A likelihood method for jointly estimating the selection coefficient and the allele age for time serial data

Anna-Sapfo Malaspinas[1], Orestis Malaspinas[2], Steven N. Evans[3], and Montgomery Slatkin[4]

[1]Centre for Geogenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen, Denmark

[2]Institut Jean le Rond d'Alembert - Université Pierre et Marie Curie, 4 place Jussieu - case 162, F-75252 Paris cedex 5, France Comp. Sci. Dept., Centre Universitaire d'Informatique, Université de Genève, CUI - SPC, 1227 Carouge, Switzerland

[3]Department of Statistics, 367 Evans Hall, University of California, Berkeley CA 94720-3860, USA

[4]Department of Integrative Biology, University of California, 3060 Valley Life Sciences Bldg, Berkeley, CA 94720-3140, USA

running title: Inferring allele age and selection coefficient

keywords: allele age, ancient DNA, population genetics, selection, time-serial data

corresponding author:

Anna-Sapfo Malaspinas

mailing-address: Centre for Geogenetics, Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark,

tel: +4535321225

e-mail: anna.sapfo.malaspinas@snm.ku.dk

Abstract

Recent advances in sequencing technologies have made available an ever-increasing amount of ancient genomic data. In particular, it is now possible to target specific single nucleotide polymorphisms in several samples at different time points. Such time series data are also available in the context of experimental or viral evolution. Time-series data should allow for a more precise inference of population genetic parameters, and to test hypotheses about the recent action of natural selection. In this manuscript, we develop a likelihood method to jointly estimate the selection coefficient and the age of an allele from time serial data. Our method can be used for allele frequencies sampled from a single diallelic locus. The transition probabilities are calculated by approximating the standard diffusion equation of the Wright-Fisher model with a one step process. We show that our method produces unbiased estimates. The accuracy of the method is tested via simulations. Finally, the utility of the method is illustrated with an application to several loci encoding coat color in horses, a pattern that has previously been linked with domestication. Importantly, given our ability to estimate the age of the allele, it is possible to gain traction on the important problem of distinguishing selection on new mutations from selection on standing variation. In this coat color example for instance, we estimate the age of this allele, which is found to predate domestication.

1 Introduction

Time series analysis is widespread in several fields, such as meteorology, economics and physics (e.g. Hamilton (1994)) with the relation being statistical models designed to deal with a time ordered sequence of observations. Such observations are also prevalent in several areas of biology. Until recently, however, time series molecular data were only available for time spanning a few generations in higher organisms. Therefore, in the context of population genetics, time serial data were mostly limited to viral or experimental evolution (e.g. Wichman et al. (2005); Bollback and Huelsenbeck (2007); Nelson and Holmes (2007); Gresham et al. (2008)).

However, with recent advances in DNA sequencing and DNA preparation techniques, the study of extinct and long dead organisms is now entering a new era, an era in which time-sampled measurements may be obtained spanning hundreds or thousands of generations for even mammalian species. For example, while previous studies were limited to short segments of mitochondrial DNA, whole nuclear genomes are now available from several extinct species, (e.g. Rasmussen et al. (2010); Reich et al. (2010)) and it is now additionally possible to target specific DNA regions in ancient organisms (e.g. Lalueza-Fox et al. (2007); Ludwig et al. (2009); Rusk (2009)). Therefore, time serial data will become increasingly available for a whole range of organisms allowing one to test evolutionary questions using not only present day samples, but also samples from extinct populations.

The relevant theory to describe such temporal changes in allele frequency has existed since the advent of population genetics (e.g. Fisher (1922); Wright (1931)). Although not very common, several statistical methods and estimators to deal with time serial data have been developed and applied to, for example, estimate historical changes in population size (e.g. Waples (1989); Williamson and Slatkin (1999); Anderson et al. (2000); Drummond and Rambaut (2007)). More recently, in 2008, Bollback et al. developed a method to co-estimate the effective population size, N_e, and the selection coefficient, s, from temporal allele frequency data. They model the evolution of the allele frequency of a diallelic locus

with a diffusion process that approximates a Wright-Fisher population genetic model (WF), under the assumption that the locus is under constant natural selection acting on diploid individuals.

Our work is a natural extension of Bollback et al.'s method to also allow for the estimation of the allele age (i.e., the time since the mutational event), t_0. Allele age is an omnipresent parameter in population genetics and along with the selection coefficient it plays a crucial role in determining the sojourn time of a beneficial mutation (see Slatkin and Rannala (2000) for a review). Additionally, given the recent focus on the important question of distinguishing between models of selection on new versus standing mutations - a phenomenon which speaks to the fundamental mode and tempo of the process of adaptation - the ability to estimate the time of a mutational event is of paramount importance (see review of Barrett and Schluter (2008)).

Our extension allows these competing models to be addressed, unlike the previous work of Bollback et al. (2008) that assumed that at the first time of sampling the population allele frequency was uniformly distributed. It follows from this latter assumption that even if the allele was not sampled at the oldest sampling time, it had to be present in the population. Here, we present an approach to co-estimate s, N_e and t_0 by computing the likelihood of these parameters for a suitable model.

In Section 2.1 we explain how we approximate the WF model with a one step process. We then discuss the numerical details of the implementation in Sections 2.2 and 3.1. We show how our method performs based on simulations in Sections 2.3 and 3.2. We analyze a dataset of horses for the *ASIP* locus for samples dating from the Pleistocene up to the present in Sections 2.4 and 3.3. We conclude and offer some further perspectives in Section 4.

2 Materials and Methods

2.1 Theory

We assume that there is a single, panmictic population evolving according to a WF population genetic model. Under this model, the frequency of an allele A is a homogeneous discrete-time Markov chain. We denote the Markov chain describing the frequency of the allele A through time by X_t. We assume that selection is constant from the time the allele arose up to present. The allele under selection arises only once and there is no recurrent mutation. In other words, the only evolutionary forces acting on that allele are genetic drift and selection.

Selection is modeled as acting on diploid individuals. If we denote the two alleles by A and a, we can choose the genotypic fitness to be $w_{AA} = 1 + s$, $w_{Aa} = 1 + sh$ and $w_{aa} = 1$ where s is the selection coefficient and h is the dominance coefficient ($s > -1$ and $h \in [0, 1]$, see e.g. Ewens (2004)). If N_e is the effective population size, the states of X_t are the allelic frequencies, with respect to the population size $x_j = \frac{j}{2N_e}$ for $0 \leq j \leq 2N_e$. Therefore, the state space is $\{0, \frac{1}{2N_e}, ..., \frac{2N_e-1}{2N_e}, 1\}$. We define the rescaled selection coefficient $\gamma = 2N_e s$.

We would like to compute the likelihood of the allele age t_0, the rescaled selection coefficient γ, and the effective population size N_e. To simplify the notation, define $\theta \equiv (\gamma, N_e, t_0)$ the parameters of interest. Assume that we have samples from m distinct sampling time points. We suppose that $M = (n_1, n_2, \ldots, n_m)$ chromosomes were collected, among which $I = (i_1, i_2, \ldots, i_m)$ are of the A type, and that the chromosomes where drawn at times $T = (t_1, t_2, \ldots, t_m)$, where time is measured in generations with $t_{k-1} < t_k$ (see Figure 1). The likelihood function of the parameters, for a given M and h, is $\ell(\theta) = p(i_1, \ldots, i_m | \theta, T)$.

To compute the likelihood, we can condition and sum over all the population allelic frequencies, $x_{j_1}, \ldots, x_{j_m}$, at each sampling time $t_1, t_2, \ldots, t_m$. We can then rewrite the likelihood:

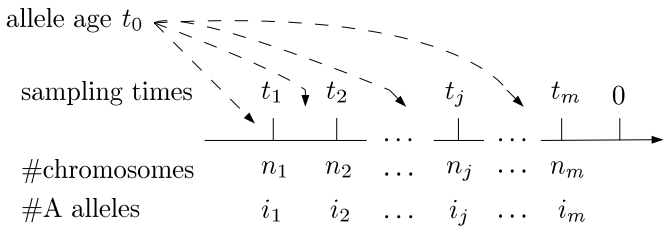

Figure 1: Notation used throughout the text. The chromosomes $M = (n_1, n_2, \ldots, n_m)$ are sampled at times $T = (t_1, t_2, \ldots, t_m)$ and there are $I = (i_1, i_2, \ldots, i_m)$ A alleles at each sampling time.

$$\ell(\theta) = \sum_{j_1} \ldots \sum_{j_m} p(i_1, \ldots, i_m | \theta, T, x_{j_1}, x_{j_2}, \ldots, x_{j_m}) p(x_{j_1}, x_{j_2}, \ldots, x_{j_m} | \theta, T). \quad (1)$$

Conditional on the population allelic frequencies, the number of A alleles i_j at each sampling time are independent of one other. The first term of the summation of equation 1 becomes

$$p(i_1, \ldots, i_m | \theta, T, x_{j_1}, x_{j_2}, \ldots, x_{j_m}) = p(i_1 | x_{j_1}) \ldots p(i_m | x_{j_m}). \quad (2)$$

In the WF model the population is large and panmictic, therefore we can assume that we sample the chromosomes with replacement, for $k \in \{0, .., m\}$ and write:

$$p(i_k | x_{j_k}) = \begin{pmatrix} n_k \\ i_k \end{pmatrix} x_{j_k}^{i_k} (1 - x_{j_k})^{n_k - i_k}. \quad (3)$$

Since X_t is a Markov chain, the second term of the summation of equation 1 is given by:

$$p(x_{j_1}, x_{j_2}, \ldots, x_{j_m} | \theta, T) = p(x_{j_m} | x_{j_{m-1}}, \theta, T) p(x_{j_{m-1}} | x_{j_{m-2}}, \theta, T) \ldots p(x_{j_1} | x_{j_0}, \theta, T), \quad (4)$$

where x_{j_0} is the frequency of the allele when it first arose in the population, i.e. $x_{j_0} = \frac{1}{2N_e}$. We can rewrite the transition probabilities of X_t $p(x_{j_k} | x_{j_{k-1}}, \theta, T) = p_{t_k - t_{k-1}}(x_{j_{k-1}}, x_{j_k})$, for a given θ and T. By substituting equation 2 and 4 into 1 we get:

$$\ell(\theta) = p(i_1, \ldots, i_m|\theta, T) =$$
$$\sum_{j_m=0}^{2N_e} p(i_m|\frac{j_m}{2N_e}) \sum_{j_{m-1}=0}^{2N_e} p_{t_m - t_{m-1}}(\frac{j_{m-1}}{2N_e}, \frac{j_m}{2N_e}) \cdot$$
$$\sum_{j_{m-1}=0}^{2N_e} p(i_{m-1}|\frac{j_{m-1}}{2N_e}) \sum_{j_{m-2}=0}^{2N_e} p_{t_{m-1} - t_{m-2}}(\frac{j_{m-2}}{2N_e}, \frac{j_{m-1}}{2N_e}) \cdots$$
$$p(i_2|\frac{j_2}{2N_e}) \sum_{j_1=0}^{2N_e} p_{t_2 - t_1}(\frac{j_1}{2N_e}, \frac{j_2}{2N_e}) \cdot$$
$$p(i_1|\frac{j_1}{2N_e}) p_{t_1 - t_0}(\frac{1}{2N_e}, \frac{j_1}{2N_e}). \quad (5)$$

The solution for the transition probabilities for the non-neutral case of the WF model is elaborate (Ewens (2004) and citations therein). But if we rescale the time by $2N_e$, the Markov chain, X_t, can be approximated by a diffusion process ("WF diffusion process"), Y_τ (see e.g. Durrett (2008)). Time is now in units of $2N_e$ generations and is continuous and we replace T by $\mathcal{T} = (\tau_1, \ldots, \tau_m)$ where $\tau_i = \frac{t_i}{2N_e}$. The state space is also continuous with states denoted by $y \in [0, 1]$. This holds in the limit of large N_e, where $X_{[\tau 2N_e]} \simeq Y_\tau$. The transition densities of the diffusion process are denoted $p(y_k|y_{k-1}, \theta, \mathcal{T}) = p_{\tau_k - \tau_{k-1}}(y_{k-1}, y_k)$. In this paper we further approximate the diffusion process itself by a one step process that we denote by Z_τ (see e.g. Van Kampen (1992)). A one step process is a continuous-time Markov chain (i.e. discrete in space and continuous in time) where jumps are only allowed between two states that are adjacent to each other. As before, the states of the process Z_τ are certain population allelic frequencies that we denote by $\{z_0, z_1, \ldots, z_{H-1}\}$, where H is an integer. The states are chosen such that z_0 and z_{H-1} are respectively the 0 and 1 allelic frequencies. These are absorbing states since there is no recurrent mutation. The other states are chosen such that $0 < z_k < 1$ and $z_{k-1} < z_k$ for $0 < k < H - 1$. The infinitesimal generator Q of such a process is a tridiagonal $H \times H$ matrix. By denoting β_i (respectively δ_i) the rate of jumping to the right (respectively the left) of state i, we have that:

$$Q = \begin{pmatrix} 0 & & & \dots & & & 0 \\ \delta_1 & \eta_1 & \beta_1 & 0 & & & \\ 0 & \ddots & \ddots & \ddots & 0 & \vdots & \vdots \\ & 0 & \delta_k & \eta_k & \beta_k & 0 & \\ \vdots & & 0 & \ddots & \ddots & \ddots & 0 \\ & & & 0 & \delta_{H-2} & \eta_{H-2} & \beta_{H-2} \\ 0 & & & \dots & & 0 & 0 \end{pmatrix} \tag{6}$$

where $\eta_k = -(\beta_k + \delta_k)$. The transition probability between two states $z_{j_{k-1}}$ and z_{j_k} of the process is $p_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k}) = (\exp(Q(\tau_{k+1} - \tau_k)))_{j_{k-1}, j_k}$. With the appropriate choice of β_i and δ_i (see Supplementary Material A), one can show that for large H, $Z_\tau \simeq Y_\tau$. In particular, β_i and δ_i will be functions of z_j, z_{j-1}, z_{j+1}, γ and h. Note that Y_τ is a continuous variable whereas Z_τ is discrete. Therefore, choosing $y_{k-1} = z_{j_{k-1}}$ and $y_k = z_{j_k} \notin \{0, 1\}$ we have that:

$$p_{\tau_k - \tau_{k-1}}(y_{k-1}, y_k) \backsimeq \frac{p_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k})}{\left(\frac{z_{j_k+1} - z_{j_k-1}}{2}\right)} = \frac{(\exp(Q(\tau_k - \tau_{k-1})))_{j_{k-1}, j_k}}{\left(\frac{z_{j_k+1} - z_{j_k-1}}{2}\right)}, \tag{7}$$

where the denominator is necessary since Y_τ has a continuous state space and Z_τ has a discrete state space. We can approximate the likelihood described in equation 5 by replacing the original process X_t by the one step process Z_τ. We then have:

$$\ell(\theta) = p(i_1, \ldots, i_m|\theta, \mathcal{T}) =$$
$$\sum_{j_m=0}^{H-1} p(i_m|z_{j_m}) \sum_{j_{m-1}=0}^{H-1} p_{\tau_m-\tau_{m-1}}(z_{j_{m-1}}, z_{j_m}) \cdot$$
$$\sum_{j_{m-1}=0}^{H-1} p(i_{m-1}|z_{j_{m-1}}) \sum_{j_{m-2}=0}^{H-1} p_{\tau_{m-1}-\tau_{m-2}}(z_{j_{m-2}}, z_{j_{m-1}}) \cdots$$
$$p(i_2|z_{j_2}) \sum_{j_1=0}^{H-1} p_{\tau_2-\tau_1}(z_{j_1}, z_{j_2}) \cdot$$
$$p(i_1|z_{j_1}) p_{\tau_1-\tau_0}(\frac{1}{2N_e}, z_{j_1}). \qquad (8)$$

where $p(i_k|z_{j_k}) = \begin{pmatrix} n_k \\ i_k \end{pmatrix} z_{j_k}^{i_k}(1 - z_{j_k})^{n_k - i_k}$ from equation 3.

In the case of experimental evolution this unconditional process should be realistic since in principle one might want to estimate the selection coefficient for any locus. We will now consider one special case of what is presented above, motivated by ancient DNA data. We will assume that the allele is segregating at the last sampling time (i.e., the process has not reached states 0 or 1). This case corresponds to what we think is a realistic scenario for how ancient DNA data would be collected, where presumably the locus of interest is polymorphic at present. Indeed, only such loci would be selected for inference.

We can rewrite the resulting likelihood as follows:

$$\ell^C(\theta) = p(i_1, \ldots, i_m|\theta, \mathcal{T}, z_{j_m} \notin \{0, 1\}) \quad = \quad \frac{p(i_1, \ldots, i_m, z_{j_m} \notin \{0, 1\}|\theta, \mathcal{T})}{\sum_{j_m=1}^{H-2} p_{\tau_m-\tau_0}(\frac{1}{2N_e}, z_{j_m})}, \qquad (9)$$

where

$$
\begin{aligned}
&p(i_1, \ldots, i_m, z_{j_m} \notin \{0,1\} | \theta, \mathcal{T}) = \\
&\sum_{j_m=1}^{H-2} p(i_m|z_{j_m}) \sum_{j_{m-1}=1}^{H-2} p_{\tau_m - \tau_{m-1}}(z_{j_{m-1}}, z_{j_m}) \cdots \\
&p(i_2|z_{j_2}) \sum_{j_1=0}^{H-2} p_{\tau_2 - \tau_1}(z_{j_1}, z_{j_2}) \cdot \\
&p(i_1|z_{j_1}) p_{\tau_1 - \tau_0}(\frac{1}{2N_e}, z_{j_1}).
\end{aligned} \tag{10}
$$

We consider the subprocess Z_τ^C defined on the reduced state space $\{z_1, \ldots, z_{H-2}\} \subset \{z_0, z_1 \ldots z_{H-2}, z_{H-1}\}$. The infinitesimal generator q^C of such a process is the matrix Q without the first and last rows and columns, i.e.:

$$
q^C = \begin{pmatrix}
\eta_1 & \beta_1 & 0 & & \ldots & & 0 \\
\delta_2 & \eta_2 & \beta_2 & 0 & & & \\
0 & \ddots & \ddots & \ddots & 0 & & \vdots \\
 & 0 & \delta_k & \eta_k & \beta_k & 0 & \\
\vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\
 & & & 0 & \delta_{H-3} & \eta_{H-3} & \beta_{H-3} \\
0 & & \ldots & & 0 & \delta_{H-2} & \eta_{H-2}
\end{pmatrix}. \tag{11}
$$

Denoting $p^C_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k})$ the transition probabilities of this subprocess we have that $p_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k}) = p^C_{\tau_k - \tau_{k-1}}(z_{j_{k-1}}, z_{j_k})$ for $\forall\, j_{k-1}, j_k \notin \{0, H-1\}$ (see Supplementary Material B for more details).

Finally, in order to compute the likelihood of equations 8 and 9, all that remains is to compute the matrix exponentiation $e^{Q\tau}$ and $e^{q^C \tau}$, respectively.

2.2 Numerics

We evaluate numerically the matrix exponentiation. The advantage of the current approach compared to Bollback et al.'s is that we do not need to do a numerical integration step since the state space is already finite. The description of the matrix exponentiation is given in Supplementary Material B.

Although asymptotically the one step process is equivalent to the WF model, since the state space of Z_τ has a finite number of states, the accuracy of the approximation will depend on the choice of the states, or what we call from now on the "grid". We investigate three grids strongly inspired by Gutenkunst et al. (2009). The first one is a uniform grid with a point added at $\frac{1}{2N_e}$. The second and third grid are a "quadratic grid" and an "exponential grid". The last two grids were chosen to be refined around the boundaries in such a way that the distance between adjacent points changes smoothly. The details for the grids are given in Supplementary Material B. All three grids have a point at $\frac{1}{2N_e}$.

Since the likelihood function is complex, we were not able to compute the maximum of the function analytically. Therefore, in order to find the maximum, we first computed the likelihood over a large range of parameters. We verified that there is a single maximum for each time interval defined by adjacent sampling times, i.e., if $t_0 < t_1$, the time intervals are $(-\infty, t_0)$, (t_1, t_2),...,(t_{m-1}, t_m), and that the likelihood surface is smooth. We used the *SciPy* (Jones et al., 2001) implementation of the Nelder-Mead simplex algorithm (Nelder and Mead 1965) to find the maximum for each time interval.

Our implementation is written in *Python* and *C++* making use of the *Numpy* (Oliphant, 2006), *SciPy* and *mpack* (Nakata, 2010) libraries for computations and of the *Matplotlib* library (Hunter, 2007) for plotting and is available upon request.

2.3 Simulations

In order to test our model, we simulate several datasets with the WF model forward in time. Simulating the WF model can be time consuming if the population size is large, so

we picked a small population size ($N_e = 500$). In principle, however, the conclusions hold for higher population size. We then infer the Maximum Likelihood Estimates (MLEs) using our one step method. We use two different sampling schemes. The first one is similar to the real dataset we analyze below, i.e., 6 sampling times each with 50 chromosomes. The second one corresponds to having twice as many sampling times with half the number of chromosomes, i.e., 12 sampling times and 25 chromosomes. We searched for the MLEs across a finite domain, i.e., $N_e \in [100, 1000]$, $t_0 \in [-3000, 0]$, and $\gamma \in [-200, 200]$. We assess the accuracy of our estimator and compare the sampling schemes by looking at the bias of the estimates and the root mean square error (RMSE).

2.4 Real data

In 2009, Ludwig et al. sequenced several loci encoding coat color in horses. Each locus has been shown to be linked with a color phenotype in present day horses. In other words, the phenotype associated with each locus is segregating in present populations. We re-analyze in this paper one of the loci encoding for the agouti-signaling-protein (ASIP), that controls the distribution of the black pigment (Rieder et al. (2001)). We investigate the hypothesis that at the beginning of domestication some coat colors in horse were positively selected for.

The samples sequenced were obtained from Siberia, Middle and Eastern Europe, China, and the Iberian Peninsula. As in Ludwig et al. (2009), we group the samples into six sampling times, $t_1 \simeq 20000$, $t_2 \simeq 13100$, $t_3 \simeq 3700$, $t_4 \simeq 2800$, $t_5 \simeq 1100$ and $t_6 \simeq 500$ years BC. We assume that the generation time of horses is 5 years, following Ludwig et al. (2009). The wild type horses are presumed to have been of bay color. The mutation of interest is recessive, since only horses homozygous for the *ASIP* locus will be black. So, in this case $h = 0$.

To compute a possible range for the population sizes we use data from Cieslak et al. (2010). They sequenced part of the control region of the mtDNA for 78 samples that are part of Ludwig et al. (2009)'s dataset. The control region of the mtDNA is a non coding region. One way to compute the population size N_e is to compute the diversity π of the

samples. Then, assuming the region is neutral and ignoring hitchhiking effects due to nearby selected sites, we use the relationship that relates the diversity of a sample to the population size, $\pi = 2N_e\mu \Rightarrow N_e = \frac{\pi}{2\mu}$, where μ is the mutation rate per base pair per generation. To get an estimate of the mean and standard error of π of the mtDNA sample, we use the maximum likelihood method implemented in *MEGA* (Tamura K et al., 2011) with default parameters. We approximate the standard error for the diversity by performing 1000 bootstraps. We use Jazin et al. (1998)'s estimate for the mutation rate (i.e., $\mu \in (3.0 \cdot 10^{-6}, 4.4 \cdot 10^{-5})$). Those authors used human families to get direct estimates of the mutation rate for mtDNA control region for a single generation. Although the mutation rate is an important parameter, we do not have direct estimate in horses and we have to rely on results for other species. To get conservative upper/lower bounds for N_e we use the 95% confidence interval (CI) bounds of the mutation rate and the diversity. If the CIs for μ and π are denoted (μ_{low}, μ_{up}) and (π_{low}, π_{up}) respectively, we defined $N_{e_{low}} = \frac{\pi_{low}}{2\mu_{up}}$ and $N_{e_{up}} = \frac{\pi_{up}}{2\mu_{low}}$.

In order to find the MLEs we use a domain defined by $N_e \in [200, 5000]$, $t_0 \in [-10000, 0]$, and $\gamma \in [-200, 200]$ for the parameters. We fix $H = 400$ for this computation.

For the CIs, there exist several asymptotic results that apply for maximum likelihood, especially for a time serial Markov chain. Our sample sizes are generally small, however, so we chose to compute the CIs with a parametric bootstrap approach.

Note that several assumptions of our model are violated with this dataset, such as constant population size, potentially random mating (since the samples are taken from all around the world), moreover the *MC1R* locus, encoding a melanocortin receptor and related to the black pigment production, is known to have an epistatic interaction with *ASIP* (Rieder et al., 2001). Nevertheless, we decided to analyze these data in the way we have indicate to be able to compare our results with those obtained with Bollback et al.'s method on the same dataset.

3 Results and Discussion

3.1 Numerics

In order to validate the method, we compared several known analytical results for the WF model with the one step process. For the neutral case, it is possible to compute the likelihood since the transition probabilities are known for the diffusion process (see e.g. Ewens (2004)). As can be seen in Supplementary Figure 3, even for a grid of size 100 the one step process is a very good approximation of the diffusion process.

We also compare the relative error between the diffusion and the one step process and demonstrate that when we increase the grid size the one step process converges towards the diffusion process. The results for a particular choice of parameters is shown in Supplementary Figure 4. We see that the one step process does converge as expected with increasing grid size. In general we see that a quadratic grid and an exponential grid perform better than a uniform grid in general (see Supplementary Material C for details). In the applications below we will use a quadratic grid of size between 100 and 400.

3.2 Simulations

We picked a population size of $N_e = 500$ and set the allele age to $t_0 = -1400$. We fix the selection coefficient to seven potential values: $\gamma \in \{-10, -5, 0, 5, 10, 15, 20\}$.

First, we fix the sampling times to $T = (-1000, -800, -600, -400, -200, 0)$ generations and sample 50 chromosomes at each time point. Then we look at a scheme where the samples are taken every 100 generations from -1100 up to 0 (i.e. 12 samples). At each sampling time we sample 25 chromosomes. The intent is to quantify whether it is better to sample more chromosomes at fewer time points, or the opposite.

The boxplot results for the MLEs for these simulations are shown on Figure 2. They are standard boxplots showing the five point summary (the minimum, the first quartile, the median, the third quartile, and the maximum). The plots for the bias and the RMSE are

shown on Supplementary Figure 5 for both schemes.

For the population size, the MLEs span all the potential range of N_e values, but the bulk of the results exclude very low population sizes. This suggests nevertheless that it is hard to estimate N_e with our method, at least with a precision higher than one order of magnitude. Our estimator is biased upwards for both schemes but this might be explained by the presence of outliers since the median is largely accurate. Moreover, the second scheme, with less chromosomes and more sampling, leads to a smaller bias and a smaller RMSE for most cases. Intuitively, we think that most of the information to estimate s comes from the general trend of change in allele frequency, while most of the information to estimate N_e comes from the oscillations around that general trend. In other words, to get a precise estimate of N_e, we need a dense sampling over time, which is not the case for our simulations that we chose to match the real data setting.

In contrast, the results for the selection coefficient are essentially unbiased, with a symmetric distribution, and the median matching the mean of the distribution. The variance remains large, and only when γ is quite high can one reject neutrality. In particular, the higher the selection coefficient, the higher the variance. The RMSE this time is worse for the second sampling scheme.

The results for the allele age also exhibit a large variance. The tail of the distribution is large. This can be explained by the use of the conditional process. Indeed for weak selection, if the number of derived alleles is high at the first sampling time the likelihood becomes uninformative for the allele age (i.e., the likelihood is flat for older allele ages; Supplementary Figure 3). This leads to difficulties for the optimization algorithm to converge to the global maximum. The results seem to be systematically biased upwards, although the median is accurate. For strong selection the likelihood is more informative and the estimator is unbiased. Also, for strong selection the scheme with more samples through time performs considerably better.

In conclusion, sampling fewer chromosomes over more sampling times will lead to better

results especially for strong selection.

3.3 Real data

The change in allelic frequency of this locus is shown in Figure 3. Although the frequency is increasing in around 3,000 generations from 0 to ∼0.8 between the first and the third sampling time, suggesting positive selection, it then drops down to 0.4 in around 500 generations. It is interesting to note that the archaeological evidence for domestication suggests a date of 3500 years BC (Outram et al., 2009), which would correspond to the third sampling time (i.e. when the sample frequencies start decreasing).

The first step is to choose a potential range for the population size. We found $\pi = 0.024$ with a 95% CI of $(0.018, 0.030)$. Together with the 95% CI of the mutation rate, this leads to a range for N_e of $(200, 5000)$. This is a small population size. It might be explained by the fact that the horses are a domesticated species and most samples are taken after the beginning of domestication, resulting in a small N_e. On the other hand it might be that the mutation rate calculated for the human population for the control region is not appropriate for horses.

In Supplementary Figure 6 we plot the likelihood surface for 4 values of N_e. This helps us confirm that we have found a global maximum. We note that the higher the population size the higher the selection coefficient and the older the allele age that maximize the likelihood. For example, if the population is fixed at $N_e = 200$ then $\gamma^{max} = -1.5$ and $t_0^{max} = -2567$. In contrast, if we fix $N_e = 5000$, then $\gamma^{max} = 9.1$ and $t_0^{max} = -3550$. In other words, if the mutation rate is overestimated by say an order of magnitude, our potential range for the population size will also be much higher.

Since there is no mutant allele at the first time of sampling, the allele might have arisen after the first sampling time. We denote *"dom1"* the range between $(-\infty, -3893]$ generations, and *"dom2"* the range $(-3893, -2516]$. As discussed before, the likelihood is therefore discontinuous as a function of the allele age with discontinuities at the sampling times. It

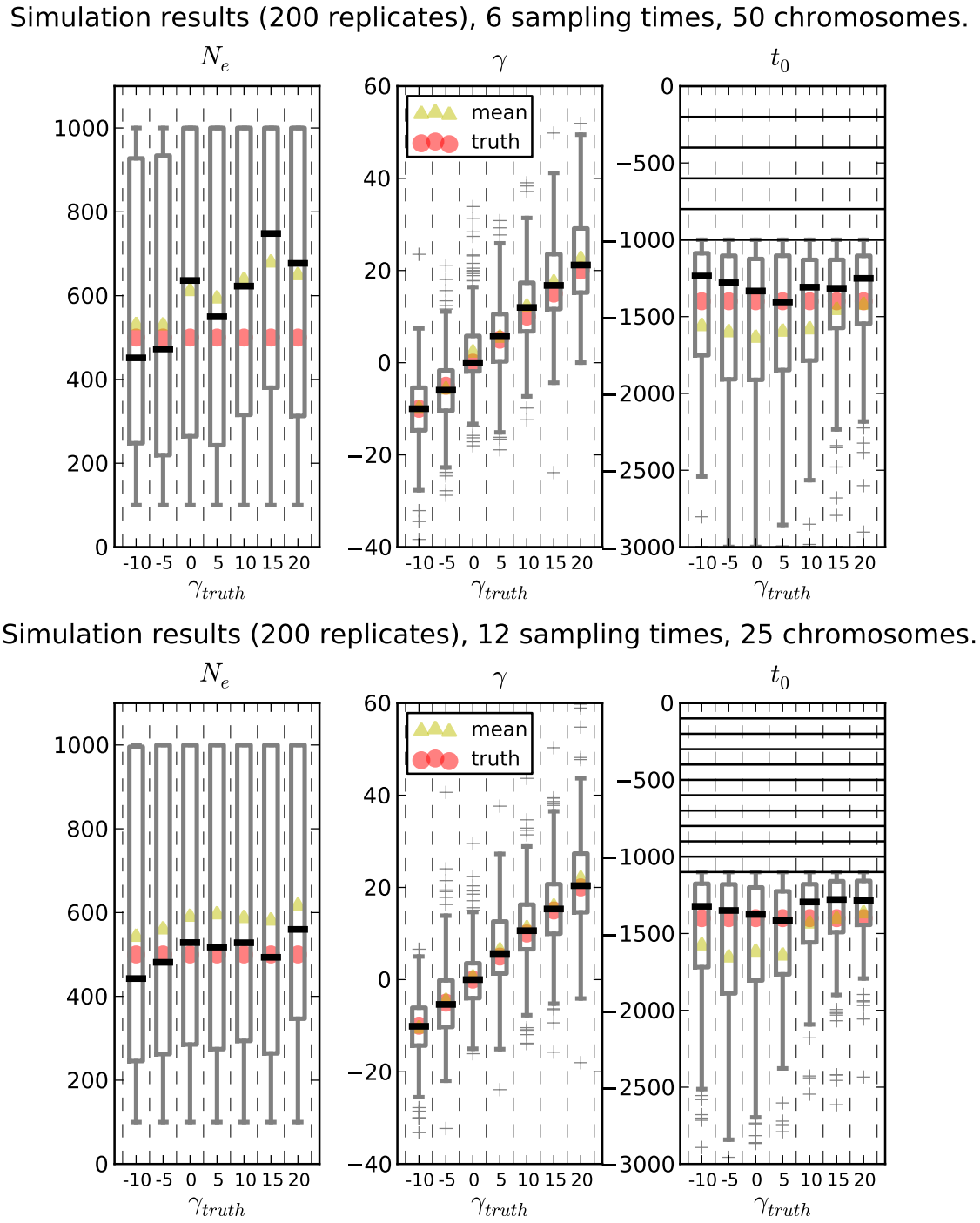

Figure 2: Boxplots for the MLEs of each simulation replicate, for $\gamma \in \{-10, -5, ..., 20\}$, $N_e = 400$ and $t_0 = -1400$. At the top is the scheme with 6 sampling times and 50 chromosomes sampled. At the bottom, the scheme with 12 sampling times and 25 chromosomes sampled. On each plot, the estimates for the population size, N_e (left), the rescaled selection coefficient, γ (middle), and the allele age, t_0 (right). For all subplots the triangle represents the mean of the estimates, and the circle the true value. The rectangles of the boxplots are for the first and third quartiles and the black line represents the median. The outliers are also indicated by crosses.

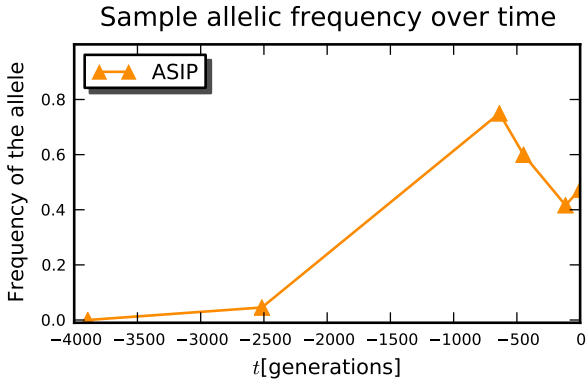

Figure 3: Change in allelic frequency over time for the *ASIP* locus. The sample sizes are $M = (10, 22, 20, 20, 36, 38)$ and the number of derived alleles $I = (0, 1, 15, 12, 15, 18)$. The times have been offset so that the last sampling time is 0. Domestication is thought to have happened around 3500 years BC which would correspond to around -600 generations on this plot, i.e., the 3rd sampling time.

is important to look for the global maximum in *dom1* and *dom2* separately. Moreover, we compute the 95% CI in *dom1* and in *dom2* separately. We build the confidence interval as a union of (potentially) disconnected domains.

The values for the MLEs and 95% CI are shown in Table 1. The first thing to note is that they are compatible with the results of Figure 6. The MLEs were found in *dom2*: $t_0^{mle} \cong -2600$ with CI $(-4760, -3893] \bigcup (-3893, -2516]$, $\gamma^{mle} \cong -1.3$ with CI $[-27.7, 60.7]$, and $N_e^{mle} \cong 600$ with CI $[200, 5000]$.

In Figure 4 we plot the distribution for the bootstrap replicates for each parameter and for the maximum likelihood values. The confidence interval was constructed as the interval between 2.5th and 97.5th percentile. We ran a total of 1400 replicates. For about 30 of those simulations, the optimizer did not converge. Among successful runs, ∼500 did not have an MLE in *dom1* or *dom2* and were discarded. From the remaining, about 823 were found in *dom2* and 34 in *dom1*.

The MLEs and the bootstrap results have several implications. First, we do not find evidence for positive selection as could be anticipated by the archaeological evidence for domestication. The discrepancy between this study and Ludwig et al. (2009) is first the

method used and second the parameter range assumed. Indeed, the results in Ludwig et al. (2009) were obtained using Bollback et al. (2008)'s method. Since our t_0^{mle} is in *dom2*, and Bollback et al. 2008 assume that the allele was already present in the first time of sampling, it is to be expected that our results will be very different. Moreover, the potential range for the population size in Ludwig et al. 2009 is from 10,000 to 100,000, i.e., it does not overlap with the range for N_e that we assume here. As noted above, if we had assumed a larger population size, the γ^{mle} would be larger.

The distribution of each parameter from the bootstrap replicates are almost unbiased relative to the true value (as could be expected from the results in the simulation section). The distribution for γ is close to a normal distribution while the distribution for N_e and t_0 are not as simple. For N_e, the distribution is bimodal with a second mode at the upper bound. This mode is a reflection of the finite domain we impose on the search for the MLE rather than an actual mode. Similarly, for t_0 there is a mode at the lower bound for *dom2*, an artifact of the bounds from the sampling times.

As could be expected from the simulations above, the 95% CI for N_e suggests that with these data we have little ability to estimate N_e, which we would expect from a sparse sampling over time as discussed earlier. Similarly, we cannot distinguish between negative and positive selection as γ's CI is between -27.7 and 60.7. On the other hand, the bootstrap replicates suggest that the allele arose in *dom2*. We can indeed test the hypothesis that the allele age is not in *dom2*; that is, we can test the null hypothesis $H_0 : t_0 \notin dom2$ versus the alternative hypothesis that the allele age is in *dom2*, $H_1 : t_0 \in dom2$. We reject the null hypothesis H_0 with p-value $1 - \frac{823}{823+34} = 0.04$.

The domain *dom2* corresponds to 20,000 to 13,100 years BC. In other words, from the data, one could have already deduced that the allele had to be present before -13,100 years (i.e., before the presumed start of domestication). Indeed, domestication in horses is thought to have started about 3,500 years BC (Outram et al., 2009). Our analysis shows that it is likely to have arisen within the last 20,000 years, thus clearly indicating that it was present

	dom1	*dom2*
	local optimum	local optimum
ℓ	14.9	**13.1**
t_0	-3893	**-2577**
γ	-0.61	**-1.3**
N_e	1617	**652**

Table 1: Maxima for the *ASIP* locus sequenced in (Ludwig et al., 2009). The MLEs are on the right most column.

as a standing variant at the time of domestication.

4 Conclusion

The allele age, the strength of selection and the population size are all crucial parameters in population genetics. Although the volume of molecular data is growing exponentially in recent years, it often remains a challenge to estimate those key parameters.

We develop a maximum likelihood approach to estimate these parameters that deals with a particular type of data - temporal data. Our method is based on an approximation to the WF diffusion process, and has the advantage of being quite flexible and appropriate for hypothesis testing. Moreover, it is fast for small γ: as one evaluation of the likelihood function takes $\sim$0.1 seconds for $\gamma \lesssim 40$ on a laptop with a i5 2.53 GHz CPU, for a dataset like the one we analyze here.

We show through simulations that, for a sample of realistic size, although the variance of our estimator is quite large, our MLE is unbiased for estimating selection and is nearly unbiased for the age of the allele and the effective population size. On the other hand, our method is not appropriate for estimating the population size, even for simulations where the model used to simulate the data match the method used to infer the parameters. Indeed, for a realistic sampling scenario, the MLEs for N_e, although unbiased, can span several orders of magnitude. This is not surprising. The effective population size is a parameter notoriously difficult to estimate, and our method considers only a single locus.

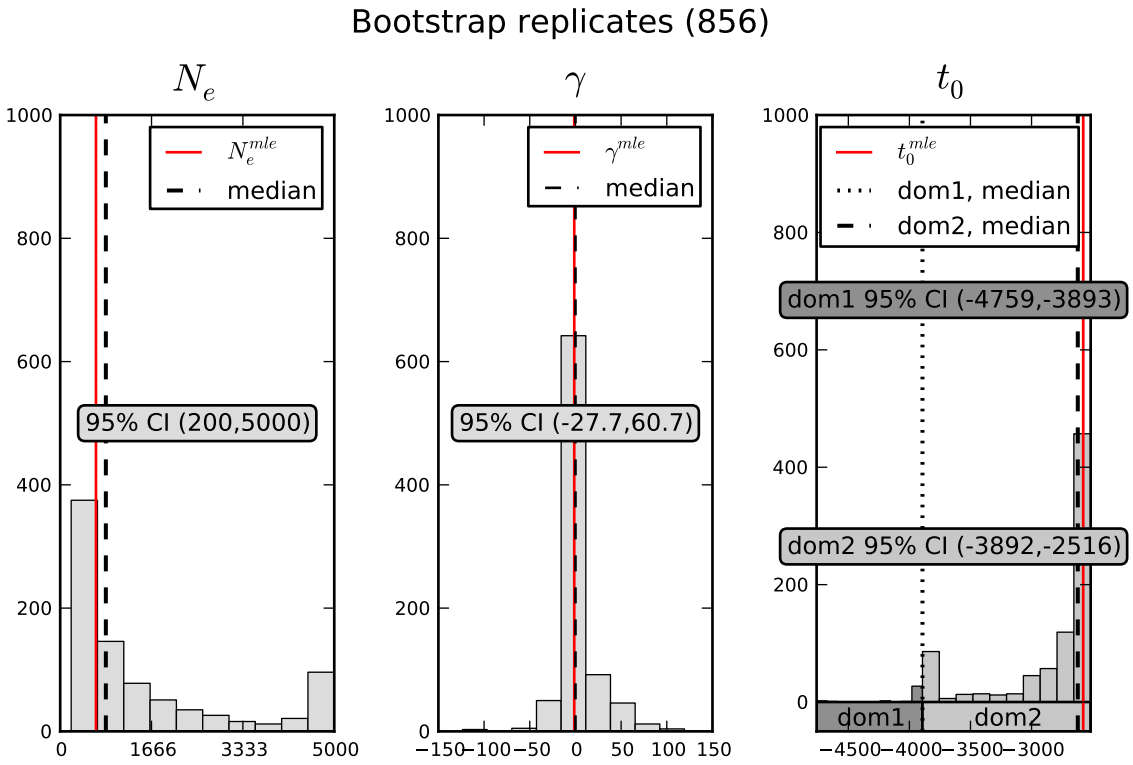

Figure 4: Bootstrap estimate of the sampling distribution of ML estimators of the three variables N_e, γ and t_0 for the parametric bootstrap. Of 1400 simulations, 856 were compatible with the data (i.e. the maximum for t_0 was in *dom1* or *dom2*. In each case the local maximum is indicated.

The sampling scheme has of course an impact on the accuracy of the estimator. We investigated two different sampling strategies and concluded that, in the cases considered, it is better to increase the number of sampling times rather than the number of samples per time point. It is indeed intuitive that in order to be able to estimate the allele age, for the conditional process, it is necessary to have a sample close to the allele age. Indeed, in the conditional process, an allele will never get fixed or lost. Thus, after several units of rescaled time, the likelihood is flat.

We re-analyze a locus that was previously found to be under positive selection, *ASIP*, by evaluating samples ranging from the Pleistocene to the present. In this study, we do not have sufficient resolution to distinguish positive from negative selection for this locus. This may be due to an insufficient amount of data, but it could also be due to an underestimate of the effective population size, or a violation of one or more assumptions of our null model, as discussed earlier. Although we are not able to estimate the selection coefficient as precisely as we would like, we find the age of the ASIP mutation to range between 20,000 to 13,100 years BC with an MLE at 13,400 years BC, which well predates domestication.

Even though we analyze a mammalian dataset, our method can in principle be applied to datasets obtained in experimental evolution or viral data. But, it is important to note that our approximation to the WF model will only be valid when the diffusion approximation to the WF model is appropriate, and hence only when $2N_e s$ is not too large.

Importantly, our framework readily lends itself to being extended to multiple loci, the topic of future investigation. This extension is anticipated to provide greatly improved estimates of N_e and permit the inference of fluctuations in historical population size - both issues of outstanding importance in gaining refined estimates of selection coefficients and allele age.

5 Acknowledgements

ASM would like to thank Fernando Perez for help with the numerical analysis and *Numpy* and *Scipy* and Philip Johnson and Emilia Huerta-Sanchez for helpful discussion. This work was part of ASM's PhD thesis in the Department of Integrative Biology, UC Berkeley. It was funded in part by an Ernst and Lucie Schmidheiny fellowship to ASM, by the French Ministry of Industry (DGCIS) and the Region Ile-de-France in the framework of the LaBS Project, by an NSF grant DMS-0907630 to SNE and by a National Institute of Health grant (R01-GM40282) to MS.

Supplementary Material

A One step process, Q matrix

We denote by L the generator of the diffusion process Y_τ. We have that

$$L = \frac{1}{2}a(y)\frac{d^2}{dy^2} + b(y)\frac{d}{dy} \tag{12}$$

where $a(y)$ and $b(y)$ are the infinitesimal variance and mean of our diffusion process. For the WF model with additive selection (see main text) those functions are:

$$a(y) = y(1-y) \tag{13}$$

$$b(y) = \gamma y(1-y)(y + h(1-2y)). \tag{14}$$

By definition, the generator can also be written as

$$\lim_{\tau \downarrow 0} \frac{\mathbb{E}^y[f(Y_\tau)] - f(y)}{\tau} = Lf(y). \tag{15}$$

Ignoring the $\Delta\tau^2$ terms, we have for the infinitesimal mean:

$$\mathbb{E}^y[Y_{s+\Delta\tau} - Y_s \mid Y_s] \cong \gamma Y_s(1-Y_s)(Y_s + h\cdot(1-2Y_s))\Delta\tau = b(Y_s)\cdot\Delta\tau. \tag{16}$$

Similarly, the infinitesimal variance is:

$$\mathbb{E}^y\left[\{Y_{s+\Delta t} - Y_s - \gamma Y_s(1-Y_s)(Y_s + h\cdot(1-2Y_s))\}^2 \mid Y_s\right] \cong Y_s(1-Y_s)\Delta\tau = a(Y_s)\cdot\Delta\tau. \tag{17}$$

We want to choose the Markov chain Z such that $Z \simeq Y$, in the sense that the probability distribution governing the samples of Z is close to the probability distribution governing the

samples of Y. To achieve that, we can match the infinitesimal mean and variance of Z and Y (see Durrett (2008)). By definition of the generator of Z_τ (see equation 6), we know the probabilities of transition in time $\Delta\tau$. Assuming the process starts at $Z_s = z_i$:

$$Z_{s+\Delta\tau} = \begin{cases} z_i & \text{with probability } 1 - (\beta_i + \delta_i)\Delta\tau + \mathcal{O}(\Delta\tau^2) \\ z_{i+1} & \text{with probability } \beta_i\Delta\tau + \mathcal{O}(\Delta\tau^2) \\ z_{i-1} & \text{with probability } \delta_i\Delta\tau + \mathcal{O}(\Delta\tau^2) \end{cases} \tag{18}$$

We can rewrite equations 16 and 17 replacing Y_τ by Z_τ. We have for the infinitesimal mean

$$\begin{aligned} \mathbb{E}^{z_i}\left[\{Z_{s+\Delta t} - z_i\}\right] &\cong z_i \cdot (1 - (\beta_i + \delta_i)) + z_{i+1}(\beta_i\Delta\tau) + z_{i-1}(\delta_i\Delta\tau) - z_i \\ &= (\beta_i(z_{i+1} - z_i) + \delta_i(z_i - z_{i-1}))\Delta\tau \\ &= b(z_i)\cdot\Delta\tau, \end{aligned} \tag{19}$$

and for the infinitesimal variance:

$$\begin{aligned} \mathrm{Var}(Z_{s+\Delta t} - z_i) &= \mathbb{E}^{z_i}\left[\{Z_{\Delta t} - z_i\}^2\right] - \mathbb{E}^{z_i}\left[\{Z_{\Delta t} - z_i\}\right]^2 \\ &\cong (z_{i+1} - z_i)^2 \cdot \beta_i\Delta\tau + (z_{i-1} - z_i)^2 \cdot (\delta\Delta\tau - (z_i - z_i)^2(1 - \beta_i - \delta_i)\Delta\tau \\ &= (\beta_i(z_{i+1} - z_i)^2) + (\delta_i(z_i - z_{i-1})^2)\Delta\tau \\ &= a(z_i)\cdot\Delta\tau. \end{aligned} \tag{20}$$

We have therefore two equations 19 and 20 with two unknowns δ_i and β_i. Solving the system we get:

$$\beta_i = \frac{(-1 + z_i)\cdot z_i \cdot (-1 - z_i^2\cdot\gamma + h\cdot(-1 + 2\cdot z_i)\cdot(z_i - z_{i-1})\cdot\gamma + z_i\cdot z_{i-1}\cdot\gamma)}{(z_i - z_{i+1})\cdot(z_{i-1} - z_{i+1})} \tag{21}$$

$$\delta_i = \frac{-((-1+z_i)\cdot z_i \cdot (-1-z_i^2\cdot\gamma + h\cdot(-1+2\cdot z_i)\cdot(z_i - z_{i+1})\cdot\gamma + z_i\cdot z_{i+1}\cdot\gamma))}{(z_i - z_{i-1})\cdot(z_{i-1}-z_{i+1})}. \quad (22)$$

Note that since we require that $\delta_i, \beta_i > 0\ \forall i$, the range of the possible parameters γ depends on the choice of the states z_{i-1}, z_i, z_{i+1}, or on the grid. In particular if we use a uniform grid we get: $\{z_0, z_1, ..., z_{H-1}\} = \{0, \frac{1}{H-1}, ..., \frac{H-2}{H-1}, 1\}$ and $\beta_i = \frac{(-1+H-k)k(1+H^2+k\gamma+H(-2+h\gamma)-h(\gamma+2k\gamma))}{2(-1+H)^2}$ and $\delta_i = \frac{(-1+H-k)k(1+H^2-k\gamma-H(2+h\gamma)+h(\gamma+2k\gamma))}{2(-1+H)^2}$. Most likely the locus of interest is either dominant, co-dominant of recessive, i.e. $h \in \{0, \frac{1}{2}, 1\}$. In those three cases for a uniform grid the range of γ is easy to compute. If $h = \frac{1}{2}$ then $-2(H-1) < \gamma < 2(H-1)$, if $h = 0$, $-(H-1) < \gamma < (H-1)$, and if $h = 1$, $-\frac{(H-1)^2}{H-2} < \gamma < \frac{(H-1)^2}{H-2}$. In other words, we will need a large grid for high values of γ.

B Numerics

B.1 Matrix exponentiation

We would like to compute the matrix exponential of the matrix Q and the matrix q^C for the conditional process. We will focus on the non conditional process as the conditional process follows easily. We use the convention of numbering the elements of a matrix starting from 0 to $H-1$ for the unconditional process, and from 1 to $H-2$ for the conditional process. We seek to compute

$$\exp(Qt),$$

where the $H \times H$ matrix Q is a tridiagonal matrix with all entries above and below the diagonal strictly positive. We implement two different approaches to compute the matrix exponentiation.

The first approach is a scaling and squaring algorithm with a Padé approximation. This approach is described in detail in Moler and Van Loan (2003) and is implemented in *SciPy*. This method works for a general matrix and takes advantage of the properties of the matrix Q.

The matrix Q is in general not symmetric ($\delta_i \neq \beta_i$ when $s \neq 0$). Nevertheless all eigenvalues are real. In particular two eigenvalues are 0 and the others are negative. Thus, when we remove the first and last column and row, the resulting matrix is the tridiagonal matrix q^C. We can transform the matrix q^C into a symmetric matrix with a similarity transformation. More precisely, there exists a diagonal matrix

$$d = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & d_{H-3} & 0 \\ 0 & 0 & 0 & d_{H-2} \end{pmatrix} \tag{23}$$

such that $s = d^{-1} q^C d$ is a symmetric matrix. The d_i can be defined recursively as follows: $d_1 = 1$,$d_2 = \sqrt{\delta_2/\beta_1} \cdot d_1$, $d_3 = \sqrt{\delta_3/\beta_2} \cdot d_2$, ... Note that the square root exists since $\beta_i, \delta_i > 0$. The matrices q^C and s have the same eigenvalues, and the eigenvalues of a symmetric matrix are all real. In particular they are also eigenvalues of the original matrix Q. The two remaining eigenvalues of Q are the two zero eigenvalues (this can be seen writing the characteristic polynomials). Therefore all eigenvalues are real. We can build a matrix D adding a first and last row and column to the matrix d:

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_2 & 0 & 0 & 0 \\ 0 & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & d_{H-3} & 0 & 0 \\ 0 & 0 & 0 & 0 & d_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{24}$$

It follows that $R = D^{-1} Q D$ symmetries the interior part of Q (the matrix q^C) and is a

tridiagonal matrix as well. Since $s = d^{-1}qd$ is symmetric, there exists an orthogonal matrix, o, such that $\ell = o^T so$ is diagonal. This matrix ℓ has the following form:

$$\ell = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ ... & ... & ... & ... \\ 0 & 0 & \lambda_{H-3} & 0 \\ 0 & 0 & 0 & \lambda_{H-2} \end{pmatrix} \tag{25}$$

We can construct the matrix O as the matrix D before, with o in its center and adding first and last rows and columns with zeros everywhere but the diagonal entries $(0,0)$ and $(H-1,H-1)$. Then we see that $T = O^T RO$ has an inner part equal to ℓ the coefficients of the first and last lines remain equal to 0, and the coefficients on the first and last columns are non-zero. We denote $T(0,j) = v_{0,j}$ with $j = 1,\ldots,H-2$ and $T(H-1,j) = v_{H-1,j}$ with $j = 1,\ldots,H-2$. That is,

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ v_{0,1} & \lambda_1 & 0 & 0 & 0 & v_{H-1,1} \\ v_{0,2} & 0 & \lambda_2 & 0 & 0 & v_{H-1,2} \\ v_{0,..} & ... & ... & ... & ... & v_{H-1,..} \\ v_{0,H-3} & 0 & 0 & \lambda_{H-3} & 0 & v_{H-1,H-3} \\ v_{0,H-2} & 0 & 0 & 0 & \lambda_{H-2} & v_{H-1,H-2} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{26}$$

where the $v_{i,j} \neq 0$. We rewrite $T = \Lambda + V$ where

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 & 0 & 0 \\ 0 & ... & ... & ... & ... & 0 \\ 0 & 0 & 0 & \lambda_{H-3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{27}$$

and

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ v_{0,1} & 0 & 0 & 0 & 0 & v_{H-1,1} \\ v_{0,2} & 0 & 0 & 0 & 0 & v_{H-1,2} \\ v_{0,..} & ... & ... & ... & ... & v_{H-1,..} \\ v_{0,H-3} & 0 & 0 & 0 & 0 & v_{H-1,H-3} \\ v_{0,H-2} & 0 & 0 & 0 & 0 & v_{H-1,H-2} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{28}$$

We note that V is nilpotent and that $V\Lambda = 0$. It follows that for $k \geq 1$, $(\Lambda + V)^k = \Lambda^k + \Lambda^{k-1}V$, which we can see by induction. There is another identity that will be useful. We define:

$$\Lambda' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1/\lambda_1 & 0 & 0 & 0 \\ 0 & 0 & ... & 0 & 0 \\ 0 & 0 & 0 & 1/\lambda_{H-2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{29}$$

where λ_1, λ_2... are the diagonal entries of l. We see that for $k \geq 2$, $\Lambda^{k-1} = \Lambda^k \Lambda'$. Since

$$\begin{aligned} T &= (DO)^{-1} Q (DO) \\ Q &= (DO) T (DO)^{-1} \\ Qt &= (DO) Tt (DO)^{-1}, \end{aligned} \tag{30}$$

we have:

$$\exp(Qt) = \sum_{k=0}^{\infty} \frac{1}{k!} (Qt)^k = \sum_{k=0}^{\infty} \frac{1}{k!} ((DO) Tt (DO)^{-1})^k = DO \left(\sum_{k=0}^{\infty} \frac{1}{k!} (Tt)^k \right) (DO)^{-1}. \tag{31}$$

Then,

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{1}{k!} (Tt)^k &= \mathtt{I} + \sum_{k=1}^{\infty} \frac{1}{k!} ((\Lambda + V)t)^k \\ &= \mathtt{I} + \sum_{k=1}^{\infty} \frac{1}{k!} (\Lambda t)^k + \sum_{k=1}^{\infty} \frac{t^k}{k!} (\Lambda^{k-1} V) \\ &= \exp(\Lambda t) + tV + \left(\sum_{k=2}^{\infty} \frac{t^k}{k!} \Lambda^{k-1} \right) V \\ &= \exp(\Lambda t) + tV + \left(\sum_{k=0}^{\infty} \frac{t^k}{k!} \Lambda^k - \mathtt{I} - \Lambda t \right) \Lambda' V \\ &= \exp(\Lambda t) + tV + (\exp(\Lambda t) - \mathtt{I} - \Lambda t) \Lambda' V. \end{aligned} \tag{32}$$

Finally:

$$\exp(Qt) = DO \left(\exp(\Lambda t) + tV + (\exp(\Lambda t) - \mathtt{I} - \Lambda t) \Lambda' V \right) (DO)^{-1}. \tag{33}$$

In terms of computing time, this requires us to compute o using an algorithm for hermitian matrices, then to compute d by recursion and the rest should follow from matrix

multiplications. The advantage compared to the Padé approach described above is that most of the work is done once D and O are computed (only once) and reused for all time intervals.

In practice, the condition number of the matrix o can be very high leading to instabilities in the matrix exponentiation. Indeed the higher the condition number, the more sensitive the matrix will be to numerical operation. The condition number of our matrix can be of the order of 10^6 for large γ and is therefore ill-conditioned. Note that for the approximation of the diffusion process to the WF model, γ has to be on the order of 1. Thus, the matrix exponentiation becomes harder when the conditions for approximating the WF model with the diffusion are not necessarily met.

In order to overcome this problem, we implemented the matrix exponentiation in *C++* using a library, *mpack* (Nakata, 2010), for multiple precision arithmetic. The library *mpack* is a multiple precision arithmetic version of *LAPACK* and *BLAS*. Although this allows us to exponentiate the matrix for any γ in principle, it makes the matrix exponentiation step much slower. We therefore empirically test for which parameter range we require more precision than the double precision of *numpy* or *SciPy* that rely on *LAPACK*.

To do so, for a particular matrix $Q = Q(H, h, \gamma)$ we compute

$$\text{test}(Q) = \text{norm}\left((D \cdot O) \cdot (O^T D^{-1})\right) - \text{trace}\left((D \cdot O) \cdot (O^T D^{-1})\right), \tag{34}$$

where $\text{norm}(A) = \text{norm}((a_{ij})) = \sum_{i,j} |a_{ij}|$. The value of $\text{test}(Q)$ should be equal to 0. We choose a threshold value ϵ such that if $\text{test}(Q) > \epsilon$, we do not trust the default *SciPy* implementation and we invoke the higher precision computation. For this paper we used $\epsilon = 10^{-5}$.

We plot on Figure 1 the Boolean $\text{test}(Q) > \epsilon$ for different values of N_e and γ for $h = 0$. We can see on those plots that the matrix instability does depend on γ but not on the population size. For all the population sizes, the default implementation becomes unstable

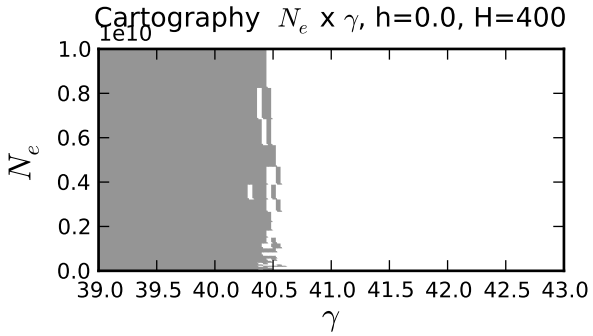

Figure 1: One example cartography of the parameter combination that require higher precision for $\epsilon = 10^{-5}$. We plot the result of the Boolean operation test($Q(H, h, \gamma, N_e) \leq \epsilon$ The legend is True for gray and white for False. We fix $H = 400$ and we plot N_e versus γ.

for $\gamma \gtrsim 40$.

To conclude, we use one existing method to exponentiate the matrix (Padé) and implemented one more method, with the possibility of increasing the double precision. Which method to use depends on the type of dataset and the parameter range one needs to explore. For high values of γ, if there are many time intervals, a method based on the spectral decomposition would be faster, otherwise the Padé approximation works well.

C Choice of grids

We investigated several grids inspired by Gutenkunst et al. (2009). No matter the parameters, to compute the likelihood we need to approximate the transition probabilities between the original frequency of the A allele, $\frac{1}{2N_e}$, and another frequency between 0 and 1. Although we could extrapolate, we decided to use grids that all include the point $\frac{1}{2N_e}$.

The first is a uniform grid with a point added at $\frac{1}{2N_e}$. We call this grid the "uniform grid". Then we investigate a quadratic grid and an exponential grid. The last two grids were chosen so that, as opposed to the uniform grid, the distance between adjacent points changes smoothly.

As before, let's denote by $\{z_0, z_1, ..., z_{H-1}\}$ the state space of the one step process or

the grid. The quadratic grid is described by a cubic equation, i.e., the difference between adjacent points is quadratic. We will assume for simplicity of notation that H is a multiple of 20 (it is straightforward to generalize), and that $G = \frac{H}{10}$. We set the first $G+1$ points to form a uniform grid between 0 and $\frac{2}{2N_e}$, so that the median of this grid is $\frac{1}{2N_e}$. In other words, $z_j = \frac{j}{N_e G}$ for $0 \leq j \leq G$. Now we assume first that $\{q_0, \ldots, q_{H-G-1}\}$ is a uniform grid between 0 and 1. In other words, $q_0 = 0$, $q_{H-G-1} = 1$ and $q_j = \frac{j}{H-G-1}$. The remaining points are described by

$$z_{G+j} = aq_j^3 + bq_j^2 + cq_j + d \tag{35}$$

where $d = \frac{2}{2N_e}$, $c = \frac{1}{2N_e G}$, $b = -3(\frac{1}{H-G-1} + c + \frac{d}{H-G-1})\frac{1}{H-G-1}$, $a = -\frac{2}{3}b$.

The exponential grid will be defined as follows. If $\{u_0, ..., u_{H-1}\}$ is a uniform grid between -1 and 1 (i.e., $u_0 = -1$, $u_{H-1} = 1$ and $u_j = -1 + j\frac{2}{H-1}$), then the grid is

$$z_j = \frac{\frac{1}{1+\exp(-\xi u_j)} - \frac{1}{1+\exp(\xi)}}{\frac{1}{1+\exp(-\xi)} - \frac{1}{1+\exp(\xi)}}, \tag{36}$$

where ξ is a parameter that defines the density of the grid around the boundaries. We pick ξ such as $z_{\left[\frac{H}{10}\right]} = \frac{1}{2N_e}$, with [] denoting the integer part. To do so, we solve numerically the equation 36 for $j = \left[\frac{H}{10}\right]$.

We plot the grids of interest versus uniform grids and the spacing between each point in Figure 2.

As noted in the main text, for the neutral case, it is possible to compute the likelihood since the transition probabilities are known for the diffusion process (see e.g. Ewens (2004)). We plot the results in Figure 3 for a quadratic grid of size 100 for two samples of size $M = (4, 4)$ and number of A alleles $I = (1, 3)$, sampled at times $T = (-200, 0)$ for several values of N_e and t_0. The plots suggest that even for a grid of size 100 the one step process is a very good approximation of the diffusion process.

We compare the relative error between the diffusion and the one step process and demonstrate that, when we increase the grid size the one step process converges towards the diffusion

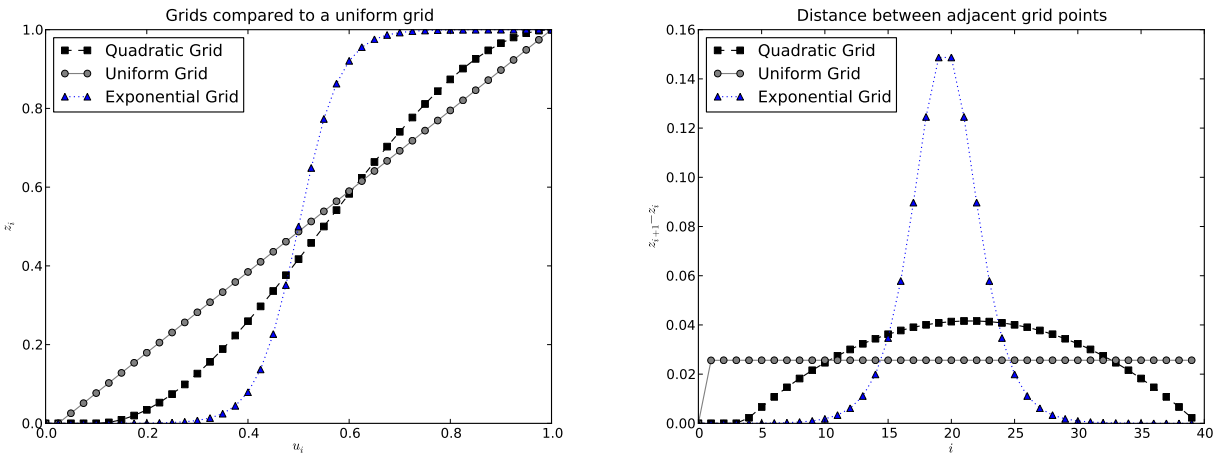

Figure 2: Description of three different grids tested of size H=41 and $N_e = 10^4$. Left: the grids are plotted against a uniform grid of points between 0 and 1. Right: the spacing of adjacent points.

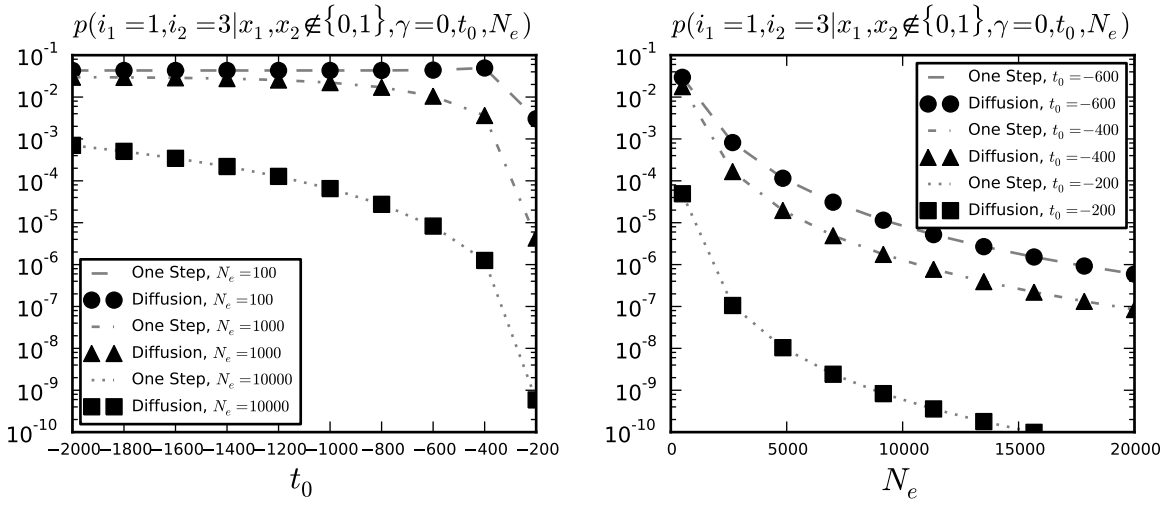

Figure 3: Likelihood for the neutral case for several values of N_e and t_0. The likelihood is for two samples taken at times -200 and 0 generations of size $M = (4, 4)$ and with $I = (1, 3)$ derived alleles. On the left (right), we fix N_e (respectively t_0) to several values and plot the likelihood versus t_0 (respectively N_e).

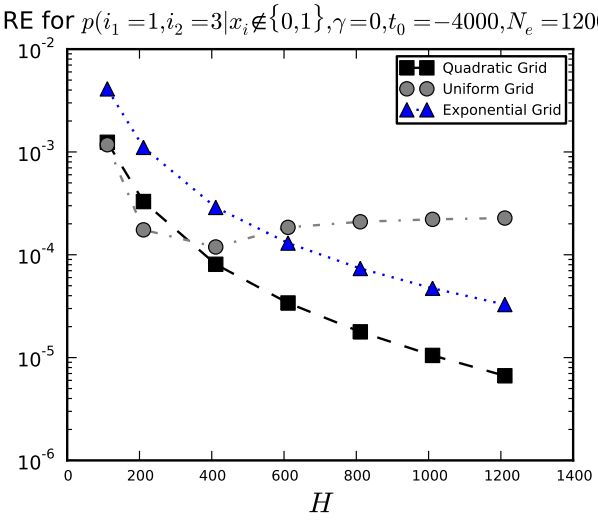

Figure 4: Relative error (RE) for the three grids discussed in 3.1 for the likelihood of 2 samples taken at times -3000 and 0, with $M = (4, 4)$. The parameter H describes the size of the grid. The y-axis is in logarithmic scale. In this example, the one step process converges towards the diffusion process faster when using the quadratic grid rather than the other two grids.

process. The results for a particular choice of parameters is shown in Figure 4 for the three grids. First we note that the one step process does converge as expected with increasing grid size. In this example, the convergence is faster for the quadratic grid. We looked at several combinations of parameters, and we observe that the quadratic grid and the exponential grid perform better than the uniform grid in general but that the ordering between the other two grids depends on the parameters. Indeed, if the allele age is close to the first sampling time a grid more refined around the frequency $\frac{1}{2N_e}$ performs better. In the applications below we will use a quadratic grid of size between 100 and 400.

D Simulations

We plot (Figure 5) here the root mean square error (RMSE) for the simulations for the two sampling schemes, *i.e.*, 6 and 12 sampling times. See main text for discussion.

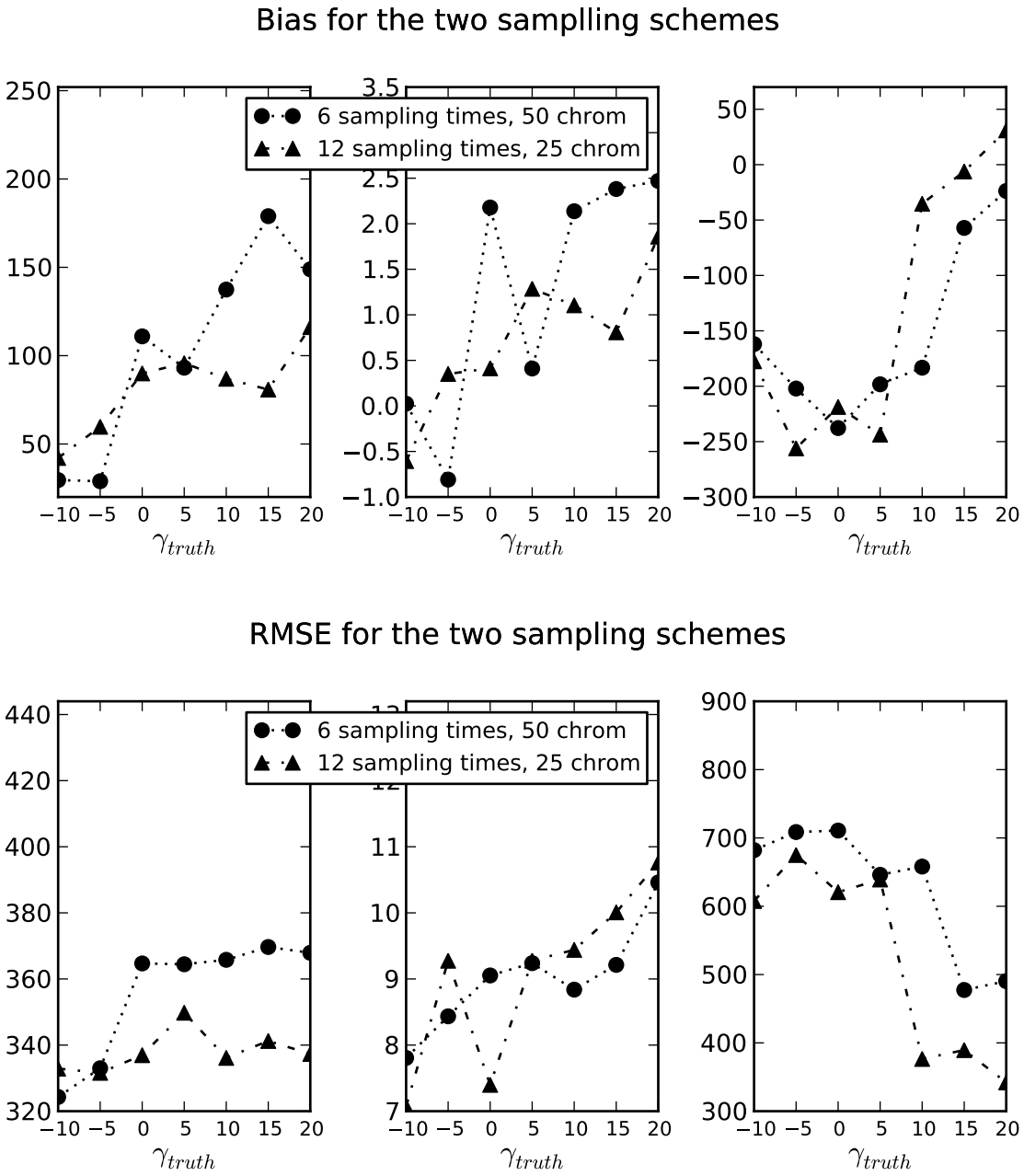

Figure 5: Bias (top plot) and RMSE (bottom plot) results for the MLEs for seven different sets of simulations also presented in Figure 2.

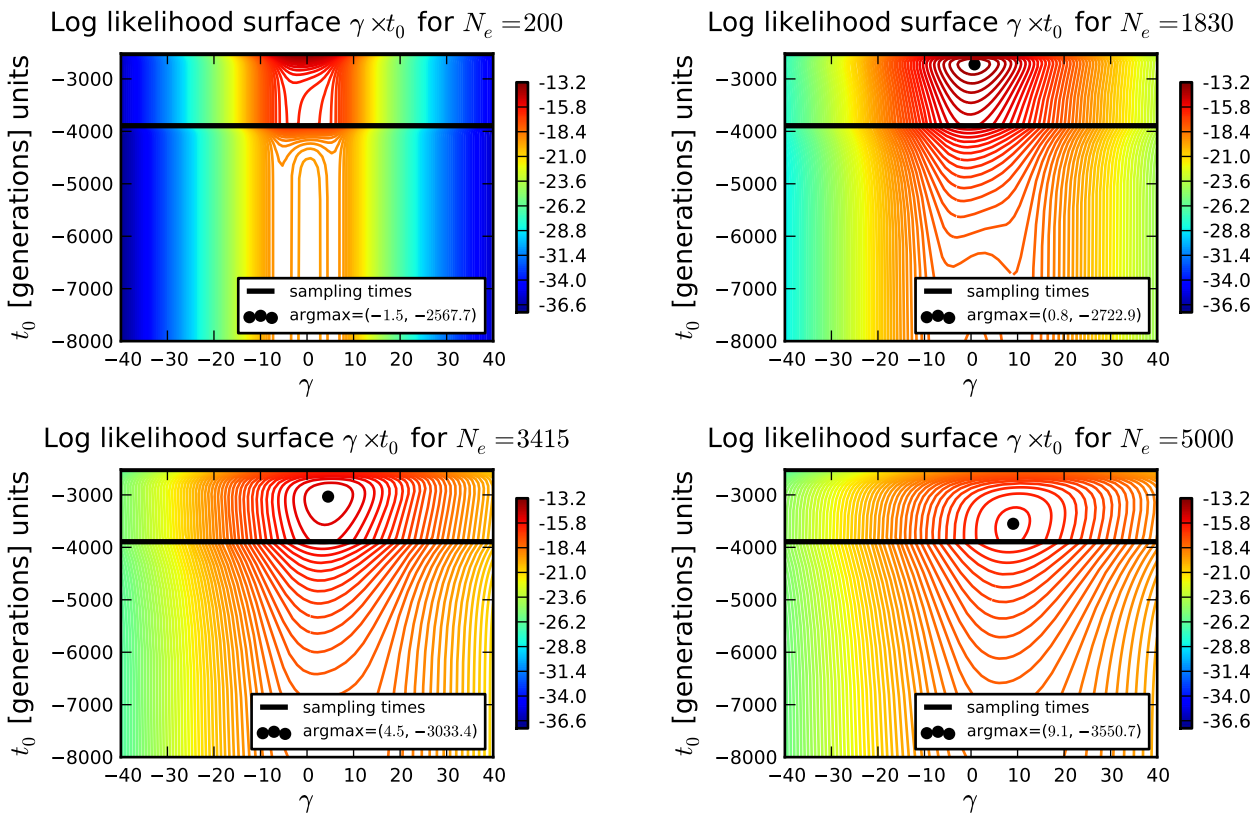

Figure 6: Likelihood surfaces for various values of N_e when analyzing the *ASIP* locus. In each case the local maximum is indicated.

E Real Data

We plot the likelihood surface for four values of N_e on Figure 6. As discussed in the main text, the higher the population size, the higher the selection coefficient, and the older the allele age that maximize the likelihood surface.

References

Anderson, E. C., E. G. Williamson, and E. A. Thompson (2000). Monte Carlo evaluation of the likelihood for N(e) from temporally spaced samples. *Genetics 156*(4), 2109–18.

Barrett, R. and D. Schluter (2008, January). Adaptation from standing genetic variation. *Trends Ecol Evol 23*(1), 38–44.

Bollback, J. P. and J. P. Huelsenbeck (2007). Clonal interference is alleviated by high mutation rates in large populations. *Mol Biol Evol 24*(6), 1397–406.

Bollback, J. P., T. L. York, and R. Nielsen (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics 179*(1), 497–502.

Cieslak, M., M. Pruvost, N. Benecke, M. Hofreiter, A. Morales, M. Reissmann, and A. Ludwig (2010). Origin and history of mitochondrial DNA lineages in domestic horses. *PLoS One 5*(12), e15311.

Drummond, A. J. and A. Rambaut (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol 7*, 214.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution* (2nd ed. ed.). Springer.

Ewens, W. J. (2004). *Mathematical Population Genetics* (second edition ed.). Springer.

Fisher, R. (1922). On the dominance ratio. *Proc. Roy. Soc. Edin. 42*, 321–341.

Gresham, D., M. M. Desai, C. M. Tucker, H. T. Jenq, D. A. Pai, A. Ward, C. G. DeSevo, D. Botstein, and M. J. Dunham (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet 4*(12), e1000303.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet 5*(10), e1000695.

Hamilton, D. (1994). *Time Series Analysis*. Princeton University Press.

Hunter, J. D. (2007, May-Jun). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering 9*(3), 90–95.

Jazin, E., H. Soodyall, P. Jalonen, E. Lindholm, M. Stoneking, and U. Gyllensten (1998). Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat Genet 18*(2), 109–10.

Jones, E., T. Oliphant, P. Peterson, and & others (2001). {SciPy}: Open source scientific tools for {Python}.

Lalueza-Fox, C., H. Rompler, D. Caramelli, C. Staubert, G. Catalano, D. Hughes, N. Rohland, E. Pilli, L. Longo, S. Condemi, M. de la Rasilla, J. Fortea, A. Rosas, M. Stoneking, T. Schoneberg, J. Bertranpetit, and M. Hofreiter (2007). A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science 318*(5855), 1453–5.

Ludwig, A., M. Pruvost, M. Reissmann, N. Benecke, G. A. Brockmann, P. Castanos, M. Cieslak, S. Lippold, L. Llorente, A. S. Malaspinas, M. Slatkin, and M. Hofreiter (2009). Coat color variation at the beginning of horse domestication. *Science 324*(5926), 485.

Moler, C. and C. Van Loan (2003). Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*. *SIAM Review 45*(1), 3–000.

Nakata, M. (2010, 6 August). The MPACK (MBLAS/MLAPACK); a multiple precision arithmetic version of BLAS and LAPACK. URL: http://mplapack.sourceforge.net/. Enter text here.

Nelson, M. I. and E. C. Holmes (2007). The evolution of epidemic influenza. *Nat Rev Genet 8*(3), 196–205.

Oliphant, T. (2006). *Guide to NumPy*. Trelgol Publishing.

Outram, A. K., N. A. Stear, R. Bendrey, S. Olsen, A. Kasparov, V. Zaibert, N. Thorpe, and R. P. Evershed (2009). The earliest horse harnessing and milking. *Science 323*(5919), 1332–5.

Rasmussen, M., Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. Guo, J. Zhao, X. Zhang, H. Zhang, Z. Li, M. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Gronnow, M. Meldgaard, C. Andreasen, S. A. Fedorova, L. P. Osipova, T. F. Higham, C. B. Ramsey, T. V. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Ponten, R. Villems, R. Nielsen, A. Krogh, J. Wang, and E. Willerslev (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature 463*(7282), 757–62.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J. J. Hublin, J. Kelso, M. Slatkin, and S. Paabo (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature 468*(7327), 1053–60.

Rieder, S., S. Taourit, D. Mariat, B. Langlois, and G. Guerin (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mamm Genome 12*(6), 450–5.

Rusk, N. (2009). Targeting ancient DNA. *Nature Methods 6*, 629.

Slatkin, M. and B. Rannala (2000). Estimating allele age. *Annu Rev Genomics Hum Genet 1*, 225–49.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution In press.* in press.

Van Kampen, N. (1992). *Stochastic processes in physics and chemistry*. Elsevier.

Waples, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics 121*(2), 379–91.

Wichman, H. A., J. Millstein, and J. J. Bull (2005). Adaptive molecular evolution for 13,000 phage generations: a possible arms race. *Genetics 170*(1), 19–31.

Williamson, E. G. and M. Slatkin (1999). Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics 152*(2), 755–61.

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