An Approximate Markov Model for the Wright-Fisher Diffusion and its Application to Time Series Data

Anna Ferrer-Admetlla*,†,‡**, Christoph Leuenberger**^{S}**, Jeffrey D. Jensen**†,** **and Daniel Wegmann**^{*,‡,1}

*Department of Biology, University of Fribourg, Fribourg, Switzerland, †Department of Life Science, Ecole polytechnique federal de Lausanne, Lausanne, Switzerland, ‡Swiss Institute of Bioinformatics, Fribourg, Switzerland, ^{S}Department of Mathematics, University of Fribourg, Fribourg, Switzerland, **Swiss Institute of Bioinformatics, Lausanne, Switzerland

ABSTRACT The joint and accurate inference of selection and demography from genetic data is considered a particularly challenging question in population genetics, since both process may lead to very similar patterns of genetic diversity. However, additional information for disentangling these effects may be obtained by observing changes in allele frequencies over multiple time points. Such data is common in experimental evolution studies, as well as in the comparison of ancient and contemporary samples. Leveraging this information, however, has been computationally challenging, particularly when considering multi-locus data sets. To overcome these issues, we introduce a novel, discrete approximation for diffusion processes, termed *mean transition time approximation*, which preserves the long-term behavior of the underlying continuous diffusion process. We then derive this approximation for the particular case of inferring selection and demography from time series data under the classic Wright-Fisher model and demonstrate that our approximation is well suited to describe allele trajectories through time, even when only a few states are used. We then develop a Bayesian inference approach to jointly infer the population size and locus-specific selection coefficients with high accuracy, and further extend this model to also infer the rates of sequencing errors and mutations. We finally apply our approach to recent experimental data on the evolution of drug resistance in Influenza virus, identifying likely targets of selection and finding evidence for much larger viral population sizes than previously reported.

KEYWORDS Wright-Fisher Model; Diffusion Approximation; Discrete Markov Model; Hidden Markov Model; Time Series Data

Detecting signatures of past selective events gives insights into the evolutionary history of a species and elucidates the interaction between genotype and phenotype, providing important functional information. Unfortunately, a population's demographic history is a major confounding factor when inferring past selective events, particularly because demographic events can mimic many of the molecular signatures of selection (Andolfatto and Przeworski 2000; Nielsen 2005). Despite efforts to create statistics robust to demography, all currently available methods to detect selection are prone to misinference under non-equilibrium demography.

Some of these issues can potentially be overcome by using multi-time point data, as the trajectory of even a single allele contains valuable information about the underlying selection coefficient. Owing to advances in sequencing technologies, such multi-time point data are becoming increasingly common from experimental evolution (Foll *et al.* 2014), from longitudinal medical or ecological studies (Wei *et al.* 1995; Renzette *et al.* 2014), and through ancient samples (Wilde *et al.* 2014; Sverrisdóttir *et al.* 2014). However, computationally efficient and accurate methods to infer demography and selection jointly from such data sets are still limited.

A natural and common way of modeling such time series data is in a Hidden Markov-Model (HMM) framework, which allows efficient integration over the distribution of unobserved states of the true population frequencies, thus allowing calculation of the likelihood based on the observed samples. Williamson & Slatkin (1999) (Williamson and Slatkin 1999), for instance, developed a maximum-likelihood approach based on such an HMM to infer the population size N from samples taken at different time

[1] Department of Biology, University of Fribourg, Chemin du Musée 10, 1200 Fribourg, Switzerland, daniel.wegmann@unifr.ch

points. More recently, similar approaches have been developed to infer population size along with the selection coefficient of a selected locus for which time series data is available (Bollback *et al.* 2008; Malaspinas *et al.* 2012).

All such approaches, however, are plagued by the problem that the number of hidden frequency states is equal to the population size, which renders HMM applications computationally unfeasible for large populations. Different routes have been taken to overcome this. One approach is to model the underlying Wright-Fisher process as a continuous diffusion process, which is then discretized for numerical integration using a numerical difference scheme (Bollback *et al.* 2008). Since, this approach remains computationally expensive, it was later suggested to directly model the diffusion process on a more coarse-grained grid (Malaspinas *et al.* 2012). Under this approach, their generator matrix for the transition between the coarse-grained states is then approximated by fitting the first and second infinitesimal moments. Unfortunately, the minimum number of states required is still computationally prohibitive for large values of $\gamma = 2Ns$ (Malaspinas *et al.* 2012). For this reason, the most recent reported method resorted to simulation based Approximate Bayesian Computation (ABC), which allowed the joint inference of locus-specific selection coefficients for many loci (Foll *et al.* 2014, 2015). However, this method requires first estimating the population size under the assumption that all loci are neutral, and thus may be biased when many loci are under selection.

Here we introduce a novel framework by approximating the WF-process with a coarse-grained Markov-Model that exactly preserves the expected waiting times for transition between states. This is achieved by exploiting the theory of Green's function for diffusion processes. Contrary to previous approaches, our approximation matches the WF-process closely even when only very few states are considered, regardless of $\gamma = 2Ns$. As we show with extensive simulations and a data application from experimental evolution, our method allows for accurate joint inference of both population size and locus-specific selection coefficients even in the presence of pervasive selection. Further, it is readily extended to incorporate population size changes, sequencing errors or the appearance of novel mutations.

Models

Mean Transition Time Approximation

Let $X(t)$ be a diffusion process on the state space $[0,1]$. This is a continuous-time Markov process with continuous sample paths and with infinitesimal generator

$$Lf = \frac{1}{2}a(x)\frac{d^2}{dx^2}f + b(x)\frac{d}{dx}f. \tag{1}$$

For general information about diffusion processes we refer to Durrett (2008, ch.7) and Etheridge (2011, ch.3).

The classical example in population genetics is the Fisher-Wright diffusion which we discuss below. We seek to find a discrete-state Markov process $U(t)$ which approximates $X(t)$. For this purpose, we subdivide the unit interval $[0,1]$ into, not necessarily equidistant, frequencies

$$u_0 = 0 < u_1 < \ldots < u_{K-1} < u_K = 1.$$

These form the states of $U(t)$. For two states u_i, u_j, consider the transition time to the first visit of u_j when starting at u_i:

$$T^U_{u_i \to u_j} = \inf\{t : U(t) = u_j \text{ for } U(0) = u_i\}.$$

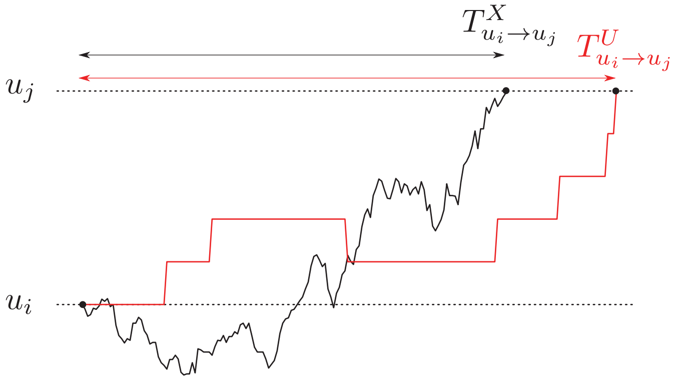

Figure 1 Mean transition time approximation of Markov processes. Shown are the realizations of a continuous diffusion process $X(t)$ (black) and a discrete-state Markov process $U(t)$ (red) starting at u_i until they reach u_j for the first time. If the expected waiting time for such a transition is the same for both processes for all pairs of states u_i, u_j, we say that $U(t)$ is a mean transition time approximation of $X(t)$.

Similarly we define the transition time for the diffusion process $X(t)$. We say that $U(t)$ is a *mean transition time approximation* of $X(t)$ if

$$\mathbb{E}\left[T^U_{u_i \to u_j}\right] = \mathbb{E}\left[T^X_{u_i \to u_j}\right] \tag{2}$$

for all pairs of states u_i, u_j (see Fig. 1). This condition guarantees that the paths of $X(t)$ and $U(t)$ exhibit comparable long-time behavior. In the following we show how to construct the Markov process $U(t)$ from the diffusion process $X(t)$ using the theory of Green's function.

We begin by recalling some notions for diffusion processes. The natural scale of the process $X(t)$ is given by

$$\phi(x) = \int^x \psi(y)dy, \tag{3}$$

where $\psi(y) = \exp(-2\int^y \frac{b(z)}{a(z)}dz)$, see Durrett (2008, p.264). The so-called speed measure is defined by

$$m(y) = \frac{1}{a(y)\psi(y)}. \tag{4}$$

According to Theorem 7.16 in Durrett (2008), the Green's function for an interval $(u,v) \subseteq [0,1]$ is given by

$$G(x,y) = \begin{cases} 2m(y)\frac{(\phi(v)-\phi(x))(\phi(y)-\phi(u))}{\phi(v)-\phi(u)}, & u \leq y \leq x \\ 2m(y)\frac{(\phi(x)-\phi(u))(\phi(v)-\phi(y))}{\phi(v)-\phi(u)}, & x < y \leq v. \end{cases} \tag{5}$$

Denote by $T_{x\to u}$ or $T_{x\to v}$ the time to first visit of u or v, respectively, starting at x. Then $T^v_u = \min(T_{x\to u}, T_{x\to v})$ is the exit time from the interval (u_1, u_2), given the process is at x at time $t = 0$. One can show (see Durrett (2008, p.279))

$$\mathbb{E}(T^v_u) = \int\limits_u^v G(x,y)dy. \tag{6}$$

Moreover, the probability of exiting at the lower limit u is

$$\mathbb{P}(T_{x\to v} > T_{x\to u}) = \frac{\phi(v)-\phi(x)}{\phi(v)-\phi(u)}. \tag{7}$$

We now want to determine the instantaneous transition rates $q_{i,j}$ of the discrete-state Markov process $U(t)$. Recall the definitions

$$\mathbb{P}\left[U(t+dt)=u_k|U(t)=u_k\right]=1-q_{k,k}dt+o(dt),$$
$$\mathbb{P}\left[U(t+dt)=u_{k+1}|U(t)=u_k\right]=q_{k,k+1}dt+o(dt)$$

and

$$\mathbb{P}\left[U(t+dt)=u_{k-1}|U(t)=u_k\right]=q_{k,k-1}dt+o(dt).$$

The sojourn time of state u_k, i.e. the time interval of $U(t)$ spent in state u_k, is an exponential random variable with parameter $q_{k,k}$. Since the expectation of this exponential variable is $1/q_{k,k}$ our condition (2) enforces

$$q_{k,k}=\frac{1}{\mathbb{E}(T_{k-1}^{k+1})},$$

where we write $k-1$ instead of u_k etc. in order to unburden the notation. From this we get

$$q_{k,k+1}=\frac{\mathbb{P}(T_{k\to k+1}<T_{k\to k-1})}{\mathbb{E}(T_{k-1}^{k+1})}$$

and

$$q_{k,k-1}=\frac{\mathbb{P}(T_{k\to k+1}>T_{k\to k-1})}{\mathbb{E}(T_{k-1}^{k+1})}.$$

We can now form the tridiagonal generator matrix

$$\boldsymbol{Q}=\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ q_{1,0} & -q_{1,0}-q_{1,2} & q_{1,2} & 0 & \cdots \\ 0 & q_{2,1} & -q_{2,1}-q_{2,3} & q_{2,3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

The transition matrix of the Markov process $U(t)$ is given by

$$\mathbf{P}(t)=\exp t\boldsymbol{Q}. \tag{8}$$

Application to Wright-Fisher Models

We will consider a classic Wright-Fisher Model of two alleles that segregate in a population of size $2N$. Time t is measured in generations of the Wright-Fisher process. In the presence of a non-vanishing dominance coefficient h the fitnesses of the three genotypes are given by $w_{AA}=1+s$, $w_{Aa}=1+hs$, and $w_{aa}=1$. Under such a model, the infinitesimal mean, which corresponds to the change in allele frequency, is then given by (Ewens 2004, p.13)

$$\begin{aligned} b(x) &= \frac{w_{AA}x^2+w_{Aa}x(1-x)}{w_{AA}x^2+2w_{Aa}x(1-x)+w_{aa}(1-x)^2}-x \\ &= \frac{x(1-x)s(x+h-2hx)}{1+sx(x+2h-2hx)}. \end{aligned} \tag{9}$$

Let $X(t)$ be a diffusion process corresponding to the frequency of allele A. As shown by Lacerda and Seoighe (2014), an excellent approximation of the Wright-Fisher process is obtained by setting

$$a(x)=\frac{x(1-x)}{2N} \tag{10}$$

and

$$b(x)=\frac{\bar{\sigma}_k x(1-x)}{1+\sigma_k x} \tag{11}$$

in the infinitesimal generator (1), where

$$\sigma_k=s(2h+u_k(1-2h)) \tag{12}$$

and

$$\bar{\sigma}_k=s(h+u_k(1-2h)) \tag{13}$$

when $u_{k-1}\le x\le u_{k+1}$

Note that in the standard diffusion approximation the denominator term in $b(x)$ is often omitted. But the above choice yields a much more accurate approximation to the WF process (Lacerda and Seoighe 2014).

From (3) and (4) we get

$$\psi(y)=\exp\left(-2\int^y\frac{2N\bar{\sigma}_k}{1+\sigma_k x}dx\right)=(1+sy)^{-4N\bar{\sigma}_k/\sigma_k} \tag{14}$$

and

$$\phi(x)=\int^x\psi(y)dy=-\frac{1}{M_k\sigma_k}(1+\sigma_k y)^{-M_k}, \tag{15}$$

where we have set

$$M_k=4N\frac{\bar{\sigma}_k}{\sigma_k}-1. \tag{16}$$

For the speed measure we obtain

$$m(y)=\frac{1}{a(y)\psi(y)}=\frac{2N}{y(1-y)}(1+\sigma_k y)^{M_k+1}. \tag{17}$$

Consider three consecutive states u_{k-1}, u_k and u_{k+1}. For the probability to exit at the lower state we get

$$\begin{aligned} P_\downarrow := \mathbb{P}(T_{k\to k+1}>T_{k\to k-1}) &= \frac{\phi(u_{k+1})-\phi(u_k)}{\phi(u_{k+1})-\phi(u_{k-1})} \\ &= \frac{(1+\sigma_k u_k)^{-M_k}-(1+\sigma_k u_{k+1})^{-M_k}}{(1+\sigma_k u_{k-1})^{-M_k}-(1+\sigma_k u_{k+1})^{-M_k}} \\ &= \frac{\left(\frac{1+\sigma_k u_{k+1}}{1+\sigma_k u_k}\right)^{M_k}-1}{\left(\frac{1+\sigma_k u_{k+1}}{1+\sigma_k u_{k-1}}\right)^{M_k}-1}. \end{aligned} \tag{18}$$

The probability for exit at the upper state is

$$P_\uparrow := \mathbb{P}(T_{k\to k+1}<T_{k\to k-1})=1-P_\downarrow.$$

Observe that the Green's function is calculated by

$$G(u_k,y)=\begin{cases} G_\downarrow(u_k,y):=2P_\downarrow m(y)(\phi(y)-\phi(u_{k-1})), & u_{k-1}\le y\le u_k \\ G_\uparrow(u_k,y):=2P_\uparrow m(y)(\phi(u_{k+1})-\phi(y)), & u_k<y\le u_{k+1}. \end{cases}$$

Using the quantities calculated above we get for the two parts of the Green's function:

$$\begin{aligned} G_\downarrow(u_k,y) &= \frac{4NP_\downarrow}{\sigma_k M_k y(1-y)}(1+\sigma_k y)^{M_k+1}\cdot\ldots \\ &\quad\cdot\left((1+\sigma_k u_{k-1})^{-M_k}-(1+\sigma_k y)^{-M_k}\right) \\ &= \frac{4NP_\downarrow}{\sigma_k M_k}\frac{1+\sigma_k y}{y(1-y)}\left(\left(\frac{1+\sigma_k y}{1+\sigma_k u_{k-1}}\right)^{M_k}-1\right) \end{aligned} \tag{19}$$

and

$$\begin{aligned} G_\uparrow(u_k,y) &= \frac{4NP_\uparrow}{\sigma_k M_k y(1-y)}(1+\sigma_k y)^{M_k+1}\cdot\ldots \\ &\quad\cdot\left((1+\sigma_k y)^{-M_k}-(1+\sigma_k u_{k+1})^{-M_k}\right) \\ &= \frac{4NP_\uparrow}{\sigma_k M_k}\frac{1+\sigma_k y}{y(1-y)}\left(1-\left(\frac{1+\sigma_k y}{1+\sigma_k u_{k+1}}\right)^{M_k}\right). \end{aligned} \tag{20}$$

With numerical integration we can determine

$$\mathbb{E}(T_{k-1}^{k+1}) = \mathcal{E}_{\downarrow} + \mathcal{E}_{\uparrow} = \int_{u_{k-1}}^{u_k} G_{\downarrow}(u_k, y)dy + \int_{u_k}^{u_{k+1}} G_{\uparrow}(u_k, y)dy.$$

Specifically, we use the extended Simpson's rule for the numerical integration (Press 2007), which we found to give accurate results with typically only 8 or 10 intervals.

If $\gamma = 2Ns$ is large, we get approximations for the Green's function which allow for analytic expressions of the integrals (see Appendix). Similarly, analytic expressions can be found in the special case $s = 0$ (see Appendix).

Bayesian Inference

Consider that at the times T_t, $t = 0, \ldots, T$, samples of sizes M_t were taken from the population and m_t alleles A were observed in these samples. In this section, we describe how the mean transition time approximation introduced above can be embedded into a Bayesian inference scheme to estimate the population size $2N$ and the locus-specific selection coefficient jointly from time series data.

As has been noted previously (Williamson and Slatkin 1999; Bollback *et al.* 2008; Malaspinas *et al.* 2012; Mathieson and McVean 2013; Steinrücken *et al.* 2014; Lacerda and Seoighe 2014), a natural way of modeling both the underlying evolutionary process as well as the process of sampling is a Hidden Markov Model (HMM). Under the assumption that the population size between two time points T_t and T_{t+1} is constant at N_t, the transition matrix of such an HMM from state $U(T_t)$ to state $U(T_{t+1})$ is calculated by

$$\mathbf{P}^t = \exp(\Delta_t \boldsymbol{Q}^t),$$

where $\Delta_t = T_{t+1} - T_t$ and the generator matrix $\boldsymbol{Q}^t$ is determined as explained above using $N = N_t$. We note here that this framework allows for instantaneous population size changes to occur at every time t during the HMM. However, we will henceforth only deal with situations in which the population size is assumed to be constant across the whole sampling period.

Following previous implementations (e.g. (Williamson and Slatkin 1999; Bollback *et al.* 2008; Malaspinas *et al.* 2012; Mathieson and McVean 2013; Steinrücken *et al.* 2014; Lacerda and Seoighe 2014)), we will assume that the sampling of alleles from the underlying population frequency is binomial, i.e.

$$\mathbb{P}(m_t = m | U(T_t) = u_k) = \binom{M_t}{m} u_k^m (1 - u_k)^{M_t - m}.$$

However, for large sample sizes, the few states u_k may be too coarse grained to capture the region of high emission probability. We thus propose to integrate the emission probabilities against a smoothing kernel. We chose to implement a beta distribution kernel, which is the conjugate prior to the binomial emission probabilities. As a result, this choice leads to a beta-binomial emission probability that can be evaluated analytically. Specifically, we chose to use a beta-kernel with mean u_k and standard deviation $\sigma_k = (u_{k+1} - u_{k-1})/4$, such that the interval $[u_{k-1}, u_{k+1}]$ corresponds to $u_k \pm 2\sigma_k$ in the case of equidistant states. Under this choice, the emission probabilities are then calculated by

$$\mathbb{P}(m_t = m | U(T_t) = u_k) = \binom{M_t}{m} \frac{B(m + \alpha_k, M_t - m + \beta_k)}{B(\alpha_k, \beta_k)},$$

where $B(\cdot, \cdot)$ is the Beta function and the parameters α_k and β_k are determined via the moment estimators for a beta distribution

$$\alpha_k = u_k \left(\frac{u_k(1 - u_k)}{\sigma_k^2} - 1 \right), \quad \beta_k = \frac{\alpha_k(1 - u_k)}{u_k}.$$

With both transition and emission probability matrices at hand, we calculate the likelihood of the full data using the standard forward recursion. To be specific, let us first define for $t = 0, \ldots, T$ the $(K+1) \times (M_t + 1)$ emission probability matrices

$$\boldsymbol{B}_{k,m}^t = \mathbb{P}(m_t = m | U(T_t) = u_k), \; k = 0, \ldots, K, \; m = 0, \ldots, M_t.$$

Denoting $\boldsymbol{m}_{1:t} = (m_1, \ldots, m_t)$, we define the total probability

$$\alpha_k(t) = \mathbb{P}(\boldsymbol{m}_{1:t}, U(T_t) = u_k).$$

This total probability can be determined efficiently with the forward recursion (Murphy 2012, p.609)

$$\alpha_k(t) = \sum_{i=0}^{K} \alpha_i(t-1) \mathbf{P}_{k,i}^{t-1} \boldsymbol{B}_{i,m_t}^t \tag{21}$$

and $\alpha_k(0) = \boldsymbol{B}_{k,m_0}^0$. Then one has

$$\mathbb{P}(\boldsymbol{m}_{1:T} | \boldsymbol{\theta}) = \sum_{k=0}^{K} \alpha_k(T), \tag{22}$$

where we made explicit the dependence of this probability on the parameters

$$\boldsymbol{\theta} = (s, h, N_0, \ldots, N_{T-1}).$$

If we impose priors $\pi(\boldsymbol{\theta})$ on the parameters then we can simulate the posterior probability $\pi(\boldsymbol{\theta} | \boldsymbol{m}_{1:T})$ with the usual MCMC scheme using (22) and the Hastings ratio

$$h(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min \left(\frac{\mathbb{P}(\boldsymbol{m}_{1:T} | \boldsymbol{\theta}') \pi(\boldsymbol{\theta}')}{\mathbb{P}(\boldsymbol{m}_{1:T} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})} \frac{q(\boldsymbol{\theta}' \to \boldsymbol{\theta})}{q(\boldsymbol{\theta} \to \boldsymbol{\theta}')}, 1 \right).$$

Extension of basic model

Sequencing errors Generally, sequencing errors are overcome with sufficient coverage. However, in many applications of next generation sequencing to experimental evolution, the goal of the sequencing is not to infer individual genotypes, but rather allele frequencies directly. Under such a setting, each sequencing read is assumed to be from a different individual. In such cases, sequencing errors may lead to false inference, especially when allele frequencies are very small.

Incorporating sequencing errors into our framework is straight forward. Under the assumption that there are only two alleles present at the locus (achieved by, for instance, pooling all non-selected alleles into one class) and symmetric error rates ϵ between those classes, we can approximate the probability that $m_t^{(i)}$, the i-th allele surveyed at time t, is A in the presence of sequencing errors as

$$\mathbb{P}(m_t^{(i)} = \mathrm{A} | U(T_t) = u_k) = (1 - \epsilon) u_k + \epsilon(1 - u_k).$$

Mutational Input We allow for the production of mutant alleles only when the process is in state $u_0 = 0$ or $u_K = 1$. The production of new alleles proceeds at a rate of $2N\mu dt$. Once a new allele is produced, say when the system is in state u_0, it must get from state $1/2N$ into state u_1. This happens with probability $P_\uparrow = \mathbb{P}(T_{1/2N \to u_1} < T_{1/2N \to 0})$ which is calculated according to (7). This yields the transition rate

$$q_{0,1} = 2N\mu\, \mathbb{P}(T_{1/2N \to u_1} < T_{1/2N \to 0}).$$

Since u_1 is close to 0, we can assume that $\sigma \approx 2sh$ and $M \approx 2N$, see (16). Using (7) and (15) we obtain

$$\begin{aligned} q_{0,1} &= 2N\mu \frac{\phi(0) - \phi(1/2N)}{\phi(0) - \phi(u_1)} \\ &= 2N\mu \frac{1 - (1 + \sigma/2N)^{-M}}{1 - (1 + \sigma u_1)^{-M}} \\ &\approx 2N\mu \frac{1 - (1 + sh/N)^{-2N}}{1 - (1 + 2shu_1)^{-2N}} \approx 2N\mu \frac{1 - \exp(-2sh)}{1 - \exp(-4Nshu_1)}. \end{aligned}$$

For the production of a new allele in state $u_K = 1$ an analogous argument yields the approximations $\sigma \approx s$, $M \approx 4N(1-h)$ and by (18)

$$\begin{aligned} q_{K,K-1} &= 2N\mu\, \mathbb{P}(T_{(2N-1)/2N \to 1} > T_{(2N-1)/2N \to u_{K-1}}) \\ &= \frac{(1 + \sigma(2N-1)/2N)^{-M} - (1+\sigma)^{-M}}{(1 + \sigma u_{K-1})^{-M} - (1+\sigma)^{-M}} \\ &\approx 2N\mu \frac{\exp\left(2s(1-h)\right) - 1}{\exp\left(4Ns(1-h)(1-u_{K-1})\right) - 1}. \end{aligned}$$

In the selection-free case, i.e. in the limit $s \to 0$, the transition probabilities simplify to

$$q_{0,1} = \frac{\mu}{u_1}, \quad q_{K,K-1} = \frac{\mu}{1 - u_{K-1}}.$$

Implementation

We have implemented the proposed model and the Bayesian inference scheme in an easy-to-use `C++` program available on our lab website (http://www.unifr.ch/biology/research/wegmann). While we use standard implementations for most aspects, we note the matrix exponentiation in Eq. 8, which is a numerically very demanding problem. A classic algorithm for matrix exponentiation is by diagonalization of the matrix (Moler and Van Loan 1978). While computationally efficient, this algorithm may be numerically unstable for matrices with large condition numbers, which are typically observed when $\gamma = 2Ns$ becomes large. This was previously observed by (Malaspinas *et al.* 2012), who addressed this issue using multiple precision arithmetics. Unfortunately, such arithmetics are computationally very demanding, leading to slow performance of their implementation.

Here, we propose to alleviate this problem using the approximation

$$\exp \boldsymbol{Q} \approx \left(\boldsymbol{I} + \frac{1}{2^n} \boldsymbol{Q} \right)^{2^n},$$

which can be calculated by successive quadration. Such matrix multiplications are generally demanding, but can be implemented in a computationally efficient manner for generator matrices that are tridiagonal, as each quadration step only adds two additional diagonals and such band matrices can be multiplied efficiently (see chapter 7.4 in (Dahlquist and Björk 2008)).

We further mention the choice of frequency bins. Malaspinas *et al.* (2012) report that for their approach, a tighter spacing of frequencies towards the boundaries led to more accurate results, in particular with what they call a "quadratic grid". We thus chose to implement, apart from a uniformly spaced grid, also a "quadratic grid" with $u_0 = 0$ and

$$u_k = u_{k-1} + x * (1 - x), x = k - \frac{1}{2},$$

scaled such that $u_K = 1$. The major difference of this choice to the quadratic grid proposed by Malaspinas *et al.* (2012) is that we do not force $u_1 = 1/2N$ and $u_{K-1} = 1 - 1/2N$, as this would force us to change the frequency bins as a function of N during the MCMC, and hence to recalculate emission probabilities.

Application to Influenza data

We analyzed allele frequency data from whole genome data sets of Influenza H1N1 obtained in a recent evolutionary experiment (Renzette *et al.* 2014). While we refer the reader to the original study for a detailed description of the experimental set-up, we summarize the key point briefly here: Influenza A/Brisbane/59/2007 (H1N1) was serially amplified on Madin-Darby canine kidney (MDCK) cells for 12 passages of 72 hours each to prevent any freeze-thaw cycles. After the three initial cycles, samples were passed either in the absence of drug, or in the presence of increasing concentrations of oseltamivir, a neuraminidase inhibitor for another 9 passages. At the end of each passage, samples were collected for whole-genome high-throughput population sequencing up to a median coverage of more than 50,000x.

For our analysis here we only considered the time-points taken during drug treatment (passages 4 to 12), but considered all 13,395 sites for which data was available (Foll *et al.* 2014). For each site, we first identified the two alleles having the highest frequencies over all passages and considered the minor allele to be the one with the lower frequency at the beginning of the experiment (passage 0). To avoid any bias, all other alleles were treated collectively as the major allele. We estimated N along with locus specific selection coefficients s, the sequencing error rate ϵ and the per site mutation rate μ. We assumed log-uniform priors on N, ϵ and μ such that $log_{10}(N)$ $U[1,5]$, $log_{10}(\epsilon) = U[-4, -0.3]$ and $log_{10}(\mu) = U[-7, -1]$, and a normal prior on the selection coefficients such that s $\mathcal{N}(0, 0.05)$. Since viruses are haploid, we fixed the dominance coefficient at $h = 0.5$. We then run an MCMC using 51 states for 25000 iterations during which each parameter was updated in turn. The first 2000 such iterations were discarded as burn-in phase.

Simulations To assess the accuracy of our approximation, we simulated trajectories under the discrete Wright-Fisher process, the diffusion process, as well as under the *mean transition time* approximation and the approximation proposed by Malaspinas *et al.* (2012). All simulations under the discrete Wright-Fisher model and the diffusion process were performed using binomial sampling and the Euler–Maruyama method, respectively. Those under approximations using frequency states were generated by simulating transitions according to the transition matrices calculated under the specific approximations.

In order to evaluate the power of our method to infer population sizes and selection coefficients, we also simulated data for 20 or 100 unlinked loci with N of 100, 1000 or 10000. For

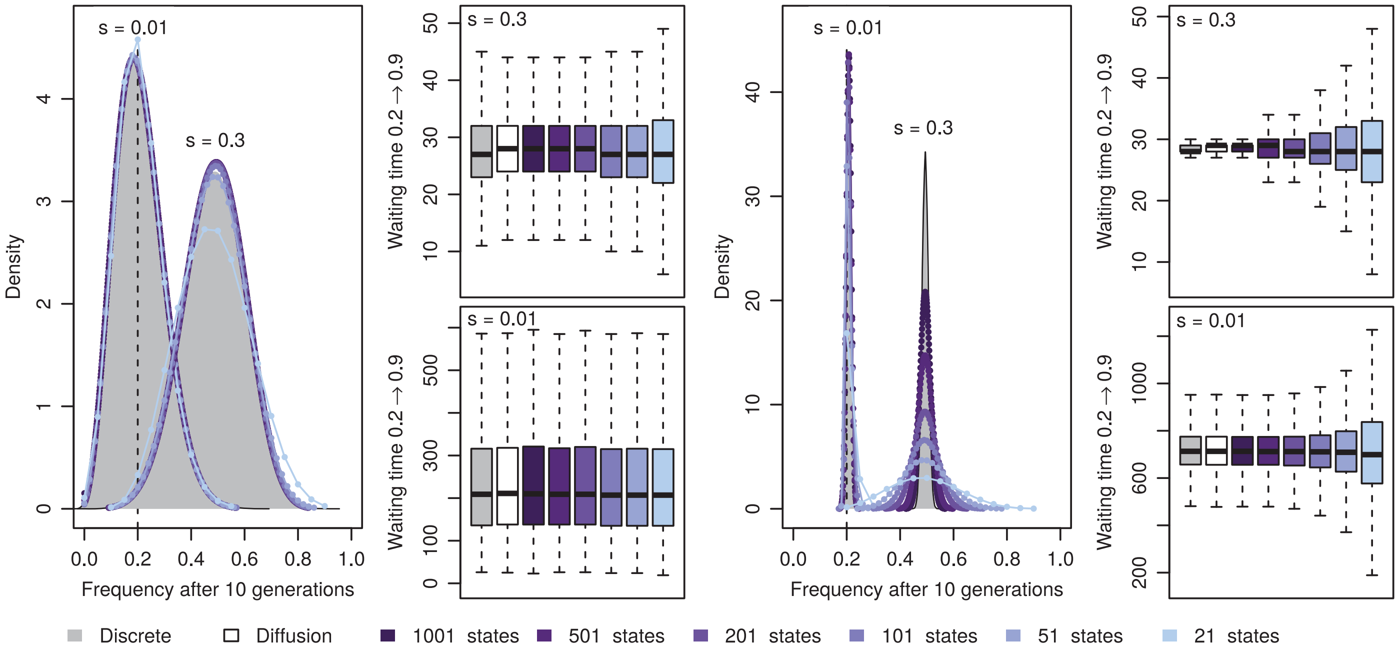

Figure 2 Mean transition time approximation of Markov processes. For both small (N=100, left panel) and large (N=10,000, right panel) population sizes as well as weak (s=0.01) and strong (s=0.3) selection, we show the allele frequency distributions after 10 generations of selection and random drift starting from a frequency of 0.2, as well as the waiting times for a transition from a frequency 0.2 to 0.9. Results obtained under the discrete Wright-Fisher process are given in gray, those obtained under the diffusion approximation as black solid lines or empty boxes, and those under our approximation in shades of blue, with darkness increasing with higher number of frequency states considered.

each of these settings, we set either 20% or 80% of the loci to be under selection, with an equal representation of four selection coefficients: -0.1, -0.01, 0.01 and 0.1. All loci, both selected and neutral, had the starting allele frequency set at random. The change in allele frequency from one time point to the next was calculated under the Wright-Fisher model matching the experimental set-up of our application. Specifically, we simulated a total of 117 generations and took a sample of 1000 sequences every 13 generations, unless otherwise stated.

Results and Discussion

Mean Transition Time Approximation

Comparisons of the long term behavior of the here introduced *mean transition time approximation* of the Wright-Fisher process with its discrete realization demonstrate the power of our approximation. In Fig. 2 we show the frequency distribution of alleles with an initial frequency of 0.2 after 10 generations of selection and random drift under the discrete Wright-Fisher process for different population sizes and different selection strength. As expected from our assumptions, the distributions obtained under our approximation have identical means and show only a slightly increased variance for large selection coefficients and a small number of states. This finding is further strengthened when comparing this distribution over larger time scales up to 1000 generations (Fig. 3), which also illustrates that our approximation leads to accurate loss / fixation ratios. In the situations studied here, all loci correctly fix in the case of strong selection or when N is large. In the case of $N = 100$ and $s = 0.01$, however, we estimate that 62.00% or 61.99% of all loci will be lost when using 1001 or 21 states, respectively. This is very similar to the proportion of 62.02% obtained among $3 \cdot 10^5$ simulations with the diffusion approximated here.

A more direct illustration of our assumption is the comparison of the distribution of waiting times for a specific transition. As shown in Fig. 2, our approximation indeed captures the mean transition time perfectly, while again exhibiting an increased variance for large selection coefficients and small number of states. Based on these results, and in order to keep the computational effort minimal, we will use 51 states for all our inference shown below.

Choice of grid

All results shown above were obtained with a uniform grid of frequency states. Following Malaspinas *et al.* (2012), we also implemented a quadratic grid, but we found the choice of grids not to affect our approximation noticeably. In general, we found the quadratic grid to describe probabilities close to boundaries more accurately, but to be less accurate for intermediate frequencies than a uniform grid. These differences, however, are small, do not inflate with increasing number of generations, and are only visible for very low number of states (Supplementary Fig. S1). This suggests that our approximation is very robust to the choice of the grid, and that the differences observed are due to the resolution in characterizing the probability distribution at different frequencies rather than an effect of the approximation itself. We will thus continue using a uniform grid in the following.

Comparison with related methods

Recently, Lacerda and Seoighe (2014) proposed to approximate the probability distribution of allele frequencies after t generations by a Gaussian distribution, the mean and variance of which can be obtained iteratively using the delta method. Their approach can easily be applied to the diffusion process studied

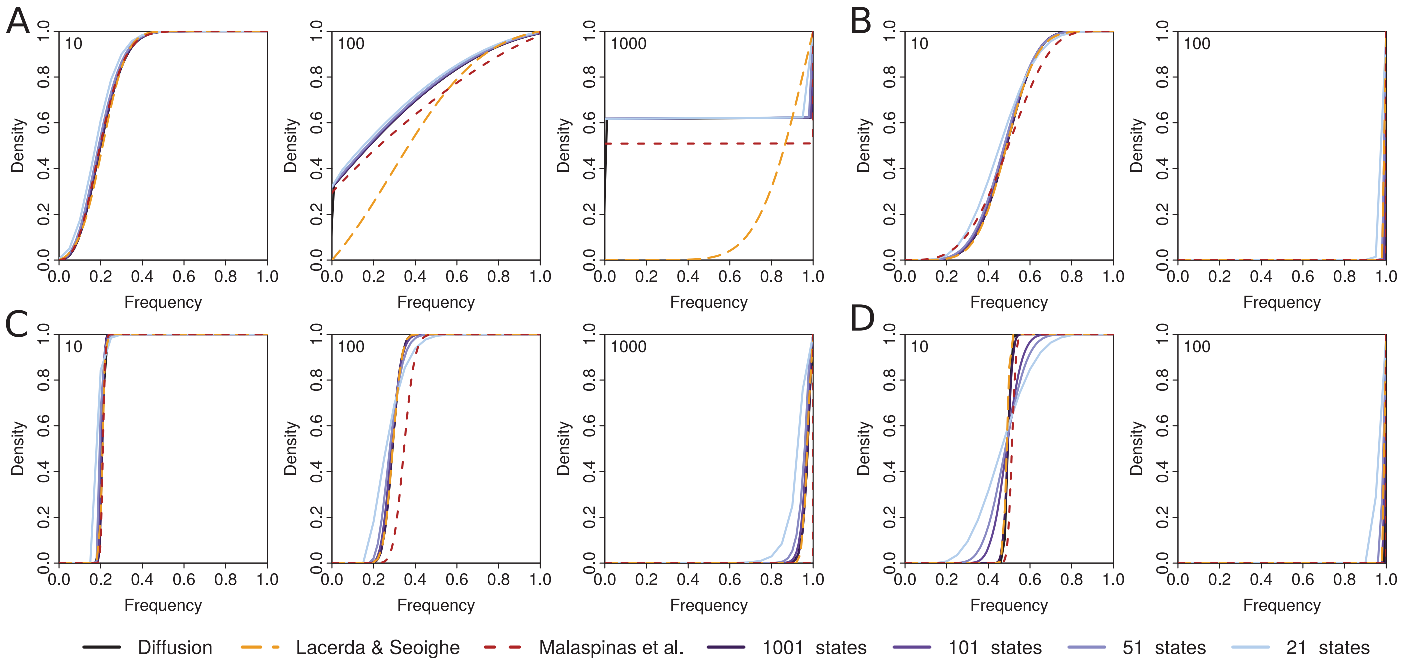

Figure 3 Comparison of different approximations. Shown are the cumulative probability density distributions (CDF) of allele frequencies after 10, 100 and 1000 generations (shown in top left corner) of selection and random drift starting from a frequency of 0.2 and obtained under the Wright-Fisher diffusion (black) and three approximations of it: the approximations introduced by Lacerda and Seoighe (2014, orange) and Malaspinas *et al.* (2012, red), and the *mean transition time* approximation introduced here (shades of blue for different number of frequency states). Results are shown for small ($N = 100$, A and B) and large ($N = 10,000$, C and D) population sizes and weak ($s = 0.01$, A and C) and strong ($s = 0.3$, B and D) section. For the approximation by Malaspinas *et al.* (2012), we used a quadratic grid as originally proposed with the minimum number of states mathematically possible, but at least 101 states (101, 101, 326 and 5813 states for A, B, C and D, respectively). The distributions for the diffusion approximation were obtained from $3 \cdot 10^5$ simulations using kernel density estimation.

here (see Appendix). As shown in Fig. 3, the approximation obtained via the delta method and our approximation are very similar over a large range of the parameter space and also agree well with the diffusion process they both approximate. However, due to the assumption of a Gaussian distribution, the approximation obtained with the delta method is less accurate than our approach in describing allele frequencies close to boundaries. This is particularly true when selection is weak enough such that the probability of fixation is < 1.0, which results in a bi-modal distribution (Fig. 3A).

A major advantage of the the delta approach, however, is its computational speed, which does not depend on the population size nor the selection strength. Our method is generally much more demanding due to its reliance on matrix calculations rather than simple recursions. But the benefit of our approach lies in the discretization of allele frequencies, without which any inference from time-series data is computationally impossible whenever N is large.

In this regard, our method is closer to that introduced by Malaspinas *et al.* (2012) that also uses a grid of discretized allele frequencies. In contrast to our method, however, their approach approximates the mean and variance of the infinitesimal transition probabilities, rather than that of the resulting waiting times. While Malaspinas *et al.* (2012) derive their approximation for the classic diffusion, it is straight forward to generalize their approach and apply it to the diffusion studied here (see Appendix). As shown in Fig. 3, their approximation holds generally well of most of the range tested, but allele frequencies appear to raise slightly too fast. More importantly, the approximation introduced by Malaspinas *et al.* (2012) requires substantially more states than our approximation due to the mathematical nature of the approximation. For the case of $N = 10,000$ and $s = 0.3$ shown in Fig. 3D, for instance, a minimum of 5,813 thousand states are required. In contrast, our approximation is computationally stable even with just a handful of states and thus allows to balance accuracy and computation effort regardless of N or s. This difference between the two approaches easily translates into a reduction in computation time of several orders of magnitude when attempting to infer parameters using a Hidden Markov Model (HMM), and essentially rendering such an analysis unfeasible for large $\gamma = 2Ns$ with the approximation introduced by Malaspinas *et al.* (2012), as has been reported recently Foll *et al.* (2015).

Power to infer population sizes

While allele trajectories are affected by both selection and drift, we aim here to disentangle these effects by integrating information from multiple loci. We first assessed the power to infer population sizes N accurately under ideal conditions, that is, for 100 unlinked loci in the absence of selection. In Fig. 4 we show the likelihood surfaces for N obtained with different number of states, for data simulated under different population sizes. While this analysis suggest high power to infer small population sizes accurately, it highlights the general issue of inferring large population sizes from changes in allele frequencies, accentuated when fewer states are used. The issue arises from the fact that

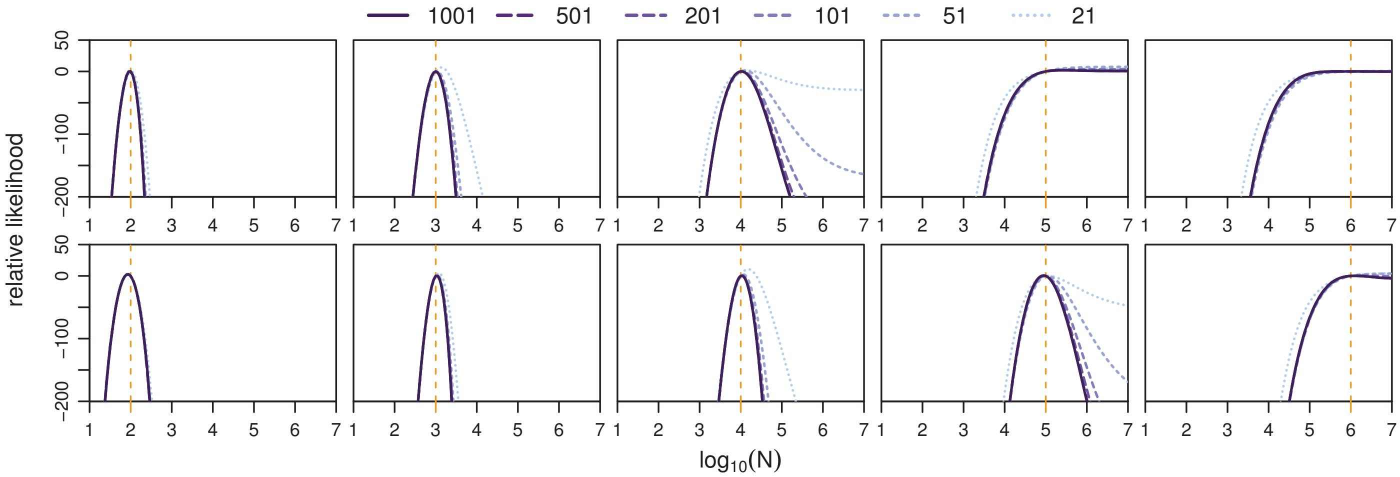

Figure 4 Power to infer population sizes. Shown are the relative likelihood surfaces obtained via our *mean transition time approximation* for a particular simulation of 100 neutral loci for different population sizes (vertical dashed lines) and different number of frequency states considered (see legend). The top row is for the case of 13 generations between time points, the bottom row for the case of 130 generations between time points.

in large populations and over the short time course of evolutionary experiments in general, the changes in allele frequencies between time points are so small, that they are compatible with almost arbitrarily large populations. While using fewer frequency states further decreases the resolution of detectable allele frequency changes, we note that this issue is more general and expected to affect all methods for inferring population sizes from such data, particularly when a small number of samples is used. The best way to overcome it is to observe changes in allele frequencies over larger intervals. Indeed, when taking samples every 130 generations instead of every 13, population sizes up to N=100,000 can be estimated accurately (Fig. 4, bottom row).

Power to infer selection

To assess the power of our framework to infer locus-specific selection coefficients, we simulated 100 unlinked loci, of which 20% experienced selection at various strengths. As shown in Fig. 5, both the population size as well as the strength of selection affects the power of this inference. For medium to large population sizes, our method infers even small selection coefficients with high accuracy. When the population size is small, however, inference of selection proves more difficult (Fig. 5). While this is generally expected due to the much larger effect of drift in small populations ($Ns = 10$ for the strongly selected alleles), it is accentuated here by our choice to simulate initial frequencies at random. Indeed, when given ideal starting frequencies (0.1 for positively and 0.9 for negatively selected alleles), our method identifies strongly selected alleles accurately even in small populations (Supplementary Fig. S2).

Remarkably, we found the power to infer population sizes as well as locus-specific selection coefficients not to be negatively affected under pervasive selection. This is illustrated by comparing the posterior distributions obtained from simulations where 80% of all loci were targeted by selection (Supplementary Fig. S3) to those shown here where only 20% were affected by selection (Fig. 5). More direct evidence is given in Table 1, where we report the posterior probability for $s > 0.0$ for different combinations of population sizes and selection coefficients and actually find higher power to identify selected loci in the case of pervasive selection than when only 20% of all loci were simulated under selection.

For computational efficiency, all results shown here were obtained using 51 states. However, we note a trade-off between power of inference and computational costs. As shown in Supplementary Fig. S4, using very few states (21) may lead to slightly broader posteriors and a small bias towards weaker values of s. Both effects are already largely overcome when using 51 states for most loci, but small improvements are still detectable with more states (Supplementary Fig. S4).

Application to Influenza data

We next applied our approach to publicly available sequencing data of Influenza H1N1 segment 6, obtained at multiple time points throughout an evolutionary experiment in which the virus was exposed to an antiviral drug (oseltamivir) (Renzette *et al.* 2014). While allele frequencies are generally estimated with high accuracy due to the very high coverage in this experiment (about 50,000x), sequencing error may contribute substantially to the observed low frequency variants. In addition, many of the observed mutations likely entered the population only during the experiment, but their exact time of origin is blurred by both the sequencing error as well as sampling. We thus extended our framework to estimate the mutation rate as well as the overall sequencing error rate jointly with the demographic and selection parameters.

We applied our extended method to each of the 8 segments of the Influenza genome individually, but obtained highly concordant results among all segments. As shown in Fig. 6, we infer the effective population size during the experiment to be around 7,000, a mutation rate of about 10^{-5} and a sequencing error rate of about $10^{-3.8}$. While our estimates of the mutation and error rates are consistent with published mutation rates for Influenza (Nobusawa and Sato 2006) and RNA viruses in general (Drake *et al.* 1998) and also with the employed quality filters on sequencing reads (Foll *et al.* 2014), our estimate of the population size is substantially larger than previous estimates of about 225 (Foll *et al.* 2014). While we found our approach to slightly overestimate larger population sizes under the spacing of time points relevant here, there are several arguments supporting a larger population size. First, the original estimates were obtained un-

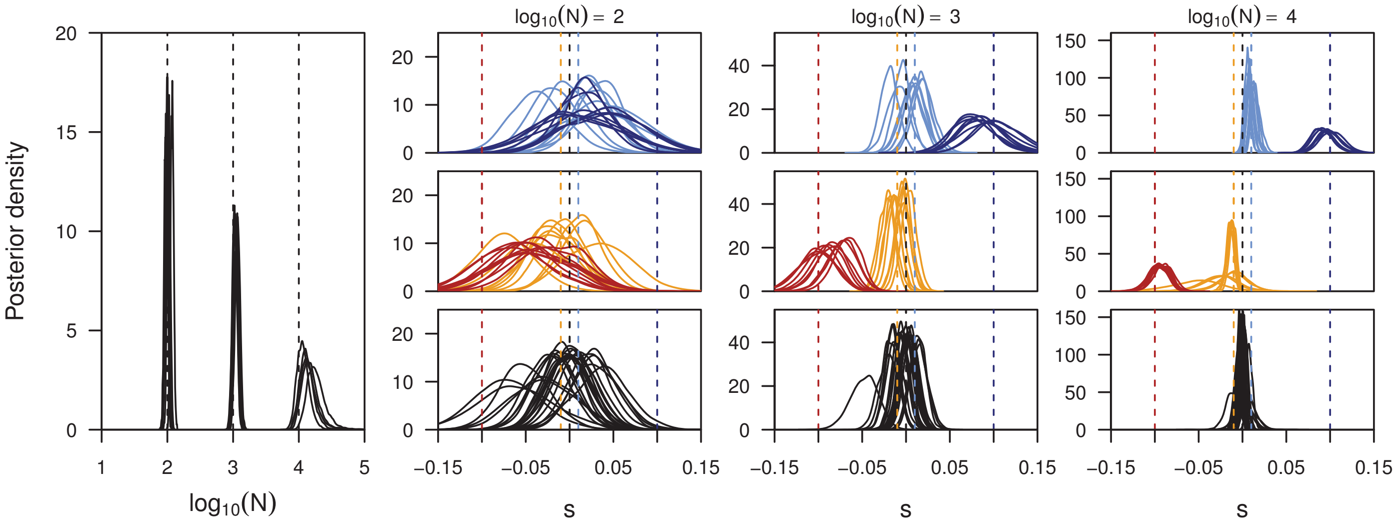

Figure 5 Power to infer selection and population size jointly. Here we show the posterior distributions on the population size (first panel) and locus-specific selection coefficients obtained for five replicate simulations for each of three different population sizes. For each replicate we plot the posteriors of all loci simulated under positive selection (blue shades, top row) and under negative selection (red shades, middle row), as well as five neutral loci picked at random (black, bottom row). In all simulations, starting frequencies were chosen randomly for each locus.

Table 1 Power to identify loci under selection We report the average and standard deviation (in parenthesis) of the posterior probability $\mathbb{P}(s > 0.0)$ obtained under various population sizes and for the cases of 20% and 80% of all loci simulated under selection.

fraction selected	$\log_{10}(N)$	s=-0.1	s=-0.01	s=0.0	s=0.01	s=0.1
0.2	2	0.27 (0.15)	0.42 (0.24)	0.50 (0.26)	0.56 (0.26)	0.78 (0.18)
0.2	3	0.00 (0.05)	0.24 (0.24)	0.50 (0.30)	0.74 (0.26)	1.00 (0.00)
0.2	4	0.00 (0.00)	0.06 (0.16)	0.50 (0.32)	0.92 (0.17)	1.00 (0.00)
0.8	2	0.12 (0.11)	0.09 (0.14)	0.50 (0.28)	0.72 (0.15)	0.89 (0.15)
0.8	3	0.00 (0.00)	0.02 (0.04)	0.50 (0.38)	1.00 (0.00)	0.99 (0.02)
0.8	4	0.00 (0.00)	0.00 (0.00)	0.50 (0.45)	1.00 (0.00)	1.00 (0.00)

der the assumption of neutrality at all loci, while our approach infers N jointly with selection. Second, the previous estimates were obtained from a small subset of the data, namely the 147 loci with an observed allele frequency $\leq$ 1% after down sampling to 1,000 reads per locus at no less than three time points. In contrast, our inference is based on the raw data at the complete set of 13,395 loci, including those with small frequencies particularly informative about drift. Third, the original inference accounted for neither sequencing errors nor mutations. In summary, our results argue for a much larger effective population size than previously reported.

Our results on selection, on the other hand, are highly concordant with previous estimates. In Fig. 6 we report the posterior distributions on the locus-specific selection coefficients for all polymorphic sites for each of the 8 segments of the Influenza genome. As expected, most mutations were found to be selectively neutral or under slight purifying selection (observe the slight asymmetry towards negative selection coefficients for many loci). For a few mutations, however, we found compelling evidence for them to be the target of positive selection (99% credible interval does not include 0). On segment NA, there were three such mutations, of which two stand out with an estimated selection coefficient around 0.2. One of these mutations (Y274H) occurred at a locus at which resistance to oseltamivir has been previously described (Collins *et al.* 2008). Many additional mutations were found to be the target of selection through out the genome, with many of those likely under negative selection. These are mutations that were found at elevated frequencies at the beginning of the experiment, yet at much lower frequencies after a few passages. The complete list of all mutations found to be under selection is given in Supplementary Table S1.

Conclusion

Here we present a novel, discrete approximation for diffusion processes. This approximation, which we term *mean transition time approximation*, is designed to preserve the long term behavior of the continuous process it approximates, which renders it particularly suitable to study time series data. Here we derived this approximation for the particular case of inferring selection and demography from such time series data under the classic Wright-Fisher model. As shown through extensive simulations, our approximation is well suited to describe allele trajectories through time, even when only a few states are used. This allowed us to develop a Bayesian inference approach to jointly

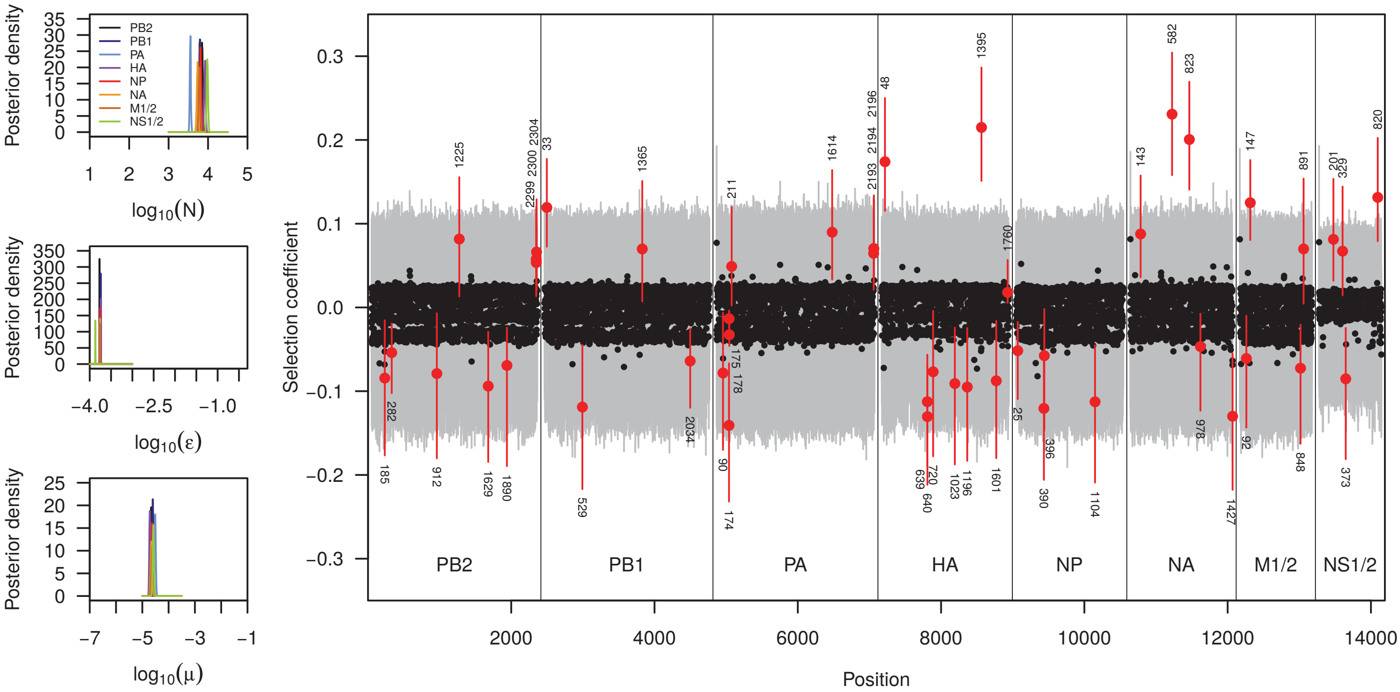

Figure 6 Evolution of drug-resistance in Influenza. Here we show the posterior distributions on the population size ($\log_{10}(N)$), sequencing error rate (ϵ), mutation rate (μ) and locus specific selection coefficients s_l estimated independently for each of the six segments of the Influenza genome. For the selection coefficients, black dots represent posterior medians and gray lines indicate the 99% credible intervals. Loci for which the 99% credible interval does not include $s = 0.0$ are shown in red and their actual position within the segment is printed.

infer the population size and locus-specific selection coefficients with high accuracy. We further extended this model to estimate the average sequencing error rate, as well as the per generation mutation rate. The approach is further readily applicable to models of instantaneous population size changes. We finally applied our approach to data from a recent experiment on the evolution of drug resistance in Influenza virus, identifying likely targets of selection and finding evidence for much larger viral population sizes than previously reported.

Acknowledgments

We thank Nicolas Renzette for his advice on how to identify the protein changes corresponding to individual mutations. We are further grateful to two anonymous reviewers for their very constructive comments on an earlier version of this work. This study was supported by Swiss National Foundation grants number PZ00P3_142643 and 31003A_149920 to DW, and grants from the Swiss National Science Foundation and a European Research Council Starting Grant to JDJ.

Literature Cited

Andolfatto, P. and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156**: 257–268.

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

Collins, P. J., L. F. Haire, Y. P. Lin, J. Liu, R. J. Russell, P. a. Walker, J. J. Skehel, S. R. Martin, A. J. Hay, and S. J. Gamblin, 2008 Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature **453**: 1258–61.

Dahlquist, G. and Å. Björk, 2008 *Numerical Methods in Scientific Computing*. Number v. 1, Society for Industrial & Applied Mathematics.

Drake, J., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148**: 1167–1186.

Durrett, R., 2008 *Probability models for DNA sequence evolution*. Springer Science & Business Media.

Etheridge, A., 2011 *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 39. Springer Science & Business Media.

Ewens, W. J., 2004 *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media.

Foll, M., Y.-P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank, H. Shim, A.-S. Malaspinas, G. Ewing, P. Liu, D. Wegmann, D. R. Caffrey, K. B. Zeldovich, D. N. Bolon, J. P. Wang, T. F. Kowalik, C. a. Schiffer, R. W. Finberg, and J. D. Jensen, 2014 Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS genetics **10**: e1004185.

Foll, M., H. Shim, and J. D. Jensen, 2015 WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Molecular ecology resources pp. 87–89.

Lacerda, M. and C. Seoighe, 2014 Population Genetics Inference for Longitudinally-Sampled Mutants Under Strong Selection. Genetics **198**: 1237–1250.

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating Allele Age and Selection Coefficient from Time-Serial Data. Genetics **192**: 599–607.

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

Moler, C. and C. Van Loan, 1978 Nineteen dubious ways to compute the exponential of a matrix. SIAM review **20**: 801–836.

Murphy, K. P., 2012 *Machine learning: a probabilistic perspective*. MIT press.

Nielsen, R., 2005 Molecular signatures of natural selection. Annual review of genetics **39**: 197–218.

Nobusawa, E. and K. Sato, 2006 Comparison of the Mutation Rates of Human Influenza A and B Viruses **80**: 3675–3678.

Press, W. H., 2007 *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Renzette, N., D. R. Caffrey, K. B. Zeldovich, P. Liu, G. R. Gallagher, D. Aiello, A. J. Porter, E. a. Kurt-Jones, D. N. Bolon, Y.-P. Poh, J. D. Jensen, C. a. Schiffer, T. F. Kowalik, R. W. Finberg, and J. P. Wang, 2014 Evolution of the influenza A virus genome during development of oseltamivir resistance in vitro. Journal of virology **88**: 272–81.

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

Sverrisdóttir, O. O., A. Timpson, J. Toombs, C. Lecoeur, P. Froguel, J. M. Carretero, J. L. Arsuaga Ferreras, A. Götherström, and M. G. Thomas, 2014 Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. Molecular biology and evolution **31**: 975–83.

Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, S. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw, 1995 Viral dynamics in human immunodeficiency virus type 1 infection. Nature **373**: 117–122.

Wilde, S., A. Timpson, K. Kirsanow, E. Kaiser, M. Kayser, M. Unterländer, N. Hollfelder, I. D. Potekhina, W. Schier, M. G. Thomas, and J. Burger, 2014 Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proceedings of the National Academy of Sciences of the United States of America **111**: 4832–7.

Williamson, E. and M. Slatkin, 1999 Using maximum likelihood to estimate population size from temporal changes in allele frequencies. Genetics **152**: 755–761.

Appendix

Approximation for large $\gamma = 2Ns$

If $\gamma = 2Ns$ is large, we get approximations for the Green's function which allow for analytic expressions of the integrals. More precisely, assume that $M_k\sigma_k = 4N\bar{\sigma}_k$ is large. We can then neglect the minus one terms in the numerator and denominator of (18) and we get the approximation

$$\begin{aligned} P_\downarrow &\approx \left(\frac{1+\sigma_k u_{k-1}}{1+\sigma_k u_k}\right)^{M_k} = \left(1 - \frac{M_k\sigma_k(u_k-u_{k-1})/(1+\sigma_k u_k)}{M_k}\right)^{M_k} \\ &\approx \exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_k}(u_k-u_{k-1})\right), \end{aligned} \tag{23}$$

which will be very small for large γ. The probability for exit at the upper state is $P_\uparrow \approx 1$. Inserting the first approximating expression or $P_\downarrow$ into (19) and using $4N/M_k \approx \sigma_k/\bar{\sigma}_k$, we get

$$\begin{aligned} G_\downarrow(u_k, y) &\approx \frac{1+\sigma_k y}{\bar{\sigma}_k y(1-y)}\left(\left(\frac{1+\sigma_k y}{1+\sigma_k u_k}\right)^{M_k} - P_\downarrow\right) \\ &\approx \frac{1+\sigma_k y}{\bar{\sigma}_k y(1-y)} \exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_k}(u_k-y)\right). \end{aligned} \tag{24}$$

The exponential term is dominant for y close to u_k. In the integral we can thus keep the factor of the exponential constant at $y = u_k$ since it does not vary much when y is close to u_k:

$$\begin{aligned} \mathcal{E}_\downarrow &= \int_{u_{k-1}}^{u_k} G_\downarrow(u_k, y)dy \\ &\approx \frac{1+\sigma_k u_k}{\bar{\sigma}_k u_k(1-u_k)} \int_{u_{k-1}}^{u_k} \exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_k}(u_k-y)\right)dy \\ &\approx \frac{(1+\sigma_k u_k)^2}{M_k\sigma_k\bar{\sigma}_k u_k(1-u_k)}\left(1-\exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_k}(u_k-u_{k-1})\right)\right) \\ &\approx \frac{(1+\sigma_k u_k)^2}{M_k\sigma_k\bar{\sigma}_k u_k(1-u_k)}. \end{aligned} \tag{25}$$

From (20) we get the approximation

$$G_\uparrow(u_k, y) \approx \frac{1+\sigma_k y}{\bar{\sigma}_k y(1-y)}\left(1-\exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_{k+1}}(u_{k+1}-y)\right)\right).$$

To get $\mathcal{E}_\uparrow$ we integrate this approximate expression. Observe that the exponential term becomes important only when y gets close to u_{k+1}. For this reason we can safely keep the factor in front of the exponential term constant when integrating the second term:

$$\begin{aligned} \mathcal{E}_\uparrow &= \int_{u_k}^{u_{k+1}} G_\uparrow(u_k, y)dy \\ &\approx \int_{u_k}^{u_{k+1}} \frac{1+\sigma_k y}{\bar{\sigma}_k y(1-y)}dy - \frac{1+\sigma_k u_{k+1}}{\bar{\sigma}_k u_{k+1}(1-u_{k+1})}\int_{u_k}^{u_{k+1}} e^{-\frac{M_k\sigma_k(u_{k+1}-y)}{1+\sigma_k u_{k+1}}}dy \\ &= \frac{1}{\bar{\sigma}_k}\left(\log\frac{u_{k+1}}{u_k} - (1+\sigma_k)\log\frac{1-u_{k+1}}{1-u_k}\right) - \frac{(1+\sigma_k u_{k+1})^2}{M_k\sigma_k\bar{\sigma}_k u_{k+1}(1-u_{k+1})}\left(1-\exp\left(-\frac{M_k\sigma_k}{1+\sigma_k u_{k+1}}(u_{k+1}-u_k)\right)\right) \\ &\approx \frac{1}{\bar{\sigma}_k}\left(\log\frac{u_{k+1}}{u_k} - (1+\sigma_k)\log\frac{1-u_{k+1}}{1-u_k} - \frac{(1+\sigma_k u_{k+1})^2}{M_k\sigma_k u_{k+1}(1-u_{k+1})}\right). \end{aligned} \tag{26}$$

Numerical experiments indicate that the approximate formulae (25) and (26) are adequate when the conditions

$$4Nsh(u_{k+1}-u_{k-1}) > 10 \quad \text{and} \quad 4Ns(1-h)(u_{k+1}-u_{k-1}) > 10 \tag{27}$$

are met. In that case we set $q_{k,k-1} = 0$ and

$$q_{k,k+1} = \frac{1}{\mathcal{E}_\downarrow + \mathcal{E}_\uparrow}.$$

Note that Formula (26) gets singular for $k = K-1$ since in that case $1 - u_{k+1} = 0$. Using the substitution $z = 1-y$, we get for that case from (20) the approximation

$$\begin{aligned} \mathcal{E}_\uparrow &\approx \frac{1}{\bar{\sigma}_{K-1}}\int_0^{1-u_{K-1}} \frac{1+\sigma_{K-1}(1-z)}{z(1-z)}\left(1-\exp\left(-\frac{M_{K-1}\sigma_{K-1}}{1+\sigma_{K-1}}z\right)\right)dz \\ &\approx \frac{1+s}{s(1-h)}\int_0^{1-u_{K-1}}\left(1-\exp\left(-\frac{4Ns(1-h)}{1+s}z\right)\right)\frac{dz}{z}. \end{aligned}$$

The last integral can be written as an exponential integral

$$\mathrm{Ei}(x) = \int_0^x \frac{1-e^{-t}}{t} dt$$

in the form

$$\mathcal{E}_\uparrow \approx \frac{1+s}{s(1-h)} \mathrm{Ei}\left[\frac{4Ns(1-h)}{1+s}(1-u_{K-1})\right].$$

Using the approximation

$$\mathrm{Ei}(x) \approx \log(x) + 0.577\ldots$$

where 0.577 . . . is the Euler-Mascheroni constant, we finally arrive at

$$\mathcal{E}_\uparrow \approx \frac{1+s}{s(1-h)} \left(\log\left(\frac{4Ns(1-h)}{1+s}(1-u_{K-1})\right) + 0.577\ldots\right). \tag{28}$$

Similarly, the case $k = 1$ deserves special attention because the denominator of (24) gets singular at $y = 0$. Since u_1 is small and y even smaller, we can set $\sigma_1 = 2sh$ and $M_1 = 2N$. From (19) we then get the approximations

$$\begin{aligned} G_\downarrow(u_1, y) &= \frac{4NP_\downarrow}{\sigma_1 M_1} \frac{1+\sigma_1 y}{y(1-y)} \left((1+\sigma_1 y)^{M_1} - 1\right) \\ &\approx \frac{P_\downarrow}{shy(1-y)} \left((1+2shy)^{2N} - 1\right) \\ &\approx \frac{P_\downarrow}{shy} \left(e^{4Nshy} - 1\right). \end{aligned}$$

From this we obtain for the downward mean transition time

$$\begin{aligned} \mathcal{E}_\downarrow &= \int_0^{u_1} G_\downarrow(u_1, y) dy \approx \frac{P_\downarrow}{sh} \int_0^{u_1} \frac{e^{4Nshy} - 1}{y} dy \\ &= \frac{P_\downarrow}{sh} \int_0^{4Nshu_1} \frac{e^t - 1}{t} \approx \frac{P_\downarrow}{sh} \cdot \frac{e^{4Nshu_1}}{4Nshu_1} \end{aligned}$$

because the integrand is very dominant at the upper integration limit. From (23) we get the approximation $P_\downarrow \approx e^{-4Nshu_1}$ and thus

$$\mathcal{E}_\downarrow \approx \frac{1}{4Ns^2h^2u_1}. \tag{29}$$

The Wright-Fisher process in the absence of selection

In the absence of selection ($s = 0$), the expressions for the generator matrix can be explicitly evaluated since $b(x) = 0$ (see Eq. 11). We have $\phi(x) = x$ and $m(y) = 2N/x(1-x)$. From this we get

$$P_\downarrow = \frac{u_{k+1} - u_k}{u_{k+1} - u_{k-1}}, \quad P_\uparrow = \frac{u_k - u_{k-1}}{u_{k+1} - u_{k-1}}. \tag{30}$$

The two parts of the Green's function are given by

$$G_\downarrow(u_k, y) = 4NP_\downarrow \left(\frac{1-u_{k-1}}{1-y} - \frac{u_{k-1}}{y}\right)$$

and

$$G_\uparrow(u_k, y) = 4NP_\uparrow \left(\frac{u_{k+1}}{y} - \frac{1-u_{k+1}}{1-y}\right).$$

These integrate to

$$\mathcal{E}_\downarrow = 4NP_\downarrow \left(u_{k-1} \log\frac{u_{k-1}}{u_k} + (1-u_{k-1}) \log\frac{1-u_{k-1}}{1-u_k}\right) \tag{31}$$

and

$$\mathcal{E}_\uparrow = 4NP_\uparrow \left(u_{k+1} \log\frac{u_{k+1}}{u_k} + (1-u_{k+1}) \log\frac{1-u_{k+1}}{1-u_k}\right). \tag{32}$$

As above we determine the transition rates by

$$q_{k,k-1} = \frac{P_\downarrow}{\mathcal{E}_\downarrow + \mathcal{E}_\uparrow}, \quad q_{k,k+1} = \frac{P_\uparrow}{\mathcal{E}_\downarrow + \mathcal{E}_\uparrow}.$$

Approximations via the delta method

Following the argument of (Lacerda and Seoighe 2014), an approximate solution to the diffusion equation can be obtained by the delta method. While their original formulation applies to the discrete Wright-Fisher process, the argument works as well for the diffusion process studied here.

As above (e.q. 1), $X(t)$ is a diffusion process on the state space $[0, 1]$ with infinitesimal generator

$$Lf = \frac{1}{2}a(x)\frac{d^2}{dx^2}f + b(x)\frac{d}{dx}f. \tag{33}$$

Recall that the infinitesimal moments of the diffusion process are given by

$$\begin{aligned} \mathbb{E}(dX(t)|X(t) = x) &= b(x)dt + o(dt), \\ \mathrm{var}(dX(t)|X(t) = x) &= a(x)dt + o(dt). \end{aligned}$$

The mean $\mu(t)$ of the process can be approximated iteratively as follows:

$$\begin{aligned} \mu(t + dt) &= \mathbb{E}\left[X(t + dt)\right] = \mathbb{E}\left[\mathbb{E}(X(t + dt)|X(t))\right] \\ &= \mathbb{E}\left[\mathbb{E}(X(t) + dX(t)|X(t))\right] = \mathbb{E}(X(t)) + \mathbb{E}[b(X(t)dt)] \\ &\approx \mu(t) + b(\mu(t))dt. \end{aligned}$$

In the last step, we used the delta approximation $\mathbb{E}(f(X)) \approx f(\mathbb{E}X)$. Similarly, we apply the delta approximation $\mathrm{var}(f(X)) \approx |f'(\mathbb{E}X)|^2\mathrm{var}(X)$ to get an iterative approximation for the variation:

$$\begin{aligned} \sigma^2(t + dt) &= \mathrm{var}(X(t + dt)) \\ &= \mathbb{E}\left[\mathrm{var}\left(X(t) + dX(t)|X(t)\right)\right] + \mathrm{var}\left[\mathbb{E}\left(X(t) + dX(t)|X(t)\right)\right] \\ &= \mathbb{E}\left[\mathrm{var}\left(dX(t)|X(t)\right)\right] + \mathrm{var}\left[X(t) + b(X(t))dt\right] \\ &\approx \mathbb{E}\left[a(X(t))dt\right] + \left(1 + b'(\mathbb{E}X(t))dt\right)^2 \mathrm{var}(X(t)) \\ &\approx a(\mu(t))dt + \left(1 + b'(\mu(t))dt\right)^2 \sigma^2(t). \end{aligned}$$

For the case of $h = \frac{1}{2}$ and by inserting (9), one gets in particular

$$\begin{aligned} \mu(t + dt) &= \mu(t) + \frac{s\mu(t)\left(1 - \mu(t)\right)}{2\left(1 + s\mu(t)\right)}dt, \\ \sigma^2(t + dt) &\approx \frac{\mu(t)\left(1 - \mu(t)\right)}{2N}dt + \left(1 + \frac{s - 2s\mu(t) - s^2\mu^2(t)}{2(1 + s\mu(t))^2}dt\right)^2 \sigma^2(t). \end{aligned}$$

Approximations as proposed by Malaspinas et al.

As in (Malaspinas *et al.* 2012), we construct the Markov chain $U(t)$ with states $u_0 = 0 < u_1 < \ldots u_K = 1$ by matching the infinitesimal mean and infinitesimal variance of U and X. This allows to determine the generator matrix $\boldsymbol{Q}$. Here we generalize their notation for any diffusion process

$$Lf = \frac{1}{2}a(x)\frac{d^2}{dx^2}f + b(x)\frac{d}{dx}f.$$

Formulas (8) and (9) from (Malaspinas *et al.* 2012) we then get

$$\begin{aligned} q_{i,i+1}(u_{i+1} - u_i) - q_{i-1,i}(u_i - u_{i-1}) &= b(u_i), \\ q_{i,i+1}(u_{i+1} - u_i)^2 + q_{i-1,i}(u_i - u_{i-1})^2 &= a(u_i). \end{aligned}$$

These can be solved for the infinitesimal generators:

$$\begin{aligned} q_{i,i+1} &= \frac{a(u_i) + b(u_i)\Delta_{i-1}}{\Delta_i^2 + \Delta_i\Delta_{i-1}}, \\ q_{i-1,i} &= \frac{a(u_i) - b(u_i)\Delta_i}{\Delta_{i-1}^2 + \Delta_i\Delta_{i-1}}, \end{aligned}$$

where we used the abbreviation $\Delta_i = u_{k+i} - u_i$. To apply these general formulas to the particular diffusion studied here we simply use $a(x)$ and $b(x)$ as given in eq. 10 and 11, respectively.