Fixation at a locus with multiple alleles: Structure and solution of the Wright Fisher model

D. Waxman

www.elsevier.com/locate/yjtbi

Fixation at a locus with multiple alleles: Structure and solution of the Wright Fisher model

D. Waxman

PII: S0022-5193(08)00613-9
DOI: doi:10.1016/j.jtbi.2008.11.025
Reference: YJTBI5378

To appear in: *Journal of Theoretical Biology*

Received date: 17 May 2008
Revised date: 21 November 2008
Accepted date: 21 November 2008

Cite this article as: D. Waxman, Fixation at a locus with multiple alleles: Structure and solution of the Wright Fisher model, *Journal of Theoretical Biology* (2008), doi:10.1016/j.jtbi.2008.11.025

21st November 2008

Fixation at a locus with multiple alleles: Structure and solution of the Wright Fisher model

D. Waxman

Centre for the Study of Evolution

School of Life Sciences

University of Sussex

Brighton BN1 9QG

Sussex UK

Running Head: Multiple Allele Fixation

Correspondence to:

Prof. D. Waxman, School of Life Sciences, University of Sussex, Brighton BN1 9QG, Sussex UK.

E-mail: D. Waxman@sussex.ac.uk

Phone: +44 (0)1273 678559

Abstract

We consider the Wright Fisher model for a finite population of diploid sexual organisms where selection acts at a locus with multiple alleles. The mathematical description of a such a model requires vectors and matrices of a multidimensional nature, and hence has a considerable level of complexity. In the present work we avoid this complexity by introducing a simple mathematical transformation. This yields a description of the model in terms of ordinary vectors and ordinary matrices, thereby allowing standard linear algebra techniques to be directly employed. The new description yields a common mathematical representation of the Wright Fisher model that applies for arbitrary numbers of alleles. Within this framework, it is shown how the dynamics decomposes into component parts that are responsible for the different possible transitions of segregating and fixed populations, thereby allowing a clearer understanding of the population dynamics. This decomposition allows expressions to be directly derived for the mean time of fixation, the mean time of segregation (i.e., the sojourn time) and the probability of fixation. Numerical methods are discussed for the evaluation of these quantities.

Key words: random genetic drift; time of fixation; probability of fixation; multiple alleles; sojourn time; theoretical population genetics

1 Introduction

Genetic drift is a stochastic process that occurs in finite populations and causes gene frequencies to undergo a form of random walk. The Wright Fisher model (Fisher 1922, Wright 1931) provides a conceptually simple and straightforward approach to the calculation of the effects of genetic drift. Beyond this, it provides a solid foundation from which we can compare approximations or other approaches to genetic drift. In the present work we consider the explicit formulation of the Wright Fisher model when there are more than two alleles at a locus of interest. The genetic drift of *multiple alleles* is of direct relevance to population genetics and evolution (see e.g., Bazykin *et al* 2004) as well as having applications in other areas such as language change (Baxter *et al* 2006).

In the present work, we focus, virtually completely, on the fixation of alleles within populations; a process which underlies evolutionary adaptation. To this end, we include selection in the dynamics of a population, but exclude processes which prevent fixation, that is to say migration and mutation.

Even with the exclusion of migration and mutation, the genetic drift of multiple alleles still has a substantial degree of complexity. For example, if two alleles at a locus are segregating in a population, and a new allele arises at the locus (by mutation), then ignoring further mutations, the final outcome is the result of the interplay of drift and selection. The three segregating alleles at the locus may interfere with one another in potentially complicated ways, because of the deterministic and stochastic processes occurring, and a variety of outcomes are generally possible.

In addition to a complexity of dynamics, multiple alleles lead to problems com-

putational complexity. For example, when the number of different types of alleles, or the number of adults maintained in a population are increased, there is a substantial combinatoric increase in the size of the problem. This motivates approximations, such as a diffusion approach (not considered here), where the combinatorially inflated size of the full model can be replaced by the possibly more tractable solution of a diffusion equation.

The present work investigates the formulation and solution of multiple allele drift problems. We note that Wright Fisher models, for different numbers of alleles, require different mathematical descriptions, such as requiring transition matrices of different dimensions. The investigation carried out here, amongst other things, shows what key features of multiple allele drift problems are common, and transcend the precise number of alleles at a locus.

2 The Wright Fisher model with multiple alleles

Consider a single locus of a panmictic diploid population with K alleles that are labelled A_1, A_2,..., A_K. The population evolves in discrete generations. In the adults of generation t $(= 0, 1, 2, 3, ...)$, the proportion of all alleles that are allele A_i is written $X_i(t)$. This is the relative frequency (henceforth termed frequency) of allele A_i in adults.

In an effectively infinite population, the frequencies of all alleles in adults change deterministically, according to the equation

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

where boldface quantities, such as $\mathbf{X}(t)$, denote K component vectors. Thus in Eq. (1), $\mathbf{X}(t) = (X_1(t), X_2(t), ..., X_K(t))$, $\mathbf{D}(\mathbf{x}) = (D_1(\mathbf{x}), D_2(\mathbf{x}), ..., D_K(\mathbf{x}))$ and the $D_i(\mathbf{x})$'s generally incorporate processes that change allele frequencies, that is to say mutation, migration and selection. In the present work we focus attention on properties of fixation, and hence concentrate on the case where the $D_i(\mathbf{x})$'s just incorporate effects of selection.

The deterministic change of allele frequencies in adults, given in Eq. (1), does not assume selection acts additively (i.e., without dominance), as is shown in the derivation of this equation in Appendix A. However, the phenomena of interest in the current work are the *stochastic* changes of allele frequencies which occur when population size is finite, and here the precise nature of selection can be significant, as we discuss below.

For definiteness, we adopt the following lifecycle for the processes taking place in

one generation.

Generation t	adults	
	$\downarrow$	gamete production
	gametes	
	$\downarrow$	random mating, followed
		by the death of all adults
	zygotes	
	$\downarrow$	viability selection
	juveniles	
	$\downarrow$	number regulation
Generation $t+1$	adults.	

Within the lifecycle we assume that a very large number zygotes of all genotypes are produced, so that viability selection is essentially deterministic in character. The juveniles surviving viability selection undergo a non-selective process of ecological thinning. This leads to an adult population containing N individuals. Thinning effectively corresponds to randomly picking a set of N individuals, without replacement, to survive and constitute the set of adults that proceed to reproduce. It results in random variation in the number of offspring produced by the individuals surviving selection and the frequencies of different alleles in adults, $X_j(t)$, becoming random variables that can take the values $0/(2N), 1/(2N), .., 2N/(2N)$.

The random sampling without replacement, that is associated with thinning, results in the frequencies of *adult genotypes* having a multivariate hypergeometric

distribution (for properties of this distribution, see e.g. Freund et. al. 1999). We assume the population size of adults, N, is a small fraction of the number of individuals present immediately after viability selection. Given this, the multivariate hypergeometric distribution of adult genotypes can be well approximated by a multinomial distribution in much the same way that a hypergeometric distribution can be approximated by a binomial distribution (Haigh 2002).

The distribution of *allele* frequencies in adults is directly determined from the distribution of adult genotypes, which is approximately multinomial. However, when thinning occurs in diploid individuals, it does not automatically follow that the distribution of alleles in adults will also be multinomial. This point is made by Nagylaki (1992, p. 252), where he points out that a multinomial distribution of allele frequencies is only guaranteed to follow from multiplicative viabilities, i.e., in the absence of dominance at the locus in question.

To illustrate how dominance can disrupt the occurrence of a "standard" Wright Fisher distribution of alleles in adults, namely a binomial distribution when there are $K = 2$ alleles and a multinomial distribution when there are $K > 2$ alleles, consider a locus with $K = 2$ alleles. At this locus, assume both types of homozygote have zero viability (i.e., there is complete overdominance). In this case, the only individuals surviving selection are heterozygotes, and randomly reducing the number of these to N, by non selective thinning, has no effect on allele frequencies, which are equal before and after thinning. This example illustrates how dominance can have significant effects on the distribution of allele frequencies in adults, and may invalidate the Wright Fisher model. Indeed, in this example, the effects of dominance

completely prevent the occurrence of genetic drift.

Generally, models of genetic drift, where *allele* frequencies in diploid adults have multinomial distribution, amounts to the absence or neglect of dominance at the locus under selection (no such absence of dominance is required if thinning occurs in a haploid stage). If dominance is a significant aspect of selection, then frequencies of genotypes, rather than alleles, need to be followed. While we do not do so here, the methodology introduced in the following Section may be employed in such a case, at a significant computational cost. In the present work, we assume a negligible level dominance (or the absence of dominance) so allele frequencies in adults have a multinomial distribution. As a consequence, when the frequencies in generation t are given by $\mathbf{X}(t)$, the frequencies in the following generation are given by

$$\mathbf{X}(t+1) = \frac{\mathbf{M}}{2N} \tag{2}$$

where $\mathbf{M} = (M_1, M_2, ..., M_K)$ denotes a set of random integers drawn from a multinomial distribution. The parameters describing this distribution are $2N$ and $\mathbf{X}(t) + \mathbf{D}(\mathbf{X}(t))$, which represent, respectively, the number of trials associated with a multinomial distribution and the probabilities of the K different outcomes. In this way we arrive at a variant of a Wright Fisher model (Fisher 1922, Wright 1931) with multiple alleles.

The statistical description of allele frequencies in such a Wright Fisher model arises from consideration of a very large number of replicate populations, each of

which maintains N adults each generation. Let

$$\mathbf{n} = (n_1, n_2, ..., n_K) \tag{3}$$

represent the numbers of the K different types of alleles present in adults, in a particular replicate population. As such, the elements of $\mathbf{n}$ can take all possible integral values in the range 0 to $2N$, subject to the restriction that their sum, $\sum_{j=1}^{K} n_j$, is the total number of alleles in a population, namely $2N$.

The possible values of the allele frequencies, $\mathbf{X}(t)$, are given by $\mathbf{x_n}$ where

$$\mathbf{x_n} = \left(\frac{n_1}{2N}, \frac{n_2}{2N}, ..., \frac{n_K}{2N}\right) = \frac{\mathbf{n}}{2N}. \tag{4}$$

Furthermore, if, in generation t, the set of frequencies, $\mathbf{X}(t)$, coincide with $\mathbf{x_m}$ then $2N \times \mathbf{X}(t+1)$ is a random variable with a multinomial distribution with parameters $2N$ and $\mathbf{x_m} + \mathbf{D}(\mathbf{x_m})$.

We write the probability that $\mathbf{X}(t)$ has the value $\mathbf{x_n}$ as $F_{\mathbf{n}}(t)$. The corresponding probability distribution in generation $t+1$ is given by

$$F_{\mathbf{n}}(t+1) = \sum\nolimits_{\mathbf{m}} V_{\mathbf{n},\mathbf{m}} F_{\mathbf{m}}(t) \tag{5}$$

where $V_{\mathbf{n},\mathbf{m}}$ is a transition probability, namely the probability that $\mathbf{X}(t+1)$ equals $\mathbf{x_n}$, given that $\mathbf{X}(t)$ equals $\mathbf{x_m}$. The transition probability takes the multinomial form

$$V_{\mathbf{n},\mathbf{m}} = (2N)! \prod\nolimits_{i=1}^{K} \frac{([\mathbf{x_m} + \mathbf{D}(\mathbf{x_m})]_i)^{n_i}}{n_i!} \tag{6}$$

where $[\mathbf{x_m} + \mathbf{D}(\mathbf{x_m})]_i$ denotes the i'th element of the vector $\mathbf{x_m} + \mathbf{D}(\mathbf{x_m})$.

When there are more than two types of alleles, we can think of Eq. (5) as an equation involving a multidimensional transition matrix (not a conventional, two dimensional matrix), where an element of this multidimensional matrix, namely the transition probability $V_{\mathbf{n},\mathbf{m}}$, is labelled by indices $\mathbf{n}$ and $\mathbf{m}$ that are not single numbers, but K dimensional *vectors*. There is the further complication that the vector "indices", such as $\mathbf{n}$, have components whose sum is constrained to have the value $2N$.

3 Conversion to an ordinary matrix equation

In order to extract the general structure of a multiple allele problem in a transparent and useful form, we convert Eq. (5) to an ordinary matrix equation. In doing so, the transformed equation does not have indices which are vectors, unlike the quantities in Eq. (5), but rather it has indices which are single numbers (scalars). This resulting ordinary matrix equation allows us to exploit standard matrix approaches or computational packages for matrices.

Our method of conversion involves the introduction of an indexing function $I(\mathbf{n})$ which yields a unique number – a scalar index – associated with each distinct "vector index" $\mathbf{n}$. The total number of distinct $\mathbf{n}$'s is given by $I_{\max}$ and for a population of size N, with K types of alleles at the locus in question, we have

$$I_{\max} = \frac{(2N + K - 1)!}{(2N)!(K - 1)!}. \tag{7}$$

As $\mathbf{n}$ ranges over all allowed values, the indexing function, $I(\mathbf{n})$, is explicitly constructed so that it takes the values $1, 2, .., I_{\max}$.

The ordinary matrices, to which the multiple allele problem is converted, are of size of $I_{\max} \times I_{\max}$ and as might be expected, the value of $I_{\max}$ rapidly increases with N and K. For example, in a two allele model ($K = 2$) we have the familiar result $I_{\max} = (2N + 1)$. By contrast, when $K = 3$ or $K = 4$, we find $I_{\max}$ takes the values $(2N + 2)(2N + 1)/2$ or $(2N + 3)(2N + 2)(2N + 1)/6$, which are of order N^2 or N^3, respectively.

The indexing function, $I(\mathbf{n})$, can be simply constructed by making a list of all of the distinct $\mathbf{n}$'s of the form $\mathbf{n} = (n_1, n_2, ..., n_K)$ whose elements are non negative integers that sum to $2N$. The location of a given n in the list is its index. To make the mathematical structure of multiple allele drift explicit, we shall arrange the list so that the first K members (i.e., the first K of the $\mathbf{n}$'s) are specified. In particular, for $i = 1, 2, ..., K$, the i'th vector $\mathbf{n}$ in the list corresponds to a population where where all $2N$ alleles are allele A_i – i.e., allele A_i is fixed. The remaining vectors in the list (at positions $K + 1$, $K + 2$, ..., $I_{\max}$) can have an arbitrary order. As an example, when there are $K = 3$ different alleles and the population size is $N = 1$, we consider the set of all $\mathbf{n}$'s with $n_1 + n_2 + n_3 = 2$ and can use the indexing scheme of Table 1, although other forms of $I(\mathbf{n})$ are obviously possible.

Table 1

The indexing function, $I(\mathbf{n})$, associates a unique integer in the range 1 to $I_{\max}$ with each possible value of the vector index, $\mathbf{n}$. It is possible to take the opposite point of

view, and associate a unique vector, namely a *vector index*, with each integer in the range 1 to $I_{\max}$. We write the vector unique *vector index* associated with integer i as $\mathbf{n}^{[i]}$. Using the example of Table 1 we then have that $\mathbf{n}^{[1]} = (2, 0, 0)$, $\mathbf{n}^{[2]} = (0, 2, 0)$, ..., $\mathbf{n}^{[6]} = (1, 1, 0)$. The general relations between $I(\mathbf{n})$ and $\mathbf{n}^{[i]}$ are

$$I(\mathbf{n}^{[i]}) = i \tag{8}$$

$$\mathbf{n}^{[I(\mathbf{m})]} = \mathbf{m} \tag{9}$$

and these bear a superficial resemblance to the relation between logarithms and exponentials.

We can use Eq. (9) to construct a column vector $\mathbf{f}(t)$, with $I_{\max}$ elements, that contains identical information to the distribution $F_{\mathbf{m}}(t)$. The elements of the vector $\mathbf{f}(t)$, namely $f_i(t)$, are given by

$$f_i(t) = F_{\mathbf{n}^{[i]}}(t), \qquad i = 1, 2, ..., I_{\max} \tag{10}$$

while the reverse transformation, from the $f_i(t)$ to the $F_{\mathbf{n}}(t)$, is simply

$$F_{\mathbf{n}}(t) = f_{I(\mathbf{n})}(t). \tag{11}$$

We can thus freely convert between quantities with vector and scalar indices, i.e., between the original form of the model, and the representation in terms of quantities with scalar indices.

As shown in Appendix B, the conversion of Eq. (5), to an ordinary matrix equation, leads to the appearance of an $I_{\max} \times I_{\max}$ matrix, which we write as $\mathbf{v}$, and has elements

$$v_{i,j} = V_{\mathbf{n}^{[i]},\mathbf{n}^{[j]}}, \qquad i,j = 1,2,...,I_{\max}. \tag{12}$$

This is used to write Eq. (5) in the completely equivalent form $f_j(t+1) = \sum_{k=1}^{I_{\max}} v_{j,k} f_k(t)$, which in matrix notation reads

$$\mathbf{f}(t+1) = \mathbf{v}\mathbf{f}(t). \tag{13}$$

This equation puts us in the familiar territory of an equation involving standard objects: column vectors and a two dimensional matrix, and has the solution

$$\mathbf{f}(t) = \mathbf{v}^t \mathbf{f}(0). \tag{14}$$

Working with this solution, which just involves the conventional matrix $\mathbf{v}$, means we can use standard linear algebra techniques to numerically determine all of the properties of $\mathbf{f}(t)$. We can then use Eq. (11) to return to the original, multiple-allele representation, to calculate quantities of interest – or, indeed, we can calculate such quantities directly from $\mathbf{f}(t)$.

4 Structure of the dynamics

While it is possible to directly work with Eq. (14) and utilise standard techniques to solve it, more information and insight can be obtained by writing the matrix $\mathbf{v}$

in a form that expresses the underlying structure of the problem. Given that only selection occurs in the dynamics, it follows that all alleles have the possibility of fixing, and the matrix $\mathbf{v}$ must, under the indexing scheme adopted, have the form

$$\mathbf{v} = \begin{pmatrix} \mathbf{1} & \mathbf{u} \\ \mathbf{0} & \mathbf{w} \end{pmatrix}. \tag{15}$$

Here $\mathbf{1}$, $\mathbf{u}$, $\mathbf{0}$ and $\mathbf{w}$ are all matrices, and with

$$K' = I_{\max} - K \tag{16}$$

$\mathbf{1}$ is a $K \times K$ identity matrix, $\mathbf{0}$ is an $K' \times K$ matrix of zeros while $\mathbf{u}$ and $\mathbf{w}$ are matrices of sizes $K \times K'$ and $K' \times K'$ respectively that can be directly read off from the form of $\mathbf{v}$, once this matrix has been calculated.

As we show below, the interpretation of the nontrivial matrices $\mathbf{u}$ and $\mathbf{w}$ appearing in Eq. (15) is that (i) the matrix $\mathbf{u}$ governs transitions of a population from being segregating to becoming fixed, and contains the probabilities of these processes and (ii) the matrix $\mathbf{w}$ describes pure "segregation" dynamics, i.e., it contains the probabilities of transitions between different states of a population that continues to segregate.

To establish why the form for $\mathbf{v}$ in Eq. (15) must arise from the dynamics, and the interpretation of the various block matrices in Eq. (15), we note that the vector

$\mathbf{f}(t)$ can be written as

$$\mathbf{f}(t) = \begin{pmatrix} \mathbf{f}^{\text{fixed}}(t) \\ \mathbf{f}^{\text{seg}}(t) \end{pmatrix} \tag{17}$$

where:

(i) $\mathbf{f}^{\text{fixed}}(t)$ is a column vector with K elements that give the probabilities of a population containing only a single type of allele *by* the end generation t, i.e., of being fixed.

(ii) $\mathbf{f}^{\text{seg}}(t)$ is a column vector with K' elements that give the frequencies of populations where more than one type of allele is segregating at generation t.

It follows that under the irreversible dynamics of fixation, the only processes that can occur amongst the replicate populations are (a) segregating populations in one generation contributing to both segregating and fixed populations in the following generation and (b) fixed populations in one generation merely producing fixed populations, from one generation to the next.

Schematically:

$$\begin{aligned} \text{fixed} &\rightarrow \text{fixed} \\ \\ \text{segregating} &\rightarrow \text{segregating} + \text{fixed}. \end{aligned}$$

The action of $\mathbf{v}$ of Eq. (15) on $\mathbf{f}(t)$ of Eq. (17), to produce $\mathbf{f}(t+1)$ is

$$\begin{pmatrix} \mathbf{f}^{\text{fixed}}(t+1) \\ \mathbf{f}^{\text{seg}}(t+1) \end{pmatrix} = \begin{pmatrix} \mathbf{f}^{\text{fixed}}(t) + \mathbf{u}\mathbf{f}^{\text{seg}}(t) \\ \mathbf{w}\mathbf{f}^{\text{seg}}(t) \end{pmatrix} \tag{18}$$

which exactly reflects how some probability flows from segregating populations to

fixed populations, but not vice versa.

Let us assume that all replicate populations start at time $t = 0$ with the same definite set of frequencies, $\mathbf{x_a}$. We shall restrict all further considerations to $\mathbf{x_a}$ which do not correspond to fixed populations. Thus $\mathbf{f}^{\rm fixed}(0)$ vanishes, while $\mathbf{f}^{\rm seg}(0)$ has only one non-zero element, corresponding to the set of frequencies $\mathbf{x_a}$, and $f^{\rm seg}_{I(\mathbf{a})}(0) = 1$.

For $t = 1, 2$, ..., the solution of Eq. (18) can be written in the form

$$\mathbf{f}^{\rm seg}(t) = \mathbf{w}^t \mathbf{f}^{\rm seg}(0) \tag{19}$$

$$\mathbf{f}^{\rm fixed}(t) = \mathbf{u} \sum_{r=1}^{t} \mathbf{w}^{r-1} \mathbf{f}^{\rm seg}(0). \tag{20}$$

5 Fixation probability and time to fixation

In Appendix C we provide derivations of three important quantities of the dynamics of the problem, namely (i) the probability of fixation of a given type of allele, (ii) the mean time that alleles in a population segregate and (iii) the mean time it takes a given type of allele to fix. All results are given in terms of the initial (unfixed) allele frequencies, $\mathbf{x_a}$, and the matrix

$$\mathbf{G} = \sum_{j=1}^{\infty} \mathbf{w}^{j-1} = (\mathbf{I} - \mathbf{w})^{-1} \tag{21}$$

where $\mathbf{I}$ is an identity matrix of size $K' \times K'$ (the size of $\mathbf{w}$).

In order to refer to elements of the various matrices and vectors appearing in this Section, we use the scalar index (that runs from 1 to $I_{\rm max}$) which was introduced in

Section 3. However, to avoid the need for any distracting offsets of the scalar index, we shall label the elements of the various matrices/vectors according to their actual position in Eq. (15) or Eq. (17). In particular, this means that the matrix $\mathbf{u}$ has elements $u_{i,j}$ where i runs from 1 to K, while j runs from $K+1$ to $I_{\max}$, while the matrix $\mathbf{w}$ has elements $w_{i,j}$ where both i and j run from $K+1$ to $I_{\max}$. Because the matrix $\mathbf{G}$ is constructed from the matrix $\mathbf{w}$ (via Eq. (21)), its elements, $G_{i,j}$, also have i and j running from $K+1$ to $I_{\max}$. Similarly, the elements of the vector $\mathbf{f}^{\text{fixed}}(t)$ are labelled by an index running from 1 to K, while those of $\mathbf{f}^{\text{seg}}(t)$ run from $K+1$ to $I_{\max}$. With this method of referring to the elements of matrices and vectors, we find:

(i) The fixation probability of allele A_i is

$$\Pi_i(\mathbf{x_a}) = [\mathbf{uG}]_{i,I(\mathbf{a})}\,, \qquad i = 1, 2, .., K. \tag{22}$$

(ii) The mean time that alleles are segregating in a population (the sojourn time) at frequencies $\mathbf{x_b}$, given an initial frequency of $\mathbf{x_a}$ is written $E\left[T(\mathbf{x_b})|\mathbf{x_a}\right]$ and given by

$$E\left[T(\mathbf{x_b})|\mathbf{x_a}\right] = G_{I(\mathbf{b}),I(\mathbf{a})}. \tag{23}$$

(iii) The mean time to fixation of allele A_i, given the starting frequencies $\mathbf{x_a}$ is given by

$$\sum_{t=1}^{\infty} tP(t|\{A_i \text{ fixes}\}, \mathbf{x_a}) = \frac{\left[\mathbf{uG}^2\right]_{i,I(\mathbf{a})}}{[\mathbf{uG}]_{i,I(\mathbf{a})}}, \qquad i = 1, 2, .., K. \tag{24}$$

where $P(t|\{A_i \text{ fixes}\}, \mathbf{x_a})$ is the probability that allele A_i segregates for t generations, given the population starts at the frequencies $\mathbf{x_a}$ and fixes at the end of generation

t.

6 Numerical considerations

The exactness of the formalism presented in the present work can be illustrated by comparing the numerical results it yields with exact results. In Appendix D we carry out such a comparison for the change of expected values of allele frequencies and their covariances that occur over one generation, namely $E\left[X_i(t+1)|\mathbf{X}(t)=\mathbf{x}\right]$ and $\mathrm{Cov}\left(X_i(t+1), X_j(t+1)|\mathbf{X}(t)=\mathbf{x}\right)$. The results of the Appendix indicate that to the numerical precision with which the calculations are carried out (1 part in 10^{13}), there is agreement between the two sets of results.

Beyond calculations that provide a numerical verification of the formalism presented here, there is also the computational problem of dealing with the large matrices that naturally arise in multiple allele drift problems. For example in calculating the quantities in Section 5, we need to calculate the matrix $\mathbf{G}$ (Eq. (21)), which plays a central role in the results (Eqs. (22), (23) and (24)) and which may have a substantial size.

We calculate the matrix $\mathbf{G}$ and related quantities from the transition probabilities $V_{\mathbf{n},\mathbf{m}}$ given in Eq. (6), which are converted to the matrix $\mathbf{v}$ via Eq. (12). From $\mathbf{v}$, we directly read off the matrices $\mathbf{u}$ and $\mathbf{w}$ of Eq. (15).

The actual size of the matrix $\mathbf{w}$ suggests the numerical procedure that is best adopted to determine the matrix $\mathbf{G}$. If $\mathbf{w}$ is of moderate size (with size defined by computer memory), then a numerical linear algebra package (e.g., Matlab) can

directly determine $\mathbf{G}$ from $\mathbf{w}$ via Eq. (21). If the size of the matrix $\mathbf{w}$ is large, as can easily be the case, then a straightforward way of proceeding is to rewrite the equation for $\mathbf{G}$ in the form $\mathbf{G} = \mathbf{I} + \mathbf{wG}$ and simply iterate this equation, starting with $\mathbf{G} = \mathbf{I}$. The iteration may be terminated when an iteration results in $\mathbf{G}$ changing by less than a predetermined level (e.g., 1 part in 10^{10}) and this has worked well in practice, with a matrix $\mathbf{G}$ with of order 10^6 elements being straightforwardly determined.

To give an idea of the computer time it takes to calculate e.g., times to fixation, we have considered several illustrative cases. When $K = 3$ and $N = 30$, this corresponds to $I_{\max} = 1891$ and ordinary matrices with $I_{\max} \times I_{\max} = 3,575,881$ elements. On a 2.66 GHz pc with a core 2 processor and 2 GB of RAM, the numerical linear algebra package Matlab took approximately 20 seconds to calculate results for this case (including construction of the matrix $\mathbf{v}$). The iterative method (outlined above), by contrast, is a relatively unsophisticated algorithm and achieves the same result in a much longer amount of computer time, of the order of 800 seconds. A larger population size of $N = 50$, with the same value of K corresponds to $I_{\max} = 5151$ (i.e., a matrix having $26,532,801$ elements) and requires approximately 150 seconds on Matlab. The rapid increase of $I_{\max}$ with the number of alleles, as follows from Eq. (7) means that when $K = 4$, a computer of the above specification is restricted, when using Matlab, to $N = 14$ (corresponding to a matrix with $20,205,025$ elements) and requires approximately 120 seconds of computation time.

7 Summary

In this work we have considered the Wright Fisher model for a locus that is under selection in a diploid sexual organism. We have assumed that there are more than two alleles at this locus. The statistical description of a Wright Fisher model with multiple alleles is, in its original formulation, given in terms of a vector. The elements of this vector are the probabilities of occurrence of different sets of allele frequencies in a population, and the indices of this vector are themselves vectors (giving the numbers of different types of alleles). This is thus a multidimensional representation. The dynamical behaviour of such a model requires a transition matrix that is also multidimensional in character. In the present work this complicated description is converted, by the introduction of a simple mathematical transformation, from multidimensional matrices and vectors into *ordinary* matrices and vectors. This has the immediate advantage that standard linear algebra packages may be used to numerically determine the dynamics. Beyond this, however, we have, by a particular requirement on the mathematical transformation, established a common form for the general structure of multiple allele Wright Fisher models that holds, irrespective of the number of alleles. This allows the transition matrix to be decomposed into fundamental blocks, that, in the language of replicate populations, are responsible for transitions from segregating to fixed populations, or describe populations that continue to segregate, or describe fixed populations that remain fixed. This common form for the structure of Wright Fisher models has allowed general expressions to be determined for key quantities, such as the mean time of segregation (which is also referred to as the sojourn time), the mean time of fixation and the probability of

fixation, which apply, irrespective of the number of alleles.

It is readily apparent, and indeed demonstrated in the present work, that the intrinsic size of the matrix and vectors needed to describe dynamics of a Wright Fisher model, rapidly grow with the number of alleles and the number of adults in a population. Using the mathematical machinery introduced here we can, for moderate population sizes and moderate numbers of alleles, take a direct, essentially exact, numerical approach in calculations. This can allow an effectively exact numerical exploration of different selection schemes.

Acknowledgements

I would like to thank the two anonymous reviewers for very constructive comments that have significantly improved this paper.

Table 1

Index, $I(\mathbf{n})$	n_1	n_2	n_3
1	2	0	0
2	0	2	0
3	0	0	2
4	0	1	1
5	1	0	1
6	1	1	0

Table 1 Caption

Table 1 contains an illustration of how the indexing function $I(\mathbf{n})$ is constructed when there are $K = 3$ alleles and the population size is $N = 1$. We list all distinct vector indices, $\mathbf{n} = (n_1, n_2, n_3)$, that represent the numbers of the $K = 3$ different types of alleles there are in adults in a population. Elements of $\mathbf{n}$ are non negative integers that satisfy $n_1 + n_2 + n_3 = 2N = 2$. The location of given $\mathbf{n}$ in the list is its index. We have specifically organised the list so that when the index j takes one of the first K values (in this example, $j = 1, 2, 3$), the $\mathbf{n}$ obtained corresponds to a population fixed for allele A_j. In Table 1 this means an index of value 1, 2 or 3 corresponds to a population that is fixed for allele A_1, A_2 or A_3 respectively. The overall length of the list is given by $I_{\max}$ of Eq. (7), and for $K = 3$ and $N = 1$, this yields $I_{\max} = 6$, as Table 1 illustrates.

Appendix A

In this Appendix, we derive Eq. (1) for the deterministic change of allele frequencies, when population size is effectively infinite. For generality, we incorporate mutation as well as viability selection. We shall make use of the Kronecker delta, $\delta_{a,b}$, which takes the value of unity when a coincides with b and vanishes otherwise.

Following Nagylaki (1992), we assume that at the start of generation t, the A_iA_j unordered genotype (where $i \leq j$) of adults has frequency $X_{i,j}(t)$ with $\sum_{i,j=1(i\leq j)}^{K} X_{i,j}(t) = 1$. Then the frequency of allele A_i in adults is $X_i(t) = X_{i,i}(t) + \frac{1}{2}\sum_{j(j>i)} X_{i,j}(t) + \frac{1}{2}\sum_{j\ (j<i)} X_{j,i}(t)$. The adults produce gametes. Taking $R_{j,i}$ as the mutation rate from allele A_i to allele A_j (with $\sum_{j=1}^{K} R_{j,i} = 1$), the frequency of A_k bearing gametes is $X_k^*(t) = \sum_{i=1}^{K} R_{k,i}X_i(t)$. Thus the frequency of A_iA_j unordered genotypes, after random union of gametes and selection is

$$X_{i,j}^{**}(t) = (2-\delta_{i,j})V_{i,j}X_i^*(t)X_j^*(t) \Big/ \sum_{k,l(k\leq l)} (2-\delta_{k,l})V_{k,l}X_k^*(t)X_l^*(t) \qquad (25)$$

where $V_{i,j}$ is the viability of A_iA_j genotypes. The corresponding frequency of A_i in adults in generation $t+1$, is $X_i(t+1) = X_{i,i}^{**}(t) + \frac{1}{2}\sum_{j\ (j>i)} X_{i,j}^{**}(t) + \frac{1}{2}\sum_{j\ (j<i)} X_{j,i}^{**}(t)$ and, using Eq. (25), this can be explicitly expressed in terms of the frequency of alleles in adults in the previous generation, $X_i(t)$. We can thus generally write $X_i(t+1) = X_i(t) + D_i(\mathbf{X}(t))$.

Appendix B

In this Appendix, we derive the form of the dynamical equation for the distribution of the numbers of alleles of different types present in the population, in terms of quantities with scalar indices.

As in the main text, the index i runs from 1 to the value $I_{\max}$ given in Eq. (7) and the quantity $\mathbf{n}^{[i]}$ is the i'th *vector* index.

Using Eqs. (10) and (11), we can rewrite Eq. (5) as

$$f_i(t+1) = \sum\nolimits_{\mathbf{m}} V_{\mathbf{n}^{[i]},\mathbf{m}} f_{I(\mathbf{m})}(t). \tag{26}$$

The sum over $\mathbf{m}$ covers all possible the vector indices for the relevant values of K and N. The sum can be replaced by $\sum_{j=1}^{I_{\max}} V_{\mathbf{n}^{[i]},\mathbf{n}^{[j]}} f_{I(\mathbf{n}^{[j]})}(t)$ since when j runs from 1 to $I_{\max}$, the quantity $\mathbf{n}^{[j]}$ covers all allowed vector indices. Using Eq. (8) to replace $I(\mathbf{n}^{[j]})$ by j then yields $f_i(t+1) = \sum_{j=1}^{I_{\max}} V_{\mathbf{n}^{[i]},\mathbf{n}^{[j]}} f_j(t)$. With the introduction of the $I_{\max} \times I_{\max}$ matrix $\mathbf{v}$ with elements $v_{i,j} = V_{\mathbf{n}^{[i]},\mathbf{n}^{[j]}}$, we arrive at $f_i(t+1) = \sum_{j=1}^{I_{\max}} v_{i,j} f_j(t)$ which is the equation given in the main text.

For completeness, we note that $V_{\mathbf{n},\mathbf{m}} = v_{I(\mathbf{n}),I(\mathbf{m})}$.

Appendix C

In this Appendix, we derive expressions for (i) the probability of fixation of a given allele type, (ii) the mean time alleles in a population segregate and (iii) the mean time it takes alleles to fix.

We make use of the Kronecker delta, $\delta_{a,b}$, which takes the value of unity, when a coincides with b and vanishes otherwise. We shall use this definition of $\delta_{a,b}$ when a and b are both scalar quantities or both vectors. We also use the convention described in Section 5, that elements of the various vectors and matrices are labelled with the scalar indices appropriate to their actual position in Eq. (15) or Eq. (17). Thus the matrix $\mathbf{u}$ has elements $u_{i,j}$ where i runs from 1 to K, while j runs from $K+1$ to $I_{\max}$; the matrix $\mathbf{w}$ has elements $w_{i,j}$ where both i and j run from $K+1$ to $I_{\max}$ and similarly for the elements the matrix $\mathbf{G}$ (because it is constructed from the matrix $\mathbf{w}$ via Eq. (21)). In the same way, the elements of the vector $\mathbf{f}^{\text{fixed}}(t)$ are labelled by an index running from 1 to K, while those of $\mathbf{f}^{\text{seg}}(t)$ run from $K+1$ to $I_{\max}$.

We begin the calculations by taking all replicate populations to start at time $t = 0$ with the same definite set of frequencies, $\mathbf{x_a}$, and we only consider $\mathbf{x_a}$ which do not correspond to fixed populations. The identical initial genetic composition of all replicate populations is expressed as $F_{\mathbf{n}}(0) = \delta_{\mathbf{n},\mathbf{a}}$ or equivalently

$$f_i^{\text{seg}}(0) = \delta_{i,I(\mathbf{a})}, \qquad i = K+1, K+2, .., I_{\max}. \tag{27}$$

Fixation probability

Taking t to tend to infinity in Eq. (20) yields

$$\mathbf{f}^{\text{fixed}}(\infty) = \mathbf{u}\mathbf{G}\mathbf{f}^{\text{seg}}(0) \tag{28}$$

where

$$\mathbf{G} = \sum_{j=1}^{\infty} \mathbf{w}^{j-1} = (\mathbf{I} - \mathbf{w})^{-1} \tag{29}$$

and $\mathbf{I}$ is an identity matrix of the same size as $\mathbf{w}$, namely $K' \times K'$.

The fixation probability of allele A_i for this case is $f_i^{\text{fixed}}(\infty)$. We write this as the conditional probability of A_i fixing, given initial frequencies of $\mathbf{x_a}$, i.e., $P(\{A_i \text{ fixes}\}|\mathbf{x_a})$ and also in the more conventional form $\Pi_i(\mathbf{x_a})$. It then follows from Eqs. (27) and (28) that

$$P(\{A_i \text{ fixes}\}|\mathbf{x_a}) = \Pi_i(\mathbf{x_a}) = [\mathbf{u}\mathbf{G}]_{i,I(\mathbf{a})}, \qquad i = 1, 2, .., K. \tag{30}$$

As a technical aside, we do not prove that $\mathbf{I} - \mathbf{w}$ is invertible, which is required for the existence of $\mathbf{G}$ (Eq. (29)), but note that this follows from all K alleles being capable of fixing, since then there will be precisely K unit eigenvalues of the matrix $\mathbf{v}$ (Eq. (12)), with all other eigenvalues smaller than unity. The form given for $\mathbf{v}$ in Eq. (15) means the $K \times K$ identity matrix $\mathbf{1}$ incorporates K unit eigenvalues of $\mathbf{v}$ and all of the remaining eigenvalues of $\mathbf{v}$ are also the eigenvalues of the matrix $\mathbf{w}$. Thus all K alleles being able to fix means all eigenvalues of $\mathbf{w}$ are smaller than unity, so $\mathbf{I} - \mathbf{w}$ is invertible and hence $\mathbf{G}$ exists.

Time of segregation (sojourn time)

The Kronecker delta, $\delta_{\mathbf{x_b},\mathbf{X}(t)}$, has the value of unity when $\mathbf{x_b} = \mathbf{X}(t)$ and vanishes otherwise. Thus the quantity

$$T(\mathbf{x_b}) = \sum_{t=0}^{\infty} \delta_{\mathbf{x_b},\mathbf{X}(t)} \tag{31}$$

is a random variable representing the total number of generations where allele frequencies equal $\mathbf{x_b}$. The mean time that alleles are segregating at frequencies $\mathbf{x_b}$, given an initial frequency of $\mathbf{x_a}$ is $E\left[T(\mathbf{x_b})|\mathbf{x_a}\right] = \sum_{t=0}^{\infty} E\left[\delta_{\mathbf{x_b},\mathbf{X}(t)}\right]$. We note that $E\left[\delta_{\mathbf{x_b},\mathbf{X}(t)}\right] = F_{\mathbf{b}}(t) = f^{\rm seg}_{I(\mathbf{b})}(t)$ and using Eq. (19) we have $E\left[T(\mathbf{x_b})|\mathbf{x_a}\right] = \sum_{t=0}^{\infty} f^{\rm seg}_{I(\mathbf{b})}(t) = \left[\sum_{t=0}^{\infty} \mathbf{w}^t \mathbf{f}^{\rm seg}(0)\right]_{I(\mathbf{b})}$. Using Eqs. (27) and (29) then allows us to write the mean time of segregation at frequencies $\mathbf{x_b}$ as

$$E\left[T(\mathbf{x_b})|\mathbf{x_a}\right] = G_{I(\mathbf{b}),I(\mathbf{a})}. \tag{32}$$

The mean time of segregating at *any* non-fixed set of frequencies is

$$\sum_{\text{all non fixed } \mathbf{b}} E\left[T(\mathbf{x_b})|\mathbf{x_a}\right] = \sum_{j=K+1}^{I_{\max}} G_{j,I(\mathbf{a})}. \tag{33}$$

Time to fixation

From Eq. (20) of the main text, we observe that $f_i^{\rm fixed}(t)$ breaks up into a sum of t terms of the form $\left[\mathbf{u}\mathbf{w}^{r-1}\mathbf{f}^{\rm seg}(0)\right]_i = \left[\mathbf{u}\mathbf{w}^{r-1}\right]_{i,I(\mathbf{a})}$. Each such term has the natural interpretation as a *joint* probability of two events: (i) the segregation of allele A_i for $r-1$ generations and (ii) the occurrence of fixation of allele A_i *at the end* of the r'th

generation. This is consistent with the interpretation that $f_i^{\text{fixed}}(t)$ is the probability that allele A_i has fixed *by* the end of generation t.

Consider the particular generation where allele A_i fixes, namely at the process of thinning, at the end of the generation. We count the presence of more than one type of allele, prior to fixation in this generation, as the occurrence of segregation. Thus if we write the *conditional* probability that: allele A_i segregates for t generations, given allele A_i fixes at the end of this time, as $P(t|\{A_i \text{ fixes}\}, \mathbf{x_a})$, it follows that this equals $[\mathbf{uw}^{t-1}]_{i,I(\mathbf{a})} / \Pi_i(\mathbf{x_a})$. That is,

$$P(t|\{A_i \text{ fixes}\}, \mathbf{x_a}) = \frac{[\mathbf{uw}^{t-1}]_{i,I(\mathbf{a})}}{[\mathbf{uG}]_{i,I(\mathbf{a})}}, \qquad t = 1, 2, ..., \quad i = 1, 2, ..., K. \tag{34}$$

The mean time to fixation of allele A_i is then given by

$$\sum_{t=1}^{\infty} tP(t|\{A_i \text{ fixes}\}, \mathbf{x_a}) = \sum_{t=1}^{\infty} t \frac{[\mathbf{uw}^{t-1}]_{i,I(\mathbf{a})}}{[\mathbf{uG}]_{i,I(\mathbf{a})}} = \frac{\left[\mathbf{uG}^2\right]_{i,I(\mathbf{a})}}{[\mathbf{uG}]_{i,I(\mathbf{a})}}, \quad i = 1, 2, ..., K. \tag{35}$$

From the above equation and Eq. (30), it follows that

$$\sum_{i=1}^{K} \sum_{t=1}^{\infty} tP(t|\{A_i \text{ fixes}\}, \mathbf{x_a}) P(\{A_i \text{ fixes}\}|\mathbf{x_a}) = \sum_{i=1}^{K} \left[\mathbf{uG}^2\right]_{i,I(\mathbf{a})} \tag{36}$$

and it is natural to interpret this result as the mean time of segregating at *any* non-fixed set of frequencies, i.e., $\sum_{\text{all non fixed } \mathbf{b}} E\left[T(\mathbf{x_b})|\mathbf{x_a}\right]$. To establish this interpretation, we note that conservation of probability entails $\sum_{i=1}^{I_{\max}} v_{i,j} = 1$. Using the two *row* vectors $\mathbf{L}_K$ and $\mathbf{L}_{K'}$ of length K and K', respectively, in which all elements are 1, we can express conservation of probability as $(\mathbf{L}_K, \mathbf{L}_{K'})\mathbf{v} = (\mathbf{L}_K, \mathbf{L}_{K'})$. This

result, using Eq. (15), leads to $\mathbf{L}_K\mathbf{u} + \mathbf{L}_{K'}\mathbf{w} = \mathbf{L}_{K'}$ which can also be written as $\mathbf{L}_K\mathbf{u} = \mathbf{L}_{K'}\mathbf{G}^{-1}$. Multiplying this last result from the right by $\mathbf{G}^2$ yields $\mathbf{L}_K\mathbf{u}\mathbf{G}^2 = \mathbf{L}_{K'}\mathbf{G}$ and is equivalent to $\sum_{i=1}^{K}\left[\mathbf{u}\mathbf{G}^2\right]_{i,I(\mathbf{a})} = \sum_{i=K+1}^{I_{\max}} G_{i,I(\mathbf{a})}$. By Eq. (33), this does indeed equal the mean time of segregating at *any* non-fixed set of frequencies, $\sum_{\text{all non fixed } \mathbf{b}} E\left[T(\mathbf{x_b})|\mathbf{x_a}\right]$.

Appendix D

In this Appendix, we test the approach of this work, by comparing its results with exact results. The exact results are for expected values of allele frequencies and their covariances, conditional on the allele frequencies in the previous generation, namely $E\left[X_i(t+1)|\mathbf{X}(t)=\mathbf{x}\right]$ and $\mathrm{Cov}\left(X_i(t+1),X_j(t+1)|\mathbf{X}(t)=\mathbf{x}\right)$. These quantities are known, for an arbitrary number of allele types, K, and an arbitrary population size, N (Nagylaki 1992; Chapter 9), when selection that has no (or negligible levels of) dominance so allele frequencies in adults have a multinomial distribution (see main text for details).

The viability of an A_iA_j genotype individual is taken to be $(1+s_i)(1+s_j)$. We incorporate mutation into the lifecycle, with $R_{i,j}$ denoting the probability of a parental A_j allele yielding an A_i allele in a gamete ($\sum_{i=1}^{K} R_{i,j}=1$). Following Nagylaki (1992; Chapter 9) we define

$$x_i^{**}=\frac{\sum_{j,k=1}^{K} R_{i,j}x_j(1+s_j)(1+s_k)x_k}{\sum_{j,k=1}^{K} x_j(1+s_j)(1+s_k)x_k}=\frac{\sum_{j=1}^{K} R_{i,j}x_j(1+s_j)}{\sum_{j=1}^{K} x_j(1+s_j)} \tag{37}$$

and obtain

$$E\left[X_i(t+1)|\mathbf{X}(t)=\mathbf{x}\right]=x_i^{**} \tag{38}$$

and

$$\mathrm{Cov}\left(X_i(t+1),X_j(t+1)|\mathbf{X}(t)=\mathbf{x}\right)=\frac{1}{2N}x_i^{**}(\delta_{i,j}-x_j^{**}) \tag{39}$$

where $\delta_{i,j}$ denotes a Kronecker delta ($\delta_{i,j}=1$ if $i=j$ and vanishes otherwise).

To obtain expressions for the corresponding quantities using the formalism in-

troduced in the present work, we use Eq. (5), which holds for a multidimensional transition matrix (with elements $V_{\mathbf{n},\mathbf{m}}$) that incorporate both mutation and selection. We proceed by deriving expressions for the expected allele frequencies, and covariances in terms of $V_{\mathbf{n},\mathbf{m}}$ and then expressing the results in terms of the ordinary matrix $\mathbf{v}$ of Eq. (12) and numerically evaluating them. The forms of selection and mutation, adopted above, yield a function $D_i(\mathbf{x})$ of Eq. (1) given by

$$D_i(\mathbf{x}) = \frac{\sum_{j,k=1}^{K} R_{i,j} x_j (1+s_j)(1+s_k) x_k}{\sum_{j,k=1}^{K} x_j (1+s_j)(1+s_k) x_k} - x_i = x_i^{**} - x_i \tag{40}$$

and this determines the $V_{\mathbf{n},\mathbf{m}}$ and hence the matrix $\mathbf{v}$.

For the expected values of allele frequencies, we have $E\left[X_i(t+1)|\mathbf{X}(t) = \mathbf{m}/(2N)\right] = \sum_{\mathbf{n}} \mathbf{n}_i/(2N) V_{\mathbf{n},\mathbf{m}}$ which can also be written $\sum_{j=1}^{I_{\max}} n_i^{[j]} V_{\mathbf{n}^{[j]},\mathbf{m}}/(2N) = \sum_{j=1}^{I_{\max}} n_i^{[j]} v_{I(\mathbf{n}^{[j]}),I(\mathbf{m})}/(2N)$ where $n_i^{[j]}$ is the i'th component of the vector index $\mathbf{n}^{[j]}$. Using Eq. (8), we thus obtain

$$E\left[X_i(t+1)|\mathbf{X}(t) = \frac{\mathbf{m}}{2N}\right] = \sum_{j=1}^{I_{\max}} \frac{n_i^{[j]}}{2N} v_{j,I(\mathbf{m})}. \tag{41}$$

In a similar way, the conditional covariances are given by

$$\begin{aligned} &\mathrm{Cov}\left(X_i(t+1), X_j(t+1)|\mathbf{X}(t) = \frac{\mathbf{m}}{2N}\right) \\ &= \sum_{k=1}^{I_{\max}} \frac{n_i^{[k]}}{2N} \frac{n_j^{[k]}}{2N} v_{k,I(\mathbf{m})} - \left(\sum_{k=1}^{I_{\max}} \frac{n_i^{[k]}}{2N} v_{k,I(\mathbf{m})}\right)\left(\sum_{k=1}^{I_{\max}} \frac{n_j^{[k]}}{2N} v_{k,I(\mathbf{m})}\right). \end{aligned} \tag{42}$$

We have compared the expressions above for the conditional expectations and covariances for a range of N from 10 to 30, for $K = 3$ and a range of s_j ranging from

-1 to 3. In all cases, without exception, the exact results and the results derived from the formalism of this work differed by less than one part in 10^{13}.

References

Baxter, G. J., Blythe, R. A., Croft, W., and McKane, A. J., 2006. Utterance selection model of language change. Phys. Rev E 73 046118.

Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S., and Kondrashov, A. S., 2004. Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. Nature 429, 558-562.

Fisher, R. A., 1922. On the dominance ratio. Proc. R. Soc. Edinb. 42, 321-341.

Freund, J. E., Miller, I. and Miller, M. 1999. John E. Freund's Mathematical Statistics. Prentice Hall, Upper Saddle River, N.J.

Griffiths, R. C., 1979. A transition density expansion for a multi-allele diffusion model, Adv. Appl. Probab. 11, 310-325

Haigh, J., 2002. Probability Models. Springer-Verlag. London

Kimura, M., 1955a. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbour Symp. Quant. Biol. 20, 33–53.

Kimura, M., 1955b. Random genetic drift in multi-allelic locus. Evolution 9, 419-435.

Kimura, M., 1956. Random genetic drift in a tri-allelic locus; exact solution with a continuous model. Biometrics 12, 57-66.

Littler, R. A., 1975. Loss of variability at one locus in a finite population, Math. Biosci. 25, 151-163.

Littler, R. A. and Good, A. J., 1978. Ages, extinction times, and first passage probabilities for a multiallele diffusion model with irreversible mutation. Theor. Popul. Biol. 13, 214-225.

Nagylaki, T. 1992. Introduction to Theoretical Population Genetics, Springer, Berlin

Wright, S. 1931. Evolution in Mendelian populations. Genetics, 16, 97-159.

Wright, S., 1945. The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. USA 31, 382–389.