

## Accepted Manuscript

Inference in Population Genetics Using Forward and Backward,  
Discrete and Continuous Time Processes

Juraj Bergman, Dominik Schrempf, Carolin Kosiol, Claus Vogl

PII: S0022-5193(17)30547-7  
DOI: [10.1016/j.jtbi.2017.12.008](https://doi.org/10.1016/j.jtbi.2017.12.008)  
Reference: YJTBI 9290



To appear in: *Journal of Theoretical Biology*

Received date: 1 June 2017  
Revised date: 23 November 2017  
Accepted date: 8 December 2017

Please cite this article as: Juraj Bergman, Dominik Schrempf, Carolin Kosiol, Claus Vogl, Inference in Population Genetics Using Forward and Backward, Discrete and Continuous Time Processes, *Journal of Theoretical Biology* (2017), doi: [10.1016/j.jtbi.2017.12.008](https://doi.org/10.1016/j.jtbi.2017.12.008)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Highlights**

- Inference of population genetic parameters from a sample of sequences represented as site frequency spectra (SFS), using concepts akin to the forward-backward algorithm of hidden Markov models is described.
- Discrete transition matrices and continuous diffusion models of iterating the population allelic proportion, forward and backward in time, are used for calculating the marginal likelihood of the data for maximum likelihood inference of parameters.
- The method is demonstrated for simulated joint site frequency spectra (i.e., data from two or more populations) under different models of mutation and for different demographic scenarios.

# Inference in Population Genetics Using Forward and Backward, Discrete and Continuous Time Processes

Juraj Bergman<sup>a,b</sup>, Dominik Schrempf<sup>a,b</sup>, Carolin Kosiol<sup>a,d</sup>, Claus Vogl<sup>c,\*</sup>

<sup>a</sup>*Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*

<sup>b</sup>*Vienna Graduate School of Population Genetics, A-1210 Wien, Austria*

<sup>c</sup>*Institut für Tierzucht und Genetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*

<sup>d</sup>*Centre of Biological Diversity, School of Biology, University of St. Andrews, St Andrews KY16 9TH, UK*

---

## Abstract

A central aim of population genetics is the inference of the evolutionary history of a population. To this end, the underlying process can be represented by a model of the evolution of allele frequencies parametrized by *e.g.*, the population size, mutation rates and selection coefficients. A large class of models use forward-in-time models, such as the discrete Wright-Fisher and Moran models and the continuous forward diffusion, to obtain distributions of population allele frequencies, conditional on an ancestral initial allele frequency distribution. Backward-in-time diffusion processes have been rarely used in the context of parameter inference. Here, we demonstrate how forward and backward diffusion processes can be combined to efficiently calculate the exact joint probability distribution of sample and population allele frequencies at all times in the past, for both discrete and continuous population genetics models. This procedure is analogous to the forward-backward algorithm of hidden Markov models. While the efficiency of discrete models is limited by the population size, for continuous models it suffices to expand the transition density in orthogonal polynomials of the order of the sample size to infer marginal likelihoods of population genetic parameters. Additionally, conditional allele trajectories and marginal likelihoods of samples from single populations or from multiple populations that split in the past can be obtained. The described approaches allow for efficient maximum likelihood inference of population genetic parameters in a wide variety of demographic scenarios.

**Keywords:** bi-allelic mutation-drift model, Markov chain, forward-backward algorithm, forward-backward diffusion, exact inference.

---

\*Corresponding author

*Email addresses:* juraj.bergman@vetmeduni.ac.at (Juraj Bergman), dominik.schrempf@vetmeduni.ac.at (Dominik Schrempf), ck202@st-andrews.ac.uk (Carolin Kosiol), claus.vogl@vetmeduni.ac.at (Claus Vogl)

---

## 1 1. Introduction

2 Most basic population genetic models, *e.g.*, the Wright-Fisher and the Moran  
3 models as well as the forward and backward diffusion models, were introduced  
4 before molecular sequence data became available [reviewed in 10]. Thus, em-  
5 phasis was on demonstrating processes over time and on qualitatively explaining  
6 observations, rather than on quantitative inference of population genetic forces  
7 given molecular data. Much later, coalescent theory [17] has been used both for  
8 demonstration of processes as well as for inference given a population sample  
9 [13, 38]. The coalescent reconstructs the genealogical history of a particular  
10 sample at a particular locus conditional on population genetic forces. However,  
11 the aim in statistical population genetics is usually the inference of evolution-  
12 ary forces or of the evolutionary trajectory of allele proportions of the whole  
13 population.

14 Population genetic parameters have often been inferred from allele frequency  
15 data of a single locus sampled at multiple time-points in the past. Due to the  
16 short time-spans, mutation can usually be neglected, while selection is impor-  
17 tant. Bollback *et al.* [4] developed a method based on a forward diffusion model  
18 to infer the strength of selection acting on an allele. This method was later ex-  
19 tended to additionally infer the age of the selected allele [22]. To calculate the  
20 likelihood of the observed trajectory, these methods rely on solving the diffu-  
21 sion equation using a numerical grid approach. On the other hand, Steinrücken  
22 *et al.* [31] use a system of orthogonal polynomials, *i.e.*, a spectral representa-  
23 tion of the transition density [29], to analytically solve the diffusion equation  
24 and model the evolution of allele frequency. Recently, Schraiber *et al.* [27] de-  
25 veloped a Bayesian approach that uses Markov chain Monte Carlo (MCMC)  
26 integration of allele frequency trajectories to provide estimates of population  
27 genetic parameters.

28 While the above-described methods deal with a single locus with data from  
29 multiple time-points, the focus of this study is to infer the demographic history  
30 and the population genetic forces acting on a whole population from present-day  
31 data. Specifically, we are interested in inference of population genetic param-  
32 eters, such as the scaled mutation rate or mutation bias given data  $y$  from the  
33 present,  $t = 0$ , that consist of an alignment of  $M$  (haploid) sequences. Nu-  
34 cleotide data are assumed to be independently and identically drawn from a  
35 population across  $L$  freely recombining nucleotide sites. The sites are assumed  
36 to be neutral, *e.g.*, in short introns, or at least nearly-neutral, *e.g.*, fourfold de-  
37 generate sites, such that the data are informative about population demography  
38 and mutation processes. Because sites are assumed independent, they can be  
39 summarized as a site frequency spectrum (SFS), also called the allele frequency  
40 spectrum. The likelihood of the population sample  $y$  can be calculated given the  
41 present population allele frequency  $x_0$  and a probability model of the sampling  
42 process. The distribution of  $x_0$  is in turn given by a population genetic model  
43 parametrized to capture mutation or the demographic history. These population

44 genetic parameters can be inferred by first integrating over  $x_0$  and subsequently  
 45 maximizing the marginal likelihood of the data  $y$  by varying the model parame-  
 46 ters; a strategy that may also be viewed as the empirical Bayes method [e.g., 5].  
 47 Under the assumption of equilibrium and given a general mutation-drift model,  
 48 this strategy leads to a beta-binomial likelihood, which can be maximized using  
 49 an expectation-maximization algorithm [34]. Assuming that mutations are rare  
 50 and arise only at fixed sites, *i.e.*, a boundary mutation model, it is possible to  
 51 derive maximum likelihood estimators of the mutation rate and bias as well as  
 52 the selection coefficient [35]. The estimator of the mutation rate in [35] is a  
 53 variant of the well-know Ewens-Watterson  $\theta$  [9, 39].

54 The assumption of equilibrium is often violated in natural populations and,  
 55 therefore, within this framework, modelling allele frequency trajectories is neces-  
 56 sary to accurately infer parameters from the observed SFS. Furthermore, even  
 57 under equilibrium, maximum likelihood inference requires modelling of allele  
 58 trajectories with data from two or more populations that split some time in the  
 59 past, represented by a joint SFS (jSFS). Herein, we mostly focus on inference  
 60 using the jSFS given the canonical model of two populations that split at some  
 61 known or unknown time in the past, from which samples of sizes  $M^{(1)}$  and  $M^{(2)}$   
 62 are obtained at the present time. Inference using jSFS has been implemented  
 63 in the well-known program *∂a∂i* by Gutenkunst *et al.* [12]. It is widely used to  
 64 infer migration rates, selection coefficients and split times given data from mul-  
 65 tiple populations using a numerical grid approach to solve the forward diffusion  
 66 equation and model allele trajectories. An alternative approach was developed  
 67 in Lukić *et al.* [21] and Lukić and Hey [20], where as in [29, 31], orthogonal poly-  
 68 nomials are used to model allele frequency evolution. A similar, but discrete  
 69 model of allele frequency evolution is presented in Jewett *et al.* [14].

70 All of these methods model the evolution of the allele frequency forward in  
 71 time. However, backward models can also be used to model allele frequency tra-  
 72 jectories and calculate the likelihood of the data  $y$  conditional on the population  
 73 allele frequency  $x_t$  at earlier times ( $t < 0$ ). Based on the Wright-Fisher model,  
 74 Zhao *et al.* [46] provide an algorithm to calculate probabilities of intermediate  
 75 states conditional on the starting and end states. This allows simulation of  
 76 conditional trajectories. Schrempf *et al.* [28] use a Moran model in phylogenetic  
 77 inference. The “pruning algorithm” [11] allows computation of the likelihood  
 78 from the tips of a phylogenetic tree down to the root, *i.e.*, backward in time. For  
 79 efficient inference of phylogenetic trees reversibility of the evolutionary process  
 80 is generally assumed.

81 In this article, we demonstrate the usefulness of backward-in-time processes  
 82 in parameter inference, while considering both discrete population genetics mod-  
 83 els and continuous diffusion. We also show parallels between discrete and  
 84 continuous models. Combining the forward and backward processes, as with  
 85 the forward-backward algorithm of hidden Markov models (HMM) [25], the  
 86 probability distribution of population allele frequencies conditional on data  
 87  $\Pr(x_t | y, \dots)$  can be inferred at time  $t$  in the past and the distribution of con-  
 88 ditional trajectories can be simulated. We therefore use forward and backward  
 89 processes to conveniently calculate probability distributions in time conditional

90 on a SFS or jSFS from the present. Furthermore, we introduce bi-allelic bound-  
 91 ary mutation models, with mutations occurring only at fixed sites. Specifically,  
 92 we present the solution to the boundary mutation-drift diffusion model, which  
 93 underlies the infinite site or Poisson-random-fields models [16, 26] and is impor-  
 94 tant in statistical inference in population genetics as a starting point to derive  
 95 maximum likelihood estimators, such as the well-known Ewens-Watterson es-  
 96 timator of the scaled mutation rate [9, 39]. The Markov chains of the models  
 97 under consideration have no absorbing states and therefore have stationary dis-  
 98 tributions. We do not always assume time-reversibility. For the discrete models,  
 99 the transition matrix must be multiplied repeatedly to obtain the distribution  
 100 of population allele frequencies forward and backward in time. As the size of  
 101 the transition matrix depends on the population size  $N$ , multiplication becomes  
 102 cumbersome if  $N$  is large. In the limit of large population sizes, the corre-  
 103 sponding Kolmogorov forward and backward diffusion equations are obtained.  
 104 Orthogonal polynomials provide a flexible and fast method to solve the diffusion  
 105 equations and calculate marginal likelihoods for inference in population genet-  
 106 ics. For most purposes, expansion of polynomials up to the order of the sample  
 107 size  $M$  suffices to accurately infer the transition density. With two populations,  
 108 it can be shown that the order of the expansion is between the minimum and the  
 109 maximum of the two sample sizes, depending on the starting distribution. As  
 110 this is usually much less than the population size, continuous diffusion models  
 111 may be much more efficient for parameter inference in population genetics than  
 112 equivalent discrete models.

## 113 2. Time-homogeneous discrete Markov chains

114 In this section we apply the forward-backward algorithm [25] to discrete  
 115 population genetic models for inference given a SFS or a jSFS. To this end, we  
 116 rephrase iteration using discrete population genetic models (Wright-Fisher or  
 117 Moran) in the terminology of the forward-backward algorithm [*e.g.*, 25]. We  
 118 mainly use matrix notation to emphasize the similarities between discrete iter-  
 119 ation and the continuous models in Sections 3 and 7.1. For completeness and  
 120 clarity, subsections include reviews of standard theory.

### 121 2.1. Assumptions

- 122 (i) Assume a haploid population of size  $N$  and a bi-allelic mutation model.  
 123 The time-dependent frequency of allele one in the population at time  $t$  is  
 124 denoted  $x_t$  ( $0 \leq x_t \leq N$ ) and is assumed to evolve as a discrete, time-  
 125 homogeneous Markov chain with a transition probability matrix  $\mathbf{T}$ , where  
 126  $(\mathbf{T})_{ij} = \Pr(x_{t+1} = j | x_t = i)$  with  $i, j \in \{0, \dots, N\}$ .  $\mathbf{T}$  is an aperiodic,  
 127 right stochastic matrix.
- 128 (ii) At a (possibly unknown) time  $t = s$  ( $s < 0$ ) in the past, a distribution  
 129 of population allele proportions is given by  $\boldsymbol{\rho}$  with entries  $(\rho_i)_{i \in \{0, \dots, N\}} =$   
 130  $\Pr(x_s = i)$ . In particular,  $\boldsymbol{\rho}$  may be the stationary distribution  $\boldsymbol{\pi} =$   
 131  $(\pi_i)_{i \in \{0, \dots, N\}}$  or may correspond to a joint distribution of some other data  
 132 and the equilibrium allele frequency distribution.

133 (iii) The population evolves until the present time  $t = 0$ , when a sample of  
 134 size  $M$  is drawn. We denote the sampled frequency of allele one as  $y$   
 135 ( $0 \leq y \leq M$ ). The probability of observing  $y$ , *i.e.*, the likelihood, is  
 136  $\Pr(y|M, x_0)$  (we may drop the dependency on  $M$  in the following) and  
 137 will be defined according to the application.

138 For two populations, assumptions (ii) and (iii) are modified:

139 (ii) At a (possibly unknown) time  $t = s$  ( $s < 0$ ) in the past,  $x_s$  is drawn from a  
 140 distribution of population allele proportions  $\boldsymbol{\rho}$ . The population separates  
 141 immediately into two populations with the same initial allele frequency  $x_s$ .  
 142 (iii) The two populations evolve independently until the present time  $t = 0$ ,  
 143 when samples of sizes  $M^{(1)}$  and  $M^{(2)}$  are drawn from each population.

144 For discrete models, iteration is more efficient if the population size  $N$  is  
 145 small.  $N$  can be decreased by increasing the mutation rate  $\mu$  such that their  
 146 product  $\theta = N\mu$  remains constant. For moderate  $N$ , the error introduced by  
 147 such scaling is small and converges to zero in the diffusion limit. Therefore,  $N$   
 148 can be set according to numerical convenience. Often, our data are from the  
 149 present and we want to condition on the configuration of allele frequencies at  
 150 earlier times.

## 151 2.2. The forward-backward algorithm

152 The forward-backward algorithm of hidden Markov models (HMMs) [*e.g.*,  
 153 25, 6, 37] is an efficient numerical method for calculating probabilities assuming  
 154 a Markovian underlying process, where key variables, the “states”, are assumed  
 155 to be unknown, *i.e.*, “hidden”. Intermediate results and the algorithm in general  
 156 can readily be interpreted probabilistically. The algorithm’s numerical efficiency  
 157 is based on the simple, acyclic conditional dependence structure of the unknown  
 158 variables, which allows for “dynamic programming”. In our case, the possible  
 159 values of the population allele frequency  $x_t$  correspond to the hidden states,  
 160 while the probability distribution  $\Pr(y|x_t = i)$  to the emission probabilities.  
 161 With the Wright-Fisher or the Moran models, allele frequencies at the next  
 162 time-point  $x_{t+1}$  depend only on the current ones, which conforms to a Markov  
 163 process. Knowing the sample allele frequencies generally does not completely  
 164 identify the population allele frequencies at any time-point; the exact state of  
 165 the underlying variable remains “hidden”.

## 166 2.3. Forward in time

167 We introduce the row vector  $\mathbf{f}_t$  with entries  $(\mathbf{f}_t)_i = \Pr(x_t = i | \boldsymbol{\rho})$ , where  
 168  $i \in \{0, \dots, N\}$ , and  $\mathbf{f}_s = \boldsymbol{\rho}$ , *i.e.*, the vector of initial probabilities of states, and  
 169 define recursively:

$$\mathbf{f}_{t+1} = \mathbf{f}_t \mathbf{T} \quad (s \leq t < 0). \quad (1)$$

170 Thus,  $\mathbf{f}_t$  can be interpreted as the probability of the allele frequency at time  $t$   
 171 conditional on the ancestral state  $\boldsymbol{\rho}$ ,  $\mathbf{f}_t = \Pr(x_t | \boldsymbol{\rho})$ . This corresponds to the  
 172 forward method in the forward-backward algorithm in the theory of HMMs [*e.g.*,

173 25, 37]. Let  $\mathbf{b}'_0$  be a column vector (the prime ' depicts matrix transposition)  
 174 corresponding to the conditional of the sampling process, such that  $(\mathbf{b}_0)_i =$   
 175  $\Pr(y | x_0 = i)$  with  $i \in \{0, \dots, N\}$ . The marginal likelihood then is

$$\Pr(y | \boldsymbol{\rho}) = \boldsymbol{\rho} \mathbf{T}^{|\mathbf{s}|} \mathbf{b}'_0. \quad (2)$$

#### 176 2.4. Backward in time

177 Using a strategy as with the backward method in the theory of HMM [25, 37],  
 178 we set

$$\mathbf{b}'_t = \mathbf{T} \mathbf{b}'_{t+1} \quad (s \leq t < 0), \quad (3)$$

179 which can also be written as

$$(\mathbf{b}_t)_i = \Pr(y | x_t = i) = \sum_j \Pr(x_{t+1} = j | x_t = i) \Pr(y | x_{t+1} = j). \quad (4)$$

180 From the definition of  $\mathbf{b}_t$ , it follows that we condition on  $x_t$ . The recursion  
 181 moves the conditioning to ever earlier times. The marginal likelihood (2) may  
 182 also be obtained as follows:

$$\begin{aligned} \Pr(y | \boldsymbol{\rho}) &= \boldsymbol{\rho} \left[ \mathbf{T}^{|\mathbf{s}|} \mathbf{b}'_0 \right] \\ &= \boldsymbol{\rho} \mathbf{b}'_s \\ &= \sum_i \rho_i \Pr(y | x_s = i). \end{aligned} \quad (5)$$

#### 183 2.5. Constant marginal distribution and adjointness

184 Considering the sampling probability, we can choose any arbitrary  $t$  such  
 185 that

$$\Pr(y | \boldsymbol{\rho}) = \mathbf{f}_t \mathbf{b}'_t = \sum_i \Pr(x_t = i | \boldsymbol{\rho}) \Pr(y | x_t = i) = \langle \mathbf{f}_t, \mathbf{b}_t \rangle, \quad (6)$$

186 holds, where  $\langle \cdot, \cdot \rangle$  denotes an inner product. It follows that the forward and  
 187 backward transition matrices, *i.e.*,  $\mathbf{T}$  and its transpose  $\mathbf{T}'$ , are adjoint since

$$\begin{aligned} \Pr(y | \boldsymbol{\rho}) &= \Pr(y | \boldsymbol{\rho}) \\ (\mathbf{f}_t \mathbf{T}) \mathbf{b}'_{t+1} &= \mathbf{f}_t (\mathbf{T} \mathbf{b}'_{t+1}) \\ \langle \mathbf{f}_t \mathbf{T}, \mathbf{b}_{t+1} \rangle &= \langle \mathbf{f}_t, \mathbf{b}_{t+1} \mathbf{T}' \rangle. \end{aligned} \quad (7)$$

188 This adjoint relationship allows movement forward and backward in time.

#### 189 2.6. Joint and conditional distribution

190 The probability of  $x_t = i$  and  $y$  conditional on the starting distribution  $\boldsymbol{\rho}$  is

$$\Pr(x_t = i, y | \boldsymbol{\rho}) = (\mathbf{f}_t)_i (\mathbf{b}_t)_i. \quad (8)$$

191 Furthermore, the probability of  $x_t = i$  conditional on the data and the starting  
 192 distribution is

$$\Pr(x_t = i | y, \boldsymbol{\rho}) = \frac{(\mathbf{f}_t)_i (\mathbf{b}_t)_i}{\mathbf{f}_t \mathbf{b}'_t}. \quad (9)$$

193 This allows calculation of the distribution of population allele frequencies con-  
 194 ditional on the data and an initial condition at any time.



195 *2.7. Sampling from conditional trajectories*

196 It is possible to simulate trajectories given the initial distribution  $\rho$  at time  
 197  $s$  and the likelihood at time  $t = 0$ . Note that Zhao *et al.* [46] provide a similar  
 198 algorithm based on the Wright-Fisher model to simulate trajectories of popula-  
 199 tion allele proportions conditional on the starting and end states. In contrast,  
 200 we start with a sample at time  $t = s$  from the conditional probabilities (9).  
 201 Given the state at time  $t - 1$  the probability of the state at time  $t$  is

$$\Pr(x_t = j | x_{t-1} = i, y) = \frac{(\mathbf{T})_{ij}(\mathbf{b}_t)_j}{(\mathbf{b}_{t-1})_i}, \quad (10)$$

202 which can be used to obtain a sample trajectory. Although the probability  
 203 distribution of trajectories depends on  $\rho$ , the transition at a given time  $t$  (10)  
 204 does not contain  $\rho$  since it is a Markov process.

205 *2.8. Left and right eigenvectors, stationary distribution*

206 Let  $\pi = (\pi_i)_{i \in \{0, \dots, N\}}$  be the stationary distribution of  $\mathbf{T}$ , if it exists.  $\pi$  is  
 207 the left eigenvector associated with the largest eigenvalue (equal to one) [10, p.  
 208 87]

$$\pi = \pi \mathbf{T}. \quad (11)$$

209 All entries of  $\pi$  are strictly greater than zero because the transition matrix  
 210 was assumed to be irreducible and  $\sum \pi_i = 1$ . Thus the entries of  $\pi$  can be  
 211 interpreted as probabilities. Since the rows of  $\mathbf{T}$  sum to one, it is obvious  
 212 that a column vector of all ones  $\mathbf{1}'$  is the right eigenvector associated with the  
 213 unit eigenvalue. In our context, this means that iterating forward in time will  
 214 converge to a vector proportional to  $\pi$  and iterating backward in time to a  
 215 vector proportional to  $\mathbf{1}'$ . Thus, every state is equally likely when  $s \rightarrow -\infty$   
 216 and we have no information about the initial distribution of states, because the  
 217 process has already reached equilibrium.

218 *2.9. Reversibility*

219 Define the diagonal matrix  $\mathbf{\Pi}$  with the entries  $\pi_i$  on the main diagonal. Since  
 220 irreducible Markov chains with finite state space have stationary distributions  
 221 with only strictly positive entries,  $\mathbf{\Pi}$  is invertible with  $\mathbf{\Pi}^{-1}$  being a diagonal  
 222 matrix with entries  $1/\pi_i$ . Set

$$\mathbf{T}^* = \mathbf{\Pi} \mathbf{T} \mathbf{\Pi}^{-1}. \quad (12)$$

223 The Markov chain is reversible, if  $\mathbf{T}^* = \mathbf{T}'$ , because then

$$\begin{aligned} \mathbf{T}' &= \mathbf{\Pi} \mathbf{T} \mathbf{\Pi}^{-1} \\ \mathbf{T}' \mathbf{\Pi} &= \mathbf{\Pi} \mathbf{T}, \end{aligned} \quad (13)$$

224 which corresponds to the condition of detailed balance.

225 If reversibility holds, we can separate  $\mathbf{f}_t$  into a product of a time dependent  
 226 row vector  $\mathbf{g}_t$  and the stationary distribution matrix  $\mathbf{\Pi}$

$$\mathbf{f}_t = \mathbf{g}_t \mathbf{\Pi}. \quad (14)$$

227 Under reversibility, we have forward in time

$$\begin{aligned} \mathbf{g}_{t+1} \mathbf{\Pi} &= \mathbf{g}_t \mathbf{\Pi} \mathbf{T} \\ \mathbf{g}_{t+1} &= \mathbf{g}_t \mathbf{\Pi} \mathbf{T} \mathbf{\Pi}^{-1} \\ \mathbf{g}_{t+1} &= \mathbf{g}_t \mathbf{T}' . \end{aligned} \quad (15)$$

228 We may interpret  $\mathbf{g}_t$  as a ‘‘projected likelihood’’ that, when multiplied with  
 229 the stationary distribution, gives the joint distribution  $\mathbf{f}_t$ . Note that with the  
 230 decomposition (14), the likelihood becomes

$$\Pr(y | \boldsymbol{\rho}) = \mathbf{g}_t \mathbf{\Pi} \mathbf{b}'_t \quad \text{for all } t. \quad (16)$$

231 The adjoint relationship (7) can be modified analogously, to result in the self-  
 232 adjoint relationship

$$\begin{aligned} \Pr(y | \boldsymbol{\rho}) &= \Pr(y | \boldsymbol{\rho}) \\ (\mathbf{g}_t \mathbf{\Pi} \mathbf{T}) \mathbf{b}'_{t+1} &= \mathbf{g}_t (\mathbf{T}' \mathbf{\Pi} \mathbf{b}'_{t+1}) \\ \langle \mathbf{g}_t \mathbf{\Pi} \mathbf{T}, \mathbf{b}_{t+1} \rangle &= \langle \mathbf{g}_t, \mathbf{b}_{t+1} \mathbf{\Pi} \mathbf{T} \rangle. \end{aligned} \quad (17)$$

### 233 2.10. Example: Conditional probabilities under irreversible mutation

234 As a particular realization of a discrete process consider a bi-allelic model,  
 235 where alleles can be labeled either as ancestral (zero) or derived (one). Mutation  
 236 rates are assumed to be small (at most one mutation is segregating per site) and  
 237 occur only at the boundary zero. When a derived allele is fixed, it immediately  
 238 becomes ancestral. This process is a variant of the infinite sites model [16], but  
 239 differs in that it allows for a stationary distribution at a particular site. Using  
 240 diffusion theory, Evans *et al.* [8] provide an analysis based on moments of the  
 241 allele proportions of a similar model with mutations from only one boundary,  
 242 assuming changing population sizes, *i.e.*, not assuming equilibrium. Zivkovic  
 243 *et al.* [48] extend the analysis to include selection.

244 The transition matrix  $\mathbf{T}$  is defined as follows. Given a time-homogeneous  
 245 mutation rate  $\mu$ , transition probabilities at the boundary zero are

$$\begin{cases} \Pr(x_{t+1} = 0 | x_t = 0) &= 1 - \mu / (1 - \theta H_{N-1}) \\ \Pr(x_{t+1} = 1 | x_t = 0) &= \mu / (1 - \theta H_{N-1}), \end{cases} \quad (18)$$

246 where  $\theta = N\mu$  and the harmonic number  $H_{N-1} = \sum_{i=1}^{N-1} 1/i$ . With this defini-  
 247 tion, we consider the Moran model where with each time-step (note that with  
 248 the Moran model  $N$  time-steps correspond to one generation with the Wright-  
 249 Fisher model), one individual sampled at random has one offspring that replaces

250 one other random individual. Within the polymorphic region, random drift is  
 251 the only force affecting allele frequencies, such that for  $2 \leq i \leq N - 2$

$$\begin{cases} \Pr(x_{t+1} = i - 1 | x_t = i) &= \frac{1}{N^2}i(N - i) \\ \Pr(x_{t+1} = i | x_t = i) &= 1 - \frac{1}{N^2}2i(N - i) \\ \Pr(x_{t+1} = i + 1 | x_t = i) &= \frac{1}{N^2}i(N - i). \end{cases} \quad (19)$$

252 For  $i = N - 1$ , drift may lead to fixation of the derived allele, which then  
 253 becomes the ancestral allele, *i.e.*,

$$\begin{cases} \Pr(x_{t+1} = N - 2 | x_t = N - 1) &= \frac{1}{N^2}(N - 1) \\ \Pr(x_{t+1} = N - 1 | x_t = N - 1) &= 1 - \frac{1}{N^2}2(N - 1) \\ \Pr(x_{t+1} = 0 | x_t = N - 1) &= \frac{1}{N^2}(N - 1). \end{cases} \quad (20)$$

254 The state  $i = N$  is never reached and is left out of the state space. The system  
 255 is not in detailed balance, as probability mass moves from state  $i = N - 1$  to  
 256 state  $i = 0$ , but not in the reverse direction.

257 The stationary distribution is

$$\pi(x) = \begin{cases} \Pr(x = 0) &= 1 - \theta H_{N-1} \\ \Pr(x = i)_{i \in \{1, \dots, N-1\}} &= \theta/i, \end{cases} \quad (21)$$

258 as can be ascertained by substitution.

259 Note that the proportion of polymorphism in equilibrium is  $\theta H_{N-1}$ . This  
 260 equilibrium proportion corresponds to the Ewens-Watterson estimator  $\theta_W$  [9,  
 261 39], which was derived using the infinite site model [16]. In formula (18), the  
 262 mutation probability per time-step  $\mu$  is weighted by the inverse of the probability  
 263 of being at the boundary  $1 - \theta H_{N-1}$ , which ensures that the average probability  
 264 of mutations per time-step is constant, irrespective of  $N$ . This in turn assures  
 265 correspondence to the infinite site model.

266 Assume a hypergeometric likelihood of  $y$ , conditional on  $N$ ,  $x_0 = i$ , and the  
 267 sample size  $M \leq N$

$$\Pr(y | N, x_0 = i, M) = \frac{\binom{i}{y} \binom{N-i}{M-y}}{\binom{N}{M}}, \quad (22)$$

268 where  $0 \leq y \leq M$  and  $0 \leq i \leq (N - 1)$ . In equilibrium, the joint distribu-  
 269 tion is obtained by multiplying the stationary distribution with the likelihood.  
 270 Summing out the population allele frequency  $x_0$ , the marginal distribution is  
 271 obtained

$$\Pr(y | M) = \begin{cases} \Pr(y = 0 | M) &= 1 - \theta H_{M-1} \\ \Pr(y = i | M)_{i \in \{1, \dots, M-1\}} &= \theta/i. \end{cases} \quad (23)$$

272 It follows that the expected heterozygosity, *i.e.*, the probability of obtaining one  
 273 derived allele and one ancestral allele in a sample of size  $M = 2$  is  $\theta$ .

274 As an example of a demographic scenario (Fig. 1A), consider a population  
 275 with a stationary allele frequency distribution (21) defined by the ancestral  
 276 mutation rate  $\mu_a$  at some time  $s$  in the past; *i.e.*,  $\rho = \pi_a$ . Furthermore, assume  
 277 an instantaneous increase in the mutation rate  $\mu$  between generations  $s$  and  
 278  $s+1$ . As  $\theta = N\mu$ , this mimicks an expansion of the population size, without the  
 279 inconvenience of having to change the dimension of the transition matrix. From  
 280 then on, the population is out of equilibrium and evolving with a new current  
 281 mutation rate  $\mu_c > \mu_a$ . At the present time ( $t = 0$ ), we sample  $M$  haplotypes  
 282 from the population. Assume that the ancestral state of the sampled haplotypes  
 283 can be determined without error. Thus, a polarized SFS may be constructed.  
 284 The transition matrix  $\mathbf{T}$  and its transpose  $\mathbf{T}'$  can be calculated conditional on  
 285  $\mu_c$ . Assume hypergeometric sampling. The conditional probabilities of allelic  
 286 states  $\Pr(x_t | y, \rho)$ , for any time  $s \leq t \leq 0$ , in a site frequency spectrum of size  
 287  $M$  can then be calculated (Fig. 2).

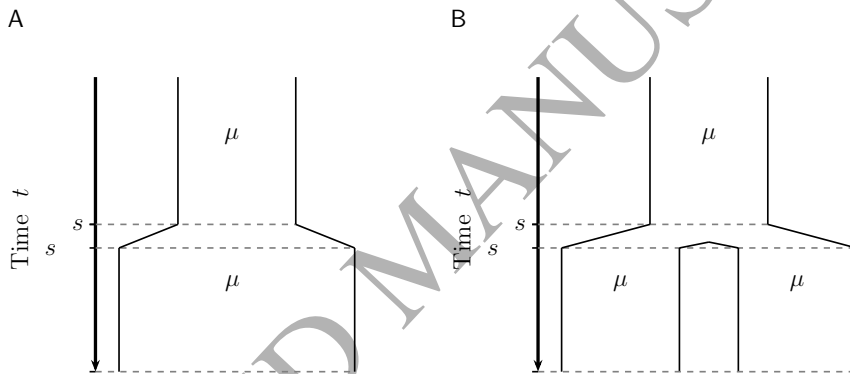


Figure 1: Demographic scenarios. A) Population expansion. B) Population split.

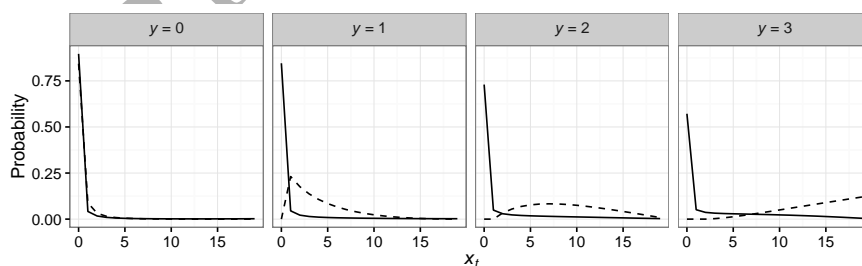


Figure 2: Conditional probabilities of allelic states in a site frequency spectrum of size  $M = 3$ . The solid lines represent the conditional probabilities of an allelic state  $x_t$  given  $y$ , at  $t = s$ , while the dashed lines represent the probabilities at  $t = 0$ . The parameters were set to  $\mu_a = 0.05$ ,  $\mu_c = 0.1$ ,  $s = -200$  and  $N = 20$ .

288 *2.11. Example: Joint site frequency spectrum under reversible mutation*

289 As another realization of a discrete process consider a bi-allelic mutation-  
 290 drift decoupled Moran model [2, 7] with haploid population size  $N$ , mutation  
 291 rate towards zero  $\mu_0$  and mutation rate towards one  $\mu_1$  ( $\mu = \mu_0 + \mu_1$ ). We  
 292 introduce the parameters  $\alpha = \mu_1/\mu$  ( $0 \leq \alpha \leq 1$ ) and  $\beta = 1 - \alpha = \mu_0/\mu$   
 293 which are the mutation biases towards allele one and zero, respectively. Let  $i$   
 294 ( $0 \leq i \leq N$ ) be the frequency of allele one. Then, the tri-diagonal transition  
 295 rate matrix  $\mathbf{T}$  depends on  $N$ ,  $\mu$  and  $\alpha$

$$\begin{cases} \Pr(x_{t+1} = i - 1 | x_t = i) &= \frac{i(N-i)}{N^2} + \beta\mu\frac{i}{N} \\ \Pr(x_{t+1} = i | x_t = i) &= 1 - \frac{2i(N-i)}{N^2} + \beta\mu\frac{i}{N} + \alpha\mu\frac{N-i}{N} \\ \Pr(x_{t+1} = i + 1 | x_t = i) &= \frac{i(N-i)}{N^2} + \alpha\mu\frac{N-i}{N}. \end{cases} \quad (24)$$

296 The stationary distribution of  $x$  is a beta-binomial

$$\Pr(x = i | N, \alpha, \theta) = \binom{N}{i} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(i + \alpha\theta)\Gamma(N - i + \beta\theta)}{\Gamma(N + \theta)}, \quad (25)$$

297 which can be verified by substitution into the equations of detailed balance (25).  
 298 As above, hypergeometric sampling at time  $t = 0$  is assumed. Assuming equi-  
 299 librium, the marginal likelihood of a single sample of size  $M$  is again a beta-  
 300 binomial, with  $M$  replacing  $N$  [34].

301 Consider an ancestral population with the stationary allele frequency dis-  
 302 tribution (25). The ancestral population splits into two at some time  $s$  in the  
 303 past (Fig. 1B). For simplicity, no change in the mutation, the drift parameter,  
 304 and the size in both populations is assumed. A jSFS is simulated from both  
 305 populations (Table 1) at  $t = 0$ . The likelihood of the split time  $s$  calculated  
 306 given the simulated jSFS (Figure 3A) has a single maximum close to the true  
 307 value of  $t = -40$ .

308 It may be instructive to calculate some marginal and conditional probabili-  
 309 ties with this example. We set for the likelihood of the second population, *i.e.*,  
 310 the conditional distribution of the data given the allele frequencies in the sec-  
 311 ond population at time  $t = 0$ ,  $\mathbf{b}_0^{(2)} = \Pr(y^{(2)} | x_0^{(2)})$ . We then iterate backward  
 312 within the second population until  $t = s$  to obtain the joint probability of the  
 313 second sample  $y^{(2)}$  and the  $i$ th allele frequency  $x_s = i$  at time  $t = s$ :

$$\Pr(x_s = i, y^{(2)} | \rho) = \rho_i (\mathbf{b}_s^{(2)})_i. \quad (26)$$

314 Note that, on the left side of the above equation, we drop the superscript to  
 315 indicate the population for  $x_s$ , because time  $t = s$  is just before the split into the  
 316 two descendant populations. Without information from the second population,  
 317 we would set the starting distribution of the first population  $\mathbf{f}_s^{(1)}$  to the prior  
 318 probability of the allele frequencies at time  $t = s$ , *i.e.*,  $\mathbf{f}_s^{(1)} = \rho$ . With infor-  
 319 mation on the second population, we instead start at time  $t = s$  from the joint  
 320 probability (26) and set  $\mathbf{f}_s^{(1)*} = \Pr(x_s, y^{(2)} | \rho)$ . As before, we iterate forward  
 321 to obtain  $\mathbf{f}_t^{(1)*}$  within the first population; we can interpret  $\mathbf{f}_t^{(1)*}$  as the joint

322 probability of the allele frequency in the first population and the data of the  
 323 second population:  $\mathbf{f}_t^{(1)*} = \Pr(x_t^{(1)}, y^{(2)} | \boldsymbol{\rho})$ . Setting now for the likelihood of  
 324 the first population  $\mathbf{b}_0^{(1)} = \Pr(y^{(1)} | x_0^{(1)})$  and iterating backward within the first  
 325 population until  $t$ , we obtain the probability of the allele frequency of the first  
 326 population at  $t$ , conditional on data from both the first and second population  
 327 as well as on the prior distribution  $\boldsymbol{\rho}$  as:

$$\Pr(x_t^{(1)} = i | y^{(1)}, y^{(2)}, \boldsymbol{\rho}) = \frac{(\mathbf{f}_t^{(1)*})_i (\mathbf{b}_t^{(1)})_i}{\mathbf{f}_t^{(1)*} \mathbf{b}_t^{(1)}}. \quad (27)$$

328 Figure 3B gives the conditional probability  $\Pr(x_t | y^{(1)}, y^{(2)}, \boldsymbol{\rho})$  for one site class  
 329 of the jSFS determined by  $y^{(1)}$  and  $y^{(2)}$  which denote the polymorphism levels  
 330 of the specific class for populations one and two, respectively; *e.g.*, the site class  
 331 determined by  $y^{(1)} = 1$  and  $y^{(2)} = 2$  contains all sites with one derived allele in  
 332 population one and two derived alleles in population two.

Table 1: A jSFS simulated with a discrete Moran model with parameters  $L = 10^5$ ,  $M^{(1)} = M^{(2)} = 3$ ,  $\alpha = 2/3$ ,  $\theta = 0.1$ ,  $s = -40$  and  $N = 20$ .

$y$	0	1	2	3
0	29037	1315	436	185
1	1276	688	539	432
2	446	529	662	1524
3	202	507	1430	60792

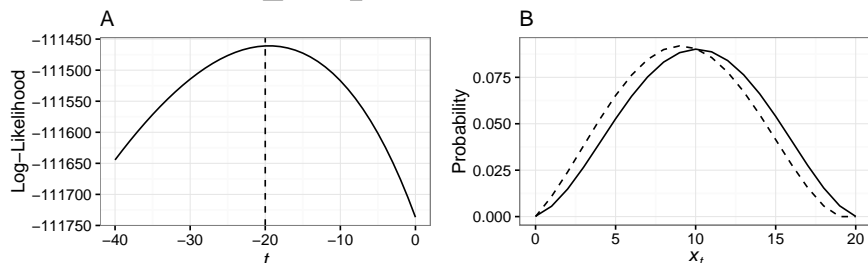


Figure 3: A) The log-likelihood of the split time  $s$ , given a jSFS (Table 1). The dashed line indicates the true split time. B) The conditional probability of the allelic state  $x_t$  given  $y^{(1)} = 1$  and  $y^{(2)} = 2$ , at  $t = s$  (solid line) and  $t = 0$  (dashed line).

### 333 2.12. Summary: discrete Markov chains

334 With standard discrete population genetic models, *e.g.*, the Wright-Fisher  
 335 or the Moran models, iteration of discrete Markov chains forward in time cor-  
 336 responds to the forward algorithm and backward in time to the backward al-  
 337 gorithm of the forward-backward algorithm [25]. With such algorithms, it is

338 straightforward to calculate exact likelihoods given SFS and jSFS from the  
 339 present. Some standard population genetic mutation models are reversible,  
 340 others are not. In contrast to phylogenetic applications [11, 28], reversibility  
 341 of the Markov chain does not simplify calculations considerably; in both cases,  
 342 iteration of an  $(N + 1) \times (N + 1)$  transition matrix is needed.

### 343 3. Forward and backward diffusion equations

344 In this section, we provide theory for the continuous analogs of the discrete  
 345 forward and backward transition probabilities both for reversible and irreversible  
 346 Markov processes and illustrate with examples. We derive the forward and  
 347 backward diffusion equations from the discrete general mutation-drift Moran  
 348 model using only the definitions of the first and second symmetric derivative  
 349 (Appendix 7.1).

350 With the forward and backward diffusion operators

$$\begin{aligned} \mathcal{L} &= -\frac{\partial}{\partial x}P(x) + \frac{\partial^2}{\partial x^2}Q(x) \\ \mathcal{L}^* &= P(x)\frac{\partial}{\partial x} + Q(x)\frac{\partial^2}{\partial x^2}, \end{aligned} \quad (28)$$

351 the forward and backward diffusion equations are written as

$$\begin{aligned} \frac{\partial}{\partial \tau}\phi(x|\tau, \rho) &= \mathcal{L}\phi(x|\tau, \rho) \\ -\frac{\partial}{\partial \tau}\psi(y|x, \tau) &= \mathcal{L}^*\psi(y|x, \tau), \end{aligned} \quad (29)$$

352 where  $\tau$  is the continuous-time analog of  $t$ , and  $\rho$  is the initial condition of  
 353 the continuous allelic frequency  $x$ . The functions  $\phi(x|\tau, \rho)$  and  $\psi(y|x, \tau)$  are  
 354 transition density functions of the forward and backward diffusion, respectively.  
 355 Obviously, these functions must be twice differentiable in the open interval  $(0, 1)$ .  
 356 The operators  $\mathcal{L}$  and  $\mathcal{L}^*$  together with the boundary conditions correspond to  
 357 the forward transition matrix  $\mathbf{T}$  and its transpose  $\mathbf{T}^T$ , respectively.

#### 358 3.1. Forward and backward in time

359 As in the discrete case, consider the situation when the distribution of  
 360 the continuous allelic proportion  $x$  at time  $\tau = s$  is given by  $\rho(x)$ . Setting  
 361  $\phi(x|\tau = s) = \rho(x)$ ,  $\phi(x|\tau = 0, \rho)$  can be calculated using the forward dif-  
 362 fusion equation (29). Assume again a discrete sample of size  $M$  with a fre-  
 363 quency of  $y$  alleles of type one at time  $\tau = 0$ . In the backward time direction,  
 364  $\psi(y|x, \tau = 0) = \Pr(y|x, \tau = 0, M)$ , which corresponds to a binomial likelihood  
 365 as the allelic proportion is now assumed to be continuous. Note that a binomial  
 366 likelihood corresponds to a polynomial of order of the sample size  $M$  and is thus  
 367 finite. With the backward diffusion equation (29), the conditioning on  $x$  may

368 be moved backward in time. The marginal likelihood of  $y$  may be obtained by  
369 integration over the product of the forward and backward functions

$$\Pr(y|\rho) = \int_0^1 \phi(x|\tau, \rho)\psi(y|x, \tau) dx \quad \text{for } s \leq \tau \leq 0, \quad (30)$$

370 analogously to equation (6). As with the discrete case, we require the marginal  
371 likelihood to be constant irrespective of time. Furthermore, for any marginal  
372 likelihood of a discrete random variable  $0 \leq \Pr(y|\rho) \leq 1$  must hold. This  
373 constrains the boundary conditions.

374 As  $\Pr(y|\rho)$  is independent of time  $\tau$ , its derivative with respect to time  $\tau$   
375 must be 0. Exchanging the order of differentiation and integration and applying  
376 the product rule to  $\Pr(y|\rho)$ , we have

$$\begin{aligned} \frac{\partial}{\partial \tau} \Pr(y|\rho) &= 0 \\ \int_0^1 \left[ \frac{\partial}{\partial \tau} \phi(x|\tau, \rho) \right] \psi(y|x, \tau) dx + \int_0^1 \phi(x|\tau, \rho) \left[ \frac{\partial}{\partial \tau} \psi(y|x, \tau) \right] dx &= 0. \end{aligned} \quad (31)$$

377 Substituting the right sides of the forward and backward diffusion equations  
378 (29) for the time derivatives, we have the adjoint relationship

$$\begin{aligned} \int_0^1 [\mathcal{L} \phi(x|\tau, \rho)] \psi(y|\tau) dx &= \int_0^1 \phi(x|\tau, \rho) [\mathcal{L}^* \psi(y|x, \tau)] dx \\ \langle \mathcal{L} \phi(x|\tau, \rho), \psi(y|x, \tau) \rangle &= \langle \phi(x|\tau, \rho), \mathcal{L}^* \psi(y|x, \tau) \rangle. \end{aligned} \quad (32)$$

379 The adjoint relationship (32) requires the boundary condition (84) to hold (Ap-  
380 pendix 7.2). At each time-point, any change to the marginal likelihood from  
381 applying the forward operator  $\mathcal{L}$  to the forward function  $\phi(x|\tau, \rho)$  is exactly  
382 matched by a change from applying the backward operator  $\mathcal{L}^*$  to the back-  
383 ward function  $\psi(y|x, \tau)$ . As in the discrete case, the adjoint relationship allows  
384 movement forward and backward in time.

### 385 3.2. Self-Adjointness and Reversibility

386 In this section, we deal with reversible Markov processes. Introduce the  
387 weight or speed function [e.g., 10, 29]

$$w(x) = \frac{1}{Q(x)} e^{\int_0^x \frac{P(z)}{Q(z)} dz}. \quad (33)$$

388 Substituting  $w(x)g(x, \tau, \rho)$  for  $\phi(x|\tau, \rho)$ , the boundary condition (84) becomes  
389 (Appendix 7.2)

$$w(x)Q(x) \left( g(x, \tau, \rho) \frac{d}{dx} \psi(y|x, \tau) - \psi(y|x, \tau) \frac{d}{dx} g(x, \tau, \rho) \right) \Big|_0^1 = 0. \quad (34)$$

390 Since  $w(x)Q(x)$  may be infinite at the boundary,  $\psi(y|x, \tau)$  and  $g(x, \tau, \rho)$  need  
391 to be finite.



392 Assume  $w(x) > 0$  for  $x \in ]0, 1[$ , and substitute  $w(x)g(x, \tau, \rho)$  for  $\phi(x | \tau, \rho)$   
 393 into the general forward equation (29)

$$\begin{aligned} \frac{\partial}{\partial \tau} w(x)g(x, \tau, \rho) &= -\frac{\partial}{\partial x} P(x)w(x)g(x, \tau, \rho) + \frac{\partial^2}{\partial x^2} Q(x)w(x)g(x, \tau, \rho) \\ w(x) \frac{\partial}{\partial \tau} g(x, \tau, \rho) &= P(x)w(x) \frac{\partial}{\partial x} g(x, \tau, \rho) + Q(x)w(x) \frac{\partial^2}{\partial x^2} g(x, \tau, \rho) \quad (35) \\ \frac{\partial}{\partial \tau} g(x, \tau, \rho) &= P(x) \frac{\partial}{\partial x} g(x, \tau, \rho) + Q(x) \frac{\partial^2}{\partial x^2} g(x, \tau, \rho). \end{aligned}$$

394 Note that the last line is identical to the backward equation (29), with the  
 395 exception of the reversed sign to the left. Note that, nevertheless,  $\phi(x | \tau, \rho)$  may  
 396 be infinite. If the stationary distribution  $\pi(x)$  exists, it is proportional to  $w(x)$ .  
 397 From substituting  $\pi(x)g(x, \tau, \rho)$  for  $\phi(x | \tau, \rho)$  into the marginal likelihood (30),  
 398 it follows that  $g$  and  $\phi$  are square integrable with respect to the weight function  
 399  $\pi(x) \propto w(x)$  [29]. The Markov process is then self-adjoint and reversible and the  
 400 relationship between the forward operator  $\mathcal{L}$  and its adjoint  $\mathcal{L}^*$  may be written  
 401 compactly

$$\mathcal{L}^* = \frac{1}{\pi(x)} [\mathcal{L}\pi(x)], \quad (36)$$

402 similar to the reversed transition matrix (eq. 12) or to the condition of detailed  
 403 balance (eq. 13) in the discrete case.

### 404 3.3. Joint and conditional distributions

405 The function corresponding to the joint distribution of the allelic proportion  
 406  $x$  and the sample allele frequency  $y$  in the discrete case (8) at time  $\tau$  ( $s \leq \tau \leq 0$ )  
 407 is

$$j(x, y | \tau) = \phi(x | \tau, \rho) \psi(y | x, \tau). \quad (37)$$

408 For the conditional distribution of the allelic proportion  $x$  given the sample  
 409 allele frequency  $y$ , corresponding to eq. (9) in the discrete case,  $j(x, y | \tau)$  must  
 410 be divided by the marginal likelihood (30)

$$p(x | \tau, \rho, y) = \frac{j(x, y | \tau)}{\text{Pr}(y | \rho)}. \quad (38)$$

### 411 3.4. General mutation and drift and orthogonal polynomials

412 The diffusion operators in this section are as in (28), with  $P(x) = \theta(\alpha - x)$   
 413 and  $Q(x) = x(1 - x)$ . In population genetics,  $Q(x)$  is generally half the genetic  
 414 variance with the bi-allelic Moran model (see also Appendix 7.1). In the context  
 415 we consider, the backward function  $\psi(y | x, \tau)$  at time  $\tau = 0$  is a binomial  
 416 likelihood, *i.e.*, a polynomial of the degree of the sample size  $M$ . Without  
 417 selection, the backward function remains a polynomial with degree  $M$  for  $s \leq$   
 418  $\tau \leq 0$ .

419 With the general bi-allelic mutation-drift model, Song and Steinrücken [29]  
 420 already demonstrated self-adjointness and showed how to use modified Jacobi

421 polynomials to obtain a solution. For the general mutation-drift model, the  
 422 weight function  $w(x, \alpha, \theta) = x^{\alpha\theta-1}(1-x)^{\beta\theta-1}$  is proportional to the stationary  
 423 distribution

$$\pi(x) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} x^{\alpha\theta-1}(1-x)^{\beta\theta-1}. \quad (39)$$

424 Since  $Q(x) = x(1-x)$ , the boundary condition (34) holds if, at both boundaries  
 425  $x = 0$  and  $x = 1$ ,  $x(1-x)w(x) = 0$  and  $\psi(y|x, \tau)$  and  $g(x, \tau, \rho)$  are finite.  
 426 Since  $x(1-x)w(x) = x^{\alpha\theta}(1-x)^{\beta\theta}$  is zero at both boundaries for the non-  
 427 degenerate case of  $\theta > 0$  and  $0 < \alpha < 1$ , the boundary condition (34) holds  
 428 if  $\frac{\partial}{\partial x}(g(x, \tau, \rho)\psi(y|x, \tau))$  is finite at the boundaries, which can be assumed for  
 429 population genetic applications.

430 The (modified) Jacobi polynomials (compare formula 22.3.2 in Abramowitz  
 431 and Stegun [1])

$$R_n^{(\alpha, \theta)}(x) = \sum_{l=0}^n (-1)^l \frac{\Gamma(n-1+l+\theta)\Gamma(n+\alpha\theta)}{\Gamma(n-1+\theta)\Gamma(l+\alpha\theta)!(n-l)!} x^l \quad (40)$$

432 are eigenvectors of the backward operator

$$-\lambda_n R_n^{(\alpha, \theta)}(x) = \mathcal{L}^* R_n^{(\alpha, \theta)}(x), \quad (41)$$

433 with eigenvalues

$$\lambda_n = n(n+\theta-1). \quad (42)$$

434 The corresponding eigenfunctions of the forward operator are  $w(x)R_n^{(\alpha, \theta)}(x)$   
 435 with identical eigenvalues.

436 Since a binomial distribution with sample size  $M$  corresponds to a polyno-  
 437 mial of order  $M$ , the likelihood can be represented by an expansion with coef-  
 438 ficients  $c_n(y)$  into the modified Jacobi polynomials up to order  $M$ . Note that  
 439 a change in the effective population size (population demography), or equiva-  
 440 lently in the scaled mutation rate from  $\theta_a$  to  $\theta_c$  needs to be accommodated with  
 441 a change in the base from  $R_n^{(\alpha, \theta_a)}(x)$  to  $R_n^{(\alpha, \theta_c)}(x)$ .

442 The orthogonality relationship of the modified Jacobi polynomials is

$$\int_0^1 R_n^{(\alpha, \theta)}(x) R_m^{(\alpha, \theta)}(x) w(x) dx = \delta_{n,m} \Delta_n^{(\alpha, \theta)}, \quad (43)$$

443 where  $\delta_{n,m}$  is the Kronecker delta, and

$$\Delta_n^{(\alpha, \theta)} = \frac{\Gamma(n+\alpha\theta)\Gamma(n+\beta\theta)}{(2n+\theta-1)\Gamma(n+\theta-1)\Gamma(n+1)}. \quad (44)$$

444 Let  $c_n(y)$  be the coefficients of the expansion of the likelihood into the mod-  
 445 ified Jacobi polynomials, which breaks off at  $n = M$ . Then the solution to the  
 446 backward equation can be written as

$$\psi(y|x, \tau) = \sum_{n=0}^M c_n(y) R_n^{(\alpha, \theta)}(x) e^{-\lambda_n \tau}, \quad (45)$$

447 with  $\psi(y | x, \tau = 0) = \Pr(y | M, x)$  corresponding to the likelihood.

448 Let  $\rho_n$  be the coefficients of the expansion of the starting distribution  $\rho(x)$   
449 at time  $\tau = s$ . The solution to the forward equation can then be represented as

450

$$\phi(x | \tau, \rho) = w(x) \sum_{n=0}^{\infty} \rho_n R_n^{(\alpha, \theta)}(x) e^{-\lambda_n(s-\tau)}. \quad (46)$$

451 The orthogonality relationship can be used to simplify the marginal likeli-  
452 hood

$$\begin{aligned} \Pr(y | \rho) &= \int_0^1 \phi(x | \tau, \rho) \psi(y | x, \tau) dx \\ &= \int_0^1 \sum_{n=0}^M \rho_n c_n(y) w(x) [R_n^{(\alpha, \theta)}(x)]^2 e^{-\lambda_n \tau} e^{-\lambda_n(s-\tau)} dx \\ &= \sum_{n=0}^M \rho_n c_n(y) \Delta_n^{(\alpha, \theta)} e^{-\lambda_n s}. \end{aligned} \quad (47)$$

453 Because of the orthogonality relation (43), the calculation of the marginal  
454 likelihood (47) requires an expansion in eigenfunctions up to order  $M$ , where  
455  $M$  is the minimum of the forward-in-time expansion of  $\rho(x)$ , say  $M_f$ , and the  
456 backward-in-time expansion of  $\Pr(y | x, \tau = 0)$ , say  $M_b$ . Therefore, for calculating  
457 the joint distribution (37) and thus also the conditional (38), an expansion up to  
458 order  $M_f \times M_b$  is needed.

#### 459 3.4.1. Example: two splitting populations and binomial likelihoods

460 Here, we apply the theory to a model with two populations and binomial  
461 likelihoods; *i.e.*, a jSFS analogous to the second example in the discrete case  
462 (subsection 2.11). The initial distribution  $\rho(x)$  is assumed to be the equilibrium  
463 distribution. Only the first eigenfunction is necessary to expand the equilibrium  
464 distribution; *i.e.*,  $\rho_0 = \frac{1}{\Delta_0^{(\alpha, \theta)}}$  while  $\rho_{n \geq 1} = 0$ . In equilibrium, the marginal like-  
465 lihood of a single-population sample of size  $M$  assuming mutation-drift equilib-  
466 rium with parameters  $\alpha$  and  $\theta$  is a beta-binomial, as in the discrete case (25),

467

$$\begin{aligned} \Pr(y | M, \alpha, \theta) &= \int_0^1 \Pr(y | M, x) \pi(x, \alpha, \theta) dx \\ &= \int_0^1 \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} x^{\alpha\theta+y-1} (1-x)^{\beta\theta+M-y-1} dx \\ &= \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(y+\alpha\theta)\Gamma(M-y+\beta\theta)}{\Gamma(M+\theta)}. \end{aligned} \quad (48)$$

468 It follows from the orthogonality relation that only the first term in the ex-  
469 pansion  $n = 0$  contributes to the marginal likelihood, *i.e.*, the inner product

470

$$\begin{aligned}
\Pr(y | M, \alpha, \theta) &= \int_0^1 c_0(y) R_0^{(\alpha, \theta)}(x) \pi(x, \alpha, \theta) dx \\
&= \int_0^1 c_0(y) R_0^{(\alpha, \theta)}(x) \frac{1}{\Delta_0^{(\alpha, \theta)}} R_0^{(\alpha, \theta)}(x) x^{\alpha\theta-1} (1-x)^{\beta\theta-1} dx \quad (49) \\
&= c_0(y).
\end{aligned}$$

471 For two populations with sample sizes  $M^{(1)}$  and  $M^{(2)}$ , the respective likeli-  
472 hoods  $\Pr(y^{(1)} | M^{(1)})$  and  $\Pr(y^{(2)} | M^{(2)})$  are similarly expanded into the modi-  
473 fied Jacobi polynomials with coefficients  $c_n(y^{(1)})$  and  $c_m(y^{(2)})$ . At time  $\tau$  back  
474 in the past, we have

$$\Pr(y^{(1)} | x, M^{(1)}, \alpha, \theta, \tau) = \sum_{n=0}^{M^{(1)}} c_n(y^{(1)}) R_n^{(\alpha, \theta)}(x) e^{-\lambda_n \tau} \quad (50)$$

475 and similarly for the second population. If the two populations join at time  
476  $\tau = s$  in the past, when the population is assumed to be in mutation-drift  
477 equilibrium, the marginal likelihood is

$$\begin{aligned}
\Pr(y^{(1)}, y^{(2)} | M^{(1)}, M^{(2)}, \alpha, \theta, \tau = s) &= \sum_{n=0}^{M^{(1)}} \sum_{m=0}^{M^{(2)}} \int_0^1 c_n(y^{(1)}) R_n^{(\alpha, \theta)}(x) e^{-\lambda_n s} \\
&\quad \times c_m(y^{(2)}) R_m^{(\alpha, \theta)}(x) \pi(x, \alpha, \theta) e^{-\lambda_m s} dx \\
&= \sum_{n=0}^M \int_0^1 c_n(y^{(1)}) c_n(y^{(2)}) \left[ R_n^{(\alpha, \theta)}(x) \right]^2 \pi(x, \alpha, \theta) e^{-2\lambda_n s} dx \\
&= \sum_{n=0}^M \frac{c_n(y^{(1)}) c_n(y^{(2)}) \Delta_n^{(\alpha, \theta)} e^{-2\lambda_n s}}{\Delta_0^{(\alpha, \theta)}}, \quad (51)
\end{aligned}$$

478 where  $M = \min(M^{(1)}, M^{(2)})$ , since higher order terms contribute zero weight  
479 to the inner product.

480 A joint site frequency spectrum is drawn (Table 2) at the present time  $\tau = 0$ .  
481 Given the jSFS, the likelihood of the population split time is readily calculated  
482 (Figure 4). The jSFSs in Tables 1 and 2 are similar because scaled mutation  
483 rates and biases under which they are simulated are identical; for the discrete  
484 model, the population size is set to 20 instead of approaching infinity as in the  
485 continuous model, which, together with sampling variation, explains the slight  
486 differences.

#### 487 3.4.2. Summary: bi-allelic general mutation-drift diffusion

488 Assuming a bi-allelic general mutation-drift model, forward and backward  
489 diffusion equations and continuous analogs to the discrete forward and backward

Table 2: A jSFS simulated with a continuous diffusion model with parameters  $L = 10^5$ ,  $M^{(1)} = M^{(2)} = 3$ ,  $\alpha = 2/3$ ,  $\theta = 0.1$ , and  $s = -0.1$ .

$y$	0	1	2	3
0	28877	1447	494	231
1	1448	570	491	557
2	497	516	543	1491
3	253	521	1506	60558

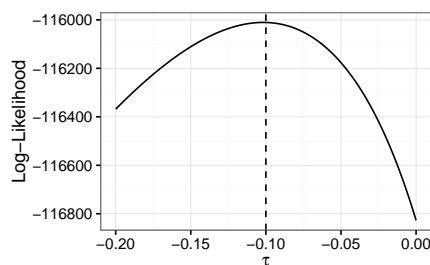


Figure 4: The log-likelihood of the split time  $s$ , given a jSFS (Table 2). The dashed line indicates the true split time.

490 algorithms, as well as the forward-backward algorithm, are derived. As with the  
 491 discrete models, it is straightforward to calculate exact likelihoods given a SFS  
 492 or a jSFS from the present. With the bi-allelic general mutation-drift model  
 493 a self-adjoint system results. Modified Jacobi polynomials  $R_n^{(\alpha, \theta)}(x)$  provide a  
 494 convenient base for calculations, both forward and backward in time. In the  
 495 discrete case, iteration of an  $(N + 1) \times (N + 1)$  transition matrix is needed to  
 496 evolve the allelic proportion; in the continuous case, only polynomials up to the  
 497 sample size  $M$  are needed with mutation-drift models. As  $M \ll N$ , this may  
 498 lead to considerably increased efficiency. A change in the effective population  
 499 size (population demography), or equivalently in the scaled mutation rate needs  
 500 to be accommodated with a change in the base of the orthogonal polynomials  
 501 as in Steinrücken *et al.* [32].

#### 502 4. Boundary mutation-drift model

503 In this section we deal with irreversible Markov processes. If mutation rates  
 504 are small relative to drift, polymorphism in a sample of moderate size originates  
 505 from a single mutation. We can therefore assume that mutations originate ex-  
 506 clusively from sites fixed for allele zero or one, *i.e.*, from the boundaries. Such  
 507 models are particularly important for statistical inference in population genet-  
 508 ics [*e.g.*, 9, 39, 12] and it is therefore worthwhile to provide solutions to the  
 509 corresponding diffusion equations. As a solution to the forward and backward  
 510 diffusion equations we present a system of orthogonal eigenfunctions. Through-

511 out the presentation, we emphasize the similarities with previous approaches.  
 512 While the solution to the forward diffusion is mainly a review, the backward  
 513 direction and the overall concepts are new.

#### 514 4.1. Pure drift model

515 We start with the pure drift model and clarify basic concepts. The forward  
 516 and backward diffusion operators are

$$\begin{aligned} \mathcal{L} &= \frac{\partial^2}{\partial x^2} Q(x) \\ \mathcal{L}^* &= Q(x) \frac{\partial^2}{\partial x^2}. \end{aligned} \quad (52)$$

517 For the pure drift model, the adjoint relationship between the forward and  
 518 backward operators holds as long as the boundary condition (84) with  $Q =$   
 519  $x(1-x)$  holds within the unit interval

$$0 = \left( x(1-x)\phi\psi' - (x(1-x)\phi)'\psi \right) \Big|_0^1. \quad (53)$$

520 Following Kimura [15], most population geneticists implicitly or explicitly re-  
 521 quire at both boundaries  $\psi(y|x, \tau)$  and  $x(1-x)\phi(x|\tau, \rho)$  to be zero [see also  
 522 10, 29]. With these assumptions, modified Gegenbauer polynomials  $U_n(x) =$   
 523  $-\frac{2}{n}C_{n-2}^{(3/2)}(2x-1)$  ( $C_k^\nu(z)$  are the Gegenbauer polynomials as defined in [1]) are  
 524 eigenfunctions of the forward diffusion equation with eigenvalues  $\lambda_n = n(n-1)$   
 525 for  $n \geq 2$ . Furthermore  $x(1-x)U_n(x)$  are eigenfunctions of the backward equa-  
 526 tion with identical eigenvalues. The forward and backward operators are then  
 527 self-adjoint with the weight function  $w(x) = x^{-1}(1-x)^{-1}$  [10, 29]. Note that  
 528 without mutation no stationary distribution exists. The orthogonality relation  
 529 of  $U_n(x)$  is

$$\int_0^1 U_n(x)U_m(x)w(x) dx = \delta_{n,m}\Delta_n, \quad (54)$$

530 with

$$\Delta_n = \frac{n-1}{(2n-1)n}. \quad (55)$$

531 However, these assumptions are too restrictive; polynomials of zeroth and  
 532 first degree, 1 and  $x$ , cannot be represented by  $x(1-x)U_n(x)$ , but both are  
 533 eigenfunctions of the pure drift backward equation with eigenvalues  $\lambda_0 = \lambda_1 = 0$ .  
 534 Importantly, assuming a binomial likelihood, these eigenfunctions are needed  
 535 when representing monomorphic samples. To address this issue, Tran *et al.*  
 536 [33] add 1 and  $x$  to the eigenfunctions of the backward equation. The two new  
 537 backward eigenfunctions require augmenting the forward eigenfunctions with  
 538 point masses at the boundaries that counterbalance the probability mass in the  
 539 interior. Additionally, point masses at the boundaries, independent of those  
 540 associated with the forward eigenfunctions, need to be introduced [33].

541 Independently from Tran *et al.* [33], we derived a boundary mutation-drift  
 542 model forward in time from probabilistic population genetic considerations [35]  
 543 with eigenfunctions proportional to those in Tran *et al.* [33]. Our approach  
 544 is similar to that presented in McKane and Waxman [23] and Waxman [40].  
 545 Furthermore, we showed that the forward eigenfunctions can be derived from  
 546 those of the general mutation-drift model, *i.e.*, from Jacobi polynomials times  
 547 the stationary beta distribution (or the proportional weight function  $w(x, \alpha, \theta)$ ),  
 548 by expanding into a Taylor series in  $\theta$  and keeping terms up to order zero [36,  
 549 Appendix A.1]. Therefore, in the context of pure drift, the set of eigenfunc-  
 550 tions, which provide the solution to the forward diffusion equation, can then be  
 551 represented in relation to Jacobi polynomials  $R_n^{(\alpha, \theta)}$  as

$$\begin{cases} F_0^{(\alpha, 0)}(x) &= \lim_{\theta \rightarrow 0} \pi(x, \alpha, \theta) = \beta \delta(x) + \alpha \delta(x-1) \\ F_1^{(\alpha, 0)}(x) &= \lim_{\theta \rightarrow 0} w(x, \alpha, \theta) R_1^{(\alpha, \theta)} = -\delta(x) + \delta(x-1) \\ F_{n \geq 2}^{(\alpha, 0)}(x) &= \lim_{\theta \rightarrow 0} w(x, \alpha, \theta) R_n^{(\alpha, \theta)} = -\frac{(-1)^n}{n} \delta(x) + U_n(x) - \frac{1}{n} \delta(x-1), \end{cases} \quad (56)$$

552 where  $\delta(x)$  is the Dirac delta functional. Note that eigenfunctions are only  
 553 defined up to a proportionality constant. The associated eigenvalues are

$$\begin{cases} \lambda_0 &= 0 \\ \lambda_1 &= \lim_{\theta \rightarrow 0} \theta = 0 \\ \lambda_{n \geq 2} &= n(n-1). \end{cases} \quad (57)$$

554 Similarly, the backward eigenfunctions can be derived by expanding the  
 555 modified Jacobi polynomials into a Taylor series in  $\theta$  and keeping terms up to  
 556 order zero.

$$\begin{cases} B_0^{(\alpha, 0)}(x) &= R_0^{(\alpha, \theta)} = 1 \\ B_1^{(\alpha, 0)}(x) &= \frac{1}{\theta} R_1^{(\alpha, \theta)} = x - \alpha \\ B_{n \geq 2}^{(\alpha, 0)}(x) &= \lim_{\theta \rightarrow 0} R_n^{(\alpha, \theta)} = x(1-x)U_n(x). \end{cases} \quad (58)$$

557 The eigenvalues correspond to those forward in time in eq. (57). The mutation  
 558 bias  $\alpha$  may obtain any value between zero and one. If  $\alpha$  is set to zero, the  
 559 backward eigenfunctions correspond to those of Tran *et al.* [33].

560 The orthogonality relation is

$$\int_0^1 F_n^{(\alpha, 0)}(x) B_m^{(\alpha, 0)}(x) dx = \delta_{n,m} \Delta_n, \quad (59)$$

561 with  $\Delta_0 = \Delta_1 = 1$  and  $\Delta_n$  as in (55). However, note that

$$\int_0^1 B_n^{(\alpha, 0)}(x) B_m^{(\alpha, 0)}(x) w(x) dx = \delta_{n,m} \Delta_n \quad (60)$$

562 only holds for pairs  $m, n \geq 2$  and the pair  $m = 0$  and  $n = 1$ , but not for the  
 563 pairs  $m = 0$  (or  $m = 1$ ) and  $n \geq 2$ ; and similarly for the forward eigenfunctions  
 564  $F_n(x)$ .

565 The forward function is then set to

$$\phi(x | \tau, \rho) = \sum_{n=0}^{\infty} \rho_n F_n^{(\alpha,0)}(x) e^{-\lambda_n(s-\tau)} \quad (61)$$

566 and the backward function to

$$\psi(y | x, \tau) = \sum_{m=0}^{\infty} c_m(y) B_m^{(\alpha,0)}(x) e^{-\lambda_m \tau}. \quad (62)$$

567 The marginal and joint distribution can now be defined as above. The time  
 568 derivative of the marginal likelihood (31) of the eigenfunctions with  $n = 0$   
 569 and  $n = 1$  is zero, because the respective eigenvalues are zero. For  $n \geq 2$ ,  
 570 the backward expansion contains only the terms  $x(1-x)U_n(x)$  as does  $w(x)$   
 571 times the forward expansion,  $w(x)F_{n \geq 2}^{(\alpha,0)}(x) = x(1-x)U_{n \geq 2}(x)$ . Indeed the  
 572 eigenfunctions with  $n \geq 2$  correspond to those usually considered [15, 29]. As  
 573 backward and forward functions are thus zero at both boundaries, the boundary  
 574 condition (53) is met. It is also straightforward to show for  $n = 0$  and  $n = 1$   
 575 that condition (32) holds, because the integrals on both sides are always zero.

#### 576 4.2. Mutation-drift model

577 Following Vogl and Bergman [36], we introduce recurrent mutations into the  
 578 pure drift model by setting the eigenvalue  $\lambda_1 = \theta$ . We consider the case where  
 579  $0 < \theta \ll 1$ , such that mutations occur at a low rate and thus, do not affect the  
 580 allele frequency dynamics of the polymorphic classes; these classes are governed  
 581 exclusively by genetic drift and therefore, eigenfunctions with  $n \geq 2$  remain as  
 582 in the pure drift model. We may thus distinguish between two classes of sites  
 583 with distinct spatial and temporal differences: the slowly evolving boundaries,  
 584 where the rate of evolution depends on  $\theta$ , and the fast evolving polymorphic  
 585 classes governed by genetic drift [*e.g.*, 42, 36]. Furthermore, we may think of  
 586 the boundary mutation-drift model as a first order Taylor series expansion in  
 587 the scaled mutation rate  $\theta$  of the general mutation-drift model.

588 Note that, with the discrete boundary mutation model, we scaled the mu-  
 589 tation rate such that, independent of the population size  $N$ , the heterozygosity  
 590 in a sample of size two is equal to  $\theta$  for the model with mutations from a single  
 591 boundary (compare the term  $\mu/(1 - \theta \sum_{i=1}^{N-1} \frac{1}{i})$  in (18)), or  $2\alpha\beta\theta$  for the model  
 592 with mutations from both boundaries. With the transition to continuous dif-  
 593 fusion,  $N \rightarrow \infty$  and thus  $\theta \sum_{i=1}^{N-1} \frac{1}{i}$  will grow logarithmically without bound.  
 594 Mutations are therefore modeled from the boundary zero at a rate  $\alpha\theta b_0(\tau)$ ,  
 595 where  $\alpha\theta$  is the mutation rate towards allele one and  $b_0(\tau)$  corresponds to the  
 596 probability mass already at boundary zero plus the probability mass to arrive  
 597 there quickly by drift, and similarly at the boundary one. The system is thus  
 598 not in detailed balance and therefore not reversible.

599 *Forward expansion.* With mutations from the boundaries and forward in time,  
 600 Vogl and Bergman [36] use the same augmented forward eigenfunctions as with



601 pure drift (56) to model the spatial part of the eigensystem. With pure drift, the  
 602 temporal parts of the eigenfunctions ( $e^{-\lambda_n(s-\tau)}$ ) with  $n \geq 2$  fulfill homogeneous  
 603 differential equations, *i.e.*, are decreasing exponentially from starting values at  
 604 rates  $\lambda_n = n(n-1)$ , while the first two eigenfunctions with  $n = 0$  and  $n = 1$  do  
 605 not change with time. With the boundary mutation model, the temporal part  
 606  $T_n(\tau)$  corresponds to a system of linear differential equations: homogeneous for  
 607  $n = 0$  and  $n = 1$  with eigenvalues  $\lambda_0 = 0$  and  $\lambda_1 = \theta$ , and inhomogenous for  
 608  $n \geq 2$  with eigenvalues  $\lambda_n = n(n-1)$ :

$$\begin{cases} \frac{d}{d\tau} T_0(\tau) &= 0 \\ \frac{d}{d\tau} T_1(\tau) &= -\theta T_1(\tau) \\ \frac{d}{d\tau} T_{n \geq 2}(\tau) &= -\lambda_n T_n(\tau) + \vartheta E_n T_0(\tau) + \theta O_n T_1(\tau), \end{cases} \quad (63)$$

609 with

$$\begin{aligned} \vartheta &= \alpha\beta\theta, \\ E_n &= -(n-1) \frac{((-1)^n + 1)}{\Delta_n}, \\ O_n &= -(n-1) \frac{(-1)^n \alpha - \beta}{\Delta_n}, \end{aligned} \quad (64)$$

610 where  $\beta = (1 - \alpha)$  and  $\Delta_n$  as in (55).

611 The forward system can be diagonalized by setting

$$\begin{cases} F_0^{(\alpha, \theta)}(x) &= F_0^{(\alpha, 0)}(x) + \vartheta \sum_{n=2}^{\infty} \frac{E_n}{\lambda_n} F_n^{(\alpha, 0)}(x) \\ F_1^{(\alpha, \theta)}(x) &= F_1^{(\alpha, 0)}(x) + \theta \sum_{n=2}^{\infty} \frac{O_n}{\lambda_n} F_n^{(\alpha, 0)}(x) \\ F_{n \geq 2}^{(\alpha, \theta)}(x) &= F_n^{(\alpha, 0)}(x), \end{cases} \quad (65)$$

612 where the polynomials with base  $(\alpha, 0)$  on the right hand side of the equations  
 613 are as in (56). The temporal parts of the system are then  $\frac{d}{d\tau} T_n(\tau) = -\lambda_n T_n(\tau)$   
 614 for all  $n$ .

With increasing  $N$ , the stationary distribution converges to the following  
 function [35, 36]

$$\pi(x, \alpha, \theta) = F_0^{(\alpha, \theta)}(x) = \lim_{N \rightarrow \infty} \begin{cases} \beta - \vartheta \int_{\frac{1}{N}}^{\frac{N-1}{N}} \frac{1}{x} dx & \text{if } 0 \leq x < 1/N \\ \vartheta \frac{1}{x(1-x)} & \text{if } 1/N \leq x \leq 1 - 1/N \\ \alpha - \vartheta \int_{\frac{1}{N}}^{\frac{N-1}{N}} \frac{1}{1-x} dx & \text{if } 1 - 1/N < x \leq 1. \end{cases} \quad (66)$$

615 This function integrates to unity, but has singularities at the boundaries, which  
 616 makes it difficult to interpret probabilistically. Moments about zero up to an  
 617 order  $m = M_{\max}$  may be defined meaningfully, by multiplying  $\pi(x, \alpha, \theta)$  with

618  $x^m$  and integrating. We have

$$\begin{aligned} \int_0^1 \pi(x)x^m dx &= \alpha - \vartheta \int_0^1 \frac{1-x^{m-1}}{1-x} dx \\ &= \alpha - \vartheta H_{m-1}, \end{aligned} \quad (67)$$

619 where  $H_{m-1}$  is the harmonic number. As this same relationship must also hold  
620 for the moments about boundary one,  $\min(\alpha, \beta)/\vartheta < H_{m-1}$ , which leads to  
621  $M_{\max} \approx e^{\min(\alpha, \beta)/\vartheta}$ . Note that a monomorphic sample from a binomial distri-  
622 bution, with sample size  $M$ , leads to terms  $x^M$  or  $(1-x)^M$ , which correspond to  
623 the moments about zero and one. Thus the sample size needs to be restricted to  
624  $M \approx e^{\min(\alpha, \beta)/\vartheta}$  to avoid negative values for probabilities. Since the boundary  
625 mutation model generally requires  $\theta < 0.1$  [35], this constraint on  $M$  should not  
626 pose practical problems.

627 Note that the same issue occurs with the closely related Ewens-Watterson  
628 estimator  $\hat{\theta}_W$  of molecular diversity [9, 39]. With the assumptions used for  
629 deriving  $\hat{\theta}_W$ , the probability of obtaining a monomorphic sample of size  $M$   
630 is  $1 - \theta \sum_{i=1}^{M-1} \frac{1}{i}$ . It is therefore necessary to restrict the sample size below  
631  $M_{\max} \approx e^{1/\theta}$ .

632 *Backward expansion.* The backward system of differential equations with eigen-  
633 functions  $B_n^{(\alpha, \theta)}(x)$  is the transpose of the forward system (65). It can also be  
634 diagonalized by setting

$$\begin{cases} B_0^{(\alpha, \theta)}(x) &= B_0^{(\alpha, 0)}(x) = 1 \\ B_1^{(\alpha, \theta)}(x) &= B_1^{(\alpha, 0)}(x) = x - \alpha \\ B_{n \geq 2}^{(\alpha, \theta)}(x) &= B_n^{(\alpha, 0)}(x) - \vartheta \frac{E_n \Delta_n}{\lambda_n} B_0^{(\alpha, 0)}(x) - \theta \frac{B_n \Delta_n}{\lambda_n} B_1^{(\alpha, 0)}(x). \end{cases} \quad (68)$$

635 It can be verified that the forward and backward eigenfunctions fulfil the  
636 orthogonality relation (59) with  $\Delta_0 = \Delta_1 = 1$  and  $\Delta_n$  as in (55). In particular,  
637 for  $n = 0$  and  $m \geq 2$ , we have

$$\begin{aligned} \int_0^1 F_0^{(\alpha, \theta)}(x) B_m^{(\alpha, \theta)}(x) dx &= \int_0^1 \left( F_0^{(\alpha, 0)}(x) + \vartheta \sum_{n=2}^{\infty} \frac{E_n}{\lambda_n} F_n^{(\alpha, 0)}(x) \right) \\ &\times \left( B_m^{(\alpha, 0)}(x) - \vartheta \frac{E_m \Delta_m}{\lambda_m} B_0^{(\alpha, 0)}(x) - \theta \frac{O_m \Delta_m}{\lambda_m} B_1^{(\alpha, 0)}(x) \right) dx \\ &= \vartheta \frac{E_m}{\lambda_m} \Delta_m - \vartheta \frac{E_m \Delta_m}{\lambda_m} \Delta_0 = 0, \end{aligned} \quad (69)$$

638 and similarly for  $m = 1$  and  $n \geq 2$ .

639 Furthermore, we have, as before, the forward function

$$\phi(x | \tau, \rho) = \sum_{n=0}^{\infty} \rho_n F_n^{(\alpha, \theta)}(x) T_n(\tau), \quad (70)$$

640 and the backward function

$$\psi(y | x, \tau) = \sum_{n=0}^{\infty} c_n(y) B_n^{(\alpha, \theta)}(x) T_n(\tau). \quad (71)$$

641 The backward function and the marginal distribution, as long as  $M < M_{\max} \approx$   
 642  $e^{\min(\alpha, \beta)/\vartheta}$ , can be interpreted probabilistically as with the general mutation-  
 643 drift or the pure drift model. As the forward function may attain negative  
 644 values, expanding it beyond the sample size  $M$  has little meaning.

#### 645 4.2.1. Example: one change in the mutation parameters

646 We present the version of the boundary mutation model with the inhomogeneous linear differential equations, *i.e.*, with the eigenfunctions  $F_0^{(\alpha, 0)}$  and  $B_n^{(\alpha, 0)}$ . With this choice, a change in the effective population size (population demography), or equivalently in the scaled mutation rate does not necessitate a change in the base. Assume a population in equilibrium at  $\tau = s$  with mutation parameters  $\theta_a$  and  $\alpha_a$ , such that the initial distribution is  $\rho(x) = \pi(x | \theta_a, \alpha_a)$ . The scaled mutation parameters then changes immediately to  $\theta$  and  $\alpha$ , respectively, and remain constant thereafter. Expanding the stationary distribution at time  $\tau = s$  into the forward eigenfunctions  $F_n^{(\alpha, 0)}(x)$  results in

$$\begin{aligned} \phi(x | \tau = s) &= F_0^{(\alpha, 0)}(x) + (\alpha_a - \alpha) e^{-\theta \tau} F_1^{(\alpha, 0)}(x) \\ &+ \sum_{n=2}^{\infty} \left( E_n(\vartheta + (\vartheta_a - \vartheta) e^{-\lambda_n(s-\tau)} \right. \\ &\left. + (\alpha_a - \alpha) \theta O_n(e^{-\theta(s-\tau)} - e^{-\lambda_n(s-\tau)}) \right) F_n^{(\alpha, 0)}(x). \end{aligned} \quad (72)$$

655 With a sample of size  $M$  with  $y$  alleles of the first type at time  $\tau = 0$ , the  
 656 binomial likelihood can be expanded into the backward eigenfunctions with

$$\psi(y | x, \tau = 0) = \sum_{n=0}^M c_n(y) B_n^{(\alpha, 0)}(x) \quad (73)$$

657 The marginal likelihood, calculated at time  $\tau = 0$ , is

$$\begin{aligned} \Pr(y) &= \int_0^1 \phi(x | \tau = 0, \rho) \psi(y | x, \tau = 0) dx = \left[ c_0(y) \cdot 1 \right] + \left[ c_1(y) (\alpha_a - \alpha) e^{-\theta s} \cdot 1 \right] \\ &+ \left[ \sum_{n=2}^M c_n(y) \left( E_n(\vartheta + (\vartheta_a - \vartheta) e^{-\lambda_n s}) \right. \right. \\ &\left. \left. + (\alpha_a - \alpha) \theta O_n(e^{-\theta s} - e^{-\lambda_n s}) \right) \cdot \Delta_n \right], \end{aligned} \quad (74)$$

658 where the terms in the successive square brackets come from the terms in the  
 659 expansion with  $n = 0$ ,  $n = 1$ , and  $2 \leq n \leq M$ , respectively, while all terms with

660  $n > M$  are zero. Within the square brackets, the terms before the dot are the  
 661 time-dependent functions of the forward expansion. The same marginal likeli-  
 662 hood is also obtained by using the backward eigenfunctions  $B_n^{(\alpha,0)}$ , multiplying  
 663 with the stationary distribution at  $\tau = s$ , and integrating:

$$\begin{aligned}
 \Pr(y) &= \int_0^1 \psi(y | x, \tau = s) \pi(x, \alpha_a, \theta_a) dx \\
 &= \left[ \left( c_0(y) + \vartheta \sum_{n=2}^M c_n(y) E_n \Delta_n (1 - e^{-n(n-1)s}) \right) \cdot 1 \right] \\
 &\quad + \left[ \left( c_1(y) e^{-\theta s} + \theta \sum_{n=2}^M c_n(y) E_n \Delta_n (e^{-\theta s} - e^{-\lambda_n s}) \right) \cdot (\alpha_a - \alpha) \cdot 1 \right] \\
 &\quad + \left[ \sum_{n=2}^M c_n(y) e^{-\lambda_n s} \cdot \vartheta_a E_n \Delta_n \right].
 \end{aligned} \tag{75}$$

664 Within the square brackets, the terms before the dot are the time-dependent  
 665 functions of the backward expansion. The two different versions of the marginal  
 666 likelihoods evaluated at  $\tau = 0$  and  $\tau = s$  are identical.

#### 667 4.2.2. Summary: boundary mutation-drift diffusion

668 Assuming a bi-allelic boundary mutation-drift model, a system of orthogonal  
 669 eigenfunctions is defined. As with Jacobi polynomials for the general mutation-  
 670 drift model, these functions provide a convenient base for calculations. While  
 671 some mathematical inconvenience compared to the modified Jacobi polynomials  
 672 is encountered, changes in the (effective) population size (*i.e.*,  $\theta$ ) are easily  
 673 accommodated, because the base of the polynomials need not be changed. As  
 674 with the general mutation-drift model, efficiency is increased compared to the  
 675 discrete models since only eigenfunction expansions up to order  $M$  instead of  
 676  $N$  are needed.

### 677 5. The order of the expansion

678 With bi-allelic diffusion models we naturally assumed a binomial likelihood.  
 679 This likelihood function corresponds to a polynomial of the order of the sample  
 680 size  $M$ . Both with the general mutation-drift model as with the boundary  
 681 mutation-drift model only orthogonal polynomials up to the order of the sample  
 682 size are needed when modeling the allele trajectory backward in time. We also  
 683 note that a change in the base of the polynomials, because the scaled mutation  
 684 parameters changed, does not change the order of the expansion.

685 Now consider two populations with sample sizes  $M^{(1)}$  and  $M^{(2)}$ . Tracing  
 686 back the allele frequency evolution to the split time requires a polynomial expan-  
 687 sion of up to  $\max(M^{(1)}, M^{(2)})$ . Integrating over the population allelic propor-  
 688 tion to obtain the marginal likelihood of the data at the split time then requires  
 689 multiplication with the starting distribution, which can also be expanded into

690 orthogonal polynomials of order  $M_a$ . If the starting population is in equilib-  
 691 rium, then  $M_a = 0$ . If we first multiply the starting distribution with the  
 692 backward orthogonal expansion of the smaller population, we obtain a forward  
 693 expansion of order at least  $M_a + \min(M^{(1)}, M^{(2)})$ . Because of the orthogonality  
 694 relation, when multiplying with the backward expansion of the second popula-  
 695 tion, only polynomials of order up to the minimum of  $M_a + \min(M^{(1)}, M^{(2)})$   
 696 and  $\max(M^{(1)}, M^{(2)})$  are needed for obtaining the marginal likelihood. Thus  
 697 the maximal expansion needed depends on the sample sizes and the starting dis-  
 698 tribution, but is always at least  $\min(M^{(1)}, M^{(2)})$  and at most  $\max(M^{(1)}, M^{(2)})$ .  
 699 Therefore, the required degree of the polynomial expansion is considerably less  
 700 than previously thought necessary [21, 20]. Similar considerations also apply to  
 701 more than two populations, where it can be shown that the required expansion  
 702 to obtain the marginal likelihood is less than the sum of the sample sizes.

## 703 6. Discussion

704 Starting from bi-allelic mutation-drift models, we use forward and backward  
 705 processes in discrete or continuous time to efficiently calculate probabilities of  
 706 population allele proportions. Given a sample from a single population, *i.e.*, a  
 707 SFS, or samples from more than one population, *i.e.*, a jSFS, from the present,  
 708 this theory may be used to infer trajectories of population allele frequencies  
 709 in the past. Integrating over the population allelic proportion, the marginal  
 710 likelihood of the data may be used to infer population genetic parameters.  
 711 The discrete-time algorithm is a variant of the forward-backward algorithm  
 712 and thus makes use of dynamic programming. The continuous time algorithm  
 713 uses orthogonal polynomials for even more convenient calculation. Further-  
 714 more, we introduce bi-allelic population genetic models that provide us with  
 715 time-reversible and irreversible transition matrices or kernels. The irreversible  
 716 models are related to the infinite site [16, 8] or Poisson-random-field models  
 717 [26]. Both reversible and irreversible models have stationary distributions.

718 Previous diffusion-based methods for inference of population genetic param-  
 719 eters are generally based on modelling allelic proportion trajectories forward-  
 720 in-time. Solutions to the forward diffusion equations are either approximated  
 721 numerically [*e.g.*, 4, 12, 22] or are provided as functions of orthogonal poly-  
 722 nomials [*e.g.*, 21, 20, 29, 31]. These methods can, in principle, accommodate many  
 723 demographic scenarios while considering general selection and continuous migra-  
 724 tion. The complexity of these models in combination with the forward-in-time  
 725 approach often results in complex likelihood functions. Herein, we demonstrate  
 726 that combining forward- and backward-in-time approaches naturally leads to  
 727 relatively simple likelihood functions for both discrete and continuous popula-  
 728 tion genetics models (compare eqs. 16 and 30, respectively).

729 Discrete models involve repeated multiplications with a transition matrix  
 730 of dimension  $(N + 1) \times (N + 1)$ , where  $N$  is the haploid population size. For  
 731 biological reasons,  $N$  should be large to model the large (effective) population  
 732 sizes usually encountered. For numerical reasons,  $N$  should be small, because

733 iteration of large matrices is time-consuming and numerical errors may accumu-  
 734 late. Mutation rates can be scaled to account for a reduction of  $N$ . Transition  
 735 matrices may be diagonalized to speed up calculations. In any case,  $N$  must  
 736 be at least as big as the sample size  $M$  to not lose information. A prior distri-  
 737 bution must be assumed at some time in the past. If this distribution is taken  
 738 as the stationary distribution of the transition matrix, calculations simplify. At  
 739 the present time, a probability model of the sampling process, generally a hy-  
 740 pergeometric likelihood, is assumed that is conditional on the sample size  $M$ .  
 741 Zhao *et al.* [46] present a similar method that is also based on the iteration of a  
 742 transition matrix (in their case, based on the Wright-Fisher model) and allows  
 743 for conditioning on the beginning and end states of the chain. They derive the  
 744 marginal distribution of states intermediate in the chain and simulate trajecto-  
 745 ries. Extending this method to distributions instead of states (in our case, the  
 746 prior at the beginning and the likelihood at the end of the chain) requires ad-  
 747 ditional considerations and diagonalizing the transition matrix seems necessary  
 748 in all but the simplest cases.

749 With continuous diffusion models, the use of orthogonal polynomials is con-  
 750 venient. The degree of the polynomials need not be higher than the sample size  
 751  $M$ , while the population size is large, which usually fits biological reality. Thus,  
 752 the diffusion approach is mostly preferable over the discrete approach.

753 Song and colleagues [29, 30, 31, 48] analyse self-adjoint continuous models,  
 754 such as the general mutation-drift model herein. These authors usually take a  
 755 Dirac delta function as starting condition instead of a prior distribution at  $\tau = s$   
 756 (but see Supplemental Information, Section D in Steinrücken *et al.* [31]). Repre-  
 757 sentation of a Dirac delta function requires an infinite expansion and modeling  
 758 an arbitrary distribution as starting condition would require a further step (see  
 759 Appendix 7.3). As these authors also consider selection, eigenfunctions with, in  
 760 principle, infinite expansions are necessary in any case. A problem with their  
 761 approach for pure drift models, however, is the restriction at the boundaries,  
 762 which allows only polymorphic samples to be analyzed (see the subsection 4.1).  
 763 Interestingly, Zhao *et al.* [45] also present a diffusion approach to calculate con-  
 764 ditional trajectories that involves the product of solutions of the forward and  
 765 backward equations. They consider a Dirac delta function as starting state  
 766 and, additionally, also as a final state. Usually in population genetics, however,  
 767 only a sample from the present is given, while the starting conditions are even  
 768 less well defined. Applying this approach to real data thus requires integration  
 769 over possible starting and final states, which adds another layer of complex-  
 770 ity avoided with our approach. In contrast, Lukić and Hey [20] also use the  
 771 equilibrium distribution as a starting condition as with the approach presented  
 772 herein.

773 Generally, using a delta function as an initial condition requires an infinite  
 774 expansion in orthogonal polynomials. Yet for calculating marginal likelihoods a  
 775 much lower expansion is needed. Lukić and Hey [20], citing [26], set the degree  
 776 of polynomial expansion to  $(M - 2)^K$ , where  $M$  is the number of haplotypes  
 777 sampled and  $K$  the number of populations. Yet we show that only an expan-  
 778 sion between  $\min(M^{(1)}, M^{(2)})$  and  $\max(M^{(1)}, M^{(2)})$  is needed, where  $M^{(1)}$  and

779  $M^{(2)}$  are the sample sizes in the two populations. With additional populations,  
 780 the expansion needed is less than  $\sum_{i=1}^K M_i$ . Furthermore, these authors use  
 781 Chebyshev polynomials, which are not orthogonal with respect to the forward  
 782 and backward operators. This necessitates numerical integration of a linear sys-  
 783 tem of differential equations to obtain the temporal part of the solution. With  
 784 orthogonal polynomials, the corresponding system of differential equations is  
 785 diagonal and thus much simpler.

786 An analysis also involving a coupled system of ordinary differential equa-  
 787 tions for the temporal evolution of moments [8, 47, 48] also provides solutions  
 788 for the forward and backward diffusions. The basic model analyzed by these  
 789 authors is the continuous version of the single-boundary mutation-drift model  
 790 presented here, where ancestral and derived alleles are differentiated. Zivkovic  
 791 and Stephan [47] also point out relations of the backward approach to coales-  
 792 cent theory. Recently, a diffusion framework of weak mutation and selection  
 793 has been incorporated in the theoretical analysis of adaptive landscapes [42], a  
 794 concept first formulated by Wright [41].

795 We note that many approaches above [8, 21, 20, 47, 48, 36] use boundary  
 796 mutation models. Indeed, much of the statistics of population genetics is based  
 797 on this model, *e.g.*, the important Ewens-Watterson  $\theta$  [9, 39]. For this model,  
 798 only the forward transition probabilities have been given so far [8, 21, 36]. For  
 799 the first time, we give the backward system of orthogonal polynomials and their  
 800 corresponding eigenvalues herein. The system of eigenfunctions of the pure drift  
 801 model [33] follows as a special case. As explained above, the possibility to move  
 802 backward simplifies inference.

803 The demographic scenarios presented here (Fig. 1) are common, *e.g.*, in nat-  
 804 ural populations of fruit flies of the *Drosophila* genus [*e.g.*, 19, 43, 24]. Addi-  
 805 tionally, the abundance of population data for *Drosophila* species makes them  
 806 especially suitable for SFS and jSFS analysis under the described framework.  
 807 Furthermore, the theory can be extended to more than two populations, *i.e.*, to  
 808 phylogenetic inference. Our methods can also be adjusted to an experimental  
 809 setting with samples from multiple time points, as *e.g.*, in evolve-and-resequence  
 810 experiments [18]. Furthermore, a setting with multiple time-points also applies  
 811 to the analysis of ancient DNA samples as noted by Steinrücken *et al.* [31].

812 Generally, the methods and models we present in this article are simple, yet  
 813 allow for maximum marginal likelihood analysis of SFS and jSFS from split-  
 814 ting populations with mutation-drift or pure drift models, and for inference of  
 815 evolutionary trajectories of population allele proportions conditional on data.

## 816 Acknowledgments

817 The authors thank Reinhard Bürger, Joachim Hermisson and other col-  
 818 leagues from the Faculty of Mathematics of the University of Vienna, all mem-  
 819 bers of the Institute of Population Genetics at the University of Veterinary  
 820 Medicine, Vienna, and Andreas Futschik (Johannes Kepler Universtiy, Linz).  
 821 All authors were supported by the Austrian Science Fund (FWF): DK W1225-  
 822 B20. DS and CK were partially funded by FWF-P24551-B25. CK has been

823 partially funded by the Vienna Science and Technology Fund (WWTF) through  
824 project MA16-061.

## 825 References

- 826 [1] Abramowitz, M. and Stegun, I., editors (1970). *Handbook of Mathematical*  
827 *Functions*. Dover, 9th ed. edition.
- 828 [2] Baake, E. and Bialowons, R. (2008). Ancestral processes with selection:  
829 branching and Moran models. *Banach center publications*, **80**, 33–52.
- 830 [3] Bayin, S. (2006). *Mathematical methods in science and engineering*. Wiley,  
831 N.Y.
- 832 [4] Bollback, J. P., York, T. L., and Nielsen, R. (2008). Estimation of  $2N_e s$   
833 from temporal allele frequency data. *Genetics*, **179**(1), 497–502.
- 834 [5] Carlin, B. and Louis, T. (2000). *Bayes and empirical Bayes methods*. Chap-  
835 man and Hall, 2nd ed. edition.
- 836 [6] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological se-*  
837 *quence analysis*. Cambridge University Press, Cambridge.
- 838 [7] Etheridge, A. and Griffiths, R. (2009). A coalescent dual process in a Moran  
839 model with genic selection. *Theoretical Population Biology*, **75**, 320–330.
- 840 [8] Evans, S., Shvets, Y., and Slatkin, M. (2007). Non-equilibrium theory of the  
841 allele frequency spectrum. *Theoretical Population Biology*, **71**, 109–119.
- 842 [9] Ewens, W. (1974). A note on the sampling theory for infinite alleles and  
843 infinite sites models. *Theoretical Population Biology*, **6**, 143–148.
- 844 [10] Ewens, W. (2004). *Mathematical Population Genetics*. Springer, N.Y., 2nd  
845 edition.
- 846 [11] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum  
847 likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.
- 848 [12] Gutenkunst, R., Hernandez, R., Williamson, S., and Bustamante, C.  
849 (2009). Inferring the Joint Demographic History of Multiple Populations  
850 from Multidimensional SNP Frequency Data. *PLoS Genetics*, **5**, e1000695.
- 851 [13] Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene genealogies, variation,*  
852 *and evolution: a primer in coalescent theory*. Oxford University Press.
- 853 [14] Jewett, E. M., Steinrücken, M., and Song, Y. S. (2016). The effects of  
854 population size histories on estimates of selection coefficients from time-series  
855 genetic data. *Molecular biology and evolution*, **33**(11), 3002–3027.
- 856 [15] Kimura, M. (1955). Solution of a process of random genetic drift with a  
857 continuous model. *Proc. Natl. Acad. Sci. USA*, **41**, 144–150.



- 858 [16] Kimura, M. (1969). The number of heterozygous nucleotide sites main-  
 859 tained in a finite population due to steady flux of mutations. *Genetics*, **61**,  
 860 893–903.
- 861 [17] Kingman, J. (1982). On the genealogy of large populations. *Journal of*  
 862 *Applied Probability*, **19A**, 27–43.
- 863 [18] Kofler, R. and Schlötterer, C. (2014). A guide for the design of evolve and  
 864 resequencing studies. *Molecular Biology and Evolution*, **31**, 474–483.
- 865 [19] Li, H. and Stephan, W. (2006). Inferring the Demographic History and  
 866 Rate of Adaptive Substitution in *Drosophila*. *PLOS Genetics*, **10**, e166.
- 867 [20] Lukić, S. and Hey, J. (2012). Demographic inference using spectral meth-  
 868 ods on SNP data, with an analysis of the human out-of-Africa expansion.  
 869 *Genetics*, **192**(2), 619–639.
- 870 [21] Lukić, S., Hey, J., and Chen, K. (2011). Non-equilibrium allele frequency  
 871 spectra via spectral methods. *Theoretical population biology*, **79**(4), 203–219.
- 872 [22] Malaspinas, A.-S., Malaspinas, O., Evans, S. N., and Slatkin, M. (2012).  
 873 Estimating allele age and selection coefficient from time-serial data. *Genetics*,  
 874 **192**(2), 599–607.
- 875 [23] McKane, A. and Waxman, D. (2007). Singular solutions of the diffusion  
 876 equation of population genetics. *Journal of Theoretical Biology*, **247**, 849–  
 877 858.
- 878 [24] Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno,  
 879 C. M., Crepeau, M. W., Duchon, P., Emerson, J. J., Saelao, P., Begun, D. J.,  
 880 and Langley, C. H. (2012). Population genomics of sub-saharan *Drosophila*  
 881 *melanogaster*: African diversity and non-African admixture. *PLOS Genet*,  
 882 **8**(12), e1003080.
- 883 [25] Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov  
 884 models. *IEEE ASSP magazine*, **3**, 4–16.
- 885 [26] Sawyer, S. and Hartl, D. (1992). Population genetics of polymorphism and  
 886 divergence. *Genetics*, **132**, 1161–1176.
- 887 [27] Schraiber, J. G., Evans, S. N., and Slatkin, M. (2016). Bayesian inference of  
 888 natural selection from allele frequency time series. *Genetics*, **203**(1), 493–511.
- 889 [28] Schrepf, D., Minh, B. Q., De Maio, N., von Haeseler, A., and Kosiol,  
 890 C. (2016). Reversible polymorphism-aware phylogenetic models and their  
 891 application to tree inference. *Journal of Theoretical Biology*, **407**, 362–370.
- 892 [29] Song, Y. and Steinrücken, M. (2012). A simple method for finding ex-  
 893 plicit analytic transition densities of diffusion processes with general diploid  
 894 selection. *Genetics*, **190**, 1117–1129.

- 895 [30] Steinrücken, M., Wang, R., and Song, Y. (2013). An explicit transition den-  
896 sity expansion for a multi-allelic WrightFisher diffusion with general diploid  
897 selection. *Theoretical Population Biology*, **83**, 1–14.
- 898 [31] Steinrücken, M., Bhaskar, A., and Song, Y. (2014). A novel method for  
899 inferring general diploid selection from time series genetic data. *Annals of*  
900 *Applied Statistics*, **8**, 2203–2222.
- 901 [32] Steinrücken, M., Jewett, E. M., and Song, Y. S. (2015). SpectralTDF:  
902 transition densities of diffusion processes with time-varying selection param-  
903 eters, mutation rates and effective population sizes. *Bioinformatics*, **32**(5),  
904 795–797.
- 905 [33] Tran, T., Hofrichter, J., and Jost, J. (2013). An introduction to the math-  
906 ematical structure of the Wright-Fisher model of population genetics. *Theory*  
907 *in Biosciences*, **132**, 73–82.
- 908 [34] Vogl, C. (2014). Estimating the Scaled Mutation Rate and Mutation Bias  
909 with Site Frequency Data. *Theoretical Population Biology*, **98**, 19–27.
- 910 [35] Vogl, C. and Bergman, J. (2015). Inference of directional selection and  
911 mutation parameters assuming equilibrium. *Theoretical Population Biology*,  
912 **106**, 71–82.
- 913 [36] Vogl, C. and Bergman, J. (2016). Computation of the likelihood of joint  
914 site frequency spectra using orthogonal polynomials. *Computation*, **4**, 6.
- 915 [37] Vogl, C. and Futschik, A. (2010). Hidden markov models in biology. In  
916 O. Carugo and F. Eisenhaber, editors, *Biological Data Mining*, Methods in  
917 Molecular Biology. Humana Press.
- 918 [38] Wakeley, J. (2009). *Coalescent theory, an Introduction*. Roberts and Co.
- 919 [39] Watterson, G. (1975). On the number of segregating sites in genetical  
920 models without recombination. *Theoretical Population Biology*, **7**, 256–276.
- 921 [40] Waxman, D. (2011). Comparison and content of the WrightFisher model  
922 of random genetic drift, the diffusion approximation, and an intermediate  
923 model. *Journal of Theoretical Biology*, **269**, 79–87.
- 924 [41] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and  
925 selection in evolution. *Proceedings of the sixth international congress of ge-*  
926 *netics*, **1**, 356–366.
- 927 [42] Xu, S., Jiao, S., Jiang, P., and Ao, P. (2014). Two-time-scale population  
928 evolution on a singular landscape. *Physical Review E*, **89**(1), 012724.
- 929 [43] Zeng, K. and Charlesworth, B. (2010). Studying patterns of recent evo-  
930 lution at synonymous sites and intronic sites in *Drosophila melanogaster*.  
931 *Journal of Molecular Evolution*, **183**, 651–662.

- 932 [44] Zhao, L., Lascoux, M., Overall, A., and Waxman, D. (2013a). The charac-  
933 teristic trajectory of a fixing allele: a consequence of fictitious selection that  
934 arises from conditioning. *Genetics*, **195**, 993–1006.
- 935 [45] Zhao, L., Yue, X., and Waxman, D. (2013b). Complete numerical solution  
936 of the diffusion equation of random genetic drift. *Genetics*, **194**, 419–426.
- 937 [46] Zhao, L., Yue, X., and Waxman, D. (2014). Exact solution of conditioned  
938 Wright-Fisher models. *Journal of Theoretical Biology*, **194**, 973–985.
- 939 [47] Zivkovic, D. and Stephan, W. (2011). Analytical results on the neutral non-  
940 equilibrium allele frequency spectrum based on diffusion theory. *Theoretical*  
941 *Population Biology*, **79**, 184–191.
- 942 [48] Zivkovic, D., Steinrücken, M., Song, Y., and Stephan, W. (2015). Transi-  
943 tion densities and sample frequency spectra of diffusion processes with selec-  
944 tion and variable population size. *Genetics*, **200**, 601–617.

945 **7. Appendices**

946 *7.1. Derivation of the forward and backward diffusion equations from the de-*  
 947 *coupled general mutation-drift Moran model*

948 In this appendix, we derive the forward and backward diffusion equation  
 949 from the forward and backward transition probabilities of the decoupled Moran  
 950 model with general mutation and drift and show the tight connections between  
 951 the discrete and continuous models. Derivations are simpler than usual [10];  
 952 terms higher than the first derivative with respect to time and second derivative  
 953 with respect to space do not occur.

954 Consider a focal bi-allelic site with the population frequency of allele one  
 955 denoted by  $i$  ( $1 \leq i \leq N-1$ ). With the transition probabilities of the decoupled  
 956 Moran model (24), the frequency  $i$  may increase or decrease by one due to  
 957 mutation or drift, or remain constant. Forward in time, the difference of the  
 958 probability at frequency  $i$  per Moran step may be written as

$$\begin{aligned}
 \Pr(x_{t+1} = i) - \Pr(x_t = i) = & \\
 & \frac{\alpha\theta}{N^2} \left( (N-i+1) \Pr(x_t = i-1) - (N-i) \Pr(x_t = i) \right) \\
 & + \frac{\beta\theta}{N^2} \left( (i+1) \Pr(x_t = i+1) - i \Pr(x_t = i) \right) \\
 & + \frac{1}{N^2} \left( (i-1)(N-i+1) \Pr(x_t = i-1) \right. \\
 & \left. + (i+1)(N-i-1) \Pr(x_t = i+1) - 2i(N-i) \Pr(x_t = i) \right), \tag{76}
 \end{aligned}$$

959 where the term within the first pair of square brackets corresponds to mutation  
 960 towards allele one, the term within the second pair to mutation towards allele  
 961 zero, and the term within the third pair to genetic drift.

962 To approximate the change in frequency as a process in continuous time  
 963 and space, the quantities  $\delta\tau = 1/N^2$  and  $\delta x = 1/N$  are introduced. Further-  
 964 more, time is rescaled as  $\tau = t\delta\tau$ , the allele proportions as  $x = i\delta x$ , such that  
 965  $\phi(x|\tau, \rho)\delta\tau\delta x = \Pr(x_t = i)$ . Taking the limit  $N \rightarrow \infty$ , eq. (76) is rewritten as

$$\begin{aligned}
 \lim_{N \rightarrow \infty} \frac{\phi(x|\tau + \delta\tau, \rho) - \phi(x|\tau, \rho)}{\delta\tau} = & \\
 \lim_{N \rightarrow \infty} \left[ \alpha\theta \left( \frac{(1-x+\delta x)\phi(x-\delta x|\tau, \rho) - (1-x)\phi(x|\tau, \rho)}{\delta x} \right) \right. & \\
 + \beta\theta \left( \frac{(x+\delta x)\phi(x+\delta x|\tau, \rho) - x\phi(x|\tau, \rho)}{\delta x} \right) & \\
 + \left( \frac{(x-\delta x)(1-x+\delta x)\phi(x-\delta x|\tau, \rho)}{\delta x^2} \right. & \\
 \left. + \frac{(x+\delta x)(1-x-\delta x)\phi(x+\delta x|\tau, \rho)}{\delta x^2} - \frac{2x(1-x)\phi(x|\tau, \rho)}{\delta x^2} \right) & \left. \right]. \tag{77}
 \end{aligned}$$

966 The term to the left of the equality sign of (77) corresponds to the definition  
 967 of the first derivative with respect to time  $\tau$  of  $\phi(x|\tau, \rho)$ ; the terms with muta-  
 968 tions correspond to the first derivatives with respect to  $x$  of  $-(1-x)\phi(x|\tau, \rho)$   
 969 and  $x\phi(x|\tau, \rho)$ , respectively; the drift term corresponds to the definition of the  
 970 second symmetric derivative with respect to  $x$  of  $x(1-x)\phi(x|\tau, \rho)$ . After minor  
 971 rearrangements, the familiar form of the forward general mutation-drift diffusion  
 972 equation is obtained

$$\frac{\partial}{\partial \tau} \phi(x|\tau, \rho) = -\frac{\partial}{\partial x} \theta(\alpha - x)\phi(x|\tau, \rho) + \frac{\partial^2}{\partial x^2} x(1-x)\phi(x|\tau, \rho). \quad (78)$$

973 Considering the Moran model backward in time (see Subsection 2.4), the  
 974 change in frequency  $i$  back in time is determined by the transpose of the forward  
 975 transition matrix (24) and can be written as

$$\begin{aligned} \Pr(y|x_t = i) - \Pr(y|x_{t+1} = i) &= \\ &= \frac{\alpha\theta(N-i)}{N^2} \left( \Pr(y|x_{t+1} = i+1) - \Pr(y|x_{t+1} = i) \right) \\ &+ \frac{\beta\theta i}{N^2} \left( \Pr(y|x_{t+1} = i-1) - \Pr(y|x_{t+1} = i) \right) \\ &+ \frac{i(N-i)}{N^2} \left( \Pr(y|x_{t+1} = i+1) + \Pr(y|x_{t+1} = i-1) \right. \\ &\left. - 2\Pr(y|x_{t+1} = i) \right). \end{aligned} \quad (79)$$

976 After rescaling time and space, considering the limit  $N \rightarrow \infty$ , and setting  
 977  $\psi(y|x, \tau) = \Pr(y|x_{t+1} = i)$ , we get the backward diffusion equation

$$-\frac{\partial}{\partial \tau} \psi(y|x, \tau) = \theta(\alpha - x) \frac{\partial}{\partial x} \psi(y|x, \tau) + x(1-x) \frac{\partial^2}{\partial x^2} \psi(y|x, \tau). \quad (80)$$

978 The minus sign on the left side of the backward diffusion equation (80) may  
 979 be unusual [compare 10], but necessary such that the time  $\tau$  runs in the same  
 980 direction in the forward and backward diffusion. Note that Zhao *et al.* [44] also  
 981 use a pair of forward and backward diffusion equations with differing signs.

## 982 7.2. Boundary condition

983 In the following, we use the prime ( $'$ ) to indicate the (partial) derivative with  
 984 respect to  $x$  and leave away the terms in brackets for  $\phi$  and  $\psi$ . Eq. (32) can  
 985 then be written as

$$\int_0^1 [-(P\phi)' + (Q\phi)'] \psi dx = \int_0^1 \phi [P\psi' + Q\psi''] dx. \quad (81)$$

986 The first term on the right side is

$$\int_0^1 \phi P \psi' dx = \phi P \psi \Big|_0^1 - \int_0^1 (\phi P)' \psi dx \quad (82)$$

987 and the second term on the right side is

$$\begin{aligned} \int_0^1 \phi Q \psi'' dx &= \phi Q \psi' \Big|_0^1 - \int_0^1 (Q\phi)' \psi' dx \\ &= \phi Q \psi' \Big|_0^1 - (\phi Q)' \psi \Big|_0^1 + \int_0^1 (Q\psi)'' \psi dx, \end{aligned} \quad (83)$$

988 Hence for eq. (81) to hold, we require the boundary condition

$$(\phi Q \psi' - (\phi Q)' \psi + \phi P \psi) \Big|_0^1 = 0. \quad (84)$$

989 Using the weight function  $w(x)$  defined in formula (33), this condition can be  
990 represented more compactly. The weight function fulfils

$$Pw = (wQ)'. \quad (85)$$

991 Substitute  $\phi(x|\tau) = w(x)g(x, \tau, \rho)$  into eq. (84) to obtain

$$\begin{aligned} 0 &= (wQg\psi' - (wQg)' \psi + Pw g \psi) \Big|_0^1 \\ &= (wQg\psi' - ((wQ)g' + wQg') \psi + Pw g \psi) \Big|_0^1 \\ &= (wQg\psi' - Pw g \psi - wQg' \psi + Pw g \psi) \Big|_0^1 \\ &= wQ(g\psi' - g' \psi) \Big|_0^1 \end{aligned} \quad (86)$$

992 Note that  $w(x)Q(x) \propto 1/\xi(x)$  where  $\xi(x)$  is the scale function defined in eq. (2)  
993 of Song and Steinrücken [29] and  $g(x)$  and  $\psi(x)$  correspond to  $f(x)$  in Song and  
994 Steinrücken [29]. This condition obviously holds if, at both boundaries, either  
995  $w(x)Q(x) = 0$  while  $(g\psi' - g'\psi)$  is finite, or  $(g\psi' - g'\psi) = 0$  while  $w(x)Q(x)$  is  
996 finite.

### 997 7.3. Propagator

998 Song and Steinrücken [29] analyze self-adjoint differential equations, with  
999 a Dirac delta function  $\delta(x-p)$  as starting point at  $\tau = s$ . Denote the eigen-  
1000 functions of the diffusion equation with the backward operator  $\mathcal{L}^*$  with  $B_n(x)$ .  
1001 Eq. (5) of Song and Steinrücken [29] defines a “propagator” [3, chap. 19]

$$p(x|p, \tau) = \sum_{n=0}^{\infty} e^{-\lambda_n \tau} \pi(x) \frac{B_n(x) B_n(p)}{\langle B_n(x) B_n(p) \rangle_\pi} \quad (87)$$

1002 as the solution of the diffusion equation with a starting state modeled by the  
1003 Dirac Delta function  $\delta(x-p)$ . If the starting condition is not a particular state  
1004 but, more usually, a distribution  $\rho(p)$ , the function

$$h(x|p, \rho, \tau) = \int_0^1 p(x|p, \tau) \rho(p) dp \quad (88)$$

1005 solves the diffusion equation. From the orthogonality relation it is evident that,  
1006 also with this indirect route, only an expansion of degree  $M$  is needed for cal-  
1007 culating the marginal likelihood.