A comprehensive representation of genotypic selection beyond two alleles, and its estimation from time-series data

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, since 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 clinically significant loci, such as 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 biallelic loci, and up to now there is no flexible expression for the force of selection which applies for arbitrary numbers of alleles (and thus any number of heterozygotes), and which accommodates diverse fitness regimes. Here, we derive an expression, for the force of selection, which explicitly separates the effects of genetic diversity from the effects of fitness of different genotypes. The result presented facilitates our understanding and analysis of selection, and applies in a variety of different situations involving multiple alleles. This includes situations where fitnesses are additive, multiplicative, randomly fluctuating, frequency-dependent, and can involve explicit gene interactions, such as heterozygote advantage. We show how our results can be used to estimate fitness effects from allele frequency trajectories and employ such an approach on existing data in the literature on experimental yeast evolution. We uncover evidence of wide-spread heterozygote advantage across the yeast genome.

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

Introduction

Genetic diversity plays a central role in evolutionary theory since it is the raw material that natural selection acts upon. In population genetics, genetic diversity is usually measured at the level of single nucleotides or single loci, with a focus on *biallelic* single nucleotide polymorphisms (SNPs) (Heyne et al., 2023). However, if loci are indeed multiallelic then restricting the analysis of genetic diversity to biallelic loci may lead to biased population genomic inferences. At a biallelic locus in a diploid population there are two possible homozygous states and one heterozygous state. By contrast, at a locus with n alleles, the population can contain up to n distinct homozygote and $n(n-1)/2$ distinct heterozygote types (Gillespie, 2004; Hartl and Clark, 1997). When $n > 2$, multiple heterozygote types emerge, and once n exceeds 3, the number of heterozygote types outnumbers that of homozygote types, demonstrating how multiallelic loci can generate substantially greater genotypic diversity than biallelic loci. Genetic diversity, arising from different alleles, is ultimately created by mutations, and is often strongly elevated in mutational hotspots, i.e., genomic regions with high mutation rates (Nesta, Tafur, and Beck, 2021), where multiple alleles may be continually segregating at loci in the region. 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; Siljestam and Rueffler, 2024; Ayala and Campbell, 1974). In general, it seems likely that in any population, some loci will have more than two alleles, and hence have the additional genetic diversity that goes along with this, i.e., more than one possible heterozygote type.

We note that 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, sometimes with *hundreds* of alleles, include the major histocompatibility complex (MHC), which is involved in vertebrate immunity (Radwan et al., 2020), the ABO blood types and their many subgroups (Yamamoto, 2021), and single locus genetic diseases, such as cystic fibrosis (Sharma and Cutting, 2020). In general, we 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, self-incompatibility loci in flowering plants are 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). Increasing the detection probability can be achieved by either using pooled sequencing with very high depth or by sequencing a larger number of individuals, which both increase the total number of reads covering a specific base. Performing *whole genome sequencing* at such a high sequencing depth used to be prohibitively expensive but is now increasingly affordable. 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).

The above indicates that there is a definite need for population genetic studies of scenarios with more than two alleles.

Here, we present a framework for the force of selection that, at *a single site or locus*, can incorporate an *arbitrary number of alleles* and *different forms of genotypic fitness*. The same framework can be applied to a closely linked set of sites or loci, and the associated set of haplotypes. The framework allows the inclusion of different selection regimes, such as frequency-dependent selection, and distinguishes the different roles played by genetic diversity and genotypic fitness. Furthermore, the framework provides the background for a numerical approach, based on allele frequency trajectories, for the estimation of fitness effects of different genotypes.

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 an 'ideal' diploid sexual population (i.e., randomly mating, having an equal sex ratio and no sexual dimorphism) with discrete generations, where a single biallelic locus is subject to weak additive or multiplicative selection. Let x denote the frequency of a focal allele at the locus in a given generation. Then by the beginning of the next generation we have 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,

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

(Gillespie, 2004).

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 above case, with two alleles, to motivate the following question, which 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.

[1]We work within the framework of ideal Wright-Fisher model, which depends on the census population size, N, and not the effective population size, N_e. However, incorporation of N_e into this model is possible (Zhao, Gossmann, and Waxman, 2016).

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 are subject to viability selection at a single 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 A_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 A_j in adults, at the start of generation t, immediately prior to reproduction.

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

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

The quantity $F_{i,j}$ is a measure of the selection acting on the A_iA_j genotype and we shall often describe $F_{i,j}$ as the 'fitness effect' of this genotype. The $F_{i,j}$ may depend on allele frequencies, in which case selection is *frequency dependent*, and in generation t the $F_{i,j}$ will depend on $\mathbf{X}(t)$. We shall assume the $F_{i,j}$ are constants unless we state otherwise.

Noting that allele frequency dynamics depend on *ratios* of fitnesses, not absolute values, this allows the use of relative fitnesses, which are defined only up to a multiplicative constant. In Appendix A we show the corresponding property of the $F_{i,j}$.

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

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
- $\mathbf{1}$ denotes an n component *column* vector with all elements 1:

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

- $\mathcal{I}$ 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

We first consider the behaviour of the population when its size is very large - effectively infinite. Let $\mathbf{X}(t)$ denote the column vector containing the n allele frequencies in the adults of generation t. The frequencies in generation $t+1$ follow from the occurrence of reproduction and selection in generation t, and are given by

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

Here $\mathbf{D}(\mathbf{X}(t))$ is an n component column vector containing the deterministic changes that occur in allele frequencies in generation t due to natural selection. We will often call $\mathbf{D}(\mathbf{X}(t))$ the 'selective force' and will shortly give its form.

If, over the timescales of interest, we need also to take into account recurrent mutation between the different alleles, then we replace $\mathbf{X}(t)$ on the right hand side of Eq. (4) by

$\mathbf{X}(t) + \mathbf{M}\mathbf{X}(t)$, where $\mathbf{M}$ is an $n \times n$ matrix containing mutation rates[2].

When the population size is *finite*, in addition to selection, we also have the effects of a non selective thinning process, where N individuals are randomly picked to constitute the N adults of the next generation. This introduces randomness, i.e., *random genetic drift*, into the dynamics, where the allele frequencies in generation $t+1$ are distributed *around* the deterministic prediction given by the right hand side of Eq. (4). Incorporation of random genetic drift entails modifying Eq. (4) to

$$\mathbf{X}(t+1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)) + \boldsymbol{\xi}(t) \tag{5}$$

where $\boldsymbol{\xi}(t)$ (also a column vector) is a random contribution to the change in allele frequencies.

To describe Eq. (5) 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 by[3]

$$E[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}] \simeq \frac{\mathbf{V}(\mathbf{x})}{2N} \tag{6}$$

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

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 C we show that it can be compactly written as

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{F}\mathbf{x}} \qquad \text{general case.} \tag{7}$$

In this representation of $\mathbf{D}(\mathbf{x})$ we see the presence of the matrix $\mathbf{F}$, which is associated with genotypic fitness - see Eq. (1). But we also see the matrix $\mathbf{V}(\mathbf{x})$, which naturally emerged as a feature of random genetic drift. Evidently, the matrix $\mathbf{V}(\mathbf{x})$ also plays a key role in the form taken by 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

[2]With $\mu_{i,j}$ the probability of mutation from allele A_j to allele A_i (with $\mu_{i,i} = 0$), the matrix $\mathbf{M}$ has the following form. For the elements $M_{i,j}$ with $i \neq j$ (off diagonal elements), we have $M_{i,j} = \mu_{i,j}$. In contrast, the diagonal elements are negative and given by $M_{i,i} = -\sum_{j=1}^{n} \mu_{j,i}$. A consequence of this form of $\mathbf{M}$ is that all of its columns sum to zero.

[3]Equation (6) is an approximation, based on the assumption that the effects of selection are small, so that on the right hand side, $\mathbf{D}(\mathbf{x})$ has been neglected in comparison with $\mathbf{x}$ (Ewens, 2004). When selection is not weak, $\mathbf{D}(\mathbf{x})$ cannot be neglected, for example, as occurs when there are lethal genotypes (Waxman and Overall, 2020).

a small time interval dt, the change in frequencies, $d\mathbf{X}$, arises from the sum of effects of selective and drift forces. With $\mathbf{D}(\mathbf{x}) \simeq \mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}$ and omitting t arguments, the diffusion approximation of Eq. (4) is motivated in Appendix A, and takes the form

$$d\mathbf{X} = \mathbf{V}(\mathbf{X})\mathbf{F}\mathbf{X}dt + \sqrt{\frac{\mathbf{V}(\mathbf{X})}{2N_e}}d\mathbf{W} \tag{8}$$

where N_e is the effective population size and the column vector $\mathbf{W} \equiv \mathbf{W}(t)$ contains random elements, representing the stochasticity of genetic drift.

Thus the same frequency-dependent matrix, $\mathbf{V}(\mathbf{X})$, appears in the term arising from selection in Eq. (8), and the corresponding term arising from drift. This matrix can be associated with the *diversity* of the population. In particular, the i'th diagonal element of $\mathbf{V}(\mathbf{X})$, namely $X_i(1 - X_i)$ is clearly recognisable as a measure of diversity, and when a single allele is at a frequency of unity (and all other alleles at zero frequency), all elements of $\mathbf{V}(\mathbf{X})$ vanish. However, the matrix $\mathbf{V}(\mathbf{X})$ is also associated with a notion of *directionality*. The matrix $\mathbf{V}(\mathbf{X})$ plays an instrumental role in allele frequency *changes* that originate in selection or drift (namely $\mathbf{V}(\mathbf{X})\mathbf{F}\mathbf{X}dt$ or $\sqrt{\mathbf{V}(\mathbf{X})/(2N_e)}d\mathbf{W}$, respectively, in Eq. (8)). These changes take place in the n dimensional space of allele frequencies, along a direction that is dependent on properties of the matrix $\mathbf{V}(\mathbf{X})$, namely its eigenvectors.

Overall, the matrix $\mathbf{V}(\mathbf{X})$ is somewhat richer than just a direct a measure of diversity, and while it contains a measure of diversity, and vanishes in the absence of diversity, it also 'directs' allele frequency changes. In multiallelic problems we will, for convenience, just refer to $\mathbf{V}(\mathbf{X})$ as the *diversity*.

With Eq. (7) or Eq. (8), 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 the matrix* $\mathbf{V}(\mathbf{x})$.

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

We next explore some special cases of the selective force given in Eq. (7).

Additive selection

Under additive selection, the A_iA_j genotype has a fitness proportional to $1+F_{i,j} = 1+s_i+s_j$ where s_i is a selection coefficient associated with allele A_i. With $\mathbf{s}$ an n component column

vector whose i'th element is $\mathbf{s}_i$, we have

$$\mathbf{s} = (s_1, s_2, \ldots, s_n)^T \tag{9}$$

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

$$\mathbf{F} = \mathbf{s}\mathbf{1}^T + \mathbf{1}\mathbf{s}^T \tag{10}$$

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

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 $\mathbf{1}^T\mathbf{x} \equiv \mathbf{x}^T\mathbf{1} = 1$, and this leads to $\mathbf{V}(\mathbf{x})\mathbf{1}$ vanishing (i.e., using Eq. (3) it becomes a vector with all elements zero). Thus $\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x} = \mathbf{V}(\mathbf{x})\left(\mathbf{s}\mathbf{1}^T + \mathbf{1}\mathbf{s}^T\right)\mathbf{x} = \mathbf{V}(\mathbf{x})\mathbf{s}$ and also $1 + \mathbf{x}^T\mathbf{F}\mathbf{x} = 1 + \mathbf{x}^T\left(\mathbf{s}\mathbf{1}^T + \mathbf{1}\mathbf{s}^T\right)\mathbf{x} = 1 + 2\mathbf{s}^T\mathbf{x}$. These lead to

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{s}}{1 + 2\mathbf{s}^T\mathbf{x}} \qquad \text{additive selection.} \tag{11}$$

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 A_1 of $D_1(\mathbf{x}) = sx(1-x)/\left(1 + 2sx\right)$.

Multiplicative selection

Under multiplicative selection, the A_iA_j genotype has a fitness proportional to $1 + F_{i,j} = (1 + s_i)(1 + s_j)$ and again we shall refer to s_i as the selection coefficient associated with allele A_i. In this case, again using Eq. (9), we can write

$$\mathbf{F} = (\mathbf{1} + \mathbf{s})(\mathbf{1} + \mathbf{s})^T - \mathbf{1}\mathbf{1}^T. \tag{12}$$

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

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{s}}{1 + \mathbf{s}^T\mathbf{x}} \qquad \text{multiplicative selection.} \tag{13}$$

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 A_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

$$F_{i,j} = \begin{cases} 0, & i = j \\ \\ \sigma, & i \neq j. \end{cases} \tag{14}$$

This corresponds to

$$\mathbf{F} = \sigma\left(\mathbf{1}\mathbf{1}^T - \boldsymbol{\mathcal{I}}\right). \tag{15}$$

Using this result in Eq. (7) yields

$$\mathbf{D}(\mathbf{x}) = -\frac{\sigma\mathbf{V}(\mathbf{x})\mathbf{x}}{1+\sigma\left(1-\mathbf{x}^{\mathbf{T}}\mathbf{x}\right)} \qquad \text{heterozygote selection.} \tag{16}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain a selective force on the frequency of allele A_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. (7).

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 elements of the matrix $\mathbf{F}$, that appears in Eq. (7), and we expand $\mathbf{D}(\mathbf{x})$ to second order in $\mathbf{F}$. This leads to $\mathbf{D}(\mathbf{x}) \simeq \mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x} - \mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}\left(\mathbf{x}^T\mathbf{F}\mathbf{x}\right)$.

As an illustrative example, we take

$$F_{i,j} = \sqrt{\lambda}(Z_{i,j} + Z_{j,i})/2 \tag{17}$$

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 $F_{i,j}$ adopted in Eq. (17) ensures that $F_{i,j} = F_{j,i}$. It also leads to the mean value of $\mathbf{F}$ vanishing (the average effect of selection on any genotype is zero) which we write as $\bar{\mathbf{F}} = \mathbf{0}$. Furthermore with $F_{i,j}$ given by Eq. (17), to calculate the average of $\mathbf{D}(\mathbf{x})$, we require knowledge of the average of $F_{i,j}F_{k,l}$ for arbitrary i, j, k and l. This is given by

$$\overline{F_{i,j}F_{k,l}} = \lambda\left(\delta_{i,k}\delta_{j,l} + \delta_{i,l}\delta_{j,k}\right)/2 \tag{18}$$

and using this result leads to[4]

$$\overline{\mathbf{D}}(\mathbf{x}) \simeq -\overline{\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}\left(\mathbf{x}^T\mathbf{F}\mathbf{x}\right)} = -\lambda\left(\mathbf{x}^T\mathbf{x}\right)\mathbf{V}(\mathbf{x})\mathbf{x}. \tag{19}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain an averaged selective force on allele A_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, $\mathbf{F}$, now becoming $\mathbf{F}(\mathbf{x})$, where $\mathbf{x}$ can be thought of as the vector of current allele frequencies. The A_iA_j genotype now has a fitness proportional to $1 + F_{i,j}(\mathbf{x})$ and we take

$$F_{i,j}(\mathbf{x}) = -cx_ix_j \tag{20}$$

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

Since we can write $\mathbf{F}(\mathbf{x}) = -\,c\mathbf{x}\mathbf{x}^T$, the form of the selective force following from Eq. (7) is

$$\mathbf{D}(\mathbf{x}) = -\frac{c\left(\mathbf{x}^T\mathbf{x}\right)}{1 - c\left(\mathbf{x}^T\mathbf{x}\right)^2}\mathbf{V}(\mathbf{x})\mathbf{x}. \tag{21}$$

In the special case of $n = 2$ alleles, and writing $x = x_1$, we obtain a selective force on the frequency of allele A_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 interesting dynamics, that naturally arises from our framework, we calculated allele frequency trajectories numerically in MATLAB (The Mathworks Inc., 2025) in the absence of mutation. The detailed scripts are available as a github repository (`https://github.com/NikolasVellnow/selection_multiallelic`). Purely for illustrative purposes, we present results for $n = 3$ alleles; the selective framework presented can apply for an arbitrary number of alleles, and indeed later we consider empirical data for a case with $n = 4$. Thus for a locus with $n = 3$ alleles, we have calculated frequency trajectories for a number of

[4]The model adopted for fluctuating selection, where $F_{i,j}$ is given by Eq. (17), leads to a simple result for the averaged force of selection in terms the parameter λ, the matrix $\mathbf{V}(\mathbf{x})$, and the vector $\mathbf{x}$. However, the model has the feature that the variance of fitness effects of homozygotes is *twice* that of heterozygotes ($\overline{F_{i,i}^2} = \lambda$, but $\overline{F_{i,j}^2} = \lambda/2$ for $i \neq j$).

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

different scenarios, over 5×10^3 generations. In each scenario, three stochastic trajectories, corresponding to population size of 10^3, and a large number of deterministic trajectories, corresponding to an effectively infinite population size, were calculated.

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 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}).

Trajectories for a single randomly generated set of fitness effects

In 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. (17). The set of fitness values, for the resulting $\mathbf{F}$, given in Eq. (22), was generated just once, at the beginning of the simulation. We then used Eq. (7) for the general case to calculate the allele frequency dynamics.

$$\boldsymbol{F} = \begin{pmatrix} 0.011\,, & 0.027\,, & -0.027 \\ 0.027\,, & 0.006\,, & -0.010 \\ -0.027\,, & -0.010\,, & 0.072 \end{pmatrix} \tag{22}$$

We simulated a scenario where three populations with finite population size 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 points of the populations depend on the initial frequencies (see Figure 1). The populations with starting frequencies in the "lower area" all reached the end point $(0.5595, 0.4405, 0)$ (see Figure 1; red trajectories).

This end point, under deterministic dynamics, is a stable polymorphism, where alleles A_1 and A_2 are maintained in the population while allele A_3 is lost. By contrast, the populations with the starting frequencies in the "upper area" reached the end point $(0, 0, 1)$, corresponding to fixation of allele A_3 (see Figure 1; blue trajectories).

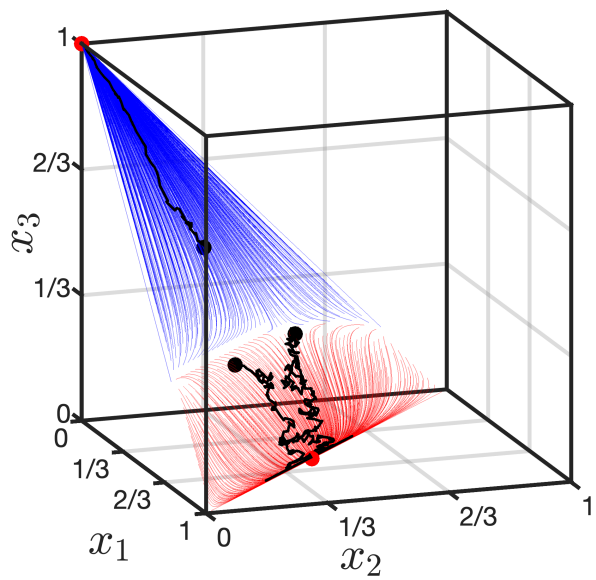


Figure 1: Plots of allele frequency trajectories for $n = 3$ alleles, for a single randomly generated set of fitness effects. In the figure, allele frequencies are denoted by x_1, x_2 and x_3. The figure contains three stochastic trajectories, which are plotted in black. These start from distinct points (allele frequency sets) indicated by the three black dots. The population size associated with these stochastic trajectories is $N = 10^3$. In addition, a large number of deterministic trajectories, associated with an effectively infinite population, are plotted. These have end points indicated by the two red dots. The deterministic trajectories arise from many different initial points and are coloured according to the final point they reach. Both stochastic and deterministic trajectories were run for a total of 5×10^3 generations.

Trajectories for a single set of additive fitness effects

In Figure 2 we illustrate the dynamics when fitnesses are additive (Eq. 10). For this, we simulated a scenario where the three finite 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, the deterministic trajectories (which all started with three alleles segregating), and the three stochastic trajectories, were all observed to reach fixation of the A_3 allele (see Figure 2).

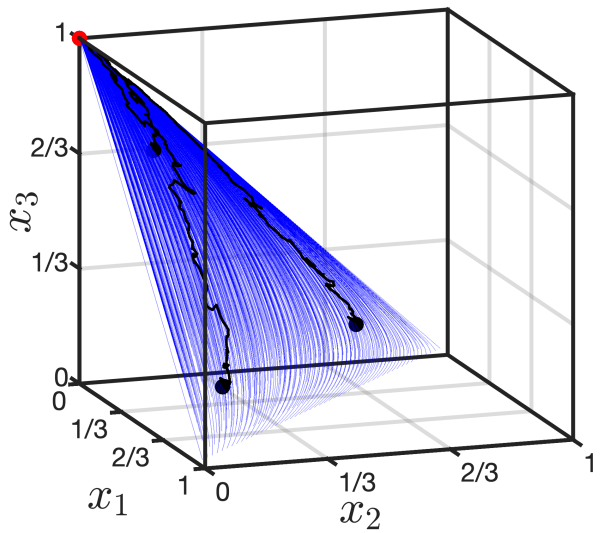


Figure 2: Plots of allele frequency trajectories for $n = 3$ alleles, for a single set of additive fitness fitness effects. In the figure, allele frequencies are denoted by x_1, x_2 and x_3. The figure contains three stochastic trajectories, which are plotted in black. These start from distinct points (allele frequency sets) indicated by the three black dots. The population size associated with these stochastic trajectories is $N = 10^3$. In addition, a large number of deterministic trajectories, associated with an effectively infinite population, are plotted. The deterministic trajectories arise from many different initial points and have a common end point, indicated by the red dot. All such trajectories are coloured blue. Both stochastic and deterministic trajectories were run for a total of 5×10^3 generations.

Trajectories for negative frequency-dependent selection

In Figure 3 we give an example of a type of balancing selection, with the dynamics calculated under frequency-dependent selection (Eq. 20). We simulated a scenario with three finite 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 following Eq. (20)). Here, all 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; blue trajectories). 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 (see Figure 3)

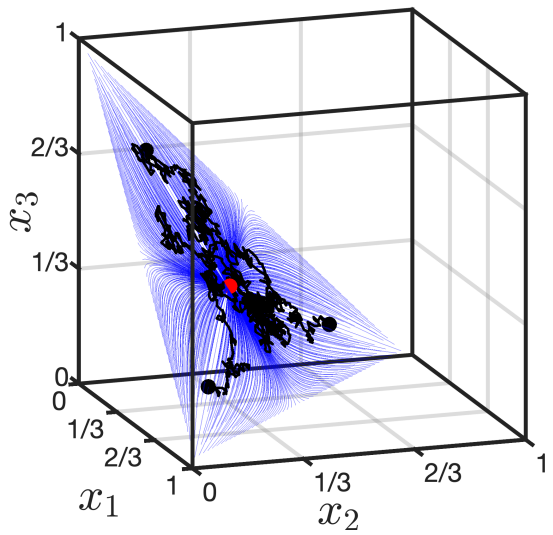


Figure 3: Plots of allele frequency trajectories for $n = 3$ alleles, for negative frequency-dependent selection. In the figure, allele frequencies are denoted by x_1, x_2 and x_3, and a single form of fitness effects, associated with negative frequency-dependent selection, was used. The figure contains three stochastic trajectories, which are plotted in black. These start from distinct points (allele frequency sets) indicated by the three black dots. The population size associated with these stochastic trajectories is $N = 10^3$. In addition, a large number of deterministic trajectories, associated with an effectively infinite population, are plotted. The deterministic trajectories arise from many different initial points and have a common end point, indicated by the red dot. All such trajectories are coloured blue. Both stochastic and deterministic trajectories were run for a total of 5×10^3 generations.

Incorporating diverse cases into the model

To generate evolutionary trajectories for other diverse cases of the selective force, it is only necessary to change the form of the matrix $\mathbf{F}(\mathbf{x})$, and decide if it is calculated only once (e.g., multiplicative selection), at the beginning of the simulation, or whether it has to be updated after each generation (e.g., frequency-dependent selection). This highlights the flexibility with which our framework can implement different forms of selection on multiallelic loci.

Applications to empirical data analysis

Our modelling framework can also be useful in understanding and analysing empirical data. In general, it can be applied to all population genomic data sets that allow the tracking of multiple allele or haplotype frequencies over several generations. For instance, evolve-and-resequence experiments are prime candidates to be analysed in this way (Schlötterer et al., 2015).

Estimating F from allele frequency trajectories

One application of our framework is to estimate the matrix of fitness effects, $\mathbf{F}$, from empirical trajectories of multiple alleles or haplotypes. To do this, we need to make certain assumptions about the properties of the $\mathbf{F}$ matrix and the evolutionary processes that created the empirical trajectories. Here, we present an algorithm to estimate the matrix of fitness effects, $\mathbf{F}$, when the following assumptions are applicable:

(i) the population is very large (effectively infinite),

(ii) the $\mathbf{F}$ matrix is constant (i.e., has no time or frequency dependence), and

(iii) over the time interval considered, mutations can be neglected.

Under these assumptions, the equation that determines allele frequency trajectories is Eq. (4).

Then with a data set consisting solely of empirical allele frequency trajectories, that cover a range of times, we proceed as follows.

First, we define a 'cost function' $C(\mathbf{F})$ which measures the mismatch between the empirical allele frequency trajectories of the data set, and the allele frequency trajectories arising from a putative matrix of fitness effects, $\mathbf{F}$. Using a suitable optimization algorithm we then find the matrix of fitness effects that minimizes the cost function, $C(\mathbf{F})$, i.e., which produces trajectories that are most similar to the empirical trajectories. Details of the estimation algorithm are given in Appendix D.1, with additional notes given in Section 1 of the Supplementary Material. Since, in real world data collections, genomic data is often not available for every generation, but instead may be given for time points with different numbers of generations between them, we also provide an algorithm that can estimate the matrix of fitness effects in such cases - see Appendix D.2.

Application to a yeast data set

As an example of estimating the matrix of fitness effects, on real world data, we analysed publicly available data from an evolve-and-resequence experiment with the outcrossing brewer's yeast *Saccharomyces cerevisiae* (Burke, Liti, and Long, 2014; Phillips et al., 2020). Briefly, the authors crossed four wild yeast strains to create a mixed outbred founder population and let replicated populations evolve for 540 generations under a laboratory domestication treatment. Sexual reproduction was enforced every 30 generations and samples of the populations were resequenced at 17 time points during the experiment. The authors provided the relative frequencies of haplotypes from the four original wild strains, calculated for 10 KB windows across 17 time points, enabling the tracking of their evolutionary trajectories (Burke, Liti, and Long, 2014; Phillips et al., 2020). Here, we take the four haplotypes as being equivalent to four alleles at a locus in a diploid organism. Furthermore, all 30 generations of mitotic reproduction, between outcrossing events, are taken as a single effective generation of a diploid organism, during which time selection occurs.

First, as an example region, we estimated $\mathbf{F}$ for the trajectories of the haplotypes in the 10 KB window around position 457,847 on chromosome C16 for replicate population 3 (see Figure 4a; cf. Fig. 6b in Phillips et al., 2020). The estimated $\mathbf{F}$ has a form which depends on which genotype is taken as the reference type (other, 'evolutionarily equivalent' forms of $\mathbf{F}$, which use different reference genotypes, can be obtained from the transformation in Appendix A). Here, we give the form of $\mathbf{F}$ when the genotype A_3A_4 is used as the reference type (in which case its fitness effect vanishes). We also give the corresponding matrix of relative fitness values, which we denote by $\mathbf{w}$. This matrix contains the fitness values of all genotypes, *relative* to that of the most fit genotype, i.e., $w_{i,j} = (1+F_{i,j})/(1+\max_{k,l}(F_{k,l}))$:

$$\mathbf{F} = \begin{pmatrix} 1.513\,, -0.887\,, -0.195\,, & 1.511 \\ -0.887\,, -1.000\,, & 1.138\,, & 1.244 \\ -0.195\,, & 1.138\,, & 0.841\,, & 0.000 \\ 1.511\,, & 1.244\,, & 0.000\,, -1.000 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} 1.000\,, 0.045\,, 0.320\,, 0.999 \\ 0.045\,, 0.000\,, 0.851\,, 0.893 \\ 0.320\,, 0.851\,, 0.733\,, 0.398 \\ 0.999\,, 0.893\,, 0.398\,, 0.000 \end{pmatrix}. \tag{23}$$

Using the above form of $\mathbf{F}$ we can create trajectories that reasonably resemble the empirical ones (see Figure 4).

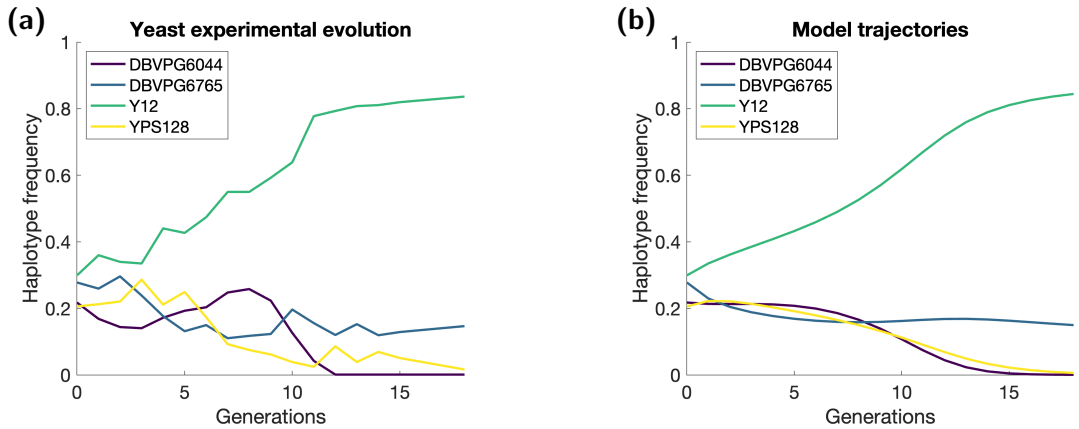


Figure 4: Comparison of haplotype frequency trajectories from a yeast experiment and from estimated parameters using our new model. Data is from a 10 KB window around position 457847 on chromosome C16 for replicate 3 (Burke, Liti, and Long, 2014; Phillips et al., 2020). Haplotypes are DBVPG6044, West African in dark blue, DBVPG6765, wine/European in purple, Y12, sake/Asian in green, and YPS128, North American in yellow. a) Empirical trajectories from yeast experimental evolution. b) Trajectories produced by Eq. (4) where the estimated $\mathbf{F}$ matrix, Eq. (23), was used.

Once $\mathbf{F}$, and hence $\mathbf{w}$, have been estimated from the data, they can be used to address questions of interest. For instance, in the example region considered, we found evidence for general heterozygote advantage as heterozygotes have, on average, a fitness 35 % higher than the homozygotes. However, pairwise heterozygote advantage appears to be highly variable as some heterozygotes (A_1A_4, A_2A_3, and A_2A_4) exhibit strong advantage while others (A_1A_2 and A_1A_3) show a disadvantage compared to the respective homozygotes.

The results we obtain, for the estimate of $\mathbf{F}$, are sensitive to the validity of the assumptions we have made. When applying our estimation method to regions of another chromosome (chromosome C12), we do not see such a good fit to the data (see Figure S1). We suspect that one or more of the assumptions we have made may not apply. However, the effective population size of $N_e = 10^6$ that was quoted for the evolve-and-resequence experiments with the outcrossing brewer's yeast (Burke, Liti, and Long, 2014) do not lead us to suspect that random genetic drift is playing a significant role.

Next, we analyzed $\mathbf{F}$ across 5725 sliding 10 KB windows (step size: 2 KB) spanning the 16 major chromosomes of the *S. cerevisiae* genome. For each of these windows, we identified the most fit genotype. Remarkably, in 99.2 % of the windows, a heterozygote emerged as the most fit genotype, with A_4A_3 being most fit in 19.1 % of cases (see Figure 5).

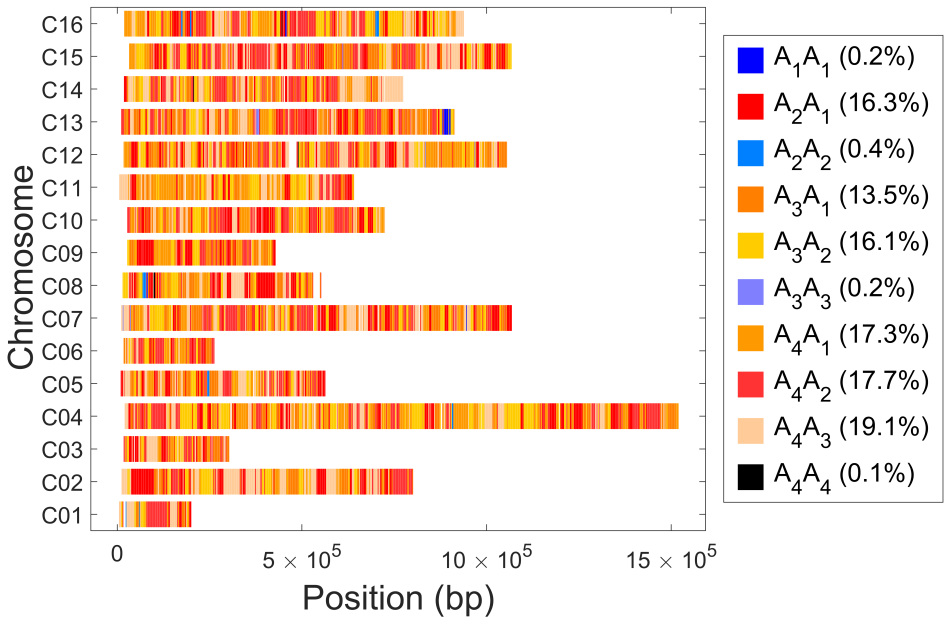


Figure 5: Illustration of the most fit genotypes across the major 16 S. cerevisiae chromosomes based on the experimental evolution experiment by Burke, Liti, and Long (2014) and Phillips et al. (2020). Each colored point indicates which genotype was the fittest for a sliding 10 KB window (plotted at its mid-point) with step size of 2 KB. Heterozygotes are colored in warm colors and homozygotes are colored in shades of blue and black. Relative frequencies with which each genotype was found to be the fittest one are given in brackets in the legend. Note that the data provided does not always start at the first possible mid-point (5 KB) for all chromosomes so that our plot starts at different positions on the left hand side.

Extracting multiallelic trajectories from pooled sequencing data

Multiallelic variation can be identified from next-generation sequencing (NGS) data using appropriate variant calling and filtering strategies. This includes the detection of small haplotypes (combinations of nearby SNPs) and structural variation, as well as estimation of allele frequencies in pooled sequencing samples. To illustrate this, we analyzed pooled sequencing data from a population of *Drosophila melanogaster* that was sampled at multiple time points. Variant calling was performed using FreeBayes (Garrison and Marth, 2012), which supports the identification of complex haplotypes and structural variants. Our analysis focused on chromosome arm 3R across three distinct sampled time points (July 2009, July 2012, and October 2015) from the "Linvilla" population, which is part of the "Drosophila in Time and Space" (DEST) project (Kapun et al., 2021). We applied a stringent filter, retaining only sites where all detected alleles were supported by at least three reads at every time point. This approach uncovered a wide range of multiallelic variation: we found 17,678 sites with three alleles, 1,602 with four, 180 with five, 14 with six, and 2 with seven alleles

(see Figure 6). Considerable numbers of multiallelic sites were also found when we used different filters (see Figure S3). We provide the code for this extraction pipeline in our github repository `https://github.com/NikolasVellnow/selection_multiallelic`.

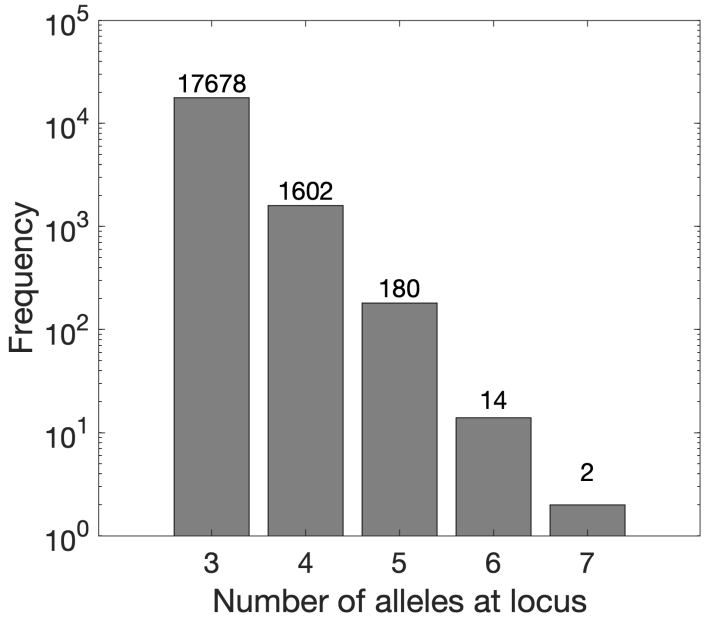


Figure 6: The number of sites that harbor more than two alleles in a natural D. melanogaster population. Data are shown for the chromosome 3R of the 'Linvilla' population obtained for three different time points (July 2009, July 2012, and October 2015), which is part of the DEST dataset (Kapun et al., 2021). We included only sites for which each allele was supported by at least three reads at each of the three sampled time points, i.e. we applied a hard filter.

Discussion

In this work we have presented a generalised framework to model multiallelic genetic diversity that incorporates complex forms of selection, as well as a method to estimate the selective force from empirical data. We have provided 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 when 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; Vellnow, Gossmann, and Waxman, 2024). 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; Siljestam and Rueffler, 2024). 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. For instance, 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, 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 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. (16), (19), and (21), 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.

In our description of the dynamics of the model, we mentioned the way recurrent mutation is incorporated into the evolutionary forces that generate allele frequency changes. There is another issue associated with mutation that we should mention. Consider the situation where we have allele frequency trajectory data, but mutation plays a significant role. Then, from mutation, we may have new alleles arising. The simplest description of such a situation is where we look at the set of *all* alleles that have been present in the population at *some time* for which we have data. A description that encompasses *all* such alleles can then be used from the outset. In this way, we never have to deal with a description that changes dimensionality. Some of the allele frequencies may be zero during some of the time that we have data, but our framework can still be applied. All quantities, such as the force of selection, are well-defined even when some allele frequencies are zero.

Average mutation rates can vary between eukaryotic species ranging from 0.01×10^{-9} to 55.58×10^{-9} (Wang and Obbard, 2023). However, mutation rates can be substantially higher in certain regions of the genome than in others (Gossmann, Woolfit, and Eyre-Walker, 2011; Nesta, Tafur, and Beck, 2021). We envisage that the primary applications of the present work are to evolutionary trajectories (describing allele or haplotype frequencies) on timescales of the order of tens or hundreds of generations. For such times scales, with typical mutation rates, we expect the influence of mutation on allele frequency trajectories to be very limited. Most new mutations—that are rare—will become lost after a very short time. If, however, mutations are considered to be important, e.g. because the locus of interest lies in a mutational hotspot or the timescale of interest is very long, they can be incorporated into our framework. Recurrent mutations, that transform one allele into another, can be incorporated by adding the term $\mathbf{MX}(t)$ to the right hand side of Eq. (4) (neglecting 'cross terms' from selection and mutation, which we assume are very small). New mutations can be incorporated by simply adding more components to the vector of allele frequencies.

Our model offers a versatile framework for application to empirical data, particularly through the estimation of the matrix of fitness effects, $\mathbf{F}$, from allele trajectory data. In this study, we estimated fitness effects for a genomic region on chromosome 16 of *S. cerevisiae* (and directly from this the matrix of relative fitnesses, $\mathbf{w}$). The estimated fitness effects were quite large in magnitude. In part, this may be attributable to the action of strong selection in a non

naturalistic environment during the experimental evolution that was reported. Additionally, the fitness effects, that we calculated, represent the concatenation of fitness effects over 30 mitotic generations. We note that two homozygote genotypes (A_2A_2 and A_4A_4) had estimated fitness effects of -1 which corresponds to complete lethality. However, these estimated values are close to the lowest possible value of the $F_{i,j}$, and may be sensitive to small numerical errors. For example, replacing such a fitness effect by $-1 + 10^{-6}$ would lead to a fitness effect associated with a *single mitotic generation* of approximately -0.3 which represents a quite strongly deleterious effect but not lethality. We therefore suggest caution in interpreting such results as being precisely lethal, but rather as being strongly deleterious, and susceptible to error. Generally, we expect the fitness effects of multicellular organisms in a 'natural' environment, to which they are reasonably adapted, to be much smaller. Beyond this issue, estimating $\mathbf{F}$ for four haplotypes allowed us to find evidence for heterozygote advantage, 'on average', but also to appreciate variability of the heterozygote fitnesses (some were advantageous, others disadvantageous) that would simply not be resolved within the framework of a biallelic model. With our approach, researchers can address key evolutionary questions about the prevalence of heterosis and the strength of selection that is acting on multiple alleles.

As demonstrated, our method can also be applied to genome-wide scans, where $\mathbf{F}$ is estimated for sliding windows across the entire genome. In an experimental evolution study with *S. cerevisiae* (Burke, Liti, and Long, 2014; Phillips et al., 2020), we found evidence for widespread heterozygote advantage throughout the genome. Notably, instances where a *homozygote* exhibited the highest fitness appeared to be confined to highly localized genomic regions, such as the terminal region of chromosome C13 (see Figure 5). The observed heterozygote advantage may stem from strong artificial selection and/or inbreeding experienced by the original strains prior to the experimental evolution. Consequently, these strains may have undergone intense selection against homozygotes when exposed to a novel environment during the experiment. Interestingly, heterozygote advantage have been reported in other studies on baker's yeast (Sellis et al., 2016; Aggeli et al., 2022).

To estimate $\mathbf{F}$ from empirical data, we recommend using trajectories with a minimum of 5 time points. While trajectories shorter than 5 time points may already capture the rank ordering of genotype fitnesses reasonably well (see Figure S2b), at least 5 time points are required for a robust estimation of $\mathbf{F}$ (see Figure S2a). For highly accurate estimates where genotype fitnesses converge closely to their true values, analyzing trajectories with 30 time points seems advisable. However, the benefit of adding more time points diminishes noticeably

after approximately 10 time points (see Figure S2). This makes our method suitable for experimental evolution studies involving organisms with short generation times, such as arthropods or nematodes. Importantly, it enables researchers to address diverse questions in evolutionary biology within the typical timeframe of a PhD project.

Our model is highly relevant for analyzing genomic datasets spanning multiple generations of wild populations (Kapun et al., 2021) or evolve-and-resequence experiments (Schlötterer et al., 2015). As genomic technologies continue to advance, we anticipate the availability of an increasing number of such datasets in the near future. Notably, our framework could be applied to novel cases, such as where allele frequencies are inferred from sequenced fossils or lake sediments (Decaestecker et al., 2007; Burger et al., 2007) or applied to the analysis of the ongoing evolution of insecticide resistance, e.g. in *Anopheles* mosquitoes (Edi et al., 2014). Furthermore, our model can facilitate genome scans to identify loci under balancing selection. For instance, it can identify loci where the estimated F-matrices make the long-term maintenance of multiple alleles more likely (Lewontin, Ginzburg, and Tuljapurkar, 1978; Siljestam and Rueffler, 2024). This in turn might enable predictions about evolutionary trajectories beyond the observed time frame (Wortel et al., 2023).

There are a number of practical considerations 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 S4a). Thus, dependent on the expected population size, it might be necessary to sequence a pooled sample at an enormous coverage. However, since sequencing costs have dropped substantially in recent years, 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 a 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 S4b). 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 it is evident that much diversity and evolutionary dynamics may be missed because it occurs below threshold.

Third, if SNPs are considered in a population-wide sample, then in principle a single base position can, at most, have four alleles. However, we would expect selection to act at the 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). Such natural assemblages will lead, effectively, 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 not only leads to a realistic description of the variation present, but may also be of high importance, since it may allow capture of 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, Zhao, 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 will, potentially, need to be moderate in order to identify heterozygous sites, but it would then need to be scaled with the number of individuals tested (see Figure S4a). However, often bioinformatic pipelines take population-wide signatures of polymorphic sites into account, in order 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.

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

Aggeli, Dimitra et al. (June 2022). "Overdominant and Partially Dominant Mutations Drive Clonal Adaptation in Diploid Saccharomyces Cerevisiae". In: Genetics 221.2, iyac061. ISSN: 1943-2631. DOI: 10.1093/genetics/iyac061.

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.

Burger, J. et al. (Mar. 2007). "Absence of the Lactase-Persistence-Associated Allele in Early Neolithic Europeans". In: Proceedings of the National Academy of Sciences 104.10, pp. 3736–3741. DOI: 10.1073/pnas.0607187104.

Burke, Molly K., Gianni Liti, and Anthony D. Long (Dec. 2014). "Standing Genetic Variation Drives Repeatable Experimental Evolution in Outcrossing Populations of Saccharomyces Cerevisiae". In: Molecular Biology and Evolution 31.12, pp. 3228–3239. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msu256.

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.

Decaestecker, Ellen et al. (Dec. 2007). "Host–Parasite 'Red Queen' Dynamics Archived in Pond Sediment". In: Nature 450.7171, pp. 870–873. ISSN: 1476-4687. DOI: 10.1038/nature06291.

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.

Edi, Constant V. et al. (Mar. 2014). "CYP6 P450 Enzymes and ACE-1 Duplication Produce Extreme and Multiple Insecticide Resistance in the Malaria Mosquito Anopheles Gambiae". In: PLOS Genetics 10.3, e1004236. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004236.

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.

Kapun, Martin et al. (Dec. 2021). "Drosophila Evolution over Space and Time (DEST): A New Population Genomics Resource". In: Molecular Biology and Evolution 38.12, pp. 5782–5805. ISSN: 1537-1719. DOI: 10.1093/molbev/msab259.

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.

Phillips, Mark A. et al. (2020). "Increased Time Sampling in an Evolve-and-Resequence Experiment with Outcrossing Saccharomyces Cerevisiae Reveals Multiple Paths of Adaptive Change". In: Molecular Ecology 29.24, pp. 4898–4912. ISSN: 1365-294X. DOI: 10.1111/mec.15687.

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.

Sellis, Diamantis et al. (July 2016). "Heterozygote Advantage Is a Common Outcome of Adaptation in Saccharomyces Cerevisiae". In: Genetics 203.3, pp. 1401–1413. ISSN: 1943-2631. DOI: 10.1534/genetics.115.185165.

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.

Siljestam, Mattias and Claus Rueffler (Nov. 2024). "Heterozygote Advantage Can Explain the Extraordinary Diversity of Immune Genes". In: eLife 13. Ed. by Yaroslav Ispolatov, Detlef Weigel, and Yaroslav Ispolatov, e94587. ISSN: 2050-084X. DOI: 10.7554/eLife.94587.

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.

The Mathworks Inc. (2025). MATLAB. The MathWorks Inc. Natick, Massachusetts, United States.

Torczon, Virginia (Feb. 1997). "On the Convergence of Pattern Search Algorithms". In: SIAM Journal on Optimization 7.1, pp. 1–25. ISSN: 1052-6234. DOI: 10.1137/S1052623493250780.

Tuckwell, Henry C. (Dec. 2019). Elementary Applications of Probability Theory: With an Introduction to Stochastic Differential Equations. 2nd edition. Chapman and Hall/CRC. ISBN: 978-0-367-44905-6.

Vellnow, Nikolas, Toni I. Gossmann, and David Waxman (Apr. 2024). "The Pseudoentropy of Allele Frequency Trajectories, the Persistence of Variation, and the Effective Population Size". In: BioSystems 238, p. 105176. ISSN: 0303-2647. DOI: 10.1016/j.biosystems.2024.105176.

Wang, Yiguan and Darren J Obbard (Aug. 2023). "Experimental Estimates of Germline Mutation Rate in Eukaryotes: A Phylogenetic Meta-Analysis". In: Evolution Letters 7.4, pp. 216–226. ISSN: 2056-3744. DOI: 10.1093/evlett/qrad027.

Wang, Yunhao, Yue Zhao, 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.

Waxman, David and Andrew D. J. Overall (Apr. 2020). "Influence of Dominance and Drift on Lethal Mutations in Human Populations". In: Frontiers in Genetics 11. ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00267.

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

Wortel, Meike T. et al. (2023). "Towards Evolutionary Predictions: Current Promises and Challenges". In: Evolutionary Applications 16.1, pp. 3–21. ISSN: 1752-4571. DOI: 10.1111/eva.13513.

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 Evolutionarily equivalent fitness effects

In this appendix we relate a property of fitness effects to a property of relative fitnesses. Relative fitnesses are defined only up to a multiplicative constant. In other words, if $w_{i,j}$ represents fitness, or relative fitness, of the A_iA_j genotype, then with k a positive constant, replacing $w_{i,j}$ by $k \times w_{i,j}$, for all i and j, will produce identical allele frequency trajectories. The corresponding property of the fitness effects (the $F_{i,j}$) is that replacing $F_{i,j}$ by $(k-1)+kF_{i,j}$, for all i and j, leads to identical allele frequency trajectories. Thus the matrix of fitness effects with elements $F_{i,j}$ is *evolutionarily equivalent* to the corresponding matrix with elements $(k-1)+kF_{i,j}$.

To show the above, we note the feature that allele frequency dynamics depends only on *ratios* of fitnesses. This allows the use of *relative fitnesses*, which are defined only up to a multiplicative constant. This means if the fitness of the A_iA_j genotype is $w_{i,j}$, then, for all i and j, replacing $w_{i,j}$ by $k \times w_{i,j}$ where k is a positive constant, leads to identical allele frequency dynamics. It also means if, e.g., we designate A_1A_1 as the reference genotype, then we can choose $k = 1/w_{1,1}$ and the resulting fitness of all genotypes are relative to the fitness of the A_1A_1 genotype.

In terms of the $F_{i,j}$, the fitness of the A_iA_j genotype is proportional to $1 + F_{i,j}$. An *evolutionarily equivalent* alternative set of fitness effects, written $\mathbf{F}'$ can be defined, for all i and j, by

$$1 + F'_{i,j} = k \times (1 + F_{i,j}) \tag{24}$$

where k a positive constant. This equation can be rewritten as

$$F'_{i,j} = (k-1) + k \times F_{i,j} \tag{25}$$

and the $F'_{i,j}$ can be used in place of the $F_{i,j}$ with no effect on allele frequency dynamics.

Additionally, if we, e.g., designate A_1A_1 as the reference genotype, and choose it to have a relative fitness of unity, then this can be taken to correspond to a *vanishing* fitness effect of this genotype. Given a set of fitness effects, $F_{i,j}$, where $F_{1,1} \neq 0$, using the transformation in Eq. (25) with $k = 1/(1 + F_{1,1})$ leads to a new and evolutionarily equivalent set of fitness effects $F'_{i,j} = (k-1) + k \times F_{i,j}$ which have the property $F'_{1,1} = 0$.

Such a transformation, which makes a particular genotype the reference genotype, gives it a vanishing fitness effect. Any *non lethal genotype* (i.e., that has a positive fitness, or

equivalently has a fitness effect > -1) can be converted to the reference genotype.

B 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

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

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

$$2N\mathbf{X}(t+1) \sim \mathrm{Mult}(2N, \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)). \tag{27}$$

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

$$E\left[\mathbf{M}(t)|\mathbf{X}(t) = \mathbf{x}\right] = 2N\left[\mathbf{x} + \mathbf{D}(\mathbf{x})\right]. \tag{28}$$

It follows that equivalent to Eq. (27) is the equation $\mathbf{X}(t+1) = \mathbf{M}(t)/(2N)$ which we write as

$$\mathbf{X}(t+1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t)) + \boldsymbol{\xi}(t) \tag{29}$$

[6]We note that Eq. (27) is an approximation when fitnesses are not multiplicative, however for $|F_{i,j}| \ll 1$ it is a good approximation (Nagylaki, 1992) (P. 252).

where

$$\boldsymbol{\xi}(t) = \frac{\mathbf{M}(t)}{2N} - \left[\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))\right]. \tag{30}$$

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.,

$$E\left[\boldsymbol{\xi}(t)|\mathbf{X}(t) = \mathbf{x}\right] = \mathbf{0} \tag{31}$$

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

$$E\left[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}\right] = \frac{\mathbf{V}(\mathbf{x} + \mathbf{D}(\mathbf{x}))}{2N} \tag{32}$$

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

$$\mathbf{V}(\mathbf{x}) = \mathrm{diag}(\mathbf{x}) - \boldsymbol{x}\boldsymbol{x}^T \tag{33}$$

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)

$$E\left[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}\right] \simeq \frac{\mathbf{V}(\mathbf{x})}{2N}. \tag{34}$$

We thus arrive at a description of the dynamics given by Eq. (29) with the primary properties of $\boldsymbol{\xi}(t)$ given by Eqs. (31) and (34).

We note that under a diffusion approximation, both time and allele frequencies are approximated as continuous quantities (see e.g., (Kimura, 1955)). This simply entails Eq. (29) becoming $d\mathbf{X}(t) = \mathbf{D}(\mathbf{X}(t))dt + \boldsymbol{\xi}(t)$. The form of $\boldsymbol{\xi}(t)$, for continuous time and continuous frequencies, has to have properties that lead to $E\left[\boldsymbol{\xi}(t)|\mathbf{X}(t) = \mathbf{x}\right] = \mathbf{0}$ and $E\left[\boldsymbol{\xi}(t)\boldsymbol{\xi}^T(t)|\mathbf{X}(t) = \mathbf{x}\right] = \frac{\mathbf{V}(\mathbf{x}(t))}{2N}dt$. These are closely analogous Eqs. (31) and (34) and can both be achieved by the replacement $\boldsymbol{\xi}(t) = \sqrt{\frac{\mathbf{V}(\mathbf{X}(\mathbf{t}))}{2N}}d\mathbf{W}(\mathbf{t})$, where $\mathbf{W}(t)$ is a column vector with n elements that are 'Gaussian noise', i.e., independent Wiener processes (Tuckwell, 2019).

Lastly, deviations of a population from being 'ideal' are often incorporated into the diffusion approximation by replacing the census population size, N, by an 'effective' population size N_e. This leads to $\mathbf{X}(t)$ obeying the stochastic differential equation $d\mathbf{X}(t) = \mathbf{D}(\mathbf{X}(t))dt + \sqrt{\frac{\mathbf{V}(\mathbf{X}(\mathbf{t}))}{2N_e}}d\mathbf{W}(\mathbf{t})$.

C 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 A_i with $i = 1, 2, \ldots, n$ and assume the A_iA_j genotype has a fitness proportional to $1 + F_{i,j}$ with $F_{i,j} = F_{j,i}$.

Let x_i be the frequency of allele A_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 A_iA_i genotype and (ii) offspring with the A_iA_j genotype, with $j \neq i$. With all sums running from 1 to n unless otherwise stated, this leads to

$$x_i' = \frac{2(1+F_{i,i})x_i^2 + \sum_{j(\neq i)}(1+F_{i,j})2x_ix_j}{2\left[(1+F_{i,i})x_i^2 + \sum_{j(\neq i)}(1+F_{i,j})2x_ix_j + \sum_{j(\neq i),k(\neq i)}\left(1+F_{j,k}\right)x_jx_k\right]}. \tag{35}$$

This can be simplified to

$$x_i' = \frac{x_i\left(1 + \sum_j F_{i,j}x_j\right)}{1 + \sum_{j,k} F_{j,k}x_jx_k}. \tag{36}$$

Alternatively, we have

$$x_i' = x_i + \frac{x_i\left(\sum_j F_{i,j}x_j - \sum_{j,k} F_{j,k}x_jx_k\right)}{1 + \sum_{j,k} F_{j,k}x_jx_k} \tag{37}$$

and we rewrite this equation as

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

$$= x_i + \frac{\sum_{j,k} \left(\delta_{i,j}x_j - x_ix_j\right) F_{j,k}x_k}{1 + \sum_{j,k} x_jF_{j,k}x_k}. \tag{38}$$

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

$$V_{i,j}(\mathbf{x}) = x_i\delta_{i,j} - x_ix_j \tag{39}$$

we can write Eq. (38) as $x_i' = x_i + \frac{[\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}]_i}{1+\mathbf{x}^T\mathbf{F}\mathbf{x}}$ where $\mathbf{F}$ is a matrix with elements $F_{i,j}$. From this we obtain

$$\mathbf{x}' = \mathbf{x} + \frac{\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{F}\mathbf{x}} \tag{40}$$

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

$$\mathbf{D}(\mathbf{x}) = \frac{\mathbf{V}(\mathbf{x})\mathbf{F}\mathbf{x}}{1 + \mathbf{x}^T\mathbf{F}\mathbf{x}}. \tag{41}$$

D Estimating the $\mathbf{F}$ matrix

Let the exact fitness of the A_iA_j genotype, for all i and j, be proportional to $1 + F^{(0)}_{i,j}$ and let $\mathbf{F}^{(0)}$ denote the $n \times n$ symmetric matrix containing the $F^{(0)}_{i,j}$. In this appendix we give a method for estimating $\mathbf{F}^{(0)}$ from allele frequency trajectory data via an optimisation procedure.

We present the method under the assumption that $\mathbf{F}^{(0)}$ is independent of allele frequencies. Extensions, to e.g., frequency dependent selection, may be possible if an $\mathbf{F}^{(0)}$ matrix with a specific frequency dependent form is known or adopted, and parameters within this form are included within the optimisation procedure.

We shall first present a method of estimating $\mathbf{F}^{(0)}$ when allele frequencies are known at consecutive generations. We then extend this to the case where allele frequencies are only known only for particular, non-consecutive, generations.

D.1 Allele frequencies known every generation

We begin with data from an assumed effectively infinite population, for a locus with n alleles. The data takes the form of a trajectory

$$\mathbf{X}|_0^T \equiv [\mathbf{X}(0), \mathbf{X}(1), \mathbf{X}(2),..., \mathbf{X}(T)] \tag{42}$$

where for $t = 0, 1, ..., T$, the quantity $\mathbf{X}(t)$ is a column vector of length n, and contains n allele frequencies. Thus $\mathbf{X}|_0^T$ is an $n \times (T + 1)$ matrix.

We work under the key assumption that the trajectory $\mathbf{X}|_0^T$ originates from purely selective dynamics, as described in the main text (Eq. (26)), and given by[7]

$$\mathbf{X}(t + 1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t), \mathbf{F}^{(0)}). \tag{43}$$

The form of the matrix $\mathbf{F}^{(0)}$ is unknown and our objective is to estimate it from the *data* (i.e., from the trajectory, $\mathbf{X}|_0^T$).

To proceed, we define another trajectory

$$\mathbf{Y}|_1^T \equiv [\mathbf{Y}(1), \mathbf{Y}(2),..., \mathbf{Y}(T)] \tag{44}$$

[7] In the main text we wrote the equation for $\mathbf{X}(t)$ as $\mathbf{X}(t + 1) = \mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$. In this appendix we find it helpful to be more explicit, and shall additionally list, as an argument of $\mathbf{D}$, the $\mathbf{F}$ matrix that it depends upon.

which is a matrix of size $n \times T$. We determine $\mathbf{Y}|_1^T$ from the known trajectory data (i.e., $\mathbf{X}|_0^T$) and an arbitrarily chosen $\mathbf{F}$ matrix via

$$\begin{aligned}
\mathbf{Y}(1) &= \mathbf{X}(0) + \mathbf{D}(\mathbf{X}(0), \mathbf{F}) \\
\mathbf{Y}(2) &= \mathbf{X}(1) + \mathbf{D}(\mathbf{X}(1), \mathbf{F}) \\
&\vdots \\
\mathbf{Y}(T) &= \mathbf{X}(T-1) + \mathbf{D}(\mathbf{X}(T-1), \mathbf{F}).
\end{aligned} \tag{45}$$

That is, each $\mathbf{Y}(t)$ is determined from: (i) the data frequencies of one generation earlier, namely $\mathbf{X}(t-1)$, and (ii) the $\mathbf{F}$ matrix. We can say $\mathbf{Y}(t)$ is the *predicted value* of $\mathbf{X}(t)$, given $\mathbf{X}(t-1)$ and the matrix $\mathbf{F}$.

We now define a 'cost function' $C(\mathbf{F})$ that is minimised when $\mathbf{F} = \mathbf{F}^{(0)}$. With the norm of an n component column vector with elements x_i $(i = 1, 2, ..., n)$ given by

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2 \tag{46}$$

we take the cost of $\mathbf{F}$ to be

$$C(\mathbf{F}) = \sum_{t=1}^{T} \|\mathbf{X}(t) - \mathbf{Y}(t)\|^2 . \tag{47}$$

If $\mathbf{F} = \mathbf{F}^{(0)}$ then $\mathbf{Y}(t) = \mathbf{X}(t)$ and the cost vanishes:

$$C(\mathbf{F}^{(0)}) = 0. \tag{48}$$

The cost also vanishes for $\mathbf{F}$ matrices that are *evolutionarily equivalent* to $\mathbf{F}^{(0)}$ (see Appendix A) and hence produce the same allele frequency trajectories. For any $\mathbf{F}$ that is not evolutionarily equivalent to $\mathbf{F}^{(0)}$ we have $C(\mathbf{F}) > 0$ since the $\mathbf{Y}(t)$ will not generally coincide with the $\mathbf{X}(t)$.

When we carry out the minimisation, we require that two points are satisfied.

(i) The matrix $\mathbf{F}$ is unambiguous.

We achieve this by always taking $F_{1,1} = 0$, which sets A_1A_1 to be the reference genotype, and the fitnesses of all genotypes are then given relative to the fitness of this genotype (see Appendix A).

(ii) When changes of the matrix $\mathbf{F}$ are made, it is always kept a symmetric matrix.

When there are n alleles, there are $n(n+1)/2$ independent elements of $\mathbf{F}$. Satisfying point (i) reduces this to $n(n+1)/2 - 1$ independent elements. We satisfy point (ii) by expressing

results in terms of a vector $\mathbf{v}$ with $n(n+1)/2-1$ elements, and using a function $\mathbf{S}(\mathbf{v})$ that produces, from the elements of $\mathbf{v}$, a symmetric matrix with a vanishing $(1,1)$ element. Thus the actual procedure carried out is a minimisation of $C(\mathbf{S}(\mathbf{v}))$ over $\mathbf{v}$.

With $\hat{\mathbf{v}}$ the vector that minimises $C(\mathbf{S}(\mathbf{v}))$, the estimate of $\mathbf{F}^{(0)}$ is $\mathbf{S}(\hat{\mathbf{v}})$.

D.2 Allele frequencies not sampled every generation

We now assume the allele frequencies are only known in generations T_1, T_2, ..., T_g where generally $T_{i+1}-T_i \neq 1$ and this interval may be variable. We define the 'sampled' trajectory by

$$\widetilde{\mathbf{X}} \equiv [\mathbf{X}(T_1), \mathbf{X}(T_2),..., \mathbf{X}(T_g)] \tag{49}$$

which is a matrix of size $n \times g$.

To proceed, we define another sampled trajectory

$$\widetilde{\mathbf{Y}} \equiv [\mathbf{Y}(T_2), \mathbf{Y}(T_3),..., \mathbf{Y}(T_g)] \tag{50}$$

which is a matrix of size $n \times (g-1)$. We determine $\widetilde{\mathbf{Y}}$ from $\widetilde{\mathbf{X}}$ and an arbitrarily chosen $\mathbf{F}$ matrix, with the calculations as follows.

To determine $\mathbf{Y}(T_2) \equiv \mathbf{Y}(T_1+(T_2-T_1))$ we carry out the T_2-T_1 calculations[8]

$$\begin{aligned}
\mathbf{Y}(T_1+1) &= \mathbf{X}(T_1) + \mathbf{D}(\mathbf{X}(T_1), \mathbf{F}) \\
\mathbf{Y}(T_1+2) &= \mathbf{Y}(T_1+1) + \mathbf{D}(\mathbf{Y}(T_1+1), \mathbf{F}) \\
\mathbf{Y}(T_1+3) &= \mathbf{Y}(T_1+2) + \mathbf{D}(\mathbf{Y}(T_1+2), \mathbf{F}) \\
\mathbf{Y}(T_1+4) &= \mathbf{Y}(T_1+3) + \mathbf{D}(\mathbf{Y}(T_1+3), \mathbf{F}) \\
&\vdots \\
\mathbf{Y}(T_1+(T_2-T_1)) &= \mathbf{Y}(T_1+(T_2-T_1)-1) + \mathbf{D}(\mathbf{Y}(T_1+(T_2-T_1)-1), \mathbf{F}).
\end{aligned} \tag{51}$$

[8]Note that in the first equation, $\mathbf{X}(T_1)$ is used on the right hand side, while in all subsequent equations $\mathbf{Y}$'s are used.

Equivalently, to determine $\mathbf{Y}(T_3) \equiv \mathbf{Y}(T_2 + (T_3 - T_2))$ we carry out the $T_3 - T_2$ calculations

$$\begin{aligned}
\mathbf{Y}(T_2+1) &= \mathbf{X}(T_2) + \mathbf{D}(\mathbf{X}(T_2), \mathbf{F}) \\
\mathbf{Y}(T_2+2) &= \mathbf{Y}(T_2+1) + \mathbf{D}(\mathbf{Y}(T_2+1), \mathbf{F}) \\
\mathbf{Y}(T_2+3) &= \mathbf{Y}(T_2+2) + \mathbf{D}(\mathbf{Y}(T_2+2), \mathbf{F}) \\
\mathbf{Y}(T_2+4) &= \mathbf{Y}(T_2+3) + \mathbf{D}(\mathbf{Y}(T_2+3), \mathbf{F}) \\
&\vdots \\
\mathbf{Y}(T_2 + (T_3 - T_2)) &= \mathbf{Y}(T_2 + (T_3 - T_2) - 1) + \mathbf{D}(\mathbf{Y}(T_2 + (T_3 - T_2) - 1), \mathbf{F}).
\end{aligned} \tag{52}$$

We proceed in this way until we have calculated all $\mathbf{Y}(T_1), \mathbf{Y}(T_2), ..., \mathbf{Y}(T_g)$.

The 'cost' of matrix $\mathbf{F}$ is then taken to be

$$C(\mathbf{F}) = \sum_{j=2}^{g} \|\mathbf{X}(T_j) - \mathbf{Y}(T_j)\|^2 . \tag{53}$$

As in Section D.1, we parameterise $\mathbf{F}$ in terms of an $n(n+1)/2 - 1$ component vector, $\mathbf{v}$, and use a function $\mathbf{S}(\mathbf{v})$ to construct $\mathbf{F}$ from $\mathbf{v}$:

$$\mathbf{F} = \mathbf{S}(\mathbf{v}). \tag{54}$$

The minimisation that is carried out is that of $C(\mathbf{S}(\mathbf{v}))$ over $\mathbf{v}$. As before, with $\hat{\mathbf{v}}$ the vector that minimises $C(\mathbf{S}(\mathbf{v}))$, the estimate of $\mathbf{F}^{(0)}$ is $\mathbf{S}(\hat{\mathbf{v}})$.

Supplementary Material

A comprehensive representation of genotypic selection beyond two alleles, and its estimation from time-series data

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

In Sections 1 and 2 of this Supplementary Material we give details of the estimation of fitness effect matrices from frequency trajectories. In Section 3 we explore how the length of a frequency trajectory influences the estimate of the fitness effect matrix. In Section 4 we explore the sensitivity of the number of alleles at a locus to the level of filtering, and in Section 5 we provide calculations and simulations to illustrate detection and threshold effects.

1 Details of estimating the $\mathbf{F}$ matrix from empirical data

For the estimation of $\mathbf{F}$ with our 'cost function' $C(\mathbf{F})$ (see Eq. 53 in Appendix D.2), we used the 'fminunc' optimisation function in MATLAB R2024b (The Mathworks Inc., 2025), which implements gradient-based methods for unconstrained minimization. This method performed well, but sometimes lead to an estimated $\mathbf{F}$ with some elements having values less than -1, i.e., $F_{i,j} < -1$. Such values correspond to negative fitness values, which are not possible. In the cases where this occurred, we switched to a constrained minimization, using the patternsearch algorithm, which is based on derivative-free direct search methods (Torczon, 1997) and which implemented the constraint $F_{i,j} \geq -1$.

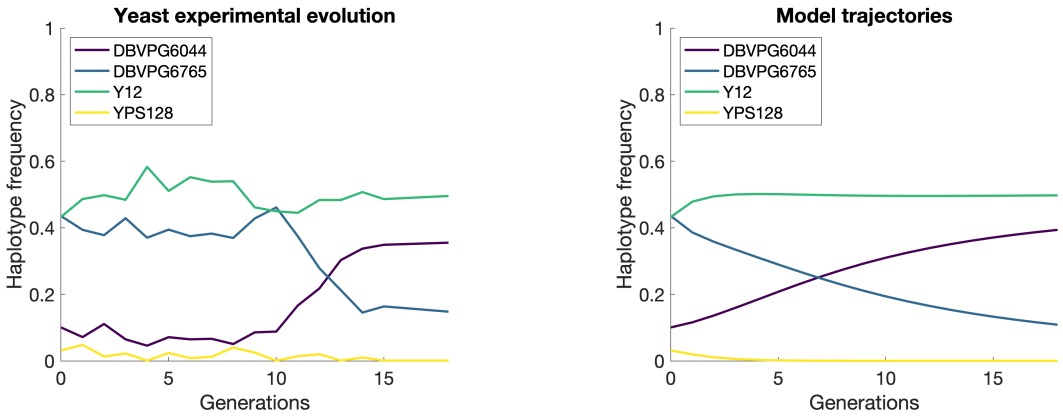
For our scan of the entire yeast genome, we started with 5779 sliding 10 KB windows (step size: 2 KB). In 25 windows the 'fminunc' optimiser failed to converge. Furthermore, in 29 windows all trajectories remained constant at 0.25 throughout the experiment, which led to an estimated $\mathbf{F}$ containing only fitness effects of zero. These estimated $\mathbf{F}$-matrices are uninformative (they do not allow us find a most fit genotype as all genotypes have the same fitness). Consequently, we excluded those windows with $\mathbf{F}$-matrices that did not converge or were uninformative. We thus retained data from 5725 windows for further analysis.

2 Estimating F matrix from empirical data: Another yeast example

As another example, we also estimated $\mathbf{F}$ (with constrained minimization) for trajectories of the haplotypes in the 10 KB window around position 436424 on chromosome C12, for replicate population 3 (see Figure S1a), from the data set of (Burke, Liti, and Long, 2014; Phillips et al., 2020). The estimated $\mathbf{F}$ and the corresponding matrix of relative fitnesses $\mathbf{w}$ are:

$$\mathbf{F} = \begin{pmatrix} 0.000\,, & 0.925\,, & 6.908\,, & 2.877 \\ 0.925\,, & -0.648\,, & 5.836\,, & 3.672 \\ 6.908\,, & 5.836\,, & 0.234\,, & -1.000 \\ 2.877\,, & 3.672\,, & -1.000\,, & -1.000 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} 0.127\,, 0.243\,, 1.000\,, 0.490 \\ 0.243\,, 0.045\,, 0.864\,, 0.591 \\ 1.000\,, 0.864\,, 0.156\,, 0.000 \\ 0.490\,, 0.591\,, 0.000\,, 0.000 \end{pmatrix}. \tag{S1}$$

The estimated $\mathbf{F}$ leads to trajectories that resemble the empirical ones although the resemblance is not very strong (see Figure S1b). In particular, we hypothesize that after generation 9 the selective forces may have changed. Since under that scenario $\mathbf{F}$ would not have been constant, our estimation method may not find a well-matching $\mathbf{F}$.



(a) Empirical haplotype frequencies from yeast data.

(b) Trajectories modelled based on the estimated $\mathbf{F}$ matrix (right).

Figure S1: Estimating the $\mathbf{F}$ matrix from brewer's yeast haplotype frequency trajectories. Data is from a 10 KB window around position 436424 on chromosome C12 for replicate 3 (Burke, Liti, and Long, 2014; Phillips et al., 2020). Haplotypes are DBVPG6044, West African in dark blue, DBVPG6765, wine/European in purple, Y12, sake/Asian in green, and YPS128, North American in yellow.

3 How the length of allele frequency trajectories affects the estimate of $\mathbf{F}$

In order to investigate how the length of allele frequency trajectories affects the estimate of $\mathbf{F}$, we proceeded as follows.

We generated trajectories based on a known $\mathbf{F}^{(\mathbf{0})}$. Using trajectories produced in this way, but of different length (i.e., different time duration), yielded different estimates of the $\mathbf{F}$ matrix. We compared the estimates with the actual fitness effect matrix, namely $\mathbf{F}^{(\mathbf{0})}$.

We began by first generating an $\mathbf{F}^{(\mathbf{0})}$ with elements, drawn form a normal distribution with mean 0 and standard deviation 0.1, along with random starting frequencies. From these, we calculated deterministic trajectories of length 50 generations. We next used our method to estimate $\mathbf{F}$, based on trajectories of different lengths (corresponding to different experimental observation periods), ranging from 1 to 50 generations. Next, we calculated, for each estimated $\mathbf{F}$ matrix, the distance to the true fitness effect matrix (namely $F^{(0)}$), along with the corresponding Spearman rank correlation.

In Appendix D.1 we introduced the $\mathbf{v}$ vector, which contains the independent elements of the matrix of fitness effects. The distance was calculated as

$$distance(\mathbf{F}^{(\mathbf{0})}, \mathbf{F}) = \frac{\|\mathbf{v}_{\text{opt}} - \mathbf{v}_0\|}{\|\mathbf{v}_0\|} \tag{S2}$$

where $\|\cdot\|$ represents the Euclidean norm, while $\mathbf{v}_{\text{opt}}$ and $\mathbf{v}_0$ represent the vector $\mathbf{v}$ describing the matrix of fitness effects at the start, and at the end, of the optimisation, respectively.

Finally we repeated this procedure for 100 independently generated $\mathbf{F}^{(\mathbf{0})}$ matrices. From these matrices, for each of the different observation lengths, we calculated the median, along with the 25% and 75% quantiles of the distance between $\mathbf{F}^{(\mathbf{0})}$ and $\mathbf{F}$. In addition to calculating distances and rank correlation for $\mathbf{F}$, we also calculated the distance and rank correlation for $\mathbf{w}$, the matrix of relative fitnesses. Here, the distance is given by

$$distance(\mathbf{w}^{(\mathbf{0})}, \mathbf{w}) = \|\mathbf{v}_{\text{opt,w}} - \mathbf{v}_{0,w}| \tag{S3}$$

where $\|\cdot\|$ represents the Euclidean norm, and $\mathbf{v}_{\text{opt,w}}$ and $\mathbf{v}_{0,w}$ represent the vector $\mathbf{v}$ at the start and at the end of the optimization, respectively, when transformed to relative fitness values.

We observe good convergence of the distance and correlation measures within 50 generations (see Figure S2). A substantial part of the improvement in the $\mathbf{F}$-estimation occurs within the first 10 generations and a very well estimated $\mathbf{F}$ is reached after 35 generations (see Figure S2a). The $\mathbf{w}$-estimation seems to improve at an even faster rate, with substantial improvements occurring within the first 5 generations. The rank ordering of the estimated $\mathbf{F}$- and $\mathbf{w}$-values correlate strongly with the rank ordering of $\mathbf{F}^{(\mathbf{0})}$- and $\mathbf{w}^{(\mathbf{0})}$-values after just 10 generations and reach an almost perfect correlation after 20 generations (see Figures S2b and S2d). This suggests that statements about the rank ordering of estimated fitness values of different genotypes will typically be correct after relatively short observation periods. By contrast, for the same observation periods, the estimated fitness value of a genotype may have a larger uncertainty attached to it.

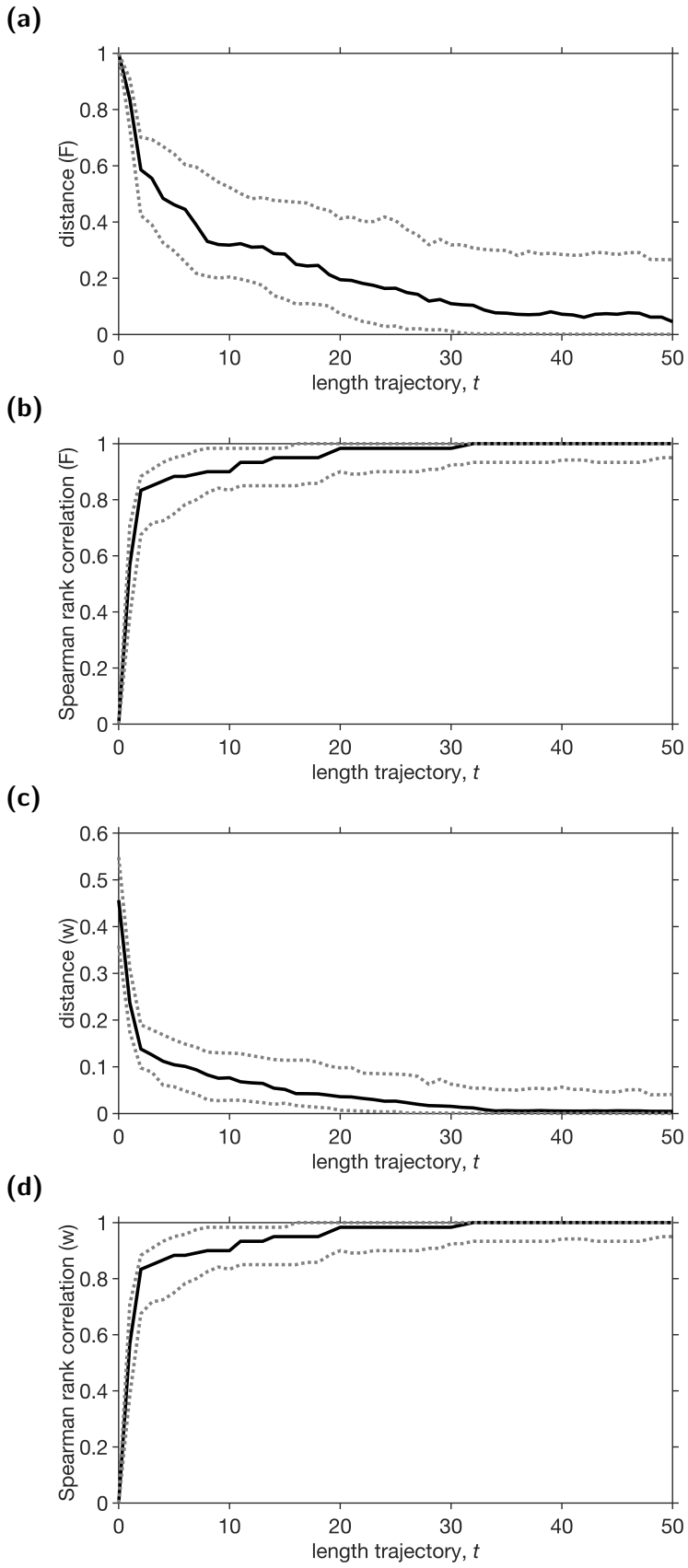


Figure S2: Convergence of the estimated $\mathbf{F}$ to the actual matrix, namely, $\mathbf{F}^{(\mathbf{0})}$, as a function of length of the observation period. Data is from 100 replicate simulations with random $\mathbf{F}^{(\mathbf{0})}$ matrices. Shown are the median (black line), along with 25% and 75%-quantiles (grey dotted lines), across these 100 replicate simulations. a) Distance between estimated $\mathbf{F}$ and $\mathbf{F}^{(\mathbf{0})}$, b) Spearman rank correlation coefficient, ρ, between estimated $\mathbf{F}$ and $\mathbf{F}^{(\mathbf{0})}$, c) distance between estimated $\mathbf{w}$ and $\mathbf{w}^{(\mathbf{0})}$, and d) Spearman rank correlation coefficient, ρ, between estimated $\mathbf{w}$ and $\mathbf{w}^{(\mathbf{0})}$ for different observation period lengths.

4 Multiallelic sites in DEST and their filtering

Whether an allele is considered present in a population may depend on many factors, e.g., the importance attributed to the avoidance of false-positives vs. false-negatives. This in turn will influence how strictly filtering thresholds are implemented. A simple way of filtering the sites with called haplotypes, from the pooled sequencing data we used in our main manuscript, is hard filtering based on the number of reads that support a haplotype. In the main manuscript we chose to only include sites at which at least three reads supported the presence of a haplotype at each of the three time points. However, setting the filter threshold differently, may lead to a different number of included sites with multiple alleles (more for a lower threshold, fewer for a higher threshold - see Figure S3). Note, however, that even when applying a high threshold of at least seven reads, we still detect a substantial number of sites that harbour three or four alleles (see Figure S3, bottom right).

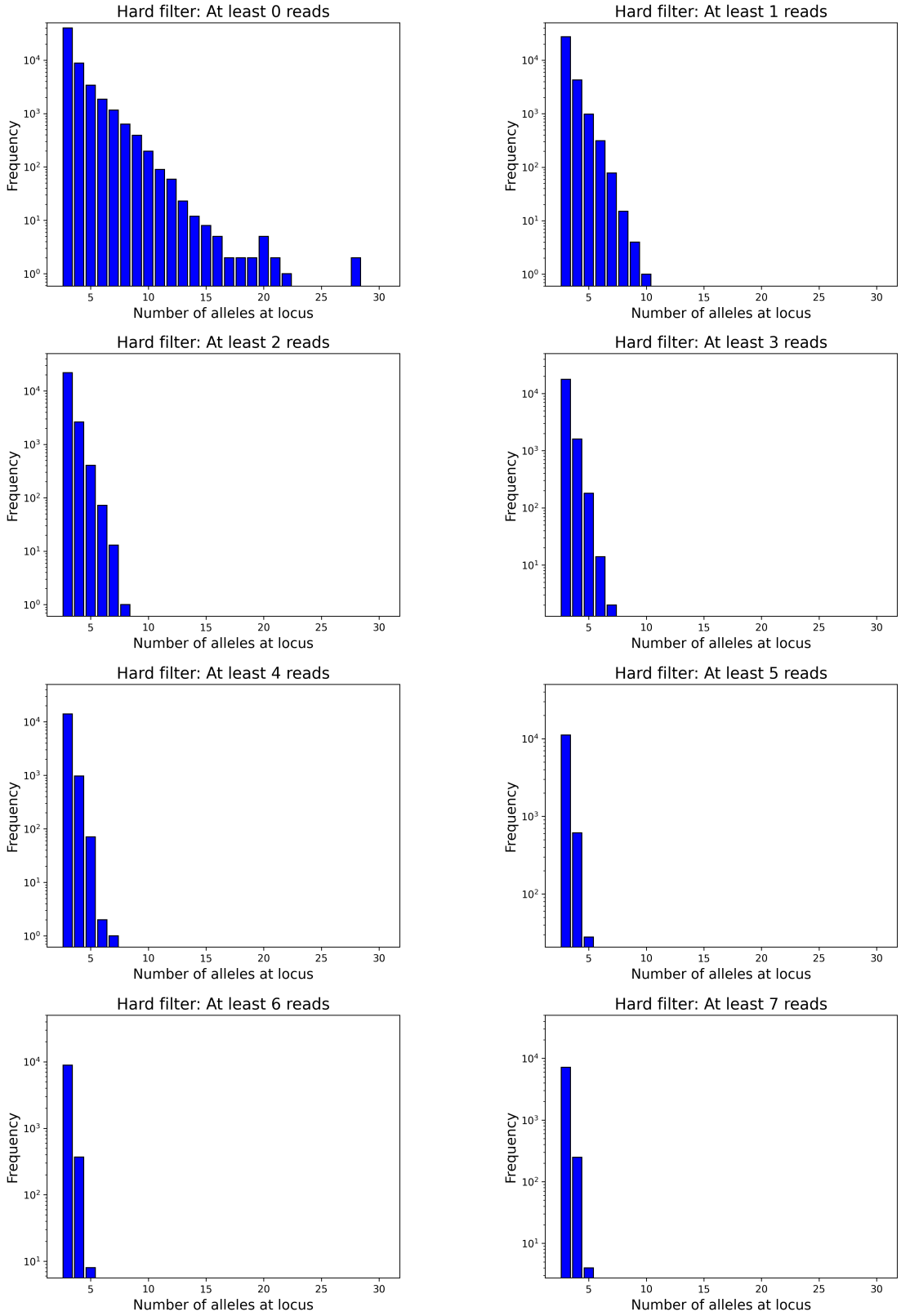


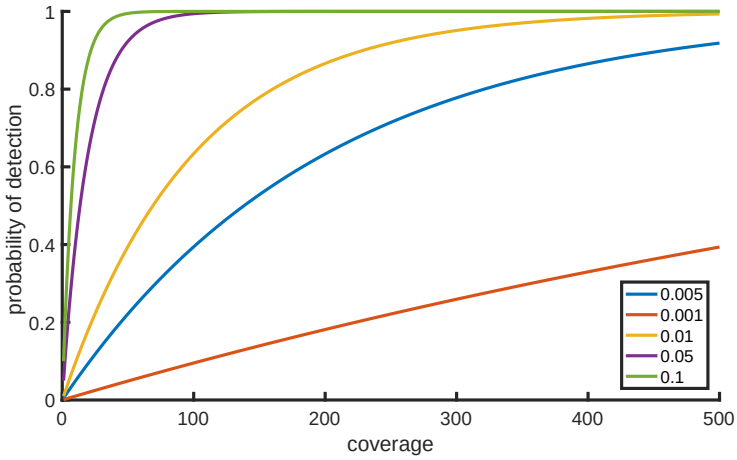
Figure S3: Sensitivity of number of alleles on hard filtering. Histograms are shown for the number of sites harbouring different numbers of alleles. The only sites included were those in which all alleles/haplotypes equalled or exceeded a minimum number of reads, at each of the three time points. This filtering criterion, for the minimum number of reads, was varied in the different subplots.

5 Probability of detection and thresholds

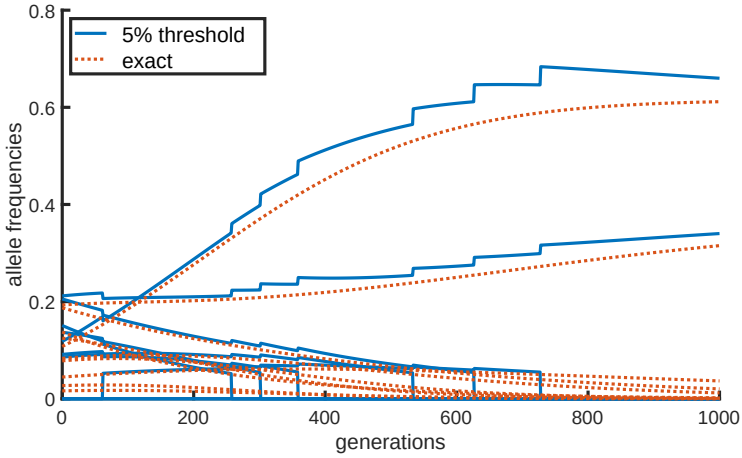
Here we provide calculations and simulations to substantiate some of the arguments given in the Discussion of the main text, concerning detection and threshold effects.

We used MATLAB for this purpose.

First, we modelled the probability of detection, for a coverage of x, as 1 minus the probability that the allele will not be contained in any of the x sequenced reads, i.e., probability of detection $= 1 - (1 - p)^x$, where p is the relative frequency of the allele. We then plotted the probability of detection against the coverage for several values of p (see Figure S4a). Second, we simulated evolution of a locus with ten alleles, with a randomly generated $\mathbf{F}$ matrix. We then imitated the effect that a detection threshold would have on the allele frequency trajectories to be analyzed (see Figure S4b).



(a)



(b)

Figure S4: 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, as applied to pooled samples. (b) Plots of the exact allele frequency trajectories (red), and trajectories 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 the remaining frequencies so that they sum to unity. A consequence of (i) and (ii) are the 'steps' in the blue curves.