computation

MDPI

Article

Computation of the Likelihood of Joint Site Frequency Spectra Using Orthogonal Polynomials

Claus Vogl [1,*] **and Juraj Bergman** [2,3]

[1] Institute of Animal Breedings and Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria
[2] Institut für Populationsgenetik, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria; juraj.bergman@vetmeduni.ac.at
[3] Vienna Graduate School of Population Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria
* Correspondence: claus.vogl@vetmeduni.ac.at; Tel.: +43-1-25077-5631

Academic Editor: Rainer Breitling
Received: 28 September 2015; Accepted: 26 January 2016; Published: 4 February 2016

Abstract: In population genetics, information about evolutionary forces, e.g., mutation, selection and genetic drift, is often inferred from DNA sequence information. Generally, DNA consists of two long strands of nucleotides or sites that pair via the complementary bases cytosine and guanine (C and G), on the one hand, and adenine and thymine (A and T), on the other. With whole genome sequencing, most genomic information stored in the DNA has become available for multiple individuals of one or more populations, at least in humans and model species, such as fruit flies of the genus *Drosophila*. In a genome-wide sample of L sites for M (haploid) individuals, the state of each site may be made binary, by binning the complementary bases, e.g., C with G to C/G, and contrasting C/G to A/T, to obtain a "site frequency spectrum" (SFS). Two such samples of either a single population from different time-points or two related populations from a single time-point are called joint site frequency spectra (joint SFS). While mathematical models describing the interplay of mutation, drift and selection have been available for more than 80 years, calculation of exact likelihoods from joint SFS is difficult. Sufficient statistics for inference of, e.g., mutation or selection parameters that would make use of all the information in the genomic data are rarely available. Hence, often suites of crude summary statistics are combined in simulation-based computational approaches. In this article, we use a bi-allelic boundary-mutation and drift population genetic model to compute the transition probabilities of joint SFS using orthogonal polynomials. This allows inference of population genetic parameters, such as the mutation rate (scaled by the population size) and the time separating the two samples. We apply this inference method to a population dataset of neutrally-evolving short intronic sites from six DNA sequences of the fruit fly *Drosophila melanogaster* and the reference sequence of the related species *Drosophila sechellia.*

Keywords: bi-allelic mutation-drift model; small-scaled mutation rate; orthogonal polynomials; transition probability

1. Introduction

Evolutionary forces, e.g., mutation, selection and genetic drift, shape DNA sequence information. Typically, the evolutionary processes that have influenced the data reach back millions of generations or years. Mathematical theory that describes these processes has been available for more than 80 years (e.g., [1,2]), yet inference of population genetic parameters using probabilistic models is difficult, and only few analytical maximum-likelihood estimators are available; those based on diffusion theory, so far, assume independence among sites and are briefly reviewed in Vogl [3], Vogl and Bergman [4] and in the theory section below.

A DNA molecule is a string (or strand) of nucleotides (or sites) that usually pairs with a complementary strand to form a double-stranded chromosome. Pairing of sites is accomplished by hydrogen bonds between the complementary base pairs adenine (A) and thymine (T), on the one hand, and cytosine (C) and guanine (G), on the other. An assortment of chromosomes forms a genome, which is specific for a species. The main functional units of genomes are genes that often code for proteins. Proteins provide structure, catalyze metabolism or mediate physiological pathways in all living organisms. While the single-celled Bacteria and Archaea generally have compact genomes, genes of the more complex eukaryotic organisms are often interrupted by non-coding introns. Introns are spliced out, *i.e.*, eliminated, during maturation of the messenger RNA, which is then translated into the chain of amino acids that makes up proteins.

A point mutation at a certain nucleotide or site creates a new genetic variant, *i.e.*, an allele. Mutation is not strand-specific, but may be biased towards A or T (A/T) over C or G (C/G) or *vice versa*, because mutation rates between these two allelic classes may vary. Genome-wide sequence data may be made bi-allelic (binary), by considering A/T nucleotides as Allele 0 and C/G nucleotides as Allele 1. This simplifies mathematical analysis, such that maximum-likelihood inference becomes possible. Mutations introduce new variants into the genome and, thus, increase genomic variation. Conversely, stochastic fluctuations of the allelic proportion due to finite population sizes, *i.e.*, random genetic drift, eventually cause fixation of an allelic type, thus eliminating variation. An equilibrium between mutation and drift may establish with time.

Recently, relatively inexpensive, high throughput DNA sequencing methods have made available population data from whole genomes (in multicellular organisms typically comprising $10^7 - 10^{11}$ sites), at least for humans and model species, such as fruit flies of the genus *Drosophila*. These data provide the basis for inference of population genetic forces, such as random genetic drift, the mutation rate scaled by the population size, directional selection and the time of the split between two populations.

In this article, we focus on inferring population genetic parameters using a mutation-drift model of the allelic proportion x. For mathematical convenience, genomic sites are classified as binary with respect to their nucleotide (C/G *vs.* A/T). A total of L sites are classified into categories, depending on the count y of Allele 1 (C/G) among M aligned genomic sequences. Together, these counts form a site frequency spectrum (SFS) of size $(M + 1)$ with $0 \leq y \leq M$. Joint SFS may be constructed considering the allelic states of sites within a single population at two different time points or two related populations at a single time point.

The solution of the diffusion equation describing the evolution of x conditional on mutation and drift parameters has previously been represented as a series expansion of orthogonal polynomials (e.g., [5–9]). In this article, we extend the mathematical theory to a boundary-mutation model [4], which describes the evolution of x when the scaled mutation rate θ is small, *i.e.*, on the order 0.1 or smaller [10]. Using this model, a method for the inference of θ and the time of split t is derived and applied to both simulated and empirical *Drosophila* population data. The empirical data are joint SFS of short introns, as the nucleotide composition of this site class is considered to not be affected by selection, but only by the joint forces of mutation and drift [11–13]. Therefore, the study of these sites likely provides an accurate estimate of the population demography and the genome-wide scaled mutation rate. The joint SFS from a sample of six individuals from the Malawian *D. melanogaster* population [14] and the *D. sechellia* reference sequence (Release 1.0; [15]) is used to infer mutation and drift parameters.

1.1. Inference with a Single Site Frequency Spectrum Assuming Equilibrium

Assume that a sample of L genomic loci or sites is available for M haploid individuals. The sample space of the allelic count for each locus l is then $y_l = (0, \ldots, M)$ copies of Allele 1, with $1 \leq l \leq L$. In regions of high recombination rates relative to mutation rates, sites may be assumed to be independently and identically distributed, such that the probabilities given the model parameters of all L sites can be multiplied. In this case, the theory developed below can be considered maximum likelihood. In regions of relatively low recombination rates, estimators are still consistent and may be considered a composite likelihood. For notational convenience, the index l is often dropped in the following.

1.2. Inference Based on the Beta Equilibrium Distribution

In a classical study, Wright [2] proposed a model for the evolution of a bi-allelic locus under the influence of the population genetic forces: mutation, directional selection and drift. He also derived the equilibrium distribution of the allelic proportion, conditional on the scaled mutation rate, the mutation bias and the scaled strength of directional selection. In the absence of selection, the equilibrium distribution of the population allelic proportion x of Allele 1 is a beta:

$$p(x|\alpha,\theta) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{\alpha\theta-1}(1-x)^{\beta\theta-1}\,, \tag{1}$$

where α is the mutation bias towards Allele 1 and $\beta = (1-\alpha)$ and θ the overall scaled mutation rate, *i.e.*, the product of the per-generation mutation rate μ and the effective population number or size N.

Conditional on x, the distribution of the allelic count y is assumed to be binomial. Especially with genome-wide samples, the allelic proportions at particular sites are not interesting and "integrated out", which leads to the beta-binomial distribution of the allelic count:

$$p(y\,|\,\alpha,\theta) = \binom{M}{y}\frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\frac{\Gamma(\alpha\theta+y)\Gamma(\beta\theta+M-y)}{\Gamma(\theta+M)}\,. \tag{2}$$

Given a sample of L independent loci for M individuals for each locus and a common α and θ, let L_y represent the counts of sites with y alleles of Type 1. Set $q_y = p(y\,|\,\alpha,\theta)$. The likelihood is then:

$$\ell(L_0,\ldots,L_M\,|\,\alpha,\theta,L) = \frac{L!}{L_0!\cdots L_M!}\, q_0^{L_0}\cdots q_M^{L_M}\,. \tag{3}$$

For arbitrarily large values of θ, only iterative algorithms have been derived to obtain maximum likelihood estimates of α and θ [3], even in the simple case without selection. Note that this model corresponds to the canonical model of the empirical Bayes method, and maximizing this marginal likelihood corresponds to a parametric empirical Bayes approach [16].

In the limit of small θ, the beta-binomial compound distribution (2) can be expanded into a Taylor series in θ at $\theta = 0$, up to first order. The "folded" site frequency spectrum (folded SFS) is derived from the general site frequency spectrum by lumping the samples L_y with L_{M-y}, such that the state space becomes $0 \le y \le [M/2]$ per locus. For a polymorphic sample of such a folded SFS, the Taylor series expansion in θ of the beta-binomial compound distribution has been derived by RoyChoudhury and Wakeley [17]. In the general situation with $0 \le y \le M$, the series expansion of the beta-binomial compound distribution leads to the general "RoyChoudhury–Wakeley" distribution [3]:

$$p(y\,|\,\alpha,\theta) = \begin{cases} \beta - \alpha\beta\theta\sum_{y=1}^{M-1}\frac{1}{y} & \text{for } y = 0,\\ \alpha\beta\theta\frac{M}{y(M-y)} = \alpha\beta\theta\,(\frac{1}{y}+\frac{1}{M-y}) & \text{for } 1 \le y \le M,\\ \alpha - \alpha\beta\theta\sum_{y=1}^{M-1}\frac{1}{y} & \text{for } y = M. \end{cases} \tag{4}$$

With this first order expansion of the beta-binomial in θ, approximate maximum likelihood (ML) estimators of α and θ and their posterior distributions can be obtained easily [3]. In particular, the approximate ML estimator for the scaled mutation rate is:

$$\hat{\vartheta} = \frac{L_p}{2L\sum_{y=1}^{M-1} 1/y}\,, \tag{5}$$

where $\hat{\vartheta} = \hat{\alpha}\hat{\beta}\hat{\theta}$ and $L_p = \sum_{y=1}^{M-1} L_y$, *i.e.*, the sum over all polymorphic sites in the sample, while the approximate ML estimator for α is:

$$\hat{\alpha} = \frac{L_M + L_p/2}{L}\,. \tag{6}$$

If the boundary mutation model is assumed, these estimators are maximum likelihood, rather than approximations in the limit of small-scaled mutation rates θ [4].

1.3. Inference Based on the Assumptions of Equilibrium and Rare Mutations

The estimator $\hat{\vartheta}$ of Formula (5) is a variant of the well-known Ewens–Watterson estimator of the scaled mutation rate [18,19], $\hat{\theta}_w = L_p/(L\sum_{y=1}^{M-1} 1/y)$. The latter was originally derived assuming the infinite sites model [20,21], which in turn was based on a model with irreversible mutation [2]. With the infinite sites model, infinitely many sites may be hit by mutation at a finite rate, such that each site is hit only once [19,21]. Furthermore, it is usually assumed that the ancestral and derived allelic states can be inferred with outgroup information, *i.e.*, with information from closely-related species or populations (e.g., [12,13,22–24]). Then, segregating mutations are assumed to only arise once from the ancestral background. Alleles at a site are thus not defined as having bases A/T *versus* C/G, but as being ancestral *versus* derived. Note that the factor of two difference between $\hat{\vartheta}$ of Formula (5) and the Ewens–Watterson estimator reflects that mutations arise from both boundaries in the former and only one boundary in the latter. Obviously, the mutation bias cannot be inferred with polarization. Yet, irrespective of whether or not data are polarized, a polymorphic site is scored as polymorphic.

The Ewens–Watterson estimator is generally unbiased. If sites are unlinked, it can be shown that it is also the maximum likelihood estimator of θ and, thus, a sufficient statistic (e.g., [3]). Ewens [18] neglected to show this explicitly, while he earlier showed that the estimator for the corresponding infinite alleles model is maximum likelihood [25]. Furthermore, it can been shown that the estimators are unbiased (e.g., [3,18]). Note that if assumptions are met, $\hat{\theta}_w$ corresponds to the "expected heterozygosity", *i.e.*, the expected proportion of polymorphism in a sample of size $M = 2$.

Similar to the infinite sites model, applications of the Poisson random field (PRF) model to population genetics generally assume small mutation rates. The PRF theory is often based on irreversible mutation models and, like the infinite sites model of Kimura [21], usually assumes the presence of directional selection. For an equilibrium distribution to exist, an inexhaustible and unvarying supply of sites must be assumed. Furthermore, the ancestral state of all sites must be known without errors and conditioned on. This is because, as discussed above, the rates of mutation from A/T to C/G generally differ, and the force of directional selection is reversed if an A/T mutates to a C/G or *vice versa*. The above assumptions are not met with real datasets: genomes are finite, and inference of the ancestral state is error-prone. Nevertheless, if appropriate outgroup information is available and quasi-equilibrium is assumed, the approach is sensible [26–29].

While RoyChoudhury and Wakeley [17] also use the PRF approach, they do not assume outgroup information, but rather start from a Taylor series expansion in θ of the equilibrium beta distribution (1). As shown above (Equation (5)), the estimator RoyChoudhury and Wakeley [17] derived is essentially identical to the Ewens–Watterson estimator of θ. Starting from a Moran model that only allows for mutations from the boundaries, Vogl and Clemente [10] derive a generalization of the estimator also for the case with directional selection, without assuming outgroup information. Vogl and Bergman [4] derive ML estimators for all three parameters: mutation bias α, scaled mutation rate θ and scaled selection strength γ, with the same assumptions, but base the analysis on a diffusion model.

2. Mathematical Theory and Algorithms

The Ewens–Watterson estimator $\hat{\theta}_w$ [18,19] or its varieties are sufficient statistics for the analysis of site frequency spectra assuming the infinite sites model, equilibrium and unlinked sites. With real datasets, however, changes in demography or mutation parameters usually invalidate the equilibrium assumption. Moreover, the approach to equilibrium is dominated by the scaled mutation rate θ. Since θ is often on the order of 10^{-2} per unit of diffusion time, which is scaled in N generations, it takes on the order of $100 \cdot N$ generations to reach equilibrium. This is on the order of $10^7 - 10^9$ generations. Even with the short-lived fruit flies, equilibrium is thus usually not reached before a change in population demography, the selection regime or the mutation bias. For probabilistic analysis of datasets that have not yet reached equilibrium, calculation of transition probabilities or densities is necessary. This is also necessary for joint site frequency spectra, where samples are drawn from a single population at two different time points or two closely-related populations at a single time point, which we will present in this article.

Consider a population of haploid population size $N(t)$, where t is time. The dynamics are governed by only two population genetic forces: mutation and drift. Generally, the diffusion limit, *i.e.*, $N \to \infty$, is considered such that, at each time point only two quantities matter: the scaled mutation rate $\theta(t) = \mu(t)N(t)$ and the mutation bias $\alpha(t) = (1 - \beta(t)) = \mu_1(t)/(\mu_0(t) + \mu_1(t))$. Let $x(t)$ denote the proportion of the first allelic type, which in our case may be identified with the proportion of C/G at this site in the population at time t. Assume now that the parameters $N(t)$, $\mu(t)$ and $\alpha(t)$ are piecewise constant and consider only a single such epoch of constant parameters. Usually, the following forward operator is obtained [3,30]:

$$\mathcal{L}_f = \frac{\partial^2}{\partial x^2} x(1-x) - \frac{\partial}{\partial x}\theta(\alpha - x)\,. \tag{7}$$

The corresponding forward diffusion or Kolmogorov equation is:

$$\frac{\partial}{\partial t}\phi(x,t) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\phi(x,t)\Big) - \frac{\partial}{\partial x}\Big(\theta(\alpha - x)\phi(x,t)\Big)\,, \tag{8}$$

where $\phi(x,t)$ is the transition density of the allelic proportion x at any time t. To solve this equation, Song and Steinrücken [7] employ a series expansion with the modified Jacobi polynomials:

$$R_i^{(\theta,\alpha)}(x) = P_i^{(\beta\theta-1,\alpha\theta-1)}(2x-1)\,, \tag{9}$$

where $P_i^{(a,b)}(z)$ are the classical Jacobi polynomials [31]. Note that Song and Steinrücken [7] primarily analyze the backward diffusion equation (but also use the forward diffusion equation in the section: "Empirical Transition Densities and Stationary Distributions"). However, the relationship between the adjoint backward and forward diffusion equations is such that adaption of the theory concerning the backward equation to the forward equation is minimal (compare [9]).

2.1. The Boundary-Mutation Model

Further in the text, we will follow Vogl and Bergman [4] and model mutations as only affecting the boundaries. Then, $\phi(x,t)$ must be interpreted as a generalized probability measure that integrates to one over the unit interval, but may contain point masses at the boundaries (compare [32]). Within the polymorphic region, $1/N \le x \le (N-1)/N$, the dynamics are purely governed by drift, such that the diffusion generator is:

$$\mathcal{L}_f = \frac{\partial^2}{\partial x^2} x(1-x)\,, \tag{10}$$

and the corresponding Kolmogorov forward (or forward diffusion) equation is:

$$\frac{\partial}{\partial t}\phi(x,t) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\phi(x,t)\Big)\,. \tag{11}$$

Mutations are assumed to arise at the boundaries and correspond to a transition from $x = 0$ to $x = 1/N$, for a mutation from A/T to C/G, or from $x = 1$ to $x = (N-1)/N$, for a mutation from C/G to A/T.

The Wright–Fisher model is most familiar to population geneticists. With this model, the transition between subsequent generations due to drift is modeled via binomial sampling, such that transitions between distant states are possible. The slightly less familiar Moran model only allows transitions between neighboring states, which simplifies the math. This simplification pertains also to the boundaries [4]. With the boundary-mutation model, mutations are assumed to arise only at the boundaries; a transition from $x = 0$ to $x = 1/N$ corresponds to a mutation from a monomorphic state with only A/T to a polymorphic state with a single C/G in the population; conversely, a transition from $x = 1$ to $x = (N-1)/N$ corresponds to a mutation from C/G to A/T. The reverse transitions from the polymorphic region to the boundaries at $x = 0$ and $x = 1$ are caused by drift. In particular, the flow from $x = 1/N$ towards zero is proportional to drift times the probability mass at $x = 1/N$, and similarly at the other boundary.

With a change from the Moran to the diffusion model, the formulas for the flow towards the boundaries due to drift are:

$$\frac{d\,\mathrm{F}(\alpha,\theta)}{dt} = \begin{cases} -\frac{N-1}{N}\phi(x = \frac{1}{N}, t\,|\,\alpha,\theta) & \text{for } x = 1/N \text{ to } x = 0, \\ \frac{N-1}{N}\phi(x = \frac{N-1}{N}, t\,|\,\alpha,\theta) & \text{for } x = (N-1)/N \text{ to } x = 1, \end{cases} \tag{12}$$

where the sign of the flow represents the direction. Conversely, the mutational flow from the boundaries to the interior is given by:

$$\frac{d\,\mathrm{F}(\alpha,\theta)}{dt} = \begin{cases} N\alpha\theta \int_0^1 (1-x)\phi(1-x,t\,|\,\alpha,\theta)\,dx & \text{for } x = 0 \text{ to } x = 1/N, \\ N\beta\theta \int_0^1 x\phi(x,t\,|\,\alpha,\theta)\,dx & \text{for } x = 1 \text{ to } x = (N-1)/N\,. \end{cases} \tag{13}$$

2.2. Modified Gegenbauer Polynomials

We will first analyze the situation without mutations, *i.e.*, $\theta = 0$. With pure drift, the transition density $\phi(x,t)$ can be expanded into a series of Gegenbauer polynomials (e.g., [5,7–9,33]). Define:

$$U_{i+2}(x) = x^{-1}(1-x)^{-1}G_i(x) = -\frac{2}{i+2}C_i^{(3/2)}(2x-1)\,, \tag{14}$$

where the $G_i(x)$ are the modified Gegenbauer polynomials of Song and Steinrücken [7], and the $C_i^{(\alpha)}(z)$ correspond to the classical ultraspherical or Gegenbauer polynomials with $\alpha = 3/2$ ([31], Chapter 22), also used by Kimura [5] and Tran *et al.* [8]. The forward and backward diffusion generators are adjoint, such that the modified Gegenbauer polynomials from Song and Steinrücken [7] can also be used to solve the forward diffusion equation. Multiplication of the weight function $x^{-1}(1-x)^{-1}$ and $G_i(x)$ in (14) transforms a solution of the backward equation into that of the forward equation (compare [9]).

The first two polynomials are $U_2(x) = -1$ and $U_3(x) = (2-4x)$; the recurrence relation to calculate all other polynomials is [7]:

$$U_{i+1}(x)\frac{(i+1)(i-1)}{2i(2i-1)} = U_i(x)\left(x - \tfrac{1}{2}\right) - U_{i-1}(x)\frac{(i-1)}{2(2i-1)}\,. \tag{15}$$

The $U_i(x)$ solve the differential equation:

$$-\lambda_i U_i(x) = \frac{d^2}{dx^2}\Big(x(1-x)U_i(x)\Big)\,, \tag{16}$$

with:

$$\lambda_i = i(i-1)\,. \tag{17}$$

The $U_i(x)$ are orthogonal with the weight function:

$$w(x) = x(1-x)\,, \tag{18}$$

and the proportionality constant:

$$\Delta_i = \frac{i-1}{(2i-1)i}\,. \tag{19}$$

A function $f(x)$ defined within $]0,1[$ can be represented by an expansion of the $U_i(x)$. The coefficients c_i can be calculated using:

$$c_i = \frac{1}{\Delta_i}\int_0^1 x(1-x)U_i(x)f(x)\,dx\,. \tag{20}$$

2.2.1. Solution of the Pure Drift Forward Equation with Gegenbauer Polynomials

Substituting $\phi(x,t) = \sum_{i=2}^{\infty} \tau_i(t)U_i(x)$, where $\tau_i(t)$ is a function pertaining to the temporal part of the transition density, into the diffusion Equation (11) leads to:

$$\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\tau_i(t)U_i(x)\Big) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\sum_{i=2}^{\infty}\tau_i(t)U_i(x)\Big)\,. \tag{21}$$

For each i, we have:

$$\frac{\partial}{\partial t}\Big(\tau_i(t)U_i(x)\Big) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\tau_i(t)U_i(x)\Big)\,, \tag{22}$$

which can be rearranged to:

$$\frac{\frac{\partial}{\partial t}\tau_i(t)}{\tau_i(t)} = \frac{\frac{\partial^2}{\partial x^2}\Big(x(1-x)U_i(x)\Big)}{U_i(x)}\,, \tag{23}$$

Observing Equation (16), we obtain the eigenvalue equations:

$$\begin{cases} -\lambda_i &= \frac{\frac{d^2}{dx^2}\left(x(1-x)U_i(x)\right)}{U_i(x)}\,, \\ -\lambda_i &= \frac{\frac{d}{dt}\tau_i(t)}{\tau_i(t)}\,. \end{cases} \tag{24}$$

As stated above, the $U_i(x)$ solve the spatial differential Equation in (24) for each i, while $\tau_i(t) = c_i e^{-\lambda_i t}$ solves the temporally homogeneous, linear differential Equation in (24). The c_i depend on the starting conditions and can be obtained from Formula (20). Since any function in $]0,1[$ can be represented by $\sum_{i=2}^{\infty} c_i U_i(x)$,

$$\phi(x,t) = \sum_{i=2}^{\infty}\tau_i(t)U_i(x) = \sum_{i=2}^{\infty} c_i e^{-\lambda_i t}U_i(x) \tag{25}$$

solves the diffusion Equation (11) for any starting condition.

Integrating (16) from zero to one for the symmetric eigenvectors with even i, we obtain:

$$\begin{aligned}
-\lambda_i \int_0^1 U_i(x)\,dx &= \int_0^1 \frac{d^2}{dx^2}\Big(x(1-x)U_i(x)\Big)\,dx \\
&= \int_0^1 \frac{d}{dx}\left(x(1-x)\frac{d}{dx}U_i(x) + (1-2x)U_i(x)\right)\,dx \\
&= \left(x(1-x)\frac{d}{dx}U_i(x) + (1-2x)U_i(x)\right)\Big|_0^1 \\
&= -U_i(1) - U_i(0)\,.
\end{aligned} \tag{26}$$

Note that, for odd i, we have $U_i(x) = -U_i(1-x)$, such that $U_i(0) = -U_i(1)$, $U_i(1/2) = 0$, and $\int_0^1 U_i(x)\,dx = 0$. Equation (26) is therefore trivially true for odd i.

Following Kimura [5], we substitute $\phi(x,t) = \sum_{i=2}^{\infty} \tau_i(t)U_i(x)$ into the differential Equation (11) and integrate from zero to one observing Equation (26):

$$\begin{aligned}
\frac{\partial}{\partial t}\Big(\lim_{N\to\infty}\int_{1/N}^{(N-1)/N} \phi(x,t)\,dx\Big) &= \lim_{N\to\infty}\int_{1/N}^{(N-1)/N} \frac{\partial^2}{\partial x^2}\Big(x(1-x)\phi(x,t)\Big)\,dx \\
\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\tau_i(t)\int_0^1 U_i(x)\,dx\Big) &= \sum_{i=2}^{\infty}\tau_i(t)\int_0^1 \frac{\partial^2}{\partial x^2}\Big(x(1-x)U_i(x)\Big)\,dx \\
\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\frac{\tau_i(t)}{\lambda_i}\left(U_i(0)+U_i(1)\right)\Big) &= -\sum_{i=2}^{\infty}\tau_i(t)\left(U_i(0)+U_i(1)\right)\,.
\end{aligned} \tag{27}$$

As in the temporal part of Formula (24), substituting $\tau_i(t) = c_i e^{-\lambda_i t}$ solves the system of differential equations:

$$\frac{d}{dt}\tau_i(t) = -\lambda_i\tau_i(t)\,. \tag{28}$$

For even i, the flow out at the boundaries zero and one per unit time is symmetric and corresponds to what is present just inside the boundaries, *i.e.*, $U_i(0)$ and $U_i(1)$. The rate of loss is the eigenvalue λ_i.

Now, multiply $\phi(x,t)$ with x and integrate again from zero to one:

$$\begin{aligned}
\frac{\partial}{\partial t}\Big(\lim_{N\to\infty}\int_{1/N}^{(N-1)/N} x\phi(x,t)\,dx\Big) &= \lim_{N\to\infty}\int_{1/N}^{(N-1)/N} \frac{\partial^2}{\partial x^2}\Big(x(1-x)x\phi(x,t)\Big)\,dx \\
\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\tau_i(t)\int_0^1 xU_i(x)\,dx\Big) &= \sum_{i=2}^{\infty}\tau_i(t)\int_0^1 \frac{\partial^2}{\partial x^2}\Big(x^2(1-x)U_i(x)\Big)\,dx \\
\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\frac{\tau_i(t)}{\lambda_i}U_i(1)\Big) &= -\sum_{i=2}^{\infty}\tau_i(t)U_i(1)\,.
\end{aligned} \tag{29}$$

Thus, $x\phi(x,t)$ will eventually drift to Boundary 1, and conversely $(1-x)\phi(x,t)$ to Boundary 0. This is expected, since drift is symmetric, such that the probability of eventual fixation is equal to the proportion x, for the boundary at one, and $(1-x)$, for the boundary at zero (see also [30], Chapter 4.3). As in the temporal part of Formula (24), substituting $\tau_i(t) = c_i e^{-\lambda_i t}$ solves the system of differential Equations (28).

The solution of the diagonal system of differential Equations (28) thus fulfills Equation (11) for all t, as well as the boundary conditions at both zero and one (12).

Note that:

$$\begin{cases} U_i(0)/\lambda_i = \int_0^1 (1-x)U_i(x)\,dx \\ U_i(1)/\lambda_i = \int_0^1 xU_i(x)\,dx\,. \end{cases} \tag{30}$$

The above results suggest augmenting the $U_i(x)$ with the boundary terms:

$$\begin{cases} -U_i(0)/\lambda_i = (-1)^i/i \\ -U_i(1)/\lambda_i = 1/i\,. \end{cases} \tag{31}$$

Furthermore, the result (30) shows that the boundary terms derived above correspond to those defined by Tran *et al.* [8].

We therefore define the following set of orthogonal polynomials augmented with boundary terms:

$$H_i(x) = \frac{(-1)^i\,\delta(x) + \delta(x-1)}{i} + U_i(x)\,, \tag{32}$$

where $\delta(x)$ is the Dirac delta function (compare Tran *et al.* [8], who arrive at the corresponding set of augmented eigenfunctions). With this definition of eigenfunctions, the probability mass that leaves the polymorphic region for each i at $1/N$ and $(N-1)/N$ is added to the monomorphic boundaries at $x = 0$ and $x = 1$. The integral over the closed interval between zero and one thus remains unity for all times. In Appendix A.1, we show that these augmented orthogonal polynomials can also be obtained by a Taylor series expansion of the general eigensystem solving the diffusion Equation (11) using the modified Jacobi polynomials $R^{(\alpha,\theta)}(x)$.

2.2.2. Starting and Prior Distributions

We base the following description on the theory of hierarchical Bayesian models and the empirical Bayes method [16] that we also employed earlier [3,4]. In a frequentist context, one would rather use the context of marginal likelihoods.

Traditionally, a Dirac delta function at a certain position p has been used as a starting condition [33]. With a site frequency spectrum, however, the joint density of the population allelic proportion x and the observed allelic count y in a sample of size M_0 must be used as starting density. Most naturally, the conditional distribution of the data y given the allelic proportion x is modeled as a binomial:

$$p(y \,|\, x, M_0) = \binom{M_0}{y} x^y (1-x)^{M_0-y}\,. \tag{33}$$

With the pure drift model, we are generally interested in the polymorphic region, since probability mass at a boundary remains there due to the absence of mutations.

For "integrating out" the population allelic proportions x, a prior distribution for x must be assumed. With small-scaled mutation rates, an "improper prior" proportional to $x^{-1}(1-x)^{-1}$ within $1/N$ and $(N-1)/N$ is appropriate, as this is proportional to the equilibrium distribution (see also Subsections 2.3.4 and 2.3.5 below). Note that this prior corresponds to the inverse of the weight function. Thus, the inner product (20) to calculate the initial coefficients becomes:

$$\begin{aligned} c_i &= \frac{1}{\Delta_i} \lim_{N\to\infty} \int_{1/N}^{(N-1)/N} x(1-x) U_i(x)\, p(y \,|\, x, M_0) x^{-1}(1-x)^{-1}\, dx \\ &= \frac{1}{\Delta_i} \lim_{N\to\infty} \int_{1/N}^{(N-1)/N} U_i(x)\, p(y \,|\, x, M_0)\, dx\,, \end{aligned} \tag{34}$$

where the limit notation indicates that the integration includes only the polymorphic region, *i.e.*, no point masses at the boundaries.

We can thus specify a general algorithm that also includes the boundaries.

2.2.3. Algorithm 1: Allelic Proportions x with Pure Drift for All Times t, Conditional on Initial Values

- A measure $f(x)$ between zero and one, which may have point masses m_0 and m_1 at Boundaries 0 and 1, is represented by an expansion of the $H_i(x)$ up to $i = n$. The coefficients c_i are calculated according to Equation (34). The expansion of $g(x)$ times the prior, up to the order n, is then:

$$g(x) = \left(m_0 - \sum_{i=2}^{n} c_i \frac{(-1)^i}{i} \right) \delta(x) + \left(m_1 - \sum_{i=2}^{n} \frac{c_i}{i} \right) \delta(x-1) + \sum_{i=2}^{n} (c_i H_i(x)) + O(n+1) \,. \quad (35)$$

- The solution of Equation (28) for all t conditional on the initial distribution can be represented by a series expansion up to n:

$$g(x,t) = \left(m_0 - \sum_{i=2}^{n} c_i \frac{(-1)^i}{i} \right) \delta(x) + \left(m_1 - \sum_{i=2}^{n} \frac{c_i}{i} \right) \delta(x-1) + \sum_{i=2}^{n} \left(c_i H_i(x) e^{-\lambda_i t} \right) + O(n+1) \,, \quad (36)$$

  with $\lambda_i = i(i-1)$.

Note that the $H_i(x)$ contain the boundary terms that balance the probability masses at zero and one. This is obvious if the initial probability measure $f(x)$ does not contain point masses at the boundaries, *i.e.*, if $m_0 = m_1 = 0$.

2.3. Modified Gegenbauer Polynomials and the Boundary-Mutation Model

In this subsection, we will use the expansion in orthogonal polynomials with boundary terms to model both mutation and drift.

2.3.1. Mutation and Drift: Slowly Evolving Dynamics

For the slowly evolving dynamics at the boundaries, we augment the system with two eigenfunctions. Starting from the system for general θ, which can be expanded in a series of modified Jacobi polynomials (see Equation (9) in Song and Steinrücken [7]), we note that the eigenfunction for $i = 0$ does not change with time, *i.e.*, $\lambda_0 = 0$. The eigenfunction for $i = 1$ has the eigenvalue $\lambda_1 = \theta$ (compare: [7]) and reflects the slow change in allele frequencies through mutation. Expressing the Jacobi polynomials as beta distributions and taking the limit $\theta \to 0$, such that only probability masses at the boundaries remain, the first two eigenvectors become:

$$\begin{cases} H_0^{(\alpha)}(x) = \beta\delta(x) + \alpha\delta(x-1)\,, \\ H_1(x) = -\delta(x) + \delta(x-1). \end{cases} \quad (37)$$

(see Appendix A.1). Obviously, these two eigenfunctions are unaffected by the dynamics in the polymorphic region inside $[1/N, (N-1)/N]$.

These two eigenvectors have no probability mass within the polymorphic region, such that only eigenvectors with $i \geq 2$ have nonzero probability masses in the polymorphic region. Hence, the model separates two spatial regions: the monomorphic boundaries and the polymorphic interior. As $\lambda_1 = \theta \ll 1$ while the $\lambda_i > 1$ for all eigenvalues with $i > 2$, two different temporal regions can also be distinguished, in addition to the two spatial regions. Thus, evolution is modeled as a two-time process, where the slow dynamics captured by the eigenfunctions $i = 0$ and $i = 1$ are evolving independently from the polymorphic region, while the fast dynamics in the polymorphic region are in quasi-equilibrium with the slow dynamics at the boundaries. Generally, we are thus looking at a system of differential equations that for the slowly evolving part of the system is:

$$\begin{cases} \tau_0(t) = 1\,, \\ \frac{d}{dt}\tau_1(t) = -\theta\tau_1(t)\,. \end{cases} \quad (38)$$

Initially at $t = 0$, the boundary values are $b_0(0) = \int_0^1 x f(1-x \mid t=0)\, dx$ and $b_1(0) = \int_0^1 x f(x \mid t = 0)\, dx = (1 - b_0(0))$. The solution over time is $\tau_1(t) = (b_1(t) - \alpha)e^{-\theta t}$, such that the boundary values will slowly, at a rate of θ, approach the equilibrium values:

$$b_1(t) = \alpha + (b_1(0) - \alpha)e^{-\theta t} = 1 - b_0(t) \tag{39}$$

If $f(x)$ does not integrate to one, *i.e.*, $b_0(t) + b_1(t) \neq 1$, modifications are trivial. Note that $b_0(t = 0)$ and $b_1(t = 0)$ correspond to the probability mass currently at the boundaries plus the part of the probability mass within the polymorphic region that is expected to be fixed by drift (*i.e.*, without any further mutations) at the respective boundaries. They would only be identical to the probability mass currently at the boundaries if there were no probability mass in the polymorphic region.

2.3.2. Mutation and Drift: Quickly Evolving Dynamics

The slowly evolving part of the system is given in (39). For the quickly evolving part, note that, from Equation (13), mutation moves probability mass from the boundary at zero $x = 0$ to $x = 1/N$ and from $x = 1$ to $x = (N-1)/N$, respectively. We can model this with a Dirac delta function at $x = 1/N$ and $x = (N-1)/N$:

$$\begin{cases} N\alpha\theta\delta(x - 1/N)b_0(t) & \text{for } x = 0 \text{ to } x = 1/N, \\ N\beta\theta\delta(x - (N-1)/N)b_1(t) & \text{for } x = 1 \text{ to } x = (N-1)/N, \end{cases} \tag{40}$$

with $b_0(t)$ and $b_1(t)$ as above. Combined with the pure drift diffusion Equation (11), we thus obtain the following diffusion equation within the interval between $1/N$ and $(N-1)/N$:

$$\frac{\partial}{\partial t}\phi(x,t) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\phi(x,t)\Big) + N\alpha\theta\delta(x - 1/N)b_0(t) + N\beta\theta\delta(x - (N-1)/N)b_1(t)\,, \tag{41}$$

Equation (41) is an extension of Equation (21) to mutations from the boundaries.

2.3.3. Mutation and Drift: Slowly and Quickly Evolving Dynamics

Theorem 1. *Starting from a generalized probability measure $f(x)$ within the unit interval (Equation (11)), with the boundary Conditions (12) and (13), and letting $N \to \infty$, the following function provides the general solution for all times of the Kolmogorov forward equation of boundary-mutation drift diffusion:*

$$\phi(x,t) = H_0^{(\alpha)}(x) + \sum_{i=1}^{\infty} \tau_i(t)\, H_i(x)\,, \tag{42}$$

with the previously-defined eigenfunctions (Equations (37) and (32)); the $\tau_i(t)$ are given by a system of linear inhomogenous first order differential equations:

$$\begin{cases} \frac{d}{dt}\tau_1(t) = -\theta\tau_1(t) \\ \frac{d}{dt}\tau_i(t) = -\lambda_i\tau_i(t) + \theta(2i-1)i\left((-1)^i\alpha b_0(t) + \beta b_1(t)\right)\ , \textit{for } i \geq 2. \end{cases} \tag{43}$$

The starting values, $\tau_i(t = 0)$ for $i \geq 1$, are given by the initial probability masses at the boundaries and by the expansion of the initial density $f(x)$ in the interior into the eigenvectors.

Proof. The slowly evolving part of the system is given in (39). The coefficients for expanding the delta function in (40) are (compare Equation (20)) for the boundary at zero:

$$\begin{aligned}\frac{1}{\Delta_i}\lim_{N\to\infty}\int_{1/N}^{(N-1)/N} Nx(1-x)U_i(1/N)\delta(x-1/N)\,dx = \frac{U_i(0)}{\Delta_i} &= \frac{(2i-1)iU_i(0)}{(i-1)}\\ &= -\frac{(2i-1)i(-1)^i(i-1)}{i-1}\\ &= -(-1)^i(2i-1)i\,,\end{aligned} \tag{44}$$

and similarly for the boundary at one. Substituting the Gegenbauer expansion into Equation (41), we obtain:

$$\frac{\partial}{\partial t}\Big(\sum_{i=2}^{\infty}\tau_i(t)U_i(x)\Big) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\sum_{i=2}^{\infty}\tau_i(t)U_i(x)\Big) - \theta(2i-1)i\left((-1)^i\alpha b_0(t)+\beta b_1(t)\right)U_i(x)\,. \tag{45}$$

For each i, we have:

$$\frac{\partial}{\partial t}\Big(\tau_i(t)U_i(x)\Big) = \frac{\partial^2}{\partial x^2}\Big(x(1-x)\tau_i(t)U_i(x)\Big) - \theta(2i-1)i\left((-1)^i\alpha b_0(t)+\beta b_1(t)\right)U_i(x)\,, \tag{46}$$

which can be rearranged to:

$$\frac{\frac{\partial}{\partial t}\tau_i(t)+\theta(2i-1)i\left((-1)^i\alpha b_0(t)+\beta b_1(t)\right)}{\tau_i(t)} = \frac{\frac{\partial^2}{\partial x^2}\Big(x(1-x)U_i(x)\Big)}{U_i(x)}\,, \tag{47}$$

Observing Equation (16), we obtain eigenvalue equations corresponding to those in (24):

$$\begin{cases} -\lambda_i &= \frac{\frac{d^2}{dx^2}\left(x(1-x)U_i(x)\right)}{U_i(x)}\,,\\ -\lambda_i &= \frac{\frac{d}{dt}\tau_i(t)+\theta(2i-1)i\left(\alpha(-1)^ib_0(t)+\beta b_1(t)\right)}{\tau_i(t)}\,. \end{cases} \tag{48}$$

Compared to the case without mutation, the spatial part is unchanged, while the temporal part becomes a system of linear inhomogenous first order differential equations:

$$\frac{d}{dt}\tau_i(t) = -\lambda_i\tau_i(t) - \theta(2i-1)i\left((-1)^i\alpha b_0(t)+\beta b_1(t)\right)\,. \tag{49}$$

All other considerations correspond to the case without mutation. □

Note that, substituting $b_0(t)$ and $b_1(t)$, Equation (49) can also be written as:

$$\frac{d}{dt}\tau_i(t) = -\lambda_i\tau_i(t) + A_i + B_i\,e^{-\theta t}\,, \tag{50}$$

with constants:

$$\begin{aligned}A_i &= -\alpha\beta\theta(2i-1)i\left((-1)^i+1\right)\,,\\ B_i &= -\theta(2i-1)i(b_0(0)-\beta)\left((-1)^i\alpha-\beta\right)\,.\end{aligned} \tag{51}$$

Furthermore, note that we assumed that the probability measure $f(x)$ integrates to one over the closed interval between zero and one, *i.e.*, $\int_0^1 f(x)\,dx = 1$. If this is not the case, the constants A_i and B_i must be multiplied by $\int_0^1 f(x)\,dx$.

2.3.4. Boundary-Mutation-Drift Equilibrium Distribution

In earlier work [4], we show that the equilibrium solution of the boundary-mutation model is the measure:

$$BME(x\,|\,\alpha,\theta,N) = \begin{cases} \beta - \alpha\beta\theta\ \int_{1/N}^{(N-1)/N} \frac{1}{x}\,dx & \text{for } x = 0, \\ \alpha\beta\theta\ \frac{1}{x(1-x)} & \text{for } 1/N \le x \le (N-1)/N, \\ \alpha - \alpha\beta\theta\ \int_{1/N}^{(N-1)/N} \frac{1}{1-x}\,dx & \text{for } x = 1\,, \end{cases} \tag{52}$$

where the interior region is bounded by $1/N$ and $(N-1)/N$ and BME stands for boundary-mutation equilibrium. $BME(x\,|\,\alpha,\theta,N)$ integrates to one over the unit interval, irrespective of N. However, note that for large N, it integrates to more than one inside the interval $[1/N,(N-1)/N]$, while assuming negative values at the boundaries. In this limit, it therefore must be considered an "improper distribution" [4,34].

In Appendix A.2, we show that, with time, Solution (43) converges to the BME (Equation (52)).

2.3.5. Prior Distribution

With the BME as prior and a binomial likelihood $p(y\,|\,x,M_0)$ with $0 \le y \le M_0$, the coefficients of the joint distribution $p(x,y\,|\,M_0,\alpha,\theta) = p(y\,|\,x,M_0)x^{-1}(1-x)^{-1}$ become:

$$c_i = \alpha\beta\theta\frac{1}{\Delta_i}\lim_{N\to\infty}\int_{1/N}^{(N-1)/N} x(1-x)U_i(x)\,p(y\,|\,x,M_0)x^{-1}(1-x)^{-1}\,dx\,. \tag{53}$$

where the limit notation indicates that the integration includes only the polymorphic region. Note that already Ewens used the same limit for inference [18,25]. For polymorphic data, *i.e.*, $1 \le y \le (M_0-1)$, this function is a polynomial and, thus, can be represented accurately as a series of Gegenbauer polynomials as long as $n > M_0$. The boundary terms can also be derived easily because the probability of drifting to boundary one corresponds to the current proportion x (and to $(1-x)$ to the boundary zero), such that:

$$\begin{cases} b_1(1 \le y \le M_0, t=0) = \alpha\beta\theta\ \int_0^1 p(y\,|\,x,M_0)(1-x)^{-1}\,dx = \alpha\beta\theta\frac{1}{y} \\ b_0(1 \le y \le M_0, t=0) = \alpha\beta\theta\ \int_0^1 p(y\,|\,x,M_0)x^{-1}\,dx = \alpha\beta\theta\frac{1}{M_0-y}\,, \end{cases} \tag{54}$$

where the limit notation is not used for brevity.

For monomorphic y, *i.e.*, $y=0$ or $y=M_0$, the c_i for the probability mass in the interior are also given by Equation (53) with $i \le n$. The corresponding boundary terms are:

$$\begin{cases} b_1(y=M_0, t=0) = \alpha - \alpha\beta\theta\ \sum_{y=1}^{M_0}\frac{1}{y} \\ b_0(y=M_0, t=0) = \alpha\beta\theta\frac{1}{M_0} \end{cases} \tag{55}$$

and analogously for $y=0$.

2.3.6. Algorithm 2: Allelic Proportions x with Boundary-Mutations and Drift for All Times t, Conditional on Initial Values

- The interior of a joint distribution $p(x, y \,|\, M_0, \alpha, \theta)$ is represented as a Gegenbauer series (53).
- The slowly evolving part of the system consists of the dynamics at the boundaries. Set the boundary terms at $t = 0$ to $b_0(t = 0)$ and $b_1(t = 0)$ as in Equations (54) and (55). With time, the boundary terms $b_0(t)$ and $b_1(t)$ then change slowly at the rate of θ according to the exponential function in Equation (39).
- Set $\omega = \int_0^1 p(x, y \,|\, M_0, \alpha, \theta)\, dx = b_0(t) + b_1(t)$. The solution of Equation (41) for all t conditional on $f(x)$ can be represented by a series expansion up to n:

$$f(x,t) = b_0(t)\delta(x) + b_1(t)\delta(x-1) + \sum_{i=2}^{n} \left(\tau_i(t) H_i(x)\right) + O(n+1)\,, \tag{56}$$

with:

$$\tau_i(t) = \frac{A_i'}{\lambda_i} + \left(c_i - \frac{A_i'}{\lambda_i} - \frac{B_i'}{\lambda_i - \theta}\right) e^{-\lambda_i t} + \frac{B_i'}{\lambda_i - \theta} e^{-\theta t}\,, \tag{57}$$

and constants analogous to (51):

$$\begin{aligned} A_i' &= -\omega\alpha\beta\theta(2i-1)i\left((-1)^i + 1\right), \\ B_i' &= -\theta(2i-1)i(b_0(0) - \omega\beta)\left((-1)^i\alpha - \beta\right). \end{aligned} \tag{58}$$

Equation (57) is a solution to (50), as can be shown by substitution.

3. Applications

In this section, we illustrate the calculation of the marginal likelihoods of a mock dataset and an empirical fruit fly dataset using the expansion of Gegenbauer polynomials up to degree $n = 52$.

3.1. A Joint Site Frequency Spectrum under Pure Drift

With the pure drift model, the time between two time points $t_0 = 0$ and $t_1 > 0$ is assumed to be so small that newly arising mutations can be neglected. Moreover, sites where the samples from both time points are monomorphic for the same allele are usually ignored with such data analysis. For simplicity, assume that the sample size of the initial sample at time t_0 is $M_0 = 3$ and that of time t_1 also $M_1 = 3$. Four different cases need to be considered: (i) a site is polymorphic in both samples; (ii) a site is polymorphic in the first sample and monomorphic in the second; (iii) a site is monomorphic in the first sample and polymorphic in the second; and (iv) a site is monomorphic in the first sample for one allelic type and polymorphic in the second sample for the other allelic type. For Cases (i) and (ii), assume, e.g., a sample with two alleles of a certain type (zero or one), *i.e.*, $y_0 = 2$. Thus, the joint density of the sample y_0 and the allelic proportions x become:

$$p(y_0 = 2, x \,|\, M_0 = 3, t = 0) \propto \binom{3}{2} x^2(1-x)\, x^{-1}(1-x)^{-1} = 3x\,. \tag{59}$$

This is represented by a sum of the modified Gegenbauer polynomials of degree up to three with $c_3 = -\frac{3}{4}$ and $c_2 = -\frac{3}{2}$. At time t_1, before considering the second sample, the probability mass of the joint interior distribution has diminished:

$$p(y_0 = 2, 0 < x < 1 \,|\, M_0 = 3, t = t_1) \propto -\frac{3}{2} e^{-2t_1}(-1) - \frac{3}{4} e^{-6t_1}(2 - 4x)\,, \tag{60}$$

while it has grown at the boundaries:

$$p(y_0 = 2, x = 0 \mid M_0 = 3, t = t_1) \propto \frac{3}{2} \cdot \frac{1}{2}\,(1 - e^{-2t_1}) - \frac{3}{4} \cdot \frac{1}{3}\,(1 - e^{-6t_1}) \tag{61}$$

and:

$$p(y_0 = 2, x = 1 \mid M_0 = 3, t = t_1) \propto \frac{3}{4}\,(1 - e^{-2t_1}) + \frac{1}{4}\,(1 - e^{-6t_1})\,. \tag{62}$$

For Case (i), the likelihood of a second sample of size $M_1 = 3$ with $y_1 = 1$ alleles of the first type is binomial: $3x(1-x)^2$. The joint distribution consists only of an interior part, from which x can be integrated out to obtain the marginal likelihood:

$$\begin{aligned} \ell(y_0 = 2, y_1 = 1 \mid M_0 = 3, M_1 = 3, t = t_1) &\propto \int_0^1 3x(1-x)^2 \left(\frac{3}{2}\,e^{-2t_1} - \frac{3}{4}\,e^{-6t_1}(2 - 4x) \right) dx \\ &= \frac{3}{8}\,e^{-2t_1} - \frac{3}{40}\,e^{-6t_1}. \end{aligned} \tag{63}$$

For Case (ii), the likelihood of a second sample of, e.g., size $M_1 = 3$ with $y_1 = 3$ alleles of the first type, is binomial: x^3. The joint distribution consists of an interior part, from which x can be integrated out:

$$\begin{aligned} p(y_0 = 2, y_1 = 3 \mid M_0 = 3, M_1 = 3, t = t_1, 0 < x < 1) &\propto \int_0^1 x^3 \left(\frac{3}{2}\,e^{-2t_1} + \frac{3}{4}\,e^{-6t_1}(-2 + 4x) \right) dx \\ &= \frac{3}{8}\,e^{-2t_1} + \frac{9}{40}\,e^{-6t_1}\,. \end{aligned} \tag{64}$$

Summing the interior and the boundary part, the marginal likelihood of the two samples is obtained:

$$\ell(y_0 = 2, y_1 = 3 \mid M_0 = 3, M_1 = 3, t = t_1) \propto \frac{3}{8}\,e^{-2t_1} + \frac{9}{40}\,e^{-6t_1} + \frac{3}{4}\,(1 - e^{-2t_1}) + \frac{1}{4}\,(1 - e^{-6t_1})\,. \tag{65}$$

For Cases (i) and (ii), thus, a finite expansion of the Gegenbauer polynomials was sufficient.

For Cases (iii) and (iv), the product of the likelihood and prior at time $t = 0$ results in an infinite series of Gegenbauer polynomials. Note that the monomorphic term at $t = 0$ does not need to be included since, without new mutation, a polymorphism or a monomorphic alternative allele at $t = t_1$ implies that the population allelic proportion x must have been in the polymorphic region already at $t = 0$. Take, e.g., $M_0 = 3$ and $y_0 = 3$, which results in:

$$p(y_0 = 3, x \mid M_0 = 3, t = 0) \propto x^3\,x^{-1}(1-x)^{-1} = x^2(1-x)^{-1}\,. \tag{66}$$

While Equation (20) can be used in this case, it results in a rather "wiggly" function of x (Figure 1). With the expansion $(1-x)^{-1} = \sum_{i=0}^{n-2} x^i$ a much smoother polynomial of degree n will be obtained, which can be expressed without loss as a sum of Gegenbauer polynomials up to that degree. Even with only moderate n, the two expansions will produce nearly indistinguishable likelihoods. In either case, the algebra cannot easily be reproduced here, but likelihoods corresponding to transitions between all possible allelic states with arbitrary (but moderate) M_0, M_1 and t_1 can be easily calculated using computers. We implemented such an algorithm and tested inference using simulated datasets of $L = 10^5$ and $M_0 = M_1 = 3$, while varying the parameter t_1. As expected, the modes of the likelihood curves closely coincide with the true values of t_1 (Figure 2).

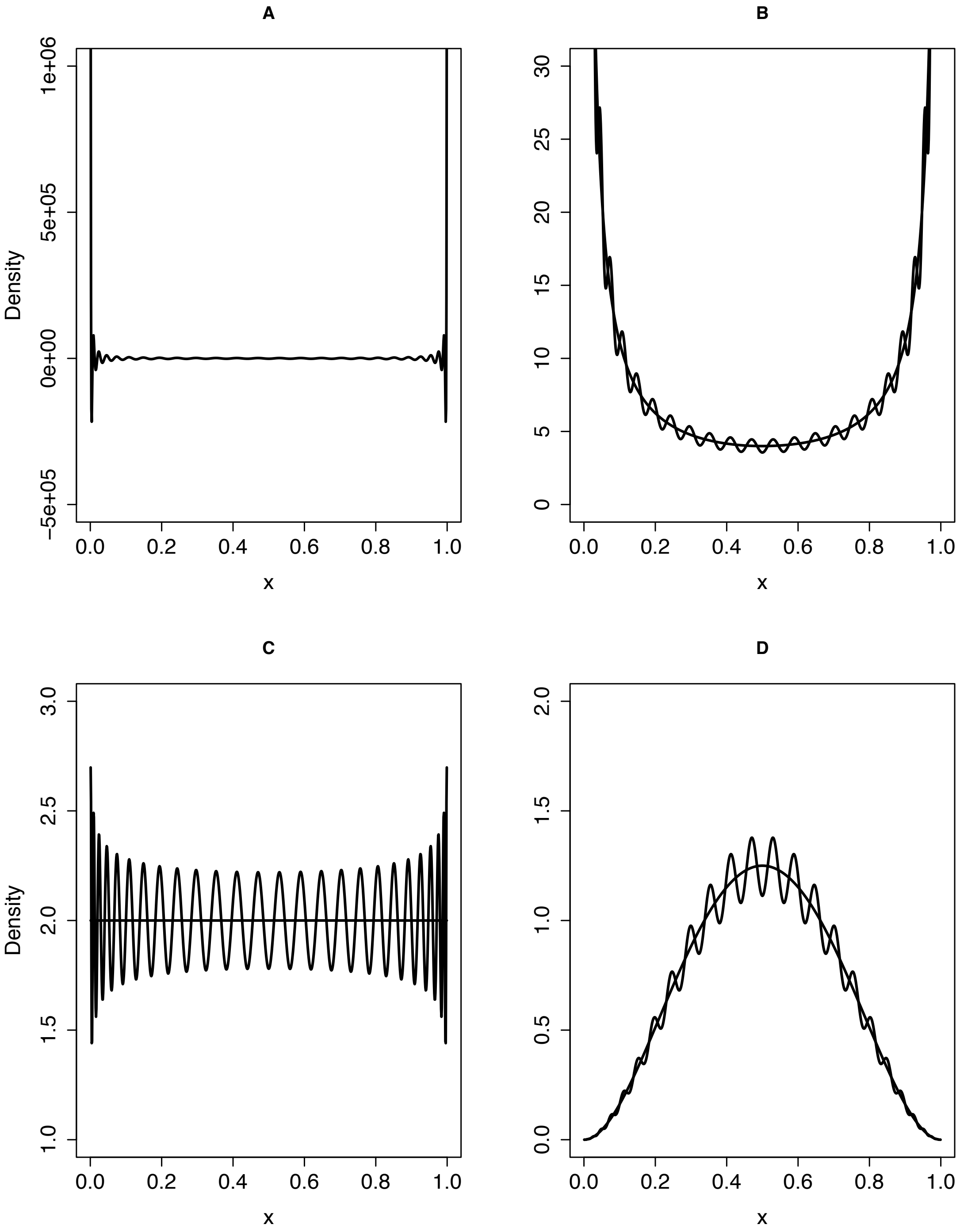

Figure 1. Approximate densities using the Gegenbauer polynomial expansion with terms up to $n = 52$. (**A**) Approximation to point masses at both boundaries, but without mass in the interior region; (**B**) approximation to the equilibrium improper density overlying the function $x^{-1}(1-x)^{-1}$; (**C**) approximation to the joint posterior distribution for a sample with $y = 1$, $M = 1$ overlying the joint distribution $2\,x^{1-1}(1-x)^{1-1}$; (**D**) approximation to the joint posterior distribution for a sample with $y = 3$, $M = 6$ overlying the joint distribution $\binom{6}{3}\,x^{3-1}(1-x)^{3-1}$.

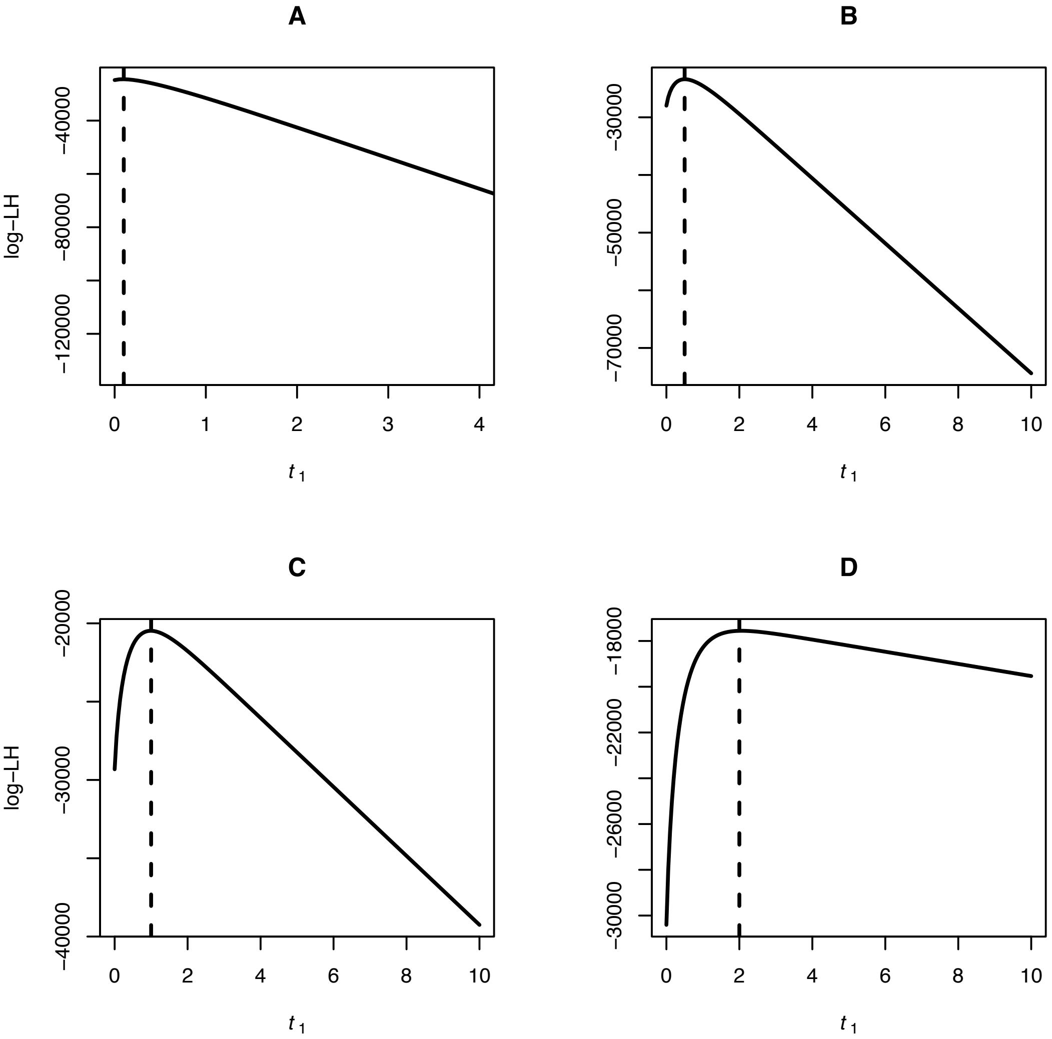

Figure 2. Pure drift model. Likelihood curves of the parameter t_1 given a sample of $L = 10,000$, $M_0 = M_1 = 3$ and true t_1 (dashed vertical lines) equal to 0.1 (**A**), 0.5 (**B**), 1 (**C**) and 2 (**D**).

3.2. Application to Drosophila Population Data

The joint site frequency spectrum of putatively neutral short intronic sites (positions 8–30 of introns less than 66 bp in length [11]) was used for inference (Table 1). The interior part of short introns is unlikely to contain selectively-constrained sequences. Short introns also show the highest intra- and inter-species diversity of any sequence class within the *Drosophila* genome [11]. Furthermore, short introns are the most abundant intron type within the *Drosophila* genome. It is therefore assumed that mutation-drift dynamics shape the nucleotide composition of short intronic sites, and since polymorphism within a single intron is rare and linkage disequilibria decrease quickly, free recombination among sites may be assumed. Sites were classified as binary by lumping A and T nucleotides together as Allele 0 and C and G nucleotides as Allele 1. The reference sequence from *D. sechellia* (Release 1.0; [15]) was taken as ancestral, *i.e.*, the initial sample of size $M_0 = 1$ at time $t_0 = 0$. While the states of closely related species are routinely taken as ancestral (e.g., [12,13,22–24]), this practice violates the model assumptions that data are from a single populations and two time points. A *D. melanogaster* Malawian population sample [14] of size $M_1 = 6$ was considered as a sample from a later time point t_1. The sequences were annotated by aligning the *D. sechellia* reference and the *D. melanogaster* population sample to the *D. melanogaster* reference sequence (Release 5.9; [15]).

Table 1. A joint site frequency spectrum of *Drosophila* short intronic sites with $M_0 = 1$ and $M_1 = 6$. The left-most column and the upper row of the table represent the possible allelic states of sites for the sample M_0 and M_1, respectively. The interior entries of the table are the counts of sites with a specific allelic state with respect to Allele 1.

	0	**1**	**2**	**3**	**4**	**5**	**6**
0	84,294	862	369	59	233	293	5121
1	5637	259	276	310	475	1168	41,531

Interest is centered on inferring the time point t_1, *i.e.*, the time in N generations since the split of the two *Drosophila* species and the scaled mutation rate θ_1, corresponding to the current mutation rate of the *D. melanogaster* population sample. Firstly, a prior distribution of allelic counts needed to be determined by setting initial parameters, α_0 and θ_0. The ancestral mutation bias α_0 was inferred from the *D. sechellia* data to be $\hat{\alpha}_0 = 0.35$ and is assumed to not change, *i.e.*, $\alpha_1 = \alpha_0$. We estimate the ancestral scaled mutation rate to be about $\hat{\theta}_0 = 0.079$ [4] from *D. simulans* data, as this closely related species most likely reflects the ancestral state of both *D. sechellia* and *D. melanogaster* species due to its relatively constant (over the evolutionary times considered) and large effective population size [35].

We implemented a direct grid search algorithm, with the likelihood calculated as in Subsection 2.3.6, to obtain maximum likelihood estimates of parameters t_1 and θ_1 (Figure 3). The maximum likelihood estimates $\hat{t}_1 = 4.5$ and $\hat{\theta}_1 = 0.03$ correspond closely to previously published estimates [36,37].

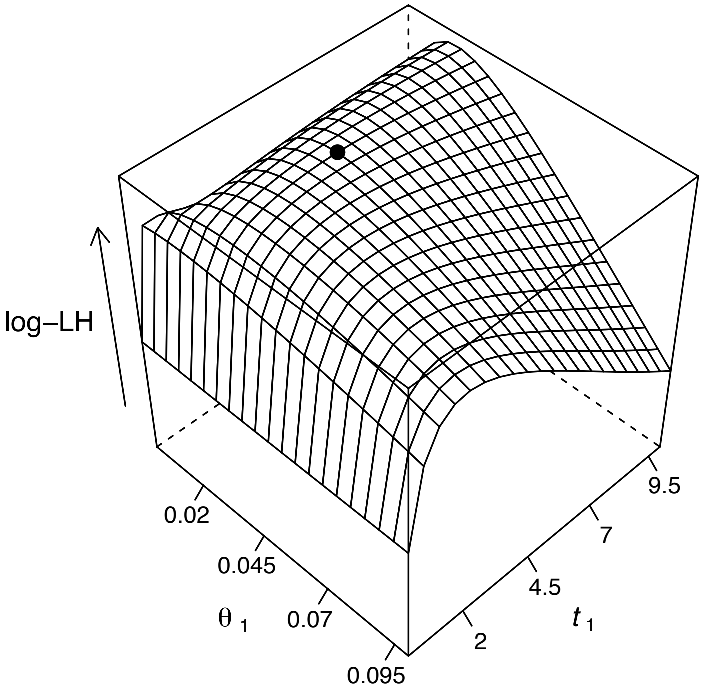

Figure 3. Likelihood surface with respect to parameters θ_1 and t_1 estimated from the joint site frequency spectrum in Table 1. The point on the likelihood surface corresponds to ML estimates: $\hat{\theta}_1 = 0.03$ and $\hat{t}_1 = 4.5$.

4. Discussion

Most population genetic models, e.g., the Wright–Fisher and Moran models and the corresponding (forward) Wright–Fisher or Moran diffusion models, do not restrict the number of mutations segregating in a population at a given site and time. By contrast, most population genetic methods that allow for analytical maximum likelihood estimators assume that variation at a specific site originates from only a single mutation, such that no more than two alleles can be present at any given site and time. This assumption is made for the Ewens–Watterson estimators of the scaled mutation rate [18,19] and the Poisson random field (PRF) models [26–29]. These approaches thus implicitly assume low scaled mutation rates. Usually, it is also assumed that the ancestral state of a site is known. Both of these assumptions are made explicit with the infinite sites model. When introducing the infinite sites model, Kimura [20,21] assumed selection against the mutant allelic variant, such that only the favored ancestral allele would exist in the monomorphic state, while a mutant allele is eventually lost from the population by the joint action of adverse selection and drift. Kimura based the mathematics on a model of irreversible mutation [2].

In later developments of the infinite sites model [18,19], the assumption of selection was dropped. Without selection, sites may become monomorphic for all possible allelic states (usually two states are assumed, such that the model is bi-allelic). In practical applications, the ancestral state has to be inferred via "outgroup" information. For this inference, which is also called "polarization", data from an extant closely related population or species, *i.e.*, an "outgroup", are used (e.g., [12,13,22–24]). Polarization can only be successful if the outgroup is related closely enough to the focal population, such that double mutations are improbable, but distantly enough, such that allelic polymorphism is not shared with the focal population. These biological assumptions are rather restrictive.

With unrestricted mutations, *i.e.*, with the assumptions of the Wright–Fisher diffusion or the Moran models, allelic proportions in a population converge to a beta equilibrium distribution. For a sample of moderate size, a beta-binomial equilibrium distribution is obtained. It seems that RoyChoudhury and Wakeley [17] were the first to expand the beta-binomial equilibrium distribution into a Taylor series up to first order in the scaled mutation rate θ to derive the sample distribution of a bi-allelic locus. With this approach, polarization is not necessary. Rather, DNA sequence data can be made binary by grouping together sites with the bases adenine (A) and thymine (T) to A/T and cytosine (C) and guanine (G) to C/G. In spite of this difference of the original infinite sites model, the sample distribution of polymorphic sites is a variant of the infinite sites model [18], and the maximum likelihood estimator of the scaled mutation rate derived from this distribution is a variant of the Ewens–Watterson estimator. Based on the Moran model, Vogl and Clemente [10] arrived at the same distribution of polymorphic sites if mutations only occur at the monomorphic boundaries. A Moran model with mutations only at the boundaries approximates a Moran model with mutations at any allelic state sufficiently well if the scaled mutation rate θ is below 0.1 [10]. Obviously, the critical assumption that allows for analytical derivation of the maximum likelihood estimator is not that the ancestral state is known, but rather that only a single mutation segregates. Indeed, without selection equivalence of the estimators derived from the Taylor series expansion in θ of the general mutation model, estimators assuming the boundary mutation model can be shown [4]. In the latter case, the ML estimators can be considered exact. Note that mutation bias has generally been ignored when analyzing DNA sequence data. In contrast to the usual infinite sites model, which assumes that ancestral and derived states can be inferred, a mutation bias α creating an imbalance between A/T and C/G sites can be modeled naturally with our approach. As deviations in the A/T:C/G ratio from 1:1 are generally observed and as inference of ancestral states is difficult, such theory is practically useful. Vogl [3] derived a maximum likelihood estimator for not only the scaled mutation rate θ, but also the mutation bias α.

Computation-intensive, probabilistic methods for estimating population genetic parameters, such as the ones implemented in the LAMARC software package [38], are suitable for the analysis of populations governed by non-equilibrium dynamics. Our method also considers non-equilibrium population dynamics, such as changes in the effective size of the population, the mutation rate and mutation bias between different time points, but poses less computational burden. This is achieved by extending the boundary mutation model [4] to joint site frequency spectra and using modified Gegenbauer polynomials to solve the transition density $\phi(x,t)$ of the allelic proportion x at any time t. If these model assumptions hold, the method also provides maximum likelihood estimates.

Gutenkunst *et al.* [39] estimate migration rates, selection coefficients and split times from joint site frequency spectra in their program $\partial a \partial i$ by approximating the diffusion process using a numerical grid of population proportions x. Influx of mutations is modeled by "injecting ϕ density at low frequency in each population (at a rate proportional to the total mutation flux θ)". This presumably corresponds to the boundary mutation model, but seems to assume influx in equal proportions from the boundaries. Furthermore, this method is directed towards evaluating population sizes, growth rates and migration rates from joint site frequency spectra, rather than scaled mutation rates. The difference in mutation rates of the different allelic classes, *i.e.*, mutation bias, is not taken into account.

With mutations arising only at the boundaries, evolution of the allelic proportion x separates into a slowly evolving part, where the proportions of alleles at the boundaries change at a rate of θ, while the interior dynamics adjust relatively quickly to the slowly evolving boundary proportions. This process leads to a system of inhomogeneous linear differential equations. With this theory, changes in the parameters, *i.e.*, the scaled mutation rate θ or the mutation bias α, do not necessitate a recalculation of the eigensystem, unlike the approach described in Song and Steinrücken [7]. Thus, our approach speeds up computation, such that more complicated population genetic scenarios may be modeled, e.g., growing or shrinking population sizes that are commonly observed in nature. Since the equilibrium is reached at the rate of the scaled mutation rate θ, natural populations are rarely in equilibrium, and non-equilibrium dynamics need to be considered when inferring population genetic parameters.

We note that Evans *et al.* [40] also arrived at a system of inhomogeneous linear differential equations assuming the infinite sites model as defined above. Furthermore, they assume directional selection. Their analysis is based on iteration of moments, rather than on orthogonal polynomials, which leads to a recursive inhomogeneous system of differential equations. Nevertheless, the similarities between their and our approaches are readily apparent. In a follow-up study, Zivkovic *et al.* [41] apply their algorithm to human and fruit fly data.

So far, we have only discussed approaches based on the Kolmogorov (or diffusion) forward or backward equations. Another approach successfully employed for analyzing joint site frequency spectra is based on Kingman's coalescence [42,43]. With the coalescence, the starting point is the sample. Inference proceeds by summing and integrating over the sample's genealogical history. It is also possible to derive the beta-binomial equilibrium distribution with the coalescent. In more complicated cases, one or the other of the two approaches may be more suited. Generally, the coalescence seems preferable if the sample distribution can be derived fairly easily compared to the population distribution. This is the case for the infinite sites model without recombination, where only the coalescence approach has been used to derive joint site frequency spectra, as far as we are aware. Recently, Chen [44,45] and Kamm *et al.* [46] improved the computational efficiency of the coalescence approach to joint site frequency spectra. We note that changes in the mutation bias, as incorporated into our approach, have not yet been incorporated into the coalescence approach.

Our algorithm allows for inference of the mutation bias α, the scaled mutation rate θ and the time separating the samples of joint site frequency spectra. A fruit fly dataset consisting of a joint site frequency spectrum of two *Drosophila* species is analyzed to illustrate the method.

5. Conclusions

In this article, we present a method for inferring population genetic parameters from joint site frequency spectra, *i.e.*, from allelic frequencies at two (or more) time points. The parameters inferred are the time between the samples t and the mutation rate θ, scaled by the (effective) population size N. In contrast to earlier approaches [7,47,48], we assume a small mutation rate θ, such that only a mutation of single origin may segregate per site at any given time point in a sample. This assumption simplifies the mathematical treatment because, unlike with earlier approaches, changes in the parameters do not lead to a change in the eigensystem. Rather, the spatial expansion in orthogonal polynomials, specifically in Gegenbauer polynomials, remains unaffected. Compared to the case without mutations, *i.e.*, to the pure drift model, the temporal part changes from a system of homogeneous to one of inhomogenous linear differential equations. In effect, the system separates into the slowly evolving boundaries, which change at a rate of the scaled mutation rate θ, and a fast evolving interior polymorphic region, which changes at the rate of drift, conditional on the dynamics at the boundaries. We show that the eigenvectors can be derived from the general Jacobi polynomials by a Taylor series expansion. Furthermore, the equilibrium distribution corresponds to the Taylor series expansion of the equilibrium distribution for general θ. Due to the underlying boundary mutation model, parameter estimation is computationally efficient, and the method can be expanded to accommodate analysis of more complex population genetic scenarios.

Acknowledgments: The authors thank all members of the Institute of Population Genetics for support, especially Andrea Betancourt and Dominik Schrempf for discussions and suggestions. We also thank two anonymous reviewers whose suggestions helped to improve the article. The study was funded by the Austrian Science Fund (FWF): DK W1225-B20.

Author Contributions: CV conceived of the project, provided the mathematical derivations and wrote a first draft of the manuscript. JB improved the mathematical notation, downloaded, prepared and analyzed the *Drosophila* data. JB and CV jointly implemented the algorithms and wrote the final version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendices

A.1. Appendix: Modified Gegenbauer Polynomials as the Limit of Modified Jacobi Polynomials

Lemma 1. *The set of eigenvectors $H_i(x)$, for $i \geq 2$, can be derived from the modified Jacobi polynomials in Equation (9) [7] multiplied by the weight function, $w^{(\theta,\alpha)}(x)R_i^{(\theta,\alpha)}(x)$ if: (i) only terms in a Taylor expansion in θ up to zeroth order are kept in the polymorphic region $]0,1[$, while (ii) terms that, for $\theta \to 0$, vanish in the interior and converge to point masses at the boundaries are set to point masses at the boundaries; compactly,*

$$w^{(\theta,\alpha)}(x)R_i^{(\theta,\alpha)}(x) = H_i(x) + O(\theta)\,. \tag{A1}$$

Proof. The modified Jacobi polynomials times the weight functions can be expressed as a sum of polynomials, which in turn can be expressed as a sum of beta distributions:

$$\begin{aligned} w^{(\theta,\alpha)}(x)R_i^{(\theta,\alpha)}(x) &= \sum_{m=0}^{i} \frac{(-1)^{i-m}\Gamma(i+\alpha\theta)\Gamma(i+\beta\theta)}{\Gamma(m+1)\Gamma(i-m+1)\Gamma(m+\alpha\theta)\Gamma(i-m+\beta\theta)} \\ &\quad \cdot x^{m+\alpha\theta-1}(1-x)^{i-m+\beta\theta-1} \\ &= \sum_{m=0}^{i} \frac{(-1)^{i-m}\Gamma(i+\alpha\theta)\Gamma(i+\beta\theta)}{\Gamma(m+1)\Gamma(i-m+1)\Gamma(i+\theta)}\, beta(x\,|\,m+\alpha\theta, i-m+\beta\theta)\,. \end{aligned} \tag{A2}$$

For $m = 0$, the beta distribution converges to a delta function for small θ and $i \geq 1$:

$$\begin{aligned} \frac{(-1)^i\Gamma(i+\alpha\theta)\Gamma(i+\beta\theta)}{\Gamma(1)\Gamma(i+1)\Gamma(i+\theta)}\, beta(x\,|\,\alpha\theta, i+\beta\theta) &= \frac{(-1)^i\Gamma(i)\Gamma(i)}{\Gamma(1)\Gamma(i+1)\Gamma(i)}\delta(x) + O(\theta) \\ &= \frac{(-1)^i}{i}\delta(x) + O(\theta)\,, \end{aligned} \tag{A3}$$

and analogously for $m = i$.

For $i = 1$, we therefore have:

$$w^{(\theta,\alpha)}(x)R_1^{(\theta,\alpha)}(x) = -\delta(x) + \delta(x-1) + O(\theta)\,. \tag{A4}$$

For $i \geq 2$, we have:

$$\begin{aligned} w^{(\theta,\alpha)}(x)R_i^{(\theta,\alpha)}(x) &= \sum_{m=0}^{i} \frac{(-1)^{i-m}\Gamma(i+\alpha\theta)\Gamma(i+\beta\theta)}{\Gamma(m+1)\Gamma(i-m+1)\Gamma(m+\alpha\theta)\Gamma(i-m+\beta\theta)} \\ &\quad \cdot x^{m+\alpha\theta-1}(1-x)^{i-m+\beta\theta-1} \\ &= \sum_{m=1}^{i-1} \frac{(-1)^{i-m}\Gamma(i)\Gamma(i)}{\Gamma(m+1)\Gamma(i-m)\Gamma(i-m+1)\Gamma(m)} \\ &\quad \cdot x^{m-1}(1-x)^{i-m-1} + (-1)^i\delta(x)/i + \delta(x-1)/i + O(\theta) \\ &= \sum_{m=0}^{i-2} \frac{(-1)^{i-m-1}\Gamma(i)\Gamma(i)}{\Gamma(m+2)\Gamma(i-m-1)\Gamma(i-m)\Gamma(m+1)} \\ &\quad \cdot x^{m}(1-x)^{i-m-2} + (-1)^i\delta(x)/i + \delta(x-1)/i + O(\theta) \\ &= U_i(x) + (-1)^i\delta(x)/i + \delta(x-1)/i + O(\theta) \\ &= H_i(x) + O(\theta)\,. \end{aligned} \tag{A5}$$

□

Remark 1. The $H_i(x)$ are obviously independent of θ and α for $i \geq 1$.

Note that the integrals over the whole region, including the boundary terms, are:

$$\begin{cases} -\int_0^1 xH_i(x)\,dx = 0 \\ -\int_0^1 (1-x)H_i(x)\,dx = 0\,; \end{cases} \tag{A6}$$

such that the probability masses at the boundaries exactly offset that in the interior.

A.2. Appendix: Mutation-Drift Equilibrium

Theorem 2. *The equilibrium solution of the dynamic system with the slowly evolving part given by Equation (39) and the boundary Conditions (12) and (13) is given by (52) in the limit of $N \to \infty$.*

Proof. For any starting value, $b_0(t)$ will converge to $b_0(\infty) = \beta$ and similarly $b_1(\infty) = \alpha$. Substituting these values into Equation (50) and setting the derivatives to zero, we obtain:

$$\tau_i(\infty) = \frac{A_i}{\lambda_i} \tag{A7}$$

It follows that, for all odd i, $\tau_i(\infty) = 0$, and, for all even i,

$$\tau_i(\infty) = -\alpha\beta\theta(4i-2)/(i-1)\,. \tag{A8}$$

The function:

$$\phi(x,\infty) = \beta\delta(x) + \alpha\delta(x-1) + \alpha\beta\theta\sum_{i=1}^{\infty} c_{2i}\,H_{2i}(x) \tag{A9}$$

corresponds to the modified Gegenbauer expansion of the equilibrium solution for $N \to \infty$ where:

$$c_{2i} = \frac{1}{\Delta_{2i}}\int_0^1 x(1-x)U_{2i}(x)x^{-1}(1-x)^{-1}\,dx = -\frac{(4i-1)2i}{2i-1}\frac{2}{2i} = -\frac{4(2i)-2}{2i-1}\,. \tag{A10}$$

□

References

1. Fisher, R. *The Genetical Theory of Natural Selection*; Clarendon Press: Oxford, UK, 1930.
2. Wright, S. Evolution in Mendelian populations. *Genetics* **1931**, *16*, 97–159.
3. Vogl, C. Estimating the scaled mutation rate and mutation bias with site frequency data. *Theor. Popul. Biol.* **2014**, *98*, 19–27.
4. Vogl, C.; Bergman, J. Inference of directional selection and mutation parameters assuming equilibrium. *Theor. Popul. Biol.* **2015**, *106*, 71–82.
5. Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **1955**, *41*, 144–150.
6. Griffiths, R.; Spanò, D. Diffusion processes and coalescent trees. In *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*; Cambridge University Press: Cambridge, UK, 2010; pp. 358–375.
7. Song, Y.; Steinrücken, M. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* **2012**, *190*, 1117–1129.
8. Tran, T.; Hofrichter, J.; Jost, J. An introduction to the mathematical structure of the Wright-Fisher model of population genetics. *Theory Biosci.* **2013**, *132*, 73–82.
9. Vogl, C. Computation of the likelihood in biallelic diffusion models using orthogonal polynomials. *Computation* **2014**, *2*, 199–220.
10. Vogl, C.; Clemente, F. The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theor. Popul. Genet.* **2012**, *81*, 197–209.
11. Parsch, J.; Novozhilov, S.; Saminadin-Peter, S.; Wong, K.; Andolfatto, P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol. Biol. Evol.* **2010**, *27*, 1226–1234.
12. Clemente, F.; Vogl, C. Unconstrained evolution in short introns?—An analysis of genome-wide polymorphism and divergence data from *Drosophila*. *J. Evol. Biol.* **2012**, *25*, 1975–1990.
13. Clemente, F.; Vogl, C. Evidence for complex selection on four-fold degenerate sites in *Drosophila melanogaster*. *J. Evol. Biol.* **2012**, *25*, 2582–2595.
14. Lack, J.; Cardeno, C.; Crepeau, M.; Taylor, W.; Corbett-Detig, R.B.; Stevens, K.; Langley, C.; Pool, J. The *Drosophila* Genome Nexus: A population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* **2015**, *199*, 1229–1241.
15. NCBI Updates of Drosophila Annotations. Available online: http://www.flybase.org/ (accessed on 21 October 2015).
16. Carlin, B.; Louis, T., Eds. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed.; Chapman and Hall: Boca Raton, FL, USA, 2000.
17. RoyChoudhury, A.; Wakeley, J. Sufficiency of the number of segregating sites in the limit under finite-sites mutation. *Theor. Popul. Biol.* **2010**, *78*, 118–122.
18. Ewens, W. A note on the sampling theory for infinite alleles and infinite sites models. *Theor. Popul. Biol.* **1974**, *6*, 143–148.
19. Watterson, G. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **1975**, *7*, 256–276.
20. Kimura, M. Diffusion models in population genetics. *J. Appl. Probab.* **1964**, *1*, 177–232.
21. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **1969**, *61*, 893–903.
22. Chan, A.; Jenkins, P.; Song, Y. Genome-Wide Fine-Scale Recombination Rate Variation in *Drosophila melanogaster*. *PLoS Genet.* **2012**, *8*, e1003090.
23. Campos, J.L.; Zeng, K.; Parker, D.; Charlesworth, B.; Haddrill, P. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in *Drosophila melanogaster*. *Mol. Biol. Evol.* **2013**, *30*, 811–823.
24. Campos, J.L.; Halligan, D.L.; Haddrill, P.R.; Charlesworth, B. The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. *Mol. Biol. Evol.* **2014**, *31*, 1010–1028.
25. Ewens, W. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **1972**, *3*, 87–112.
26. Sawyer, S.; Hartl, D. Population genetics of polymorphism and divergence. *Genetics* **1992**, *132*, 1161–1176.

27. Bustamante, C.; Wakeley, J.; Sawyer, S.; Hartl, D. Directional selection and the site-frequency spectrum. *Genetics* **2001**, *159*, 1779–1788.
28. Bustamante, C.; Nielsen, R.; Hartl, D. Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* **2003**, *63*, 91–103.
29. Williamson, S.; Fledel-Alon, A.; Bustamante, C. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* **2004**, *168*, 463–475.
30. Ewens, W. *Mathematical Population Genetics*; Springer: New York, NY, USA, 1979.
31. Abramowitz, M.; Stegun, I. (Eds.) *Handbook of Mathematical Functions*, 9th ed.; Dover: Mineola, NY, USA, 1970.
32. Zhao, L.; Yue, X.; Waxman, D. Complete numerical solution of the diffusion equation of random genetic drift. *Genetics* **2013**, *194*, 973–985.
33. Ewens, W. *Mathematical Population Genetics*, 2nd ed.; Springer: New York, NY, USA, 2004.
34. Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. *Bayesian Data Analysis*; Chapman & Hall: London, UK, 1995.
35. Lachaise, D.; Cariou, M.; David, J.; Lemeunier, F.; Tsacas, L.; Ashburner, M. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* **1988**, *22*, 159–225.
36. Russo, C.; Takezaki, N.; Nei, M. Molecular phylogeny and divergence times of *Drosophilid* species. *Mol. Biol. Evol.* **1995**, *12*, 391–404.
37. Cutter, A. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol. Biol. Evol.* **2008**, *25*, 778–786.
38. Kuhner, M. LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **2006**, *15*, 768–770.
39. Gutenkunst, R.; Hernandez, R.; Williamson, S.; Bustamante, C. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet.* **2009**, *5*, e1000695.
40. Evans, S.; Shvets, Y.; Slatkin, M. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.* **2007**, *71*, 109–119.
41. Zivkovic, D.; Steinrücken, M.; Song, Y.; Stephan, W. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics* **2015**, *200*, 601–617.
42. Hein, J.; Schierup, M.; Wiuf, C. *Gene Genealogies, Variation, and Evolution: A Primer in Coalescent Theory*; Oxford University Press: Oxford, UK, 2005.
43. Wakeley, J. *Coalescent Theory: An Introduction*; Roberts and Co.: Greenwood Village, CO, USA, 2009.
44. Chen, H. The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theor. Popul. Biol.* **2012**, *81*, 179–195.
45. Chen, H. Intercoalescence time distribution of incomplete gene genealogies in temporally varying populations and applications in population genetic inference. *Ann. Hum. Genet.* **2013**, *77*, 158–173.
46. Kamm, J.; Terhorst, J.; Song, Y. Efficient computation of the joint sample frequency spectra for multiple populations. 2015, arXiv: 1503.01133. Available online: http://arxiv.org/abs/1503.01133 (accessed on 3 March 2015).
47. Steinrücken, M.; Wang, R.; Song, Y. An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. *Theor. Popul. Biol.* **2013**, *83*, 1–14.
48. Steinrücken, M.; Bhaskar, A.; Song, Y. A novel method for inferring general diploid selection from time series genetic data. *Ann. Appl. Stat.* **2014**, *8*, 2203–2222.