Genetics: Early Online, published on April 14, 2015 as 10.1534/genetics.115.175265

Transition densities and sample frequency spectra of diffusion processes with selection and variable population size

Daniel Živković[a,*], Matthias Steinrücken[b,e], Yun S. Song[b,c,d], Wolfgang Stephan[a]

[a]*Section of Evolutionary Biology, Department of Biology, Ludwig-Maximilian University Munich, Munich, Germany*
[b]*Department of Statistics, University of California, Berkeley, CA 94720, USA*
[c]*Computer Science Division, University of California, Berkeley, CA 94720, USA*
[d]*Department of Integrative Biology, University of California, Berkeley, CA 94720, USA*
[e]*Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA 01003, USA*

Abstract

Advances in empirical population genetics have made apparent the need for models that simultaneously account for selection and demography. To address this need, we here study the Wright-Fisher diffusion under selection and variable effective population size. In the case of genic selection and piecewise-constant effective population sizes, we obtain the transition density by extending a recently developed method for computing an accurate spectral representation for a constant population size. Utilizing this extension, we show how to compute the sample frequency spectrum in the presence of genic selection and an arbitrary number of instantaneous changes in the effective population size. We also develop an alternate, efficient algorithm for computing the sample frequency spectrum using a moment-based approach. We apply these methods to answer the following questions: If neutrality is incorrectly assumed when there is selection, what effects does it have on demographic parameter estimation? Can the impact of negative selection be observed in populations that undergo strong exponential growth?

[*] Corresponding author: zivkovic@bio.lmu.de
Running title: Transition densities and frequency spectra for selection and demography

Introduction

Advances in empirical population genetics have pointed out the need for models that simultaneously account for selection and demography. Studies on samples from various species including humans (e.g., Williamson et al. 2005; Tennessen et al. 2012) and *Drosophila melanogaster* (Glinka et al. 2003; Duchen et al. 2013) have shown that demographic processes such as population size changes shape in large part the patterns of polymorphism among genomes and estimated the impact of selection on top of such underlying neutral conditions. Thus far, most theoretical papers considered selective and demographic forces independently of each other for the sake of simplicity (e.g., Stephan and Li 2007).

Theoretical studies of neutral models of time-varying population size have been accomplished within the diffusion and the coalescent frameworks. Kimura (1955a) derived the transition density of the Wright-Fisher (WF) diffusion with a constant population size that characterizes the neutral evolution of allele frequencies over time. Shortly thereafter, Kimura (1955b) noted how to rescale time to generalize this result to a deterministically changing population size. Nei et al. (1975) derived the average heterozygosity under this general condition by applying a differential equation method, before studies on time-varying population size started to utilize the coalescent. Watterson (1984) derived the probability distribution and the moments of the total number of alleles in a sample using models of one or two sudden changes in population size. Slatkin and Hudson (1991) considered the distribution of pairwise differences in exponentially growing populations, before Griffiths and Tavaré (1994) provided the coalescent for arbitrary deterministic changes in population size. The allele frequency spectrum, which is the distribution of the number of times a mutant allele is observed in a sample of DNA sequences, has been utilized in many theoretical and empirical studies. It can be further distinguished into the allelic spectrum and the sample frequency spectrum (SFS) according to whether absolute or relative frequencies are meant. Fu (1995) derived the first- and second-order moments of the allelic spectrum for a constant population size, which has been generalized to time-varying population size by Griffiths and Tavaré (1998) and Živković and Wiehe (2008). Although deterministic fluctuations in population size are commonly considered for the interpretation of biological data, studies have also examined stochastic changes in population size (e.g., Kaj and Krone 2003).

The mathematical modeling of natural selection is mostly carried out within the diffusion framework, whereas coalescent approaches have proved to be analytically challenging (e.g., Krone and

Neuhauser 1997). Fisher (1930) derived the equilibrium solution for the allelic spectrum of a population, which became particularly useful when Sawyer and Hartl (1992) modeled the frequencies of mutant sites via a Poisson random field approach. Kimura (1955c) employed a perturbation approach to obtain a series representation of the transition density that is accurate for scaled selection coefficients smaller than one. However, as noted in Williamson et al. (2005), an appropriate use of this result with respect to the analysis of whole-genome data is even difficult for a constant population size. In a recent paper, Song and Steinrücken (2012) devised an efficient method to accurately compute the transition density of the WF diffusion with recurrent mutations and general diploid selection. This nonperturbative approach that can be applied to scaled selection coefficients substantially greater than one finds the eigenvalues and the eigenfunctions of the diffusion generator and leads to an explicit spectral representation of the transition density. The results for this biallelic case have been extended to an arbitrary number of alleles by Steinrücken et al. (2013). The process dual to this multi-allelic diffusion has been analyzed earlier by Barbour et al. (2000). While providing theoretical insight, their approach does not straightforwardly allow computation of the transition density.

In recent years, several researchers have started to investigate the combined effect of natural selection and demography. The majority of these studies have utilized finite difference schemes to enable tractable computation. Williamson et al. (2005) employed such a scheme to obtain a numerical solution of the SFS for a model with genic selection and one instantaneous population size change. The authors applied this result within a likelihood-based method to infer population growth and purifying selection at non-synonymous sites across the human genome. Evans et al. (2007) investigated the forward diffusion equation with genic selection and deterministically varying population size and incorporated the effect of point mutations via a suitable boundary condition. They derived a system of ODEs for the moments of the allelic spectrum, but had to resort to a numerical scheme to make their results applicable. Gutenkunst et al. (2009) considered population substructure and selection to obtain the joint allele frequency spectrum of up to three populations by approximating the associated diffusion equation by a finite difference scheme as well. Lukić and Hey (2012) applied spectral methods that even account for a fourth population in the otherwise same setting as Gutenkunst et al. (2009). Recently, and again with respect to a single population, Zhao et al. (2013) provided a numerical method to solve the diffusion equation for random genetic drift that can incorporate the forces of mutation and selection. The authors illustrated the accuracy of their discretization approach by determining the probability of fixation in the presence of selection for both an instantaneous population size change and a linear increase in population size. In general, such methods require an appropriate discretization of grid points, which may depend strongly on the parameters. This makes it difficult, however, to predict if a particular discretization will produce accurate results.

In this study, we use the polynomial approach by Song and Steinrücken (2012) to obtain the transition density for genic selection and instantaneous changes in population size. First, we focus on a single time period during which the population has a different size relative to a fixed reference population size. We compute the eigenvalues and the eigenfunctions of the diffusion operator with respect to the modified drift term of the underlying diffusion equation. Similarly to a constant population size, the eigenfunctions are given as a series of orthogonal functions. The eigenvalues and eigenfunctions facilitate a spectral representation of the transition density describing the change in allele frequencies across this time period. Such transition densities for single time periods can then be folded over various instantaneous population size changes to obtain the overall transition density for such a multi-epoch model with genic selection. After illustrating the applicability of this approach, we derive the SFS by means of the transition density. While the transition density proves useful for the analysis of time-series data that are mostly gathered from species with short generation times as bacteria (e.g., Lenski 2011) but also from species with long generation times (Steinrücken et al. 2014), the SFS can also be applied to whole-genome data collected at a single time point. As an alternative approach to employing the transition density for the SFS, we modify the moment-based approach by Evans et al. (2007) to efficiently compute allele frequency spectra for genic selection, point mutations and piecewise changes in population size.

We then employ a maximum likelihood method to estimate the demographic and selective parameters of a given bottleneck model. After examining the accuracy of parameter estimation, we discuss how the estimates change when selection is ignored or a simpler demographic model is assumed. We investigate the demography of an African population of *Drosophila melanogaster* (Duchen et al. 2013), allowing for selection coefficients that are either constant or vary according to a given distribution of fitness effects. Furthermore, we answer an other, important question arising in human population genetics (Tennessen et al. 2012): Can the impact of negative selection be observed in populations that undergo strong exponential growth? We investigate, how strong selection would have to be to leave a signature in the SFS.

The transition density for genic selection and piecewise-constant population sizes with K epochs

Model and notation

We assume that the diploid effective population size changes deterministically, with $N(t)$ denoting the size at time t. Here, time is measured in units of $2N_{\text{ref}}$ generations, where N_{ref} is a fixed reference population size. Unless stated otherwise, the initial population size will be used as the reference population size in the various numerical examples. In the diffusion limit, the relative population size $N(t)/N_{\text{ref}}$ converges to a scaling function which we denote by $\rho(t)$.

We assume the infinitely-many-sites model (Kimura 1969) with A_0 and A_1 denoting the ancestral and derived allelic types, respectively. The relative fitnesses of A_1/A_1 and A_1/A_0 genotypes over the A_0/A_0 genotype are respectively given by $1+2s$ and $1+s$. The population-scaled selection coefficient is denoted by $\sigma = 2N_{\text{ref}} \cdot s$. The frequency of the derived allele A_1 at time t is denoted by X_t. Let f be a twice continuously differentiable, bounded function over $[0, 1]$. The backward generator of a time-inhomogeneous one-dimensional WF diffusion process on $[0, 1]$ is denoted by $\mathscr{L}$, which acts on f as

$$\mathscr{L} f(x) = \frac{1}{2} b(x;t) \frac{\partial^2}{\partial x^2}\{f(x)\} + a(x)\frac{\partial}{\partial x}\{f(x)\}, \tag{1}$$

where the diffusion and drift terms are given by $b(x;t) = x(1-x)/\rho(t)$ and $a(x) = \sigma x(1-x)$, respectively. While selection operates on a natural time scale as represented by the drift term, changes in population size require an appropriate rescaling of time within the diffusion term. Thus, the relative strength of natural selection and genetic drift is time-inhomogeneous. This prohibits classical time-rescaling approaches and introduces considerable challenges in obtaining analytic results. To gain insights, we here focus on the case where ρ is piecewise constant. In this case, the diffusion and drift terms differ by a constant factor within each piece, thus simplifying the analysis.

Throughout, we assume that ρ has K constant pieces (or epochs) in the time interval $[\tau_0, \tau)$. The change points are denoted by $t_1, \ldots, t_{K-1}$, and for convenience we define $t_0 = \tau_0$ and $t_K = \tau$. Then, for $t_i \leq t < t_{i+1}$, with $0 \leq i \leq K-1$, we assume $\rho(t) = c_i$, where c_i is some positive constant. For the epoch $t_i \leq t < t_{i+1}$, the diffusion term is thus given by $b_i(x) = x(1-x)/c_i$ and the corresponding generator is denoted by $\mathscr{L}^i$. The scale density ξ_i (Karlin and Taylor 1981, Ch. 15) for the epoch is given by

$$\xi_i(x) \;\;=\;\; \exp\left[-\int_0^x \frac{2a(z)}{b_i(z)}dz\right] = \exp(-2c_i\sigma x),$$

while the speed density π_i is given (up to a constant) by

$$\pi_i(x) \;\;=\;\; [b_i(x)\xi_i(x)]^{-1} = \frac{c_i \exp(2c_i\sigma x)}{x(1-x)}. \tag{2}$$

Given real-valued functions f and g on $[0,1]$ that satisfy appropriate boundary conditions and are square integrable with respect to some real positive density h, we use $\langle f,g\rangle_h$ to denote

$$\langle f,g\rangle_h = \int_0^1 f(x)g(x)h(x)dx.$$

The transition density within each epoch $[t_i, t_{i+1})$

For the epoch $[t_i, t_{i+1})$, let the transition density be denoted by $p_i(t;x,y)$, where $t \in [t_i, t_{i+1})$, $X_{t_i} = x$ and $X_t = y$. Under the initial condition $p_i(t_i;x,y) = \delta(x-y)$, the spectral representation of $p_i(t;x,y)$ is given by

$$p_i(t;x,y) = \sum_{n=0}^{\infty} \exp[-\Lambda_n^i(t-t_i)]\pi_i(y)\Phi_n^i(x)\Phi_n^i(y)\frac{1}{\langle \Phi_n^i, \Phi_n^i\rangle_{\pi_i}}, \tag{3}$$

where $-\Lambda_n^i$ and Φ_n^i are the eigenvalues and eigenfunctions of $\mathscr{L}^i$, respectively. That is,

$$\mathscr{L}^i\Phi_n^i(x) = -\Lambda_n^i\Phi_n^i(x).$$

It can be shown that the eigenvalues are all real and non-positive. Furthermore,

$$0 \leq \Lambda_0^i < \Lambda_1^i < \Lambda_2^i < \cdots,$$

with $\Lambda_n^i \to \infty$ as $n \to \infty$. The associated eigenfunctions $\{\Phi_n^i(x)\}_{n=0}^{\infty}$ form an orthogonal basis of $L^2([0,1],\pi_i)$, the space of real-valued functions on $[0,1]$ that are square integrable with respect to the speed density π_i, defined in (2).

Song and Steinrücken (2012) recently developed a method for finding Λ_n^i and Φ_n^i in the case of $c_i = 1$. We will give a brief description of their method and modify it accordingly to incorporate an arbitrary $c_i > 0$. Let $\mathscr{L}_0^i$ denote the diffusion generator under neutrality (i.e., $\sigma = 0$). The eigenfunctions of $\mathscr{L}_0^i$ are modified Gegenbauer polynomials $\{G_n(x)\}_{n=0}^{\infty}$ (cf. *Appendix*), and the corresponding eigenvalues are $-\lambda_n^i$, with

$$\lambda_n^i = \binom{n+2}{2} \frac{1}{c_i}. \tag{4}$$

Similar to Song and Steinrücken (2012), define $H_n^i(x)$ as

$$H_n^i(x) = \frac{\exp(-c_i \sigma x)}{\sqrt{c_i}} G_n(x). \tag{5}$$

Then, $\{H_n^i(x)\}_{n=0}^{\infty}$ form an orthogonal system with respect to the weight function $\pi_i(x)$. By directly applying the full generator $\mathscr{L}^i$ to $H_n^i(x)$, we observe that $H_n^i(x)$ are not eigenfunctions of $\mathscr{L}^i$. Instead, we obtain

$$\mathscr{L}_i H_n^i(x) \quad = -[\lambda_n^i + c_i \, Q(x;\sigma)] H_n^i(x), \tag{6}$$

where $Q(x;\sigma) = 1/2 \cdot \sigma^2 x(1-x)$. However, since both $\{H_n^i(x)\}_{n=0}^{\infty}$ and $\{\Phi_n^i(x)\}_{n=0}^{\infty}$ are orthogonal with respect to the same weight function $\pi_i(x)$, and $\{H_n^i(x)\}_{n=0}^{\infty}$ form a basis of $L^2([0,1], \pi_i)$, we can represent $\Phi_n^i(x)$ as a linear combination of $H_m^i(x)$:

$$\Phi_n^i(x) \quad = \sum_{m=0}^{\infty} u_{n,m}^i H_m^i(x). \tag{7}$$

Furthermore, the fact that $\Phi_n^i(x)$ is an eigenfunction of $\mathscr{L}^i$ with eigenvalue $-\Lambda_n^i$ implies that $\{u_{n,m}^i\}_{m=0}^{\infty}$ and Λ_n^i satisfy the following equation:

$$\begin{pmatrix} \lambda_0^i + c_i a_0^{(0)} & 0 & c_i a_2^{(-2)} & 0 & 0 & \cdots \\ 0 & \lambda_1^i + c_i a_1^{(0)} & 0 & c_i a_3^{(-2)} & 0 & \cdots \\ c_i a_0^{(+2)} & 0 & \lambda_2^i + c_i a_2^{(0)} & 0 & c_i a_4^{(-2)} & \cdots \\ 0 & c_i a_1^{(+2)} & 0 & \lambda_3^i + c_i a_3^{(0)} & 0 & \cdots \\ 0 & 0 & c_i a_2^{(+2)} & 0 & \lambda_4^i + c_i a_4^{(0)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} u_{n,0}^i \\ u_{n,1}^i \\ u_{n,2}^i \\ u_{n,3}^i \\ u_{n,4}^i \\ \vdots \end{pmatrix} = \Lambda_n^i \begin{pmatrix} u_{n,0}^i \\ u_{n,1}^i \\ u_{n,2}^i \\ u_{n,3}^i \\ u_{n,4}^i \\ \vdots \end{pmatrix}, \tag{8}$$

where λ_n^i is as defined in (4) and $a_m^{(-2)}, a_m^{(0)}, a_m^{(+2)}$ are known constants that depend on σ and m

(cf. Song and Steinrücken 2012 for details).

The transition density expansion (3) can be obtained by numerically solving the eigensystem (8). Denote the infinite-dimensional matrix on the left hand side of (8) by W_i. The eigenvalues Λ_n^i of W_i correspond (up to a sign) to the eigenvalues of $\mathscr{L}^i$, and the associated eigenvectors $\boldsymbol{u}_n^i = \left(u_{n,0}^i, u_{n,1}^i, u_{n,2}^i, \ldots\right)^T$ of W_i determine the eigenfunctions of $\mathscr{L}^i$ via (7). Let $W_i^{[D]}$ denote the $D \times D$ matrix obtained by taking the first D rows and D columns of W_i, and let $\Lambda_n^{i,[D]}$ and $\boldsymbol{u}_n^{i,[D]} = \left(u_{n,0}^{i,[D]}, u_{n,1}^{i,[D]}, u_{n,2}^{i,[D]}, \ldots\right)^T$ denote the eigenvalues and eigenvectors of $W_i^{[D]}$, respectively. The truncated eigensystem

$$W_i^{[D]} \boldsymbol{u}_n^{i,[D]} = \Lambda_n^{i,[D]} \boldsymbol{u}_n^{i,[D]} \tag{9}$$

can then be used to approximate (8). This finite-dimensional linear system can be easily solved numerically. Since the truncated versions of the eigenvalues and eigenvectors converge rapidly as D increases, an accurate approximation of the transition density (3) can be efficiently obtained. The truncation level D required for convergence is higher when modeling a large population compared to the basic selection model, and lower when the population size is small. The reason for this is that the necessary truncation level depends on the effective strength of selection, which is higher in large populations and lower in small populations. Therefore, for a fixed selection coefficient s, large populations are computationally more demanding than small populations. Furthermore, we observed that positive selection coefficients require higher values for D than negative ones.

The transition density for the entire period $[\tau_0, \tau)$ with K epochs

Suppose $X_{\tau_0} = x$ and $X_\tau = y$. The transition density $p(\tau_0, \tau; x, y)$ for the entire period $[\tau_0, \tau)$ is obtained by combining the transition densities for the K epochs as follows:

$$p(\tau_0, \tau; x, y) \;\; = \int\limits_{[0,1]^{K-1}} p_0(t_1; x, x_1) \left[\prod_{i=1}^{K-2} p_i(t_{i+1}; x_i, x_{i+1})\right] p_{K-1}(\tau; x_{K-1}, y)\, dx_1 \ldots dx_{K-1}, \tag{10}$$

where x_i denotes the allele frequency at the change point t_i. Using (3), we can write (10) as

$$p(\tau_0, \tau; x, y) \;\; = \;\; \boldsymbol{\Phi}_0(x)^T \boldsymbol{E}_0 \boldsymbol{S}_0 \boldsymbol{E}_1 \boldsymbol{S}_1 \cdots \boldsymbol{E}_{K-2} \boldsymbol{S}_{K-2} \boldsymbol{E}_{K-1} \boldsymbol{\Phi}_{K-1}(y) \pi_{K-1}(y), \tag{11}$$

where $\boldsymbol{\Phi}_i(x) = \big(\Phi_0^i(x), \Phi_1^i(x), \Phi_2^i(x), \dots\big)^T$ is an infinite-dimensional column vector, while $\boldsymbol{E}_i$ and $\boldsymbol{S}_i$ are infinite-dimensional matrices defined as

$$\boldsymbol{E}_i = \operatorname{diag}\left(\frac{e^{-\Lambda_0^i(t_{i+1}-t_i)}}{\langle \Phi_0^i, \Phi_0^i\rangle_{\pi_i}}, \frac{e^{-\Lambda_1^i(t_{i+1}-t_i)}}{\langle \Phi_1^i, \Phi_1^i\rangle_{\pi_i}}, \dots\right)$$

and

$$\boldsymbol{S}_i = \int_0^1 \pi_i(z)\boldsymbol{\Phi}_i(z)\boldsymbol{\Phi}_{i+1}(z)^T dz.$$

In general, $\boldsymbol{S}_i$ is not a diagonal matrix since $\Phi_n^i(z)$ and $\Phi_m^{i+1}(z)$ are not orthogonal with respect to $\pi_i(z)$ if $c_i \neq c_{i+1}$. In *Appendix*, we show that the entry (n, m) of $\boldsymbol{S}_i$ is given by

$$\begin{aligned}\int_0^1 \pi_i(z)\Phi_n^i(z)\Phi_m^{i+1}(z)dz \quad &= \quad \sqrt{\frac{c_i}{c_{i+1}}}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} u_{n,k}^i u_{m,l}^{i+1} \sum_{j=1}^{k+l+2} (-1)^{j+1} \frac{e^{\sigma(c_i - c_{i+1})} - (-1)^{k+l+j}}{[\sigma(c_i - c_{i+1})]^{j+1}} \\ &\times \frac{(k+1)(l+1)j!}{(k+2)(l+2)} \sum_{r=0}^{j-1} \binom{k+2}{j-r}\binom{k+j-r}{j-r-1}\binom{l+r+2}{r+1}\binom{l}{r}. \qquad (12)\end{aligned}$$

Note that the last line of (12) does not depend on n or m, so it needs to be computed only once. The overall computational time for evaluating $p(\tau_0, \tau; x, y)$ scales linearly with the number K of epochs.

To better understand the joint impact of selection and demography on the transition density, we consider two scenarios, where $p(0, \tau; x, y)$ is simply denoted as $p(\tau; x, y)$. Figure 1 illustrates the density in a scenario in which the selection coefficient is fixed and various K-epoch demographic models are considered. In comparison to the case of a constant population size (cf. Figure 1a), an instantaneous expansion (cf. Figure 1b) narrows the distribution around the mean, whereas an additional phase of a reduced population size (cf. Figure 1c) increases the variance relative to a population of a constant size. Figure 2 illustrates the same scenarios with a fixed transition time and varying selection coefficients. Note that all theoretical results and the corresponding applications in this paper were implemented in *Mathematica*. The implementation is available from the authors upon request.

The sample frequency spectrum

The transition density approach

The transition density derived in the previous section can be employed to obtain the sample frequency spectrum (SFS) of a sample. Consider a sample of size n obtained at time $t = \tau$. The probability that the A_1 allele with frequency x at time $t = \tau_0$ is observed b times in the sample is (Griffiths 2003)

$$p_{n,b}(x;\tau_0,\tau) \quad = \quad \int\limits_0^1 \binom{n}{b} y^b(1-y)^{n-b} p(\tau_0,\tau;x,y)dy. \tag{13}$$

For piecewise-constant population size models with K epochs, a spectral representation of $p(\tau_0,\tau;x,y)$ can be found via (11) and evaluating (13) involves computing the integral $\int_0^1 y^b(1-y)^{n-b}\pi_{K-1}(y)\mathbf{\Phi}_{K-1}(y)dy$. For $l \geq 0$, using (2), (5), and (7), we obtain

$$\begin{aligned} &\int_0^1 y^b(1-y)^{n-b}\pi_{K-1}(y)\Phi_l^{K-1}(y)dy \\ &= \sum_{m=0}^{\infty} \sqrt{c_{K-1}} u_{l,m}^{K-1} \int_0^1 y^{b-1}(1-y)^{n-b-1} e^{c_{K-1}\cdot\sigma y} G_m(y)dy \\ &= \sum_{m=0}^{\infty} \sqrt{c_{K-1}} u_{l,m}^{K-1} \frac{1}{b+1} \sum_{h=0}^{m} (-1)^{h+1} \frac{\binom{m+1}{h+1}\binom{h+m+2}{h}}{\binom{n+h+1}{b+1}} \cdot {}_1F_1(b+1;n+h+2;c_{K-1}\cdot\sigma), \end{aligned} \tag{14}$$

where ${}_1F_1(a;b;z) = \sum\limits_{j\geq 0} a_{(j)}/b_{(j)} z^j/j!$ is the confluent hypergeometric function of the first kind. The descending factorials $d_{(j)}$ are defined in *Appendix*.

The SFS $q_{n,b}(\tau)$ is the probability distribution on the number b of mutant alleles in a sample of size n taken at time τ, conditioned on segregation. For $1 \leq b \leq n-1$, $q_{n,b}(\tau)$ is given by

$$q_{n,b}(\tau) \quad = \quad \lim_{x\to 0} \frac{\int_{-\infty}^{\tau} p_{n,b}(x;\tau_0,\tau)d\tau_0}{\int_{-\infty}^{\tau} \sum_{a=1}^{n-1} p_{n,a}(x;\tau_0,\tau)d\tau_0}. \tag{15}$$

In (15), the SFS at a single site is obtained by averaging over sample paths. This is equivalent to the frequency spectrum distribution over a large number of independent mutant sites in the Poisson random field model of Sawyer and Hartl (1992). Using (11), (12), (13) , and (14), we can approximate (15) numerically. If it is unknown which allele is derived, a folded version of (15) can be obtained as $[q_{n,b} + q_{n,n-b}]/(1+\delta_{b,n-b})$, where $\delta_{b,n-b}$ denotes the Kronecker delta.

A moment-based approach

As detailed above, the transition density can be employed to obtain the SFS. However, the specific solution for the transition density is not required to obtain the less complex and thus computationally less demanding SFS. Here, we utilize the work of Evans et al. (2007) to develop an efficient algorithm for computing the allele frequency spectrum in the case of genic selection and piecewise-constant population sizes.

Suppose mutations arise at rate $\theta/2$ (per sequence per $2N_{\text{ref}}$ generations) and according to the infinitely-many-sites model (Kimura 1969). Evans et al. (2007) use the forward diffusion equation to describe population allele frequency changes and introduce mutations by an appropriate boundary condition. Slightly modifying their notation, we use $f(y,t)dy$ to denote the expected number of sites where the mutant allele has a frequency in $(y, y+dy)$, with $0 < y < 1$, at time t. The forward equation is

$$\frac{\partial}{\partial t} f(y,t) \quad = \quad \frac{1}{2}\frac{\partial^2}{\partial y^2}\{b(y;t)f(y,t)\} - \frac{\partial}{\partial y}\{a(y)f(y,t)\}, \tag{16}$$

where the diffusion term $b(y;t) = y(1-y)/\rho(t)$, the drift term $a(y) = \sigma y(1-y)$, the scaled selection coefficient σ, and the population size function $\rho(t)$ are defined as before. The influx of mutations is incorporated into this process via the boundary conditions

$$\lim_{y \downarrow 0} y f(y,t) \quad = \quad \theta\rho(t) \quad \text{and} \quad \lim_{y\uparrow 1} f(y,t) \text{ finite.} \tag{17}$$

The resulting polymorphic sites follow the dynamics of (16) thereafter. Note that this differs from the diffusion process studied in the previous section, as the influx of mutations is now explicitly modeled.

Again, it is analytically more practical to consider the corresponding backward equation, which is obtained by setting $g(y,t) := y(1-y)f(y,t)$. This substitution transforms the forward equation for $f(y,t)$ into a backward equation for $g(y,t)$, which is essentially given by (1) up to the sign of the drift term. Evans et al. (2007) derived a coupled system of ordinary differential equations (ODEs) for the moments $\mu_j(t) = \int_0^\infty y^j g(y,t)dy$:

$$\mu_0'(t) \quad = \quad \frac{\theta}{2} - \frac{1}{\rho(t)}\mu_0(t) + \sigma[\mu_0(t) - 2\mu_1(t)], \tag{18}$$

$$\mu_j'(t) = \frac{1}{\rho(t)}\left[\binom{j+1}{2}\mu_{j-1}(t) - \binom{j+2}{2}\mu_j(t)\right] + \sigma\Big[(j+1)\mu_j(t) - (j+2)\mu_{j+1}(t)\Big], \quad j \geq 1, \tag{19}$$

where $\mu_j'(t) = d\mu_j(t)/dt$. A similar system of ODEs was derived and solved by Kimura (1955a) for a neutral scenario with a constant population size and without mutations. For $\sigma = 0$, the above system is finite and can be solved explicitly (Živković and Stephan 2011). In the case of selection ($\sigma \neq 0$), on the other hand, the system is infinite and obtaining an explicit solution for an arbitrary ρ is a challenging problem, even if the system is truncated by setting $\mu_j(t) = 0$ for $j \geq D$.

From now on, assume $\mu_j(t) \equiv 0$ for $j \geq D$ and rewrite the truncated system of ODEs in matrix form as

$$\boldsymbol{M}'(t) = \left[\frac{1}{\rho(t)}\boldsymbol{B} + \sigma\boldsymbol{A}\right]\boldsymbol{M}(t) + \boldsymbol{\Theta}, \tag{20}$$

where $\boldsymbol{M}(t) = \left(\mu_0^{[D]}(t), \mu_1^{[D]}(t), \ldots, \mu_{D-1}^{[D]}(t)\right)^T$, $\boldsymbol{M}'(t) = d\boldsymbol{M}(t)/dt$, $\boldsymbol{\Theta} = (\theta/2, 0, \ldots, 0)^T$ are D-dimensional column vectors, and $\boldsymbol{B} = (b_{kl})$ and $\boldsymbol{A} = (a_{kl})$ are $D \times D$ matrices with entries

$$b_{kl} = \begin{cases} -\binom{k+2}{2}, & \text{if } l = k, \\ \binom{k+1}{2}, & \text{if } l = k-1, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{and} \qquad a_{kl} = \begin{cases} k+1, & \text{if } l = k, \\ -(k+2), & \text{if } l = k+1, \\ 0, & \text{otherwise,} \end{cases}$$

for $0 \leq k, l \leq D-1$. The formal solution of (20) cannot be written in terms of a matrix exponential but only as a Peano-Baker series (Baake and Schlägel 2011) for arbitrary ρ, which can be numerically quite demanding. Therefore, we focus on the case of piecewise constant ρ and develop an efficient method to solve the truncated system of ODEs.

We first consider $\rho(t) \equiv c_0$ (i.e., a constant population size), for which the solution of (20) takes the form of a matrix exponential given by

$$\begin{aligned} \boldsymbol{M}(t) &= \exp\left[\int_0^t \left(\frac{\boldsymbol{B}}{c_0} + \sigma\boldsymbol{A}\right) ds\right]\boldsymbol{M}(0) + \left\{\int_0^t \exp\left[\int_s^t \left(\frac{\boldsymbol{B}}{c_0} + \sigma\boldsymbol{A}\right) du\right] ds\right\}\boldsymbol{\Theta} \\ &= \exp\left[\left(\frac{\boldsymbol{B}}{c_0} + \sigma\boldsymbol{A}\right)t\right]\boldsymbol{M}(0) + \left\{\exp\left[\left(\frac{\boldsymbol{B}}{c_0} + \sigma\boldsymbol{A}\right)t\right] - \boldsymbol{I}\right\}\left(\frac{\boldsymbol{B}}{c_0} + \sigma\boldsymbol{A}\right)^{-1}\boldsymbol{\Theta}. \end{aligned} \tag{21}$$

Let $-\lambda_k$, $(l_{k,0}, \ldots, l_{k,D-1})$, and $(r_{0,k}, \ldots, r_{D-1,k})^T$ respectively denote the eigenvalues, row eigenvectors, and column eigenvectors of $\boldsymbol{B}/c_0 + \sigma\boldsymbol{A}$. Then, (21) implies

$$\mu_j^{[D]}(t) \quad = \quad \sum_{i=0}^{D-1} \mu_i^{[D]}(0) \sum_{k=0}^{D-1} r_{jk} l_{ki} e^{-\lambda_k t} + \frac{\theta}{2} \sum_{k=0}^{D-1} r_{jk} l_{k0} \frac{1 - e^{-\lambda_k t}}{\lambda_k}. \tag{22}$$

It is intractable to find closed-form expressions of $-\lambda_k$, l_{ki}, and r_{jk}, but, for a given truncation level D, they can be computed numerically. Depending on the details of the model under consideration, it might be more efficient to solve (21) numerically rather than applying the more analytic form given in (22).

We now investigate the equilibrium solution of (22), since it can be applied as an initial condition in a model in which the population size remains constant over a longer period of time before instantaneous population size changes occur. Assuming that all alleles are monomorphic at time zero, i.e. $\mu_i^{[D]}(0) \equiv 0$, and letting $t \to \infty$, we obtain the moments at equilibrium as

$$\hat{\mu}_j^{[D]} \quad = \quad \frac{\theta}{2} \sum_{k=0}^{D-1} \frac{r_{jk} l_{k0}}{\lambda_k}.$$

For D sufficiently large, this result is numerically close to the exact solution $\hat{\mu}_j$. The latter can also be obtained as follows. The equilibrium population frequency spectrum is given by (Fisher 1930)

$$\hat{f}(y) = \frac{\theta c_0 \left[1 - e^{-2c_0\sigma(1-y)}\right]}{y(1-y)(1 - e^{-2c_0\sigma})}. \tag{23}$$

The sampled version can be easily found via binomial sampling as in (13):

$$\hat{f}_{n,b} \quad = \quad \theta c_0 \frac{n}{b(n-b)} \frac{1 - {}_1F_1(b; n; 2c_0\sigma) e^{-2c_0\sigma}}{1 - e^{-2c_0\sigma}}. \tag{24}$$

For $\sigma \neq 0$, the moments $\hat{\mu}_j$ of $\hat{g}(y) = y(1-y)\hat{f}(y)$ are given by

$$\hat{\mu}_j \quad = \quad \theta c_0 \frac{1}{1 - e^{-2c_0\sigma}} \left\{ \frac{e^{-2c_0\sigma}[\Gamma(j+1, -2c_0\sigma) - j!]}{(-2c_0\sigma)^{j+1}} + \frac{1}{j+1} \right\},$$

where $\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t} dt$ is the incomplete gamma function.

Now, consider the piecewise-constant model with K epochs in the time interval $[\tau_0, \tau]$ defined earlier. For $t_i \leq t < t_{i+1}$,

$$\boldsymbol{M}'(t) \quad = \quad \left(\frac{\boldsymbol{B}}{c_i} + \sigma \boldsymbol{A}\right) \boldsymbol{M}(t) + \boldsymbol{\Theta}, \tag{25}$$

which can be solved as in (21). For $\tau > t_{K-1}$,

$$\begin{aligned} \boldsymbol{M}(\tau) \quad = \quad & \exp\left[\Big(\frac{\boldsymbol{B}}{c_{K-1}} + \sigma \boldsymbol{A}\Big)(\tau - t_{K-1})\right] \boldsymbol{M}(t_{K-1}) + \\ & \left\{\exp\left[\Big(\frac{\boldsymbol{B}}{c_{K-1}} + \sigma \boldsymbol{A}\Big)(\tau - t_{K-1})\right] - \boldsymbol{I}\right\} \Big(\frac{\boldsymbol{B}}{c_{K-1}} + \sigma \boldsymbol{A}\Big)^{-1} \boldsymbol{\Theta}, \end{aligned} \tag{26}$$

where $M(t_i)$, for $1 \leq i \leq K-1$, is recursively given by

$$\begin{aligned} \boldsymbol{M}(t_i) \quad = \quad & \exp\left[\left(\frac{\boldsymbol{B}}{c_{i-1}} + \sigma \boldsymbol{A}\right)(t_i - t_{i-1})\right] \boldsymbol{M}(t_{i-1}) + \\ & \left\{\exp\left[\left(\frac{\boldsymbol{B}}{c_{i-1}} + \sigma \boldsymbol{A}\right)(t_i - t_{i-1})\right] - \boldsymbol{I}\right\} \left(\frac{\boldsymbol{B}}{c_{i-1}} + \sigma \boldsymbol{A}\right)^{-1} \boldsymbol{\Theta}. \end{aligned}$$

The initial condition $\boldsymbol{M}(t_0)$ is either chosen as the equilibrium solution described above or the zero vector, which corresponds to the case of all loci being monomorphic at time $t_0 = \tau_0$.

The accuracy of the above framework depends on how fast the truncated moments $\mu_j^{[D]}(\tau)$ converge to zero as D increases. Similar to the transition density approach, the truncated moments converge faster for negative than for positive σ, and for instantaneous declines compared to instantaneous expansions. For a large positive σ, a higher truncation level D may be required to achieve the desired accuracy. Finally, the allelic spectrum $f_{n,b}(\tau)$, for $1 \leq b \leq n-1$, of a sample of size n taken at time τ can be obtained from the moments $\mu_j(\tau)$ by using the relationship

$$f_{n,b}(\tau) \quad = \quad \binom{n}{b} \sum_{l=0}^{n-b-1} (-1)^l \binom{n-b-1}{l} \mu_{l+b-1}(\tau). \tag{27}$$

The SFS $q_{n,b}(\tau)$ at time τ is then given by

$$q_{n,b}(\tau) \quad = \quad \frac{f_{n,b}(\tau)}{\sum_{a=1}^{n-1} f_{n,a}(\tau)}. \tag{28}$$

Substituting the truncated moments obtained from (26) into (27) provides numerical approximations of (27) and (28).

The joint impact of a population bottleneck and selection on the SFS is illustrated in Figure 3 for various points in time. As expected, negative and positive selection result in a skew of the SFS towards low- and high-frequency derived variants, respectively, when compared to a model without selection, across all sampling times. Moreover, this skew varies in intensity at different points in time. In the neutral demographic model (cf. Figure 3b), the relative frequency of singletons at time τ_3 is higher than at time τ_4, whereas under the same demographic model with negative selection (cf. Figure 3c) this relation is inverted. This is because the amount of singletons that is caused by demographic forces decreases after the expansion from τ_3 to τ_4, while negative selection is still increasing the low-frequency derived classes in this time interval.

Applications

Here, we discuss biologically relevant questions that can be addressed using our theoretical framework. This section consists of the following parts:

1. We first consider models with negative selection and bottlenecks of medium strength at different time points. We examine the SFS under such models and try to estimate the demographic parameters while taking selection into account. We also carry out demographic inference ignoring selection. Whereas the former demonstrates how well the demographic and selective parameters can be estimated jointly, the latter mimics the common practice of assuming genome-wide polymorphic sites as putatively neutral (due to the difficulty of jointly estimating the impact of selection and demography using existing tools). We finally examine the consequences of assuming a too simple underlying demography on parameter estimation.
2. We then analyze an African sample of *Drosophila melanogaster* to investigate its demographic history and possible selective effects.
3. Lastly, we examine a model of strong exponential population growth (mimicking human evolution) and superimpose negative selection of various strengths to understand if and when selection can be inferred for such a model.

Throughout, the first population size change will occur after the allele frequencies have reached an equilibrium according to (24).

Joint inference of population bottleneck and purifying selection

A maximum likelihood approach

Under the assumption that the considered sites are independent, the log-likelihood of a model $\mathcal{M}$ given data $\mathcal{D}$ is $\log[L(\mathcal{D};\mathcal{M})] = \sum_{i=1}^{n-1} d_i \log(q_i) +$ constant, where d_i is the observed number of sites at which the derived allele occurs i times in the sample, and q_i is the probability that the derived allele occurs i times in the sample at a segregating site under model $\mathcal{M}$ (e.g., Wooding and Rogers 2002). Recall that q_i can be either obtained via the transition density or the moment-based approach. The latter is preferable here, since the transition density is not explicitly required.

Consider the bottleneck model illustrated in Figure 4. Note that the present relative size c_S is fixed to 1, i.e., here the present population size is used as the reference population size N_{ref}. First, we consider the scenario where the ancestral population size c_0 prior to the bottleneck is allowed to vary. In this case, the model has five free parameters: c_0, the initial population size; c_B, the population size during the bottleneck; t_B, the duration of the bottleneck; $t_S = \tau - t_B$, the time since recovery from the bottleneck; and σ, the scaled selection coefficient. We then also consider the scenario where the ancestral population size is the same as the present population size, i.e., $c_0 = c_S$, resulting in a model with four free parameters.

We adopted a grid search in our estimation procedure, with $\sigma \in [-10, 0]$ and $c_B, t_B, t_S \in [0.001, 1]$. For the 5-parameter model, c_0 was chosen from the range $[0.01, 10]$. In total, 110,000 grid points were chosen in the selected case and 10,000 in the neutral case. Note that the grid search also accounts for models of one or two successive instantaneous population expansions. For the 4-parameter model, 11,000 grid points were chosen in the selected case and 1000 in the neutral case. The grid points are summarized in Table 1.

Estimation of bottleneck and selection parameters

We first evaluated the SFS for a sample of size $n = 50$ in the following twelve scenarios, all with $c_S = 1$ and $\sigma \in \{0, -1/2, -2\}$:

1. Constant population size (i.e., $c_0 = c_B = c_S = 1$).
2. Bottleneck models with $c_0 = 1/2, c_B = 1/10$, $t_B = 1/10$, and $t_S \in \{1/200, 1/20, 1/2\}$.

First, to test how well the demographic and selective parameters can be estimated jointly from sampled data, we focused on the bottleneck demography with $t_S = 1/20$ and considered two scenarios: The neutral case ($\sigma = 0$) and the selected case with $\sigma = -2$. To mimic the limited availability of independent polymorphic sites across the genome, we sampled 10,000 sites according to the SFS for the two chosen scenarios, and repeated this procedure 200 times. For each of these 200 datasets, we maximized the log-likelihood over the grid of parameter values described earlier, assuming (A1) neutrality when the true model has $\sigma = 0$, (A2) neutrality when the true model has $\sigma = -2$, (A3) presence of selection when the true model has $\sigma = -2$, and (A4) presence of selection when the true model has $\sigma = 0$.

The estimated parameters are shown in Table 2. For inference under correct model assumptions (A1 and A3), the median estimates are equal to the true parameters. When selection is ignored although present in the dataset (A2), the ancestral population size (c_0) and the duration of the bottleneck (t_B) are underestimated, whereas the bottleneck size (c_B) and the time since the bottleneck (t_S) are accurately estimated. When the true model is neutral but the inference procedure allows for selection (A4), a neutral demographic model is accurately inferred. We calculated likelihood-ratio statistics for each of the 200 datasets to compare the two nested models of selection and neutrality. The null hypothesis of neutrality can be rejected at the 5% significance level with a power of 55%.

We further analyzed all twelve scenarios using the expected SFS directly, assuming that the amount of data is sufficiently large such that the observed SFS closely approximates the expected value. Our goal in this case is to study the effect of model misspecification on parameter estimation; specifically, assuming selection when the true model is neutral or assuming neutrality when there is selection. In the former case, the maximum likelihood estimates (MLEs) always coincided with the true parameters. Therefore, it is useful to allow for selection in an analysis even when putatively neutral regions are considered. In the latter case, our results are summarized in Table 3. For a constant population size, two rather old instantaneous expansions are estimated. For the bottleneck models, ignoring selection leads to the largest errors for the most recent bottleneck and $\sigma = -1/2$ and the least recent bottleneck and $\sigma = -2$, for which an instantaneous expansion is estimated. The time since the bottleneck was robustly estimated in many cases.

To assess the impact of assuming a slightly simplified model for parameter estimation, we carried out an analogous study where the ancestral population size c_0 was incorrectly assumed to equal the current size $c_S = 1$, while the true model had $c_0 = 1/2$ and $c_S = 1$. For the resampling analysis, we considered the same bottleneck scenarios as before with $\sigma = 0$ or -2, and maximized the log-likelihood values over a grid in the parameter space (as described earlier) for each of the 200 simulated datasets each containing 10,000 polymorphic sites. The parameter estimates are shown in Table 4. The time since the bottleneck (t_S) is accurately estimated irrespective of correct or wrong assumptions regarding selection. Incorrectly assuming $c_0 = c_S$ results in either an overestimation of the duration of the bottleneck (t_B) in most of the cases (A1–A3) or an inference of selection when $\sigma = 0$ (A4). Selection was poorly estimated even under (A3).

Again, we also analyzed all twelve scenarios under the assumption that the observed SFS is a close approximation to the expected value, to study the effect of model misspecification on parameter estimation. The results are shown in Table 5. The biases caused by incorrectly assuming $c_0 = c_S$ are largest for the scenario that captures the youngest bottleneck ($t_S = 1/200$). Here, not only the selection coefficients are strongly misestimated but also the time since the bottleneck (t_S) is largely underestimated. In all the other scenarios, at least the time since the bottleneck (t_S) is accurately estimated. The estimation accuracy of the other demographic parameters and selection coefficients increases with bottleneck age and the concomitant decreasing impact of the ancestral population size on the SFS. In summary, we note that assuming a too simplistic demographic model can lead to large errors in parameter estimation.

Testing a dataset of Drosophila melanogaster

Here, we apply our method to analyze a dataset which has been recently used to estimate the joint demographic history of several populations of *Drosophila melanogaster* (Duchen et al. 2013). The dataset consists of 12 sequences from a Zimbabwe population comprising 197 non-coding loci; and within each locus there are between 1 and 41 segregating sites (3234 polymorphic sites in total). We focused on the effects of weak selection and used all segregating sites in our analysis, treating them as independent. We note that whereas the 197 loci are scattered over the genome, at least tens of thousands of bases apart, the sites within each locus are tightly linked and hence not independent. We have tried a bootstrap resampling procedure to study the effect of assuming independence, but the strong stochasticity among the small subsets of presumably independent sites, which were generated by sampling one site from each locus, prevented a reliable inference.

The empirical SFS of the data shows an uptick of high-frequency derived alleles (cf. Figure 5a). As explained in *Discussion*, this is likely to be caused by ancestral misidentification, not by positive selection. This effect is also unlikely to be caused by linkage, since the uptick is still observed in the previously mentioned subsamples of widely separated sites. To assess the effect of presumably misoriented sites on inference, we compare results for the unfolded SFS with those obtained from a partly folded version, where only singletons and doubletons are folded with their high-frequency counterparts, since these classes appear to be affected the most (cf. Baudry and Depaulis 2003).

We carried out our analysis based on the bottleneck model of the previous section allowing the current and the ancestral population size to differ. To account for varying selection pressures across the genome, sites are usually subdivided into various genomic categories (e.g., exons, introns, UTRs), often assuming a constant selection coefficient for each category. Alternatively, or even combined with such a categorization, selection coefficients are assumed to follow some distribution; a gamma distribution (Kimura 1979) is a popular choice due to its flexibility to fit empirical data. Since neutrality and purifying selection are considered to be prevalent in intronic and intergenic regions of African *Drosophila*, we focused on negative selection coefficients in our analysis. A non-coding dataset can be classified as a single functional category. Therefore, we analyzed the dataset first by either assuming constant selection or neutrality, followed by an analysis where the selection coefficients were allowed to vary according to a given distribution.

We initially computed an MLE for the unfolded and the partly folded SFS under the constant selection and the neutral bottleneck model on the coarse parameter grid given in Table 1. For each model, we investigated the accuracy of the parameter estimates via parametric bootstrap, using 200 bootstrap samples each consisting of 3234 polymorphic sites. We obtained rather narrow confidence intervals for the selection coefficient and the time since the bottleneck, whereas the other details of the bottleneck were less confidently estimated. To improve the parameter estimates, we further refined the grid as follows: Nine values for c_0 were chosen from the range $[0.5, 10]$, 20 values for σ from $[-2, 0]$, 10 values for c_B from $[0.001, 0.1]$, 25 values for t_B/c_B from $[0.84, 3.31]$, and 25 values for t_S from $[0.05, 0.22]$. This gives in total 1,125,000 parameter combinations for selection and 56,250 for neutrality. As before, the ratio of two consecutive values in each parameter range was kept roughly constant. Focusing on rescaled time t_B/c_B instead of t_B relies on the observation that t_B and c_B correlate strongly and has the advantage that unlikely combinations of t_B and c_B can be omitted. More values were chosen for time parameters, since these are more sensitive than the population size parameters.

The MLEs are given in Table 6 and both versions of the SFS are illustrated in Figure 5. The analysis based on the partly folded SFS shows a better fit than the unfolded version, since negative selection combined with any demographic model is incompatible with the uptick of high-frequency derived variants in the empirical SFS. Interestingly, a neutral model was inferred for the unfolded SFS, while the model with selection fits better for the partly folded version. Since an excess of high-frequency derived variants favors demographic models that capture a strong population decline, a much smaller estimate of the bottleneck population size (c_B) was obtained for the unfolded SFS.

In accordance with the previous section, the time since the bottleneck (t_S) was robustly estimated in both cases, as illustrated by the 10 and 100 most likely parameter estimates. However, partially folding the SFS led to a smaller estimate $\hat{t}_S$. A further refinement of the grid barely changed the estimates $\hat{t}_S$ and $\hat{c}_B$. The estimates of bottleneck duration (t_B) and ancestral population size (c_0) appeared to be strongly correlated.

We now relax the assumption of a fixed σ for all sites, and allow a distribution of fitness effects by introducing gamma distributed selection coefficients. For $\sigma > 0$, the probability density of the gamma distribution with shape and rate parameters α and β is given by $\gamma(\sigma) = \beta(\beta\sigma)^{\alpha-1}e^{-\beta\sigma}/\Gamma(\alpha)$, where $\Gamma(\cdot)$ denotes the gamma function. The allelic spectrum for gamma distributed selection coefficients is then obtained by integrating the allelic spectrum for constant selection coefficients given by (27) against a gamma distribution, i.e.,

$$\tilde{f}_{n,b}(\tau) \quad = \quad \int_{-\infty}^{0} f_{n,b}(\tau,\sigma)\gamma(-\sigma)d\sigma. \tag{29}$$

The SFS for gamma distributed selection coefficients is then given by $\tilde{q}_{n,b}(\tau) = \frac{\tilde{f}_{n,b}(\tau)}{\sum_{a=1}^{n-1}\tilde{f}_{n,a}(\tau)}$.

Even when the allelic spectrum is in equilibrium and the population size is constant, the integral in (29) cannot be solved explicitly, so we needed to employ numerical integration. Previous studies (e.g., Boyko et al. 2008, Racimo and Schraiber 2014) on the distribution of fitness effects in the presence of population size changes first inferred a demographic history using putatively neutral sites, and then estimated the parameters α and β based on that fixed demography. Since we do not have a separately inferred demographic model here, we considered several σ values along a variety of demographic parameter combinations. We used a coarser grid for the demographic parameters due to the larger number of σ values needed for the numerical integration step, which adds additional computational burden. While the evaluation of the allelic spectrum takes less than half a second for a given σ value with high numerical precision, the numerical integration over the range of σ values according to (29) takes a few seconds. Thus, to further reduce computational cost, we restricted the analysis to exponentially distributed selection coefficients by setting $\alpha = 1$ and compared the MLEs for various values of β. See Table 7 for results. The MLE was found for $\beta = 1$, so the average σ equals $-\alpha/\beta = -1$. This finding and the associated demographic estimates are consistent with the result found for a fixed selection coefficient. However, this result may change if one allows for more general shape and rate parameters.

A model of human exponential population growth

We now demonstrate the utility of our method to investigate population size histories containing epochs of exponential growth in combination with selection. To this end, we adopted the following demographic history of a sample of African human exomes that had been estimated by Tennessen et al. (2012) as a modification of a model by Gravel et al. (2011). The population had an ancestral size of 7310 individuals until 5920 generations ago (assuming a generation time of 25 years), when it increased instantaneously in size to 14,474 individuals. After this increase, the population remained constant in size until 205 generations ago, when it started to grow exponentially until reaching 424,000 individuals at present. The relative population size function for this model can be described by

$$\rho(t) = \begin{cases} 1, & t < 0, \\ c, & 0 \leq t < t_e, \\ c\exp[R(t-t_e)], & t_e \leq t \leq \tau, \end{cases} \tag{30}$$

where c is the ratio of population sizes after and before the instantaneous expansion, which can be dated arbitrarily, so we set the time of this expansion to zero. R is the scaled exponential growth rate, t_e is the time at which the expansion started, and τ is the time of sampling (the present). Times are given in units of $2N_{\text{ref}}$, where the reference population size N_{ref} is the initial size before time zero (the ancestral size). Since the theoretical framework presented above assumes a history of piecewise constant population sizes, the phase of exponential growth in this model had to be adequately discretized to obtain a suitable piecewise approximation. The following piecewise function can be chosen to approximate the exponential growth phase via a geometric growth function:

$$q(t) = \begin{cases} 1, & t < 0, \\ c, & 0 \leq t < t_1, \\ c(1+\delta)^i, & t_i \leq t < t_{i+1}, \end{cases} \tag{31}$$

with times $t_i = t_e + \log\left[(1+\delta)^{i-1}(2+\delta)/2\right]/R$, $i = 1, \ldots, i_\tau$. Here, the number of population size changes during the phase of exponential growth is given by

$$i_\tau \quad := \quad \left\lfloor \frac{R(\tau - t_e) - \log(\delta/2+1)}{\log(\delta+1)} \right\rfloor + 1.$$

Varying the growth rate δ determines the number of discretization intervals used.

The SFS (28) of the discretized version is obtained straightforwardly from (26) and (27). For the demographic parameters given above, we computed the SFS for various sample sizes up to 200 and we used $\delta = 1/4$, which was chosen large enough to provide reasonably fast computation times but sufficiently small to provide a good approximation of the exponential growth model. In the neutral case, the goodness of the approximation can be verified via the explicit solution of the SFS (Živković and Stephan 2011), which can be applied to the continuous and the discretized model. As shown in Figure 6a, where a sample size of $n = 200$ is chosen, the spectra of both continuous and piecewise-constant models agree very well with each other; the percentage error is 0.57% based on the l^2-norm, while the Kullback-Leibler divergence is about 1.76×10^{-7}.

Using our method, selection can then be incorporated into the piecewise-constant population size model. The effect of various negative selection coefficients (scaled with respect to the ancestral population size) is illustrated again for sample size $n = 200$ in Figure 6b, and the same trend can be observed for smaller sample sizes as well. It is probably not surprising that the resolution in distinguishing the selective and the neutral model rises with σ. More interestingly, differences between the neutral and the selective models are apparently more pronounced among derived alleles in intermediate to high frequency. Therefore, for large datasets where intermediate- to high-frequency derived alleles are present in sufficient numbers, one may focus more strongly on these allelic classes than on low-frequency derived ones for the statistical analysis of purifying selection.

Discussion

In this article, we extended the approach of Song and Steinrücken (2012) to develop a method for finding the transition density of a WF diffusion under genic selection and piecewise-constant effective population sizes. It can be used to obtain the SFS, but explicit knowledge of the transition density is actually not required for the computation of the SFS. To that end, we revisited and simplified the moment-based method by Evans et al. (2007) in the case of a constant population size, and utilized the result to obtain an efficient method for computing the SFS for a model with piecewise-constant population sizes.

The transition density for a variable population size can be incorporated into a hidden Markov model framework to analyze time series genetic data, as done by Steinrücken et al. (2014) in the case of a constant population size. However, in this article we focused on biological questions that can be investigated using the SFS and sampling at a single time point. The SFS has been employed into a maximum likelihood framework that can be applied to *simultaneously* infer selection coefficients and the parameters of a multi-epoch demographic model. The importance of methods that enable the joint estimation of selective and demographic parameters becomes particularly apparent in large populations, for which the scaled selection coefficient can take considerable values across large regions of the genome, so that demography and selection cannot be estimated independently.

We tested our inference method on simulated data, generated by sampling a large number of sites from the SFS of a bottleneck model for a range of selection strengths. In our parameter estimation procedure, we assumed the same model as the one used in simulations, as well as a slightly less complex model. We demonstrated that our method can accurately estimate the parameters in the majority of the bottleneck scenarios, but less so when the simpler model is assumed. The time since the bottleneck was retrieved in most of the cases even when assuming the simpler model or when the datasets simulated with selection were analyzed under neutrality. This result is encouraging for the many published demographic estimates that have been obtained assuming neutrality, but further investigation is warranted to consider more realistic models, e.g., including phases of exponential growth. Our results encourage the application of not too simple demographic models anyway.

In the African *Drosophila* sample, no or barely any negative selection was inferred when the possible impact of misoriented sites was ignored. To account for ancestral misidentification while maintaining sufficient information for inference, we applied a partly folded spectrum, where only the first two classes were folded with the corresponding last two classes. Using this partly folded spectrum, a negative selection coefficient of about $\sigma = -1$ was estimated, irrespective of assuming constant or exponentially distributed selection coefficients.

Our analyses were performed based on the bottleneck model illustrated in Figure 4. The maximum number of piecewise changes that can be incorporated into a demographic model is a function of sample size (cf. Bhaskar and Song 2014 for the neutral case), so more elaborate demographic models would have been barely accessible for this dataset, especially given the limited amount of segregating sites. It indeed turned out to be difficult to pinpoint the ancestral population size and the duration of the bottleneck, whereas the time since the bottleneck was again robustly estimated. Comparing both versions of the SFS obtained using our parameter estimates and the ones given in Duchen et al. (2013), we obtained an improved goodness-of-fit to the observed SFS from the data, and date the bottleneck as about half as old (in rescaled, but also in calendar time) based on the partly folded SFS. This discrepancy is not surprising, since primarily summary statistics of the SFS were used in their study while accounting for linkage to some extent.

We also applied a grid search to test if weak positive selection could explain the uptick of high-frequency derived variants in the unfolded empirical SFS. However, we did not obtain estimates being plausible from a biological point of view. When, as in this example, an excess of low- and high-frequency derived variants is simultaneously observed in comparison to a standard neutral model, unrealistically large estimates for σ are needed to explain the data. Positive selection on its own (and of some appreciable strength) causes a decline of low-frequency derived variants and an excess of high-frequency derived alleles, whereas an expansion (as embedded in the bottleneck model) acts in the opposite way. Therefore, both forces have to severely counteract each other so that the requirements of both ends of the SFS can be met.

We analyzed an example of exponential human population growth (Tennessen et al. 2012) to see the effect of purifying selection in the context of this model. As illustrated in Figure 6b for a sample of size 200 and various selection coefficients, intermediate- and high-frequency derived variants are more affected by exponential growth and negative selection than the low-frequency derived ones. A plausible explanation is that both exponential growth and negative selection enforce an increase of low-frequency derived variants until these classes are saturated and their impact can be observed in the complimentary high-frequency allelic classes. In general, this example illustrates nicely that even more elaborated models that include various phases of exponential growth and population declines can be computationally efficiently treated via an appropriate discretization of phases of continuous population size change, using the methods presented in this paper.

Acknowledgments

We thank the generous support of the Simons Institute for the Theory of Computing, where much of this work was carried out while we were participating in the 2014 program on “Evolutionary Biology and the Theory of Computing.” We thank valuable comments and suggestions from two reviewers. DZ thanks Anand Bhaskar, Steven N. Evans and Andreas Wollstein for helpful discussions. YSS thanks the Miller Institute for providing a Research Professorship while this paper was completed. This research is supported in part by DFG grant STE 325/14 from the Priority Program 1590 (DZ, WS), the Volkswagen Foundation grant I/84232 (DZ), an NIH grant R01-GM094402 (MS, YSS), and a Packard Fellowship for Science and Engineering (YSS).

References

Baake, M., Schlägel, U., 2011. The Peano-Baker series. Proceedings of the Steklov Institute of Mathematics 275, 155–159.

Barbour, A., Ethier, S., Griffiths, R., 2000. A transition function expansion for a diffusion model with selection. Annals of Applied Probability 10, 123–162.

Baudry, E., Depaulis, F., 2003. Effect of misoriented sites on neutrality tests with outgroup. Genetics 165, 1619–1622.

Bhaskar, A., Song, Y. S., 2014. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Annals of Statistics 42, 2469–2493.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., et al., 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genetics 4, e1000083.

Duchen, P., Živković, D., Hutter, S., Stephan, W., Laurent, S., 2013. Demographic inference reveals African and European admixture in the North American *Drosophila melanogaster* population. Genetics 191, 291–301.

Evans, S. N., Shvets, Y., Slatkin, M., 2007. Non-equilibrium theory of the allele frequency spectrum. Theoretical Population Biology 71, 109–119.

Fisher, R. A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Fu, Y.-X., 1995. Statistical properties of segregating sites. Theoretical Population Biology 48, 172–197.

Glinka, S., Ometto, L., Mousset, S., Stephan, W., De Lorenzo, D., 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics 165, 1269–1278.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., The 1000 Genomes Project, Bustamante, C. D., 2011. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108, 11983–11988.

Griffiths, R. C., 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. Theoretical Population Biology 64, 241–251.

Griffiths, R. C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society B: Biological Sciences 344, 403–410.

Griffiths, R. C., Tavaré, S., 1998. The age of a mutation in a general coalescent tree. Stochastic Models 14, 273–295.

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

Kaj, I., Krone, S. M., 2003. The coalescent process in a population of stochastically varying size. Journal of Applied Probability 40, 33–48.

Karlin, S., Taylor, H., 1981. A Second Course in Stochastic Processes. Academic Press.

Kimura, M., 1955a. Solution of a process of random genetic drift with a continuous model. Proceedings of the National Academy of Sciences of the United States of America 41, 144–150.

Kimura, M., 1955b. Random genetic drift in multi-allelic locus. Evolution 9, 419–435.

Kimura, M., 1955c. Stochastic processes and distribution of gene frequencies under natural selection. In: Cold Spring Harbor Symposia on Quantitative Biology. Vol. 20. Cold Spring Harbor Laboratory Press, pp. 33–53.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893–903.

Kimura, M., 1979. Model of effectively neutral mutations in which selective constraint is incorporated. Proceedings of the National Academy of Sciences of the United States of America 76, 3440–3444.

Krone, S. M., Neuhauser, C., 1997. Ancestral processes with selection. Theoretical Population Biology 51, 210–237.

Lenski, R. E., 2011. Evolution in action: a 50,000-generation salute to Charles Darwin. Microbe 6, 30–33.

Lukić, S., Hey, J., 2012. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. Genetics 192, 619–639.

Nei, M., Maruyama, T., Chakraborty, R., 1975. The bottleneck effect and genetic variabiliy in populations. Evolution 29, 1–10.

Racimo, F., Schraiber, J. G., 2014. Approximation to the distribution of fitness effects across functional categories in human segregating polymorphisms. PLoS Genetics 10, e1004697.

Sawyer, S. A., Hartl, D. L., 1992. Population genetics of polymorphism and divergence. Genetics 132, 1161–1176.

Slatkin, M., Hudson, R. R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562.

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

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

Steinrücken, M., Wang, Y., Song, Y. S., 2013. An explicit transition density expansion for a multi-allelic Wright–Fisher diffusion with general diploid selection. Theoretical Population Biology 83, 1–14.

Stephan, W., Li, H., 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98, 65–68.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Eimear, E., *et al.*, 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69.

Watterson, G. A., 1984. Allele frequencies after a bottleneck. Theoretical Population Biology 26, 387–407.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., Bustamante, C. D., 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of Sciences of the United States of America 102, 7882–7887.

Wooding, S., Rogers, A., 2002. The matrix coalescent and an application to human single-nucleotide polymorphisms. Genetics 161, 1641–1650.

Zhao, L., Yue, X., Waxman, D., 2013. Complete numerical solution of the diffusion equation of random genetic drift. Genetics 194, 973–985.

Živković, D., Stephan, W., 2011. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. Theoretical Population Biology 79, 184–191.

Živković, D., Wiehe, T., 2008. Second-order moments of segregating sites under variable population size. Genetics 180, 341–357.

Appendix. Derivation of (12)

Here, we derive the expression shown in (12). Using (2), (5), and (7), note that

$$\begin{aligned}\int_0^1 \pi_i(z)\Phi_n^i(z)\Phi_m^{i+1}(z)dz &= \int_0^1 \frac{c_i e^{2c_i\sigma z}}{z(1-z)} \sum_{k=0}^{\infty} u_{n,k}^i H_k^i(z) \sum_{l=0}^{\infty} u_{m,l}^{i+1} H_l^{i+1}(z)dz \\ &= \sqrt{\frac{c_i}{c_{i+1}}} \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} u_{n,k}^i u_{m,l}^{i+1} \int_0^1 \frac{e^{\sigma z(c_i-c_{i+1})}}{z(1-z)} G_k(z)G_l(z)dz. \qquad \text{(A.1)}\end{aligned}$$

Without loss of generality, assume $c_i \neq c_{i+1}$. (If $c_i = c_{i+1}$, the integral in (A.1) is trivial to evaluate using orthogonality.) Since $z^{-1}(1-z)^{-1}G_k(z)G_l(z)$ is a polynomial of order $k+l+2$, its jth derivative vanishes for $j \geq k+l+3$. Using integration by parts recursively $k+l+2$ times, we obtain

$$\int_0^1 \frac{e^{\sigma z(c_i-c_{i+1})}}{z(1-z)} G_k(z)G_l(z)dz = \sum_{j=0}^{k+l+2} (-1)^j \left[\frac{e^{\sigma z(c_i-c_{i+1})}}{[\sigma(c_i-c_{i+1})]^{j+1}} \frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\right]_0^1.$$

Note that the summand for $j=0$ in the previous equation is equal to zero and will be omitted in the remainder. Since $G_k(1-z) = (-1)^k G_k(z)$, we have

$$\frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\Bigg|_{z=0} = (-1)^{k+l+j}\frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\Bigg|_{z=1},$$

so that

$$\int_0^1 e^{\sigma z(c_i-c_{i+1})}\frac{G_k(z)G_l(z)}{z(1-z)}dz = \sum_{j=1}^{k+l+2} (-1)^j \frac{e^{\sigma(c_i-c_{i+1})} - (-1)^{k+l+j}}{\{\sigma(c_i-c_{i+1})\}^{j+1}} \frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\Bigg|_{z=1}. \qquad \text{(A.2)}$$

The modified Gegenbauer polynomials are defined as

$$G_n(x) = -x(1-x)(n+1)\cdot {}_2F_1(-n, n+3; 2; 1-x),$$

where ${}_2F_1(a,b;c;z) = \sum_{j\geq 0} a_{(j)}b_{(j)}/c_{(j)}z^j/j!$ is the Gauss hypergeometric function, $d_{(0)} = 1$, and $d_{(j)} = d(d+1)\cdots(d+j-1)$, $j \geq 1$. Applying this definition, we obtain

$$\begin{aligned}\frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\bigg|_{z=1} &= (k+1)(l+1)\sum_{u=0}^{k}\sum_{v=0}^{l}\frac{(-k)_{(u)}(k+3)_{(u)}}{2_{(u)}u!}\frac{(-l)_{(v)}(l+3)_{(v)}}{2_{(v)}v!}\\ &\qquad\times\frac{\partial^j}{\partial z^j}\left\{z(1-z)^{u+v+1}\right\}\Big|_{z=1}.\end{aligned}$$

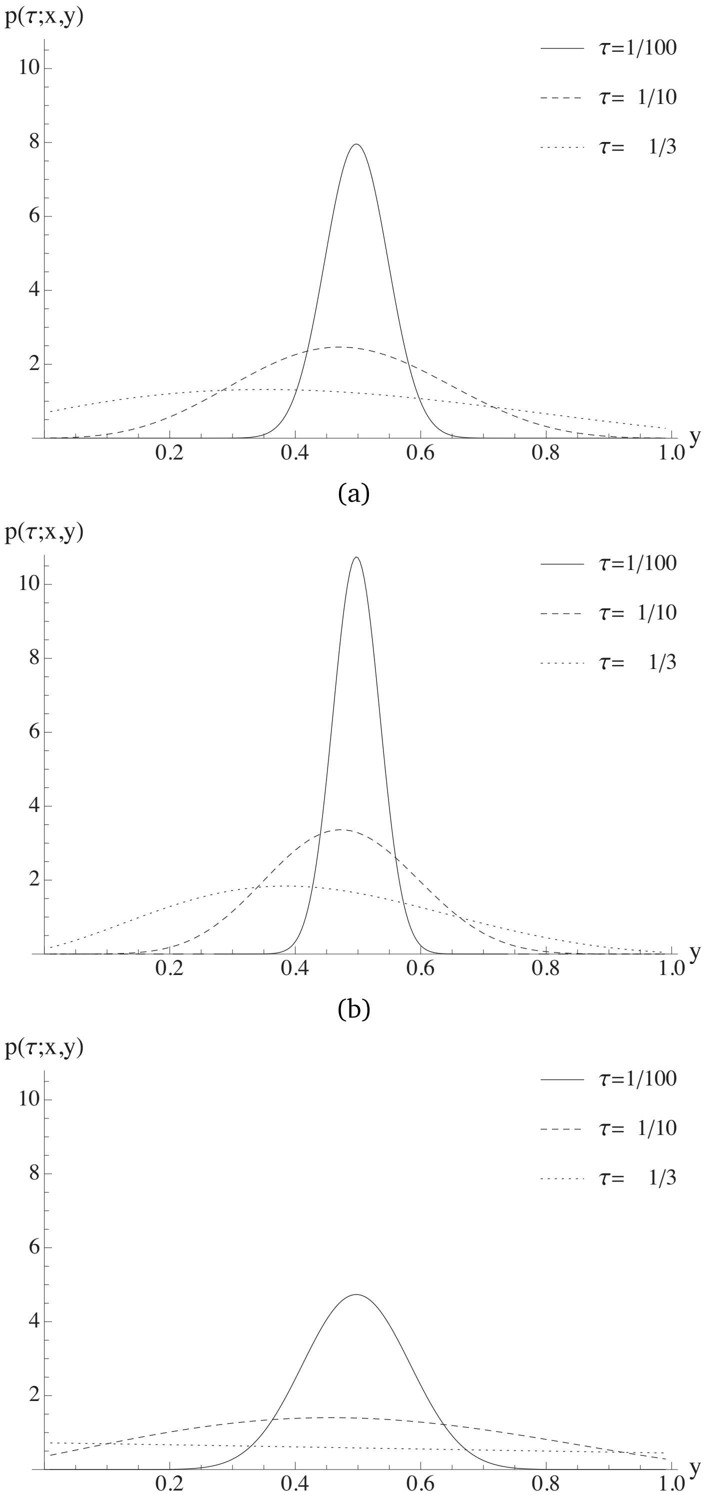
Note that the sums are finite, since $(-a)_{(b)} = 0$ for integers $a < b$. It is simple to show that

$$\frac{\partial^j}{\partial z^j}\left\{z(1-z)^{u+v+1}\right\}\Big|_{z=1} = \begin{cases}(-1)^j j!, & j = u+v+1,\\ (-1)^{j-1} j!, & j = u+v+2,\\ 0, & \text{otherwise}.\end{cases}$$

By applying this result we obtain, after some algebra,

$$\begin{aligned}\frac{\partial^j}{\partial z^j}\left\{\frac{G_k(z)G_l(z)}{z(1-z)}\right\}\bigg|_{z=1} &= (k+1)(k+1)(k+2)(l+1)\\ &\qquad\times(-1)^{j+1}\sum_{r=0}^{j-1}\binom{j}{r}\frac{(-k)_{(j-r-2)}(k+3)_{(j-r-2)}}{2_{(j-r-2)}}\frac{(-l)_{(r)}(l+3)_{(r)}}{2_{(r)}}\\ &= -\frac{k+1}{l+2}\sum_{r=0}^{j-1}\frac{j!(l+r+2)!(k+j-r)!}{r!(r+1)!(j-r)!(j-r-1)!(l-r)!(k-(j-r-2))!}\\ &= -\frac{(k+1)(l+1)j!}{(k+2)(l+2)}\sum_{r=0}^{j-1}\binom{k+2}{j-r}\binom{k+j-r}{j-r-1}\binom{l+r+2}{r+1}\binom{l}{r}. \qquad \text{(A.3)}\end{aligned}$$

Finally, combining (A.3), (A.2), and (A.1) yields the desired result.

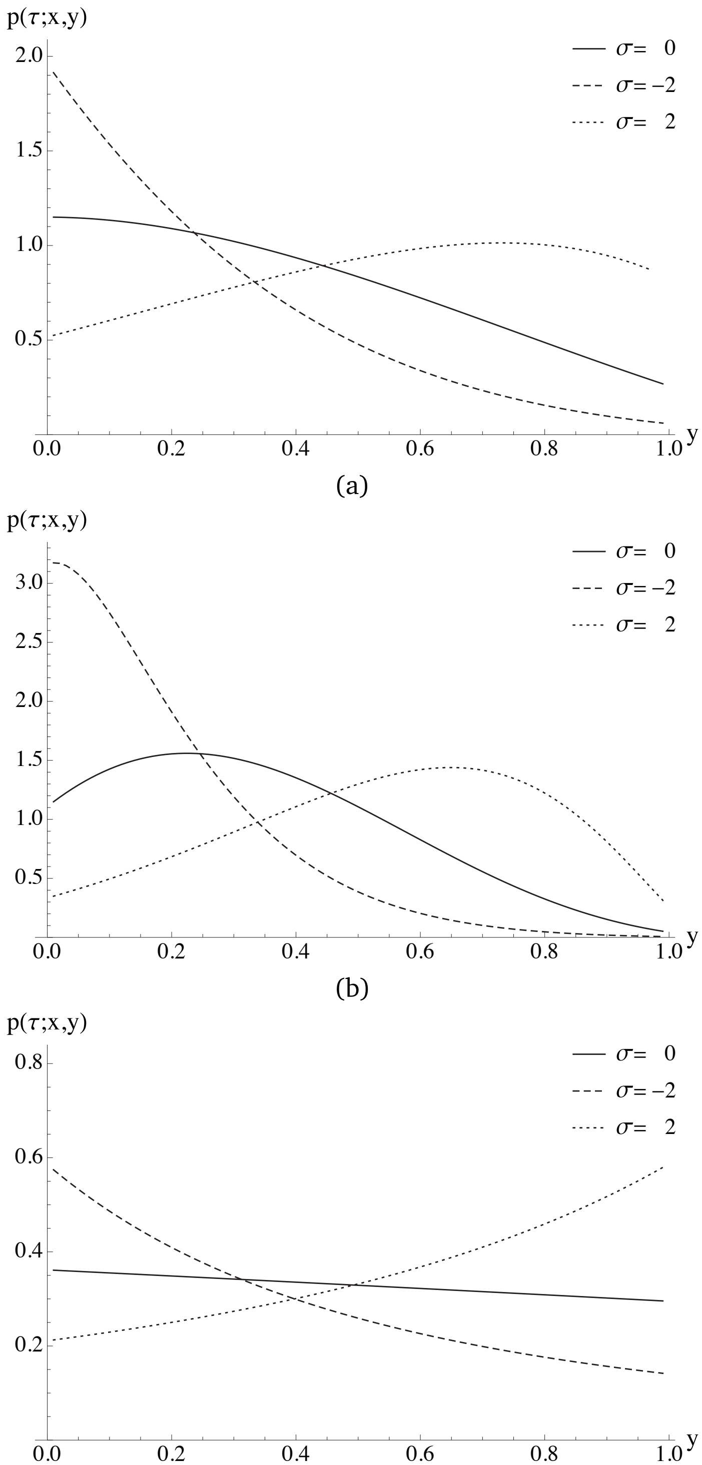


(a)

(b)

(c)

Figure 1 Transition densities for various transition times τ and a fixed selection coefficient $\sigma = -1$. In all cases, we set $x = 1/2$ and $D = 100$. (a) A single-epoch model ($K = 1$), a constant population size with $c_0 = 1$ (b) A two-epoch model ($K = 2$), with an instantaneous expansion ($c_0 = 1, c_1 = 10$, $t_1 = \tau/2$). (c) A three-epoch model ($K = 3$), with a population bottleneck followed by an expansion ($c_0 = 1, c_1 = 1/10$, $c_2 = 10$, $t_1 = \tau/4, t_2 = \tau/2$).

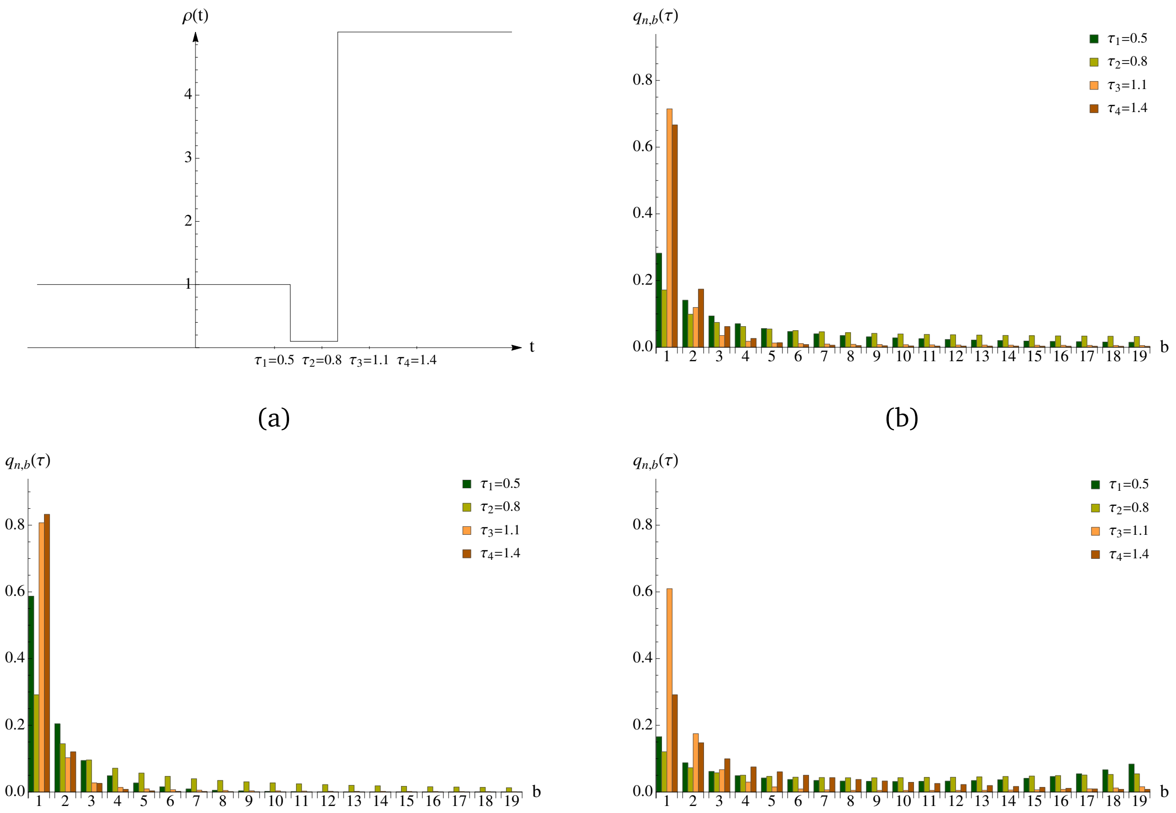


(a)

(b)

(c)

Figure 2 Transition densities for various selection coefficients σ and a fixed transition time $\tau = 1/2$. In all cases, we set $x = 1/3$ and $D = 100$. (a) A single-epoch model ($K = 1$), a constant population size with $c_0 = 1$. (b) A two-epoch model ($K = 2$), with an instantaneous expansion ($c_0 = 1, c_1 = 10$, $t_1 = \tau/2$). (c) A three-epoch model ($K = 3$), with a population bottleneck followed by an expansion ($c_0 = 1, c_1 = 1/10$, $c_2 = 10$, $t_1 = \tau/4, t_2 = \tau/2$).



(a)

(b)

(c)

(d)

Figure 3 (a) The relative population size, $\rho(t)$, is initially 1 and changes instantaneously to $1/10$ and 5 at times $6/10$ and $9/10$, respectively. The SFS of a sample of size 20 are plotted for this demography (b) without selection, (c) negative selection of $\sigma = -2$ and (d) positive selection of $\sigma = 10$. The times of sampling are illustrated in (a) and the bars are accordingly displayed from the left to the right. Truncation levels D=100 and D=500 were respectively applied for (c) negative and (d) positive selection, while the SFS was explicitly calculated for (b) neutrality.

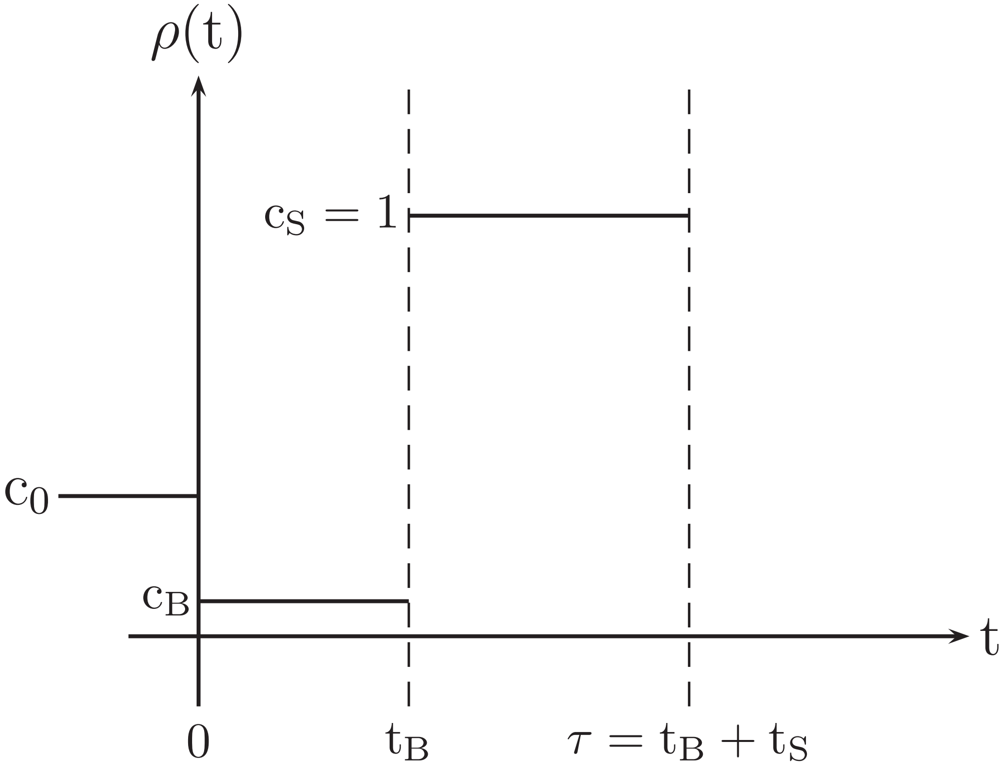
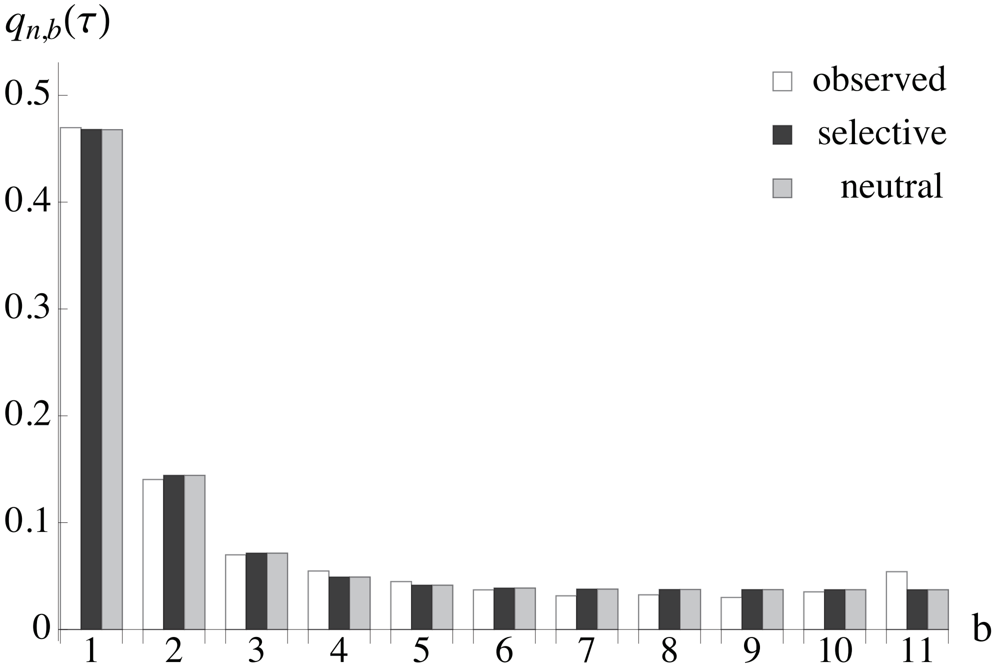
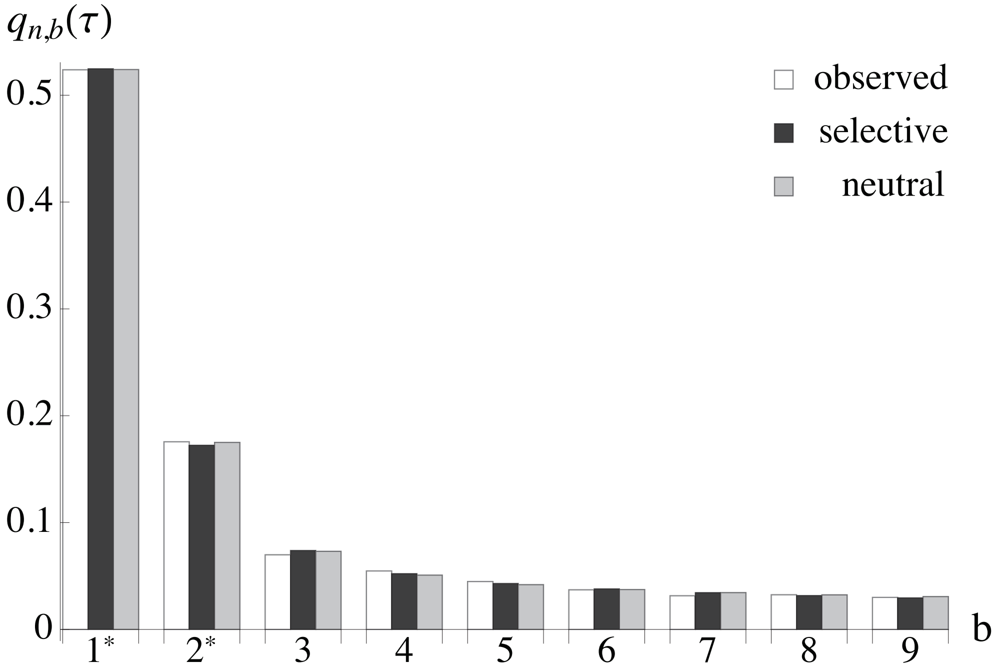


Figure 4 The population is constant in size before being instantaneously changed to relative size c_B at time zero. Then, another jump to relative population size c_S follows at time t_B, before a sample is taken at time $\tau = t_B + t_S$.

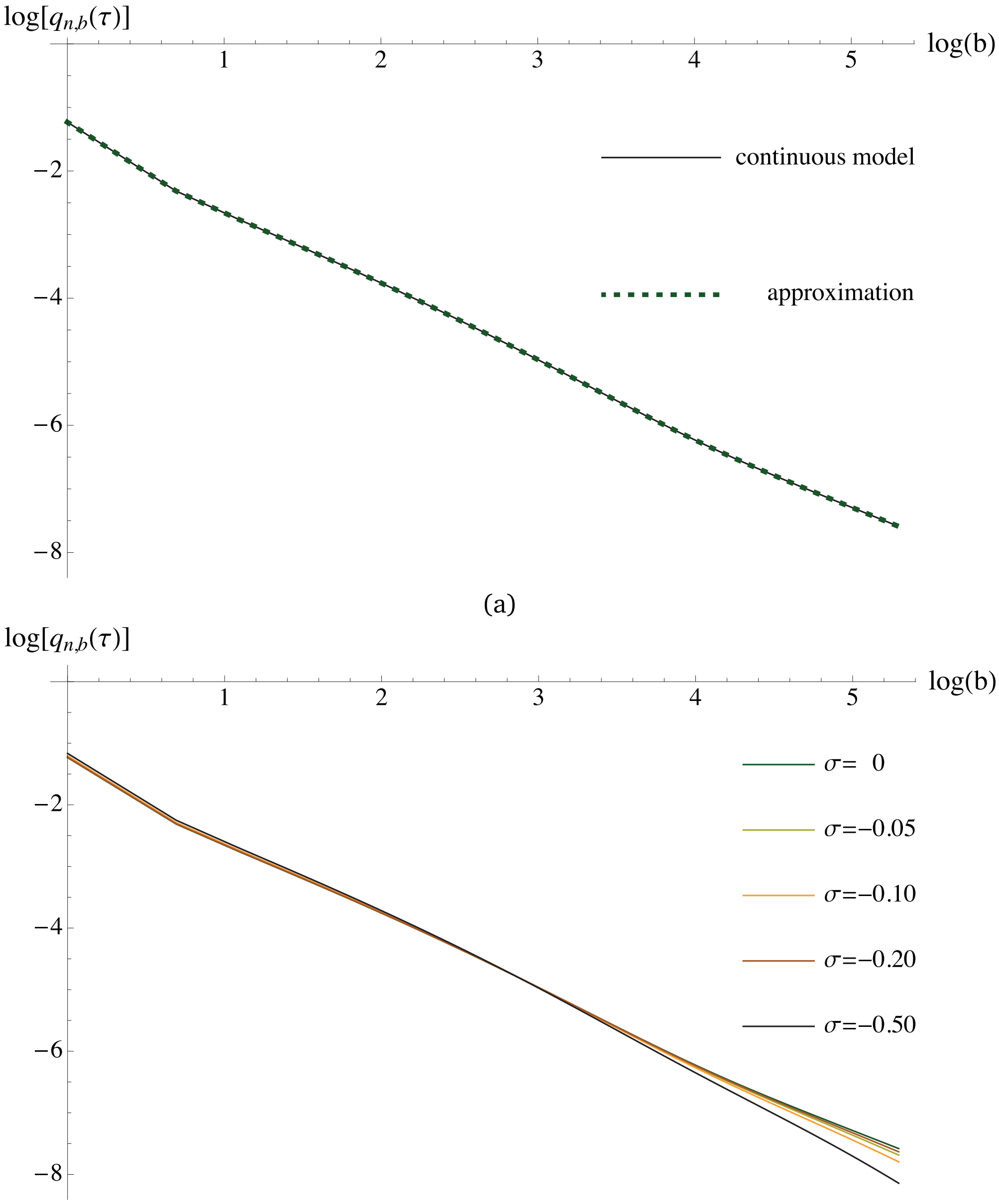


(a)



(b)

Figure 5 (a) SFS for the observed data and the most likely selective and neutral parameter estimates from left to right. (b) The same as (a) except that the allelic classes 1 and 2 were respectively folded with 11 and 10.



(a)

(b)

Figure 6 (a) Log-log plots for the SFS of the continuous and the discretized version of the estimated human African demography and neutral evolution. (b) Log-log plots for the SFS of the discretized version under various selection coefficients. The selection coefficients in the legend are ordered from top to bottom according to the function values of the high-frequency derived alleles. The sample size is given by $n = 200$ in both subfigures and a truncation level D=300 was applied in (b).

Table 1 Grid values chosen for each parameter in our optimization procedure

c_0	$\mathbf{0.011}$	0.023	$\mathbf{0.05}$	$\mathbf{0.1}$	0.224	$\mathbf{0.5}$	$\mathbf{1}$	2.154	4.642	$\mathbf{10}$	
σ	$-\mathbf{10}$	-5.848	-3.420	$-\mathbf{2}$	-1.260	-0.79	$-\mathbf{0.5}$	-0.292	-0.171	$-\mathbf{0.1}$	$\mathbf{0}$
c_B	$\mathbf{0.001}$	0.0022	$\mathbf{0.005}$	0.011	0.023	$\mathbf{0.05}$	$\mathbf{0.1}$	0.224	$\mathbf{0.5}$	$\mathbf{1}$	
t_B	$\mathbf{0.001}$	0.0022	$\mathbf{0.005}$	0.011	0.023	$\mathbf{0.05}$	$\mathbf{0.1}$	0.224	$\mathbf{0.5}$	$\mathbf{1}$	
t_S	$\mathbf{0.001}$	0.0022	$\mathbf{0.005}$	0.011	0.023	$\mathbf{0.05}$	$\mathbf{0.1}$	0.224	$\mathbf{0.5}$	$\mathbf{1}$	

The underlying bottleneck model is illustrated in Figure 4. Grid values c_0 were considered for the 5-parameter model, whereas $c_0 = c_S$ in the 4-parameter model. The grid values for the remaining parameters were applied in both scenarios. The ratio of two consecutive values remains constant between (and including the) two subsequent bold entries.

Table 2 Parameter estimation results based on 10,000 sampled sites

		$\hat{c}_0$	$\hat{\sigma}$	$\hat{c}_B$	$\hat{t}_B$	$\hat{t}_S$
True parameters		0.5	0 or -2	0.1	0.1	0.05
	5%	0.5		0.1	0.1	0.05
(A1)	Median	0.5		0.1	0.1	0.05
	95%	0.5		0.1	0.1	0.05
	5%	0.22		0.02	0.005	0.05
(A2)	Median	0.22		0.1	0.05	0.05
	95%	0.22		0.1	0.05	0.05
	5%	0.22	-2	0.05	0.01	0.05
(A3)	Median	0.5	-2	0.1	0.1	0.05
	95%	0.5	0	0.1	0.1	0.05
	5%	0.5	-0.5	0.1	0.001	0.05
(A4)	Median	0.5	0	0.1	0.1	0.05
	95%	2.15	0	0.1	0.1	0.05

SFS were computed for the true parameters and the demography illustrated in Figure 4 ($c_0 = 1/2$, $c_S = 1$). Then, 10,000 sites were sampled according to the SFS of the neutral and the selective scenario, and this procedure was repeated 200 times each. The log-likelihood values were maximized over the parameter spaces as specified in the main text, and the table reports the median, the 0.05 and the 0.95 quantiles. The four cases correspond to assuming (A1) neutrality when $\sigma = 0$, (A2) neutrality when $\sigma = -2$, (A3) presence of selection when $\sigma = -2$, and (A4) presence of selection when $\sigma = 0$.

Table 3 Parameter estimation results based on the expected SFS assuming neutrality when the true model is under selection

Selection coefficient	$\sigma = -1/2$	$\sigma = -2$
Demographic model	$(\hat{c}_0, \hat{c}_B, \hat{t}_B, \hat{t}_S)$	$(\hat{c}_0, \hat{c}_B, \hat{t}_B, \hat{t}_S)$
Constant population size	$(0.500, 1.00, 1.10 - \hat{t}_S, \quad \hat{t}_S \quad)$	$(0.100, 1.000, 0.523 - \hat{t}_S, \quad \hat{t}_S \quad)$
Bottleneck with $t_S = 1/200$	$(0.224, 0.05, 0.05 \qquad , 0.002)$	$(0.224, 0.100, 0.050 \qquad , 0.005)$
Bottleneck with $t_S = 1/20$	$(0.500, 0.10, 0.10 \qquad , 0.050)$	$(0.224, 0.100, 0.050 \qquad , 0.050)$
Bottleneck with $t_S = 1/2$	$(1.000, 0.05, 0.10 \qquad , 0.500)$	$(0.100, 1.000, 0.324 - \hat{t}_S, \quad \hat{t}_S \quad)$

SFS were computed for the following demographic scenarios and selection coefficients. In terms of the demography, either a constant population size was assumed, or a bottleneck model according to Figure 4 with parameters $c_0 = 1/2$, $c_B = 1/10$, $c_S = 1$, $t_B = 1/10$ and $t_S = 1/200$, $1/20$ or $1/2$. The selection coefficients are $\sigma = -1/2$ and -2. The parameter estimates were obtained according to the procedure and the parameter spaces described in the main text and by assuming neutrality in each case. In the first row, and in the forth row, second column, we obtained $\hat{c}_B = 1$, i.e. an instantaneous expansion occurs as the only size change $\hat{t}_B + \hat{t}_S$ before sampling.

Table 4 Parameter estimation results based on 10,000 sampled sites when the ancestral population size c_0 is incorrectly assumed to equal the current size c_S, while the true model has $c_0 = 1/2$ and $c_S = 1$.

		c_0	$\hat{\sigma}$	$\hat{c}_B$	$\hat{t}_B$	$\hat{t}_S$
True parameters		0.5	0 or -2	0.1	0.1	0.05
	5%			0.1	0.22	0.02
(A1)	Median			0.1	0.22	0.05
	95%			0.22	0.5	0.05
	5%			0.1	0.22	0.05
(A2)	Median			0.1	0.22	0.05
	95%			0.22	1	0.05
	5%		-0.79	0.1	0.22	0.05
(A3)	Median		-0.79	0.1	0.22	0.05
	95%		-0.5	0.1	0.22	0.05
	5%		-1.26	0.01	0.01	0.05
(A4)	Median		-1.26	0.05	0.05	0.05
	95%		-0.79	0.1	0.1	0.1

SFS were computed for the true parameters and the demography illustrated in Figure 4 ($c_0 = 1/2$, $c_S = 1$). Then, 10,000 sites were sampled according to the SFS of the neutral and the selective scenario, and this procedure was repeated 200 times each. The log-likelihood values were maximized over the 4-parameter space (where $c_0 = c_S$ is assumed), and the table reports the median, the 0.05 and the 0.95 quantiles. The four cases correspond to assuming (A1) neutrality when $\sigma = 0$, (A2) neutrality when $\sigma = -2$, (A3) presence of selection when $\sigma = -2$, and (A4) presence of selection when $\sigma = 0$.

Table 5 Parameter estimation results based on the expected SFS when the ancestral population size c_0 is incorrectly assumed to equal the current size c_S, while the true model has $c_0 = 1/2$ and $c_S = 1$.

Selection coefficient	$\sigma = 0$	$\sigma = -1/2$	$\sigma = -2$
Demographic model	$(\hat{\sigma}, \hat{c}_B, \hat{t}_B, \hat{t}_S)$ $(\hat{c}_B, \hat{t}_B, \hat{t}_S)$	$(\hat{\sigma}, \hat{c}_B, \hat{t}_B, \hat{t}_S)$ $(\hat{c}_B, \hat{t}_B, \hat{t}_S)$	$(\hat{\sigma}, \hat{c}_B, \hat{t}_B, \hat{t}_S)$ $(\hat{c}_B, \hat{t}_B, \hat{t}_S)$
Bottleneck with $t_S = 1/200$	$(-3.420, 0.023, 0.050, 0.001)$ $(0.224, 0.224, 0.011)$	$(-0.171, 0.224, 0.224, 0.011)$ $(0.224, 0.224, 0.011)$	$(-5.848, 0.023, 0.050, 0.001)$ $(0.023, 0.100, 0.001)$
Bottleneck with $t_S = 1/20$	$(-1.260, 0.050, 0.050, 0.050)$ $(0.100, 0.224, 0.050)$	$(-2., 0.050, 0.050, 0.050)$ $(0.100, 0.224, 0.050)$	$(-0.794, 0.100, 0.224, 0.050)$ $(0.100, 0.224, 0.050)$
Bottleneck with $t_S = 1/2$	$(-0.292, 0.224, 0.500, 0.500)$ $(0.224, 0.500, 0.500)$	$(0, 0.050, 0.100, 0.500)$ $(0.050, 0.100, 0.500)$	$(-2., 0.224, 0.500, 0.500)$ $(0.050, 0.224, 0.500)$

SFS were computed for the following demographic scenarios and selection coefficients. In terms of the demography, a bottleneck model was assumed according to Figure 4 with parameters $c_0 = 1/2$, $c_B = 1/10, t_B = 1/10$ and $t_S = 1/200$, $1/20$ or $1/2$. The selection coefficients were chosen as $\sigma = 0$, $-1/2$ and -2. The parameter estimates were obtained according to the model assuming $c_0 = c_S$ (the grid for the 4-parameter space being a subset of the grid for the 5-parameter space) and by assuming either selection or neutrality in each case.

Table 6 Parameter estimation results based on the unfolded and the partly folded SFS and constant selection coefficients

	$\hat{\sigma}$	$\hat{c}_0$	$\hat{c}_B$	$\hat{t}_B/\hat{c}_B$	$\hat{t}_S$	L
			Unfolded SFS			
MLE	0	3.162	0.001	2.633	0.164	-5962.96
Top 10	$[-0.008, 0]$	3.162	$[0.001, 0.003]$	2.633	0.164	$[-5963.01, -5962.96]$
Top 100	$[-0.063, 0]$	$[1.468, 6.813]$	$[0.001, 0.013]$	$[1.867, 3.310]$	$[0.154, 0.174]$	$[-5963.37, -5962.96]$
			Partly folded SFS			
MLE	-0.906	0.5	0.1	1.181	0.106	-5098.29
		0.5	0.1	1.402	0.113	-5098.51
Top 10	$[-1.32, -0.67]$	$[0.5, 4.642]$	0.1	$[1.181, 1.763]$	$[0.106, 0.113]$	$[-5098.31, -5098.29]$
Top 100	$[-1.74, -0.50]$	$[0.5, 10.00]$	$[0.013, 0.1]$	$[0.837, 2.348]$	$[0.099, 0.136]$	$[-5098.39, -5098.29]$

The demographic histories were estimated with and without constant selection for the demographic model illustrated in Figure 4 for the entire dataset of 3234 polymorphic sites. The estimates and their likelihood values are based on a refined grid described in the main text and shown for the unfolded and a partly folded SFS. In addition to the MLEs, the sets of the 10 and the 100 likeliest parameter combinations were also estimated. From these sets, the two outermost estimates were chosen for each single parameter and for the likelihood value L to obtain the outlined parameter ranges.

Table 7 Parameter estimation results for partly folded SFS and exponentially distributed selection coefficients

β	$\hat{c}_0$	$\hat{c}_B$	$\hat{t}_B/\hat{c}_B$	$\hat{t}_S$	L
0.1	2	0.01	0.631	0.126	-5101.36
0.2	2	0.05	1	0.158	-5098.59
0.5	1	0.1	1.584	0.1	-5098.50
1	0.5	0.1	1.259	0.1	-5098.43
2	2	0.1	2.508	0.126	-5098.69
5	0.5	0.1	1.259	0.126	-5098.67
10	0.5	0.1	1.259	0.126	-5098.73
20	0.5	0.1	1.259	0.126	-5098.79
50	0.5	0.1	1.259	0.126	-5098.84
100	0.5	0.1	1.259	0.126	-5098.86

The demographic histories were estimated based on exponentially distributed selection coefficients and for the demographic model illustrated in Figure 4 for the entire dataset of 3234 polymorphic sites. First, allelic spectra were evaluated for 12,600 different demographic parameter combinations and 100 σ values each. Then, polynomial curves of degree three were fitted between successive σ values and for every single demographic parameter combination, before a numerical integration against a gamma distribution with $\alpha = 1$ and 10 different values of β was applied. From the allelic spectra, now being corrected for varying selection coefficients, the SFS were obtained. The resultant MLEs are shown for the various choices of β.