

# UC Berkeley

## UC Berkeley Previously Published Works

### Title

A novel spectral method for inferring general diploid selection from time series genetic data

### Permalink

<https://escholarship.org/uc/item/10g5n868>

### Journal

The Annals of Applied Statistics, 8(4)

### ISSN

1932-6157

### Authors

Steinrücken, Matthias  
Bhaskar, Anand  
Song, Yun S

### Publication Date

2014-12-01

### DOI

10.1214/14-aos764

Peer reviewed



Published in final edited form as:

*Ann Appl Stat.* 2014 December ; 8(4): 2203–2222.

## A NOVEL SPECTRAL METHOD FOR INFERRING GENERAL DIPLOID SELECTION FROM TIME SERIES GENETIC DATA

Matthias Steinrücken<sup>\*,†,‡</sup>, Anand Bhaskar<sup>\*,†</sup>, and Yun S. Song<sup>†</sup>

University of California, Berkeley

### Abstract

The increased availability of time series genetic variation data from experimental evolution studies and ancient DNA samples has created new opportunities to identify genomic regions under selective pressure and to estimate their associated fitness parameters. However, it is a challenging problem to compute the likelihood of non-neutral models for the population allele frequency dynamics, given the observed temporal DNA data. Here, we develop a novel spectral algorithm to analytically and efficiently integrate over all possible frequency trajectories between consecutive time points. This advance circumvents the limitations of existing methods which require fine-tuning the discretization of the population allele frequency space when numerically approximating requisite integrals. Furthermore, our method is flexible enough to handle general diploid models of selection where the heterozygote and homozygote fitness parameters can take any values, while previous methods focused on only a few restricted models of selection. We demonstrate the utility of our method on simulated data and also apply it to analyze ancient DNA data from genetic loci associated with coat coloration in horses. In contrast to previous studies, our exploration of the full fitness parameter space reveals that a heterozygote-advantage form of balancing selection may have been acting on these loci.

### Keywords and phrases

population genetics; spectral method; transition density function; hidden Markov model

### 1. Introduction

Natural selection is a fundamental evolutionary process and finding genomic regions experiencing selective pressure has important applications, including identifying the genetic

\*These authors contributed equally to this work.

<sup>†</sup>Supported in part by NIH grant R01-GM094402 and a Packard Fellowship for Science and Engineering

<sup>‡</sup>Supported in part by DFG research fellowship STE 2011/1-1

M. Steinrücken, Department of Statistics and Computer Science Division, University of California, Berkeley, Berkeley, CA 94720, USA, steinrue@stat.berkeley.edu

A. Bhaskar, Computer Science Division, University of California, Berkeley, Berkeley, CA 94720, USA, bhaskar@eecs.berkeley.edu

Y. S. Song, Department of Statistics, Computer Science Division, and Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA, yss@stat.berkeley.edu

#### SUPPLEMENTARY MATERIAL

##### A novel spectral method for inferring general diploid selection from time series genetic data

(doi: TBD;.pdf). We provide proofs of the results stated in Section 2. The modified Jacobi polynomials appearing in this paper are defined and some of their key properties are listed. Also, the coefficients in the definition of the matrix  $M$  in equation (2.14) are provided. Lastly, we describe some alternate density functions for the allele frequency at the time when selection arises.

basis of diseases and understanding the molecular basis of adaptation. There has been a long line of theoretical and experimental research devoted to modeling and detecting selection acting at a given locus. Several earlier works have considered modeling the stationary distribution of allele frequencies in a population undergoing non-neutral evolution (Fearnhead, 2003, 2006; Genz and Joyce, 2003; Stephens and Donnelly, 2003). More recently, there has been growing interest to utilize time series genetic variation data to enhance our ability to infer allele frequency trajectories, thereby enabling better estimates of selection parameters. For example, the sequencing of samples over several generations in experimental evolution of a population (e.g., Bacteria (Wiser, Ribeck and Lenski, 2013), yeast (Lang et al., 2013), and *Drosophila* (Burke et al., 2010; Orozco Wengel et al., 2012)) under controlled laboratory environments, or direct measurements in fast evolving populations such as HIV (Shankarappa et al., 1999), has allowed us to better understand the genetic basis of adaptation to changes in the environment. Also, recent technological advances have given us the unprecedented ability to acquire ancient DNA samples (e.g., for humans (Hummel et al., 2005), ancient hominids (Green et al., 2010; Reich et al., 2010), and horses (Ludwig et al., 2009; Orlando et al., 2013)), providing useful information about allele frequency trajectories over long evolutionary timescales.

Most methods for analyzing times series DNA data model the underlying population-wide allele frequency as an unobserved latent variable in a hidden Markov model (HMM) framework, in which the sample of alleles drawn from the population at a given time is treated as a noisy observation of the hidden population allele frequency. In this framework, computing the probability of observing time series genetic variation data involves integrating over all possible hidden trajectories of the population allele frequency. For short evolutionary timescales, a discrete-time Wright-Fisher model of random mating is often used to describe the dynamics of the population allele frequency in the underlying HMM. This approach has been used to estimate the effective population size from temporal allele frequency variation, assuming a neutral model of evolution (Williamson and Slatkin, 1999). More recently, temporal and spatial variations of advantageous alleles have been investigated through an HMM framework that can incorporate migration between multiple subpopulations (Mathieson and McVean, 2013).

If the evolutionary timescale between consecutive sampling times is large, it can become computationally cumbersome to work with discrete-time models of reproduction. However, by a suitable rescaling of time, population size, and population genetic parameters, one can obtain a continuous-time process (the Wright-Fisher diffusion) which accurately approximates the population allele frequency of the discrete-time Wright-Fisher model. The key quantity needed when applying the diffusion process is the transition density function, which describes the probability density of the allele frequency changing from value  $x$  to value  $y$  in time  $t$ . This transition density function satisfies a certain partial differential equation (PDE) with coefficients that depend on the mutation and selection parameters. Bollback, York and Nielsen (2008) have used a finite-difference numerical method to approximate the solution to the PDE and incorporated the results into the aforementioned HMM framework to infer the strength of selection from time series data. Recently, an alternative approach (Malaspinas et al., 2012) based on a one-step Markov process has been

proposed to compute the necessary transition densities. In both of these approaches, the allele frequency space has to be discretized finely enough in order to reliably approximate various numerical integrals that are needed for computing the HMM likelihood. The efficiency and accuracy of these grid-based numerical methods depend critically on the spacing and distribution of the discrete grid points. Furthermore, an appropriate choice of this discretization scheme could be strongly dependent on the underlying population genetic parameters. Another limitation of these previous works is that only a few restricted models of selection have been considered. Feder, Kryazhimskiy and Plotkin (2014) recently developed a likelihood-ratio test for identifying signatures of selection from time series data, in which they combined a deterministic model and a Gaussian noise process. This approximation is less accurate than the diffusion approximation, but it facilitates computation and seems sufficiently accurate provided that the allele frequency does not get too close to the boundaries during the period of observation.

In this paper, we develop a novel algorithm based on the spectral method to circumvent the limitations mentioned above. Specifically, instead of approximating the solution to the PDE numerically, we utilize a method recently developed by Song and Steinrücken (2012) which finds an explicit spectral representation of the transition density as a function of  $x$ ,  $y$ , and  $t$ . We show that the probability of observing a given time series dataset can be computed analytically by combining the spectral representation with the forward algorithm for HMMs to *efficiently* and *analytically* integrate over all population allele frequency trajectories. The key idea in our work is to represent the intermediate densities in the forward algorithm in the basis of eigenfunctions of the infinitesimal generator of the Wright-Fisher diffusion process. Exploiting the spectral representation of the transition density, we can then efficiently compute the coefficients in this basis representation. Furthermore, since this spectral representation applies to general diploid models of selection, we are able to leverage this representation to consider more complex models of selection than previously possible. We first demonstrate the accuracy of our method on simulated data. We then apply the method to analyze time series ancient DNA data from genetic loci (ASIP and MC1R) that are associated with horse coat coloration. In contrast to the conclusions of previous studies which considered only a few special models of selection (Ludwig et al., 2009; Malaspina et al., 2012), our exploration of the full parameter space of general diploid selection reveals that a heterozygote-advantage form of balancing selection may have been acting on these loci. We implemented the algorithms described in this paper in a publicly available software package called *spectral HMM*<sup>1</sup>.

The remainder of this paper is organized as follows. In Section 2, we formally introduce the HMM framework and describe the details of our spectral algorithm. The proofs of the theoretical results underlying our algorithm are provided in the supplemental article (Steinrücken, Bhaskar and Song, 2014). In Section 3, we use simulated data to investigate the statistical properties of our maximum likelihood estimator and also apply our method to analyze the aforementioned ancient DNA data for the loci associated with horse coat

---

<sup>1</sup>Available from <http://spectralhmm.sf.net>

coloration (Ludwig et al., 2009). We conclude in Section 4 with a discussion of future extensions of our model.

## 2. Method

Here we provide a formal description of the time series data considered in this paper and present our inference method for analyzing such data.

### 2.1. Time series allele frequency data

The data we analyze consist of genotype samples obtained from individuals at  $K$  distinct times  $t_1 < \dots < t_K$  in the past (given in years). The present time is denoted by  $t_{\text{present}} = t_K$ . At each time point  $t_k$ , a sample of  $n_k \in \mathbb{N}$  individuals is randomly drawn from the population. We assume that the locus under consideration is biallelic, and that the identities of the ancestral allele  $A_0$  and the derived allele  $A_1$  are known. We also assume that the allele  $A_1$  became selected at some time  $t_0 < t_1$ . We use  $d_k$  to denote the number of derived alleles in the sample of  $n_k$  alleles drawn at time  $t_k$ , where  $0 \leq d_k \leq n_k$ . For notational convenience, we use  $o_k$  to denote the tuple  $(t_k, n_k, d_k)$ , and  $O_{[i:j]}$  to denote the partial sequence of observations  $o_i, o_{i+1}, \dots, o_j$ . Figure 1 shows an example of a time series allele frequency dataset with samples drawn at three time points.

### 2.2. The diffusion approximation

Consider a locus evolving according to a discrete Wright-Fisher model of random mating with an effective population size of  $N_e$  diploids. Let  $u_{01}$  be the per-generation probability of mutation from the ancestral allele  $A_0$  to the derived allele  $A_1$ , and  $u_{10}$  the probability of the reverse mutation. We use  $s_i$  to denote the selection coefficient of an individual with  $i$  copies of the derived allele  $A_1$ , where  $0 \leq i \leq 2$ . Without loss of generality, we can assume that  $s_0 = 0$ . In each generation of reproduction, an offspring randomly chooses a parent having  $i$  copies of the derived allele with probability proportional to  $1 + s_i$ .

Consider the scaling limit where the population size  $N_e \rightarrow \infty$  while the unit of time is rescaled by  $N_e$  and the population-scaled parameters  $(2N_e s_1, 2N_e s_2, 4N_e u_{01}, 4N_e u_{10})$  approach some constants. In this limit, the trajectory of the population frequency of allele  $A_1$  follows a Wright-Fisher diffusion process (Ewens, 2004). The unit of time  $\tau$  in this diffusion approximation is related to the physical unit of time  $t$  as

$$\tau = t / (2N_e g),$$

where  $g$  is the average number of years per generation of reproduction. Similarly, we let  $\tau_k$  denote the population-scaled versions of the physical times  $t_k$ , where

$$\tau_k = t_k / (2N_e g). \quad (2.1)$$

The population-scaled selection and mutation parameters of the Wright-Fisher diffusion process are related to the corresponding parameters in physical units as

$$\sigma_i = 2N_e s_i, \quad (2.2)$$

$$\alpha = 4N_e u_{01}, \quad (2.3)$$

$$\beta = 4N_e u_{10}. \quad (2.4)$$

From here on, we use the above population-scaled parameters when describing our analysis of the Wright-Fisher diffusion. The initial population frequency of the allele  $A_1$  when it became selected at time  $\tau_0$  is distributed according to the density function  $\rho(y)$ . In this paper, we are interested in estimating the selection coefficients of the heterozygote and  $A_1$ -homozygote ( $s_1$  and  $s_2$ , respectively) given the other population genetic parameters and assuming that the allele  $A_1$  became selected at time  $\tau_0$ .

### 2.3. Hidden Markov model framework

To analyze the time series data described earlier, we employ a hidden Markov model (HMM) framework as in Bollback, York and Nielsen (2008). In this approach, the population-wide frequency  $Y(\tau)$  of the  $A_1$  allele at time  $\tau$  is modeled as an unobserved hidden variable (see Figure 1). We denote a realization of the frequencies at the sampling times  $\tau_k$  by  $y_k \equiv Y(\tau_k)$ . The initial frequency at time  $\tau_0$  is distributed according to the density function  $\rho$ , i.e.  $Y(\tau_0) \sim \rho$ . For example, the density function  $\rho(y) = \delta(y - 1/(2N_e))$  models the case where the selected allele  $A_1$  arose as a *de novo* mutation in one individual of the population at time  $\tau_0$ .

The probability of transitioning from frequency  $y_{k-1}$  at time  $t_{k-1}$  to frequency  $y_k$  at time  $t_k$  is described by the transition density function  $p_{\Theta}(\tau_k - \tau_{k-1}; y_{k-1}, y_k)$  of the Wright-Fisher diffusion process, where  $\Theta = (\sigma_1, \sigma_2, \alpha, \beta, \tau_0, N_e)$  and  $\tau_k$  are population-scaled parameters as given in equations (2.1)–(2.4). The observations in the HMM are the number of copies  $d_k$  of the allele  $A_1$  among the  $n_k$  alleles in the sample drawn at time  $t_k$ . The probability of such an observation at time  $t_k$  with population allele frequency  $y_k$  is given by the probability mass function  $\xi(d_k; n_k, y_k)$  of a binomial distribution

$$\xi(d_k; n_k, y_k) := \binom{n_k}{d_k} y_k^{d_k} (1 - y_k)^{n_k - d_k}.$$

To compute the probability  $\mathbb{P}_{\Theta}\{O_{[1:K]}\}$  of observing the data  $O_{[1:K]}$  under the model parameters  $\Theta$ , we introduce the forward density functions  $f_k$ , given by

$$f_k(y) dy := \mathbb{P}_{\Theta}\{O_{[1:k]}, Y(\tau_k) \in dy\}, \quad k \in \{0, 1, \dots, K\}. \quad (2.5)$$

The function  $f_k$  is the joint density of the observed data up to time  $\tau_k$  and the hidden population allele frequency at time  $\tau_k$ . We also find it convenient to consider a second auxiliary density function,  $g_k$ , given by

$$g_k(y)dy := \mathbb{P}_\Theta\{O_{[1:k-1]}, \mathbf{Y}(\tau_k) \in dy\}, \quad k \in \{1, \dots, K\}. \quad (2.6)$$

This function  $g_k$  is the joint density of the observed data up to time  $\tau_{k-1}$  and the hidden frequency at  $\tau_k$ . The forward density function  $f_0$  is given by the density function for the initial allele frequency as

$$f_0(y) = \rho(y).$$

Since we approximate the time evolution of the hidden population allele frequency by the Wright-Fisher diffusion, we can get a recurrence relation between the density functions  $g_k$  and  $f_{k-1}$  by integrating over all possible allele frequencies at  $\tau_{k-1}$ :

$$g_k(y) = \int_0^1 f_{k-1}(x) p_\Theta(\tau_k - \tau_{k-1}; x, y) dx, \quad (2.7)$$

where  $k \in \{1, \dots, K\}$ . Using the binomial distribution for sampling  $d_k$  derived alleles out of  $n_k$  individuals at time  $\tau_k$ , we get another recurrence relation between the density functions  $f_k$  and  $g_k$  as follows:

$$f_k(y) = g_k(y) \xi(d_k; n_k, y). \quad (2.8)$$

Finally, the probability  $\mathbb{P}_\Theta\{O_{[1:K]}\}$  of observing the data is computed by integrating over all possible hidden frequencies at the last sampling time:

$$\mathbb{P}_\Theta\{O_{[1:K]}\} = \int_0^1 f_K(y) dy. \quad (2.9)$$

Note that the equations above describe a forward-in-time procedure for computing the probability of the data  $O_{[1:K]}$ , where the intermediate density functions have a natural interpretation.

While (2.7), (2.8) and (2.9) succinctly describe the sampling probability of the data  $O_{[1:K]}$ , no analytic solutions to the integrals in (2.7) and (2.9) are known. In the previous approaches mentioned in Introduction, these integrals were approximated numerically by discretizing the allele frequency state space. The accuracy of these approximations depends critically on the careful choice of the discretization grid. We present an analytical solution to this problem which obviates the need for such a discretization.

#### 2.4. Spectral representation of the transition density

The biallelic Wright-Fisher diffusion with general diploid selection has the infinitesimal generator  $\mathcal{L}$  given by

$$\mathcal{L} = \mathcal{L}_0 + 2x(1-x) [\sigma_1(1-2x) + \sigma_2x] \frac{\partial}{\partial x}, \quad (2.10)$$

where  $\mathcal{L}_0$  is the infinitesimal generator of the diffusion process without selection, given by

$$\mathcal{L}_0 = \frac{1}{2}x(1-x)\frac{\partial^2}{\partial x^2} + \frac{1}{2}[\alpha(1-x) - \beta x]\frac{\partial}{\partial x}. \quad (2.11)$$

We refer the reader to Ewens (2004) for more details about the Wright-Fisher diffusion. Song and Steinrücken (2012) developed an efficient method to compute the eigenvalues and eigenfunctions of  $\mathcal{L}$ , and we utilize that method here. A brief summary of their approach is provided below.

To approximate the spectral decomposition of the operator  $\mathcal{L}$ , consider the functions

$$H_m^{(\Theta)}(x) := e^{-\bar{\sigma}(x)/2} R_m^{(\alpha, \beta)}(x), \quad (2.12)$$

where  $\bar{\sigma}(x) := 4\sigma_1 x(1-x) + 2\sigma_2 x^2$  is the mean fitness of the population and  $R_m^{(\alpha, \beta)}(x)$  are a rescaled version of the classical orthogonal Jacobi polynomials and are defined in Section B of the supplemental article (Steinrücken, Bhaskar and Song, 2014). The  $\alpha$  and  $\beta$  parameters in (2.12) are the population-scaled mutation rates given in (2.3) and (2.4). The set

$\{H_m^{(\Theta)}(x)\}_{m \in \mathbb{N}_0}$  forms a basis for the Hilbert space  $L^2([0, 1], \pi)$  of real-valued functions on  $[0, 1]$  that are square integrable with respect to the stationary density  $\pi$  of the diffusion generator  $\mathcal{L}$ . Specifically,

$$\pi(x) = e^{\bar{\sigma}(x)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (2.13)$$

The basis elements  $H_m^{(\Theta)}(x)$  are orthogonal with respect to the inner product  $\langle \cdot, \cdot \rangle_\pi$  defined by  $\langle f, g \rangle_\pi = \int_0^1 f(x)g(x)\pi(x)dx$ .

In the basis  $\{H_m^{(\Theta)}(x)\}_{m \in \mathbb{N}_0}$  the operator  $\mathcal{L}$  is given by the matrix

$$M := - \left( \mathbf{\Lambda}^{(\alpha, \beta)} + \sum_{l=0}^4 q_l^{(\Theta)} \mathbf{G}^l \right), \quad (2.14)$$

where  $\mathbf{\Lambda}^{(\alpha, \beta)} := \text{diag}(\lambda_0^{(\alpha, \beta)}, \lambda_1^{(\alpha, \beta)}, \dots)$  is a diagonal matrix containing the eigenvalues of the neutral diffusion generator  $\mathcal{L}_0$ ,  $\mathbf{G} := (G_{n,m}^{(\alpha, \beta)})_{n,m \in \mathbb{N}_0}$  is the matrix of coefficients from the three-term recurrence relation for the Jacobi polynomials  $R_m^{(\alpha, \beta)}(x)$ , and  $q_l^{(\Theta)}$  are constant coefficients defined in Section C of the supplemental article (Steinrücken, Bhaskar and Song, 2014). Explicit expressions for the entries of  $\mathbf{\Lambda}^{(\alpha, \beta)}$  and  $\mathbf{G}$  are provided in equations (B.3) and (B.5), respectively, in Section B of the supplemental article.



The eigenvalues  $\lambda_n$  of the full diffusion generator  $\mathcal{L}$  are given by the eigenvalues of  $\mathbf{M}$ , and the coefficients of the eigenfunctions of  $\mathcal{L}$  in the basis  $\{H_m^{(\Theta)}(x)\}_{m \in \mathbb{N}_0}$  are given by the eigenvectors of  $\mathbf{M}$ . In particular, the eigenfunction  $B_n$  of  $\mathcal{L}$  is given by

$$B_n(x) = \sum_{m=0}^{\infty} w_{n,m} H_m^{(\Theta)}(x), \quad (2.15)$$

where  $\mathbf{w}_n = (w_{n,0}, w_{n,1}, \dots)$  is the eigenvector of  $\mathbf{M}$  corresponding to eigen-value  $\lambda_n$ . We use  $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \dots)$  to denote the diagonal matrix of eigenvalues of  $\mathbf{M}$ , and  $\mathbf{W}$  to denote the matrix with rows given by the eigenvectors  $\mathbf{w}_n$ . As can be seen from (2.15),  $\mathbf{W}$  is the change-of-basis matrix between the basis of eigenfunctions  $B_n$  of  $\mathcal{L}$  and the basis  $\{H_m^{(\Theta)}(x)\}_{m \in \mathbb{N}_0}$ .

The leading eigenvalues and the associated eigenvectors of the infinite matrix  $\mathbf{M}$  can be approximated by the eigenvalues and eigenvectors of sufficiently large submatrices of  $\mathbf{M}$ . We refer the reader to Song and Steinrücken (2012) for a more detailed empirical discussion on how the approximation accuracy varies for different submatrix sizes and different parameter regimes. The transition density function  $p_{\Theta}(\tau; x, y)$  for the probability density of the allele changing frequency from  $x$  to  $y$  in time  $\tau$  is given by the following spectral decomposition,

$$p_{\Theta}(\tau; x, y) = \sum_{n=0}^{\infty} e^{-\lambda_n \tau} \pi(y) \frac{B_n(x) B_n(y)}{\langle B_n, B_n \rangle_{\pi}}. \quad (2.16)$$

## 2.5. Incorporating the spectral representation into the HMM

Using the spectral decomposition of the transition density function in (2.16), we devise a dynamic programming algorithm to compute the likelihood  $\mathbb{P}_{\Theta} \{O_{[1:K]}\}$ . This algorithm recursively computes the density functions  $f_k$  and  $g_k$  given in (2.5) and (2.6), respectively. To update these density functions efficiently, we represent them in the basis of scaled eigenfunctions  $\{\pi(y) B_n(y)\}_{n \in \mathbb{N}_0}$  of the diffusion generator  $\mathcal{L}$ . More precisely, we express  $f_k$  and  $g_k$  as

$$f_k(y) = \pi(y) \mathbf{b}_k \mathbf{B}(y) = \sum_{n=0}^{\infty} b_{k,n} \pi(y) B_n(y), \quad (2.17)$$

$$g_k(y) = \pi(y) \mathbf{a}_k \mathbf{B}(y) = \sum_{n=0}^{\infty} a_{k,n} \pi(y) B_n(y), \quad (2.18)$$

where we employ the vector notation

$$\mathbf{b}_k := (b_{k,0}, b_{k,1}, \dots), \quad (2.19)$$

$$\mathbf{a}_k := (a_{k,0}, a_{k,1}, \dots), \quad (2.20)$$

$$\mathbf{B}(y) := (B_0(y), B_1(y), \dots)^T. \quad (2.21)$$

We now describe how the coefficient vectors  $\mathbf{a}_k$ ,  $\mathbf{b}_k$  and the probability  $\mathbb{P}_\pi \{O_{[1:K]}\}$  can be computed efficiently. All proofs can be found in Section A of the supplemental article (Steinrücken, Bhaskar and Song, 2014). First, the following proposition determines the vector  $\mathbf{b}_0$  of coefficients for the initial forward density function  $f_0$ :

**Proposition 1**—If the allele frequency at  $\tau_0$  is distributed according to the density function  $\rho(y) = \delta(y-x)$ , then the initial forward density function  $f_0$  in the basis  $\{e^{(\pi(y)B_n(y))}\}_{n \in \mathbb{N}_0}$  has the vector of coefficients

$$\mathbf{b}_0 = \left( \frac{B_0(x)}{c_0}, \frac{B_1(x)}{c_1}, \dots \right),$$

where  $B_n(x)$  is given by (2.15), and  $c_n$  are the squared norms of  $B_n$  given by

$$c_n = \langle B_n, B_n \rangle_\pi = \sum_{m=0}^{\infty} (w_{n,m})^2 c_n^{(\alpha,\beta)}, \quad (2.22)$$

where  $c_n^{(\alpha,\beta)}$  denote the squared norms of the Jacobi polynomials given in equation (B.2) in Section B of the supplemental article (Steinrücken, Bhaskar and Song, 2014).

In the case where the selected allele  $A_1$  arises from *de novo* mutation at  $t_0$  in one of the individuals in the population, we set  $x = 1/(2N_e)$  in Proposition 1. We note that our framework allows us to easily model other distributions for the frequency of the mutant allele  $A_1$  when it became selected. For example, the initial distribution of mutation-drift balance can be used to model selection arising from standing genetic variation. Some of these initial distributions are described in Section D of the supplemental article (Steinrücken, Bhaskar and Song, 2014).

The following theorem establishes how the representations of the densities  $f_k$  and  $g_k$ , for  $k > 0$ , can be computed algebraically in a recursive fashion:

**Theorem 2**—Let  $\mathbf{C} := \text{diag}(c_0^{(\alpha,\beta)}, c_1^{(\alpha,\beta)}, \dots)$  and  $\mathbf{D} := \text{diag}(c_0, c_1, \dots)$  denote diagonal matrices with entries  $c_n^{(\alpha,\beta)}$  and  $c_n$  defined as in Proposition 1. Then, for each  $k \in \{1, \dots, K\}$ , the coefficients in the representation of the densities  $g_k(y)$  and  $f_k(y)$  in (2.17) and (2.18) can be computed recursively as

$$\mathbf{a}_k = \mathbf{b}_{k-1} \exp[-\mathbf{\Lambda}(\tau_k - \tau_{k-1})], \quad (2.23)$$

$$\mathbf{b}_k = \mathbf{a}_k \mathbf{W} \mathbf{G}^{d_k} (\mathbf{1} - \mathbf{G})^{n_k - d_k} \mathbf{W}^{-1}, \quad (2.24)$$

where  $\mathbf{W}^{-1}$  is given by

$$\mathbf{W}^{-1} = \mathbf{C} \mathbf{W}^T \mathbf{D}^{-1}. \quad (2.25)$$

Combining Proposition 1 and Theorem 2, we obtain a dynamic programming algorithm for calculating the coefficients  $\mathbf{b}_k$  and  $\mathbf{a}_k$  in the representations for  $f_k$  and  $g_k$  given in (2.17) and (2.18), respectively. The vectors and matrices appearing in the above results are infinite dimensional. As in previous works (Song and Steinrücken, 2012; Steinrücken, Wang and Song, 2013) on the spectral representation of the transition density, when applying the above results we truncate the infinite vectors and matrices by choosing cutoffs for the dimensions. We provide more practical details in Section 3.3.

Finally, the probability of observing the full data  $O_{[1:K]}$  can be computed using the following proposition:

**Proposition 3**—The probability  $\mathbb{P}_\Theta \{O_{[1:K]}\}$  of observing the data  $O_{[1:K]}$  given the population genetic parameters  $\Theta$  is

$$\mathbb{P}_\Theta \{O_{[1:K]}\} = \frac{c_0}{B_0(0)} b_{K,0}, \quad (2.26)$$

where  $B_0(0)$  is given by

$$B_0(0) = \sum_{m=0}^{\infty} (-1)^m w_{0,m} \frac{\Gamma(m+\alpha)}{\Gamma(m+1)\Gamma(\alpha)}.$$

### 3. Results

In this section, we perform parametric inference via the maximum likelihood framework, using a finite grid in the parameter space. We first test the accuracy on simulated data and then apply it to analyze an ancient DNA dataset related to coat coloration in domesticated horses (Ludwig et al., 2009).

Since ancient DNA data are often collected from only those loci which are segregating at the present time, in our empirical study we condition on observing at least one copy of the derived allele at the last sampling time  $\tau_K$ . In particular, the likelihood of the parameters is given by  $L(\Theta) := \mathbb{P}_\Theta \{O_{[1:K]} \mid d_K > 0\}$ . We chose to maximize this function on a grid, since the algorithm described in the previous section can be parallelized, thus allowing to efficiently evaluate the likelihood under given parameters for several datasets at once.

### 3.1. Performance on simulated data

We simulated data under a discrete-time Wright-Fisher model with several values for the effective population size and selection coefficients. We chose the mutation probabilities to be  $u_{01} = u_{10} = 10^{-6}$ , and the number of years per generation to be five years. These parameters are similar to those considered by previous works that analyzed time series allelic samples from the ASIP and MC1R loci in horses (Ludwig et al., 2009; Malaspina et al., 2012). In our simulations, 5% of the population carried the mutant allele when it first became positively selected. We sampled 40 individuals at each of 10 time points over the course of 32,000 years.

We investigated the performance of our maximum likelihood estimator in various scenarios of selection. Here, we present the results for the following four particular selection schemes:

1. Genic selection, in which the selective fitness of the heterozygote is the arithmetic mean of the fitness of the two homozygotes, i.e.  $s_1 = s/2$  and  $s_2 = s$ .
2. Heterozygote advantage selection, in which  $s_1 = s$  and  $s_2 = 0$ .
3. Recessive selection, in which  $s_1 = 0$ ,  $s_2 = s$ .
4. Dominant selection, in which  $s_1 = s$ ,  $s_2 = s$ .

For each scenario, we considered  $s \in \{0, 0.001, 0.0025, 0.005, 0.01\}$  and simulated 200 datasets for each value of  $s$ .

Figure 2 shows the performance of the maximum likelihood estimator under a model of genic selection with an effective population size of  $N_e = 2,500$  and  $N_e = 10,000$ . It illustrates empirical boxplots of the maximum likelihood estimates, where the tips of the whiskers denote the 2.5%-quantile and the 97.5%-quantile, and the boxes represent the upper and lower quartile. As the figure shows, our maximum likelihood estimates are unbiased. The uncertainty of the estimate tends to increase with increasing values of  $s$ , while the uncertainty decreases as the population size increases, illustrating the fact that for larger population sizes, selection acts more efficiently and is easier to detect. In the case of  $N_e = 10,000$ , if the true selection coefficient is 0.0025 or more, all our maximum likelihood estimates are higher than the 97.5%-quantile of the empirical distribution of the maximum likelihood estimates for  $s = 0$ . Hence, there is high power to reject neutrality in these scenarios.

The performance of our maximum likelihood estimator for several additional selection schemes and parameter regimes can be found in Figure 3, where we also consider a scenario with fewer sampling time points. The figure shows that our maximum likelihood estimates are unbiased across the different parameter ranges and scenarios. In general, the low variance of the empirical distribution of the maximum likelihood estimates shows that our method can be used to accurately infer the selection parameters of interest in a wide range of scenarios.

### 3.2. Analysis of ancient DNA data: coat coloration in domesticated horses

Ludwig et al. (2009) extracted genotype data at several loci from ancient horse DNA obtained from various sites in Eurasia. In particular, they extracted temporal allele frequency data at eight loci that are known to play a role in coat color determination in contemporary horses. Only the locus encoding for the Agouti signaling peptide (ASIP) and the locus for the melanocortin 1 receptor (MC1R) showed strong fluctuations in the sample allele counts. Table 1 shows the time series data for the ASIP and the MC1R loci in the curated form of the original work (Ludwig et al., 2009).

Using the method of Bollback, York and Nielsen (2008) for the model of genic selection ( $s_1 = s/2$ ,  $s_2 = s$ ), Ludwig et al. (2009) established that selection acted significantly on only the ASIP and the MC1R loci. However, another recent analysis (Malaspinas et al., 2012) of the same dataset considered the model of recessive selection ( $s_1 = 0$ ,  $s_2 = s$ ) and did not find a significant signal of selection at the ASIP locus.

To investigate the dependence of the previous conclusions on the assumed selection scheme, we applied our method to reanalyze the ASIP and the MC1R data under a general selection scheme with arbitrary selection coefficients  $s_1$  and  $s_2$ . We set the mutation probability to  $u_{01} = u_{10} = 10^{-6}$  and the average length of a generation to 5 years. Table 1 shows that the derived allele is absent in both datasets at time 20,000 BCE. Thus, we set the initial frequency of the derived allele as  $1/2N_e$ , corresponding to the case where the selected allele arises as a *de novo* mutation at time  $t_0$ . We tried a range of values for  $N_e$  and  $t_0$ .

Figure 4(a) shows the likelihood surface for the temporal allele frequency data from the ASIP locus, for  $N_e = 2,500$  and  $t_0 = 17,000$  BCE. The empirical maximum of the likelihood surface is located at  $(s_1, s_2) = (0.0025, 0)$ , indicated by the 'x' in Figure 4(a). This maximum suggests that a selective scheme of heterozygote advantage best explains the data, where both the ancestral and derived allele homozygotes are of equal fitness, while the heterozygous genotype confers a selective advantage over the homozygotes. To establish the significance of this finding, we performed the following bootstrap procedure: We resampled the ASIP dataset 100 times to obtain sub-sampled datasets  $\{O_{[1:K]}^{(j)}\}_{j=1}^{100}$ . For each bootstrapped dataset  $1 \leq j \leq 100$ , we resampled  $n_k^{(j)} = n_k$  alleles at each time  $t_k^{(j)} = t_k$ . The number of derived alleles for dataset  $j$  was obtained by binomial sampling from the empirical frequency of derived alleles in the original ASIP dataset, i.e.,

$$d_k^{(j)} \sim \xi \left( \cdot; n_k^{(j)}, \frac{d_k}{n_k} \right).$$

We then reported the empirical maximum of the likelihood surface for each of these resampled datasets. Figure 4(b) shows the empirical maximum likelihood estimates and marginal histograms of the maxima for the 100 re-sampled datasets. The marginal 2.5% and 97.5% quantiles of the empirical distribution are  $[0.0025, 0.0235]$  for the heterozygote fitness  $s_1$  and  $[0, 0.0045]$  for the derived allele homozygote fitness  $s_2$ , thus providing further evidence that the data are significantly better explained by a selection model where a

heterozygous individual is selectively advantageous over the homozygous individuals. As Figure 5 shows, changing  $N_e$  from 1,000 to 10,000, or changing  $t_0$  from 19,000 BCE to 15,000 BCE has only a minimal effect on the shape of the likelihood surface and maximum likelihood estimate, again supporting that a selective scheme of heterozygote advantage best explains the data.

A similar analysis of the MC1R locus can be found in Figures 6 and 7. For this dataset, the maximum of the likelihood surface is attained at  $(s_1, s_2) = (0.004, 0.0015)$ , and the empirical marginal 2.5% and 97.5%-quantiles are  $[0.001, 0.025]$  for the heterozygote fitness and  $[-0.009, 0.0135]$  for the derived allele homozygote fitness. Together with the results shown in Figure 7, this suggests that the data at the MC1R locus is also best explained by a selection model of heterozygote advantage. However, although the marginal quantiles for the homozygote fitness cover  $s_2 = 0$ , they are rather far apart, so the evidence of heterozygote advantage for the MC1R locus is weaker than that for the ASIP locus.

### 3.3. Computational performance

The running time of our algorithm for computing the likelihood of a given set of population-scaled parameters is dependent on the dimensions of the truncation  $\tilde{\mathbf{M}}$  of the infinite matrix  $\mathbf{M}$  given in (2.14). In particular, the time complexity of computing a single likelihood is the cost of computing the eigenvalues and eigenvectors of  $\tilde{\mathbf{M}}$  plus the cost of computing the coefficients  $b_k$  in Theorem 2, where  $k \in \{1, \dots, K\}$ . To compute the eigenvalues and eigenvectors of  $\tilde{\mathbf{M}}$  to high precision, we first used LAPACK<sup>2</sup> to compute them to double precision, and then refine them by using inverse iteration (Press et al., 2007, Chapter 11.8). Each step of the inverse iteration involves solving a linear system with matrix  $\tilde{\mathbf{M}} - \mu\mathbf{I}$ , where  $\mu$  is an estimate for an eigenvalue of  $\tilde{\mathbf{M}}$ . Since this matrix has bandwidth at most 9, this linear system can be solved in  $O(D)$  time, where  $D$  is the dimension of  $\tilde{\mathbf{M}}$ . By using the repeated squaring algorithm for taking powers of the matrices  $\mathbf{G}$  and  $\mathbf{1} - \mathbf{G}$  and exploiting the fact that  $\mathbf{G}$  and  $\mathbf{1} - \mathbf{G}$  are tridiagonal matrices, each coefficient  $b_k$  can be computed in  $O(D^2 + D \min(D, n_k)^2 \log n_k)$  time, where the first  $O(D^2)$  term comes from the matrix-vector multiplications in (2.24).

For the analysis of the ASIP and MC1R datasets reported in Figures 4 and 6, we approximated the eigenvalues and eigenvectors of  $\mathbf{M}$  defined in (2.14) using a  $600 \times 600$  submatrix. Furthermore, we used the first 590 terms in (2.15) to approximate the eigenfunctions, and the dimensions of the vectors of coefficients in (2.19) and (2.20) were set to 580. We empirically verified that these cutoffs produced a stable approximation of the likelihood. Using these values the computation time for a single point of the grid in Figure 4(a) was approximately 95 seconds. We adjusted the cutoffs appropriately for the other analyses reported in Section 3.

<sup>2</sup>Available from <http://www.netlib.org/lapack/>

## 4. Discussion

In this paper, we have developed a novel, efficient spectral algorithm to analyze time series allele frequency data under a general diploid selection model. We have demonstrated that our method can be used to accurately estimate selection parameters on simulated data.

We have also applied our method to investigate loci involved in horse coat coloration. Our inferred selection coefficients show that the data are best explained by a heterozygote advantage model of balancing selection. As mentioned earlier, Ludwig et al. (2009) provided evidence for slightly positive selection at the ASIP locus, assuming a model of genic selection (where  $s_1 = s_2/2$ ). More precisely, they obtained a point estimate of  $s_2 = 0.0007$  and a 95% confidence interval of  $[0.0001, 0.0015]$ . However, using a model of selection where the derived allele homozygote is recessive (i.e.,  $s_1 = 0$ ), a subsequent re-analysis (Malaspina et al., 2012) of the same data found that  $s_2$  has a point estimate of  $-0.001$  with a 95% confidence interval of  $[-0.02, 0.051]$ , thus not rejecting neutrality at the ASIP locus. In our work, we have allowed our method to explore the two-dimensional parameter space of general diploid selection models and presented evidence for a selection mode where heterozygous individuals are advantageous over homozygous individuals. It is possible that previous analyses have only been able to infer very weak selection acting at the ASIP locus because they have restricted the model of selection to certain one-dimensional models. Indeed, if we restrict our analysis to a model of genic selection, we get results similar to those reported by Ludwig et al. (2009). Our analysis does not conclusively prove that individuals that were heterozygous at the ASIP locus had a constant evolutionary advantage since 17,000 BCE, because we have ignored the interaction of selection and demographic history, epistatic interactions between loci, time-varying models of selection, and other factors. However, our results suggest the possibility that some mode of heterozygote advantage balancing selection has maintained polymorphism at the ASIP locus that is involved in horse coat coloration.

Although we have focused on time series samples taken at a biallelic locus, the mathematical framework presented here could be readily extended to handle an arbitrary number of alleles using the spectral representation derived by Steinrücken, Wang and Song (2013). Further, changes in the population size and selection coefficients could be modeled by suitably combining the spectral representations for different population genetic parameters at the change points. It is also possible to extend the method to multiple populations and to incorporate samples taken from extinct ancestral populations. In light of emerging ancient DNA sequence data for ancient hominids (Green et al., 2010; Reich et al., 2010), such temporal sequence data and inference methods present novel opportunities to gain insight into adaptation in humans. For a more adequate modeling of biologically relevant scenarios, it is also necessary to incorporate the exchange of migrants into the model (Gutenkunst et al., 2009; Luki, Hey and Chen, 2011), and extend the framework to incorporate variation at linked loci. By taking advantage of genetic hitchhiking at closely linked sites during the course of selective sweeps, one might be able to further improve the inference of selection coefficients.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

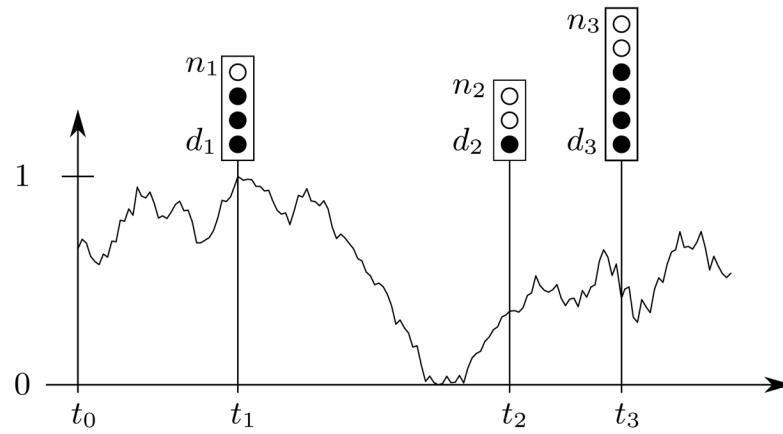
We thank Rasmus Nielsen, Joshua Schraiber, and Montgomery Slatkin for helpful comments and discussions. We also thank Karen Kafadar and two anonymous referees for suggestions that improved the exposition of this paper. Moreover, we thank Richard J. Mathar (Mathar, 2009) for making his sourcecode available to us.

## References

- Bollback JP, York TL, Nielsen R. Estimation of  $2N_e s$  From Temporal Allele Frequency Data. *Genetics*. 2008; 179:497–502. [PubMed: 18493066]
- Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD. Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature*. 2010; 467:587–590. [PubMed: 20844486]
- Ewens, W. *Mathematical Population Genetics: I. Theoretical Introduction*. 2. Springer; 2004.
- Fearnhead P. Ancestral processes for non-neutral models of complex diseases. *Theoretical Population Biology*. 2003; 63:115–130. [PubMed: 12615495]
- Fearnhead P. The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theoretical Population Biology*. 2006; 70:376–386. [PubMed: 16563450]
- Feder AF, Kryazhimskiy S, Plotkin JB. Identifying Signatures of Selection in Genetic Time Series. *Genetics*. 2014; 196:509–522. [PubMed: 24318534]
- Genz A, Joyce P. Computation of the normalizing constant for exponentially weighted Dirichlet distribution integrals. *Computing Science and Statistics*. 2003; 35:181–212.
- Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. A draft sequence of the Neandertal genome. *Science*. 2010; 328:710–722. [PubMed: 20448178]
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*. 2009; 5:e1000695. [PubMed: 19851460]
- Hummel S, Schmidt D, Kremeyer B, Herrmann B, Oppermann M. Detection of the CCR5-Delta32 HIV resistance gene in Bronze Age skeletons. *Genes and Immunity*. 2005; 6:371–374. [PubMed: 15815693]
- Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*. 2013; 500:571–574. [PubMed: 23873039]
- Ludwig A, Pruvost M, Reissmann M, Benecke N, Brockmann GA, Castaños P, Cieslak M, Lippold S, Llorente L, Malaspinas A-S, Slatkin M, Hofreiter M. Coat Color Variation at the Beginning of Horse Domestication. *Science*. 2009; 324:485. [PubMed: 19390039]
- Luki S, Hey J, Chen K. Non-equilibrium allele frequency spectra via spectral methods. *Theoretical Population Biology*. 2011; 79:203–219. [PubMed: 21376069]
- Malaspinas AS, Malaspinas O, Evans SN, Slatkin M. Estimating allele age and selection coefficient from time-serial data. *Genetics*. 2012; 192:599–607. [PubMed: 22851647]
- Mathar, RJ. A Java Math.BigDecimal Implementation of Core Mathematical Functions. ArXiv e-prints. 2009. <http://arxiv.org/abs/0908.3030v1>
- Mathieson I, McVean G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics*. 2013; 193:973–984. [PubMed: 23307902]
- Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*. 2013; 499:74–78. [PubMed: 23803765]

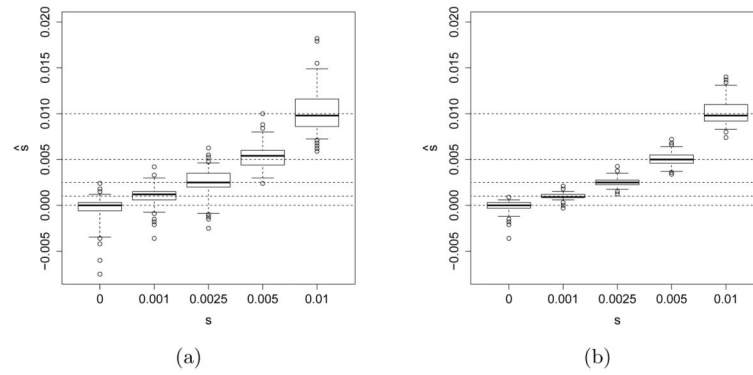


- Orozcooter Wengel P, Kapun M, Nolte V, Kofler R, Flatt T, Schlötterer C. Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular Ecology*. 2012; 21:4931–4941. [PubMed: 22726122]
- Press, W.; Teukolsky, S.; Vetterling, WT.; Flannery, BP. *Numerical Recipes: The Art of Scientific Computing*. 3. Cambridge University Press; 2007.
- Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin J-J, Kelso J, Slatkin M, Pääbo S. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010; 468:1053–1060. [PubMed: 21179161]
- Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of Virology*. 1999; 73:10489–10502. [PubMed: 10559367]
- Song YS, Steinrücken M. A Simple Method for Finding Explicit Analytic Transition Densities of Diffusion Processes with General Diploid Selection. *Genetics*. 2012; 190:1117–1129. [PubMed: 22209899]
- Steinrücken M, Bhaskar A, Song YS. Supplement to “A novel spectral method for inferring general diploid selection from time series genetic data”. 2014
- Steinrücken M, Wang YXR, Song YS. An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. *Theoretical Population Biology*. 2013; 83:1–14. [PubMed: 23127866]
- Stephens M, Donnelly P. Ancestral inference in population genetics models with selection. *Australian & New Zealand Journal of Statistics*. 2003; 45:395–423.
- Williamson EG, Slatkin M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*. 1999; 152:755–761. [PubMed: 10353915]
- Wiser MJ, Ribeck N, Lenski RE. Long-term dynamics of adaptation in asexual populations. *Science*. 2013; 342:1364–1367. [PubMed: 24231808]

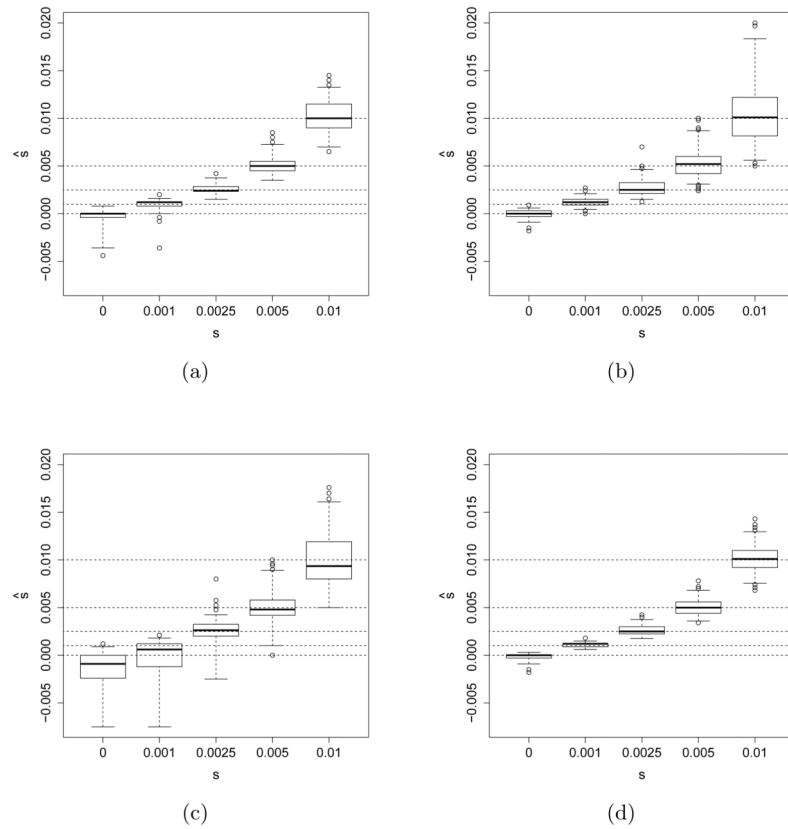


**Fig 1.**

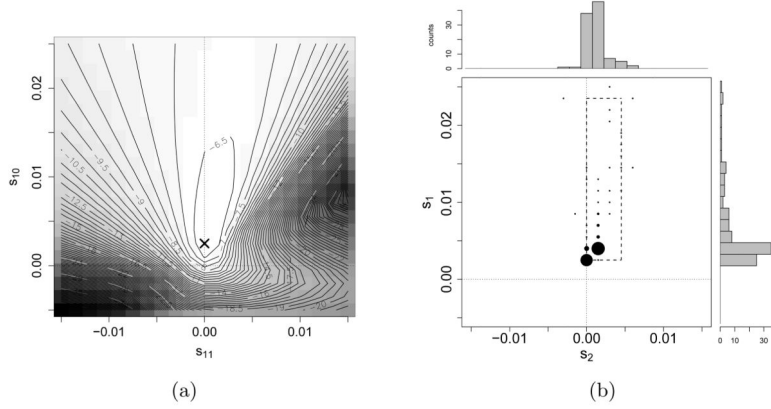
In this example, samples of size  $n_1 = 4$ ,  $n_2 = 3$ , and  $n_3 = 6$  (illustrated by the total number of circles) are taken at times  $t_1$ ,  $t_2$ , and  $t_3$  respectively. The observed number of derived alleles (filled circles) is  $d_1 = 3$ ,  $d_2 = 1$ , and  $d_3 = 4$ . The initial time is  $t_0$ , and the curve indicates a particular trajectory of the underlying population allele frequency  $Y(t) \in [0, 1]$ .



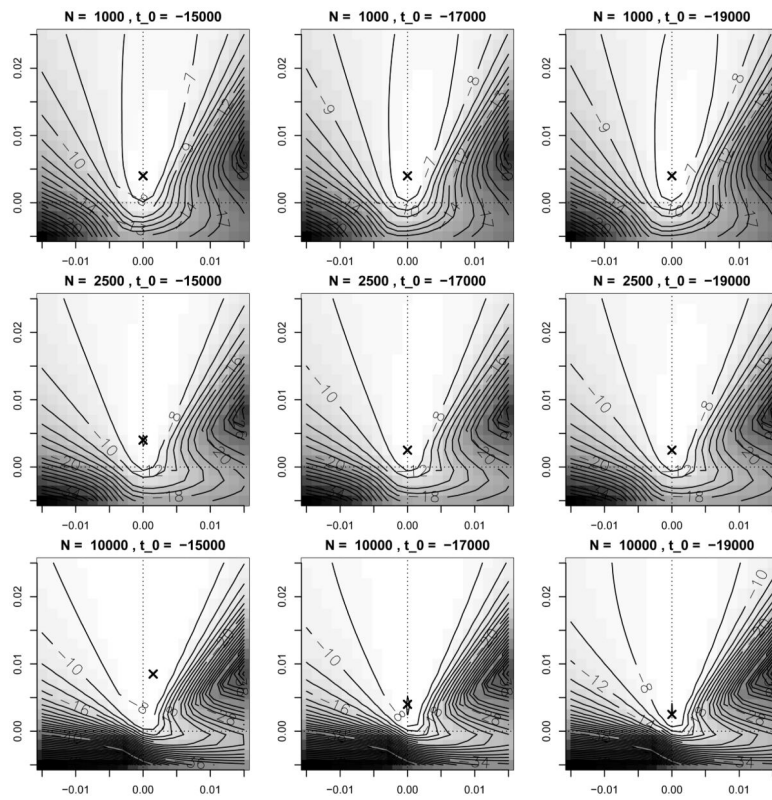
**Fig 2.** Empirical distribution of the maximum likelihood estimates for 200 datasets simulated under a model of genic selection, with heterozygote fitness  $s_1 = s/2$  and derived allele homozygote fitness  $s_2 = s$ , for each of several different values of selection strength  $s$ . The dashed lines indicate the true values. (a) The effective population size  $N_e$  is 2,500 individuals. (b)  $N_e = 10,000$  individuals.



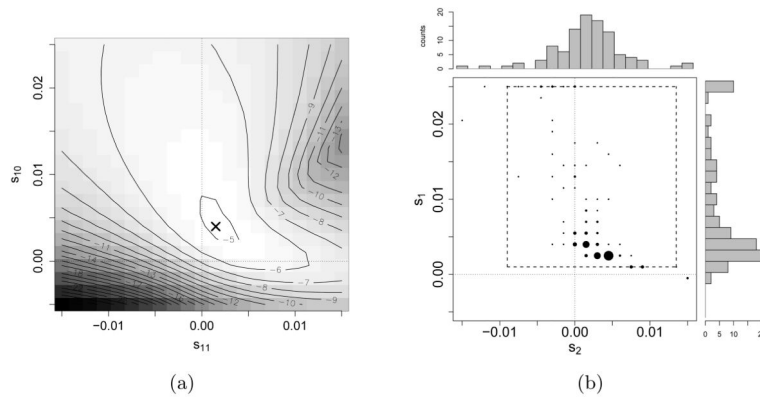
**Fig 3.** Empirical distribution of the maximum likelihood estimates of 200 simulated datasets each under different modes of selection of differing strength with  $N_e = 10,000$ . The dashed lines indicate the true values of  $s$ . (a) Genic selection ( $s_1 = s/2$ ,  $s_2 = s$ ) with only five sampling time points. (b) Heterozygote advantage model of selection ( $s_1 = s$ ,  $s_2 = 0$ ) with ten sampling time points. (c) Recessive selection ( $s_1 = 0$ ,  $s_2 = s$ ) with ten sampling time points. (d) Dominant selection ( $s_1 = s$ ,  $s_2 = s$ ) with ten sampling time points.



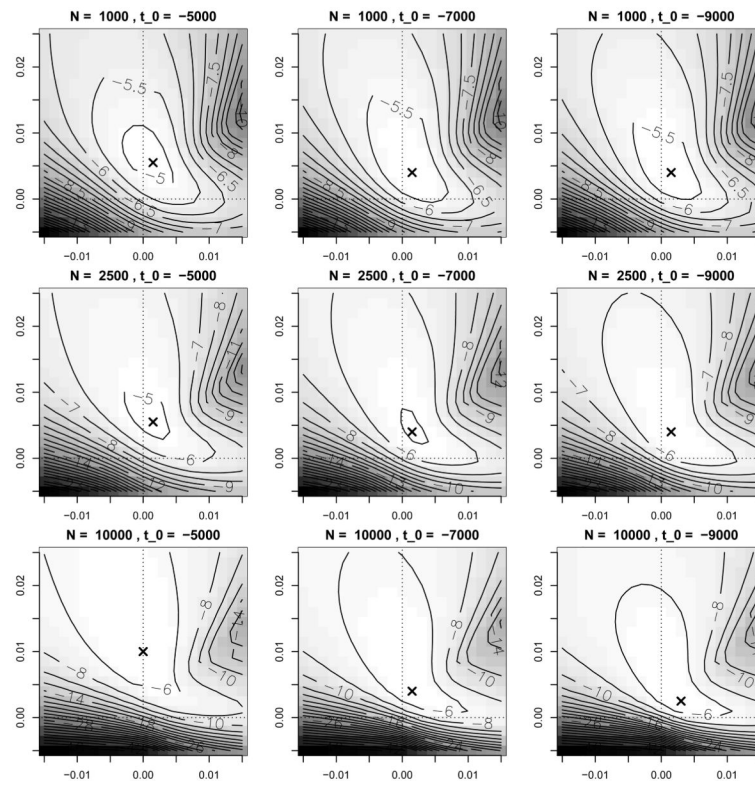
**Fig 4.** Analysis of the ASIP locus. (a) Empirical values of the likelihood  $L(\Theta)$  for temporal samples from the ASIP locus where the likelihood is computed over a  $21 \times 21$  grid. The maximum is attained at  $(s_1, s_2) = (0.0025, 0)$ , indicated by the 'x'. (b) A joint density plot and marginal histograms of the maximum likelihood estimates for 100 bootstrap resampled datasets of the temporal data at the ASIP locus. The circles are centered on the grid points at which the likelihood function is evaluated, and the sizes of the circles indicate the proportion of maximum likelihood estimates that occupy the same grid point. The marginal empirical 2.5% and 97.5%-quantiles are  $[0.0025, 0.0235]$  for the heterozygote fitness  $s_1$ , and  $[0, 0.0045]$  for the derived allele homozygote fitness  $s_2$ , as indicated by the dashed box.



**Fig 5.** Likelihood surfaces for the ASIP dataset under various combinations of  $N_e \in \{1\,000, 2\,500, 10\,000\}$  and  $t_0 \in \{15\,000\text{ BCE}, 17\,000\text{ BCE}, 19\,000\text{ BCE}\}$ . The respective maxima are indicated by an 'x'.



**Fig 6.** Analysis of the MC1R locus using the parameters  $N_e = 2,500$  and  $t_0 = 7,000$  BCE. (a) Likelihood surface for the MC1R locus. The maximum likelihood estimate is at  $(s_1, s_2) = (0.004, 0.0015)$  and is indicated by the 'x'. (b) A joint density plot and marginal histograms of the maximum likelihood estimates for 100 bootstrap resampled datasets obtained from the MC1R data as described in Section 3.2. The marginal 2.5% and 97.5%-quantiles are  $[0.001, 0.025]$  for the heterozygote fitness  $s_1$  and  $[-0.009, 0.0135]$  for the derived allele homozygote fitness  $s_2$ , as indicated by the dashed box.



**Fig 7.** Likelihood surfaces for the MC1R dataset under various combinations of  $N_e \in \{1\,000, 2\,500, 10\,000\}$  and  $t_0 \in \{5\,000\text{ BCE}, 7\,000\text{ BCE}, 9\,000\text{ BCE}\}$ . The respective maxima are indicated by an 'x'.



The temporal allele frequency datasets for the ASIP and MC1R loci associated with coat coloration in domesticated horses (Ludwig et al., 2009, Figure S. 3). For each sampling time  $t_k$  (given in years BCE), the table lists the number  $d_k$  of derived alleles among the sampled  $n_k$  alleles.

**Table 1**

time of sampling [ $t_k$ ] (BCE)	20,000	13,100	3,700	2,800	1,100	500
# of samples [ $n_k$ ]	10	22	20	20	36	38
ASIP (# der. alleles) [ $d_k$ ]	0	1	15	12	15	18
MC1R (# der. alleles) [ $d_k$ ]	0	0	1	6	13	24