A comprehensive representation of selection at loci with multiple alleles that allows complex forms of genotypic fitness

Nikolas Vellnow[1], **Toni I. Gossmann**[1,*] **and David Waxman**[2,*]

[1]TU Dortmund University, Computational Systems Biology, Faculty of Biochemical and Chemical Engineering, Emil-Figge-Str. 66, 44227 Dortmund, Germany.

✉ nikolas.vellnow@tu-dortmund.de
https://orcid.org/0000-0002-7054-3268

✉ toni.gossmann@tu-dortmund.de, *corresponding author
https://orcid.org/0000-0001-6609-4116

[2]Fudan University, Centre for Computational Systems Biology, ISTBI, 220 Handan Road, Shanghai 200433, People's Republic of China.

✉ davidwaxman@fudan.edu.cn, *corresponding author
https://orcid.org/0000-0001-9093-2108

Abstract

Genetic diversity is central to the process of evolution. Both natural selection and random genetic drift are influenced by the level of genetic diversity of a population; selection acts on diversity while drift samples from it. At a given locus in a diploid population, each individual carries only two alleles, but the population as a whole can possess a much larger number of alleles, with the upper limit constrained by twice the population size. This allows for many possible types of homozygotes and heterozygotes. Moreover, there are biologically important loci, for example those related to the MHC complex, the ABO blood types, and cystic fibrosis, that exhibit a large number of alleles. Despite this, much of population genetic theory, and data analysis, are limited to considering biallelic loci. However, to the present, what is lacking is a flexible expression for the force of selection that allows an arbitrary number of alleles (and hence an arbitrary number of heterozygotes), along with a variety of forms of fitness. In this work, we remedy this absence by giving an analytical representation of the force of selection that emphasises the very different roles played by the diversity of the population, and the fitnesses of different genotypes. The result presented facilitates our understanding and applies in a variety of different situations involving multiple alleles. This includes situations where fitnesses are: additive, multiplicative, randomly fluctuating, frequency-dependent, and it allows fitnesses which involve explicit gene interactions, such as heterozygote advantage.

Keywords: heterozygosity, genetic variation, microsatellites, evolutionary dynamics, balancing selection

Significance statement

Genetic diversity is a cornerstone of evolutionary processes, influencing both natural selection and genetic drift. While many biologically significant loci, like those associated with the MHC complex, ABO blood types, and cystic fibrosis, contain a large number of alleles, population genetics often simplifies analysis to biallelic loci. This study provides a new, flexible framework to quantify the force of selection in populations with an arbitrary number of alleles. This approach broadens our understanding of complex genetic scenarios, including those involving additive, multiplicative, and frequency-dependent fitness, as well as heterozygote advantage. This advancement enables more nuanced modeling of real-world evolutionary dynamics across diverse biological contexts.

Introduction

Genetic diversity plays a central role in evolutionary theory since it is the raw material that natural selection can act upon. In population genetics, it is usually measured at the level of single nucleotides or loci, with a focus on biallelic single nucleotide polymorphisms (SNPs) (Heyne et al., 2023). However, restricting the analysis of genetic diversity to biallelic loci may lead to biased population genomic inferences, if loci are indeed multiallelic. In a given diploid individual a biallelic locus can be in one of two homozygous states, or in one heterozygous state. However, in the population as a whole, when there are n possible alleles there can be n different homozygotes and $n(n-1)/2$ different heterozygotes (Gillespie, 2004; Hartl and Clark, 1997). This genetic diversity, which is ultimately created by mutations, may be strongly elevated in mutational hotspots, i.e., genomic regions with high mutation rates (Nesta, Tafur, and Beck, 2021), that lead to multiple alleles continually segregating at the same locus. Furthermore, genetic diversity, in the form of multiple alleles, may be both selected for, and maintained by some form of balancing selection, such as heterozygous advantage or frequency-dependent selection (Hedrick, 2012; Ayala and Campbell, 1974). In general, it seems likely that in any population, some loci will have more than two alleles, and hence will have the additional genetic diversity that goes along with this, i.e., more than one possible heterozygote type.

Indeed, there are many empirical examples of multiple alleles occurring at biologically relevant loci. For example, the amylase gene, which is involved in the digestion of starch-rich foods, exhibits multiple copy number variations in dogs, pigs, geladas and humans (Axelsson et al., 2013; Ollivier et al., 2016; Paudel et al., 2013; Caldon et al., 2024; Perry et al., 2007). Other examples, with sometimes *hundreds* of alleles include the major histocompatibility complex (MHC) involved in vertebrate immunity (Radwan et al., 2020), the ABO blood types (Yamamoto, 2021), and single locus genetic diseases like cystic fibrosis (Sharma and Cutting, 2020). In general, we may expect loci involved in the interaction between parasites and hosts to have a particularly high allelic richness since the dynamic of host-parasite co-evolution is thought to lead to the maintenance of many alleles via balancing selection (Ebert and Fields, 2020; Cornetti et al., 2024). Furthermore, the self-incompatibility locus in flowering plants is known to have an extremely high allelic richness (Takayama and Isogai, 2005), a phenomenon that is thought to be caused by strong negative frequency-dependent selection (Wright, 1939; Yokoyama and Nei, 1979; Charlesworth, Vekemans, et al., 2005).

Recent methodological advances in the field of genome sequencing and genome analysis have made the consideration of multiple alleles much more practical. In order to reliably detect alleles segregating at low frequencies (e.g., < 0.05) it is necessary to sequence at very high depth (Gossmann and Waxman, 2022). Performing *whole genome sequencing* at such a high sequencing depth used to be prohibitively expensive but is now affordable by many researchers. Such deep sequencing will make the detection of rare alleles,

and therefore the study of loci with multiple alleles, possible. We also expect that many loci, previously thought to be biallelic, will be discovered to be multiallelic when subject to deep sequencing. However, population genetic theory mostly focuses on biallelic loci, e.g. when using Watterson's θ or Tajima's D (Watterson, 1975; Tajima, 1989), and many genomics tools, and the associated software, assume (as indicated by their default settings) the presence of only two alleles (Campbell et al., 2016). Therefore, focusing on population genetic studies in scenarios with more than two alleles is clearly needed.

Here, we present a framework for the force of selection that can incorporate an arbitrary number of alleles, different forms of genotypic fitness, and which clearly differentiates the different roles played by genetic diversity and genotypic fitness. This framework allows the inclusion of interesting selection regimes, e.g. frequency-dependent selection, and gives insights into how the interplay between diversity and fitness is influenced by such selection regimes.

Motivation from a simple model

A very basic form of diversity is the genetic variation which exists at a single locus, and in what follows we shall often use *diversity* and *variation* as synonymous terms. Generally, natural selection acts on (biallelic or multiallelic) variation within a population and generates changes within the population. In addition to this, random genetic drift samples amongst existing variation within a population, and produces random changes. Thus both natural selection and random genetic drift only operate when there is variation within a population. This basic observation suggests that there may be a connection between the mathematical representation of these two distinct evolutionary 'forces' - selection and drift, which both vanish in the absence of variation. In this work we shall establish this connection by co-opting an expression involving population variability, that appears in formulae for random genetic drift, to obtain a compact formula for the selective force when there are multiple alleles.

For the simplest model involving a locus with two alleles, the connection between selection and drift can be readily seen, as we now show.

Consider a randomly mating diploid population with discrete generations, where a single biallelic locus is subject to multiplicative selection. Let x denote the frequency of a focal allele at the locus in a given generation. By the beginning of the next generation we have (Gillespie, 2004) that:

(i) selection contributes the amount $sx(1-x)$ to the change in frequency, where s is a selection coefficient associated with the focal allele (we have assumed $|s| \ll 1$),

(ii) random genetic drift contributes an amount $x(1-x)/(2N)$ to the variance of the frequency, where N is the population size[1].

[1]We work within the framework of ideal Wright-Fisher model, which depends on the census population size, N, and

We note that these effects, from selection and drift, both contain an *identical function of frequency*, namely $x(1-x)$, which vanishes in the absence of variation (at $x=0$ and $x=1$). This shows a common dependency of these evolutionary forces on a basic measure of diversity of the population at the locus, namely $x(1-x)$.

We use the case above with two alleles to motivate a question that leads to a formula for selection at a locus with multiple alleles:

> *When there are more than two alleles at a locus, do the effects of natural selection and random genetic drift have a common dependence on the diversity of the population?*

To answer this question, we first establish the model that we shall use.

Model

We adopt a Wright-Fisher model for a population of diploid individuals in which there is an equal sex ratio and no sexual dimorphism. Generations are discrete, and labelled by $t = 0, 1, 2, \ldots$. At the start of a generation there are N adult individuals in the population. The life cycle is:

(i) reproduction of adults by random mating, with all adults making the same very large contribution to the number of offspring, independent of parental genotype;
(ii) death of all adults;
(iii) offspring subject to selection at a single unlinked locus with n different alleles;
(iv) non-selective thinning of the population to N individuals - who constitute the adults of the next generation.

There is no restriction on the number of alleles at the locus, so that n can have the values 2, 3, 4, $\ldots$, and we label the j'th allele as B_j, with $j = 1, 2, ..., n$.

To describe the population, we use an n component column vector, $\mathbf{X}(t)$, whose j'th element, $X_j(t)$, is the frequency of allele B_j in adults, at the start of generation t, immediately prior to reproduction.

We take the B_iB_j genotype to have a fitness proportional to $1 + A_{i,j}$, i.e.,

$$\text{fitness of } B_iB_j \text{ genotype} \quad \propto \quad 1 + A_{i,j}. \tag{1}$$

The quantity $A_{i,j}$ is a measure of the selection acting on the B_iB_j genotype[2], and could be described as the 'fitness effect' of this genotype

not the effective population size, N_e. However, incorporation of N_e into this model is possible (Zhao, Gossmann, and Waxman, 2016).

[2]The $A_{i,j}$ may depend on allele frequencies, in which case selection is *frequency dependent*, and in generation t the $A_{i,j}$ will depend on $\mathbf{X}(t)$. In what follows we shall treat the $A_{i,j}$ as constants, unless we state otherwise.

In what follows we shall use $\mathbf{A}$ to denote the $n \times n$ matrix whose (i, j) element is $A_{i,j}$. Assuming the absence of genomic imprinting or other epigenetic phenomena (Peters, 2014) we have $A_{i,j} = A_{j,i}$, i.e., the matrix $\mathbf{A}$ is *symmetric*.

Notation

For convenience we present here the remainder of the notation that we will use in the rest of this work.

Generally, the vector of allele frequencies at any time t, namely $\mathbf{X}(t)$, is a random variable with the properties of being non-negative and normalised to unity in the sense that $X_j(t) \geq 0$ and $\sum_{j=1}^{n} X_j(t) = 1$. We write realisations (or non-random examples) of $\mathbf{X}(t)$ as $\mathbf{x}$ and these are also non-negative and normalised to unity.

With a T superscript denoting the transpose of a matrix, we use the notation where:

- $\delta_{i,j}$ denotes a Kronecker delta, which takes the value 1 when the indices i and j are equal, and is zero otherwise
- $\mathbf{F}$ denotes an n component column vector with all elements 1:

$$\mathbf{F} = (1, 1, \ldots, 1)^T \tag{2}$$

- $\mathbf{I}$ denotes the $n \times n$ identity matrix
- $\mathbf{V}(\mathbf{x})$ denotes an $n \times n$ matrix which depends on the vector $\mathbf{x}$, and has elements given by

$$V_{i,j}(\mathbf{x}) = x_i \delta_{i,j} - x_i x_j \qquad i, j = 1, 2, \ldots, n. \tag{3}$$

Dynamics of the model

The frequency of *all* alleles in generation $t + 1$, i.e., after reproduction, selection and number thinning have occurred in generation t, can be written in the form of the single equation

$$\mathbf{X}(t + 1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)) + \boldsymbol{\xi}(t). \tag{4}$$

where $\mathbf{X}(t)$, $\mathbf{D}(\mathbf{X}(t))$ and $\boldsymbol{\xi}(t)$ are all n component column vectors. The quantities $\mathbf{D}(\mathbf{X}(t))$ and $\boldsymbol{\xi}(t)$ can be called 'evolutionary forces' since they both change $\mathbf{X}(t)$. They arise from natural selection and random genetic drift, respectively.

To describe Eq. (4) in greater detail, we begin with the force of genetic drift, namely the vector $\boldsymbol{\xi}(t)$, with elements $\xi_j(t)$. For any realisation of $\mathbf{X}(t)$, the expected value of $\boldsymbol{\xi}(t)$ over replicate populations vanishes, and we write this as $E\left[\boldsymbol{\xi}(t) | \mathbf{X}(t) = \mathbf{x}\right] = \mathbf{0}$.

The variance-covariance matrix of $\boldsymbol{\xi}(t)$, conditional on $\mathbf{X}(t) = \mathbf{x}$, is given approximately[3] by

$$E[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}] \simeq \frac{\mathbf{V}(\mathbf{x})}{2N} \tag{5}$$

where the matrix $\mathbf{V}(\mathbf{x})$ of Eq. (3) enters - see Appendix A for details. Thus the (i, j) element of the variance-covariance matrix of $\boldsymbol{\xi}(t)$ is given by $E[\xi_i(t)\xi_j(t)|\mathbf{X}(t) = \mathbf{x}] \simeq V_{i,j}(\mathbf{x})/(2N)$.

When there are multiple alleles, the matrix $\mathbf{V}(\mathbf{x})$ in Eq. (5) explicitly exposes the dependence of genetic drift on the diversity of the population. The only time $\mathbf{V}(\mathbf{x})$ vanishes (has all elements zero) is when a single allele is at a frequency of unity (and all other alleles are at zero frequency).

Consider now the selective force in Eq. (4), namely $\mathbf{D}(\mathbf{X}(t))$. There are different ways to write $\mathbf{D}(\mathbf{x})$ but in Appendix B we show that it can be compactly written as

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{A}\mathbf{x}} \qquad \text{general case.} \tag{6}$$

In this representation of $\mathbf{D}(\mathbf{x})$ we see the presence of the matrix $\mathbf{A}$, which is associated with genotypic fitness - see Eq. (1). We also see the matrix $\mathbf{V}(\mathbf{x})$, which naturally emerged as the feature of random genetic drift that exposed its dependence on the population's diversity. Evidently, the matrix $\mathbf{V}(\mathbf{x})$ also plays a key role in the form of the evolutionary force of selection. This connection is even clearer under the continuous time, continuous state, diffusion approximation of Eq. (4), which says that over a small time interval dt, the change in frequencies, $d\mathbf{X}$, arises from the sum of effects of selective and drift forces. With $\mathbf{D}(\mathbf{x}) \simeq \mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}$ and omitting t arguments, the diffusion approximation of Eq. (4) takes the form

$$d\mathbf{X} = \mathbf{V}(\mathbf{X})\mathbf{A}\mathbf{X}dt + \sqrt{\frac{\mathbf{V}(\mathbf{X})}{2N_e}}d\mathbf{W} \tag{7}$$

where we have introduced the effective population size, N_e, and also the column vector $\mathbf{W} \equiv \mathbf{W}(t)$ that has random elements, representing the stochasticity of genetic drift. Thus the same frequency-dependent matrix, $\mathbf{V}(\mathbf{X})$ appears in the terms arising from selection and drift in Eq. (7) and it is natural to generally associate this matrix with a description of *diversity* of the population.

With Eq. (6) or Eq. (7), we have an answer to the question in the Introduction:

> *When there are multiple alleles, selection and drift have a common dependence on the diversity of the population which, when allele frequencies are given by* $\mathbf{x}$*, is described by* $\mathbf{V}(\mathbf{x})$.

Special cases of $\mathbf{D}(\mathbf{x})$

We next explore some special cases of the selective force given in Eq. (6).

[3]The approximation neglects the effect of selection on the right hand side of Eq. (5) (Ewens, 2004).

Additive selection

Under additive selection, the B_iB_j genotype has a fitness proportional to $1 + A_{i,j} = 1 + s_i + s_j$ where s_i is a selection coefficient associated with allele B_i. With $\mathbf{s}$ an n component column vector whose i'th element is $\mathbf{s}_i$, we have

$$\mathbf{s} = (s_1, s_2, \ldots, s_n)^T \tag{8}$$

and under additive selection we can write the matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{s}\mathbf{F}^T + \mathbf{F}\mathbf{s}^T \tag{9}$$

because its elements are precisely given by $A_{i,j} = s_i + s_j$.

Given that $\mathbf{x}$ represents a column vector of possible allele frequencies, it has the property $\sum_{j=1}^{n} x_j = 1$, which can be written as $\mathbf{F}^T\mathbf{x} = 1$, and this leads to $\mathbf{V}(\mathbf{x})\mathbf{F}$ being a vector with all elements zero (using Eq. (3)). Thus

$$\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x} = \mathbf{V}(\mathbf{x})\left(\mathbf{s}\mathbf{F}^T + \mathbf{F}\mathbf{s}^T\right)\mathbf{x} = \mathbf{V}(\mathbf{x})\mathbf{s}. \tag{10}$$

Also we have

$$1 + \mathbf{x}^T\mathbf{A}\mathbf{x} = 1 + \mathbf{x}^T\left(\mathbf{s}\mathbf{F}^T + \mathbf{F}\mathbf{s}^T\right)\mathbf{x} = 1 + 2\mathbf{s}^T\mathbf{x}. \tag{11}$$

These lead to

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{s}}{1 + 2\mathbf{s}^T\mathbf{x}} \qquad \text{additive selection.} \tag{12}$$

In the special case of $n = 2$ alleles, with $s_1 = s$, $s_2 = 0$, and writing $x = x_1$, we obtain a selective force on the frequency of allele B_1 of $D_1(\mathbf{x}) = sx(1-x)/\left(1 + 2sx\right)$.

Multiplicative selection

Under multiplicative selection, the B_iB_j genotype has a fitness proportional to $1 + A_{i,j} = (1 + s_i)(1 + s_j)$ and again we shall refer to s_i as the selection coefficient associated with allele B_i. In this case, again using Eq. (8), we can write

$$\mathbf{A} = (\mathbf{F} + \mathbf{s})(\mathbf{F} + \mathbf{s})^T - \mathbf{F}\mathbf{F}^T. \tag{13}$$

We then have $\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x} = \mathbf{V}(\mathbf{x})\mathbf{s}\left(1 + \mathbf{s}^T\mathbf{x}\right)$ and $1 + \mathbf{x}^T\mathbf{A}\mathbf{x} = \left(1 + \mathbf{s}^T\mathbf{x}\right)^2$ and hence

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{s}}{1 + \mathbf{s}^T\mathbf{x}} \qquad \text{multiplicative selection.} \tag{14}$$

In the special case of $n = 2$ alleles, with $s_1 = s$, $s_2 = 0$, and writing $x = x_1$, we obtain a selective force on the frequency of allele B_1 of $D_1(\mathbf{x}) = sx(1-x)/\left(1 + sx\right)$.

Heterozygote advantage

We next consider one of the simplest forms of selection that acts genuinely at the level of genotypes, and cannot be devolved to properties of individual alleles. This is selection where all heterozygotes have a common fitness advantage of σ relative to all homozygotes. That is

$$A_{i,j} = \begin{cases} 0, & i = j \\ \\ \sigma, & i \neq j. \end{cases} \tag{15}$$

This corresponds to

$$\mathbf{A} = \sigma \left(\mathbf{F}\mathbf{F}^T - \mathbf{I} \right). \tag{16}$$

Using this result in Eq. (6) yields

$$\mathbf{D}(\mathbf{x}) = -\frac{\sigma \mathbf{V}(\mathbf{x})\mathbf{x}}{1 + \sigma \left(1 - \mathbf{x}^{\mathbf{T}}\mathbf{x}\right)} \qquad \text{heterozygote selection.} \tag{17}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain a selective force on the frequency of allele B_1 of $D_1(\mathbf{x}) = \sigma x(1-x)(1-2x)/\left[1 + 2\sigma x(1-x)\right]$.

Fluctuating selection

We consider the situation where fitnesses randomly change (Huerta-Sanchez, Durrett, and Bustamante, 2008; Gossmann, Waxman, and Eyre-Walker, 2014), from one generation to the next, due, for example, to random environmental effects on the population. Here we give one example based on Eq. (6). Assuming fitnesses are not correlated over time, it is then appropriate to directly average over the fluctuating fitnesses. The standard way of proceeding, *as far as the selective force is concerned* (we do not consider the change in the variance of allelic effects), is to first expand the selective force to second order in selection coefficients, and then average the result (Jensen and Pollak, 1969). Here, we shall interpret the analogue of selection coefficients as the elements of the matrix $\mathbf{A}$, that appears in Eq. (6), and we will expand $\mathbf{D}(\mathbf{x})$ to second order in $\mathbf{A}$. This leads to $\mathbf{D}(\mathbf{x}) \simeq \mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x} - \mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)$.

As an illustrative example, we take

$$A_{i,j} = \sqrt{\lambda}(Z_{i,j} + Z_{j,i})/2 \tag{18}$$

where the $Z_{i,j}$ are, for all i and j, independent normal random variables with mean zero and variance one, while λ is a measure of variance of fitness effects. The form of $A_{i,j}$ adopted in Eq. (18) ensures that $A_{i,j} = A_{j,i}$. It also leads to the mean value of $\mathbf{A}$ vanishing (the average effect of selection on any genotype is zero) which we write as $\bar{\mathbf{A}} = \mathbf{0}$. Furthermore with $A_{i,j}$ given by Eq. (18), to calculate the average of $\mathbf{D}(\mathbf{x})$, we require knowledge of the average of $A_{i,j}A_{k,l}$ for arbitrary i, j, k and l. This is given by

$$\overline{A_{i,j}A_{k,l}} = \lambda \left(\delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k}\right)/2 \tag{19}$$

and using this result leads to[4]

$$\overline{\mathbf{D}}(\mathbf{x}) \simeq -\overline{\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right)} = -\lambda\left(\mathbf{x}^T\mathbf{x}\right)\mathbf{V}(\mathbf{x})\mathbf{x}. \tag{20}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain an averaged selective force on allele B_1 of $\overline{D}_1(\mathbf{x}) = \lambda x(1-x)(1-2x)\left[x^2 + (1-x)^2\right]$.

Frequency dependent selection

We can model negative frequency-dependent selection by allowing the fitness of a genotype to depend on the genotype's frequency, such that the fitness is a *decreasing function* of genotype frequency. This results in the matrix of fitness effects, $\mathbf{A}$, becoming $\mathbf{A}(\mathbf{x})$, where $\mathbf{x}$ can be thought of as the vector of current allele frequencies. The B_iB_j genotype now has a fitness proportional to $1 + A_{i,j}(\mathbf{x})$ and we take

$$A_{i,j}(\mathbf{x}) = -cx_ix_j \tag{21}$$

where c is a constant[5] that lies in the range $0 \leq c < 1$.

Since we can write $\mathbf{A}(\mathbf{x}) = -\,c\mathbf{x}\mathbf{x}^T$, the form of the selective force following from Eq. (6) is

$$\mathbf{D}(\mathbf{x}) = -\frac{c\left(\mathbf{x}^T\mathbf{x}\right)}{1 - c\left(\mathbf{x}^T\mathbf{x}\right)^2}\mathbf{V}(\mathbf{x})\mathbf{x}. \tag{22}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain a selective force on the frequency of allele B_1 of $D_1(\mathbf{x}) = cx(1-x)(1-2x)\frac{\left[x^2+(1-x)^2\right]}{1-c\left[x^2+(1-x)^2\right]^2}$.

Illustrations of the dynamics

To illustrate the dynamics that naturally arises from our framework in some interesting cases, we calculated allele frequency trajectories numerically. The detailed scripts are available as a github repository (`https://github.com/NikolasVellnow/selection_multiallelic`). In particular, for a locus with $n = 3$ alleles, we calculated three pairs of frequency trajectories, for a number of different scenarios, over 5×10^3 generations. In each pair of trajectories, one was deterministic and the other stochastic, corresponding to population sizes of infinity and 10^3, respectively.

In the results we present, the finite (but large) number of generations we consider has the consequence that the deterministic trajectories closely approach, but never actually reach the end point they would

[4]The model adopted for fluctuating selection, where $A_{i,j}$ is given by Eq. (18), leads to a simple result for the averaged force of selection, in terms of one parameter, λ, along with the matrix $\mathbf{V}(\mathbf{x})$ and vector $\mathbf{x}$ - both in unaltered form. However, the model has the feature that the variance of fitness effects of homozygotes is *twice* that of heterozygotes ($\overline{A_{i,i}^2} = \lambda$, but $\overline{A_{i,j}^2} = \lambda/2$ for $i \neq j$).

[5]A more general form of negative frequency-dependent selection is given by $A_{i,j}(\mathbf{x}) = -c_{i,j}x_ix_j$ with $c_{i,j} = c_{j,i}$ and $0 \leq c_{i,j} < 1$.

ultimately achieve. In what follows, we shall make statements about 'achieving fixation' or 'achieving a stable polymorphism' that should be interpreted to hold to an accuracy of machine precision (1 part in 10^{14}).

Furthermore, to aid visualisation of the trajectories given in the figures of this section, we have included a coloured planar triangular region, that all trajectories are confined to lie in (as a consequence of their three allele frequencies summing to unity at all times).

For Figure 1, we proceeded by first calculating the dynamics with a randomly generated set of fitness effects, based on a normal distribution, as given in Eq. (18). The set of fitness values was generated just once, at the beginning of the simulation, and was then used throughout. We simulated a scenario where three populations had the starting allele frequencies $(0.50, 0.25, 0.25)$, $(0.25, 0.50, 0.25)$, and $(0.25, 0.25, 0.50)$. We observe that the deterministic end point of the three populations may depend on the initial frequencies (see Figure 1). The populations with starting frequencies $(0.50, 0.25, 0.25)$, $(0.25, 0.50, 0.25)$ both reached the end point $(0.5595, 0.4405, 0)$. This end point, under deterministic dynamics, is a stable polymorphism, where alleles B_1 and B_2 are maintained in the population while allele B_3 is lost. By contrast, the population with the starting frequencies $(0.25, 0.25, 0.50)$ reached the end point $(0, 0, 1)$, corresponding to fixation of allele B_3.

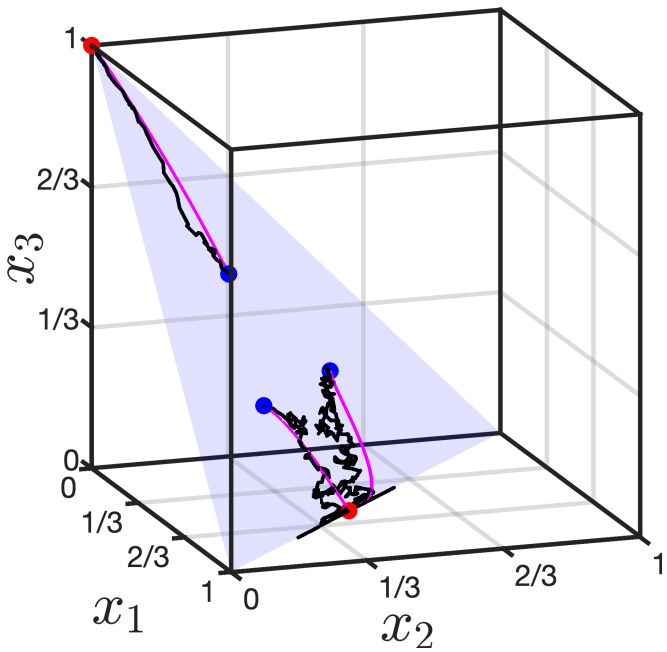


Figure 1: Allele frequency trajectories for the case of a single randomly generated set of fitness effects for $n = 3$ alleles, with frequencies x_1, x_2 and x_3. Three pairs of trajectories are plotted, each starting from a common point (blue dot). In each pair, one trajectory is deterministic (magenta) and is appropriate to an infinite population, while the other is stochastic (black), and is appropriate to a population of size $N = 10^3$. The trajectories were run for a total of 5×10^3 generations. End points of deterministic trajectories are given by red dots.

In Figure 2 we illustrate the dynamics when fitnesses are additive (Eq. 9). For this, we simulated a scenario where three populations had the starting allele frequencies $(0.70, 0.15, 0.15)$, $(0.15, 0.07, 0.15)$, and $(0.15, 0.15, 0.70)$. The set of fitness values was calculated from the allelic selection coefficients $(0.01, 0.03, 0.06)$ which were then used throughout. In this case, whether following deterministic or stochastic trajectories, allele B_3 reached fixation in all three pairs of populations, independent of the starting frequencies (Figure 2).

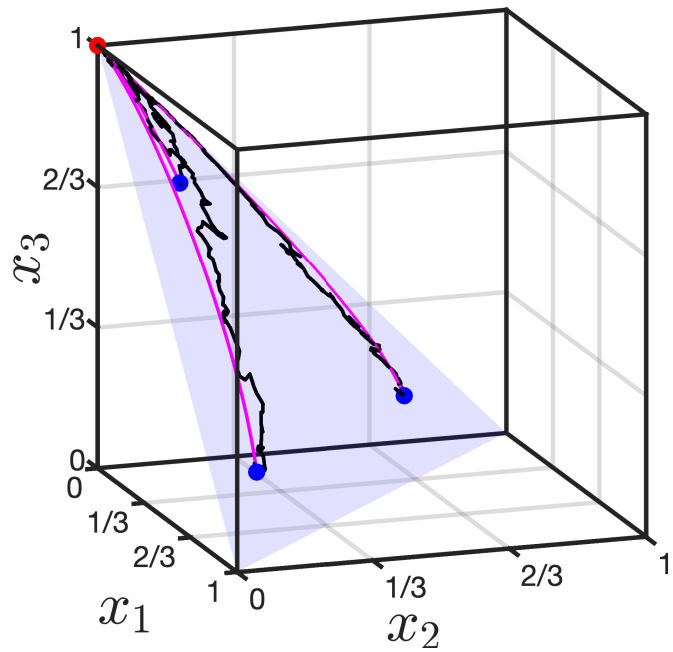


Figure 2: Allele frequency trajectories for the case of additive fitness, for $n = 3$ alleles, with frequencies x_1, x_2 and x_3. Three pairs of trajectories are plotted, each starting from a common point (blue dot). In each pair, one trajectory is deterministic (magenta) and is appropriate to an infinite population, while the other is stochastic (black), and is appropriate to a population of size $N = 10^3$. The trajectories were run for a total of 5×10^3 generations, with the common end point of all deterministic trajectories given by the red dot.

In Figure 3 we give an example of a type of balancing selection. We calculated the dynamics for frequency-dependent selection (Eq. 21). We simulated a scenario with three populations, that had the starting allele frequencies $(0.70, 0.15, 0.15)$, $(0.15, 0.70, 0.15)$, and $(0.15, 0.15, 0.70)$, where the strength of frequency-dependent selection was $c = 0.05$. The set of fitness values was updated every generation, based on the current allele frequencies (cf. Eq. (21)). Here, all three populations, following a deterministic trajectory, reached the same stable polymorphism, that had the intermediate end point frequencies $(1/3, 1/3, 1/3)$, independent of the starting frequencies (see Figure 3). All populations following a stochastic trajectory drifted around the point $(1/3, 1/3, 1/3)$ and, over the time of simulation, exhibited a polymorphism that included all alleles (Figure 3)

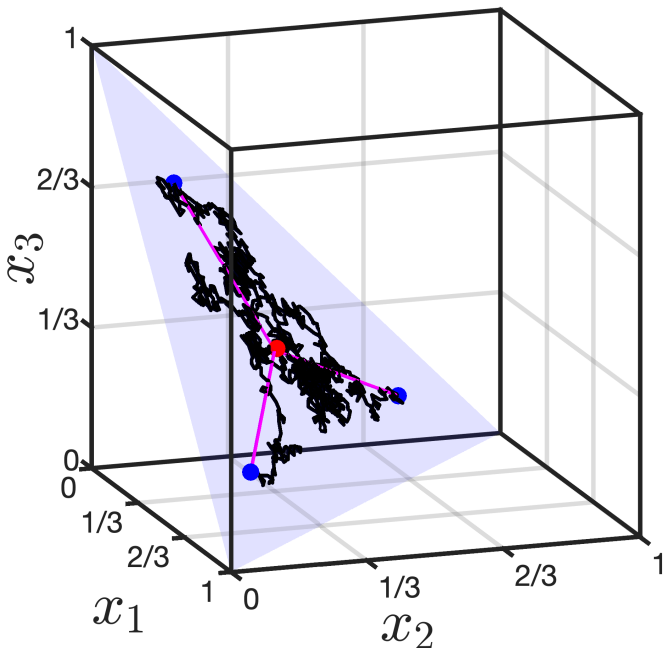


Figure 3: Allele frequency trajectories for the case of negative frequency-dependent selection for $n = 3$ alleles, with frequencies x_1, x_2 and x_3. Three pairs of trajectories are plotted, each starting from a common point (blue dot). In each pair, one trajectory is deterministic (magenta) and is appropriate to an infinite population, while the other is stochastic (black), and is appropriate to a population of size $N = 10^3$. The trajectories were run for a total of 5×10^3 generations, and the common end point of all deterministic trajectories is given by the red dot.

To calculate diverse cases of fitness, it is only necessary to change the form of the matrix $\mathbf{A}(\mathbf{x})$, and decide if it is calculated only once, at the beginning of the simulation, or whether it has to be updated after each generation (e.g., as in the case of frequency-dependent selection). This highlights the flexibility with which our framework can implement different forms of selection on multiallelic loci.

Discussion

In this work we have presented a generalised framework to model multiallelic genetic diversity that incorporates complex forms of selection. We provide detailed scripts as supplemental material to this manuscript, which allow interested researchers to tailor this powerful population genetic framework to their specific research questions. Our approach not only allows the incorporation of complex forms of selection, that act at the level of genotypes, but most importantly, it may be applied when there are an

arbitrary number of alleles. Therefore, unlike many treatments that assume biallelic loci, our approach can incorporate additional alleles that may be at (very) low frequencies. This is in particular relevant when: (i) the genetic diversity of a single locus is very high and many alleles exist (Radwan et al., 2020; Yamamoto, 2021; Sharma and Cutting, 2020) or (ii) the effective population size at a particular locus is very high (Gossmann, Woolfit, and Eyre-Walker, 2011; Smith and Haigh, 1974; Charlesworth, Morgan, and Charlesworth, 1993).

The comprehensive form for the deterministic force of selection, that we have presented in this work, can be integrated into a broader framework involving stochastic variation. It is possible that in a multiallelic sample (many) alleles segregate at low frequencies and for these alleles stochastic effects play a pronounced role. Rare alleles are more likely to be lost, due to random fluctuations in allele frequencies, than relatively common alleles. With multiple alleles, allele frequency changes due to drift can also result in complex fluctuations. Since there are more alleles, the overall variance in allele frequencies across generations can be higher than in a simple two-allele system. In loci with many alleles, genetic drift should gradually reduce the overall genetic diversity over many generations (Kimura, 1955; Latter and Novitski, 1969). However, in larger populations or populations with balancing selection (such as heterozygote advantage or frequency-dependent selection), genetic diversity may be maintained despite the random effects of drift (Kimura and Crow, 1964; Lewontin, Ginzburg, and Tuljapurkar, 1978; Spencer and Marks, 1988). Discovering alleles that may contribute to such processes is dependent on correct identification of the total genetic variation at a given genetic locus. Our approach can help in the understanding of the tight interplay between increased mutations, drift and the action of balancing selection. An example, where only certain types of allelic combinations lead to a heterosis effect, could be straightforwardly modelled within our framework.

Since our model allows the incorporation of complex forms of selection, which act at the level of the genotype, it can be used to study a wide range of questions. For this, it is only necessary to define the matrix of fitness effects, $\mathbf{A}(\mathbf{x})$, in accordance with the specific research question. For instance, we think that the interesting model used by Haasl and Payseur, in which they considered genotypic fitness to be dependent on the length of the microsatellite allele at a focal locus (Haasl and Payseur, 2013; Haasl, Johnson, and Payseur, 2014), could also be modelled as a special case of our framework. Furthermore, arbitrarily complex forms of incomplete dominance could also be modelled.

There are some direct spin-offs from the general framework that we have presented. For example, it exposes relations between the selective forces arising from different forms of fitness. Consider the selective forces in Eqs. (17), (20), and (22), that arise, respectively, from fitnesses that act on heterozygotes, or are fluctuating, or that lead to negative frequency-dependent selection. While all these forms of fitness are different, the selective forces that arise all have very similar forms. In particular all three expressions

for the force are of the form $-\kappa \mathbf{V}(\mathbf{x})\mathbf{x}$, where κ depends on $\mathbf{x}$ but has a purely numerical value (it is not a vector or matrix). This means all three selective forces push allele frequencies in precisely the same 'direction', but by different amounts, because the coefficient, κ, generally has different forms and different values in the three cases.

There are a number of practical concerns that we wish to highlight when obtaining data on genetic diversity at loci with many alleles. First, obtaining low frequency variants from population genetic data, by pooling data, requires deep sampling to obtain alleles that are segregating at low frequencies. For example, even at 500X coverage the probability of observing a variant that segregates with a frequency of 0.1 percent (e.g. a new mutant in a diploid population of 500 individuals) is less than 50% (see Figure 4a). So depending on the expected population size it might be necessary to sequence a pooled sample at an enormous coverage. However, sequencing costs have dropped substantially in recent years and therefore deep sequencing may now allow testing of more complex scenarios of selection, due to the identification of rare genetic variants at multiallelic sites.

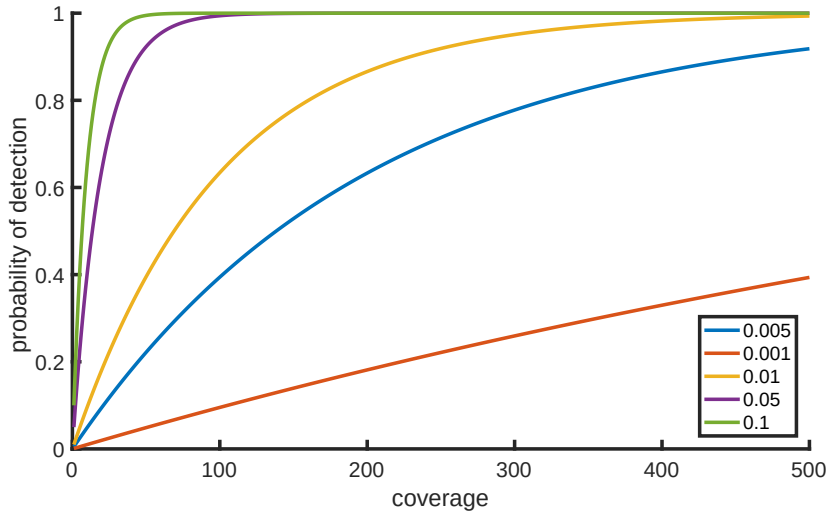
Second, incorporating low frequency variants is also crucial when considering time-series experiments since this would allow much improved resolution over time (i.e., avoiding the omission of allele frequency trajectories because they appear hidden at certain time points). This is potentially relevant for evolve and re-sequencing experiments (Schlötterer et al., 2015) or when tracing strains in metagenomic samples over time (Faust et al., 2015; Hildebrand, Moitinho-Silva, et al., 2019; Hildebrand, Gossmann, et al., 2021). However, if low-frequency variation is omitted, e.g., as a consequence of a threshold (Gossmann and Waxman, 2022), then the allele-frequencies of the remaining (above threshold) alleles might be misleading (see Figure 4b). In time-series experiments this may lead to "jumps" that can be misidentified as signatures of rapid change (e.g. selection). In general, as observation thresholds are to be expected due to sampling effects alone (see Figure 4a) it is evident that much diversity and evolutionary dynamics may be missed because it occurs below the threshold (Figure 4b).

Third, if SNPs are considered in a population-wide sample, then in principle a single base position can have four alleles, at most. However, we would expect selection to act at a level of DNA sequences that segregate together and are rarely broken up by recombination events (Williams, 1966), e.g., at the level of haplotypes (Clark, 2004; Fariello et al., 2013), or whole genes (Colbran, Ramos-Almodovar, and Mathieson, 2023), which will lead to a higher number of alleles. Furthermore, when studying epigenetic variation we can expect an even higher diversity of epialleles because there are more combinations of epigenetic modifications possible than there are unmodified nucleobases (Liu and Zhong, 2024; Kosel et al., 2024). Considering population genetic models with an arbitrary number of alleles is therefore not only realistic, but potentially very important, in order to capture all potential targets of selection (Overall and Waxman, 2024). While it is possible to obtain haplotype information from short read

sequencing data (Garrison and Marth, 2012), a better resolution of haplotypes and potential epimarks can be reached when third generation sequencing technologies are applied, such as PacBio (Rhoads and Au, 2015) or Oxford Nanopore (Wang et al., 2021). A general approach to obtain complex haplotypes might be to use barcoding approaches from long range sequencing data (Sedlazeck et al., 2018).

Fourth, low frequency variation from a population genetic sample might be more difficult to obtain when individuals are sampled, which is not uncommon for vertebrate or other multicellular species. Individual based sampling depth potentially will need to be moderate to identify heterozygous sites, but would then need to be scaled up with the number of individuals tested (see Figure 4a). However, often bioinformatic pipelines take population-wide signatures of polymorphic sites into account to filter out supposed sequencing errors (McKenna et al., 2010; Korneliussen, Albrechtsen, and Nielsen, 2014), which potentially leads to the wrong classification of rare or previously unknown variation. On the other hand, somatic mutations can be mislabeled as germline mutations (Dou et al., 2018; Lupski, 2013; Oota, 2020). Because somatic variation does not propagate to subsequent generations, time-series experiments may be used to identify such types of genetic variation, and therefore may help to accurately identify germline mutations segregating at low frequencies.

(a)



(b)

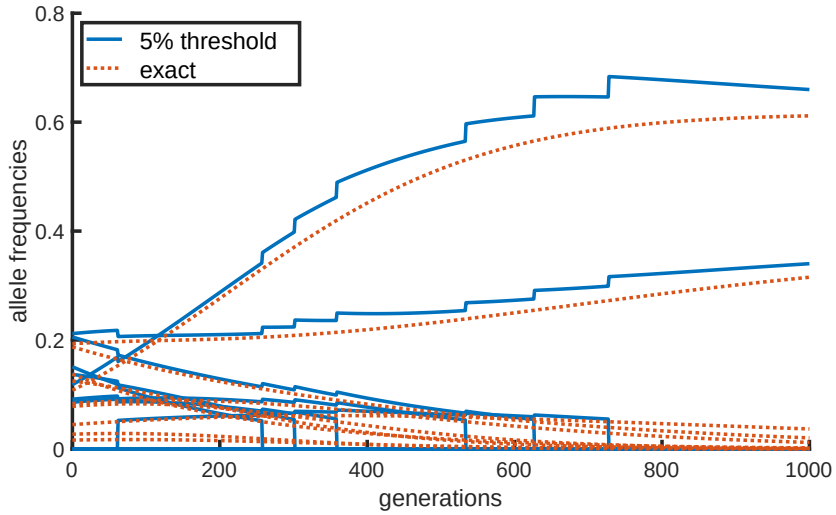


Figure 4: Illustration of observation thresholds in the analysis of multiallelic trajectories. (a) Probability to detect a genetic variant at a given frequency assuming different levels of sequencing coverage. This applies to pooled samples. (b) Shown are the exact allele frequency trajectories (red), and those after applying a 5% threshold (blue) for a locus with ten alleles and an arbitrarily chosen scheme of selection. The threshold (i) sets allele frequencies below 5% to zero and (ii) normalizes frequency so that they sum up to unity resulting in the 'steps' in the blue curves.

Data availability

All scripts are available in the github repository `https://github.com/NikolasVellnow/selection_multiallelic`

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No. 947636.

Author contributions

All three authors contributed to the conception of this work. DW carried out the mathematical analysis in consultation with NV and TIG. NV conducted the simulations, in consultation with TIG and DW. All three authors drafted and finalized the manuscript.

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.

References

Axelsson, Erik et al. (Mar. 2013). "The Genomic Signature of Dog Domestication Reveals Adaptation to a Starch-Rich Diet". In: Nature 495.7441, pp. 360–364. ISSN: 1476-4687. DOI: 10.1038/nature11837.

Ayala, Francisco J. and Cathryn A. Campbell (1974). "Frequency-Dependent Selection". In: Annual Review of Ecology and Systematics 5, pp. 115–138. ISSN: 0066-4162. JSTOR: 2096883.

Caldon, Matteo et al. (2024). "Gelada Genomes Highlight Events of Gene Flow, Hybridisation and Local Adaptation That Track Past Climatic Changes". In: Molecular Ecology n/a.n/a, e17514. ISSN: 1365-294X. DOI: 10.1111/mec.17514.

Campbell, Ian M. et al. (Mar. 2016). "Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses". In: Human mutation 37.3, pp. 231–234. ISSN: 1059-7794. DOI: 10.1002/humu.22944.

Charlesworth, B., M. T. Morgan, and D. Charlesworth (Aug. 1993). "The Effect of Deleterious Mutations on Neutral Molecular Variation". In: Genetics 134.4, pp. 1289–1303. ISSN: 0016-6731. DOI: 10.1093/genetics/134.4.1289.

Charlesworth, Deborah, Xavier Vekemans, et al. (2005). "Plant Self-Incompatibility Systems: A Molecular Evolutionary Perspective". In: New Phytologist 168.1, pp. 61–69. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.2005.01443.x.

Clark, Andrew G. (2004). "The Role of Haplotypes in Candidate Gene Studies". In: Genetic Epidemiology 27.4, pp. 321–333. ISSN: 1098-2272. DOI: 10.1002/gepi.20025.

Colbran, Laura L, Fabian C Ramos-Almodovar, and Iain Mathieson (June 2023). "A Gene-Level Test for Directional Selection on Gene Expression". In: Genetics 224.2, iyad060. ISSN: 1943-2631. DOI: 10.1093/genetics/iyad060.

Cornetti, Luca et al. (June 2024). "Long-Term Balancing Selection for Pathogen Resistance Maintains Trans-Species Polymorphisms in a Planktonic Crustacean". In: Nature Communications 15.1, p. 5333. ISSN: 2041-1723. DOI: 10.1038/s41467-024-49726-8.

Dou, Yanmei et al. (July 2018). "Detecting Somatic Mutations in Normal Cells". In: Trends in Genetics 34.7, pp. 545–557. ISSN: 0168-9525. DOI: 10.1016/j.tig.2018.04.003.

Ebert, Dieter and Peter D. Fields (Dec. 2020). "Host–Parasite Co-Evolution and Its Genomic Signature". In: Nature Reviews Genetics 21.12, pp. 754–768. ISSN: 1471-0064. DOI: 10.1038/s41576-020-0269-1.

Evans, Merran, N. A. J. Hastings, and Brian Peacock (June 1993). Statistical Distributions. Subsequent Edition. New York, NY: John Wiley & Sons Inc. ISBN: 978-0-471-55951-1.

Ewens, Warren J. (Jan. 2004). Mathematical Population Genetics 1: Theoretical Introduction. 2nd ed. 2004 Edition. New York: Springer. ISBN: 978-0-387-20191-7.

Fariello, María Inés et al. (Mar. 2013). "Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations". In: Genetics 193.3, pp. 929–941. ISSN: 1943-2631. DOI: 10.1534/genetics.112.147231.

Faust, Karoline et al. (June 2015). "Metagenomics Meets Time Series Analysis: Unraveling Microbial Community Dynamics". In: Current Opinion in Microbiology. Environmental Microbiology • Extremophiles 25, pp. 56–66. ISSN: 1369-5274. DOI: 10.1016/j.mib.2015.04.004.

Garrison, Erik and Gabor Marth (July 2012). Haplotype-Based Variant Detection from Short-Read Sequencing. DOI: 10.48550/arXiv.1207.3907. arXiv: 1207.3907.

Gillespie, John H. H. (Aug. 2004). Population Genetics: A Concise Guide. 2nd ed. Baltimore, Md.: Johns Hopkins University Press. ISBN: 978-0-8018-8009-4.

Gossmann, Toni I, Megan Woolfit, and Adam Eyre-Walker (Dec. 2011). "Quantifying the Variation in the Effective Population Size Within a Genome". In: Genetics 189.4, pp. 1389–1402. ISSN: 1943-2631. DOI: 10.1534/genetics.111.132654.

Gossmann, Toni I. and David Waxman (Apr. 2022). "Correcting Bias in Allele Frequency Estimates Due to an Observation Threshold: A Markov Chain Analysis". In: Genome Biology and Evolution 14.4, evac047. ISSN: 1759-6653. DOI: 10.1093/gbe/evac047.

Gossmann, Toni I., David Waxman, and Adam Eyre-Walker (Jan. 2014). "Fluctuating Selection Models and Mcdonald-Kreitman Type Analyses". In: PLOS ONE 9.1, e84540. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0084540.

Haasl, Ryan J., Ross C. Johnson, and Bret A. Payseur (June 2014). "The Effects of Microsatellite Selection on Linked Sequence Diversity". In: Genome Biology and Evolution 6.7, p. 1843. DOI: 10.1093/gbe/evu134.

Haasl, Ryan J. and Bret A. Payseur (Feb. 2013). "Microsatellites as Targets of Natural Selection". In: Molecular Biology and Evolution 30.2, pp. 285–298. ISSN: 1537-1719. DOI: 10.1093/molbev/mss247.

Hartl, Daniel L. and Andrew G. Clark (1997). Principles of Population Genetics. Third edition. Hartl, Daniel L.; Harvard Univ., Cambridge, MA, USA: Sinauer Associates, Inc. ISBN: 0-87893-306-9.

Hedrick, Philip W. (Dec. 2012). "What Is the Evidence for Heterozygote Advantage Selection?" In: Trends in Ecology & Evolution 27.12, pp. 698–704. ISSN: 0169-5347. DOI: 10.1016/j.tree.2012.08.012.

Heyne, H. O. et al. (Jan. 2023). "Mono- and Biallelic Variant Effects on Disease at Biobank Scale". In: Nature 613.7944, p. 519. DOI: 10.1038/s41586-022-05420-7.

Hildebrand, Falk, Toni I. Gossmann, et al. (July 2021). "Dispersal Strategies Shape Persistence and Evolution of Human Gut Bacteria". In: Cell Host & Microbe 29.7, 1167–1176.e9. ISSN: 1931-3128. DOI: 10.1016/j.chom.2021.05.008.

Hildebrand, Falk, Lucas Moitinho-Silva, et al. (Oct. 2019). "Antibiotics-Induced Monodominance of a Novel Gut Bacterial Order". In: Gut 68.10, pp. 1781–1790. ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2018-317715.

Huerta-Sanchez, Emilia, Rick Durrett, and Carlos D Bustamante (Jan. 2008). "Population Genetics of Polymorphism and Divergence Under Fluctuating Selection". In: Genetics 178.1, pp. 325–337. ISSN: 1943-2631. DOI: 10.1534/genetics.107.073361.

Jensen, Louis and Edward Pollak (1969). "Random Selective Advantages of a Gene in a Finite Population". In: Journal of Applied Probability 6.1, pp. 19–37. ISSN: 0021-9002. DOI: 10.2307/3212274. JSTOR: 3212274.

Kimura, Motoo (1955). "Random Genetic Drift in Multi-Allelic Locus". In: Evolution 9.4, pp. 419–435. ISSN: 0014-3820. DOI: 10.2307/2405476. JSTOR: 2405476.

Kimura, Motoo and James F. Crow (Apr. 1964). "The Number of Alleles That Can Be Maintained in a Finite Population". In: Genetics 49.4, pp. 725–738. ISSN: 0016-6731.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen (Nov. 2014). "ANGSD: Analysis of Next Generation Sequencing Data". In: BMC Bioinformatics 15.1, p. 356. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0356-4.

Kosel, Brinja et al. (2024). "Evolved Readers of 5-Carboxylcytosine CpG Dyads Reveal a High Versatility of the Methyl-CpG-Binding Domain for Recognition of Noncanonical Epigenetic Marks". In: Angewandte Chemie International Edition 63.17, e202318837. ISSN: 1521-3773. DOI: 10.1002/anie.202318837.

Latter, B. D. H. and C. E. Novitski (Aug. 1969). "Selection in Finite Populations with Multiple Alleles I. Limits to Directional Selection". In: Genetics 62.4, p. 859. DOI: 10.1093/genetics/62.4.859.

Lewontin, R C, L R Ginzburg, and S D Tuljapurkar (Jan. 1978). "HETEROSIS AS AN EXPLANATION FOR LARGE AMOUNTS OF GENIC POLYMORPHISM". In: Genetics 88.1, pp. 149–169. ISSN: 1943-2631. DOI: 10.1093/genetics/88.1.149.

Liu, Jie and Xuehua Zhong (Feb. 2024). "Epiallelic Variation of Non-Coding RNA Genes and Their Phenotypic Consequences". In: Nature Communications 15.1, p. 1375. ISSN: 2041-1723. DOI: 10.1038/s41467-024-45771-5.

Lupski, James R. (July 2013). "Genome Mosaicism—One Human, Multiple Genomes". In: Science 341.6144, pp. 358–359. DOI: 10.1126/science.1239503.

McKenna, Aaron et al. (Sept. 2010). "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data". In: Genome Research 20.9, pp. 1297–1303. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.107524.110.

Nagylaki, Thomas (June 1992). Introduction to Theoretical Population Genetics. 1st ed. Berlin Heidelberg: Springer. ISBN: 978-3-540-53344-3.

Nesta, Alex V., Denisse Tafur, and Christine R. Beck (Aug. 2021). "Hotspots of Human Mutation". In: Trends in Genetics 37.8, pp. 717–729. ISSN: 0168-9525. DOI: 10.1016/j.tig.2020.10.003.

Ollivier, Morgane et al. (Nov. 2016). "Amy2B Copy Number Variation Reveals Starch Diet Adaptations in Ancient European Dogs". In: Royal Society Open Science 3.11, p. 160449. DOI: 10.1098/rsos.160449.

Oota, Satoshi (Apr. 2020). "Somatic Mutations – Evolution within the Individual". In: Methods. RNA-Seq: Methods and Applications 176, pp. 91–98. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2019.11.002.

Overall, A. D. J. and D. Waxman (Aug. 2024). "Influence of Selection on the Probability of Fixation at a Locus with Multiple Alleles". In: BMC Genomics 25.1, p. 819. ISSN: 1471-2164. DOI: 10.1186/s12864-024-10733-0.

Paudel, Yogesh et al. (July 2013). "Evolutionary Dynamics of Copy Number Variation in Pig Genomes in the Context of Adaptation and Domestication". In: BMC Genomics 14.1, p. 449. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-449.

Perry, George H. et al. (Oct. 2007). "Diet and the Evolution of Human Amylase Gene Copy Number Variation". In: Nature Genetics 39.10, pp. 1256–1260. ISSN: 1546-1718. DOI: 10.1038/ng2123.

Peters, Jo (Aug. 2014). "The Role of Genomic Imprinting in Biology and Disease: An Expanding View". In: Nature Reviews Genetics 15.8, pp. 517–530. ISSN: 1471-0064. DOI: 10.1038/nrg3766.

Radwan, Jacek et al. (Apr. 2020). "Advances in the Evolutionary Understanding of MHC Polymorphism". In: Trends in Genetics 36.4, pp. 298–311. ISSN: 0168-9525. DOI: 10.1016/j.tig.2020.01.008.

Rhoads, Anthony and Kin Fai Au (Oct. 2015). "PacBio Sequencing and Its Applications". In: Genomics, Proteomics & Bioinformatics 13.5, pp. 278–289. ISSN: 1672-0229. DOI: 10.1016/j.gpb.2015.08.002.

Schlötterer, C. et al. (May 2015). "Combining Experimental Evolution with Next-Generation Sequencing: A Powerful Tool to Study Adaptation from Standing Genetic Variation". In: Heredity 114.5, pp. 431–440. ISSN: 1365-2540. DOI: 10.1038/hdy.2014.86.

Sedlazeck, Fritz J. et al. (June 2018). "Piercing the Dark Matter: Bioinformatics of Long-Range Sequencing and Mapping". In: Nature Reviews Genetics 19.6, pp. 329–346. ISSN: 1471-0064. DOI: 10.1038/s41576-018-0003-4.

Sharma, N. and G. R. Cutting (Mar. 2020). "The Genetics and Genomics of Cystic Fibrosis". In: Journal of Cystic Fibrosis. ECFS Cystic Fibrosis Research 19, S5–S9. ISSN: 1569-1993. DOI: 10.1016/j.jcf.2019.11.003.

Smith, J. M. and J. Haigh (Feb. 1974). "The Hitch-Hiking Effect of a Favourable Gene". In: Genetical Research 23.1, pp. 23–35.

Spencer, H. G. and R. W. Marks (Oct. 1988). "The Maintenance of Single-Locus Polymorphism. I. Numerical Studies of a Viability Selection Model". In: Genetics 120.2, pp. 605–613. ISSN: 0016-6731. DOI: 10.1093/genetics/120.2.605.

Tajima, F. (Nov. 1989). "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism". In: Genetics 123.3, pp. 585–595. ISSN: 0016-6731.

Takayama, Seiji and Akira Isogai (June 2005). "SELF-INCOMPATIBILITY IN PLANTS". In: Annual Review of Plant Biology 56.Volume 56, 2005, pp. 467–489. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev.arplant.56.032604.144249.

Wang, Yunhao et al. (Nov. 2021). "Nanopore Sequencing Technology, Bioinformatics and Applications". In: Nature Biotechnology 39.11, pp. 1348–1365. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01108-x.

Watterson, G. A. (Apr. 1975). "On the Number of Segregating Sites in Genetical Models without Recombination". In: Theoretical Population Biology 7.2, pp. 256–276. ISSN: 0040-5809. DOI: 10.1016/0040-5809(75)90020-9.

Williams, George Christopher (1966). Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton University Press. ISBN: 978-0-691-18550-7.

Wright, Sewall (June 1939). "The Distribution of Self-Sterility Alleles in Populations". In: Genetics 24.4, p. 538. DOI: 10.1093/genetics/24.4.538.

Yamamoto, Fumiichiro (Sept. 2021). "Molecular Genetics and Genomics of the ABO Blood Group System". In: Annals of Blood 6.0. ISSN: 2521-361X. DOI: 10.21037/aob-20-71.

Yokoyama, S. and M. Nei (Mar. 1979). "Population Dynamics of Sex-Determining Alleles in Honey Bees and Self-Incompatibility Alleles in Plants". In: Genetics 91.3, pp. 609–626. ISSN: 0016-6731. DOI: 10.1093/genetics/91.3.609.

Zhao, Lei, Toni I. Gossmann, and David Waxman (Mar. 2016). "A Modified Wright–Fisher Model That Incorporates Ne: A Variant of the Standard Model with Increased Biological Realism and Reduced Computational Complexity". In: Journal of Theoretical Biology 393, pp. 218–228. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2016.01.002.

APPENDICES

A Dynamics of the model

In this appendix we derive the dynamical equation that the frequencies of different alleles obey under a Wright-Fisher model.

We start with an effectively infinite population.

With $\mathbf{X}(t)$ an n component column vector containing the frequencies of all n alleles in generation t, the processes of reproduction and selection of the model described in Section 3 of the main text lead to the frequency obeying the equation

$$\mathbf{X}(t+1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)). \tag{23}$$

Here $\mathbf{D}(\mathbf{X}(t))$ is an n component column vector. It depends on the set of frequencies, $\mathbf{X}(t)$, of generation t and contains the changes of the frequencies that occur, due to *natural selection* in generation t. In Appendix B we give details of the form of $\mathbf{D}(\mathbf{x})$, which we interpret as the selective evolutionary force that acts when the set of allele frequencies is $\mathbf{x}$.

In a finite population, apart from the occurrence of reproduction and selection, the process of thinning of the population occurs. This amounts to *non selective sampling* that leads to N adults being present at the start of each generation. Under the appropriate generalisation of the Wright-Fisher model (Nagylaki, 1992), this has the result that the set of allele frequencies in generation $t+1$, namely $\mathbf{X}(t+1)$, does not, generally, coincide with $\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$. Rather, $\mathbf{X}(t+1)$ is *distributed around* the value $\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$ according to

$$2N\mathbf{X}(t+1) \sim \mathrm{Mult}(2N, \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)). \tag{24}$$

Here $\mathrm{Mult}(m, \mathbf{p})$ denotes a *multinomial distribution*, where the parameter m is the number of independent trials, while the parameter $\mathbf{p}$ is a column vector containing the probabilities of falling into different categories, on each trial[6].

Let $\mathbf{M}(t)$ denote a multinomial random variable with distribution $\mathrm{Mult}(2N, \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$. The form of $\mathbf{M}(t)$ is an n component column vector that has a conditional expectation of

$$E\left[\mathbf{M}(t)|\mathbf{X}(t) = \mathbf{x}\right] = 2N\left[\mathbf{x} + \mathbf{D}(\mathbf{x})\right]. \tag{25}$$

It follows that equivalent to Eq. (24) is the equation $\mathbf{X}(t+1) = \mathbf{M}(t)/(2N)$ which we write as

$$\mathbf{X}(t+1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)) + \boldsymbol{\xi}(t) \tag{26}$$

[6]We note that Eq. (24) is an approximation when fitnesses are not multiplicative, however for $|A_{i,j}| \ll 1$, as we shall assume, it is a good approximation (Nagylaki, 1992) (P. 252).

where

$$\boldsymbol{\xi}(t) = \frac{\mathbf{M}(t)}{2N} - [\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))]\,. \tag{27}$$

The quantity $\boldsymbol{\xi}(t)$ represents statistical fluctuations, due to non selective sampling, around the 'deterministic prediction' of $\mathbf{X}(t+1)$, namely $\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$, and $\boldsymbol{\xi}(t)$ can be interpreted as a stochastic 'force' on allele frequencies that arises from random genetic drift.

From properties of the multinomial distribution (Evans, Hastings, and Peacock, 1993), it follows that the conditional expectation of $\boldsymbol{\xi}(t)$ vanishes, i.e.,

$$E\left[\boldsymbol{\xi}(t)|\mathbf{X}(t) = \mathbf{x}\right] = \mathbf{0} \tag{28}$$

(the right hand side is an n component column vector of zeros). Additionally, the variance-covariance matrix of $\boldsymbol{\xi}(t)$ is

$$E\left[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}\right] = \frac{\mathbf{V}(\mathbf{x} + \mathbf{D}(\mathbf{x}))}{2N} \tag{29}$$

where $\mathbf{V}(\mathbf{x})$ is the matrix

$$\mathbf{V}(\mathbf{x}) = \mathrm{diag}(\mathbf{x}) - \boldsymbol{x}\boldsymbol{x}^T. \tag{30}$$

whose elements are given by $V_{i,j}(\mathbf{x}) = x_i\delta_{i,j} - x_i x_j$.

On the assumption that all elements of $\mathbf{D}(\mathbf{x})$ are small compared with the corresponding element of $\mathbf{x}$, which amounts to saying selection causes a very small *fractional* change in allele frequencies, as commonly holds, we have the standard approximation (Ewens, 2004)

$$E\left[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}\right] \simeq \frac{\mathbf{V}(\mathbf{x})}{2N}. \tag{31}$$

We thus arrive at a description of the dynamics given by $\mathbf{X}(t+1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)) + \boldsymbol{\xi}(t)$ with the primary properties of $\boldsymbol{\xi}(t)$ given by Eqs. (28) and (31).

B Deriving the form of the selective force

In this appendix we derive a compact representation of the evolutionary force associated with selection.

We have n alleles labelled B_i with $i = 1, 2, \ldots, n$ and assume the B_iB_j genotype has a fitness proportional to $1 + A_{i,j}$ with $A_{i,j} = A_{j,i}$.

Let x_i be the frequency of allele B_i in the adults of a particular generation, and let x_i' be the corresponding frequency in offspring, after random mating, followed by selection, has occurred.

Contributions to x_i' arise from: (i) offspring with the B_iB_i genotype and (ii) offspring with the B_iB_j genotype, with $j \neq i$. With all sums running from 1 to n unless otherwise stated, this leads to

$$x_i' = \frac{2(1+A_{i,i})x_i^2 + \sum_{j(\neq i)}(1+A_{i,j})2x_ix_j}{2\left[(1+A_{i,i})x_i^2 + \sum_{j(\neq i)}(1+A_{i,j})2x_ix_j + \sum_{j(\neq i),k(\neq i)}\left(1+A_{j,k}\right)x_jx_k\right]}. \tag{32}$$

This can be simplified to

$$x_i' = \frac{x_i \left(1 + \sum_j A_{i,j} x_j\right)}{1 + \sum_{j,k} A_{j,k} x_j x_k}. \tag{33}$$

Alternatively, we have

$$x_i' = x_i + \frac{x_i \left(\sum_j A_{i,j} x_j - \sum_{j,k} A_{j,k} x_j x_k\right)}{1 + \sum_{j,k} A_{j,k} x_j x_k} \tag{34}$$

and we rewrite this equation as

$$x_i' = x_i + \frac{\sum_{j,k} \delta_{i,j} x_j A_{j,k} x_k - \sum_{j,k} x_i x_j A_{j,k} x_k}{1 + \sum_{j,k} x_j A_{j,k} x_k}$$

$$= x_i + \frac{\sum_{j,k} \left(\delta_{i,j} x_j - x_i x_j\right) A_{j,k} x_k}{1 + \sum_{j,k} x_j A_{j,k} x_k}. \tag{35}$$

Using $[\mathbf{y}]_i$ to denote the i'th element of the vector $\mathbf{y}$, and with $\mathbf{V}(\mathbf{x})$ the $n \times n$ matrix with elements

$$V_{i,j}(\mathbf{x}) = x_i \delta_{i,j} - x_i x_j \tag{36}$$

we can write Eq. (35) as $x_i' = x_i + \frac{[\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}]_i}{1+\mathbf{x}^T\mathbf{A}\mathbf{x}}$ where $\mathbf{A}$ is a matrix with elements $A_{i,j}$. From this we obtain

$$\mathbf{x}' = \mathbf{x} + \frac{\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{A}\mathbf{x}} \tag{37}$$

and this equation allows us to identify the second term on the right hand side with the evolutionary force of selection, i.e., we identify

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{A}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{A}\mathbf{x}}. \tag{38}$$