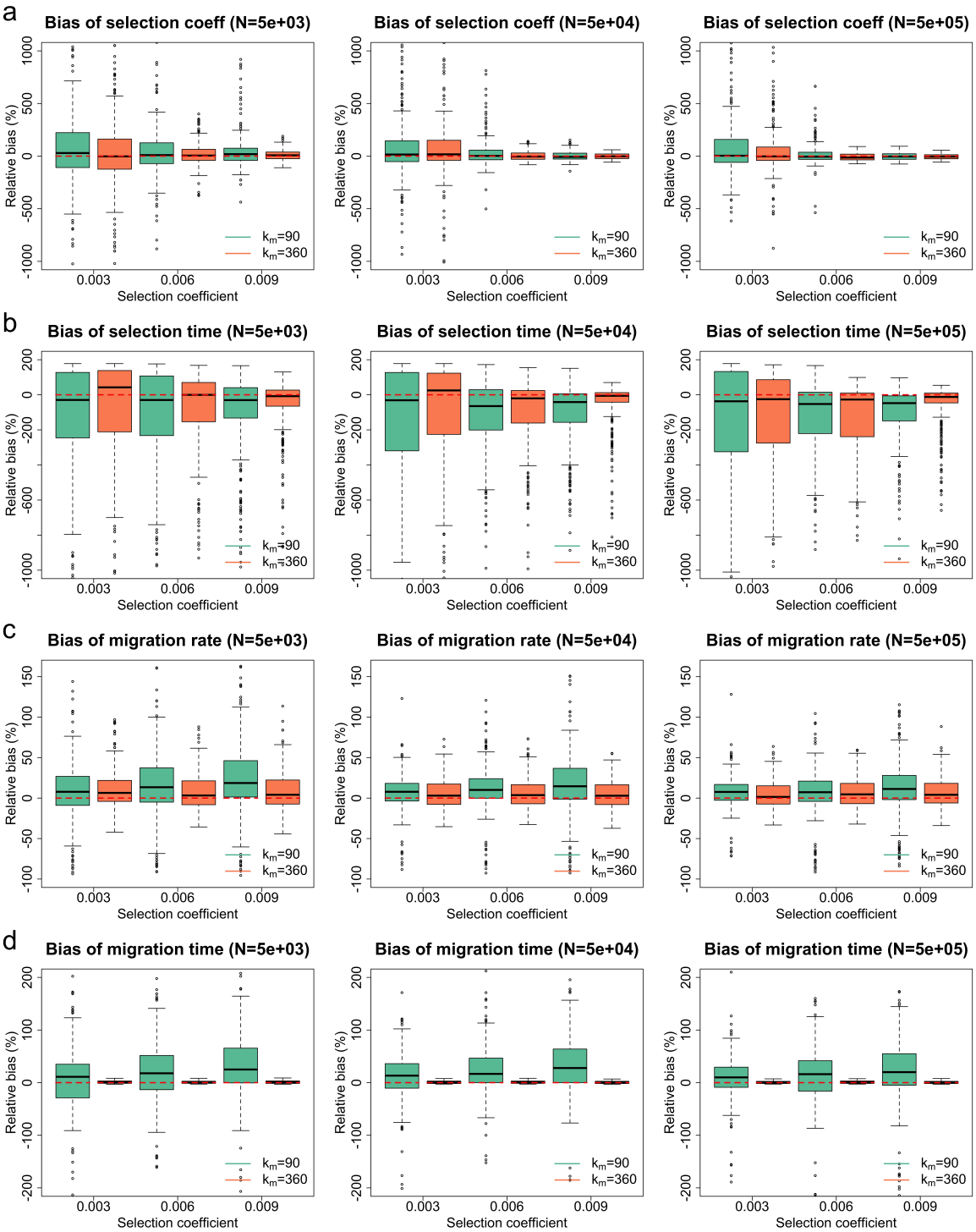

Figure 5: Empirical distributions of the estimates for 300 datasets simulated for the case of continent allele counts unavailable at the first seven sampling time points. Aquamarine boxplots represent the estimates produced for the case of natural selection starting after gene migration, and coral boxplots represent the estimates produced for the case of natural selection starting before gene migration. Boxplots of the relative bias of (a) the selection coefficient estimates (b) the selection time estimates (c) the migration rate estimates and (d) the migration time estimates. To aid visual comparison, we have picked the y axes here so that boxes are of a relatively large size. This causes some outliers to lie outside the plots. Boxplots containing all outliers can be found in Supplemental Material, Figure S2.

Sample time	Sample size	Mutant allele
-128	12	8
-25	8	5
82	8	3
200	32	14
256	14	3
1067	6	0
1309	20	18
1650	2	1
1850	2	2
1975	14	14
1995	334	328

Table 2: Time serial European chicken samples of segregating alleles at the *TSHR* locus. The unit of the sampling time is the AD year.

until the present, was historically well documented (Dana et al., 2011; Flink et al., 2014; Lyimo et al., 2015). Unlike Loog et al. (2017), we will jointly estimate the migration rate along with the selection coefficient and time by incorporating the estimate reported in Loog et al. (2017) that approximately 15% of modern European chicken have Asia origin. This allows us to obtain the sample frequency of the allele in European chicken at the most recent sampling time point that resulted from immigration from Asia. We take the average length of a generation of chicken to be one year.

In our analyses, we adopt the dominance parameter $h = 1$ since the derived *TSHR* allele is recessive, and set the population size $N = 180000$ (95% HPD 26000-460000) estimated by Loog et al. (2017). We pick a uniform prior over the interval $[-1, 1]$ for the selection coefficient s and a uniform prior over the set $\{-9000, -8999, \ldots, 0\}$ for the selection time k_s, which covers the time period of chicken domestication dated to about 8000 years ago (95% CI 7014–8768 years) (Lawal et al., 2020). We choose a uniform prior over the interval $[0, 1]$ for the migration rate m and take the migration time to be $k_m = -250$. All settings in the Euler-Maruyama scheme and the blockwise PMMH algorithm are the same as we applied in Section 3.1. The posteriors of the selection coefficient, the selection time and the migration rate are shown in Figure 6, as well as estimates for the mutant and Asian allele frequency trajectories of the underlying European population. The MAP estimates, as well as 95% HPD intervals, are summarised in Table 3.

From Table 3, we observe that our estimate of the selection coefficient for the mutant allele is 0.005120 with 95% HPD interval $[0.003591, 0.007064]$, strong evidence to support the derived

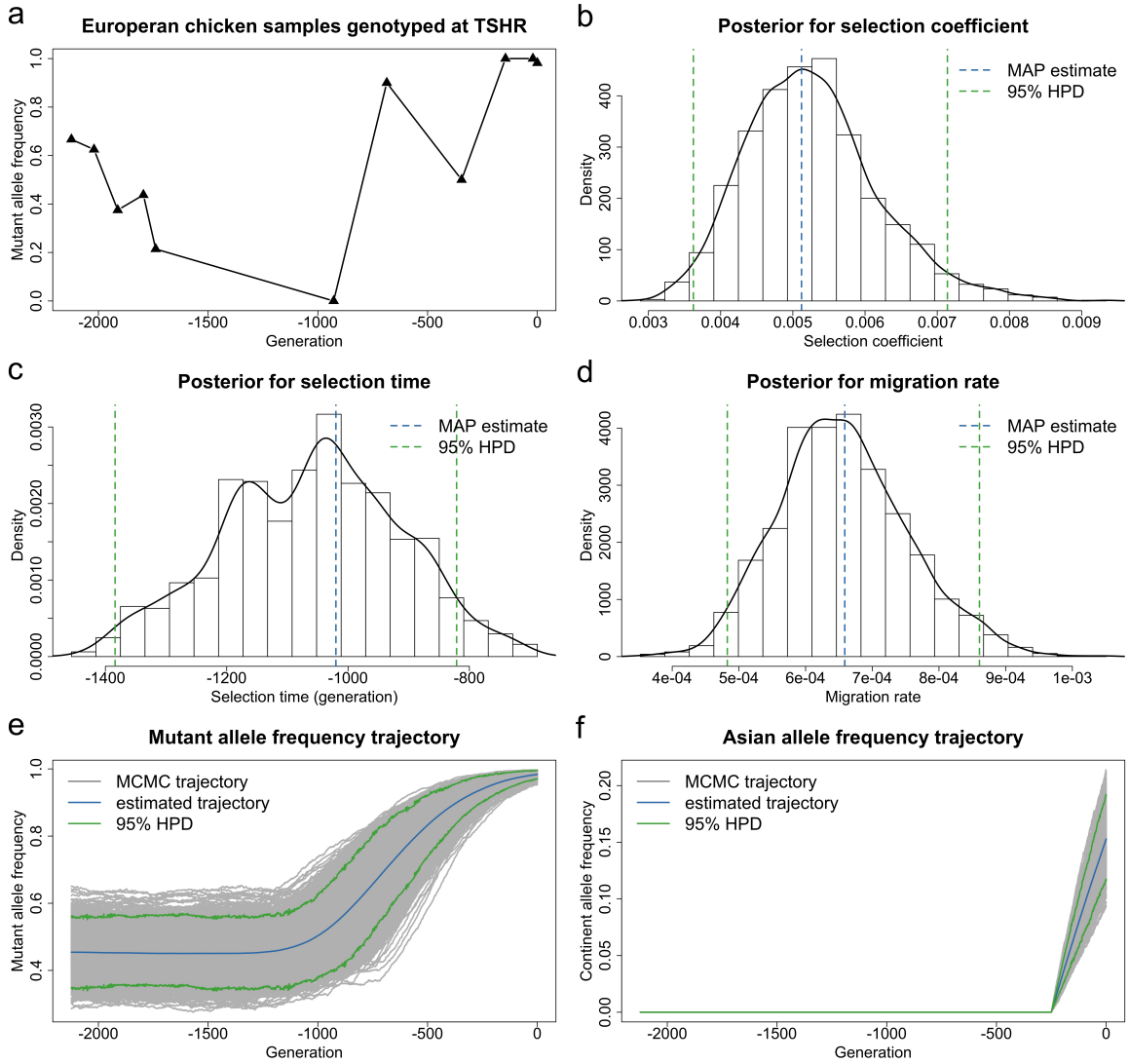

Figure 6: Bayesian estimates for aDNA data of European chicken genotyped at the *TSHR* locus from Loog et al. (2017) for the case of the population size $N = 180000$. (a) Temporal changes in the mutant allele frequencies of the sample, where the sampling time points have been offset so that the most recent sampling time point (1995 AD) is generation 0. Posteriors for (b) the selection coefficient (c) the selection time and (d) the migration rate. Estimated underlying trajectories of (e) the mutant allele frequency and (f) the Asian allele frequency in the European chicken population. The MAP estimate is for the joint posterior, and may not correspond to the mode of the marginals.

TSHR allele being selectively advantageous in the European chicken population. Such positive selection leads to an increase over time in the mutant allele frequency, starting from 975 AD with 95% HPD interval $[611, 1174]$ AD (see Figure 6e). The starting frequency of the mutant allele in 128 BC is 0.454200 with 95% HPD interval $[0.349024, 0.562094]$, which is similar to that estimated in a red junglefowl captive zoo population in Rubin et al. (2010). Our estimate of the migration rate for the Asian allele is 0.000659 with 95% HPD interval $[0.000483, 0.000861]$. This gene migration, starting about 250 years ago, leads to 15.2848% of European chicken with Asian

	Population size	MAP	95% HPD
Selection coefficient	26000	0.005109	[0.003622, 0.007141]
	180000	0.005120	[0.003591, 0.007064]
	460000	0.005122	[0.003648, 0.006578]
Selection time	26000	-1047	[-1659, -857]
	180000	-1020	[-1384, -821]
	460000	-1047	[-1327, -893]
Migration rate	26000	0.000712	[0.000448, 0.000918]
	180000	0.000659	[0.000483, 0.000861]
	460000	0.000620	[0.000478, 0.000837]

Table 3: MAP estimates of the selection coefficient, the selection time and the migration rate, as well as their 95% HPD intervals, for *TSHR* achieved with the population size $N = 26000$, $N = 180000$ and $N = 460000$.

ancestry in 1995 AD, with 95% HPD interval $[0.116412, 0.191382]$ (see Figure 6f). Our findings are consistent with those reported in Loog et al. (2017). This is further confirmed by the results obtained with different values of the population size (*i.e.*, $N = 26000$ and $N = 460000$, the lower and upper bounds of 95% HPD interval for the population size in Loog et al. (2017)). These results are shown in Supplemental Material Figures S3 and S4 and summarised in Table 3.

To further evaluate the performance of our approach when samples are sparsely distributed with small uneven sizes, such as the European chicken samples at the *TSHR* locus we have studied above, we generate 300 simulated datasets that mimic the *TSHR* data, *i.e.*, we use the sample times and sizes as given in Table 2, the timing and strength of natural selection and gene migration as given by MAP estimates found in Table 3, and population size $N = 180000$. From Figure 7, we find that our simulation studies based on the *TSHR* data yield median-unbiased estimates for the selection coefficient, the selection time and the migration rate, similar to the performance of our procedure applied in the simulation studies in Section 3.1. Additionally, the signature of natural selection can be identified in all these 300 replicates. This further shows that our method can achieve a good performance with time serial samples that are sparsely distributed with small uneven sizes, which is highly desirable for aDNA data.

In summary, our approach works well on the ancient chicken samples, even though they are sparsely distributed with small uneven sizes. Our estimates show strong evidence for the derived *TSHR* allele being positively selected between the 7th and 12th centuries AD, which coincides exactly with the time period of changes in dietary preferences and husbandry practices across northwestern Europe. This again demonstrates possible links established by Loog et al. (2017) between the selective advantage of the derived *TSHR* allele and a historically attested cultural

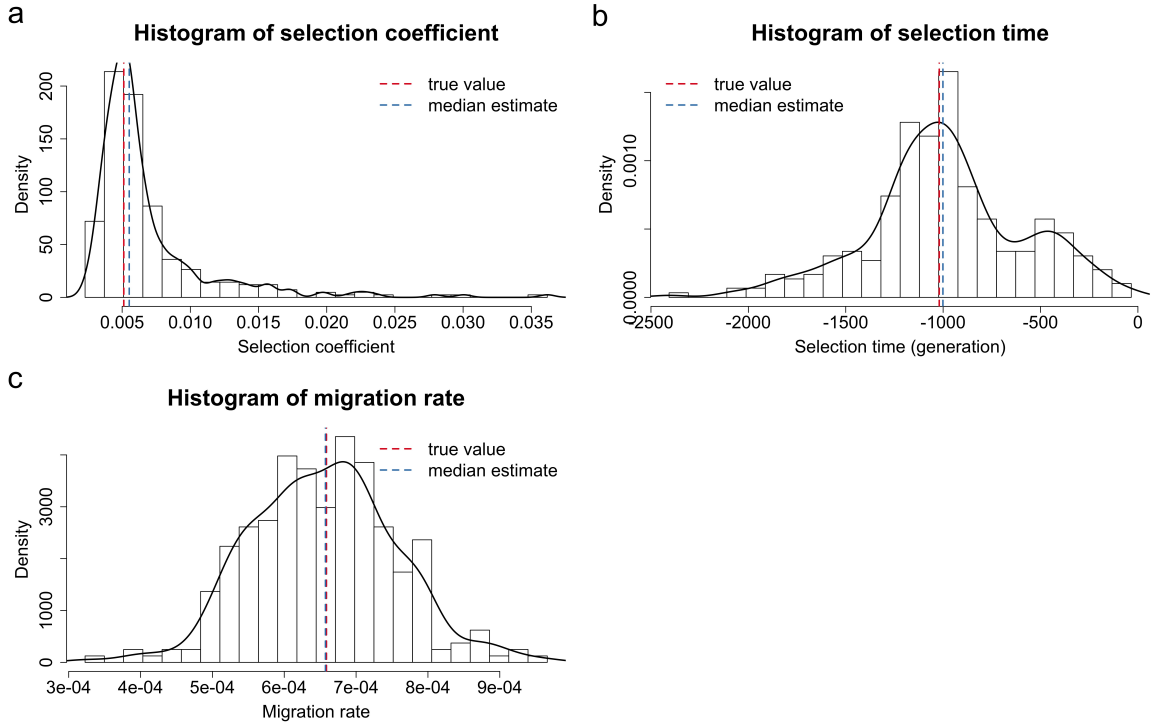

Figure 7: Empirical distributions of the estimates for 300 datasets simulated for *TSHR* based on the aDNA data shown in Table 2. We simulate the underlying population dynamics with the timing and strength of natural selection and gene migration estimated with the population size $N = 180000$ shown in Table 3. Histograms of (a) the selection coefficient estimates (b) the selection time estimates and (c) the migration rate estimates. To aid visual comparison, we have picked the x axis in (a) not to cover all 300 estimates. The histogram containing all 300 estimates can be found in Supplemental Material, Figure S5.

shift in food preference in Medieval Europe.

4. Discussion

In this work, we introduced a novel MCMC-based approach for the joint inference of the timing and strength of natural selection and gene migration from aDNA data. To our knowledge, Mathieson & McVean (2013) and Loog et al. (2017) described the only existing approaches that can jointly infer natural selection and gene migration from time series data of allele frequencies. However, the method of Mathieson & McVean (2013) is unable to estimate the time of the onset of natural selection and gene migration. Loog et al. (2017) only showed the applicability of their method in the case where timing and strength of gene migration were both given. In addition, their approach is restricted by the assumption of infinite population sizes. These undesirable features highly limit the application of their method to aDNA data.

Our Bayesian inference procedure was built on an HMM framework incorporating a multi-allele Wright-Fisher diffusion with selection and migration. Our estimates for the timing and

strength of natural selection and gene migration were achieved through the PMMH algorithm with blockwise sampling, which enables joint estimation of the allele frequency trajectories of the underlying population as well. This is a highly desirable feature for aDNA because it allows us to infer the drivers of natural selection and gene migration by correlating patterns of genetic variation with potential evolutionary events such as changes in the ecological context in which an organism has evolved.

We showed through extensive simulation studies that our approach could deliver reasonably accurate estimates for the timing and strength of natural selection and gene migration, including the estimates for allele frequency trajectories of the underlying population. In our simulation studies, the estimates for the selection rate and time were largely unbiased, while the estimates for the migration rate and time showed a slight positive bias. We applied our Bayesian inference procedure to re-analyse ancient European chicken samples genotyped at the *TSHR* locus from previous studies of Flink et al. (2014) and Loog et al. (2017). We found that the derived *TSHR* allele became selectively advantageous from 975 AD (95% HPD 611-1174 AD), which was similar to that reported in Loog et al. (2017). Our results further confirmed the findings of Loog et al. (2017) that positive selection acting on the *TSHR* locus in European chicken could be driven by chicken intensification and egg production in Medieval Europe as a result of Christian fasting practices (*i.e.*, the consumption of birds, eggs and fish became allowed (Venarde, 2011)). Except for religiously inspired dietary preferences, this could also be caused by changes in Medieval husbandry practices along with population growth and urbanisation in the High Middle Ages (around 1000-1250 AD). See Loog et al. (2017) and references cited therein for more details.

Unlike Loog et al. (2017), our approach take the influence of genetic drift into account. From Table 3, we see that our estimates from aDNA data for *TSHR* are close to each other regardless of what population size we pick from the 95% HPD interval for European chicken population size reported in Loog et al. (2017). This implies that ignoring genetic drift might have little effect on the inference of natural selection from aDNA data such as those in Loog et al. (2017). To explore the effect of genetic drift, we further simulate 300 datasets based on the aDNA data for *TSHR*, where the timing and strength of natural selection and gene migration are set to our estimates found in Table 3 but the true population size is picked to be $N = 4500$. We run our method with an incorrect population size $N = 180000$ for these 300 replicates and find that this

incorrect (larger) population size results in significant overestimation of the selection coefficient and time (see Figure 8). Moreover, the misspecified population size causes a reduction of 9.67% in the proportion of the replicates where the signature of natural selection can be identified. This implies the necessity of modelling genetic drift when we estimate the timing and strength of natural selection from aDNA data to avoid possible bias.

Furthermore, we explore the effect of gene migration in a similar manner. We simulate 300 datasets based on the aDNA data for *TSHR* with the migration rate $m = 0.00001$ and $m = 0.01$, respectively, but run our procedure with a misspecified migration rate $m = 0.000659$ (*i.e.*, the migration rate estimated with the population size $N = 180000$ shown in Table 3). We can find from Figure 9 that an incorrect migration rate does not dramatically alter the posterior median of the selection-related parameters, but can change the overall shape of the posterior surface for the selection time. This is possibly due to incorrect information used for gene migration also causing incorrect information for natural selection, therefore disturbing the resulting posterior surface.

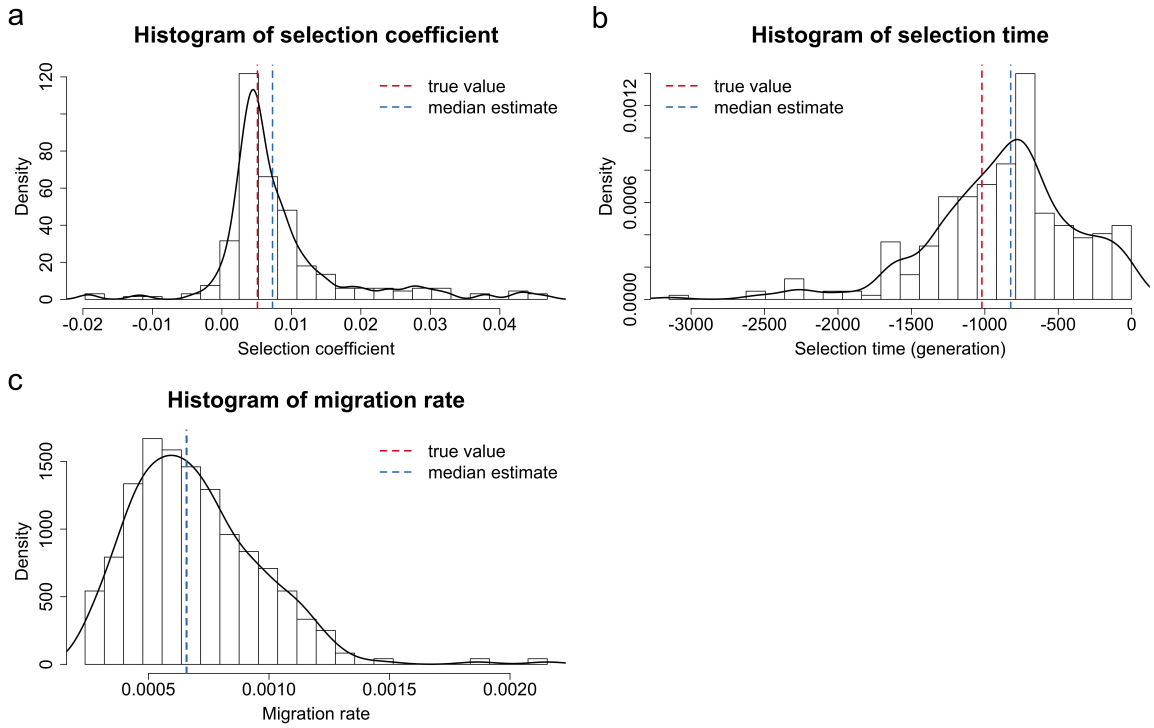

Figure 8: Empirical distributions of the estimates for 300 datasets simulated for *TSHR* based on the aDNA data shown in Table 2. We take the timing and strength of natural selection and gene migration to be those estimated with the population size $N = 180000$ shown in Table 3, but the true population size in the simulation is taken to be $N = 4500$. Histograms of (a) the selection coefficient estimates (b) the selection time estimates and (c) the migration rate estimates. To aid visual comparison, we have picked the x axis in (a) not to cover all 300 estimates. The histogram containing all 300 estimates can be found in Supplemental Material, Figure S6.

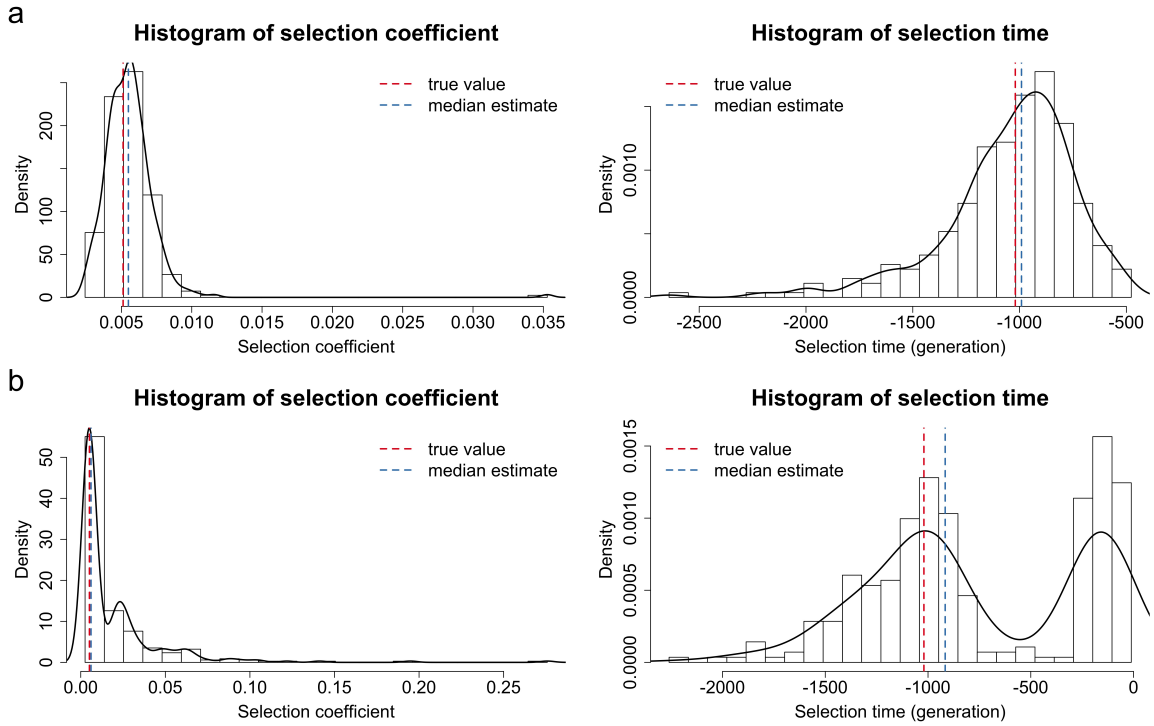

Figure 9: Empirical distributions of the estimates for 300 datasets simulated for *TSHR* based on the aDNA data shown in Table 2. We take the timing and strength of natural selection and gene migration to be those estimated with the population size $N = 180000$ shown in Table 3, but the true migration rate in the simulation is taken to be (a) $m = 0.0001$ and (b) $m = 0.001$. Histograms of the selection coefficient estimates and the selection time estimates for the migration rate (a) $m = 0.0001$ and (b) $m = 0.001$.

Although we have focused on the continent-island model in this work, our Bayesian statistical framework lends itself to being extended to more complicated models of gene migration, *e.g.*, multiple islands. With an increase in the number of demes, our method will be more computationally demanding, but improvements to exact-approximate particle filtering techniques like the PMMH continue to be developed. See *e.g.* Yıldırım et al. (2018). Our approach is also readily applicable to the analysis of multiple (independent) loci, where we observe that updating the selection-related parameters for each locus can be proceeded in parallel on different cores. Additionally, it is possible to extend our procedure to handle the case of non-constant demographic histories like Schraiber et al. (2016) and He et al. (2020c). In order to achieve accurate estimation of relevant population genetic quantities of interest, it is important to taking into account local linkage among loci, which has been illustrated to be capable of further improving the inference of natural selection (He et al., 2020b). Our Bayesian statistical framework can be readily extended to the scenario of two linked loci by incorporating the approach of He et al. (2020b) but such an extension will probably be computationally prohibitive in the case of mul-

tiple linked loci. As a tractable alternative for multiple linked loci, we can apply our two-locus method in a pairwise manner by adding additional blocks in blockwise sampling.

References

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

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

Belyaev, D. K. (1979). Destabilizing selection as a factor in domestication. *Journal of Heredity*, *70*, 301–308.

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

Dana, N., Megens, H.-J., Crooijmans, R. P. M. A., Hanotte, O., Mwacharo, J. et al. (2011). East Asian contributions to Dutch traditional and western commercial chickens inferred from mtDNA analysis. *Animal Genetics*, *42*, 125–133.

Dehasque, M., Ávila-Arcos, M. C., Díez-del Molino, D., Fumagalli, M., Guschanski, K. et al. (2020). Inference of natural selection from ancient DNA. *Evolution Letters*, *4*, 94–108.

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

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

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

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

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

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

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

Hamilton, M. (2011). *Population Genetics*. Chichester: Wiley-Blackwell.

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

He, Z., Dai, X., Beaumont, M. A., & Yu, F. (2020b). Detecting and quantifying natural selection at two linked loci from time series data of allele frequencies with forward-in-time simulations. *Genetics*, *216*, 521–541.

He, Z., Dai, X., Beaumont, M. A., & Yu, F. (2020c). Estimation of natural selection and allele age from time series allele frequency data using a novel likelihood-based approach. *Genetics*, *216*, 463–480.

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

Karlsson, A.-C., Fallahshahroudi, A., Johnsen, H., Hagenblad, J., Wright, D. et al. (2016). A domestication related mutation in the thyroid stimulating hormone receptor gene (*TSHR*) modulates photoperiodic response and reproduction in chickens. *General and Comparative Endocrinology*, *228*, 69–78.

Karlsson, A.-C., Svemer, F., Eriksson, J., Darras, V. M., Andersson, L. et al. (2015). The effect of a mutation in the thyroid stimulating hormone receptor (TSHR) on development, behaviour and TH levels in domesticated chickens. *PLoS One*, *10*, e0129040.

Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability*, *1*, 177–232.

Lawal, R. A., Martin, S. H., Vanmechelen, K., Vereijken, A., Silva, P. et al. (2020). The wild species genome ancestry of domestic chickens. *BMC Biology*, *18*, 1–18.

Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H. et al. (2017). Evolutionary patterns and processes: lessons from ancient DNA. *Systematic Biology*, *66*, e1–e29.

Loog, L., Thomas, M. G., Barnett, R., Allen, R., Sykes, N. et al. (2017). Inferring allele frequency trajectories from ancient DNA indicates that selection on a chicken gene coincided with changes in medieval husbandry practices. *Molecular Biology and Evolution*, *34*, 1981–1990.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A. et al. (2009). Coat color variation at the beginning of horse domestication. *Science*, *324*, 485–485.

Lyimo, C. M., Weigend, A., Msoffe, P. L., Hocking, P. M., Simianer, H. et al. (2015). Maternal genealogical patterns of chicken breeds sampled in Europe. *Animal Genetics*, *46*, 447–451.

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

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

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N. et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*, 499–503.

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

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

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E. et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*, *464*, 587–591.

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

Schraiber, J. G., Griffiths, R. C., & Evans, S. N. (2013). Analysis and rejection sampling of Wright-Fisher diffusion bridges. *Theoretical Population Biology*, *89*, 64–74.

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

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

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

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

Venarde, B. L. (2011). *The Rule of Saint Benedict*. Cambridge, Massachusetts: Harvard University Press.

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

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

Data Accessibility Statement

The authors state that all data necessary for confirming the conclusions of this work are represented fully within the article. Source code implementing the method described in this work is available at `https://github.com/zhangyi-he/WFM-1L-DiffusApprox-PMMH-Chicken`.

Author Contributions

F.Y. and Z.H. designed the project and developed the method; W.L. and Z.H. implemented the method; W.L. and X.D. analysed the data under the supervision of M.B., F.Y. and Z.H.; W.L., X.D. and Z.H. wrote the manuscript; M.B. and F.Y. reviewed the manuscript.