BMC Genomics

RESEARCH

Open Access

Influence of selection on the probability of fixation at a locus with multiple alleles

A. D. J. Overall[1] and D. Waxman[2*]

Abstract

Background Genes exist in a population in a variety of forms (alleles), as a consequence of multiple mutation events that have arisen over the course of time. In this work we consider a locus that is subject to either multiplicative or additive selection, and has n alleles, where n can take the values 2, 3, 4, $\ldots$. We focus on determining the probability of fixation of each of the n alleles. For $n = 2$ alleles, analytical results, that are 'exact', under the diffusion approximation, can be found for the fixation probability. However generally there are no equally exact results for $n \geq 3$ alleles. In the absence of such exact results, we proceed by finding results for the fixation probability, under the diffusion approximation, as a power series in scaled strengths of selection such as $R_{i,j} = 2N_e(s_i - s_j)$, where N_e is the effective population size, while s_i and s_j are the selection coefficients associated with alleles i and j, respectively.

Results We determined the fixation probability when all terms up to second order in the $R_{i,j}$ are kept. The truncation of the power series requires that the $R_{i,j}$ cannot be indefinitely large. For magnitudes of the $R_{i,j}$ up to a value of approximately 1, numerical evidence suggests that the results work well. Additionally, results given for the particular case of $n = 3$ alleles illustrate a general feature that holds for $n \geq 3$ alleles, that the fixation probability of a particular allele depends on that allele's initial frequency, but generally, this fixation probability also depends on the initial frequencies of *other* alleles at the locus, as well as their selective effects.

Conclusions We have analytically exposed the leading way the probability of fixation, at a locus with multiple alleles, is affected by selection. This result may offer important insights into CDCV traits that have extreme phenotypic variance due to numerous, low-penetrance susceptibility alleles.

Keywords Random genetic drift, Selective effect, Diffusion analysis, Nearly neutral alleles, Stochastic population dynamics

Introduction

Standard population genetics theory often treats loci as biallelic [1]. Variants of a gene are then often described as 'wild-type' and 'mutant' or 'major allele' and 'minor allele'. For the most part, this treatment suffices for exploring the roles of natural selection and genetic drift (e.g., [2]) and is often assumed in data analysis [3]. However, in general, genes exist in populations in a variety of forms, as a consequence of multiple mutation events that have arisen over the course of the gene's evolution, but with the discovery that tri-allelic single nucleotide polymorphisms are far more prevalent than previously believed [4], additional mechanisms may also be occurring.

The applicability of 'binning' of multiple alleles into two groups, e.g., with all deleterious alleles placed in one group, depends on context. Take, for example, the single-gene disorder cystic fibrosis. The gene responsible for the coding of the transmembrane regulator

*Correspondence:
D. Waxman
davidwaxman@fudan.edu.cn
[1] School of Applied Sciences, Huxley Building, University of Brighton, BN2 4GJ Brighton, East Sussex, UK
[2] Centre for Computational Systems Biology, ISTBI, Fudan University, 220 Handan Road, 200433 Shanghai, People's Republic of China

protein is known to have in excess of 700 mutations. Despite this, cystic fibrosis (CF) is typically described by having a wild-type allele and recessive lethal allele. Of the many hundreds of known deleterious mutations in the CFTR gene, 70% of CF cases are due to the presence of just one of these, the $\Delta F508$ mutation [5]. However, alternative alleles are brought into the story when the focus alters to, for example, compound heterozygosity [6]. Cystic fibrosis, is, however, unrepresentative of most genetic disorders that afflict human populations. Most detrimental genetic conditions fall in to either the common disease - common variants (CDCV) or common disease - rare variants (CDRV) models of complex genetic disorders, which are typically polygenic and possibly multifactorial. Although there may be large phenotypic differences that result from genotypic variation (e.g., disease vs no disease), selection is expected to play out between lots of alleles, with each contributing only small phenotypic effects, leading to small changes in allele frequencies over time [7]. There is growing evidence that common disorders, such as schizophrenia, tend to fall into the CDRV model, with multiple mutations at a single gene [8] and where there is some evidence of selective sweeps [9]. Either way, there could be numerous, potentially distinct, selection coefficients pertaining to each genotype, producing a much more complex dynamic than that described by binning alleles.

In this work we address the nature of the multiple allele dynamics by considering the probability of fixation at a locus where there are generally more than two alleles segregating. In particular, our objective is to derive an analytical approximation to the probability of fixation for each of the n alleles at the locus.

In almost any situation, other than neutrality of the locus [10], having more than two alleles makes the problem significantly harder to analyse. Even numerical treatments of exact models, for modestly sized populations, can become overwhelmed by the size of the matrices that are required [11].

In order to make some progress, we have restricted considerations to basic problems of biological interest that involve a locus with multiple alleles, where selection acts. We have taken the population under consideration to be diploid and sexual, with reproduction occurring by random mating. We assume there is an equal sex ratio and no sexual dimorphism. Selection acts at a single unlinked locus, at which there are n alleles, with no restriction on the number of alleles (thus n can take the values 2, 3, 4, $\cdots$). Selection is assumed to be either *multiplicative* or *additive* in character, but weak in the sense that selection coefficients have magnitudes that are small; with s a typical selection coefficient, we assume $|s| \ll 1$.

For the diploid locus under selection, the assumptions just made about selection correspond to the effective *absence of dominance*. This allows the diploid results to be converted to results for a haploid population (with multiple types/alleles), by simply halving the effective population size in the diploid results.

The occurrence of selection coefficients of small magnitude seems to be very common [12]. However, there are some important cases where selection coefficients do not have small magnitudes. For example, Alzheimer's disease has the common $\epsilon 4$ allele of the APOE (apolipoprotein E) which is estimated to incur a 3 to 4-fold increase in the risk of developing late-onset Alzheimer's disease [13]. However, it is this allele's influence on vitamin D levels in *early* life that mark it out as an example of a CDCV [14] with a potentially high (positive) selection coefficient. In light of this, our results are currently of relevance to a class of genetic traits with multiple small-effect alleles. Despite the assumption of $|s| \ll 1$, there may still be some value in exploring the dynamics of multi-allelic genes that also contain a lethal mutation ($|s| \approx 1$) [15]. For example, lethal alleles tend to exist in equilibrium at the frequency of the order of the mutation rate (i.e., very low), and arguably play a minor role at the scale of the population. The relative influence on fitness of the other non-lethal alleles, albeit with much smaller selection coefficients may then have an important bearing on the evolution of the gene's diversity.

Methods

The fixation probability, in a multi-allelic context with selection, has previously been explored via a numerically exact Wright-Fisher approach [11], and at first-order in selection via a coalescent approach [16]. Here we work under the *diffusion approximation* (see e.g., [17] or [18]) where both frequencies and time are approximated as quantities that take continuous values. In what follows, we take it to be understood that any 'exact' results we give are exact *within the framework of the diffusion approximation*, and that any 'approximate' results we give are approximations *relative* to results of the diffusion approximation.

For $n = 2$ alleles, the form of selection adopted allows analytical results that are 'exact' under the diffusion approximation. However, we do not have equally 'exact' results for $n \geq 3$ alleles except in a rather special case, which we shall present in this work. We will use this special case as a testing ground of the results we derive. In the absence of exact results, we proceed by finding approximate results. We determine the fixation probabilities of the different alleles in the form of a power series in *scaled strengths of selection*, such as $R_{i,j} = 2N_e(s_i - s_j)$, where N_e is the effective population size, while s_i and s_j

are the selection coefficients associated with alleles i and j, respectively. We determine the form of the fixation probabilities, when keeping all terms to second order in the $R_{i,j}$. Higher order approximations can be obtained, if required, with more calculations along the lines presented in this work.

Of course, truncating a series in the scaled strength of selection implies that the $R_{i,j}$ cannot be indefinitely large. Analytical evidence that we shall later present, leads us to conservatively suggest that the approximate results presented here hold reasonably for magnitudes of the $R_{i,j}$ up to a value of approximately 1. This means the set of $R_{i,j}$ values that are accessible by the results presented in this work cover a range of scaled selection strengths, allowing an allele to be deleterious ($R_{i,j} \sim -1$), neutral ($R_{i,j} \sim 0$), or beneficial ($R_{i,j} \sim 1$) *relative* to other alleles at the locus.

Another way of thinking about restrictions on the values of the $R_{i,j}$ is that the results we present can apply in very large populations, *provided* selection coefficients are very similar to each other, in particular, only differing by an amount of order $1/N_e$ from each other.

Dynamics

To determine fixation probabilities we need to consider the detailed dynamics of the population.

We consider a locus with n alleles, labelled $i = 1, 2, \ldots, n$. We use $X_i(t)$ to denote the frequency of allele i at time t, and collect the frequencies of all alleles into a column vector, $\mathbf{X}(t)$, that has n components.

The set of allele frequencies, contained in $\mathbf{X}(t)$, randomly changes over time, but always maintains a sum of unity, as befits a set of proportions. We denote a possible value of $\mathbf{X}(t)$ by lower case bold symbols, such as $\mathbf{x}$.

Under the diffusion approximation, the set of allele frequencies, $\mathbf{X}(t)$, obeys the following equation, which determines how the frequencies change[1] from time t to time $t + dt$ due to selection and drift:

$$d\mathbf{X}(t) = \mathbf{F}(\mathbf{X}(t))dt + \mathbf{M}(\mathbf{X}(t))d\mathbf{W}(t). \tag{1}$$

Here we briefly summarise the key quantities in this equation, with a full description given in Sect. 1.1 of the Supplementary Material.

1. The quantity $\mathbf{F}(\mathbf{x})$ is a column vector representing the effect of selection on the n allele frequencies, when their values are given by $\mathbf{x}$. The components of $\mathbf{F}(\mathbf{x})$ are written as $F_i(\mathbf{x})$, with $i = 1, 2, ..., n$. We assume selection is *multiplicative* (genic) or *additive*, where the genotype containing alleles i and j has a fitness proportional to $(1 + s_i)(1 + s_j)$ or $(1 + s_i + s_j)$, respectively, and we refer to the s_i as *selection coefficients*. We assume all selection coefficients are small ($|s_i| \ll 1$), in which case multiplicative and additive selection are very similar, and to leading order in the s_i, both schemes of selection lead to
$$F_i(\mathbf{x}) = x_i\left(s_i - \sum\nolimits_{j=1}^{n} s_j x_j\right). \tag{2}$$
2. The quantity $\mathbf{M}(\mathbf{x})$ in Eq. (1) is an $n \times n$ matrix that contains properties of random genetic drift. It depends on allele frequencies and the effective population size, N_e. The matrix $\mathbf{M}(\mathbf{x})$ multiplies the quantity $d\mathbf{W}(t)$, which is a vector containing n independent random functions of time, that capture randomness of the action of genetic drift on allele frequencies.
3. Taking the initial time to be zero ($t = 0$), we take the initial frequency of allele i to be y_i thus
$$X_i(0) = y_i. \tag{3}$$
We collect all of the y_i into a column vector that we write as $\mathbf{y}$.
4. We use $P_i(\mathbf{y})$ to denote the probability that allele i ultimately achieves fixation, when the set of all initial frequencies are given by $\mathbf{y}$. Generally, the fixation probability of allele i, namely $P_i(\mathbf{y})$, depends on the initial frequencies of *all* alleles, and not just on the initial frequency of allele i. Noting that since fixation of one allele drives the frequencies of all other alleles to zero, fixation of different alleles are mutually exclusive events. Furthermore, fixation of one of the alleles is the only possible outcome at large t. This leads to the set of n fixation probabilities constituting a *probability distribution* that follows, at large times, from the averaged solution of Eq. (1).
5. The fixation probability $P_i(\mathbf{y})$ depends on N_e and the s_j, but only in the combination $N_e(s_i - s_j)$. We define scaled selection coefficients by
$$R_j = 2N_e s_j \tag{4}$$
and write their differences as
$$R_{j,k} = R_j - R_k \equiv 2N_e(s_j - s_k) \tag{5}$$
which we often call *scaled strengths of selection*.

Results, based on an expansion of the $P_i(\mathbf{y})$

We shall now present the results of this work, for the probabilities of fixation of the n different alleles. The results are expressed as a power series in scaled

[1] We have formulated and analysed the problem in terms of random frequencies. Equivalently, we could have worked in terms of the distribution of frequencies. We would then have been dealing with the *diffusion equation* that the distribution obeys.

strengths of selection (the $R_{j,k}$). Such an approximation applies for an arbitrary number of alleles, thus n can take the values $2, 3, \ldots$. By its very nature, an approximation based on a truncated expansion is expected to work well for a limited range of the $R_{j,k}$. For reasons we explain later, based on analytical results, we take the magnitudes of the $R_{i,j}$ to be limited to a value of approximately 1 (i.e., $|R_{j,k}| \lesssim 1$). Within this constraint on the $R_{j,k}$, there can be an arbitrary pattern of values of the selection coefficients, $s_1, s_2, \ldots, s_n$.

The approximate form of $P_i(\mathbf{y})$ follows from the diffusion approximation, and originates in Eq. (1). We note that if $P_i(\mathbf{y})$ is approximated by $P_i(\mathbf{y}) \simeq c$, where c is independent of the $R_{j,k}$ (but c depends on the initial frequencies, $\mathbf{y}$), then we say the approximation is *zeroth order* in the $R_{j,k}$, while if the approximation takes the form $P_i(\mathbf{y}) \simeq c + \sum_{a,b=1}^{n} c_{a,b} R_{a,b}$ where the coefficients c and $c_{a,b}$ are independent of the $R_{j,k}$, then we say the approximation is *first order* in the $R_{j,k}$ etc.

Here, we present the *second order approximation* of the fixation probability of allele i ($i = 1, 2, \ldots, n$). The approximation depends on the scaled strengths of selection (the $R_{j,k}$) in the form of the quantity

$$Q_i \equiv Q_i(\mathbf{y}) = \sum_{j=1}^{n} R_{i,j} y_j. \tag{6}$$

We find that the approximation of the fixation probability of allele i is given by

$$P_i(\mathbf{y}) \simeq y_i \times \left[1 + Q_i + \frac{1}{3} \left(Q_i^2 - \sum_{j=1}^{n} y_j Q_j^2 \right) \right], \quad i = 2, 3, \ldots, n \tag{7}$$

- see Sect. 1.2 of the Supplementary Material for details.

The form of $P_i(\mathbf{y})$ in Eq. (7) is the complete second order approximation to the fixation probabilities, and contains all quadratic terms in the scaled strengths of selection.

Since the Q's are linear in the scaled strengths of selection, it follows from Eq. (7) that the *zeroth order* approximation to the fixation probability contains no Q terms and is given by $P_i(\mathbf{y}) = y_i$. This is the *neutral* result, where the probability of fixation of allele i, for $i = 1, 2, \ldots, n$, is simply its initial frequency, y_i. The first order approximation of the fixation probability contains terms in Q^0 and Q^1 and is given by $P_i(\mathbf{y}) \simeq y_i \times (1 + Q_i)$, while the second order approximation is directly given in Eq. (7), and involves terms up to Q^2.

The question of the range of the $R_{j,k}$, over which the approximation in Eq. (7) reasonably holds, is addressed in Sect. 1.3 of the Supplementary Material. While we do not have a definitive answer, consideration of special cases indicate that if the magnitude of the $R_{j,k}$ are too large then the approximate fixation probability does not exhibit appropriate dependence on the initial frequencies. These considerations lead us to conservatively suggest that the approximation in Eq. (7) will reliably hold for magnitudes of the $R_{i,j}$ up to approximately 1.

Comparison with two exact results

We now consider two cases of the fixation probability, for multiplicative or additive selection, that are 'exact' within the framework of the diffusion approximation. These two cases allow some understanding (and testing) of the relation between the series approximation of the fixation probability, that is derived in this work (Eq. (7)), and the 'exact' fixation probability.

Two alleles

The first exact result for the fixation probability that we consider was obtained by Kimura, for $n = 2$ alleles [19]. We write Kimura's probability of fixation of allele 1 as

$$P_{1,K}(\mathbf{y}) = \frac{1 - e^{-2R_{1,2} y_1}}{1 - e^{-2R_{1,2}}}, \qquad n = 2. \tag{8}$$

Note that the way we have defined selection coefficients, so that e.g., under additive selection, the genotype containing alleles i and j has a fitness proportional to $(1 + s_i + s_j)$, it follows that $P_{1,K}(\mathbf{y})$ in Eq. (8) depends on the *difference* of the scaled selection coefficients of allele 1 and allele 2, namely $R_{1,2} \equiv 2N_e(s_1 - s_2)$, along with the initial frequency of allele 1, namely y_1.

To have a useful comparison with the results we establish, we expand the result for $P_{1,K}(\mathbf{y})$, in Eq. (8), up to and including terms that are second order in $R_{1,2}$, with the result

$$P_{1,K}(\mathbf{y}) \simeq y_1 \times \left[1 + (1 - y_1) R_{1,2} + \frac{1}{3} \left(1 - 3y_1 + 2y_1^2 \right) R_{1,2}^2 \right]. \tag{9}$$

We point out that this approximation works very reasonably, for a range of $R_{1,2}$. For example, for $y_1 = 1/200$, the above approximation is illustrated in Fig. 1 for $R_{1,2}$ lying in the range −1 to 1.

A noticeable aspect of Fig. 1 is that the plot of $P_{1,K}(\mathbf{y})$ against $R_{1,2}$ exhibits an appreciable level of *curvature*, and hence differs from a straight line. The first order approximation of Kimura's result (in Eq. (8)) is $P_{1,K}(\mathbf{y}) \simeq y_1 \times \left[1 + (1 - y_1) R_{1,2} \right]$. For a given set of initial frequencies (i.e., holding $\mathbf{y}$ constant), the plot of this approximation, against $R_{1,2}$, is a straight line, as shown in Fig. 1, and a first order approximation cannot capture the observed curvature. By contrast, a

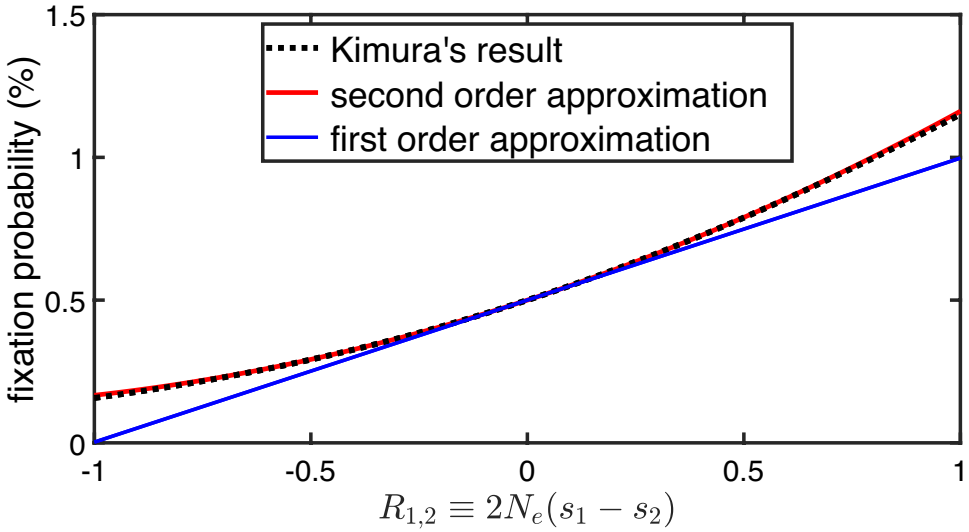

Fig. 1 Fixation probability, and its approximations, at a locus with $n = 2$ alleles. For the figure, the initial frequencies of the two alleles at the locus are held constant at the values $y_1 = 1/200$ and $y_2 = 1 - 1/200$. Kimura's diffusion approximation for the probability of fixation of allele 1 (Eq. (8)) is plotted as a function of the scaled strength of selection, $R_{1,2} \equiv 2N_e(s_1 - s_2)$. This plot is accompanied by plots of 'series approximations' of Kimura's result, one of which is first order in $R_{1,2}$, while the other is second order in $R_{1,2}$ (see Eq. (9))

second order approximation can, for a range of selection strengths, very reasonably capture the curvature, as shown in Fig. 1. This more generally motivates a second order approximation of the fixation probability.

We have also determined a third order approximation. In particular, the third order *term* that could be added to Eq. (9) is given by $-y_1^2(1-y_1)^2R_{1,2}^3/3$. This third order term, divided by the corresponding second order term, has a magnitude that to leading order in y_1 is given by $|R_{1,2}|y_1$. The third order term will thus make only a very small *additional* contribution in Eq. (9) if $|R_{1,2}|$ is in the vicinity of 1 and y_1 is very small compared with 1.

Comparison

For $n = 2$ alleles, we can determine the form of the approximate fixation probability given in Eq. (7), as follows.

The Q_i are given by $Q_1 = R_{1,2}y_2 \equiv R_{1,2}(1-y_1)$ and $Q_2 = -R_{1,2}y_1$. Then $Q_1^2 - \sum_{j=1}^2 y_j Q_j^2 = R_{1,2}^2(1-y_1)(1-2y_1)$. Using these results in Eq. (7) leads to an identical result to Eq. (9), which was obtained by expanding Kimura's result to second order in $R_{1,2}$.

Two selection coefficients, multiple alleles

The above two-allele case is useful to test and illustrate matters, but this work is generally about fixation when there are multiple (> 2) alleles. We now consider such a multiple allele case, and to allow a clear test/illustration of the result for the fixation probability (in Eq. (7)), we consider a problem that has an *exact solution* under the diffusion approximation. For definiteness, we shall restrict our considerations to additive selection.

We note that while Eq. (7) holds for n alleles, which generally have different additive selection coefficients, our requirement of an exact solution suggests the following simplified scenario. The n alleles (n arbitrary), at a locus subject to weak additive selection, are divided into two groups that we call Group 1 and Group 2. All alleles in Group 1 have the common additive selection coefficient S_1, while all alleles in Group 2 have the common additive selection coefficient S_2. This means an individual containing *any* two Group 1 alleles has fitness proportional to $1 + 2S_1$, while an individual containing *any* Group 1 allele and *any* Group 2 allele has a fitness proportional to $1 + S_1 + S_2$, and an individual containing *any* two Group 2 alleles has fitness proportional to $1 + 2S_2$. We take there to be m alleles in Group 1, labelled $1, 2, \ldots, m$, and $n - m$ alleles in Group 2, labelled $m + 1, m + 2, \ldots, n$. Thus the corresponding additive selection coefficients of the different alleles are $s_1 = s_2 = \ldots s_m = S_1$, and $s_{m+1} = s_{m+2} = \ldots s_n = S_2$.

All of the alleles in a group can be thought of as jointly constituting a 'collective allele', whose frequency is the sum of the frequencies of all alleles in that group. The probability of fixation of any of the n alleles at the locus can be obtained from the *product* of *two* probabilities: (i) the probability of fixation of the *entire group* containing the allele, and (ii) the probability of fixation of the allele, *within* this group, where all alleles within the group can

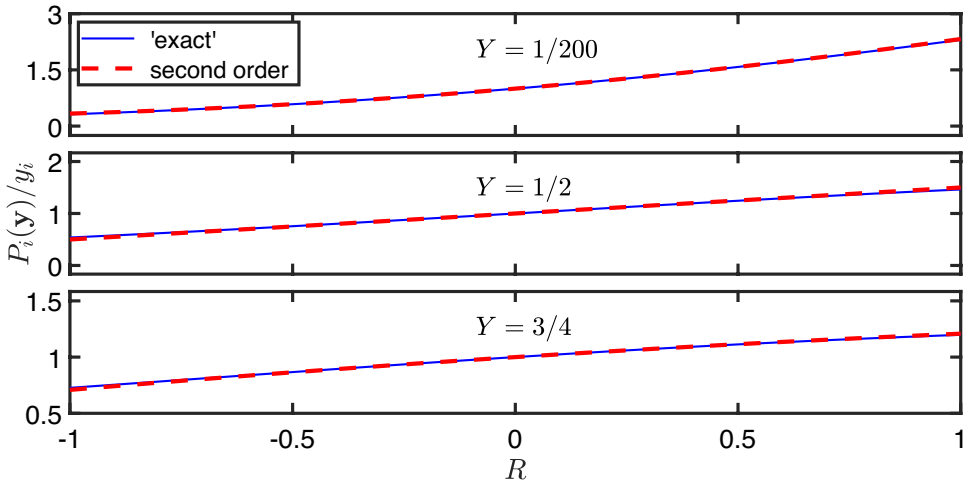

Fig. 2 Comparing results that apply for $n > 2$ alleles. We compare the quantity $P_i(\mathbf{y})/y_i$ whose 'exact' form under the diffusion approximation follows from Eq. (12), and whose second order approximation follows from Eq. (13). The ratio $P_i(\mathbf{y})/y_i$ takes the same value for all alleles (i.e., is independent of i), and depends just on the 'group' frequency, Y (Eq. (10)), and the scaled strength of selection, R (Eq. (11))

be treated as being *neutral*, since they all have the same effect on the fitness of their carriers.

Proceeding along these lines, we consider the fixation probability of allele i, which lies in Group 1. Let the m alleles in Group 1 have initial frequencies $y_1, y_2, \ldots, y_m$. Then with Y the sum of these frequencies:

$$Y = \sum_{i=1}^{m} y_i \tag{10}$$

the probability of Group 1 being the group that achieves fixation, under the diffusion approximation, follows from Kimura's formula (Eq. (8)) in the form $\left(1 - e^{-2RY}\right)/\left(1 - e^{-2R}\right)$ where, as in Eq. (8), a difference of selection coefficients, here

$$R = 2N_e(S_1 - S_2) \tag{11}$$

appears in the fixation probability:

Next, purely within Group 1, the probability that allele i is the allele that fixes is given by the neutral fixation probability - applied to Group 1. This probability is the *initial relative frequency* of allele i within Group 1, i.e., y_i/Y.

The overall probability of allele i achieving fixation is then given by the product of the two probabilities just calculated:

$$P_i(\mathbf{y}) = \frac{y_i}{Y} \times \frac{1 - e^{-2RY}}{1 - e^{-2R}} \qquad i = 1, 2, \ldots, m. \tag{12}$$

A closely related formula holds for the fixation of any of the $n - m$ alleles in Group 2.

Approximating Eq. (12) to second order in R yields

$$P_i(\mathbf{y}) \simeq y_i \times \left[1 + (1 - Y)R + \frac{1}{3}\left(1 - 3Y + 2Y^2\right)R^2\right]. \tag{13}$$

In this case, the ratio of the third order term in R (not shown) to the second order term has a magnitude that to leading order in Y is given by $|R|Y$.

Comparison

We now compare the result in Eq. (13), which was derived by expansion in R of an 'exact' fixation probability in a multiple allele model with two selection coefficients, with the approximate result given in Eq. (7), which was derived in a more general context.

We again use $R = 2N_e(S_1 - S_2)$. Then for $i = 1, 2, \ldots, m$, we have $Q_i = R\sum_{k=m+1}^{n} y_k = R(1 - Y)$, and for $j = m + 1, m + 2, \ldots, n$, we have $Q_j = -R\sum_{k=1}^{m} y_k = -RY$. These results lead, for $i = 1, 2, \ldots, m$, to $Q_i^2 - \sum_{j=1}^{n} y_j Q_j^2 = (1 - Y)(1 - 2Y)R^2$. Using these results in Eq. (7) leads to an identical result to Eq. (13).

This example allows us to make a straightforward comparison of an 'exact' result for $n > 2$ alleles and its second order approximation. We observe that the results in Eqs. (12) and (13), when divided by y_i, only depend on two parameters, namely Y and R. Thus, in Fig. 2 we plot $P_i(\mathbf{y})/y_i$ against R, for several values of Y to obtain an indication of the accuracy of the second order approximation.

From Fig. 2, we see, for essentially all values of Y, that the second order approximation of the fixation probability performs very well.

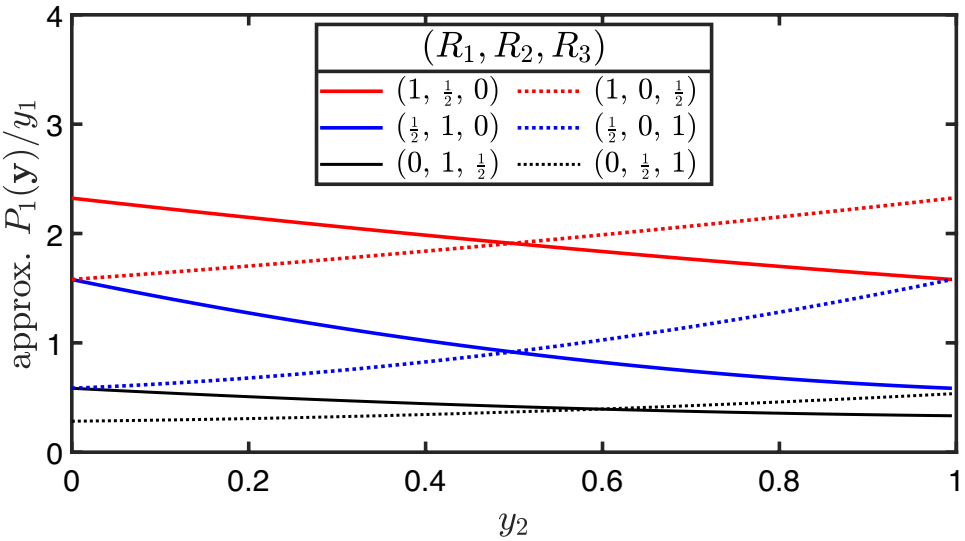

Fig. 3 The fixation probability of allele 1 at a locus with $n = 3$ alleles. For the figure, the initial frequency of allele 1 was held constant at the value $y_1 = 1/200$. The second order approximation of the fixation probability of allele 1 (Eq. (14)), relative to the neutral result, i.e., $P_1(\mathbf{y})/y_1$, is plotted as a function of the initial frequency of allele 2, namely y_2. This frequency is allowed to range from $y_2 = 0$ to $y_2 = 1 - y_1$ (the value of y_3 is set equal to $1 - y_1 - y_2$, i.e., fully determined by the value of y_2). For the purpose of illustration, we allowed the scaled selection coefficients, (R_1, R_2, R_3), to be given by all six combinations of 0, $\frac{1}{2}$ and 1, as reflected in the legend of the figure

Three alleles

We now investigate how Eq. (7), for the approximate fixation probability, applies to the case of three alleles. We generally have no analytically exact results for this case.

The scaled selection coefficients associated with the three alleles are R_1, R_2 and R_3, respectively, and initially we do not assume any relation between these.

Using Eq. (6), we readily find that $Q_1 = R_{1,2}y_2 + R_{1,3}y_3$, and using this equation, along with $y_1 = 1 - y_2 - y_3$ and $R_{2,3} = -R_{1,2} + R_{1,3}$, allows us to obtain a simple form for $Q_1^2 - \sum_{j=1}^3 y_j Q_j^2$, namely $Q_1^2 - \sum_{j=1}^3 y_j Q_j^2 = 2(R_{1,2}y_2 + R_{1,3}y_3)^2$. Equation (7) then yields the following approximation for the fixation probability of allele 1:

$$P_1(\mathbf{y}) \simeq y_1 \times \Big[1 + \big(R_{1,2}y_2 + R_{1,3}y_3\big) + \frac{2}{3}\big(R_{1,2}y_2 + R_{1,3}y_3\big)^2 - \frac{1}{3}\Big(R_{1,2}^2 y_2 + R_{1,3}^2 y_3\Big)\Big]. \tag{14}$$

The result in Eq. (14) depends on two scaled strengths of selection, namely $R_{1,2}$ and $R_{1,3}$, and these mediate the effects of the initial frequencies of alleles 2 and 3 (which are y_2 and y_3, respectively).

The presence of y_2 and y_3 in $P_1(\mathbf{y})$ signals a significant departure from the result for $n = 2$ alleles. In the $n = 2$ case, specifying the initial frequency of an allele of interest (say allele 1) fully determines the frequency of the other allele (allele 2), and hence fully determines the fixation probability. However, for $n = 3$, and higher n, specifying the initial frequency of allele 1 does not fully determine the frequencies of the other alleles. For the case of three alleles, the sum of the three initial frequencies is unity, hence specifying the value of y_1 corresponds to $y_2 + y_3$ being constrained to equal $1 - y_1$. However, this means that y_2 can lie anywhere in the range $0 \le y_2 \le 1 - y_1$. There is thus an additional level of freedom in the problem with three alleles, compared with two.

In Fig. 3 we illustrate how the additional freedom in the $n = 3$ allele problem affects the fixation probability. For the figure, we held y_1 constant, at the value 1/200, and investigated how changing y_2 affected[2] the fixation probability of allele 1, i.e., $P_1(\mathbf{y})$. To give an idea of the possible behaviours that can occur, we determined the fixation probability of allele 1 for different patterns of selection, corresponding to the scaled selection coefficients, (R_1, R_2, R_3), being given by all combinations of the values 0, $\frac{1}{2}$ and 1. That is, (R_1, R_2, R_3) was set equal to the six different sets $(1, \frac{1}{2}, 0)$, $(1, 0, \frac{1}{2})$, $(\frac{1}{2}, 1, 0)$, $(\frac{1}{2}, 0, 1)$, $(0, 1, \frac{1}{2})$ and $(0, \frac{1}{2}, 1)$. The six curves in Fig. 3 represent the fixation probabilities associated with these six different sets of scaled selection coefficients.

Figure 3 illustrates that when there are more than 2 alleles at a locus, even when the initial frequency of a

[2] For the figure, y_1 was held constant, and we allowed y_2 to vary from 0 to $1 - y_1$. Whatever value was assigned to y_2, the value of y_3 was given by $y_3 = 1 - y_1 - y_2$, i.e., it was fully determined.

focal allele is specified, the fixation probability of that allele can have a *range of possible values*. The actual value of the fixation probability generally depends on relevant selective differences, along with 'cryptic' features, namely the initial frequencies of all non-focal alleles.

While we have no analytic test of the results presented in Fig. 3, which follow from Eq. (14), we have carried out simulations of the exact Wright-Fisher model that underlies our analysis. Results given in Sect. 1.4 of the Supplementary Material, including Table 1, indicate reasonable agreement between the approximation given in Eq. (14) and simulations.

Discussion

Genes typically range in the many thousands of base pairs, presenting numerous opportunities for mutation events that plausibly make the multi-allelic state of a gene the norm. However, in many cases, interest is restricted to a single non-reference mutation of large effect and the wild-type, reference allele. Reducing a gene down to a bi-allelic entity simplifies its modelling without apparently sacrificing too much detail (e.g., [20]). However the implications of this simplification are rarely considered. This motivated the current work, of obtaining an analytical result for the probability of fixation of the n alleles of a gene, at a locus when there is selection acting, in order to obtain insight into this phenomenon.

The result we have presented for the fixation probability was derived under the diffusion approximation. The result applies to a locus where selection coefficients associated with the different alleles are small in magnitude ($|s_i| \ll 1$), as is common [12], to the extent that multiplicative and additive selection are essentially indistinguishable in their action. The result obtained indicates that the fixation probability depends only on *differences* of selection coefficients. Furthermore, the fixation probability of a specific allele depends on the initial frequencies of essentially *all* alleles, not just the initial frequency of the specific allele.

Our main result is an approximation of the fixation probability of allele i, namely $P_i(\mathbf{y})$, in the form of a power series of scaled selection strengths (scaled selection coefficient differences) (see Eq. (7)). We give the fixation probability to second order in scaled selection coefficient differences, with the result applying for any number of alleles, n. We compared the formula for the second order result with Kimura's result for the fixation probability for $n = 2$ alleles, Eq. (8), when expanded to second order in scaled selection strengths, and found full agreement at this order of expansion.

Extending our consideration to multiple alleles ($n > 2$) a number of scenarios are plausible. One such scenario is where the multiple alleles can be partitioned into two groups, where each group is associated with a distinct selection coefficient. The fixation probability of a particular allele can be calculated from the probability that its group fixes, multiplied by the neutral probability the allele fixes within its group. For this multiple ($n > 2$) allele problem, we compared the formula for the second order approximation with the expansion of the fixation probability of this 'two group' model to second order in scaled selection strengths, and again found full agreement at this order of expansion.

A simple scenario where the multiple alleles of a locus can be partitioned into two groups, each of which is associated with a distinct selection coefficient, is for a gene that has numerous mutations which *broadly* fall into one of two groups: e.g., the multiple deleterious or benign mutations within the cystic fibrosis CFTR gene [21]. However, most scenarios will involve multiple alleles each with an associated selection coefficient. To the best of our knowledge, a somewhat general *analytic* extension of Kimura's solution for the two-allele fixation probability (i.e., Eq. (8), which is derived in [1]) to more than two alleles, has yet to be found. By contrast, an exact treatment of fixation when there are multiple alleles is known for the Wright-Fisher model [11]. The fixation probability, in this reference, is given in terms of a matrix which can be very large, and while allowing numerical calculation of results for particular parameter choices, such a result does not expose, in a transparent way, dependence on parameters, unlike an analytical result.

As an example of the possible outcomes when a locus has more than two alleles, and the alleles are all associated with different selection coefficients, we have presented the approximation for the fixation probability for a locus with three alleles (see Eq. (14) and Fig. 3). This example explicitly illustrates how the probability of allele 1 fixing, for a given value of its initial frequency, y_1, is significantly influenced by the initial frequencies and selective effects of alleles 2 and 3.

As illustrated in Figs. 1 and 2, the second order approximation of the fixation probability performs very reasonably when the relevant scaled selection coefficient differences falls within the range -1 to 1. We have limited our exploration of the three allele approximation to this range. Figure 3 illustrates that the fixation probability of allele 1 can increase considerably, and non-linearly, relative to neutral expectations, as the frequency of allele 2 declines (see Fig. 3, for $(R_1, R_2, R_3) = (0.5, 1.0, 0.0)$. This result may offer important insights into traits, such as CDCV, that have extreme phenotypic variance (disease/no-disease) due to numerous, low-penetrance susceptibility alleles, for example those that have recently been identified for breast cancer [22].

Let us now consider possible future work. We reiterate that in terms of the scaled strengths of selection,

$R_{i,j} = 2N_e(s_i - s_j)$, the results obtained in this work reliably hold in a regime where the $R_{i,j}$ are not large in magnitude, in the sense $|R_{i,j}| \lesssim 1$. This is not negligibly weak selection, but it also is not strong selection, which would correspond to at least some of the $R_{i,j}$ satisfying $|R_{i,j}| \gg 1$, as can occur. We note that for the case of $n = 2$ alleles, there are existing results for strong selection that follow from Haldane's approach, based on a branching process [23], and these results can be further understood from the viewpoint of a Wright-Fisher model [24]. However, at the time of writing, we do not see how to extend these methods, nor the methods of the present work, which are based on the diffusion approximation, to the case of multiple alleles ($n > 2$) and strong selection. Useful future work would determine fixation probabilities in such a *strong selection regime*.

The mean time it takes an allele to fix, along with related temporal properties, are important aspects of the fixation process, that are often discussed at the same time as the probability of fixation. The methods we have adopted in this work are based on an analysis of trajectory statistics, and we have shown how to extract fixation probabilities from the long time values of the mean allele frequencies. Generally, frequency trajectories contain more information about the evolutionary process than just that of fixation. Existing approaches for $n = 2$ alleles have discussed this and have given *indications* of the connection between trajectory statistics and the mean time to fixation [25]. Additional useful future work would be to relate the mean time to fixation, when there are multiple alleles, to the strength of selection, possibly using the methods of the present work.

In summary, we have analytically exposed the leading way the probability of fixation, at a locus with multiple alleles, is affected by the locus not being neutral but being subject to selection.

Abbreviations

N_e	Effective population size
Eq	Equation
CDCV	Common disease, common variant

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10733-0.

Supplementary Material 1.

Acknowledgements
We thank Toni Gossmann for invaluable comments.

Authors' contributions
DW contributed to the conception and writing of the manuscript, and performed the analysis. AO contributed to the conception and writing of the manuscript. Allauthors reviewed the manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the Github repository, https://github.com/AndyOverall/Multiple-Alleles.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 April 2024 Accepted: 22 August 2024
Published online: 30 August 2024

References

1. Weinreich DM. The Foundations of Population Genetics. Cambridge: The MIT Press; 2023.
2. Santangelo JS, Johnson MTJ, Ness RW. Modern spandrels: the roles of genetic drift, gene flow and natural selection in the evolution of parallel clines. Proc R Soc B. 1878;2018(285):20180230. https://doi.org/10.1098/rspb.2018.0230.
3. Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, et al. Are minor alleles more likely to be risk alleles? BMC Med Genomics. 2018;11:3. https://doi.org/10.1186/s12920-018-0322-5.
4. Phillips C, Amigo J, Tillmar AO, Peck MA, de la Puente M, Ruiz-Ramírez J, et al. A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. Forensic Sci Int Genet. 2020;46:102232. https://doi.org/10.1016/j.fsigen.2020.102232.
5. Okay TS, Oliveira WP, Raiz-Júnior R, Rodrigues JC, Del Negro GMB. Frequency of the deltaF508 mutation in 108 cystic fibrosis patients in Sao Paulo: comparison with reported Brazilian data. Clinics. 2005;60(2). https://doi.org/10.1590/S1807-59322005000200009.
6. Strom CM, Crossley B, Redman JB, Buller A, Quan F, Peng M, et al. Cystic fibrosis screening: Lessons learned from the first 320,000 patients. Genet Med. 2004;6:136–40.
7. Yeaman S. Evolution of polygenic traits under global vs local adaptation. Genetics. 2022;220(1):iyab134. https://doi.org/10.1093/genetics/iyab134.
8. McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatr. 2007;190(3):194–9. https://doi.org/10.1192/bjp.bp.106.025585.
9. Höllinger I, Pennings PS, Hermisson J. Polygenic adaptation: From sweeps to subtle frequency shifts. PLoS Genet. 2019;15(3):e1008035. https://doi.org/10.1371/journal.pgen.1008035.
10. Baxter GJ, Blythe RA, McKane AJ. Exact solutions of the multi-allele diffusion model. Math Biosci. 2007;209:124–70.
11. Waxman D. Fixation at a locus with multiple alleles: Structure and solution of the Wright Fisher model. J Theor Biol. 2009;257:245–51.
12. Eyre-Walker AC, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8:610–8.
13. Bird TD. Genetic aspects of Alzheimer disease. Genet Med. 2008;10(4):231–9. https://doi.org/10.1097/GIM.0b013e31816b64dc.
14. Huebbe P, Nebel A, Siegert S, Moehring J, Boesch-Saadatmandi C, Most E, et al. APOE ϵ4 is associated with higher vitamin D levels in targeted replacement mice and humans. FASEB J. 2011;25(9):3262–70. https://doi.org/10.1096/fj.11-180935.
15. Waxman D, Overall ADJ. Influence of Dominance and Drift on Lethal Mutations in Human Populations. Front Genet. 2020;11. https://doi.org/10.3389/fgene.2020.00267.

16. Lessard S, Lahaie P. Fixation probability with multiple alleles and projected average allelic effect on selection. Theor Popul Biol. 2009;75(4):266–77. https://doi.org/10.1016/j.tpb.2009.01.009.
17. Kimura M. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp Quantative Biol. 1955;20:33–53.
18. Ewens WJ. Mathematical Population Genetics I. Theoretical Introduction. 2nd ed. New York: Springer-Verlag; 2004.
19. Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47:713–9.
20. De Geyter J, Gallati-Kraemer S, Zhang H, De Geyter C. Identification and selection of healthy spermatozoa in heterozygous carriers of the Phe508del-variant of the CFTR-gene in assisted reproduction. Sci Rep. 2022;12(1):1866. https://doi.org/10.1038/s41598-022-05925-1.
21. Salinas DB, Sosnay PR, Azen C, Young S, Raraigh KS, Keens TG, et al. Benign and Deleterious Cystic Fibrosis Transmembrane Conductance Regulator Mutations Identified by Sequencing in Positive Cystic Fibrosis Newborn Screen Children from California. PLoS ONE. 2016;11(5):e0155624. https://doi.org/10.1371/journal.pone.0155624.
22. Dalivandan ST, Plummer J, Gayther SA. Risks and Function of Breast Cancer Susceptibility Alleles. Cancers (Basel). 2021;13(16):3953. https://doi.org/10.3390/cancers13163953.
23. Haldane JBS. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. Math Proc Camb Philos Soc. 1927;23(7):838–44. https://doi.org/10.1017/s0305004100015644.
24. Mavreas K, Gossmann TI, Waxman D. Loss and fixation of strongly favoured new variants: Understanding and extending Haldane's result via the Wright-Fisher model. Biosystems. 2022;221:104759. https://doi.org/10.1016/j.biosystems.2022.104759.
25. Mavreas K, Waxman D. Information encoded in gene-frequency trajectories. Biosystems. 2023;231:104982. https://doi.org/10.1016/j.biosystems.2023.104982.

Publisher's Note