BAYESIAN INFERENCE OF NATURAL SELECTION FROM ALLELE FREQUENCY TIME SERIES

JOSHUA G. SCHRAIBER, STEVEN N. EVANS, AND MONTGOMERY SLATKIN

Abstract. The advent of accessible ancient DNA technology now allows the direct ascertainment of allele frequencies in ancestral populations, thereby enabling the use of allele frequency time series to detect and estimate natural selection. Such direct observations of allele frequency dynamics are expected to be more powerful than inferences made using patterns of linked neutral variation obtained from modern individuals. We developed a Bayesian method to make use of allele frequency time series data and infer the parameters of general diploid selection, along with allele age, in non-equilibrium populations. We introduce a novel path augmentation approach, in which we use Markov chain Monte Carlo to integrate over the space of allele frequency trajectories consistent with the observed data. Using simulations, we show that this approach has good power to estimate selection coefficients and allele age. Moreover, when applying our approach to data on horse coat color, we find that ignoring a relevant demographic history can significantly bias the results of inference. Our approach is made available in a C++ software package.

1. Introduction

The ability to obtain high-quality genetic data from ancient samples is revolutionizing the way that we understand the evolutionary history of populations. One of the most powerful applications of ancient DNA (aDNA) is to study the action of natural selection. While methods making use of only modern DNA sequences have successfully identified loci evolving subject to natural selection [Nielsen et al., 2005, Voight et al., 2006, Pickrell et al., 2009], they are inherently limited because they look indirectly for selection, finding its signature in nearby neutral variation. In contrast, by sequencing ancient individuals, it is possible to directly track the change in allele frequency that is characteristic of the action of natural selection. This approach has been exploited recently using whole genome data to identify candidate loci under selection in European humans [Mathieson et al., 2015].

To infer the action of natural selection rigorously, several methods have been developed to explicitly fit a population genetic model to a time series of allele frequencies obtained via aDNA. Initially, Bollback et al. [2008] extended an approach devised by Williamson and Slatkin [1999] to estimate the population-scaled selection coefficient, $\alpha = 2N_e s$, along with the effective size, N_e. To incorporate natural selection, Bollback et al. [2008] used the continuous diffusion approximation to the discrete Wright-Fisher model. This required

Date: Started on December 10, 2013. Compiled on January 18, 2016.

JGS supported by NSF grant DBI-1402120, SNE supported in part by NSF grant DMS-0907630, NSF grant DMS-1512933, and NIH grant 1R01GM109454-01, MS supported by NIH grant R01-GM40282.

them to use numerical techniques to solve the partial differential equation (PDE) associated with transition densities of the diffusion approximation to calculate the probabilities of the population allele frequencies at each time point. Ludwig et al. [2009] obtained an aDNA time series from 6 coat-color-related loci in horses and applied the method of Bollback et al. [2008] to find that 2 of them, ASIP and MC1R, showed evidence of strong positive selection.

Recently, a number of methods have been proposed to extend the generality of the Bollback et al. [2008] framework. To define the hidden Markov model they use, Bollback et al. [2008] were required to posit a prior distribution on the allele frequency at the first time point. They chose to use a uniform prior on the initial frequency; however, in truth the initial allele frequency is dictated by the fact that the allele at some point arose as a new mutation. Using this information, Malaspinas et al. [2012] developed a method that also infers allele age. They also extended the selection model of Bollback et al. [2008] to include fully recessive fitness effects. A more general selective model was implemented by Steinrücken et al. [2014], who model general diploid selection, and hence they are able to fit data where selection acts in an over- or under-dominant fashion; however, Steinrücken et al. [2014] assumed a model with recurrent mutation and hence could not estimate allele age. The work of Mathieson and McVean [2013] is designed for inference of metapopulations over short time scales and so it is computationally feasible for them to use a discrete time, finite population Wright-Fisher model. Finally, the approach of Feder et al. [2014] is ideally suited to experimental evolution studies because they work in a strong selection, weak drift limit that is common in evolving microbial populations.

One key way that these methods differ from each other is in how they compute the probability of the underlying allele frequency changes. For instance, Malaspinas et al. [2012] approximated the diffusion with a birth-death type Markov chain, while Steinrücken et al. [2014] approximate the likelihood analytically using a spectral representation of the diffusion discovered by Song and Steinrücken [2012]. These different computational strategies are necessary because of the inherent difficulty in solving the Wright-Fisher partial differential equation. A different approach, used by Mathieson and McVean [2013] in the context of a densely-sampled discrete Wright-Fisher model, is to instead compute the probability of the entire allele frequency trajectory in between sampling times.

In this work, we develop a novel approach for inference of general diploid selection and allele age from allele frequency time series obtained from aDNA. The key innovation of our approach is that we impute the allele frequency trajectory between sampled points when they are sparsely-sampled. Moreover, by working with a diffusion approximation, we are able to easily incorporate general diploid selection and changing population size. This approach to inferring parameters from a sparsely-sampled diffusion is known as high-frequency path augmentation, and has been successfully applied in a number of contexts [Roberts and Stramer, 2001, Golightly and Wilkinson, 2005, 2008, Sørensen, 2009, Fuchs, 2013]. The diffusion approximation to the Wright-Fisher model, however, has several features that are atypical in the context of high-frequency path augmentation, including a time-dependent diffusion coefficient and a bounded state-space. We then apply this new

method to several datasets and find that we have power to estimate parameters of interest from real data.

2. Model and Methods

2.1. **Generative model.** We assume a randomly mating diploid population that is size $N(t)$ at time t, where t is measured in units of $2N_0$ generations for some arbitrary, constant N_0. At the locus of interest, the ancestral allele, A_0, was fixed until some time t_0 when the derived allele, A_1, arose with diploid fitnesses as given in Table 1.

[Table 1 about here.]

Given that an allele is segregating at a population frequency $0 < x_* < 1$ at some time $t_* > t_0$, the trajectory of population frequencies of A_1 at times $t \geq t_*$, $(X_t)_{t\geq t_*}$, is modeled by the usual diffusion approximation to the Wright-Fisher model (and many other models such as the Moran model), which we will henceforth call the Wright-Fisher diffusion. While many treatments of the Wright-Fisher diffusion define it in terms of the partial differential equation that characterizes its transition densities (e.g. Ewens [2004]), we instead describe it as the solution of a stochastic differential equation (SDE). Specifically, $(X_t)_{t\geq t_*}$ satisfies the SDE

$$\begin{aligned} dX_t &= X_t(1-X_t)(\alpha_1(2X_t-1)-\alpha_2 X_t)\,dt + \sqrt{\frac{X_t(1-X_t)}{\rho(t)}}\,dB_t \\ X_{t_*} &= x_*, \end{aligned} \tag{1}$$

where B is a standard Brownian motion, $\alpha_1 = 2N_0 s_1$, $\alpha_2 = 2N_0 s_2$, and $\rho(t) = N(t)/N_0$. If $X_{t_{**}} = 0$ (resp. $X_{t_{**}} = 1$) at some time $t_{**} > t_*$, then $X_t = 0$ (resp. $X_t = 1$) for all $t \geq t_{**}$.

In order to make this description of the dynamics of the population allele frequency trajectory $(X_t)_{t\geq t_0}$ complete, we need to specify an initial condition at time t_0. In a finite population Wright-Fisher model we would take the allele A_1 to have frequency $\frac{1}{2N(t_0)}$ at the time t_0 when it first arose in a single chromosome. This frequency converges to 0 when we pass to the diffusion limit, but we cannot start the Wright-Fisher diffusion at 0 at time t_0 because the diffusion started at 0 remains at 0. Instead, we take the value of X_{t_0} to be some small, but arbitrary, frequency x_0. This arbitrariness in the choice of x_0 may seem unsatisfactory, but we will see that the resulting posterior distribution for the parameters α_1, α_2, t_0 converges as $x_0 \downarrow 0$ to a limit which can be thought of as the posterior corresponding to a certain improper prior distribution, and so, in the end, there is actually no need to specify x_0.

Finally, we model the data assuming that at known times $t_1, t_2, \ldots, t_k$ samples of known sizes $n_1, n_2, \ldots, n_k$ chromosomes are taken and $c_1, c_2, \ldots, c_k$ copies of the derived allele are found at the successive time points (Figure 1). Note that it is possible that some of the sampling times are more ancient than t_0, the age of the allele.

[Figure 1 about here.]

2.2. **Bayesian path augmentation.** We are interested in devising a Bayesian method to obtain the posterior distribution on the parameters, α_1, α_2, and t_0 given the sampled allele frequencies and sample times – data which we denote collectively as D. Because we are dealing with objects that don't necessarily have distributions which have densities with respect to canonical reference measures, it will be convenient in the beginning to treat priors and posteriors as probability measures rather than as density functions. For example, the posterior is the probability measure

$$P(d\alpha_1, d\alpha_2, dt_0 \,|\, D) = \frac{P(dD \,|\, \alpha_1, \alpha_2, t_0)\, \pi(d\alpha_1, d\alpha_2, dt_0)}{P(dD)}, \tag{2}$$

where π is a joint prior on the model parameters. However, computing the likelihood $P(dD \,|\, \alpha_1, \alpha_2, t_0)$ is computationally challenging because, implicitly,

$$P(dD \,|\, \alpha_1, \alpha_2, t_0) = \int P(dD \,|\, X)\, P(dX \,|\, \alpha_1, \alpha_2, t_0),$$

where the integral is over the (unobserved, infinite-dimensional) allele frequency path $X = (X_t)_{t \geq t_0}$, $P(\cdot \,|\, \alpha_1, \alpha_2, t_0)$ is the distribution of a Wright-Fisher diffusion with selection parameters α_1, α_2 started at time t_0 at the small but arbitrary frequency x_0, and

$$P(dD \,|\, X) = \prod_{i=1}^{k} \binom{n_i}{c_i} X_{t_i}^{c_i} (1 - X_{t_i})^{n_i - c_i}$$

because we assume that sampled allele frequencies at the times $t_1, \ldots, t_k$ are independent binomial draws governed by underlying population allele frequencies at the these times. Integrating over the infinite-dimensional path $(X_t)_{t \geq t_0}$ involves either solving partial differential equations numerically or using Monte Carlo methods to find the joint distribution of population allele frequency path at the times $t_1, \ldots, t_k$.

To address this computational difficulty, we introduce a path augmentation method that treats the underlying allele frequency path $(X_t)_{t \geq t_0}$ as an additional parameter. Observe that the posterior may be expanded out to

$$P(d\alpha_1, d\alpha_2, dt_0 \,|\, D) = \frac{\int P(dD \,|\, X')\, P(dX' \,|\, \alpha_1, \alpha_2, t_0) \pi(d\alpha_1, d\alpha_2, dt_0)}{\int P(dD \,|\, X')\, P(dX' \,|\, \alpha_1', \alpha_2', t_0') \pi(d\alpha_1', d\alpha_2', dt_0')},$$

where we use primes to designate dummy variables over which we integrate. Thinking of the path $(X_t)_{t \geq t_0}$ as another parameter and taking the prior distribution for the augmented family of parameters to be

$$P(dX \,|\, \alpha_1, \alpha_2, t_0) \pi(d\alpha_1, d\alpha_2, dt_0),$$

the posterior for the augmented family of parameters is

$$P(d\alpha_1, d\alpha_2, dt_0; dX \,|\, D) = \frac{P(dD \,|\, X) P(dX \,|\, \alpha_1, \alpha_2, t_0) \pi(d\alpha_1, d\alpha_2, dt_0)}{\int P(dD \,|\, X') P(dX' \,|\, \alpha_1', \alpha_2', t_0') \pi(d\alpha_1', d\alpha_2', dt_0')}. \tag{3}$$

We thus see that treating the allele frequency path as a parameter is consistent with the initial "naive" Bayesian approach in that if we integrate the path variable out of the posterior (3) for the augmented family of parameters, then we recover the posterior (2)

for the original family of parameters. In practice, this means that marginalizing out the path variable from a Monte Carlo approximation of the augmented posterior gives a Monte Carlo approximation of the original posterior.

Implicit in our set-up is the initial frequency x_0 at time t_0. Under the probability measure governing the Wright-Fisher diffusion, any process started from $x_0 = 0$ will stay there forever. Thus, we would be forced to make an arbitrary choice of some $x_0 > 0$ as the initial frequency of our allele. However, we argue in the Appendix that in the limit as $x_0 \downarrow 0$, we can achieve an improper prior distribution on the space of allele frequency trajectories. We stress that our inference using such an improper prior is not one that arises directly from a generative probability model for the allele frequency path. However, it does arise as a limit as the initial allele frequency x_0 goes to zero of inferential procedures based on generative probability models and the limiting posterior distributions are probability distributions. Therefore, the parameters α_1, α_2, t_0 retain their meaning, our conclusions can be thought of approximations to those that we would arrive at for all sufficiently small values of x_0, and we are spared the necessity of making an arbitrary choice of x_0.

2.3. **Path likelihoods.** Most instances of Bayesian inference in population genetics have hitherto involved finite-dimensional parameters. We recall that if a finite-dimensional parameter has a diffuse prior distribution (that is, a distribution where an individual specification of values of the parameter has zero prior probability), then one replaces the prior probabilities of parameter values that would be appear when if we had a discrete prior distribution by evaluations of densities with respect to an underlying reference measure – usually Lebesgue measure in an appropriate dimension – and the Bayesian formalism then proceeds in much the same way as it does in the discrete case with, for example, ratios of probabilities replaced by ratios of densities. We thus require a reference measure on the infinite-dimensional space of paths that will play a role analogous to that of Lebesgue measure in the finite-dimensional case.

To see what is involved, suppose we have a diffusion process $(Z_t)_{t \geq t_0}$ that satisfies the SDE

$$\begin{aligned} dZ_t &= a(Z_t, t)\, dt + dB_t \\ Z_{t_0} &= z_0, \end{aligned} \tag{4}$$

where B is a standard Brownian motion (the Wright-Fisher diffusion is not of this form but, as we shall soon see, it can be be reduced to it after suitable transformations of time and space). Let $\mathbb{P}$ be the distribution of $(Z_t)_{t \geq t_0}$ – this is a probability distribution on the space of continuous paths that start from position z_0 at time t_0. While the probability assigned by $\mathbb{P}$ to any particular path is zero, we can, under appropriate conditions, make sense of the probability of a path under $\mathbb{P}$ *relative* to its probability under the distribution of Brownian motion. If we denote by $\mathbb{W}$ the distribution of Brownian motion starting from position z_0 at time t_0, then Girsanov's theorem [Girsanov, 1960] gives the density of the path segment $(Z_s)_{t_0 \leq s \leq t}$ under $\mathbb{P}$ relative to $\mathbb{W}$ as

$$\frac{d\mathbb{P}}{d\mathbb{W}}((Z_s)_{t_0 \leq s \leq t}) = \exp\left\{ \int_{t_0}^{t} a(Z_s, s)\, dZ_s - \frac{1}{2} \int_{t_0}^{t} a^2(Z_s, s)\, ds \right\}, \tag{5}$$

where the first integral in the exponentiand is an Itô integral. In order for (5) to hold, the integral $\int_{t_0}^{t} a^2(Z_s, s)\, ds$ must be finite, in which case the Itô integral $\int_{t_0}^{t} a(Z_s, s)\, dZ_s$ is also well-defined and finite.

However, the Wright-Fisher SDE (1) is not of the form (4). In particular, the factor multiplying the infinitesimal Brownian increment dB_t (the so-called diffusion coefficient) depends on both space and time. To deal with this issue, we first apply a well-known time transformation and consider the process $(\tilde{X}_\tau)_{\tau \geq 0}$ given by $\tilde{X}_\tau = X_{f^{-1}(\tau)}$, where

$$f(t) = \int_{t_0}^{t} \frac{1}{\rho(s)}\, ds, \quad t \geq t_0. \tag{6}$$

It is not hard to see that $(\tilde{X}_\tau)_{\tau \geq 0}$ satisfies the following SDE with a time-independent diffusion coefficient,

$$\begin{aligned} d\tilde{X}_\tau &= \rho(f^{-1}(\tau))\tilde{X}_\tau(1-\tilde{X}_\tau)(\alpha_1(2\tilde{X}_\tau - 1) - \alpha_2\tilde{X}_\tau)\, d\tau + \sqrt{\tilde{X}_\tau(1-\tilde{X}_\tau)}\, d\tilde{B}_\tau \\ \tilde{X}_0 &= x_0, \end{aligned}$$

where $\tilde{B}$ is a standard Brownian motion. Next, we employ an angular space transformation first suggested by Fisher [1922], $Y_\tau = \arccos(1 - 2\tilde{X}_\tau)$. Applying Itô's lemma [Itô, 1944] shows that $(Y_\tau)_{\tau \geq 0}$ is a diffusion that satisfies the SDE

$$\begin{aligned} dY_\tau &= \frac{1}{4}\left(\rho(f^{-1}(\tau))\sin(Y_\tau)(\alpha_2 + (2\alpha_1 - \alpha_2)\cos(Y_\tau)) - 2\cot(Y_\tau)\right)\, d\tau + dW_\tau \\ Y_0 &= y_0 = \arccos(1 - 2x_0), \end{aligned} \tag{7}$$

where W is a standard Brownian motion. If the process X hits either of the boundary points $0, 1$, then it stays there, and the same is true of the time and space transformed process Y for its boundary points $0, \pi$.

The restriction of the distribution of the time and space transformed process Y to some set of paths that don't hit the boundary is absolutely continuous with respect to the distribution of standard Brownian motion restricted to the same set; that is, the distribution of Y restricted to such a set of paths has a density with respect to the distribution of Brownian motion restricted to the same set. However, the infinitesimal mean in (7) (that is, the term multiplying $d\tau$) becomes singular as Y_τ approaches the boundary points 0 and π, corresponding to the boundary points 0 and 1 for allele frequencies. These singularities prevent the process Y from re-entering the interior of its state space and ensure that a Wright-Fisher path will be absorbed when the allele is either fixed or lost. A consequence is that the density of the distribution of Y relative to that of a Brownian motion blows up as the path approaches the boundary. We are modeling the appearance of a new mutation in terms of a Wright-Fisher diffusion starting at some small initial frequency x_0 at time t_0 and we want to perform our parameter inference in such a way that we get meaningful answers as $x_0 \downarrow 0$. This suggests that rather than working with the distribution $\mathbb{W}$ of Brownian motion as a reference measure it may be more appropriate to work with a tractable diffusion process that exhibits similar behavior near the boundary point 0.

To start making this idea of matching singularities more precise, consider a diffusion process $(\bar{Z}_t)_{t\geq t_0}$ that satisfies the SDE

$$\begin{aligned} d\bar{Z}_t &= b(\bar{Z}_t, t)\, dt + d\bar{B}_t \\ \bar{Z}_0 &= z_0, \end{aligned} \tag{8}$$

where $\bar{B}$ is a standard Brownian motion. Write $\mathbb{Q}$ for the distribution of the diffusion process $(\bar{Z}_t)_{t\geq t_0}$ and recall that $\mathbb{P}$ is the distribution of a solution of (4). If $(Z_s)_{t_0\leq s\leq t}$ is a segment of path such that both $\int_{t_0}^t a^2(Z_s,s)\,ds < \infty$ and $\int_{t_0}^t b^2(Z_s,s)\,ds < \infty$, then

$$\begin{aligned} \frac{d\mathbb{P}}{d\mathbb{Q}}((Z_s)_{t_0\leq s\leq t}) &= \frac{d\mathbb{P}}{d\mathbb{W}}((X_s)_{t_0\leq s\leq t}) \Big/ \frac{d\mathbb{Q}}{d\mathbb{W}}((Z_s)_{t_0\leq s\leq t}) \\ &= \exp\left\{ \int_{t_0}^t (a(Z_s,s) - b(Z_s,s))\ dZ_s - \frac{1}{2}\int_{t_0}^t \left(a^2(Z_s,s) - b^2(Z_s,s)\right)\ ds \right\}. \end{aligned} \tag{9}$$

Note that the right-hand side will stay bounded if one considers a sequence of paths, indexed by η, $(Z^\eta_s)_{t_0\leq s\leq t}$, with $\int_{t_0}^t a^2(Z^\eta_s,s)\,ds < \infty$ and $\int_{t_0}^t b^2(Z^\eta_s,s)\,ds < \infty$, provided that $\int_{t_0}^t (a^2(Z^\eta_s,s) - b^2(Z^\eta_s,s))\,ds$ stays bounded. These manipulations with densities may seem somewhat heuristic, but they can be made rigorous and, moreover, the form of $\frac{d\mathbb{P}}{d\mathbb{Q}}$ follows from an extension of Girsanov's theorem that gives the density of $\mathbb{P}$ with respect to $\mathbb{Q}$ directly without using the densities with respect to $\mathbb{W}$ as intermediaries (see, for example, [Kallenberg, 2002, Theorem 18.10]).

We wish to apply this observation to the time and space transformed Wright-Fisher diffusion of (7). Because

$$-\frac{1}{2}\cot(y) + \frac{1}{4}\rho(f^{-1}(t))\sin(y)\left((2\alpha_1 - \alpha_2)\cos(y) + \alpha_2\right) = -\frac{1}{2y} + O(y)$$

when y is small, an appropriate reference process should have infinitesimal mean $b(y,t) \approx -1/(2y)$ as $y \downarrow 0$. Following suggestions by Schraiber et al. [2013] and Jenkins [2013], we compute path densities relative to the distribution $\mathbb{Q}$ of the Bessel(0) process, a process which is the solution of the SDE

$$\begin{aligned} d\bar{Y}_t &= -\frac{1}{2\bar{Y}_t}\, dt + d\bar{B}_t, \\ \bar{Y}_0 &= y_0 = \arccos(1 - 2x_0). \end{aligned} \tag{10}$$

For the moment, write $\mathbb{P}^{y_0}$ and $\mathbb{Q}^{y_0}$ for the respective distributions of the solutions of (7) and (10) to emphasize the dependence on y_0 (equivalently, on the initial allele frequency x_0). There are σ-finite measures $\mathbb{P}^0$ and $\mathbb{Q}^0$ with infinite total mass such that for each $\epsilon > 0$

$$\lim_{y_0\downarrow 0} \mathbb{P}^{y_0}((Y_t)_{t\geq\epsilon} \in \cdot \,|\, Y_\epsilon > 0) = \mathbb{P}^0((Y_t)_{t\geq\epsilon} \in \cdot) \Big/ \mathbb{P}^0(Y_\epsilon > 0)$$

and

$$\lim_{y_0\downarrow 0} \mathbb{Q}^{y_0}((\bar{Y}_t)_{t\geq\epsilon} \in \cdot \,|\, \bar{Y}_\epsilon > 0) = \mathbb{Q}^0((\bar{Y}_t)_{t\geq\epsilon} \in \cdot) \Big/ \mathbb{Q}^0(\bar{Y}_\epsilon > 0),$$

where the numerators and denominators in the last two equations are all finite. Moreover, $\mathbb{P}^0$ has a density with respect to $\mathbb{Q}^0$ that arises by naively taking limits as $y_0 \downarrow 0$ in the functional form of the density of $\mathbb{P}^{y_0}$ with respect to $\mathbb{Q}^{y_0}$ (we say "naively" because $\mathbb{P}^{y_0}$ and $\mathbb{Q}^{y_0}$ assign all of their mass to paths that start at position $y_0 = \arccos(1-2x_0)$ at time 0, whereas $\mathbb{P}^0$ and $\mathbb{Q}^0$ assign all of their mass to paths that start at position 0 at time 0, and so the set of paths at which it is relevant to compute the density changes as $y_0 \downarrow 0$). As we have already remarked, the limit of our Bayesian inferential procedure may be thought of as Bayesian inference with an improper prior, but we stress that the resulting posterior is proper.

The notion of the infinite measure $\mathbb{Q}^0$ may seem somewhat forbidding, but this measure is characterized by the following simple properties:

$$\mathbb{Q}^0(\bar{Y}_\epsilon \in dy) = \frac{y^2}{\epsilon^2} \exp\left\{-\frac{y^2}{2\epsilon}\right\} dy, \quad y > 0,$$

so that $\mathbb{Q}^0(\bar{Y}_\epsilon > 0) = \sqrt{\frac{\pi}{2}}\frac{1}{\sqrt{\epsilon}}$, and conditional on the event $\{\bar{Y}_\epsilon = y\}$ the evolution of $(\bar{Y}_t)_{t\geq\epsilon}$ is exactly that of the Bessel(0) process started at position y at time ϵ. Moreover, conditional on the event $\{\bar{Y}_s = a, \bar{Y}_u = b\}$ for $0 \leq s < u$ and $a, b > 0$, the evolution of the "bridge" $(\bar{Y}_u)_{s\leq t\leq u}$ is the same as that of the corresponding bridge for a Bessel(4) process; a Bessel(4) process satisfies the SDE

$$d\hat{Y}_t = \frac{3}{2\hat{Y}_t}\, dt + d\hat{B}_t.$$

Very importantly for the sake of simulations, the Bessel(4) process is just the radial part of a 4-dimensional standard Brownian motion – in particular, this process started at 0 leaves immediately and never returns. Also, the Bessel(0) process arises naturally because our space transformation $x \mapsto \arccos(1-2x) = \int_0^x \frac{1}{\sqrt{w(1-w)}}\, dw$ is approximately $x \mapsto \int_0^x \frac{1}{\sqrt{w}}\, dw = 2\sqrt{x}$ when $x > 0$ is small and a multiple of the square of Bessel(0) process, sometimes called Feller's continuous state branching processes, arises naturally as an approximation to the Wright-Fisher diffusion for low frequencies [Haldane, 1927, Feller, 1951].

2.4. **The joint likelihood of the data and the path.** To write down down the full likelihood of the observations and the path, we make the assumption that the population size function $\rho(t)$ is continuously differentiable except at a finite set of times $d_1 < d_2 < \ldots < d_M$. Further, we require that that $\rho(d_i^+) = \lim_{t\downarrow d_i} \rho(t)$ exists and is equal to $\rho(d_i)$ while $\rho(d_i^-) = \lim_{t\uparrow d_i} \rho(t)$ also exists (though it may not necessarily equal $\rho(d_i)$).

Using the notation of Subsection 2.2, write

$$L(D, (Y_t)_{t\geq 0} \,|\, \alpha_1, \alpha_2, t_0) = \mathbb{P}(D \,|\, (Y_t)_{t\geq 0}, t_0)\Phi^0((Y_t)_{t\geq 0}; \alpha_1, \alpha_2, t_0)$$

for the joint likelihood of the data and the time and space transformed allele frequency path $(Y_t)_{t\geq 0}$ given the parameters α_1, α_2, t_0. In the Appendix, we show that

$$
\begin{aligned}
&L(D, (Y_s)_{0\leq s\leq t_k} \,|\, \alpha_1, \alpha_2, t_0) \\
&\quad = \exp\Bigg\{ A(Y_{f(t_k)}, t_k^-) + A(Y_{f(d_m)}, d_m^-) - (A(Y_{f(d_K)}, d_K) + A(Y_{f(t_0)}, t_0)) \\
&\qquad + \sum_{i=m}^{K} \left[A(Y_{f(d_{i+1})}, d_{i+1}^-) - A(Y_{f(d_i)}, d_i) \right] \\
&\qquad - \int_{t_0}^{t_k} B(Y_{f(s)}, s) ds - \frac{1}{2}\int_{t_0}^{t_k} C(Y_{f(s)}, s) ds - \frac{1}{2}\int_{t_0}^{t_k} D(Y_{f(s)}, s) ds \Bigg\} \\
&\qquad \times \prod_{i=1}^{k} \binom{n_i}{c_i} \left(\frac{1-\cos(Y_{f(t_i)})}{2} \right)^{c_i} \left(\frac{1+\cos(Y_{f(t_i)})}{2} \right)^{n_i - c_i},
\end{aligned}
\tag{11}
$$

where f is as in (6), $m = \min\{i : d_i > t_0\}$ and $K = \max\{i : d_i > t_k\}$, and

$$
\begin{aligned}
A(y,t) &= \frac{\log(y)}{2} - \frac{1}{8}\left(\rho(t)\cos(y)(2\alpha_2 + (2\alpha_1 - \alpha_2)\cos(y)) + 4\log(\sin(y))\right) \\
B(y,t) &= -\frac{1}{8}\frac{d\rho}{dt}(t)\cos(y)(2\alpha_2 + (2\alpha_1 - \alpha_2)\cos(y)) \\
C(y,t) &= \frac{1}{2}\left(\alpha_1\cos(y) + \frac{\csc(y)^2}{\rho(t)}\right) - \frac{1}{2y^2\rho(t)} \\
D(y,t) &= \frac{1}{16\rho(t)}\left(\rho(t)\sin(y)(\alpha_2 + (2\alpha_1 - \alpha_2)\cos(y)) - 2\cot(y)\right)^2 - \frac{1}{4y^2\rho(t)}.
\end{aligned}
$$

While this expression may appear complicated, it has the important feature that, unlike the form of the likelihood that would arise by simply applying Girsanov's theorem, it only involves Lebesgue (indeed Riemann) integrals and not Itô integrals, which, as we recall in the Appendix, are known from the literature to be potentially difficult to compute numerically.

2.5. **Metropolis-Hastings algorithm.** We now describe a Markov chain Monte Carlo method for Bayesian inference of the parameters α_1, α_2 and t_0, along with the allele frequency path $(X_t)_{t\geq t_0}$ (equivalently, the transformed path $(Y_t)_{t\geq 0}$). While updates to the selection parameters α_1 and α_2 do not require updating the path, updating the time t_0 at which the derived allele arose requires proposing updates to the segment of path from t_0 up to the time of the first sample with a non-zero number of derived alleles. Additionally, we require proposals to update small sections of the path without updating any parameters and proposals to update the allele frequency at the most recent sample time.

[Figure 2 about here.]

2.5.1. *Interior path updates.* To update a section of the allele frequency, we first choose a time $s_1 \in (t_0, t_k)$ uniformly at random, and then choose a time s_2 that is a fixed fraction of the path length subsequent to s_1. We prefer this approach of updating a fixed fraction of

the path to an alternative strategy of holding $s_2 - s_1$ constant because paths for very strong selection may be quite short. Recalling the definition of f from (6), we subsequently propose a new segment of transformed path between the times $f(s_1)$ and $f(s_2)$ while keeping the values $Y_{f(s_1)}$ and $Y_{f(s_2)}$ fixed (Figure 2a). Such a path that is conditioned to take specified values at both end-points of the interval over which it is defined is called a bridge, and by updating small portions of the path instead of the whole path at once, we are able to obtain the desirable behavior that our Metropolis-Hastings algorithm is able to stay in regions of path space with high posterior probability. If we instead drew the whole path each time, we would much less efficiently target the posterior distribution.

Noting that bridges must be sampled against the *transformed* time scale, the best bridges for the allele frequency path would be realizations of Wright-Fisher bridges themselves. However, sampling Wright-Fisher bridges is challenging (but see Schraiber et al. [2013], Jenkins and Spano [2015]), so we instead opt to sample bridges for the transformed path from the Bessel(0) process. Sampling Bessel(0) bridges can be accomplished by first sampling Bessel(4) bridges (as described in Schraiber et al. [2013]) and then recognizing that a Bessel(4) process is the same as a Bessel(0) process conditioned to never hit 0 and hence has the same bridges – in the language of the general theory of Markov processes, the Bessel(0) and Bessel(4) processes are Doob h-transforms of each other and it is well-known that processes related in this way share the same bridges. We denote by $(Y'_\tau)_{\tau \geq 0}$ the path that has the proposed bridge spliced in between times $f(s_1)$ and $f(s_2)$ and coincides with $(Y_\tau)_{\tau \geq 0}$ outside the interval $[f(s_1), f(s_2)]$.

In the Appendix, we show that the acceptance probability in this case is simply

$$\min\left\{1, \frac{L(D, (Y'_\tau)_{f(s_1)\leq\tau\leq f(s_2)} \,|\, \alpha_1, \alpha_2, t_0)}{L(D, (Y_\tau)_{f(s_1)\leq\tau\leq f(s_2)} \,|\, \alpha_1, \alpha_2, t_0)}\right\}. \tag{12}$$

Note that we only need to compute the likelihood ratio for the segment of transformed path that changed between the times $f(s_1)$ and $f(s_2)$.

2.5.2. *Allele age updates.* The first sample time with a non-zero count of the derived allele (Figure 2b) is t_s, where $s = \min\{i : c_i > 0\}$. We must have $t_0 < t_s$. Along with proposing a new value t'_0 of the allele age t_0 we will propose a new segment of the allele frequency path from time t'_0 to time t_s. Changing the allele age t_0 to some new proposed value t'_0 changes the definition of the function f in (6). Write $f'(t) = \int_{t'_0}^{t} \frac{1}{\rho(s)}\, ds$, where we stress that the prime does not denote a derivative. The proposed transformed path Y' consists of a new piece of path that goes from location 0 at time 0 to location $Y_{f(t_s)}$ at time $f'(t_s)$ and then has $Y'_{f'(t)} = Y_{f(t)}$ for $t \geq t_s$. We use the improper prior $\rho(t_0)$ for t_0, which reflects the fact that an allele is more likely to arise during times of large population size [Slatkin, 2001]. In the Appendix, we show that the acceptance probability is

$$\min\left\{1, \frac{L(D, (Y'_\tau)_{0\leq\tau\leq f'(t_s)} \,|\, \alpha_1, \alpha_2, t'_0)}{L(D, (Y_\tau)_{0\leq\tau\leq f(t_s)} \,|\, \alpha_1, \alpha_2, t_0)} \frac{\psi(Y'_{f'(t_s)}; f'(t_s))}{\psi(Y_{f(t_s)}; f(t_s)))} \frac{q(t_0|t'_0)}{q(t'_0|t_0)} \frac{\rho(t'_0)}{\rho(t_0)}\right\} \tag{13}$$

where, in the notation of Subsection 2.3,

$$\psi(y;\epsilon) = \frac{y^2}{\epsilon^2}\exp\left\{-\frac{y^2}{2\epsilon}\right\} = \frac{\mathbb{Q}^0(\bar{Y}_\epsilon \in dy)}{dy} \tag{14}$$

is the density of the so-called entrance law for the Bessel(0) process that appears in the characterization of the σ-finite measure $\mathbb{Q}^0$ and $q(t_0'|t_0)$ is the proposal distribution of t_0' (in practice, we use a half-truncated normal distribution centered at t_0, with the upper truncation occurring at the first time of non-zero observed allele frequency).

2.5.3. *Most recent allele frequency update.* While the allele frequency at sample times $t_1, t_2, \ldots, t_{k-1}$ are updated implicitly by the interior path update, we update the allele frequency at the most recent sample time t_k separately (note that the most recent allele frequency is not an additional parameter, but simply a random variable with a distribution implied by the Wright-Fisher model on paths). We do this by first proposing a new allele frequency $Y'_{f(t_k)}$ and then proposing a new bridge from $Y_{f(t_f)}$ to $Y'_{f(t_k)}$ where $t_f \in (t_{k-1}, t_k)$ is a fixed time (Figure 2c). If $q(Y'_{f(t_k)} \,|\, Y_{f(t_k)})$ is the proposal density for $Y'_{f(t_k)}$ given $Y_{f(t_k)}$ (in practice, we use a truncated normal distribution centered at $Y_{f(t_k)}$ and truncated at 0 and π), then, arguing along the same lines as the interior path update and the allele age update, we accept this update with probability

$$\min\left\{1, \frac{L(D, (Y'_\tau)_{f(t_f)\le\tau\le f(t_k)} \,|\, \alpha_1, \alpha_2, t_0)}{L(D, (Y_\tau)_{f(t_f)\le\tau\le f(t_k)} \,|\, \alpha_1, \alpha_2, t_0)} \frac{q(Y_{f(t_k)}|Y'_{f(t_k)})}{q(Y'_{f(t_k)}|Y_{f(t_k)})} \frac{Q(Y_{f(t_f)}, Y_{f(t_k)}; f(t_k)-f(t_f))}{Q(Y_{f(t_f)}, Y'_{f(t_k)}; f(t_k)-f(t_f))}\right\}, \tag{15}$$

where

$$Q(x,y;t) = \frac{y}{t}\exp\left\{-\frac{x^2+y^2}{2t}\right\} I_1\left(\frac{xy}{t}\right) \tag{16}$$

is the transition density of the Bessel(0) process (with $I_1(\cdot)$ being the Bessel function of the first kind with index 1) – see Knight [1981, Section 4.3.6]. Again, it is only necessary to compute the likelihood ratio for the segment of transformed path that changed between the times $f(t_f)$ and $f(t_k)$.

2.6. **Updates to α_1 and α_2.** Updates to α_1 and α_2 are conventional scalar parameter updates. For example, letting $q(\alpha_1' \,|\, \alpha_1)$ be the proposal density for the new value of α_1, we accept the new proposal with probability

$$\min\left\{1, \frac{L(D, (Y_\tau)_{\tau\ge 0} \,|\, \alpha_1', \alpha_2, t_0)}{L(D, (Y_\tau)_{\tau\ge 0} \,|\, \alpha_1, \alpha_2, t_0)} \frac{q(\alpha_1 \,|\, \alpha_1')}{q(\alpha_1' \,|\, \alpha_1)} \frac{\pi(\alpha_1', \alpha_2, t_0)}{\pi(\alpha_1, \alpha_2, t_0)}\right\}.$$

The acceptance probability for α_2 is similar. For both α_1 and α_2, we use a heavy-tailed Cauchy prior with median 0 and scale parameter 100, and we take the parameters α_1, α_2, t_0 to be independent under the prior distribution. In addition, we use a normal proposal distribution, centered around the current value of the parameter. Here, it is necessary to compute the likelihood across the whole path.

3. Results

We first test our method using simulated data to assess its performance and then apply it to two real datasets from horses.

3.1. **Simulation performance.** To test the accuracy of our MCMC approach, we simulated allele frequency trajectories with ages uniformly distributed between 0.1 and 0.3 diffusion time units ago, evolving with α_1 and α_2 uniformly distributed between 0 and 100. We simulate allele frequency trajectories using an Euler approximation to the Wright-Fisher SDE (1) with $\rho(t) \equiv 1$. At each time point between -0.4 and 0.0 in steps of 0.05, we simulated of 20 chromosomes.

We then ran the MCMC algorithm for $1,000,000$ generations, sampling every 1000 generations to obtain 1000 MCMC samples for each simulation. After discarding the first 500 samples from each MCMC run as burn-in, we computed the effective sample size of the allele age estimate using the R package `coda` [Plummer et al., 2006]. For the analysis of the simulations, we only included simulations that had an effective sample size greater than 150 for the allele age, resulting in retaining 744 out of 1000 simulations.

Because our MCMC analysis provides a full posterior distribution on parameter values, we summarized the results by computing the maximum *a posteriori* estimate of each parameter. We find that across the range of simulated α_1 values, estimation is quite accurate (Figure 3A). There is some downward bias for large true values of α_1, indicating the influence of the prior. On the other hand, the strength of selection in favor of the homozygote, α_2, is less well estimated, with a more pronounced downward bias (Figure 3B). This is largely because most simulated alleles do not reach sufficiently high frequency for homozygotes to be common. Hence, there is very little information regarding the fitness of the homozygote. Allele age is estimated accurately, although there is a slight bias toward estimating a more recent age than the truth (Figure 3C).

[Figure 3 about here.]

3.2. **Application to ancient DNA.** We applied our approach to real data by reanalyzing the MC1R and ASIP data from Ludwig et al. [2009]. In contrast to earlier analyses of these data, we explicitly incorporated the demography of the domesticated horse, as inferred by Der Sarkissian et al. [2015], using a generation time of 8 years. Table 2 shows the sample configurations and sampling times corresponding to each locus, where diffusion units are scaled to $2N_0$, with $N_0 = 16000$ being the most recent effective size reported by Der Sarkissian et al. [2015]. For comparison, we also analyzed the data assuming the population size has been constant at N_0.

[Table 2 about here.]

[Figure 4 about here.]

With the MC1R locus, we found that posterior inferences about selection coefficients can be strongly influenced by whether or not demographic information is included in the analysis (Figure 4). Marginally, we see that incorporating demographic information results in an inference that α_1 is larger than the constant-size model (MAP estimates of 267.6 and

74.1, with and without demography, respectively; Figure 4A), while α_2 is inferred to be smaller (MAP estimates of 59.1 and 176.2, with and without demography, respectively; Figure 4B). This has very interesting implications for the mode of selection inferred on the MC1R locus. With constant demography, the trajectory of the allele is estimated to be shaped by positive selection (joint MAP, $\alpha_1 = 87.6$, $\alpha_2 = 394.8$; Figure 4C), while when demographic information is included, selection is inferred to act in an overdominant fashion (joint MAP, $\alpha_1 = 262.5$, $\alpha_2 = 128.1$; Figure 4D).

[Figure 5 about here.]

Incorporation of demographic history also has substantial impacts on the inferred distribution of allele ages (Figure 5). Most notably, the distribution of the allele age for MC1R is significantly truncated when demography is incorporated, in a way that correlates to the demographic events (Supplementary Figure 8). While both the constant-size history and the more complicated history result in a posterior mode at approximately the same value of the allele age, the domestication bottleneck inferred by Der Sarkissian et al. [2015] makes it far less likely that the allele rose more anciently than the recent population expansion. Because the allele is inferred to be younger under the model incorporating demography, the strength of selection in favor of the homozygote must be higher to allow it to escape low frequency quickly and reach the observed allele frequencies. Hence, α_1 is inferred to be much higher when demographic history is explicitly modeled.

[Figure 6 about here.]

Incorporation of demographic history has an even more significant impact on inferences made about the ASIP locus (Figure 6). Most strikingly, while α_1 is inferred to be very large without demography, it is inferred to be close to 0 when demography is incorporated (MAP estimates of 16.3 and 159.9 with and without demography, respectively; Figure 6A). On the other hand, inference of α_2 is largely unaffected (MAP estimates of 34.7 and 39.8 with and without demography, respectively; Figure 6B). Interestingly, this has an opposite implication for the mode of selection compared to the results for the MC1R locus. With a constant-size demographic history, the allele is inferred to have evolved under overdominance (joint MAP, $\alpha_1 = 153.3$, $\alpha_2 = 47$; Figure 6C), but when the more complicated demography is modeled, the allele frequency trajectory is inferred to be shaped by positive, nearly additive, selection (joint MAP, $\alpha_1 = 16.4$, $\alpha_2 = 46.8$; Figure 6D).

[Figure 7 about here.]

Incorporating demography has a similarly opposite effect on inference of allele age (Figure 7). In particular, the allele is inferred to be much older when demography is modeled, and features a multi-modal posterior distribution on allele age, with each mode corresponding to a period of historically larger population size (Figure 9). Because the allele is inferred to be substantially older when demography is modeled, selection in favor of the heterozygote must have been weaker than would be inferred with the much younger age. Hence, the mode of selection switches from one of overdominance in a constant demography to one in which the homozygote is more fit than the heterozygote.

4. Discussion

Using DNA from ancient specimens, we have obtained a number of insights into evolutionary processes that were previously inaccessible. One of the most interesting aspects of ancient DNA is that it can provide a *temporal* component to evolution that has long been impossible to study. In particular, instead of making inferences about the allele frequencies, we can directly measure these quantities. To take advantage of this new data, we developed a novel Bayesian method for inferring the intensity and direction of natural selection from allele frequency time series. In order to circumvent the difficulties inherent in calculating the transition probabilities under the standard Wright-Fisher process of selection and drift, we used a data augmentation approach in which we learn the posterior distribution on allele frequency paths. Doing this not only allows us to efficiently calculate likelihoods, but provides an unprecedented glimpse at the historical allele frequency dynamics.

The key innovation of our method is to apply high-frequency path augmentation methods [Roberts and Stramer, 2001] to analyze genetic time series. The logic of the method is similar to the logic of a path integral, in which we average over all possible allele frequency trajectories that are consistent with the data [Schraiber, 2014]. By choosing a suitable reference probability distribution against which to compute likelihood ratios, we were able to adapt these methods to infer the age of alleles and properly account for variable population sizes through time. Moreover, because of the computational advantages of the path augmentation approach, we were able to infer a model of general diploid selection. To our knowledge, ours is the first work that can estimate both allele age and general diploid selection while accounting for demography.

Using simulations, we showed that our method performs well for strong selection and densely sampled time series. However, it is worth considering the work of Watterson [1979], who showed that even knowledge of the full trajectory results in very flat likelihood surfaces when selection is not strong. This is because for weak selection, the trajectory is extremely stochastic and it is difficult to disentangle the effects of drift and selection [Schraiber et al., 2013].

We then applied our method to a classic dataset from horses. We found that our inference of both the strength and mode of natural selection depended strongly on whether or not we incorporated demography. For the MC1R locus, a constant-size demographic model results in an inference of positive selection, while the more complicated demographic model inferred by Der Sarkissian et al. [2015] causes the inference to tilt toward overdominance, as well as a much younger allele age. In contrast, the ASIP locus is inferred to be overdominant under a constant-size demography, but the complicated demographic history results in an inference of positive selection, and a much older allele age.

These results stand in contrast to those of Steinrücken et al. [2014], who found that the most likely mode of evolution for both loci under a constant demographic history is one of overdominance. There are a several reasons for this discrepancy. First, we computed the diffusion time units differently, using $N_0 = 16000$ and a generation time of 8 years, as inferred by Der Sarkissian et al. [2015], while Steinrücken et al. [2014] used $N_0 = 2500$ (consistent with the bottleneck size found by Der Sarkissian et al. [2015]) and

a generation time of 5 years. Hence, our constant-size model has far less genetic drift than the constant-size model assumed by Steinrücken et al. [2014]. This emphasizes the importance of inferring appropriate demographic scaling parameters, even when a constant population size is assumed. Secondly, we use MCMC to integrate over the distribution of allele ages, which can have a very long tail going into the past, while Steinrücken et al. [2014] assume a fixed allele age.

One key limitation of this method is that it assumes that the aDNA samples all come from the same, continuous population. If there is in fact a discontinuity in the populations from which alleles have been sampled, this could cause rapid allele frequency change and create spurious signals of natural selection. Several methods have been devised to test this hypothesis [Sjödin et al., 2014], and one possibility would be to apply these methods to putatively neutral loci sampled from the same individuals, thus determining which samples form a continuous population. Alternatively, if our method is applied to a number of loci throughout the genome and an extremely large portion of the genome is determined to be evolving under selection, this could be evidence for model misspecification and suggest that the samples do not come from a continuous population.

An advantage of the method that we introduced is that it may be possible to extend it to incorporate information from linked neutral diversity. In general, computing the likelihood of neutral diversity linked to a selected site is difficult and many have used Monte Carlo simulation and importance sampling [Slatkin, 2001, Coop and Griffiths, 2004, Chen and Slatkin, 2013]. These approaches average over allele frequency trajectories in much same way as our method; however, each trajectory is drawn completely independently of the previous trajectories. Using a Markov chain Monte Carlo approach, as we do here, has the potential to ensure that only trajectories with a high posterior probability are explored and hence greatly increase the efficiency of such approaches.

5. Appendix

5.1. **A proper posterior in the limit as the intiial allele frequency approaches 0.**

For reasons that we explain in Subsection 2.3, we re-parametrize our model by replacing the path variable $(X_t)_{t\geq t_0}$ with a deterministic time and space transformation of it $(Y_t)_{t\geq 0}$ that takes values in the interval $[0,\pi]$ with the boundary point 0 (resp. π) for $(Y_t)_{t\geq 0}$ corresponding to the boundary point 0 (resp. 1) for $(X_t)_{t\geq t_0}$. The transformation producing $(Y_t)_{t\geq 0}$ is such that $(X_t)_{t\geq t_0}$ can be recovered from $(Y_t)_{t\geq 0}$ and t_0.

Implicit in our set-up is the initial frequency x_0 at time t_0 which corresponds to an initial value y_0 at time 0 of the transformed process $(Y_t)_{t\geq 0}$. For the moment, let us make the dependence on y_0 explicit by including it in relevant notation as a superscript. For example, $\mathbb{P}^{y_0}(\cdot \,|\, \alpha_1, \alpha_2, t_0)$ is the prior distribution of $(Y_t)_{t\geq 0}$ given the specified values of the other parameters α_1, α_2, t_0. We will construct a tractable "reference" process $(\bar{Y}_t)_{t\geq 0}$ with distribution $\mathbb{Q}^{y_0}(\cdot)$ such that the probability distribution $\mathbb{P}^{y_0}(\cdot \,|\, \alpha_1, \alpha_2, t_0)$ has a density with respect to $\mathbb{Q}^{y_0}(\cdot)$ – explicitly, $\mathbb{Q}^{y_0}(\cdot)$ is the distribution of a Bessel(0) process started at location y_0 at time 0. That is, there is a function $\Phi^{y_0}(\cdot; \alpha_1, \alpha_2, t_0)$ on path space such

that

$$\mathbb{P}^{y_0}(dy \,|\, \alpha_1, \alpha_2, t_0) = \Phi^{y_0}(y; \alpha_1, \alpha_2, t_0)\, \mathbb{Q}^{y_0}(dy) \tag{17}$$

for a path $(y_t)_{t\geq 0}$. Assuming that π has a density with respect to Lebesgue measure which, with a slight abuse of notation, we also denote by π, the outcome of our Bayesian inferential procedure is determined by the ratios

$$\frac{\mathbb{P}(dD \,|\, y^{**}, t_0^{**})\Phi^{y_0}(y^{**}; \alpha_1^{**}, \alpha_2^{**}, t_0^{**})\pi(\alpha_1^{**}, \alpha_2^{**}, t_0^{**})}{\mathbb{P}(dD \,|\, y^{*}, t_0^{*})\Phi^{y_0}(y^{*}; \alpha_1^{*}, \alpha_2^{*}, t_0^{*})\pi(\alpha_1^{*}, \alpha_2^{*}, t_0^{*})} \tag{18}$$

for pairs of augmented parameter values $(y^*, \alpha_1^*, \alpha_2^*, t_0^*)$ and $(y^{**}, \alpha_1^{**}, \alpha_2^{**}, t_0^{**})$ (*i.e.* the Metropolis-Hastings ratio).

Under the probability measure $\mathbb{P}^{y_0}(\cdot \,|\, \alpha_1, \alpha_2, t_0)$, the process $(Y_t)_{t\geq 0}$ converges in distribution as $y_0 \downarrow 0$ (equivalently, $x_0 \downarrow 0$) to the trivial process that starts at location 0 at time 0 and stays there. However, for all $\epsilon > 0$ the conditional distribution of $(Y_t)_{t\geq\epsilon}$ under the probability measure $\mathbb{P}^{y_0}(\cdot \,|\, \alpha_1, \alpha_2, t_0)$ given the event $\{Y_\epsilon > 0\}$ converges to a non-trivial probability measure as $y_0 \downarrow 0$. Similarly, the conditional distribution of the reference diffusion process $(\bar{Y}_t)_{t\geq\epsilon}$ under the probability measure $\mathbb{Q}^{y_0}(\cdot)$ given the event $\{\bar{Y}_\epsilon > 0\}$ converges as $y_0 \downarrow 0$ to a non-trivial limit. There are σ-finite measures $\mathbb{P}^0(\cdot \,|\, \alpha_1, \alpha_2, t_0)$ and $\mathbb{Q}^0(\cdot)$ on path space that both have infinite total mass, are such that for any $\epsilon > 0$ both of these measures assign finite, non-zero mass to the set of paths that are strictly positive at the time ϵ, and the corresponding conditional probability measures are the limits as $y_0 \downarrow 0$ of the conditional probability measures described above. Moreover, there is a function $\Phi^0(\cdot; \alpha_1, \alpha_2, t_0)$ on path space such that

$$\mathbb{P}^0(dy \,|\, \alpha_1, \alpha_2, t_0) = \Phi^0(y; \alpha_1, \alpha_2, t_0)\, \mathbb{Q}^0(dy). \tag{19}$$

The posterior distribution (3) converges to

$$\mathbb{P}^0(d\alpha_1, d\alpha_2, dt_0; dY \,|\, D) = \frac{\mathbb{P}(dD \,|\, Y, t_0)\mathbb{P}^0(dY \,|\, \alpha_1, \alpha_2, t_0)\pi(d\alpha_1, d\alpha_2, dt_0)}{\int \mathbb{P}(dD \,|\, Y')\mathbb{P}^0(dY' \,|\, \alpha_1', \alpha_2', t_0')\pi(d\alpha_1', d\alpha_2', dt_0')}. \tag{20}$$

Thus, the limit as $y_0 \downarrow 0$ of a Bayesian inferential procedure for the augmented set of parameters can be viewed as a Bayesian inferential procedure with the improper prior $\mathbb{P}^0(dY \,|\, \alpha_1, \alpha_2, t_0)\pi(d\alpha_1, d\alpha_2, dt_0)$ for the parameters $Y, \alpha_1, \alpha_2, t_0$. In particular, the limiting Bayesian inferential procedure is determined by the ratios

$$\frac{\mathbb{P}(dD \,|\, y^{**}, t_0^{**})\Phi^{0}(h^{**}; \alpha_1^{**}, \alpha_2^{**}, t_0^{**})\pi(\alpha_1^{**}, \alpha_2^{**}, t_0^{**})}{\mathbb{P}(dD \,|\, y^{*}, t_0^{*})\Phi^{0}(y^{*}; \alpha_1^{*}, \alpha_2^{*}, t_0^{*})\pi(\alpha_1^{*}, \alpha_2^{*}, t_0^{*})} \tag{21}$$

for pairs of augmented parameter values $(y^*, \alpha_1^*, \alpha_2^*, t_0^*)$ and $(y^{**}, \alpha_1^{**}, \alpha_2^{**}, t_0^{**})$.

5.2. **The likelihood of the data and the path.** Write $\tau_i = f(t_i)$. Note that $\tau_0 = f(t_0) = 0$. Using equation (9), the density of the distribution of the transformed allele frequency process $(Y_t)_{0\leq s\leq\tau_k}$ against the reference distribution of the Bessel(0) process $(\bar{Y}_s)_{0\leq s\leq\tau_k}$ when $Y_0 = \bar{Y}_0 = y_0$ can be written

$$\exp\left\{\int_0^{\tau_k} (a(Y_r, r) - b(Y_r))\, dY_r - \frac{1}{2}\int_0^{\tau_k} \left(a^2(Y_r, r) - b^2(Y_r)\right)\, dr\right\} \tag{22}$$

where

$$a(y,\tau) = -\frac{1}{2}\cot(Y_\tau) + \frac{1}{4}\left(\rho(f^{-1}(\tau))\sin(y)(\alpha_2 + (2\alpha_1 - \alpha_2)\cos(y))\right)$$

is the infinitesimal mean of the transformed Wright-Fisher process and

$$b(y) = -\frac{1}{2y}$$

is the infinitesimal mean of the Bessel(0) process. However, as shown by Sermaidis et al. [2012], attempting to approximate the Itô integral in (22) using a discrete representation of the path can lead to biased estimates of the posterior distribution. Instead, consider the potential functions

$$\begin{aligned} H_1(y,\tau) &= \int^y a(\xi,\tau)\,d\xi \\ &= -\frac{1}{8}\left(\rho(f^{-1}(\tau))\cos^2(y)(2\alpha_1 - \alpha_2) + 4\log(\sin(y))\right) \end{aligned}$$

and

$$\begin{aligned} H_2(y) &= \int^y b(\xi,\tau)\,d\xi \\ &= -\frac{\log(y)}{2}. \end{aligned}$$

If we assume that ρ is continuous (not merely right continuous with left limits), then Itô's lemma shows that we can write

$$\begin{aligned} \int_0^{\tau_k} (\mu_1(Y_r,r) - \mu_2(Y_r))\,dY_r = H_1(Y_{\tau_k},\tau_k) - H_2(Y_{\tau_k}) - (H_1(Y_0,0) - H_2(Y_0)) \\ - \int_0^{\tau_k} \left(\frac{\partial H_1}{\partial \tau}(Y_r,r) - \frac{\partial H_2}{\partial \tau}(Y_r)\right) dr \\ - \int_0^{\tau_k} \left(\frac{\partial^2 H_1}{\partial y^2}(Y_r,r) - \frac{\partial^2 H_2}{\partial y^2}(Y_r)\right) dr. \end{aligned}$$

To generalize this to the case where ρ is right continuous with left limits, write

$$\int_0^{\tau_k} (a(Y_r,r) - b(Y_r))\,dY_r = I_0 + \sum_{i=m}^{K} I_i,$$

where m and K are defined in the main text,

$$I_0 = \lim_{\tau \uparrow f(d_m)} \int_0^{\tau} (a(Y_r,r) - b(Y_r))\,dY_r,$$

for $m < i < K$,

$$I_i = \lim_{\tau \uparrow f(d_{i+1})} \int_{f(d_i)}^{\tau} (a(Y_r,r) - b(Y_r))\,dY_r,$$

and

$$I_K = \lim_{\tau \uparrow \tau_k} \int_{f(d_K)}^{\tau} (a(Y_r, r) - b(Y_r))\, dY_r.$$

Itô's lemma can then be applied to each segment in turn. Following the conversion of the Itô integrals into ordinary Lebesgue integrals, making the substitution $s = f^{-1}(r)$ results in the path likelihood displayed in (11).

5.3. **Acceptance probability for an interior path update.** When we propose a new path $(y'_t)_{0 \le t \le \tau_k}$ to update the current path $(y_t)_{0 \le t \le \tau_k}$ which doesn't hit the boundary, the new path agrees with the existing path outside some time interval $[v_1, v_2]$, and has a new segment spliced in that goes from y_{v_1} at time v_1 to y_{v_2} at time v_2. The proposed new path segment comes from a Bessel(0) process over the time interval $[v_1, v_2]$ that is pinned to take the values y_{v_1} and y_{v_2} at the end-points; that is, the proposed new piece of path is a bridge.

The ratio that determines the probability of accepting the proposed path is

$$\frac{P(dD \,|\, y', t_0)}{P(dD \,|\, y, t_0)} \times \frac{\mathbb{P}(dy')\kappa(dy \,|\, y')}{\mathbb{P}(dy)\kappa(dy' \,|\, y)}, \tag{23}$$

where $P(\cdot \,|\, y', t_0)$ and $P(\cdot \,|\, y, t_0)$ give the probability of the observed allele counts given the transformed allele frequency paths and initial time t_0, $\mathbb{P}(\cdot)$ is the distribution of the transformed Wright-Fisher diffusion starting from $y_0 > 0$ at time 0 (that is, the distribution we have sometimes denoted by $\mathbb{P}^{y_0}$), the probability kernel $\kappa(\cdot \,|\, y)$ gives the distribution of the proposed path when the current path is y, and $\kappa(\cdot \,|\, y')$ is similar. To be completely rigorous, the second term in the product in (23) should be interpreted as the Radon-Nikodym derivative of two probability measures on the product of path space with itself.

Consider a finite set of times $0 \equiv \tau_0 \equiv u_0 < u_1 < \ldots < u_\ell \equiv \tau_k$. Suppose that $\{v_1, v_2\} \in \{u_0, \ldots, u_\ell\}$ $v_1 = u_m$ and $v_2 = u_n$ for some $m < n$. Let $(y_t)_{0 \le t \le \tau_k}$ and $(y'_t)_{0 \le t \le \tau_k}$ be two paths that coincide on $[0, v_1] \cup [v_2, \tau_k] = [u_0, u_m] \cup [u_n, u_\ell]$. Write $P(x, y; s, t)$ for the transition density (with respect to Lebesgue measure) of the transformed Wright-Fisher diffusion from time s to time t and $Q(x, y; t)$ for the transition density (with respect to Lebesgue measure) of the Bessel(0) process. Suppose that (ξ, ζ) is a pair of random paths with $P((\xi, \zeta) \in (dy, dy')) = \mathbb{P}(dy)\kappa(dy' \,|\, y)$. Then, writing $z_t = y_t = y'_t$ for $t \in [0, v_1] \cup [v_2, \tau_k] = [u_0, u_m] \cup [u_n, u_\ell]$, we have

$$\begin{aligned}
&P(\xi_{u_1} \in dy_{u_1}, \ldots, \xi_{u_\ell} \in dy_{u_\ell}, \zeta_{u_1} \in dy'_{u_1}, \ldots, \zeta_{u_\ell} \in dy'_{u_\ell}) \\
&\quad = P(z_{u_0}, z_{u_1}; u_0, u_1) dz_{u_1} \times \cdots \times P(z_{u_{m-1}}, z_{u_m}; u_{m-1}, u_m) dz_{u_m} \\
&\qquad \times P(z_{u_m}, y_{u_{m+1}}; u_m, u_{m+1}) dy_{u_{m+1}} \times \cdots \times P(y_{u_{n-1}}, z_{u_n}; u_{n-1}, u_n) dz_{u_n} \\
&\qquad \times P(z_{u_n}, z_{u_{n+1}}; u_n, u_{n+1}) dz_{u_{m+1}} \times \cdots \times P(z_{u_{\ell-1}}, z_{u_\ell}; u_{\ell-1}, u_\ell) dz_{u_\ell} \\
&\qquad \times Q(z_{u_m}, y'_{u_{m+1}}; u_{m+1} - u_m) dy_{u_{m+1}} \times \cdots \times Q(y_{u_{n-1}}, z_{u_n}; u_n - u_{n-1}) \\
&\qquad\quad \Big/ Q(z_{u_m}, z_{u_n}; u_n - u_m),
\end{aligned}$$

where the factor in the denominator arises because we are proposing *bridges* and hence conditioning on going from a fixed location at $v_1 = u_m$ to another fixed location at $v_2 = u_n$. Thus,

$$\frac{P(\xi_{u_1} \in dy'_{u_1}, \ldots, \xi_{u_\ell} \in dy'_{u_\ell}, \zeta_{u_1} \in dy_{u_1}, \ldots, \zeta_{u_\ell} \in dy_{u_\ell})}{P(\xi_{u_1} \in dy_{u_1}, \ldots, \xi_{u_\ell} \in dy_{u_\ell}, \zeta_{u_1} \in dy'_{u_1}, \ldots, \zeta_{u_\ell} \in dy'_{u_\ell})} = \frac{\prod_{j=m}^{n-1} P(y'_{u_j}, y'_{u_{j+1}}; u_j, u_{j+1})/Q(y'_{u_j}, y'_{u_{j+1}}; u_{j+1} - u_j)}{\prod_{j=m}^{n-1} P(y_{u_j}, y_{u_{j+1}}; u_j, u_{j+1})/Q(y_{u_j}, y_{u_{j+1}}; u_{j+1} - u_j)}.$$

Therefore, the Radon-Nikodym derivative appearing in (23) is the ratio of Radon-Nikodym derivatives

$$\frac{\frac{d\tilde{\mathbb{P}}}{d\tilde{\mathbb{Q}}}(y')}{\frac{d\tilde{\mathbb{P}}}{d\tilde{\mathbb{Q}}}(y)},$$

where $\tilde{\mathbb{P}}$ (resp. $\tilde{\mathbb{Q}}$) is the distribution of the transformed Wright-Fisher diffusion (resp. the Bessel(0) process) started at location $y_{v_1} = y'_{v_1}$ at time v_1 and run until time v_2. The formula (12) for the acceptance probability associated with an interior path update follows immediately.

The above argument was carried out under the assumption that the transformed initial allele frequency y_0 was strictly positive and so all the measures involved were probability measures. However, taking $y_0 \downarrow 0$ we see that the formula (12) continues to hold. Alternatively, we could have worked directly with the measure $\mathbb{P}^0$ in place of $\mathbb{P}^{y_0}$. The only difference is that we would have to replace $P(y_0, y; 0, s)$ by the density $\phi(y; 0, s)$ of an entrance law for $\mathbb{P}^0$. That is, $\phi(y; 0, s)$ has the property that

$$\lim_{y_0 \downarrow 0} \frac{P(y_0, y'; 0, s')}{P(y_0, y''; 0, s'')} = \frac{\phi(y'; 0, s')}{\phi(y''; 0, s'')}$$

for all $y', y'' > 0$ and $s', s'' > 0$ so that

$$\int \phi(y; 0, s) P(y, z; s, t)\, dy = \phi(z; 0, t)$$

for $0 < s < t$. Such a density, and hence the corresponding entrance law, is unique up to a multiplicative constant. In any case, it is clear that the choice of entrance law in the definition of $\mathbb{P}^0$ does not affect the formula (12) as the entrance law densities "cancels out".

5.4. **Acceptance probability for an allele age update.** The argument justifying the formula (13) for the probability of accepting a proposed update to the allele age t_0 is similar to the one just given for interior path updates. Now, however, we have to consider replacing a path y that starts from y_0 at time 0 and runs until time $f(t_k)$ with a path y' that starts from y_0 at time 0 and runs until time $f'(t_k)$. Instead of removing an internal segment of path and replacing it by one of the same length with the same values at the endpoints, we replace the initial segment of path that runs from time 0 to $f(t_s) = \int_{t_0}^{t_s} \frac{1}{\rho(s)}\, ds$ by one that runs from time 0 to time $f'(t_s) = \int_{t'_0}^{t_s} \frac{1}{\rho(s)}\, ds$, with $y'_{f'(t_s)} = y_{f(t_s)}$.

By analogy with the previous subsection, we need to consider

$$\frac{P(\xi \in dy', T_0^{\xi} \in dt', \zeta \in dy, T_0^{\zeta} \in dt)}{P(\xi \in dy, T_0^{\xi} \in dt, \zeta \in dy', T_0^{\zeta} \in dt')},$$

where ξ is a transformed Wright-Fisher process starting at y_0 at time 0 and run to time $F^{\xi} = \int_{T_0^{\xi}}^{t_s} \frac{1}{\rho(s)}\, ds$, where $P(T_0^{\xi} \in dt) = \rho(t)\, dt$, and conditional on ξ, ζ is a Bessel(0) bridge run from y_0 at time 0 to $\xi_{F^{\xi}}$ at time $F^{\zeta} = \int_{T_0^{\zeta}}^{t_s} \frac{1}{\rho(s)}\, ds$, where $P(T_0^{\zeta} \in dt) = \rho(t)dt$ independent of ξ and T_0^{ξ}.

Suppose that $0 = u_0 < u_1 < \ldots < u_m = \int_{t'}^{t_s} \frac{1}{\rho(s)}\, ds$ and $0 = v_0 < v_1 < \ldots < v_n = \int_{t}^{t_s} \frac{1}{\rho(s)}\, ds$. We have for $y'_0, \ldots y'_m$ and $y_0, \ldots, y_n$ with $y_0 = y'_0$ and $y'_m = y_n$ that

$$
\begin{aligned}
&\frac{P(\xi_{u_i} \in dy'_i, 1 \le i \le m-1, T_0^{\xi} \in dt', \zeta_{v_j} \in dy_j, 1 \le j \le n, T_0^{\zeta} \in dt)}{P(\xi_{v_j} \in dy_j, 1 \le j \le n-1, T_0^{\xi} \in dt, \zeta_{u_i} \in dy'_i, 1 \le i \le m, T_0^{\zeta} \in dt')} \\
&\quad = \left\{ \prod_{i=0}^{m-1} P(y'_j, y'_{j+1}; u_i, u_{i+1})\, dy'_{i+1} \times \rho(t')\, dt' \right. \\
&\qquad \left. \times \left[\prod_{j=0}^{n-2} Q(y_j, y_{j+1}; v_{j+1} - v_j)\, dy_{j+1} \times Q_{(}(y_{n-1}, y_n; v_n - v_{n-1}) \Big/ Q_{(}(y_0, y_n; v_n) \right] \times dt \right\} \\
&\qquad \Big/ \left\{ \prod_{j=0}^{n-1} P(y_j, y_{j+1}; v_j, v_{j+1})\, dy_{j+1}) \times \rho(t)\, dt \right. \\
&\qquad\quad \left. \times \left[\prod_{i=0}^{m-2} Q(y'_i, y'_{i+1}; u_{i+1} - u_i)\, dy'_{i+1} \times Q_{(}(y'_{m-1}, y'_m; u_m - u_{m-1}) \Big/ Q_{(}(y'_0, y'_m; u_m) \right] \times dt' \right\} \\
&\quad = \left\{ \prod_{i=0}^{m-1} P(y'_j, y'_{j+1}; u_i, u_{i+1})\, dy'_{i+1} \times \rho(t')\, dt' \right. \\
&\qquad \left. \times \left[\prod_{j=0}^{n-1} Q(y_j, y_{j+1}; v_{j+1} - v_j)\, dy_{j+1} \Big/ Q_{(}(y_0, y_n; v_n) \right] \times dt \right\} \\
&\qquad \Big/ \left\{ \prod_{j=0}^{n-1} P(y_j, y_{j+1}; v_j, v_{j+1})\, dy_{j+1}) \times \rho(t)\, dt \right. \\
&\qquad\quad \left. \times \left[\prod_{i=0}^{m-1} Q(y'_i, y'_{i+1}; u_{i+1} - u_i)\, dy'_{i+1} \Big/ Q_{(}(y'_0, y'_m; u_m) \right] \times dt' \right\} \\
&\quad = \frac{\prod_{i=0}^{m-1} P(y'_j, y'_{j+1}; u_i, u_{i+1})\, dy'_{i+1} \Big/ \left[\prod_{i=0}^{m-1} Q(y'_i, y'_{i+1}; u_{i+1} - u_i)\, dy'_{i+1} \right]}{\prod_{j=0}^{n-1} P(y_j, y_{j+1}; v_j, v_{j+1})\, dy_{j+1}) \Big/ \left[\prod_{j=0}^{n-1} Q(y_j, y_{j+1}; v_{j+1} - v_j)\, dy_{j+1} \right]} \\
&\qquad \times \frac{Q_{(}(y'_0, y'_m; u_m)}{Q_{(}(y_0, y_n; v_n)} \times \frac{\rho(t')}{\rho(t)},
\end{aligned}
$$

where the second equality follows from the fact that $y_n = y'_m$.

Thus,

$$\frac{P(\xi \in dy', T_0^{\xi} \in dt', \zeta \in dy, T_0^{\zeta} \in dt)}{P(\xi \in dy, T_0^{\xi} \in dt, \zeta \in dy', T_0^{\zeta} \in dt')} = \frac{\frac{d\check{\mathbb{P}}}{d\check{\mathbb{Q}}}(y')}{\frac{d\hat{\mathbb{P}}}{d\hat{\mathbb{Q}}}(y)} \times \frac{Q(y_0, y'_{T'}; T')}{Q(y_0, y_T; T)} \times \frac{\rho(t')}{\rho(t)},$$

where $T = \int_t^{t_s} \frac{1}{\rho(s)}\, ds$ and $T' = \int_{t'}^{t_s} \frac{1}{\rho(s)}\, ds$, $\hat{\mathbb{P}}$ (resp. $\check{\mathbb{P}}$) is the distribution of the transformed Wright-Fisher diffusion starting at location y_0 at time 0 and run until time T (resp. T'), and $\hat{\mathbb{Q}}$ (resp. $\check{\mathbb{Q}}$) is the distribution of the Bessel(0) process starting at location y_0 at time 0 and run until time T (resp. T').

We have thusfar assumed that y_0 is strictly positive. As in the previous subsection, we can let $y_0 \downarrow 0$ to get an expression in terms of Radon-Nikodym derivatives of σ-finite measures and the density $\psi(y; s)$ of an entrance law for $\mathbb{Q}^0$. That is, $\psi(y; s)$ has the property that

$$\lim_{y_0 \downarrow 0} \frac{Q(y_0, y'; s')}{Q(y_0, y''; s'')} = \frac{\psi(y'; s')}{\psi(y''; s'')}$$

for all $y', y'' > 0$ and $s', s'' > 0$, so that

$$\int \psi(y; s) Q(y, z; t)\, dy = \psi(z; s + t)$$

for $s, t > 0$. Up to an irrelevant multiplicative constant, ψ is given by the expression (14), and the formula (13) for the acceptance probability follows immediately.

5.5. **Acceptance probability for a most recent allele frequency update.** The derivation of formula (15) for the probability of accepting a proposed update to the most recent allele frequency is similar to those for the other acceptance probabilities (12) and (13), so we omit the details.

6. Supplementary Figures

[Figure 1 about here.]

[Figure 2 about here.]

References

Jonathan P Bollback, Thomas L York, and Rasmus Nielsen. Estimation of 2nes from temporal allele frequency data. *Genetics*, 179(1):497–502, 2008.

Hua Chen and Montgomery Slatkin. Inferring selection intensity and allele age from multilocus haplotype structure. *G3: Genes— Genomes— Genetics*, 3(8):1429–1442, 2013.

Graham Coop and Robert C Griffiths. Ancestral inference on gene trees under selection. *Theoretical population biology*, 66(3):219–232, 2004.

Clio Der Sarkissian et al. Evolutionary genomics and conservation oft he endangered Przewalski's horse. *Current Biology*, 2015.

606 Warren J Ewens. *Mathematical population genetics: I. Theoretical introduction*, volume 27.
607 Springer, 2004.
608 Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of
609 selection in genetic time series. *Genetics*, 196(2):509–522, 2014.
610 William Feller. Diffusion processes in genetics. In *Proc. Second Berkeley Symp. Math.*
611 *Statist. Prob*, volume 227, page 246, 1951.
612 Ronald Aylmer Fisher. On the dominance ratio. *Proceedings of the royal society of Edin-*
613 *burgh*, 42:321–341, 1922.
614 Christiane Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*.
615 Springer, 2013.
616 IV Girsanov. On transforming a certain class of stochastic processes by absolutely contin-
617 uous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301,
618 1960.
619 Andrew Golightly and Darren J Wilkinson. Bayesian inference for stochastic kinetic models
620 using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.
621 Andrew Golightly and Darren J Wilkinson. Bayesian inference for nonlinear multivariate
622 diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):
623 1674–1693, 2008.
624 John Burdon Sanderson Haldane. A mathematical theory of natural and artificial selection,
625 part v: selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical*
626 *Society*, 23(07):838–844, 1927.
627 Kiyosi Itô. Stochastic integral. *Proceedings of the Japan Academy, Series A, Mathematical*
628 *Sciences*, 20(8):519–524, 1944.
629 Paul A Jenkins. Exact simulation of the sample paths of a diffusion with a finite entrance
630 boundary. *arXiv preprint arXiv:1311.5777*, 2013.
631 Paul A Jenkins and Dario Spano. Exact simulation of the wright-fisher diffusion. *arXiv*
632 *preprint arXiv:1506.06998*, 2015.
633 Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New
634 York). Springer-Verlag, New York, second edition, 2002. doi: 10.1007/978-1-4757-4015-8.
635 URL `http://dx.doi.org/10.1007/978-1-4757-4015-8`.
636 Frank B Knight. *Essentials of Brownian motion and diffusion*, volume 18 of *Mathematical*
637 *Surveys*. American Mathematical Society, Providence, R.I., 1981.
638 Arne Ludwig, Melanie Pruvost, Monika Reissmann, Norbert Benecke, Gudrun A Brock-
639 mann, Pedro Castaños, Michael Cieslak, Sebastian Lippold, Laura Llorente, Anna-Sapfo
640 Malaspinas, et al. Coat color variation at the beginning of horse domestication. *Science*,
641 324(5926):485–485, 2009.
642 Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin.
643 Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):
644 599–607, 2012.
645 Iain Mathieson and Gil McVean. Estimating selection coefficients in spatially structured
646 populations from time series data of allele frequencies. *Genetics*, 193(3):973–984, 2013.
647 Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson,
648 Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes,

649 Mario Novak, et al. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*,
650 528(7583):499–503, 2015.
651 Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and
652 Carlos Bustamante. Genomic scans for selective sweeps using snp data. *Genome research*,
653 15(11):1566–1575, 2005.
654 Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin
655 Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman,
656 et al. Signals of recent positive selection in a worldwide sample of human populations.
657 *Genome research*, 19(5):826–837, 2009.
658 Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Conver-
659 gence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006. URL
660 `http://CRAN.R-project.org/doc/Rnews/`.
661 Gareth O Roberts and Osnat Stramer. On inference for partially observed nonlinear diffu-
662 sion models using the metropolis–hastings algorithm. *Biometrika*, 88(3):603–621, 2001.
663 Joshua G Schraiber. A path integral formulation of the wright–fisher process with genic
664 selection. *Theoretical population biology*, 92:30–35, 2014.
665 Joshua G. Schraiber, Robert C. Griffiths, and Steven N. Evans. Analysis and rejection
666 sampling of Wright-Fisher diffusion bridges. *Theoretical Population Biology*, 89(0):64–
667 74, 2013.
668 Giorgos Sermaidis, Omiros Papaspiliopoulos, Gareth O Roberts, Alexandros Beskos, and
669 Paul Fearnhead. Markov chain monte carlo for exact inference for diffusions. *Scandina-*
670 *vian Journal of Statistics*, 2012.
671 Per Sjödin, Pontus Skoglund, and Mattias Jakobsson. Assessing the maximum contribution
672 from ancient populations. *Molecular biology and evolution*, page msu059, 2014.
673 Montgomery Slatkin. Simulating genealogies of selected alleles in a population of variable
674 size. *Genetical research*, 78(01):49–57, 2001.
675 Yun S Song and Matthias Steinrücken. A simple method for finding explicit analytic
676 transition densities of diffusion processes with general diploid selection. *Genetics*, 190
677 (3):1117–1129, 2012.
678 Michael Sørensen. Parametric inference for discretely sampled stochastic differential equa-
679 tions. In *Handbook of financial time series*, pages 531–553. Springer, 2009.
680 Matthias Steinrücken, Anand Bhaskar, and Yun S Song. A novel spectral method for
681 inferring general diploid selection from time series genetic data. *The annals of applied*
682 *statistics*, 8(4):2203, 2014.
683 Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A
684 map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72, 2006.
685 GA Watterson. Estimating and testing selection: the two-alleles, genic selection diffusion
686 model. *Advances in Applied Probability*, pages 14–30, 1979.
687 Ellen G Williamson and Montgomery Slatkin. Using maximum likelihood to estimate
688 population size from temporal changes in allele frequencies. *Genetics*, 152(2):755–761,
689 1999.

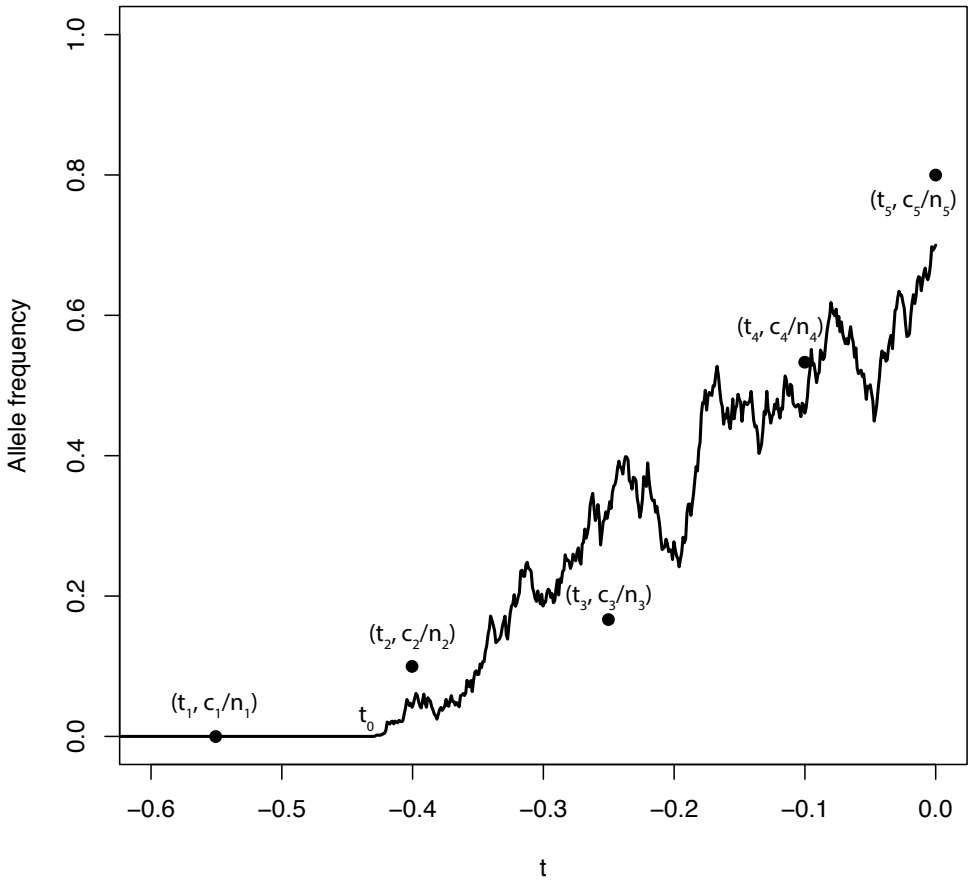

FIGURE 1. Taking samples from an allele frequency trajectory. An allele frequency trajectory is simulated from the Wright-Fisher diffusion (solid line). At each time, t_i, a sample of size n_i chromosomes is taken and c_i copies of the derived allele are observed. Each point corresponds to the observed allele frequency of sample i. Note that t_1 is more ancient than the allele age, t_0.

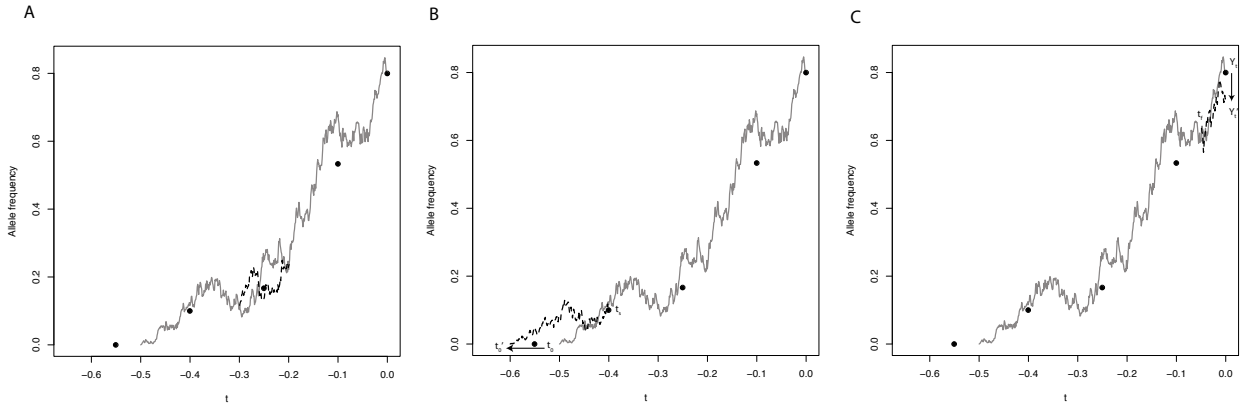

FIGURE 2. Illustration of path updates. Filled circles correspond to the same sample frequencies as in Figure 1. The solid gray line in each panel is the current allele frequency trajectory and the dashed black lines are the proposed updates. In panel a, an interior section of path is proposed between points s_1 and s_2. In panel b, a new allele age, t_0' is proposed and a new path is drawn between t_0' and t_s. In panel c, a new most recent allele frequency Y_{t_k}' is proposed and a new path is drawn between t_f and t_k.

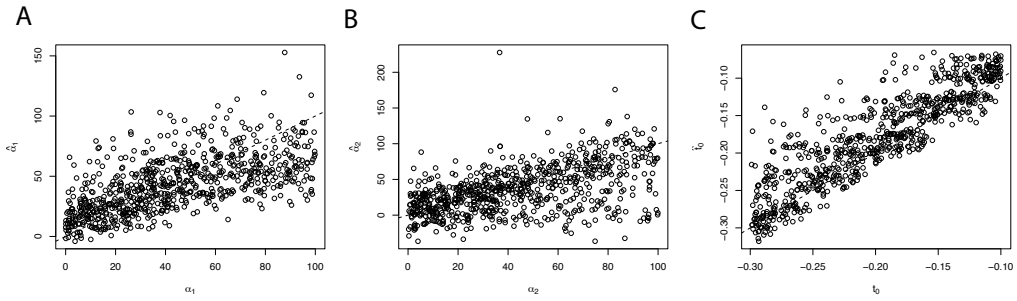

FIGURE 3. Maximum *a posterior* estimates of different parameters. Each panel shows the true value of a parameter on the x-axis, while the inferred value is on the y-axis. Dashed line is $y = x$.

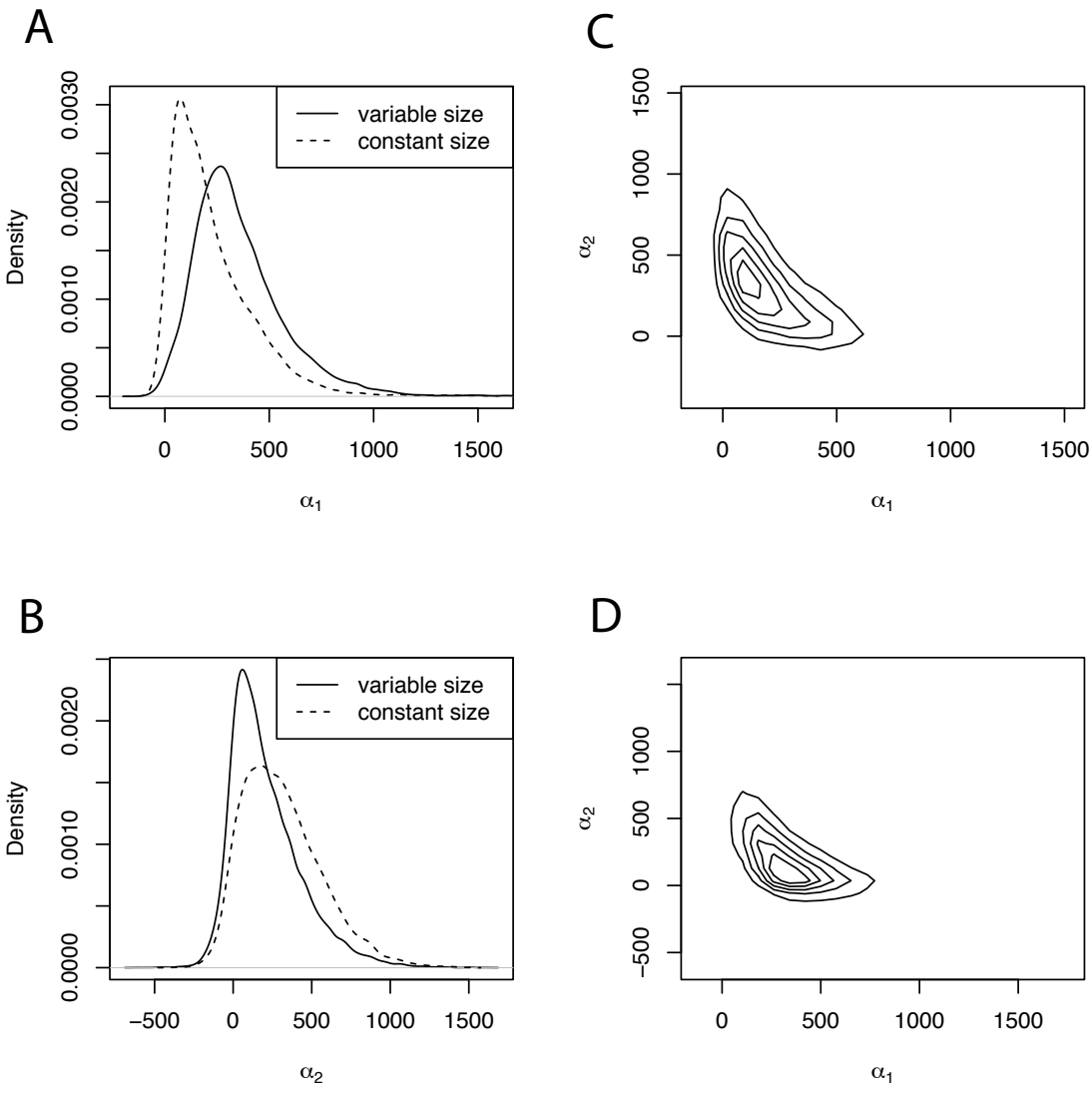

FIGURE 4. Posterior distributions of selection coefficients for the MC1R locus. Panels A and B show marginal distributions of α_1 and α_2, respectively, with the solid line indicating the posterior obtained from an analysis including the full demographic history, and the dotted line showing what would be inferred in a constant size population. Panels C and D show contour plots of the joint distribution of α_1 and α_2 without and with demography, respectively.

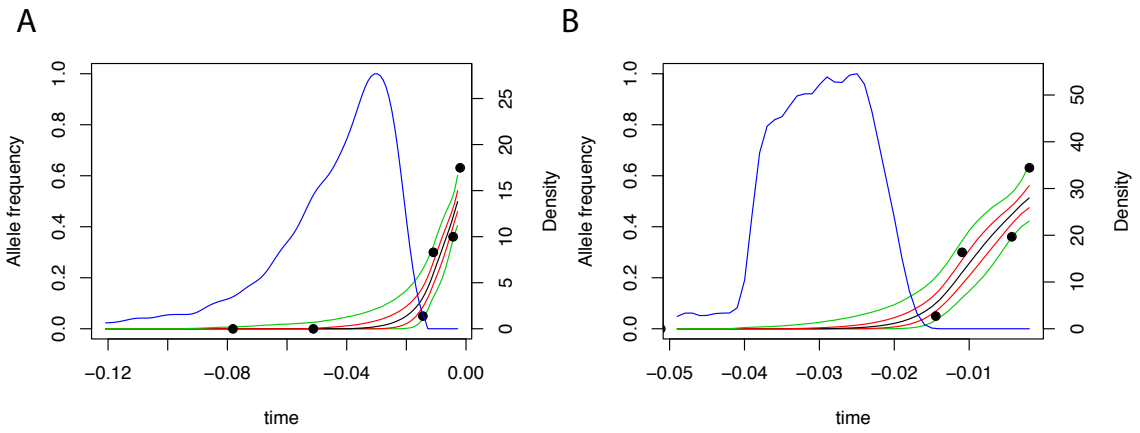

FIGURE 5. Posterior distribution on allele frequency paths for the MC1R locus. Each panel shows the sampled allele frequency data (filled circles), the point-wise median (black), 25 and 75% quantiles (red), and 5 and 95% quantiles (green) of the posterior distribution on paths, and the posterior distribution on allele age (blue). Panel A reports inference with constant demography, while panel B shows the result of inference with the full demographic history.

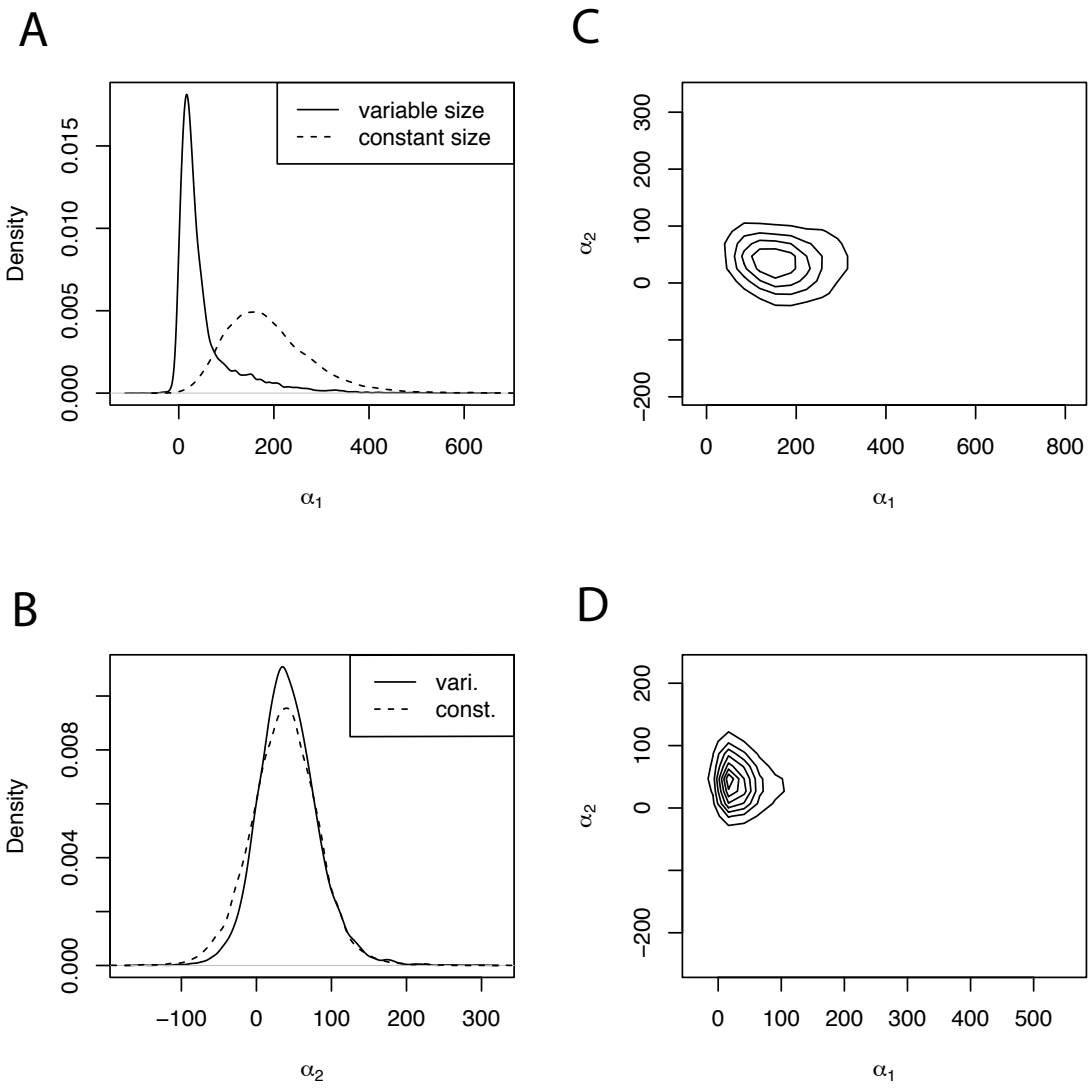

FIGURE 6. Posterior distributions of selection coefficients for the ASIP locus. Panels as in Figure 4

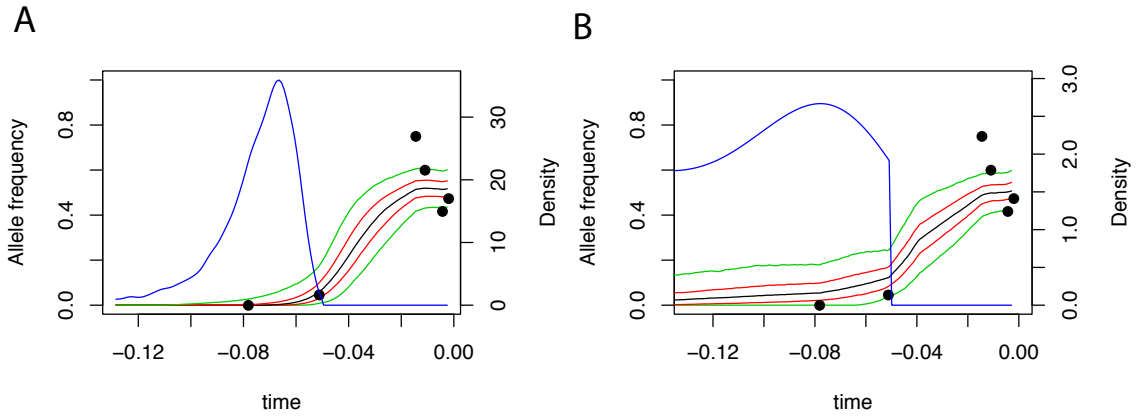

FIGURE 7. Posterior distribution on allele frequency paths for the ASIP locus. Panels are as in Figure 5.

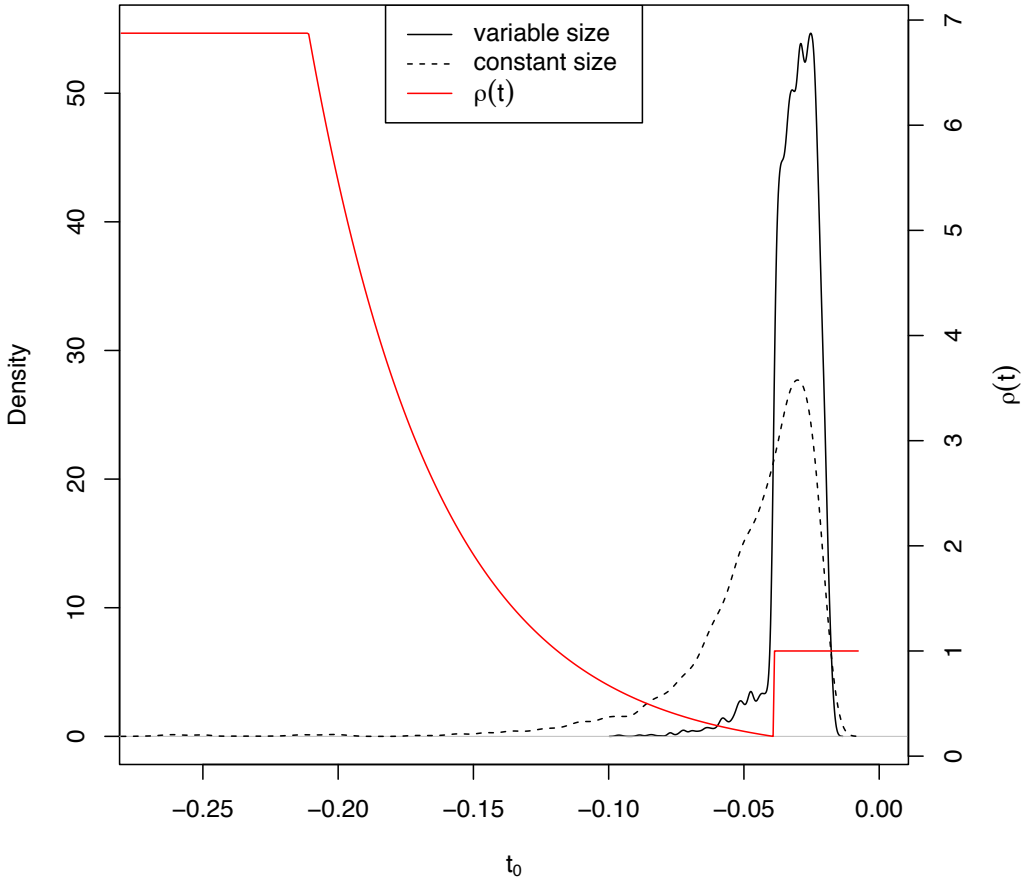

FIGURE 8. Influence of population size on age estimates of the MC1R locus. The solid and dashed lines show the posterior distribution on allele age with and without demography, respectively. In red, the demographic history inferred by Der Sarkissian et al. [2015].

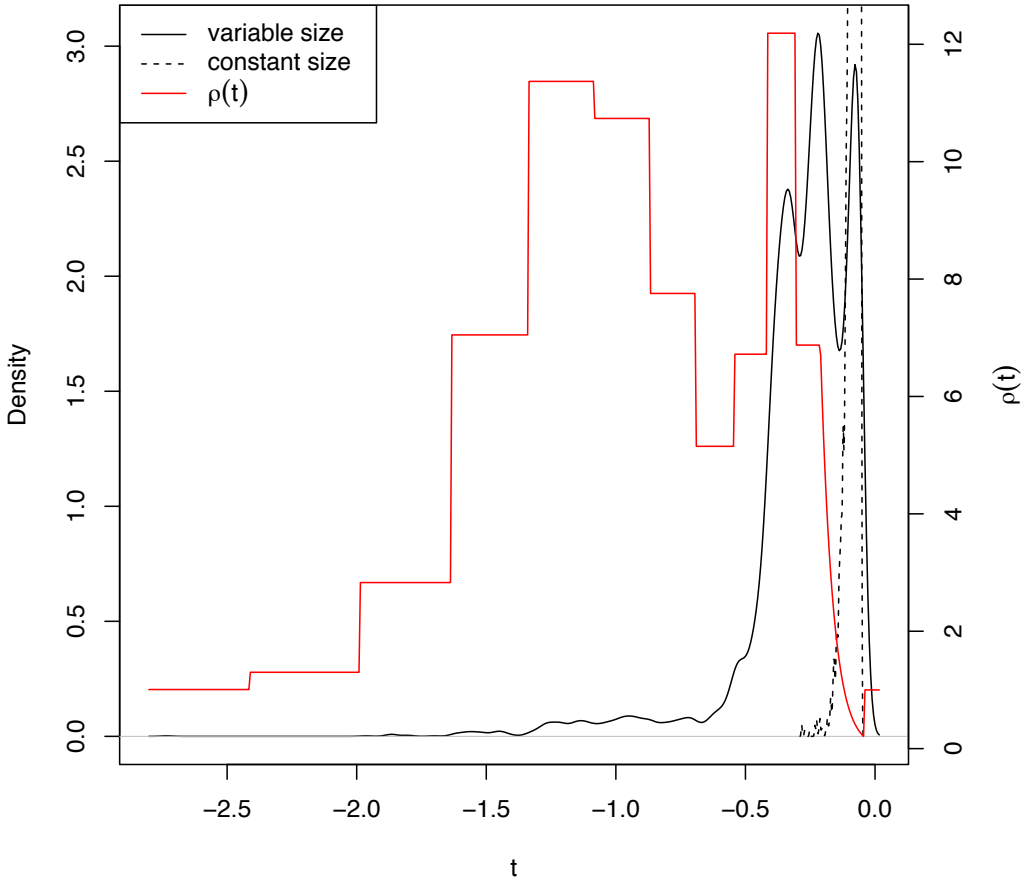

FIGURE 9. Influence of population size on age estimates of the ASIP locus. Data presented is as in Figure 8

Genotype	A_1A_1	A_1A_0	A_0A_0
Fitness	$1+s_2$	$1+s_1$	1

TABLE 1. Fitness scheme assumed in the text.

Sample time (years BCE)	20,000	13,100	3,700	2,800	1,100	500
Sample time (diffusion units)	0.078	0.051	0.014	0.011	0.004	0.002
Sample size	10	22	20	20	36	38
Count of ASIP alleles	0	1	15	12	15	18
Count of MC1R alleles	0	0	1	6	13	24

TABLE 2. Sample information for horse data. Diffusion time units are calculated assuming $N_0 = 2500$ and a generation time of 5 years.