A Wright–Fisher graph model and the impact of directional selection on genetic variation

Ingemar Kaj, Carina Mugal, Rebekka Müller-Widmann

Theoretical Population Biology 159 (2024) 13–24

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A Wright–Fisher graph model and the impact of directional selection on genetic variation

Ingemar Kaj [a,*], Carina F. Mugal [b,c], Rebekka Müller-Widmann [a]

[a] Department of Mathematics, Uppsala University, Uppsala, Sweden
[b] Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden
[c] Laboratory of Biometry and Evolutionary Biology, University of Lyon 1, UMR CNRS 5558, Villeurbanne, France

ARTICLE INFO

Keywords:
Wright–Fisher jump–diffusion process
Directional selection
Mutation bias
Genetic diversity
Effective mutation rate
Theoretical population genetics

ABSTRACT

We introduce a multi-allele Wright–Fisher model with mutation and selection such that allele frequencies at a single locus are traced by the path of a hybrid jump–diffusion process. The state space of the process is given by the vertices and edges of a topological graph, i.e. edges are unit intervals. Vertices represent monomorphic population states and positions on the edges mark the biallelic proportions of ancestral and derived alleles during polymorphic segments. In this setting, mutations can only occur at monomorphic loci. We derive the stationary distribution in mutation–selection–drift equilibrium and obtain the expected allele frequency spectrum under large population size scaling. For the extended model with multiple independent loci we derive rigorous upper bounds for a wide class of associated measures of genetic variation. Within this framework we present mathematically precise arguments to conclude that the presence of directional selection reduces the magnitude of genetic variation, as constrained by the bounds for neutral evolution.

1. Introduction

The degree of genetic variation within and among populations is determined by the interaction of a number of evolutionary processes, such as mutation, genetic drift, and natural selection. As a consequence, patterns of genetic variation contain information about the processes that have shaped them and are widely used in population genomic studies to infer the evolutionary history of the studied population(s) or species. At present, single nucleotide variants (SNVs) are most frequently applied to study patterns of genetic variation (Haasl and Payseur, 2015; Bourgeois and Warren, 2021). On the one hand, SNVs are easier accessible with conventional sequencing technologies than, for example, larger structural variants (Hu et al., 2021). On the other hand, the predominantly biallelic state of SNVs allows for mathematical models that can be efficiently implemented in population genetic inference. Several summary statistics of SNVs, as for example nucleotide diversity, have been derived to describe the degree of genetic variation in a population. Alternatively, haplotype structure, which considers associations among SNVs (Zhao et al., 2003; Garg et al., 2022), can be examined. Practically, this requires phasing of variants and a different mathematical description than the per-site focus of SNVs.

The essentially biallelic state of SNVs is a consequence of an overall small mutational input, which is consistent with observations in empirical data where multi-allelic single nucleotide variation is typically rare (Cao et al., 2015; Phillips et al., 2015). This means, the mutational input at single nucleotide sites is generally small enough to prevent additional allelic types at a locus which is already polymorphic. Hence, in the context of single nucleotide variation, mutation, the fundamental source of genetic variation, is modeled as a so-called "boundary mutation model", where mutations can only occur at monomorphic loci (Kimura, 1969; Sawyer and Hartl, 1992; McVean and Charlesworth, 1999). Traditionally, this mutation mechanism is also called "non-recurrent" (Wright, 1931) in contrast to recurrent mutation, which is ongoing during polymorphic periods and which is not restricted to biallelic states. Boundary mutation initializes the segregation of an allele in the population but otherwise does not influence the population frequency of the allele. Genetic drift and natural selection, on the other hand, control the time span over which mutations segregate in a population until eventually reaching fixation or extinction. While genetic drift ultimately acts to eliminate genetic variation, different selection mechanisms can either prolong or shorten the time to fixation or extinction. Directional selection, where one of the alleles in a given pair of allelic types has a selective advantage over the other, is commonly viewed as a force to reduce the level of genetic variation. However, as pointed out in Novak and Barton (2017), "rigorous arguments for this idea are scarce".

* Correspondence to: Department of Mathematics, Box 480, SE 751 06 Uppsala, Sweden.
E-mail address: ikaj@math.uu.se (I. Kaj).

https://doi.org/10.1016/j.tpb.2024.07.004
Received 17 February 2023
Available online 15 July 2024

Population genetics modeling for the purpose of analyzing genetic variation under the combined influence of different evolutionary processes typically builds on some version of Wright–Fisher models (Fisher, 1930; Wright, 1931, 1938) or Moran type models (Moran, 1958). As pioneered by Kimura (1964), diffusion approximation techniques under scaling of evolutionary time and large population size are instrumental and helped advance the understanding of the distribution of inherited allele frequencies, both dynamically and under steady-state (see e.g. Durrett, 2008; Etheridge, 2011). In this work we apply diffusion approximation methods to study a multi-allele and multi-locus model with non-recurrent, reversible mutation and directional selection in an isolated population assuming independence among loci. Mutation is reversible since we work with a fixed, finite number of allelic types and all mutation events involving a given pair of alleles may take place in both directions. Within this framework our objective is essentially to show that the presence of directional selection reduces the magnitude of genetic variation, as constrained by the bounds for neutral evolution. To this end, we derive stationary distributions over monomorphic and polymorphic states in mutation–selection–drift equilibrium. Closed form expressions for the expected allele frequency spectrum are obtained asymptotically under large population size scaling to interpret the behavior of the process and to extract biologically relevant conclusions. Theorem 2 provides rigorous upper bounds for a wide class of associated measures of genetic variation under the influence of directional selection, implying that directional selection constrains the number of polymorphisms in a population. Moreover, under the assumption that mutation rates are equal among types, Corollary 2 shows that genetic variation is constrained by the upper bounds for neutral evolution. In addition we discuss how the interaction of mutation bias and fixation bias impacts the results.

To put our approach in context, the extension from studying the evolutionary dynamics of genetic loci with two types to general multi-allele Wright–Fisher models with a fixed number $K \geq 2$ possible allelic states for each genetic locus, can be traced back to Wright (1949). For the case of recurrent mutation mechanisms such K-allele models have been developed in much detail. The state space for single locus frequencies is now (a subset of) the K-simplex, which presents considerable challenges in extracting useful probabilistic information. For a brief history of K-allele Wright–Fisher models with recurrent mutation, we refer to Ferguson and Buzbas (2018), and for some of the mathematical results to Etheridge (2011, Ch. 4 and 5). A recent approximation approach to multi-allele models with recurrent mutation (Burden and Tang, 2016; Ferguson and Buzbas, 2018) starts from the presumption that mutation events are rare on the time scale of evolution relevant for the diffusion approximation. Then, with sufficiently small mutation rates, the allele frequencies will be mostly concentrated either on the vertices of the K-simplex or on the edges connecting a pair of mutating alleles. Only a small fraction of probability mass remains on simplex domains that allow three or more alleles existing simultaneously. Hence, in studies for which the latter case is negligible, it suffices to consider a model with a simpler state space constituting merely the relevant graph-shaped subset of the K-simplex. Only then known properties and tools of one-dimensional diffusion theory, as for example the Green's function, can be used.

Our approach towards modeling the multi-allelic case is a jump–diffusion process, biallelic by construction, with exactly the graph-shaped subset of the K-simplex as state space. Some key features of the jump–diffusion process are already implemented in Mugal et al. (2014) and Kaj and Mugal (2016) for the simpler setting of arbitrary ancestral-and-derived alleles. In the graph model, the vertices correspond to the presence of a specific fixed type (or allele) in a genetic locus and positions on the edges between two types represent continuous polymorphic states. Similar boundary mutation multi-allele models have been discussed in the context of synonymous codon usage (Zeng, 2010) or so-called polymorphism-aware phylogenetic models (De Maio et al., 2013; Borges et al., 2019), with a focus on methodological development for statistical inference from genomic data.

2. A Wright–Fisher graph model

A continuous time Markov process, with state space constructed from a connected, directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V}$ and (continuous unit length) edges $\mathcal{E}$, captures the random change of allelic types at a single locus. Such a locus can be of abstract nature (mutant versus wild type) or specific, for instance consisting of l nucleotides in the genome, where $l = 1$ corresponds to a single site on the genome or $l = 3$ to nucleotide triplets, such as protein-coding codons. For larger l the assumption of non-recurrent mutation is likely violated. Our results therefore primarily apply to single nucleotide sites or nucleotide triplets. The term "graph" is used in a topological sense, where edges are identified with unit intervals (Hatcher, 2018). The finite vertex set represents the available allelic types at the locus, while the family of continuous edges allows for keeping track of the possible mutations among types and any polymorphic state. Each edge is a directed interval of length one, starting in a vertex $u \in \mathcal{V}$ and leading to another vertex $v \in \mathcal{V}$, such that the position on the edge records the relative frequency of type v as a mutant derived from ancestral type u. The relevant graph model is a hybrid jump and diffusion process with compact state space, in which the open edges form a continuous interior and the vertices are discrete boundary points. Mutation events occur only on the boundary. Each mutation is succeeded by a polymorphic phase of two alleles co-existing in the population, upon which the Wright–Fisher diffusion determines the frequency and subsequent extinction and fixation probabilities of the mutant. The graph in Fig. 1 illustrates an example state space on which the process moves.

Formally, we consider a Markov process $X = (X_t)_{t\geq 0}$ encoded by a triplet $X_t = (U_t, V_t, Y_t) \in \mathcal{V} \times \mathcal{V} \times [0, 1)$ with values in the compact state space D formed by a directed topological graph. Such a graph is obtained from a classical (discrete) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by identifying the vertex set with a discrete set of points and replacing edges with unit intervals $[0, 1)$ that are directed from 0 to 1 and glued together with the points in the vertex set. The boundary of the state space consists of the point set $\partial D = \{(u, u, 0) : u \in \mathcal{V}\}$, where the boundary state $(u, u, 0)$ represents a monomorphic locus at which the entire population has the same allelic type $u \in \mathcal{V}$. A state (u, v, y) within the interior of the state space

$$D^\circ = \{(u, v, y) : u \in \mathcal{V}, v \in \mathcal{V}, u \neq v, 0 < y < 1\}$$

lies on the directed edge leading from vertex state $(u, u, 0)$ to vertex state $(v, v, 0)$. It arises when a mutation from u to v occurred and brought mutants of type v to be present in the population at relative frequency in the infinitesimal interval $(y, y + dy)$. Consequently, the interior state (v, u, y), located on the complementary edge directed in the opposite direction from v to u, assigns relative frequency y to mutant type u derived from an ancestral v. Finally, the closure $D = D^\circ \cup \partial D$ is reached along the limits

$$(u, v, y) \to \begin{cases} (u, u, 0) \in \partial D, & y \to 0, \\ (v, v, 0) \in \partial D, & y \to 1, \end{cases} \qquad (u, v, y) \in D^\circ.$$

2.1. Reversible boundary mutation

The mutation mechanism of the process X is specified by a fixed entry point $x \in (0, 1)$ and a family of nonnegative mutation rate parameters $\{\lambda_{uv} : u, v \in \mathcal{V}, v \neq u\}$. Here, x represents the fraction of a population that is affected by a single mutation. We thus define λ_{uv}/x as the population mutation intensity from u to v per evolutionary time unit, where the evolutionary time unit corresponds to "x^{-1} generations". This means λ_{uv} can be thought of as the population mutation rate "per generation", which commonly represents the macroscopic mutation rate.

The graph edge set consists of those edges e_{uv} between types for which the mutation intensity is positive, $\mathcal{E} = \{e_{uv} : u, v \in \mathcal{V}, \lambda_{uv} > 0\}$. We postulate that every type is essential, that all mutations are reversible, and that the graph $\mathcal{G}$ is irreducible, by assuming that the mutation rates satisfy the conditions

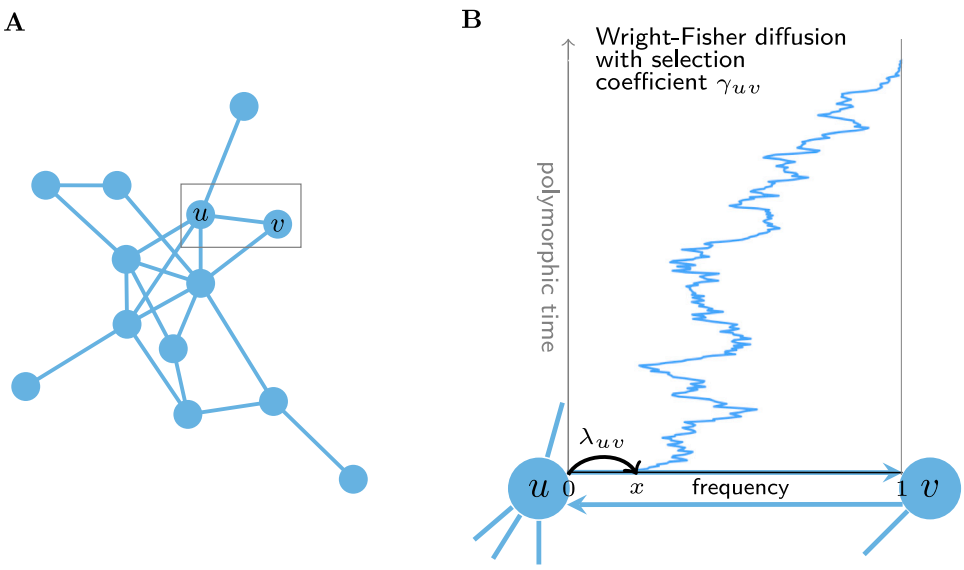

Fig. 1. Panel A: The shape of a connected graph (visualized in three-dimensional space) on which the process X moves. Between each pair of connected vertices there are two directed edges equipped with unit frequency scales. Panel B: Zoom-in of the rectangular box in panel A with directed, unit length edges between type u and v. A mutation on the boundary from type u to type v occurs with intensity λ_{uv}. The initial frequency of the mutant v is a fixed value $x \in (0,1)$ which is also the initial value for a Wright–Fisher diffusion with selection coefficient γ_{uv} that is started by the mutation. The diffusion process either goes to fixation in type v or to extinction in u.

(i) $\lambda_u := \sum_{v: v\neq u} \lambda_{uv} > 0$ for each $u \in \mathcal{V}$,
(ii) $\lambda_{uv} > 0 \iff \lambda_{vu} > 0$,
(iii) for each pair of allelic types, $u, v \in \mathcal{V}$, there is a sequence of mutations $u \to w_1 \to \cdots \to w_n \to v$, such that $\lambda_{u,w_1} \cdot \ldots \cdot \lambda_{w_n,v} > 0$.

The hybrid jump and diffusion mechanism of X is such that, starting in a boundary point $z = (u,u,0) \in \partial D$, the process holds during an exponential time with intensity λ_u/x. It then jumps to an interior point $z' = (u,v,x) \in D^\circ$ governed by jump rates λ_{uv}, which represents mutant type v entering the population at (continuous) fraction x. Assuming that a jump from $(u,u,0)$ to (u,v,x) occurs at time r, the interior trajectory of X is a continuous path

$$X_t = (u, v, Y_t^r), \quad r \le t < \tau,$$

where Y_t^r, $t \ge r$, is a diffusion in D starting from $Y_r^r = x$ such that the path $\xi_s^r = Y_{r+s}^r$, $s \ge 0$, is a Wright–Fisher process with selection, which solves

$$d\xi_s = \gamma_{uv}\, \xi_s(1-\xi_s)\, ds + \sqrt{\xi_s(1-\xi_s)}\, dB_s, \quad \xi_0 = x. \tag{1}$$

We denote by τ the first exit time of the interior state space D°, i.e. $X_\tau = (u,u,0)$ if ξ_s is absorbed in 0 and $X_\tau = (v,v,0)$ if absorbed in 1. The former case is extinction and the latter case is fixation of the allelic type v. For edges $e_{uv} \in \mathcal{E}$ the parameter γ_{uv} denotes the selection coefficient for mutations from u to v. Moreover, $(B_s)_{s\ge 0}$ is a standard Brownian motion. After absorption of the diffusion, $(Y_t)_{t\ge\tau} = 0$ and the process holds at 0 until a new mutation and a new independent diffusion segment on $[0,1)$ is initialized. The process described here is a version of the elementary return process (Feller, 1954), extended from a single edge between two types to the multi-allele graph model. Fig. 1B depicts the hybrid jump and diffusion setup.

2.2. Directional selection

To express genic selection in the model we let the family of selection coefficients γ_{uv} assigned to the edges e_{uv} in the graph satisfy the anti-symmetric condition

$$\gamma_{uv} = -\gamma_{vu}, \quad e_{uv} \in \mathcal{E}. \tag{2}$$

The central instance is directional selection based on a static fitness landscape, where each allelic type is assigned a (time-independent) fitness level F_u, $u \in \mathcal{V}$, and each polymorphic pair of alleles (u,v) has relative selection coefficient

$$\gamma_{uv} = F_v - F_u, \quad e_{uv} \in \mathcal{E}. \tag{3}$$

This implies that $\gamma_{uv} = -\gamma_{vu}$. Consequently, in each pair the selective advantage of the type with the higher fitness equals the selective disadvantage of the other type. Another relevant example of anti-symmetric selection coefficients is the preferential fixation of strong (C and G) over weak (A and T) nucleotides due to the process of GC-biased gene conversion (gBGC) (Duret and Galtier, 2009; Mugal et al., 2015), which analytically is equivalent to directional selection (Nagylaki, 1983).

We note that under assumption (2) the distribution of the process X simplifies on each pair of edges e_{uv} and e_{vu} through the equality in distribution

$$(u, v, Y_t^r) \overset{d}{=} (v, u, 1 - Y_t^r), \quad r \le t < \tau.$$

2.3. Properties of the Wright–Fisher graph process

Green's function. The graph process X restricted to a particular edge $e_{uv} \in \mathcal{E}$, is a classical Wright–Fisher diffusion, described in Eq. (1), with selection coefficient γ_{uv}. Abbreviating $\gamma = \gamma_{uv}$, we write $\mathbb{P}_x^\gamma$ for the probability measure and $\mathbb{E}_x^\gamma$ for the expectation of the process starting at x, and select the associated scale function as $S_\gamma(x) = (1-e^{-2\gamma x})/(2\gamma), \gamma \neq 0$, $S_0(x) = x$, and speed function as $m_\gamma(x) = e^{2\gamma x}/(x(1-x))$. The state space of the diffusion on the segment e_{uv} is an interval $[0,1)$ where 0 is the boundary of an elementary return process and 1 is an absorbing boundary (Feller, 1954). In terms of the graph structure, 0 is the vertex $(u,u,0)$ and 1 is identified with the vertex $(v,v,0)$ via the construction where edges are glued together. Since m_γ is not integrable near 0 or 1, both points $\{0,1\}$ are exit boundary points and attainable from the interior of the state space (Karlin and Taylor, 1981). The time τ_0 required to reach 0 is the extinction time, the time τ_1 to reach 1 the fixation time, and $\tau = \min(\tau_0, \tau_1)$ is the exit time of the interior interval $(0,1)$. The corresponding fixation probability $q_\gamma(x) = \mathbb{P}_x^\gamma(\tau_1 < \tau_0)$ equals $q_\gamma(x) = S_\gamma(x)/S_\gamma(1) = (1-e^{-2\gamma x})/(1-e^{-2\gamma})$, $\gamma \neq 0$, $q_0(x) = x$ (Kimura, 1962). The Green's function $G_\gamma(x,y)$ is defined by

$$G_\gamma(x,y) = \begin{cases} 2q_\gamma(x)(S_\gamma(1) - S_\gamma(y))m_\gamma(y), & 0 \le x \le y \le 1, \\ 2(1-q_\gamma(x))(S_\gamma(y) - S_\gamma(0))m_\gamma(y), & 0 \le y \le x \le 1. \end{cases} \tag{4}$$

Assuming we start from $z = (u, v, x) \in D^\circ$, the possible transitions from z to z' before exiting D° are those such that $z' = (u, v, y)$ for some y, $0 < y < 1$. For such a pair, the Green's function is $G(z, z') = G_{\gamma_{uv}}(x, y)$ and governs the transition of the process from z towards z'. For background on mathematical population genetics and more detailed properties of the Wright–Fisher diffusion process with selection, we refer the reader to e.g. Maruyama (1977), Karlin and Taylor (1981), Ewens (2004), Etheridge (2011).

Stationary distribution on the boundary. By replacing the polymorphic excursions of X with instantaneous jumps, we obtain an embedded continuous time Markov chain. Indeed, starting in $u \in \mathcal{V}$ the embedded chain holds during an exponential time with rate λ_u/x, then with probability λ_{uv}/λ_u picks type v and with probability $q_{\gamma_{uv}}(x)$ jumps to the new type v. Taken together, the transition rate of the embedded chain from u to v is h_{uv}/x, where $h_{uv} = \lambda_{uv} q_{\gamma_{uv}}(x)$. The stationary distribution on the boundary is a probability distribution $\eta^x = \{\eta^x(z) = \eta^x_u : z = (u, u, 0) \in \partial D\}$, which satisfies the relationship

$$\eta^x_u h_{uv} = \eta^x_u \lambda_{uv} q_{\gamma_{uv}}(x) = \eta^x_v \lambda_{vu} q_{\gamma_{vu}}(x) = \eta^x_v h_{vu}, \quad e_{uv}, e_{vu} \in \mathcal{E}, \tag{5}$$

a detailed balance equation across each pair of edges of the graph. Since the state space is finite, there exists a unique stationary and reversible distribution η^x on the boundary, typically associated with time-reversibility.

We are now in position to connect the probability weights on the boundary given by η^x with the occupation measure in the interior of the state space as provided by the Green's function. If the process starts from the boundary according to the reversible distribution η^x, i.e. the initial distribution of X_0 is η^x, then the initial position of the process after the first mutation is governed by the measure

$$\nu^x = \sum_{u,v \in \mathcal{V}} \eta^x_u \lambda_{uv} \delta_{(u,v,x)}. \tag{6}$$

In words, the intensity of the first jump is determined by the accumulation of jump intensities over all vertices and its outgoing edges, weighted by the probability to be in a specific vertex. The subsequent relative position on a particular edge is then given by the fixed entry point x. Under ν^x, the relevant Green's function contribution on the particular edge e_{uv} regarding the transition from $z = (u, v, x)$ to $z' = (u, v, y)$ is

$$G(\nu^x, z') = \eta^x_u \lambda_{uv} G_{\gamma_{uv}}(x, y), \tag{7}$$

which also justifies writing

$$G(\nu^x, dz') = \eta^x_u \lambda_{uv} G_{\gamma_{uv}}(x, y)\, dy. \tag{8}$$

The generator. We consider real-valued functions f defined on D, writing $f(z) = f_{uu}(0)$ for $z = (u, u, 0) \in \partial D$ and $f(z) = f_{uv}(y)$ for $z = (u, v, y) \in D^\circ$, and let $f'_{uv}(y)$ and $f''_{uv}(y)$ denote first and second order derivatives with respect to y defined in the interior D° of D. The infinitesimal generator of the Markov process X is the operator $\mathcal{L}$ which acts on a suitable domain $\mathcal{D}$ of functions $f : D \to \mathbb{R}$ by

$$\mathcal{L} f(z) = \sum_{v \in \mathcal{V}} \lambda_{uv}(f_{uv}(x) - f_{uu}(0)), \quad z = (u, u, 0) \in \partial D,$$

and

$$\mathcal{L} f(z) = \gamma_{uv} y(1 - y) f'_{uv}(y) + \frac{1}{2} y(1 - y) f''_{uv}(y), \quad z = (u, v, y) \in D^\circ.$$

In the following we consider functions $f \in \mathcal{F}$, where $\mathcal{F}$ is a subset of D consisting of real-valued bounded functions on D, twice continuously differentiable in the interior D°, and such that $f_{uv}(y) \to f_{uu}(0)$ as $y \to 0$ and $f_{uv}(y) \to f_{vv}(0)$ as $y \to 1$, $u, v \in \mathcal{V}$. The Wright–Fisher graph process X is therefore composed of successive segments of well-defined, classical Wright–Fisher diffusion processes alternating with exponential waiting times in the boundary points. The order of visiting vertices and edges in the graph is determined by the initial distribution and by the mutation jump rates. The resulting jump–diffusion process X inherits the strong Markov property from its components. The dynamics of X in time can be visualized as a single particle moving around on the graph elements. During those times the particle spends in a vertex, the population is monomorphic of the corresponding type. During segments of particle diffusion on a directed edge, the population is polymorphic accordingly.

2.4. Stationary distribution

In order to determine the equilibrium behavior of the Wright–Fisher graph process we will construct a stationary distribution, i.e. a distribution which is preserved under the time-dynamics of the model and hence represents the typical probability weight assigned to the various parts of the graph in steady-state. It can be shown in addition that the graph process satisfies exponential ergodicity and that the stationary distribution is the unique limit distribution. However, the proof of exponential ergodicity is extensive and outside the scope of the work at hand.

For each edge $e_{uv} \in \mathcal{E}$ let $D_{e_{uv}} = \{(u, v, y) : u, v \in \mathcal{V}, 0 \le y < 1\}$ be the subset of D which consists of the monomorphic state $(u, u, 0)$ and all polymorphic states $(u, v, y), 0 \le y < 1$, between u and v. Then $\cup_{e_{uv} \in \mathcal{E}} D_{e_{uv}} = D$ and the intersection of two edge sets contains any shared vertex. A measure μ on D is stationary for X, by definition, if the balance equations

$$\int_D \mathcal{L} f(z)\, \mu(dz) = \sum_{e_{uv} \in \mathcal{E}} \int_{D_{e_{uv}}} \mathcal{L} f(z)\, \mu(dz) = 0, \quad f \in \mathcal{F}, \tag{9}$$

hold. We say that the measure μ is edge-reversible for the Wright–Fisher graph process X, if the detailed balance edge equations

$$\int_{D_{e_{uv}}} \mathcal{L} f(z)\, \mu(dz) + \int_{D_{e_{vu}}} \mathcal{L} f(z)\, \mu(dz) = 0, \quad f \in \mathcal{F},$$

hold for every pair of edges $e_{uv}, e_{vu} \in \mathcal{E}$. By summing this relation over all pairs u and v, linked by the two edges e_{uv} and e_{vu}, we recover Eq. (9). Thus, an edge-reversible measure μ yields a stationary distribution of the Wright–Fisher graph process.

Theorem 1. *There exists an edge-reversible measure μ for X on D, which is given by*

$$\mu(z) = \frac{\eta^x(z)}{1 + \int_{D^\circ} G(\nu^x, dz')}, \quad z \in \partial D,$$

$$\mu(dz) = \frac{G(\nu^x, dz)}{1 + \int_{D^\circ} G(\nu^x, dz')}, \quad z \in D^\circ,$$

where $\eta^x(z)$, $z \in \partial D$, is the unique boundary distribution defined by the detailed balance equations in Eq. (5), ν^x is be the averaged jump measure in Eq. (6), both dependent on x, and $G(\nu^x, dz')$ is the measure on D° introduced in Eq. (8).

Proof. Put $\Omega = 1 + \int_{D^0} G(\nu^x, dz')$. To verify that μ is edge-reversible we need to establish for each pair of edges $e_{uv}, e_{vu} \in \mathcal{E}$ the identity

$$\begin{aligned} &\Omega\Big(\int_{D_{e_{uv}}} \mathcal{L} f(z)\, \mu(dz) + \int_{D_{e_{vu}}} \mathcal{L} f(z)\, \mu(dz)\Big) \\ &= \eta^x_u \lambda_{uv}(f_{uv}(x) - f_{uu}(0)) + \eta^x_u \lambda_{uv} \int_0^1 \mathcal{L} f_{uv}(y)\, G_{\gamma_{uv}}(x, dy) \\ &\quad + \eta^x_v \lambda_{vu}(f_{vu}(x) - f_{vv}(0)) + \eta^x_v \lambda_{vu} \int_0^1 \mathcal{L} f_{vu}(y)\, G_{\gamma_{vu}}(x, dy) = 0, \end{aligned} \tag{10}$$

$f \in \mathcal{F}$. Here, using Eq. (4),

$$\int_0^1 \mathcal{L} f_{uv}(y)\, G_{\gamma_{uv}}(x, dy) = q_{\gamma_{uv}}(x) A_{uv}(x) + (1 - q_{\gamma_{uv}}(x)) B_{uv}(x)$$

with

$$A_{uv}(x) = \int_x^1 \Big(f'_{uv}(y) + \frac{1}{2\gamma_{uv}} f''_{uv}(y)\Big)(1 - e^{-2\gamma_{uv}(1-y)})\, dy$$

and

$$B_{uv}(x) = \int_0^x \Big(f'_{uv}(y) + \frac{1}{2\gamma_{uv}} f''_{uv}(y) \Big)(e^{2\gamma_{uv}y} - 1)\, dy.$$

Partial integration twice in each of $A_{uv}(x)$ and $B_{uv}(x)$ yield, noticing that $f_{uv}(y) \to f_{vv}(0)$ as $y \to 1$,

$$A_{uv}(x) = f_{vv}(0) - f_{uv}(x) - \frac{1}{2\gamma_{uv}} f'_{uv}(x)(1 - e^{-2\gamma_{uv}(1-x)})$$

and

$$B_{uv}(x) = f_{uu}(0) - f_{uv}(x) + \frac{1}{2\gamma_{uv}} f'_{uv}(x)(e^{2\gamma_{uv}x} - 1).$$

Since

$$q_{\gamma_{uv}}(x) \frac{1 - e^{-2\gamma_{uv}(1-x)}}{2\gamma_{uv}} - (1 - q_{\gamma_{uv}}(x)) \frac{e^{2\gamma_{uv}x} - 1}{2\gamma_{uv}} = 0,$$

it follows that

$$\begin{aligned} &q_{\gamma_{uv}}(x) A_{uv}(x) + (1 - q_{\gamma_{uv}}(x)) B_{uv}(x) \\ &\quad = q_{\gamma_{uv}}(x)(f_{vv}(0) - f_{uv}(x)) + (1 - q_{\gamma_{uv}}(x))(f_{uu}(0) - f_{uv}(x)), \end{aligned}$$

which is

$$\int_0^1 \mathcal{L} f_{uv}(y)\, G_{\gamma_{uv}}(x, dy) = -(f_{uv}(x) - f_{uu}(0)) + q_{\gamma_{uv}}(x)(f_{vv}(0) - f_{uu}(0)).$$

Similarly, by symmetry,

$$\int_0^1 \mathcal{L} f_{vu}(y)\, G_{\gamma_{vu}}(x, dy) = -(f_{vu}(x) - f_{vv}(0)) + q_{\gamma_{vu}}(x)(f_{uu}(0) - f_{vv}(0)).$$

Thus, by combining Eq. (10) with the detailed balance Eq. (5) for the boundary distribution η^x,

$$\begin{aligned} &\Omega\Big(\int_{D_{e_{uv}}} \mathcal{L} f(z)\, \mu(dz) + \int_{D_{e_{vu}}} \mathcal{L} f(z)\, \mu(dz) \Big) \\ &= \eta_u^x \lambda_{uv} q_{\gamma_{uv}}(x)(f_{vv}(0) - f_{uu}(0)) + \eta_v^x \lambda_{vu} q_{\gamma_{vu}}(x)(f_{uu}(0) - f_{vv}(0)) = 0, \end{aligned}$$

and therefore $\int_D \mathcal{L} f(z)\, \mu(dz) = 0$ in view of Eq. (9). □

We are not aware of previous results of this type for any closely related model. Peng and Li (2013) obtain stationary distributions for diffusion processes defined on an open, bounded domain in $\mathbb{R}^d$ with holding and jumping from a regular boundary. These results, however, are not directly comparable or applicable to our situation, and are obtained using different techniques (Peng and Li, 2013, Theorem 4.2).

3. Large population size scaling

The polymorphic segments of the path of X through the interior D° run on the time scale of evolution, which is a characteristic of the Wright–Fisher diffusion. The generic re-scaling approach behind the Wright–Fisher diffusion approximation considers the change in frequency in a population of size N over the time span of N generations. Simultaneously, the relevant selection coefficient at the level of generations, s, is of the order $s \sim \gamma/N \to 0$, where γ is the selection coefficient of the limiting diffusion process. In the present model we have for each edge e_{uv} such a γ_{uv} as well as a remaining free parameter x. To properly adapt the holding time distribution to the evolutionary time scale, we introduce a parameter N as a proxy of population size and prescribe that the jumps into the interior of the state space have size $x = 1/N$. The time scale of the system is then set by the speed of mutation $\lambda_{uv}/x = \lambda_{uv} N$, the population mutation intensity per evolutionary time unit. Our goal in this section is to analyze the stationary distribution μ in Theorem 1 with $x = 1/N$ for large but fixed N. Specifically we identify the dominant terms in the asymptotic expansion of μ under scaling for large N and drop remainder terms of order $\ln N/N$ and smaller. During this procedure it is convenient to make a number of simplifying approximations valid formally in the limit $N \to \infty$. It is important to keep in mind however that the population size proxy N is kept as a finite model parameter.

3.1. Approximation of the stationary distribution

We recall that in our model two vertices u and v are always connected by two directed edges e_{uv} and e_{vu} whenever the jump rates between u and v are positive. For each edge e_{uv}, as $N \to \infty$, we introduce the scaled fixation probability ω_γ, where $\gamma = \gamma_{uv}$, by

$$q_\gamma(1/N) = \frac{\omega_\gamma}{N} + O\Big(\frac{1}{N^2}\Big), \quad \omega_\gamma = \frac{2\gamma}{1 - e^{-2\gamma}}, \quad \gamma \neq 0, \qquad \omega_0 = 1. \tag{11}$$

Due to assumption (2) on directional selection, we obtain the symmetry relation

$$\omega_{\gamma_{vu}} = \omega_{-\gamma_{uv}} = e^{-2\gamma_{uv}}\, \omega_{\gamma_{uv}}. \tag{12}$$

As before, the collection of jump rates $\{\lambda_{uv} : e_{uv} \in \mathcal{E}\}$ and selection coefficients $\{\gamma_{uv} : e_{uv} \in \mathcal{E}\}$ again define an embedded scaled continuous time Markov chain on $\mathcal{V}$. In analogy with the previous relation (5), the unique solution $\eta = \{\eta(z) = \eta_u : z = (u, u, 0) \in \partial D\}$ of the detailed balance equations

$$\eta_v \lambda_{vu} \omega_{\gamma_{vu}} = \eta_u \lambda_{uv} \omega_{\gamma_{uv}}, \quad e_{uv}, e_{vu} \in \mathcal{E}, \tag{13}$$

is the scaled stationary boundary distribution of the embedded Markov chain. The solution η of (13), that no longer depends on N, is a convenient approximation of the solution η^x of (5) with $x = 1/N$. The distribution of the first jump averaged over the scaled boundary distribution,

$$\nu^{1/N} = \sum_{u,v \in \mathcal{V}} \eta_u \lambda_{uv} \delta_{(u,v,1/N)}, \tag{14}$$

still depends on the initial mutation frequency $1/N$. The next result records the dominant terms in Theorem 1, where we have fixed all mutation and selection parameters, and then choose $x = 1/N$ and N large enough so that remainder terms of order $O(\ln N/N)$ and smaller may be removed.

Proposition 1. *The stationary single site distribution μ in Theorem 1 satisfies for large N the approximation*

$$\mu(z) = \mu_N(z) + O(1/N), \quad z \in D,$$

where the approximating distribution μ_N has monomorphic site probabilities

$$\mu_N(u, u, 0) = \frac{\eta_u}{\Omega'_N}, \quad u \in \mathcal{V},$$

for $z = (u, u, 0)$, polymorphic density given by

$$\mu_N(u, v, y)\, dy = \frac{2\eta_u \lambda_{uv}}{\Omega'_N} \Big\{ \frac{\omega_{\gamma_{uv}}(1 - e^{-2\gamma_{uv}(1-y)})}{2\gamma_{uv} y(1 - y)} 1_{\{1/N < y < 1\}} + (N - \omega_{\gamma_{uv}})\, 1_{\{0 < y < 1/N\}} \Big\}\, dy,$$

for $z = (u, v, y)$, and is normalized by $\Omega'_N = \Omega_N + O(\ln N/N)$, with

$$\Omega_N = 1 + 2 \sum_{u,v \in \mathcal{V}} \eta_u \lambda_{uv} (1 + \ln N + K_{\gamma_{uv}}), \tag{15}$$

where

$$K_\gamma = \omega_\gamma \int_0^1 (-\ln y)(e^{-2\gamma y} - e^{-2\gamma(1-y)})\, dy.$$

Before proving Proposition 1, we first elaborate on its consequences, state the approximate distribution μ_N for the case of neutral evolution in Remark 1, and comment on some properties of the function K_γ in Remark 2.

The stationary single site distribution μ_N consists of three categories that are recovered in the normalization factor Ω_N. The weight of monomorphic states is given by the stationary distribution on the boundary, $\eta_u, u \in \mathcal{V}$, summing up to one. Polymorphic states are split up into observable polymorphisms ($1/N < y < 1$) and those that are practically non-observable ($0 < y < 1/N$), where the latter only exist mathematically as a result of the diffusion approximation. The weight

of observable polymorphisms is given by the sum over $\ln N + K_{\gamma_{uv}}$ in Ω_N. Clearly, for large N the term $\ln N$, which represents low-frequency polymorphisms in the allele frequency spectrum, dominates over $K_{\gamma_{uv}}$. We note, however, that if $N \to \infty$, which corresponds to $x \to 0$, mutations could not at all be established in the population.

Remark 1. For the special case of neutral evolution, $\gamma_{uv} = 0$ for all $e_{uv} \in \mathcal{E}$, we have $\omega_0 = 1$ and $K_0 = 0$, so

$$\mu_N(u,u,0) = \frac{\eta_u}{\Omega'_N}, \quad u \in \mathcal{V},$$

and

$$\mu_N(u,v,y)\,dy = \frac{2\eta_u\lambda_{uv}}{\Omega'_N}\Big\{y^{-1}\,1_{\{1/N<y<1\}} + (N-1)\,1_{\{0<y<1/N\}}\Big\}\,dy,$$

for $(u,v,y) \in D^\circ$, where $\eta_u, u \in \mathcal{V}$, is the unique solution of the balance equations

$$\eta_v\lambda_{vu} = \eta_u\lambda_{uv}, \quad e_{uv}, e_{vu} \in \mathcal{E},$$

and $\Omega'_N = \Omega_N + O(1/N)$ the normalization factor under neutrality with

$$\Omega_N = 1 + 2(1+\ln N)\sum_{u\in\mathcal{V}}\eta_u\lambda_u.$$

Remark 2. First, we have $K_\gamma \geq 0$, $\gamma \geq 0$. Second, the function K_γ is odd, $K_{-\gamma} = -K_\gamma$. In particular, $K_{\gamma_{vu}} = K_{-\gamma_{uv}} = -K_{\gamma_{uv}}$. Third, K_γ grows logarithmically for large γ: with $\gamma_e = 0.5772\ldots$ denoting Euler's constant,

$$K_\gamma \leq \begin{cases} \gamma, & 0 \leq \gamma \leq 1, \\ \gamma_e + \ln 2\gamma, & \gamma \geq 1, \end{cases} \qquad K_\gamma \sim \gamma_e + \ln 2\gamma \quad \text{for large } \gamma.$$

Proof. For fixed $z' = (u,v,y)$,

$$G(\nu^{1/N}, z') = \eta_u\lambda_{uv}NG_{\gamma_{uv}}(1/N, y).$$

Here, using Eq. (11) with large N and $\gamma = \gamma_{uv}$,

$$\begin{aligned} NG_\gamma\Big(\frac{1}{N},y\Big) &= Nq_\gamma\Big(\frac{1}{N}\Big)\frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}\,1_{\{1/N<y<1\}} \\ &\quad + N\Big(1-q_\gamma\Big(\frac{1}{N}\Big)\Big)\frac{e^{2\gamma y}-1}{\gamma y(1-y)}\,1_{\{0<y<1/N\}} \\ &= \omega_\gamma\frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}\,1_{\{1/N<y<1\}} \\ &\quad + 2(N-\omega_\gamma)\,1_{\{0<y<1/N\}} + O\Big(\frac{1}{N}\Big), \end{aligned}$$

from which we obtain $\Omega'_N\mu_N(z)$, $z \in D^\circ$. Moreover,

$$N\int_0^1 G_\gamma\Big(\frac{1}{N},y\Big)\,dy = \omega_\gamma\int_{1/N}^1\frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}\,dy + 2 + O\Big(\frac{1}{N}\Big).$$

By partial integration the remaining integral evaluates to

$$\begin{aligned} &\int_{1/N}^1\{y+(1-y)\}\frac{1-e^{-2\gamma(1-y)}}{\gamma y(1-y)}\,dy \\ &\quad = \int_0^{1-1/N}\frac{1-e^{-2\gamma y}}{\gamma y}\,dy + \int_{1/N}^1\frac{1-e^{-2\gamma(1-y)}}{\gamma y}\,dy \\ &\quad = \frac{2\ln N}{\omega_\gamma} + 2\int_0^1(-\ln y)(e^{-2\gamma y}-e^{-2\gamma(1-y)})\,dy + O\Big(\frac{\ln N}{N}\Big). \end{aligned}$$

Hence,

$$N\int_0^1 G_\gamma\Big(\frac{1}{N},y\Big)\,dy = 2(1+\ln N + K_\gamma) + O\Big(\frac{\ln N}{N}\Big).$$

Integration over D° yields

$$\int_{D^0} G(\nu^{1/N}, z')\,dz' = 2\sum_{u,v\in\mathcal{V}}\eta_u\lambda_{uv}(1+\ln N + K_{\gamma_{uv}}) + O\Big(\frac{\ln N}{N}\Big).$$

The representation of an approximate distribution $\mu_N(z)$ as stated now follows from Theorem 1.

Remark 1 follows directly from Proposition 1 with $\gamma_{uv} = 0$ for all $e_{uv} \in \mathcal{E}$.

Finally, to verify the claims in Remark 2, for $\gamma > 0$,

$$\begin{aligned} \frac{K_\gamma}{\omega_\gamma} &= \int_0^{1/2}(-\ln y)(e^{-2\gamma y}-e^{-2\gamma(1-y)})\,dy \\ &\quad - \int_0^{1/2}(-\ln(1-y))(e^{-2\gamma y}-e^{-2\gamma(1-y)})\,dy \geq 0. \end{aligned}$$

The relation $\omega_{-\gamma} = e^{-\gamma}\omega_\gamma$ implies the symmetry $K_{-\gamma} = -K_\gamma$. Furthermore, the change-of-variable $x = 2\gamma y$ yields

$$K_\gamma = \frac{\omega_\gamma}{2\gamma}\int_0^{2\gamma}(-\ln x + \ln(2\gamma))e^{-x}\,dx - \frac{\omega_\gamma}{2\gamma}\int_0^{2\gamma} -\ln(1-x/2\gamma)\,e^{-x}\,dx.$$

The rightmost integral is positive and tends to zero as $\gamma \to \infty$, by an application of the monotone convergence theorem. Also, $\omega_\gamma/(2\gamma) \to 1$ as $\gamma \to \infty$. Thus,

$$K_\gamma - \ln(2\gamma) \sim \frac{\omega_\gamma}{2\gamma}\int_0^{2\gamma}(-\ln x)e^{-x}\,dx \to \gamma_e, \quad \gamma \to \infty. \quad \square$$

3.2. Allele frequency spectra

An immediate result of the large N approximation in Proposition 1 is the allele frequency spectrum (AFS), one of the most important summary statistics of the stationary distribution in population genetics to investigate genetic variation. The unfolded AFS describes the allele frequency distribution of the derived allele at a biallelic site and can be retrieved using the modeling setup with two directed edges between each pair of types. For a given locus, the unfolded AFS corresponds to $\sum_{u,v\in\mathcal{V}}\mu_N(u,v,y), 0 < y < 1$, representing the density of the derived allele frequency of any type. We visualize the polymorphic density on two directed edges e_{uv} and e_{vu} of a multi-allele model for $\gamma_{uv} = 1$ in Fig. 2A.

In practice, the unfolded AFS relies on a polarization of polymorphisms into derived and ancestral types, knowledge that requires additional information such as outgroup data, which is not always readily available. If this is the case, a folded AFS can be derived from data. The folded AFS takes biallelic observations and typically measures the minor allele at some frequency $y \in [0, 0.5]$, and the other allelic type at complementary frequency $1 - y \in [0.5, 1]$. To formalize representations of unfolded and folded allele frequency spectra using the stationary distribution in the current model, we introduce the set of unordered pairs of vertices $\mathcal{P} := \{\langle u,v\rangle : u,v \in \mathcal{V}\}$, with $|\mathcal{P}| = \binom{|\mathcal{V}|}{2}$. Restricting to the edges of the pair $\langle u,v\rangle$, the polymorphic density of the process when type v has frequency y and type u frequency $1-y$ equals

$$\mu_N(\langle u,v\rangle, 1-y, y) := \mu_N(u,v,y) + \mu_N(v,u,1-y).$$

The folded density of the minor allele on $\langle u,v\rangle$ is therefore

$$\mu_{\text{fold}}(\langle u,v\rangle, y) := \mu_N(\langle u,v\rangle, 1-y, y) + \mu_N(\langle v,u\rangle, 1-y, y), \quad 0 < y \leq 1/2.$$

Fig. 2B depicts the polymorphic density $\mu_N(\langle u,v\rangle, 1-y, y)$ on edges connecting a pair $\langle u,v\rangle$.

Corollary 1. *We have*

$$\begin{aligned} \mu_N(\langle u,v\rangle, 1-y, y)\,dy = \frac{2\eta_u\lambda_{uv}}{\Omega'_N}\Big\{&\frac{e^{2\gamma_{uv}y}}{y(1-y)}\,1_{\{1/N<y<1-1/N\}} \\ &+ N\,1_{\{0<y<1/N\}} + Ne^{2\gamma_{uv}}\,1_{\{1-1/N<y<1\}}\Big\}\,dy, \end{aligned}$$

and

$$\begin{aligned} \mu_{\text{fold}}(\langle u,v\rangle, y)\,dy = \frac{2\eta_u\lambda_{uv}}{\Omega'_N}\Big\{&\frac{e^{2\gamma_{uv}y}+e^{2\gamma_{uv}(1-y)}}{y(1-y)}\,1_{\{1/N<y\leq 1/2\}} \\ &+ N(1+e^{2\gamma_{uv}})\,1_{\{0<y<1/N\}}\Big\}\,dy, \end{aligned}$$

with the normalization factor Ω'_N in Proposition 1.

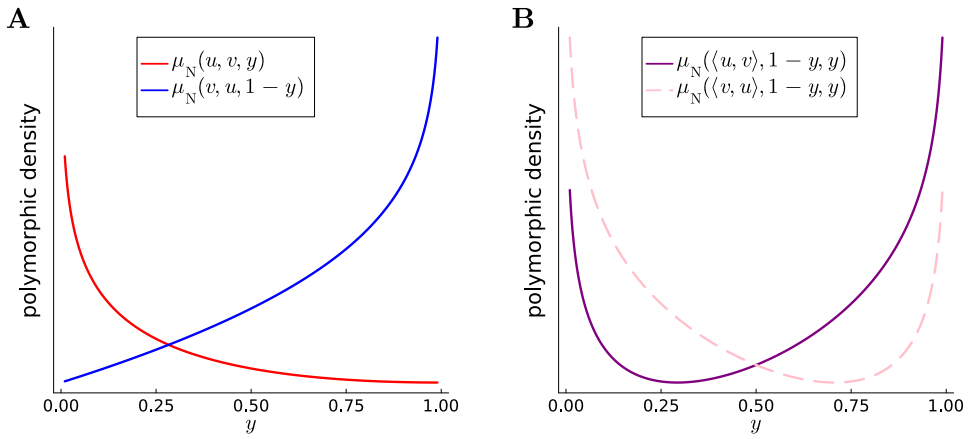

Fig. 2. Polymorphic densities of two types u and v with selection coefficient $\gamma_{uv} = 1$. Panel A: Relative log-scaled densities for type u at frequency $1-y$ on the directed edge e_{uv}, $\mu_N(u, v, y)$ in red, and on the directed edge e_{vu}, $\mu_N(v, u, 1-y)$ in blue. Panel B: The relative log-scaled density on the edges of the pair $\langle u, v\rangle$ for type u at frequency $1-y$, $\mu_N(\langle u, v\rangle, 1-y, y)$ in purple, which is the sum of the two curves in panel A. The relative log-scaled density for type v at frequency $1-y$, $\mu_N(\langle v, u\rangle, 1-y, y)$ in pink, is symmetric to the purple curve around $y = 0.5$.

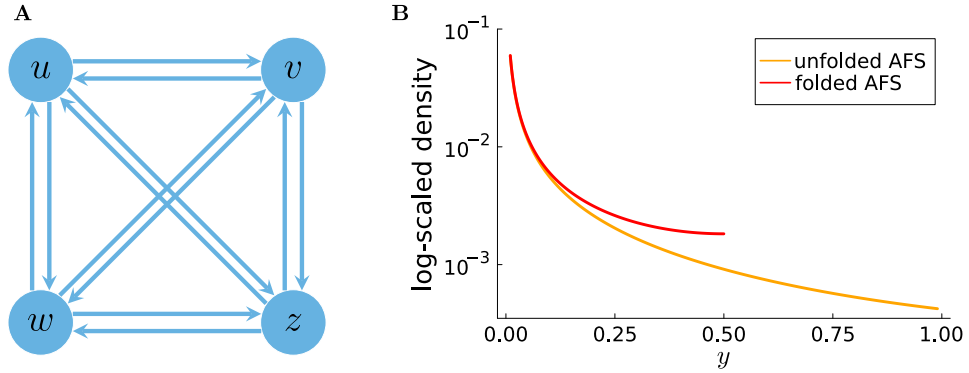

Fig. 3. Allele frequency spectra for a four type model. Panel A: State space for a model with four types. Panel B: Log-scaled unfolded (orange) and folded (red) AFS. Parameters: population size $N = 10^4$, selection coefficients defined as fitness differences with $F_u = F_z = 0$ and $F_v = F_w = 1$, and equal mutation intensity $\lambda = N \times 10^{-8}$ among all types.

Proof. Consider a fixed pair of types $\langle u, v\rangle \in \mathcal{P}$. The density of the process when v has frequency y follows from adding up the two densities on each directed edge derived in Proposition 1, $\mu_N(u, v, y) + \mu_N(v, u, 1-y)$.

Using detailed balance, Eq. (13), and the relationship $\gamma_{vu} = -\gamma_{uv}$, the contribution from the interior, $1/N < y < 1 - 1/N$, for large N is

$$\eta_u \lambda_{uv} \omega_{\gamma_{uv}} \left\{ \frac{1 - e^{-2\gamma_{uv}(1-y)}}{\gamma_{uv} y(1-y)} + \frac{1 - e^{-2\gamma_{vu} y}}{\gamma_{vu} y(1-y)} \right\} = 2\eta_u \lambda_{uv} \frac{e^{2\gamma_{uv} y}}{y(1-y)}.$$

The contributions from close to the boundaries at zero, $0 < y < 1/N$, and at one, $1 - 1/N < y < 1$, for large N, follow analogously.

As the alternative view of merging the directed edges simply entails reshuffling contributions, the normalization constant does not change. Finally, the expression for $\mu_{\text{fold}}(\langle u, v\rangle, y)\, dy$ results from similar calculations. □

An example of the unfolded AFS in a four type model with state space shown in Fig. 3A is given as the orange curve in Fig. 3B. Similarly as for the unfolded AFS, the sum $\sum_{\langle u,v\rangle \in \mathcal{P}} \mu_{\text{fold}}(\langle u, v\rangle, y), 0 < y \leq 1/2$, yields the folded, type-independent distribution of derived allele frequencies at a locus (Fig. 3B, red curve).

3.3. Extension to multiple loci

The model introduced here applies directly to a collection of L independent loci, where a locus represents either a single site or a nucleotide triplet. A collection of consecutive sites represents a DNA sequence. Thus, even though the following considerations are in general about a collection of independent loci, we may use the term sequence instead. The state of the sequence is defined by a collection of independent holding and jumping diffusion processes X^j, $1 \leq j \leq L$, with values in the direct product set $\prod_{j=1}^{L} \mathcal{G}^j$, where all graphs $\mathcal{G}^j$ have the same vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. We allow the set of selection coefficients $\{\gamma_{uv}^j : e_{uv} \in \mathcal{E}, 1 \leq j \leq L\}$ to differ from one locus to another, but assume that the transition rates $\{\lambda_{uv} : e_{uv} \in \mathcal{E}\}$ are the same among loci or along the sequence. For this it is convenient to introduce scaled intensities $\theta_{uv} = \lambda_{uv} L$, $u, v \in \mathcal{V}$. The results in Proposition 1 and Corollary 1 then apply with $\lambda_{uv} = \theta_{uv}/L$. Summing over v, the total intensity $\theta_u = \sum_{v \in \mathcal{V}} \theta_{uv}$ of a mutation from type u becomes $\theta_u = \lambda_u L$, $u \in \mathcal{V}$, and so $N\theta_u$ is the total rate in the collection of loci per time unit of a mutation affecting u. In contrast, the steady-state probabilities $\mu^j(z), z \in D$, and the boundary probabilities $\eta_u^j, u \in \mathcal{V}$ typically vary between loci, $1 \leq j \leq L$. The total mutation rate on the boundary, $\widehat{\theta}$, averaged across loci, is

$$\widehat{\theta} = \frac{1}{L} \sum_{j=1}^{L} \widehat{\theta}^j, \quad \widehat{\theta}^j = \sum_{u \in \mathcal{V}} \eta_u^j \theta_u. \tag{16}$$

The quantities $\widehat{\theta}^j$ are the jump rates of the distribution $\nu^{1/N}$ in Eq. (14). Clearly,

$$\theta_{\min} := \min_{u \in \mathcal{V}} \theta_u \leq \widehat{\theta} \leq \max_{u \in \mathcal{V}} \theta_u =: \theta_{\max}. \tag{17}$$

The closely related summation

$$\widehat{\theta}_{\text{eff}} = \sum_{u \in \mathcal{V}} \widehat{\mu}_u \theta_u, \quad \widehat{\mu}_u = \frac{1}{L} \sum_{j=1}^{L} \frac{\eta_u^j}{\Omega_N^j} \tag{18}$$

represents the *effective mutation rate* of the sequence, weighted by the average probability that loci are monomorphic. Of course, $\widehat{\theta}_{\mathrm{eff}} \le \widehat{\theta}$. We say that the mutation mechanism on the graph is *homogeneous* if the total rates in each vertex coincide, i.e. $\theta_u = \theta$, $u \in \mathcal{V}$. Under the stronger assumption of homogeneous mutation,

$$\widehat{\theta} = \theta, \qquad \widehat{\theta}_{\mathrm{eff}} = \frac{1}{L}\sum_{j=1}^{L} \frac{1}{\Omega_N^j}\,\theta. \tag{19}$$

We summarize the steady-state of the collection of loci by considering the measure-valued process $\mathcal{X}_\infty = \sum_{j=1}^{L} \delta_{X_\infty^j}$, where X_∞^j has the stationary single locus distribution μ^j on $\mathcal{G}^j$. The corresponding unfolded AFS across multiple loci is given by

$$\frac{1}{L}\sum_{j=1}^{L}\sum_{u,v\in\mathcal{V}} \mu_N^j(u,v,y), \quad 0 < y < 1.$$

For suitable functions f, $f(z) = f_{uv}(y)$, $z \in D$,

$$\langle \mathcal{X}_\infty, f\rangle = \sum_{j=1}^{L} f(X_\infty^j), \quad f \in \mathcal{F},$$

represents the sequence equilibrium distribution. The steady-state expectation under the approximate large population size distribution μ_N in Proposition 1 is

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle = \sum_{j=1}^{L}\Big\{ \sum_{u\in\mathcal{V}} f_{uu}(0)\mu_N^j(u,u,0) + \sum_{u,v\in\mathcal{V}} \int_0^1 f_{uv}(y)\,\mu_N^j(u,v,y)\,dy \Big\}. \tag{20}$$

4. Impact of directional selection on genetic variation

There exists a number of summary statistics to assess genetic variation in a population or a sample from a population. These arise as the result of evaluating functionals $\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle$ for specifically chosen functions f and can be analyzed by using Eq. (20). The first term in the sum over L in (20) provides the weight of the boundary probabilities $\mu^j(u,u,0), 1 \le j \le L$, over monomorphic loci, and the second term adds the relevant contributions from the allele frequency spectrum of the polymorphic loci. We begin with a list of the basic instances of such statistics. As a reference for each case we specialize to neutral evolution and derive the relevant neutral summary statistics. Under the assumption $\gamma_{uv} = 0$ for every $u, v \in \mathcal{V}$, Eq. (20) simplifies into

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle = \frac{L}{\Omega_N}\sum_{u\in\mathcal{V}} f_{uu}(0)\eta_u + \frac{2}{\Omega_N}\sum_{u,v\in\mathcal{V}} \eta_u\theta_{uv}\Big(f_{uv}(0+) + \int_{1/N}^1 \frac{f_{uv}(y)}{y}\,dy\Big) + \mathcal{R}_N \tag{21}$$

with

$$\Omega_N = 1 + \frac{2(1+\ln N)\widehat{\theta}^0}{L}, \quad \widehat{\theta}^0 = \sum_{u\in\mathcal{V}} \eta_u\theta_u \le \theta_{\max},$$

$\mathcal{R}_N = O(\ln N/N)$, and $f_{uv}(0+)$ is the limit of $N\int_0^{1/N} f_{uv}(y)\,dy$ as $N \to \infty$. Here, $\eta_u, u \in \mathcal{V}$, is the solution of the neutral detailed balance equation, that is, $\eta_u\theta_{uv} = \eta_v\theta_{vu}$, for every $u, v \in \mathcal{V}$, c.f. Remark 1. Under homogeneous mutation, $\widehat{\theta}^0 = \theta$.

4.1. Summary statistics under neutral evolution

The following listing is derived from (21) with the remainder term $\mathcal{R}_N$ suppressed. In order to emphasize the relative magnitude of the various terms arising from (21) we assume in addition that $\ln N/L$ is small and indicate this approximation by writing $\sim$ instead of $=$.

(i) **Average effective mutation rate**
Define f by $f_{uu}(0) = \theta_u$ and $f_{uv}(y) = 0$. The expected value $\widehat{\theta}^0_{\mathrm{eff}} = \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle/L$ is the average effective mutation rate per sequence under neutral evolution, specifically taking into effect that mutations only occur on the boundary of the graph. The first term in (21) yields

$$\widehat{\theta}^0_{\mathrm{eff}} = \frac{\widehat{\theta}^0}{1 + 2L^{-1}(1+\ln N)\widehat{\theta}^0} \sim \widehat{\theta}^0\Big(1 - \frac{2(1+\ln N)}{L}\widehat{\theta}^0\Big).$$

(ii) **Polymorphic allele functionals**
Using the effective mutation rate in (i), we observe for functions f which only depend on the frequency y and are independent of the type, i.e. with $f_{uu}(0) = 0$ and $f_{uv}(y) = f(y)$, $u, v \in \mathcal{V}$, that (21) has the form

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle = 2\widehat{\theta}^0_{\mathrm{eff}}\Big(f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big) \le 2\widehat{\theta}^0\Big(f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big).$$

(iii) **Number of monomorphic sites**
Let $f = f^\partial$ be the indicator function on the boundary ∂D. The expected number of monomorphic sites out of L is

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f^\partial\rangle = \frac{L}{\Omega_N} = \frac{L}{1 + 2L^{-1}(1+\ln N)\widehat{\theta}^0} \sim L - 2(1+\ln N)\,\widehat{\theta}^0.$$

(iv) **Number of polymorphic sites**
Let $f^\circ = 1 - f^\partial$. For L sufficiently large compared to $\ln N$ we obtain the familiar approximation of the expected number of polymorphic sites as

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f^\circ\rangle = L - \frac{L}{\Omega_N} \sim 2(1+\ln N)\widehat{\theta}^0_{\mathrm{eff}} \le 2(1+\ln N)\widehat{\theta}^0.$$

(v) **Number of segregating sites in a sample**
To obtain the number of segregating sites in a sample of size m, we take $f_{uu}(0) = 0$ and $f_{uv}(y) = g_m(y)$, where

$$g_m(y) = \sum_{k=1}^{m-1}\binom{m}{k} y^k(1-y)^{m-k} = 1 - y^m - (1-y)^m, \quad 0 \le y \le 1,$$

is the probability that a sample of size $m \ge 2$ is polymorphic when drawn from a population with derived frequency y. Then, it holds $\int_0^{1/N} g_m(y)\,dy \sim O(1/N^2)$ and

$$\int_{1/N}^1 y^{-1}g_m(y)\,dy \sim \int_0^1 y^{-1}g_m(y)\,dy = \sum_{k=1}^{m-1}\frac{1}{k}.$$

Hence

$$S^m_{N,L} = \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \sim 2\widehat{\theta}^0_{\mathrm{eff}}\sum_{k=1}^{m-1}\frac{1}{k} \le 2\widehat{\theta}^0\sum_{k=1}^{m-1}\frac{1}{k}.$$

(vi) **Pair-wise nucleotide differences**
The standard measure of genetic diversity in the population, typically denoted π, is the average number of pair-wise nucleotide differences normalized per site (Nei and Li, 1979). For sample size m we take $f_{uu}(0) = 0$ and $f_{uv}(y) = h_m(y)$, where

$$h_m(y) = \binom{m}{2}^{-1}\sum_{k=1}^{m-1} k(m-k)\binom{m}{k} y^k(1-y)^{m-k} = 2y(1-y).$$

Hence

$$\pi = \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle/L = 2\,\frac{\widehat{\theta}^0_{\mathrm{eff}}}{L} \le 2\,\frac{\widehat{\theta}^0}{L},$$

which turns out to be the average mutation load per site and is the same as the expected proportion of segregating sites in a sample of size two, $S^2_{N,L}/L$.

4.2. Allele frequency statistics and their upper bounds

We are interested in the overall effect of directional selection acting on the functionals covered above and closely related quantities, as compared to their counterparts under neutral evolution. Our main result shows that any presence of directional selection among the alleles essentially benefits monomorphic loci and constrains the number of polymorphic loci.

Theorem 2. *We consider the Wright–Fisher graph model extended to L loci with fixed, arbitrary parameters $\{\theta_{uv} : e_{uv} \in \mathcal{E}\}$ for mutation and $\{\gamma_{uv}^j : e_{uv} \in \mathcal{E}, 1 \le j \le L\}$ for selection.*

(1) With $\widehat{\theta}$, $\widehat{\theta}_{\mathrm{eff}}$, $\theta_{\min}$ and $\theta_{\max}$ defined in Eqs. (16)–(18), we have

$$\frac{\theta_{\min}}{1+2L^{-1}(1+\ln N)\theta_{\min}} \le \widehat{\theta}_{\mathrm{eff}} \le \widehat{\theta} \le \theta_{\max}.$$

(2) Let f be a function on D such that $f_{uv}(y) = f(y) \ge 0$ is a function only of the frequency y with $f(y) \to f(0+) \ge 0$, $y \to 0$, and $f(0) = 0$. Then

$$0 \le \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \le 2\widehat{\theta}_{\mathrm{eff}}\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big\} \le 2\widehat{\theta}\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big\}. \tag{22}$$

If, moreover, $\int_0^1 y^{-1} f(y)\,dy < \infty$, then

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \le 2\widehat{\theta}_{\mathrm{eff}} \int_0^1 \frac{f(y)}{y}\,dy \le 2\widehat{\theta}\int_0^1 \frac{f(y)}{y}\,dy. \tag{23}$$

While the strength and direction of selection may vary arbitrarily within and between sites, Theorem 2 illustrates that the effect of selective forces on measures of genetic variation is only channeled through to the upper bounds via the average mutation rates $\widehat{\theta}_{\mathrm{eff}}$ and $\widehat{\theta}$, respectively. It is seen furthermore that the proportionality constant $\widehat{\theta}_{\mathrm{eff}}$ is contained inside an interval that does not depend on selection parameters, namely the interval formed by the leftmost and the rightmost estimate in Theorem 2, (1). As a corollary we observe that under the stronger assumption of homogeneous mutation rates, introduced in Section 3.3, then $\widehat{\theta}$ will be independent of any selective mechanisms, and the upper bounds in Eqs. (22) and (23) will coincide with the corresponding expressions for neutral evolution in Section 4.1 (ii).

Corollary 2. *For the case when the mutation rates are homogeneous over all vertices, i.e. $\theta_u = \theta$ for all $u \in \mathcal{V}$, then*

$$\widehat{\theta} = \widehat{\theta}^0 = \theta$$

and

$$\frac{\theta}{1+2L^{-1}(1+\ln N)\theta} \le \widehat{\theta}_{\mathrm{eff}} \le \theta.$$

Hence,

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \le 2\theta\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big\}$$

and

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \le 2\theta\int_0^1 \frac{f(y)}{y}\,dy,$$

respectively.

Proof of Theorem 2. (1) For each single locus j,

$$\Omega_N^j = 1 + \frac{2}{L}\sum_{u,v\in\mathcal{V}} \eta_u^j\theta_{uv}(1+\ln N + K_{\gamma_{uv}^j}) = 1 + \frac{2\widehat{\theta}^j}{L}(1+\ln N) + \frac{2}{L}\sum_{u,v\in\mathcal{V}} \eta_u^j\theta_{uv}K_{\gamma_{uv}^j}.$$

Here, by rewriting the double sum over all vertices in $\mathcal{V}$ as the sum over all unordered pairs of vertices in $\mathcal{P}$ (see Section 3.2),

$$\sum_{u,v\in\mathcal{V}} \eta_u^j\theta_{uv}K_{\gamma_{uv}^j} = \sum_{\langle u,v\rangle\in\mathcal{P}} \{\eta_u^j\theta_{uv}K_{\gamma_{uv}^j} + \eta_v^j\theta_{vu}K_{\gamma_{vu}^j}\} = \sum_{\langle u,v\rangle\in\mathcal{P}} \eta_u^j\theta_{uv}(1-e^{2\gamma_{uv}^j})K_{\gamma_{uv}^j} \le 0,$$

for every γ_{uv}^j. Hence $1 \le \Omega_N^j \le 1 + 2\widehat{\theta}^j(1+\ln N)/L$ and therefore

$$\widehat{\theta} \ge \widehat{\theta}_{\mathrm{eff}} = \frac{1}{L}\sum_{j=1}^L \frac{\widehat{\theta}^j}{\Omega_N^j} \ge \frac{1}{L}\sum_{j=1}^L \frac{\widehat{\theta}^j}{1+2\widehat{\theta}^j(1+\ln N)/L} \ge \frac{\theta_{\min}}{1+2L^{-1}(1+\ln N)\theta_{\min}}.$$

(2) To prove the bound of $\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle$ in the second statement we take $f_{uu}(0) = f(0) = 0$ in Eq. (20) and start from the representation

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle = \sum_{j=1}^L \sum_{u,v\in\mathcal{V}} \int_0^1 f(y)\mu_N^j(u,v,y)\,dy.$$

As we apply Proposition 1 it is convenient to have the auxiliary notation $J_{uv}(y)$ (only used in this proof)

$$J_{uv}(y) = \frac{1-e^{-2\gamma_{uv}(1-y)}}{2\gamma_{uv}(1-y)}, \quad 0 < y < 1.$$

Then

$$\int_0^1 f(y)\mu_N^j(u,v,y)\,dy = \frac{2\eta_u^j\theta_{uv}}{L\Omega_N^j}\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,\omega_{\gamma_{uv}^j}J_{uv}^j(y)\,dy\Big\}.$$

We partition the right hand side as

$$\int_0^1 f(y)\mu_N^j(u,v,y)\,dy = \frac{2\eta_u^j\theta_{uv}}{L\Omega_N^j}\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big\} - R_N^j(u,v),$$

with

$$R_N^j(u,v) = \frac{2\eta_u^j\theta_{uv}}{L\Omega_N^j}\int_{1/N}^1 \frac{f(y)}{y}\Big\{1-\omega_{\gamma_{uv}^j}J_{uv}^j(y)\Big\}\,dy.$$

Letting R_N denote the sum

$$R_N = \sum_{j=1}^L \sum_{u,v\in\mathcal{V}} R_N^j(u,v),$$

these considerations imply

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle = \sum_{u,v\in\mathcal{V}} \widehat{\mu}_u 2\theta_{uv}\Big\{f(0+) + \int_{1/N}^1 \frac{f(y)}{y}\,dy\Big\} - R_N.$$

To complete the proof it remains to show that $R_N \ge 0$. Now,

$$R_N = \sum_{j=1}^L \frac{2}{L\Omega_N^j}\int_{1/N}^1 \frac{f(y)}{y}\sum_{u,v\in\mathcal{V}} \eta_u^j\theta_{uv}\Big\{1-\omega_{\gamma_{uv}^j}J_{uv}^j(y)\Big\}\,dy.$$

Thus, it suffices to show, for each site j and each frequency y,

$$\sum_{u,v\in\mathcal{V}} \eta_u^j\theta_{uv}\{1-\omega_{\gamma_{uv}^j}J_{uv}^j(y)\} \ge 0.$$

By rewriting the double summation as sum over all unordered pairs $\langle u,v\rangle \in \mathcal{P}$, the previous inequality has the equivalent representation

$$\sum_{\langle u,v\rangle\in\mathcal{P}} \Big\{\eta_u^j\theta_{uv}\{1-\omega_{\gamma_{uv}^j}J_{uv}^j(y)\} + \eta_v^j\theta_{vu}\{1-\omega_{\gamma_{vu}^j}J_{vu}^j(y)\}\Big\} \ge 0.$$

By applying the detailed balance equation to each site and each edge, the task is to show

$$\sum_{\langle u,v\rangle\in\mathcal{P}} \eta_u^j\theta_{uv}\Big\{\{1-\omega_{\gamma_{uv}^j}J_{uv}^j(y)\} + \frac{\omega_{\gamma_{uv}^j}}{\omega_{\gamma_{vu}^j}}\{1-\omega_{\gamma_{vu}^j}J_{vu}^j(y)\}\Big\} \ge 0.$$

Equivalently,

$$\sum_{\langle u,v\rangle\in\mathcal{P}} \eta_u^j\theta_{uv}\omega_{\gamma_{uv}^j}R_{uv}^j \ge 0, \quad R_{uv}^j = \omega_{\gamma_{uv}^j}^{-1} - J_{uv}^j(y) + \omega_{\gamma_{vu}^j}^{-1} - J_{vu}^j(y).$$

Next we use the anti-symmetric relation (2) for the selection coefficients. If we take a fixed site j and an edge which connects two vertices, u and v say, and let $\gamma = \gamma^j_{uv} = -\gamma^j_{vu}$ be one of the relevant selection coefficients, then it is straightforward to check that R^j_{uv} is indeed nonnegative for any signed parameter γ and $0 < y < 1$,

$$R^j_{uv} = \frac{e^{2\gamma} - e^{-2\gamma}}{2\gamma} - \frac{e^{2\gamma(1-y)} - e^{-2\gamma(1-y)}}{2\gamma(1-y)} \geq 0.$$

This verifies the claim $R_N \geq 0$ and yields

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \leq \sum_{j=1}^{L} \sum_{u,v\in\mathcal{V}} \frac{2\eta^j_u \theta_{uv}}{L\Omega^j_N} \Big\{ f(0+) + \int_{1/N}^{1} \frac{f(y)}{y}\, dy \Big\}$$
$$= 2\widehat{\theta}_{\mathrm{eff}} \Big\{ f(0+) + \int_{1/N}^{1} \frac{f(y)}{y}\, dy \Big\}.$$

We note that this proof actually provides a more general result for functions f on D that are not independent of u and v but fulfill $f_{uv}(y) = f_{vu}(y) \geq 0$. Then

$$0 \leq \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \leq 2 \sum_{u,v\in\mathcal{V}} \widehat{\mu}_u \theta_{uv} \Big\{ f_{uv}(0+) + \int_{1/N}^{1} \frac{f_{uv}(y)}{y}\, dy \Big\}.$$

The other statement in part (2) of Theorem 2 is straightforward under the additional assumption. □

Proof of Corollary 2. A simple calculation verifies that $\widehat{\theta} = \widehat{\theta}^0 = \theta$ if $\theta_u = \theta$ for every $u \in \mathcal{V}$. The rest of the corollary follows directly from Theorem 2. □

4.3. The number of segregating sites and genetic diversity

We are now in position to consider concrete measures of genetic variation in a population under the general model with selection and compare with the known properties of these measures for neutral evolution as listed in Section 4.1. Theorem 2 provides general estimates valid for arbitrary coefficients of directional selection. First, Theorem 2 applied with the functions f^∂ and f° yield bounds which directly relate to the listed items (iii) and (iv) of Section 4.1. In particular, the expected number of segregating sites under selection satisfies

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f^\circ\rangle \leq 2\widehat{\theta}_{\mathrm{eff}}\,(1 + \ln N), \quad \widehat{\theta}_{\mathrm{eff}} = \sum_{u\in\mathcal{V}} \widehat{\mu}_u \theta_u. \tag{24}$$

The parallel result for the number of segregating sites in a sample, i.e. Theorem 2 applied with the function f specified in item (v), reads

$$S^m_{N,L} = \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle \leq 2\widehat{\theta}_{\mathrm{eff}} \sum_{k=1}^{m-1} \frac{1}{k}. \tag{25}$$

Similarly, the expected genetic diversity in the population satisfies $\pi \leq 2\widehat{\theta}_{\mathrm{eff}}/L$, which extends (vi) of Section 4.1.

For specific functions f we may of course extract more detailed information in addition to the upper bounds discussed here. It is again convenient to carry out summation over unordered pairs of graph vertices. For this, let us assume that $f_{uu}(0) = 0$, $f_{uv}(y) = f(y)$, and $\int_0^1 y^{-1} f(y)\, dy < \infty$. Then

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle =$$
$$\sum_{j=1}^{L} \sum_{\langle u,v\rangle\in\mathcal{P}} \frac{2\eta^j_u \theta_{uv} \omega_{\gamma^j_{uv}}}{L\Omega^j_N} \int_0^1 \frac{f(y)}{y(1-y)} \Big(\frac{e^{2\gamma^j_{uv}(1-y)} - e^{-2\gamma^j_{uv}(1-y)}}{2\gamma^j_{uv}} \Big)\, dy.$$

In particular, for $f(y) = 2y(1-y)$,

$$\pi = \mathbb{E}_N\langle \mathcal{X}_\infty, f\rangle / L = \sum_{\langle u,v\rangle\in\mathcal{P}} \frac{2\theta_{uv}}{L} \frac{1}{L} \sum_{j=1}^{L} \frac{2}{\Omega^j_N} \frac{\eta^j_u}{\omega_{\gamma^j_{vu}}}.$$

The functional f° to obtain the expected number of segregating sites does not fulfill the condition $\int_0^1 y^{-1} f^\circ(y)\, dy < \infty$. Nevertheless we obtain an explicit representation of the expected number of segregating sites. Since the probability that a single site j is polymorphic is $(\Omega^j_N - 1)/\Omega^j_N$, the expected number of segregating sites in a sequence of length L is

$$\mathbb{E}_N\langle \mathcal{X}_\infty, f^\circ\rangle = \sum_{j=1}^{L} \frac{\Omega^j_N - 1}{\Omega^j_N}$$
$$= \sum_{j=1}^{L} \frac{2\sum_{\langle u,v\rangle\in\mathcal{P}} \eta^j_u \theta_{uv} \{1 + \ln N + K_{\gamma^j_{uv}} + e^{2\gamma^j_{uv}}(1 + \ln N - K_{\gamma^j_{uv}})\}}{L + 2\sum_{\langle u,v\rangle\in\mathcal{P}} \eta^j_u \theta_{uv} \{1 + \ln N + K_{\gamma^j_{uv}} + e^{2\gamma^j_{uv}}(1 + \ln N - K_{\gamma^j_{uv}})\}}.$$

5. Discussion

We have set up a multi-allele, multi-locus Wright–Fisher graph model to derive rigorous upper bounds for a wide class of summary statistics of genetic variation in Theorem 2. For any representative measure in this class the upper bound is a multiple of the average effective mutation rate $\widehat{\theta}_{\mathrm{eff}}$. The multiplicative factor is independent of directional selection and purely depends on the measure of genetic diversity. Hence, mutation and directional selection only affect the upper bound through $\widehat{\theta}_{\mathrm{eff}}$ or $\widehat{\theta}$. To obtain selection-independent upper bounds for arbitrary mutation rates, $\widehat{\theta}$ can be replaced with e.g. $\theta_{\max}$. The additional observation in Corollary 2 that homogeneous mutation rates make $\widehat{\theta}$ independent of directional selection shows that the upper bounds are the same as those for neutral evolution, and hence verifies the general presumption that directional selection reduces genetic variation.

There exists a number of deterministic models to verify the reduction of genetic variation due to directional selection (Feldman, 1971; Novak and Barton, 2017; Pontz and Feldman, 2020), also referred to as "constant frequency-independent selection". These models are based on replicator equations, that were initially used by Feldman (1971) in this field. Within this deterministic modeling approach analytical results on the interactions between loci due to physical linkage and/or epistasis can be derived, whereas mutation and genetic drift as additional evolutionary forces are often not taken into account. The effect of interactions among loci on genetic variation highly depends on the combination of several evolutionary processes such as the magnitude of recombination relative to the mutational input, the strength of selection, and the relative order of epistasis in comparison to the latter processes (McVean and Charlesworth, 2000; Barton, 2016; Novak and Barton, 2017). Within our current setting the effects of these phenomena are not addressed. Instead, we take a complementary approach and incorporate mutation and genetic drift, which is in particular relevant for the discussion of weak selection forces (Wright, 1931; Kimura, 1983; Ohta, 1992). In the following section we will illustrate the relevance of this setting for studying the interaction between mutation and fixation bias.

5.1. Interaction between mutation and fixation bias

We say that there is a mutation bias between two allelic types $u, v \in \mathcal{V}$, if mutation from one type to the other occurs more often than in the reverse direction, i.e. if $\theta_{uv} \neq \theta_{vu}$, and there is fixation bias between u and v whenever $\gamma_{uv} \neq 0$. In addition to selection, fixation bias can be caused by biased gene conversion. Relevant combinations of mutation bias and fixation bias may act in opposite directions and hence counterbalance their influence on the stationary distribution or act in the same direction and thus reinforce each other.

For a scenario with homogeneous mutation rates and prescribed selection parameters, Corollary 2 shows that genetic variation overall behaves much in the same way as for the case with no fixation bias. Homogeneous mutation, however, is arguably not necessarily realistic for genetic data, except perhaps on graphs of constant degree, i.e., graphs with the same number of edges attached in each vertex. Hence, whenever the total mutation rates among types differ, it is worth studying the combined impact of mutation bias and fixation bias on the upper bounds in Theorem 2. Mutation rates among the four nucleotides

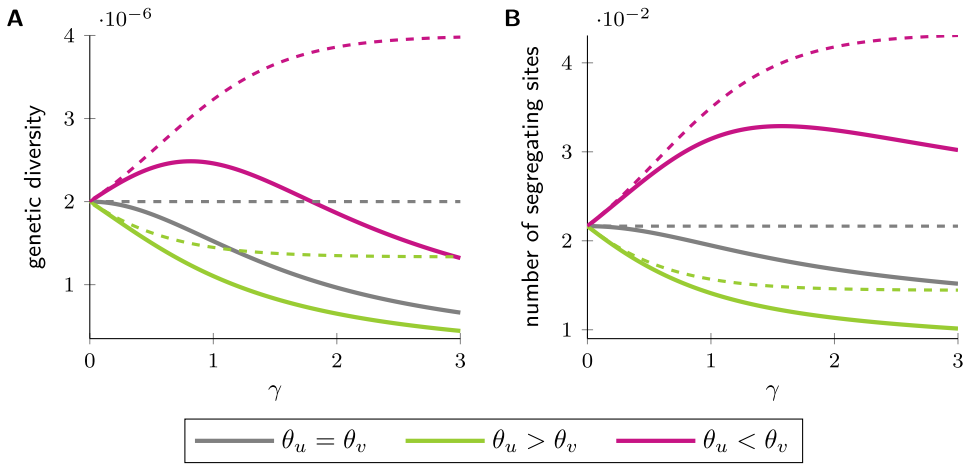

Fig. 4. Genetic diversity π (solid lines) and its upper bound $2\widehat{\theta}_{\text{eff}}/L$ (dashed lines) in panel A and the expected number of segregating sites and its upper bound $2\widehat{\theta}_{\text{eff}}(1+\ln N)$ (dashed lines) in panel B for different combinations of mutation parameters in a two-type model with equal selection coefficient among loci. Gray curves: $\theta_u = \theta_v = 1.5 \cdot 10^{-3}$, green curves: $\theta_u = 3 \cdot 10^{-3} > \theta_v = 1 \cdot 10^{-3}$, pink curves: $\theta_u = 1 \cdot 10^{-3} < \theta_v = 3 \cdot 10^{-3}$. Other parameters: population size $N = 500$ and number of loci $L = 1500$.

for example are frequently found to be different (Stoltzfus and Norris, 2015; Long et al., 2018). The four nucleotide model (Fig. 3A, nodes representing nucleotides) can be reduced to a model with two types by grouping the nucleotides into two classes: weak (A and T) and strong (C and G) bases. This classification is commonly used to describe gBGC (Duret and Galtier, 2009; Mugal et al., 2015). The fixation bias towards GC over AT nucleotides in the presence of gBGC interacts with the mutation bias between the two classes, which acts in the opposite direction in several taxa (Long et al., 2018). This illustrates that the two-type model can be relevant to describe multiple alleles that can be classified into two types.

Interaction of mutation and fixation bias in a two-type model. We consider the graph with two types, u and v, in which the dynamics at a fixed locus j are determined by three parameters θ_u, θ_v and $\gamma^j = \gamma^j_{uv}$, $1 \le j \le L$. To illustrate the results obtained in Theorem 2, the upper bounds are controlled by

$$\widehat{\theta}_{\text{eff}} = \frac{1}{L}\sum_{j=1}^{L} \frac{1}{\Omega_N^j} \frac{\theta_u\theta_v(1+e^{2\gamma^j})}{\theta_v+\theta_u e^{2\gamma^j}} \le \widehat{\theta} = \frac{1}{L}\sum_{j=1}^{L} \frac{\theta_u\theta_v(1+e^{2\gamma^j})}{\theta_v+\theta_u e^{2\gamma^j}},$$

which we can compare with the harmonic mean of the mutation rates appearing under neutral evolution, namely

$$\widehat{\theta}^0_{\text{eff}} = \frac{1}{\Omega_N}\frac{2\theta_u\theta_v}{\theta_v+\theta_u} \le \widehat{\theta}^0 = \frac{2\theta_u\theta_v}{\theta_v+\theta_u}.$$

Considering the ratio $\widehat{\theta}/\widehat{\theta}^0$ for $\gamma^j = \gamma$, $1 \le j \le L$, we obtain the relations

$$\frac{\widehat{\theta}}{\widehat{\theta}^0} = \frac{(\theta_u+\theta_v)(1+e^{-2\gamma})}{2(\theta_u+\theta_v e^{-2\gamma})} \begin{cases} > 1 & \text{if } (\theta_u > \theta_v \wedge \gamma < 0) \vee (\theta_u < \theta_v \wedge \gamma > 0), \\ = 1 & \text{if } \gamma = 0 \vee \theta_u = \theta_v, \\ < 1 & \text{if } (\theta_u < \theta_v \wedge \gamma < 0) \vee (\theta_u > \theta_v \wedge \gamma > 0). \end{cases}$$

This implies that $\widehat{\theta} > \widehat{\theta}^0$ if mutation bias is opposing fixation bias and $\widehat{\theta} < \widehat{\theta}^0$ if mutation and fixation biases enhance each other.

Such insights can be used together with the results of Theorem 2, for instance, considering the case of genetic diversity (Fig. 4),

$$\pi = \frac{1}{L\Omega_N}\frac{4\theta_u\theta_v}{\theta_v+\theta_u e^{2\gamma}}\frac{e^{2\gamma}-1}{2\gamma} \le 2\frac{\widehat{\theta}_{\text{eff}}}{L} = \frac{1}{L\Omega_N}\frac{2\theta_u\theta_v(1+e^{2\gamma})}{\theta_v+\theta_u e^{2\gamma}}.$$

Without mutation bias or with a mutation bias that enhances the fixation bias, genetic diversity decreases monotonically as selection becomes stronger (gray and green solid curves in Fig. 4A). If mutation bias counteracts fixation bias, genetic diversity first increases in the weak selection regime compared to neutral evolution until a maximum is reached for an intermediate selection coefficient, and decreases thereafter for stronger selection (pink solid curve in Fig. 4A). A similar behavior is observed and discussed in McVean and Charlesworth (1999). The upper bound (dashed lines in Fig. 4A) is constant for equal mutation rates, decreases monotonically if mutation and fixation bias reinforce each other, and increases monotonically for counterbalancing biases. The behavior of the expected number of segregating sites and its upper bounds under the different combinations of mutation rates is very akin to the curves for genetic diversity (Fig. 4B).

The scenario depicted here where all loci have equal selective pressure that can become arbitrarily large is rather artificial. In many taxa the genome-wide average of fixation bias in gBGC takes a value in the weak selection regime (De Maio et al., 2013; Glémin et al., 2015; Galtier et al., 2018; Boman et al., 2021). Likewise, according to the nearly neutral theory (Ohta, 1976, 1992) polymorphisms segregate in a population if selection is neutral or nearly neutral. In this selection regime the upper bounds capture the behavior of the measure of genetic variation well. Only when selection becomes strong, the upper bounds become more generous. However, strong selection immediately removes genetic variation and consequently, the interaction of mutation and fixation bias in the strong selection regime is less relevant when considering a large collection of loci.

CRediT authorship contribution statement

Ingemar Kaj: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Carina F. Mugal:** Supervision, Investigation, Funding acquisition, Conceptualization. **Rebekka Müller-Widmann:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors thank Nicolas Lartillot and Thibault Latrille for valuable discussions about the use of mutation-selection models for protein-coding sequence evolution. CFM has received financial support from the Knut and Alice Wallenberg Foundation, Sweden (2014/0044 to Hans Ellegren) and the Swedish Research Council (2013-8271 to Hans Ellegren).

References

Barton, N.H., 2016. How does epistasis influence the response to selection? Heredity 118 (1), 96–109.

Boman, J., Mugal, C.F., Backström, N., 2021. The effects of GC-biased gene conversion on patterns of genetic diversity among and across butterfly genomes. Genome Biol. Evol. 13 (5), evab064.

Borges, R., Szöllősi, G.J., Kosiol, C., 2019. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. Genetics 212 (4), 1321–1336.

Bourgeois, Y.X.C., Warren, B.H., 2021. An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. Mol. Ecol. 30 (23), 6036–6071.

Burden, C.J., Tang, Y., 2016. An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. Theor. Popul. Biol. 112, 22–32.

Cao, M., Shi, J., Wang, J., Hong, J., Cui, B., Ning, G., 2015. Analysis of human triallelic SNPs by next-generation sequencing. Ann. Hum. Genet. 79 (4), 275–281.

De Maio, N., Schlötterer, C., Kosiol, C., 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol. Biol. Evol. 30 (10), 2249–2262.

Duret, L., Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. 10 (1), 285–311.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution. Springer.

Etheridge, A., 2011. Some mathematical models from population genetics: École d'Été de Probabilités de Saint-Flour XXXIX-2009, vol. 2012, Springer Science & Business Media.

Ewens, W.J., 2004. Mathematical Population Genetics 1: Theoretical Introduction, Springer Verlag, Berlin.

Feldman, M.W., 1971. Equilibrium studies of two locus haploid populations with recombination. Theor. Popul. Biol. 2 (3), 299–318.

Feller, W., 1954. Diffusion processes in one dimension. Trans. Amer. Math. Soc. 77, 1–31.

Ferguson, J.M., Buzbas, E.O., 2018. Inference from the stationary distribution of allele frequencies in a family of Wright–Fisher models with two levels of genetic variability. Theor. Popul. Biol. 122, 78–87.

Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., Duret, L., 2018. Codon usage bias in animals: Disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol. Biol. Evol. 35 (5), 1092–1103.

Garg, S., Balboa, R., Kuja, J., 2022. Chromosome-scale haplotype-resolved pangenomics. Trends Genet. 38 (11), 1103–1107.

Glémin, S., Arndt, P.F., Messer, P.W., Petrov, D., Galtier, N., Duret, L., 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res. 25 (8), 1215–1228.

Haasl, R.J., Payseur, B.A., 2015. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol. Ecol. 25 (1), 5–23.

Hatcher, A., 2018. Algebraic topology, 18. printing Cambridge University Press, Cambridge.

Hu, T., Chitnis, N., Monos, D., Dinh, A., 2021. Next-generation sequencing technologies: An overview. Hum. Immunol. 82 (11), 801–811.

Kaj, I., Mugal, C.F., 2016. The non-equilibrium allele frequency spectrum in a Poisson random field framework. Theor. Popul. Biol. 111, 51–64.

Karlin, S., Taylor, H.E., 1981. A second course in stochastic processes. Academic Press, New York.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. Genetics 47 (6), 713–719.

Kimura, M., 1964. Diffusion models in population genetics. J. Appl. Probab. 1 (2), 177–232.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61 (4), 893–903.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S.F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W.R., Muscarella, M.E., Luo, H., Lennon, J.T., Brun, Y.V., Lynch, M., 2018. Evolutionary determinants of genome-wide nucleotide composition. Nat. Ecol. Evol. 2 (2), 237–240.

Maruyama, T., 1977. Stochastic Problems in Population Genetics. Springer.

McVean, G.A.T., Charlesworth, B., 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74 (2), 145–158.

McVean, G.A.T., Charlesworth, B., 2000. The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155 (2), 929–944.

Moran, P.A.P., 1958. Random processes in genetics. Math. Proc. Cambridge Philos. Soc. 54 (1), 60–71.

Mugal, C.F., Weber, C.C., Ellegren, H., 2015. GC-biased gene conversion links the recombination landscape and demography to genomic base composition. BioEssays 37 (12), 1317–1326.

Mugal, C.F., Wolf, J.B., Kaj, I., 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol. Biol. Evol. 31 (1), 212–231.

Nagylaki, T., 1983. Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. 80 (20), 6278–6281.

Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76 (10), 5269–5273.

Novak, S., Barton, N.H., 2017. When does frequency-independent selection maintain genetic variation? Genetics 207 (2), 653–668.

Ohta, T., 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Popul. Biol. 10 (3), 254–275.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23 (1), 263–286.

Peng, J., Li, W.V., 2013. Diffusions with holding and jumping boundary. Sci. China Math. 56 (1), 161–176.

Phillips, C., Amigo, J., Carracedo, Á., Lareu, M., 2015. Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data. Forensic Sci. Int. Genet. 19, 100–106.

Pontz, M., Feldman, M.W., 2020. Loss of genetic variation in the two-locus multiallelic haploid model. Theor. Popul. Biol. 136, 12–21.

Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. Genetics 132 (4), 1161–1176.

Stoltzfus, A., Norris, R.W., 2015. On the causes of evolutionary transition: Transversion bias. Mol. Biol. Evol. 33 (3), 595–602.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16 (2), 97–159.

Wright, S., 1938. The distribution of gene frequencies under irreversible mutation. Proc. Natl. Acad. Sci. 24 (7), 253–259.

Wright, S., 1949. Adaptation and selection. In: Genetics, Palaeontology and Evolution. Princeton University Press, Princeton (NJ), pp. 365–389.

Zeng, K., 2010. A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. Mol. Biol. Evol. 27 (6), 1327–1337.

Zhao, H., Pfeiffer, R., Gail, M.H., 2003. Haplotype analysis in population genetics and association studies. Pharmacogenomics 4 (2), 171–178.