Article:

Zhao, L., Gossmann, T.I. and Waxman, D. (2016) A modified Wright-Fisher model that incorporates Ne: A variant of the standard model with increased biological realism and reduced computational complexity. Journal of Theoretical Biology, 393. pp. 218-228. ISSN 0022-5193

https://doi.org/10.1016/j.jtbi.2016.01.002

31/12/2015

A modified Wright-Fisher model that incorporates N_e: A variant of the standard model with increased biological realism and reduced computational complexity

Lei Zhao[1], Toni I. Gossmann[2], and David Waxman[1]

[1]Centre for Computational Systems Biology, Fudan University, Shanghai 200433, People's Republic of China

[2]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom

Running Title: Incorporating the effective size into the Wright-Fisher model

Correspondence to:
Professor D. Waxman
Centre for Computational Systems Biology, Fudan University, 220 Handan Road, Shanghai 200433, People's Republic of China.
E-mail: davidwaxman@fudan.edu.cn

Keywords: effective population size; gene fixation and loss, site frequency spectrum, theoretical population genetics, computational methods

ABSTRACT

The Wright-Fisher model is an important model in evolutionary biology and population genetics. It has been applied in numerous analyses of finite populations with discrete generations. It is recognised that real populations can behave, in some key aspects, as though their size that is not the census size, N, but rather a smaller size, namely the effective population size, N_e. However, in the Wright-Fisher model, there is no distinction between the effective and census population sizes. Equivalently, we can say that in this model, N_e coincides with N. The Wright-Fisher model therefore lacks an important aspect of biological realism. Here, we present a method that allows N_e to be directly incorporated into the Wright-Fisher model. The modified model involves matrices whose size is determined by N_e. Thus apart from increased biological realism, the modified model also has reduced computational complexity, particularly so when $N_e \ll N$. For complex problems, it may be hard or impossible to numerically analyse the most commonly-used approximation of the Wright-Fisher model that incorporates N_e, namely the diffusion approximation. An alternative approach is simulation. However, the simulations need to be sufficiently detailed that they yield an effective size that is different to the census size. Simulations may also be time consuming and have attendant statistical errors. The method presented in this work may then be the only alternative to simulations, when N_e differs from N. We illustrate the straightforward application of the method to some problems involving allele fixation and the determination of the equilibrium site frequency spectrum. We then apply the method to the problem of fixation when three alleles are segregating in a population. This latter problem is significantly more complex than a two allele problem and since the dif-

fusion equation cannot be numerically solved, the only other way N_e can be incorporated into the analysis is by simulation. We have achieved good accuracy in all cases considered. In summary, the present work extends the realism and tractability of an important model of evolutionary biology and population genetics.

1 Introduction

The Wright-Fisher model (WFM) was introduced to describe the random genetic drift of the frequency of an allele in a finite population (Fisher 1922; Wright 1931). The model applies for populations with discrete generations, and can incorporate essentially deterministic evolutionary forces such as selection, migration and recurrent mutation (Ewens 2004). The WFM remains of current interest, with numerous applications in the recent literature involving genomic data that, to mention just two, are its use in estimating the effective population size (Hui and Burt 2015) and its use in tracking selection (Thépot et al. 2015). While the WFM is an extremely important model and has often been employed, it suffers from two drawbacks, which detract from its usefulness, and which the present work goes some way to resolve.

The first drawback is that the WFM explicitly depends on only one population size, namely the number of adults present in the population. This is a quantity we term the *census size*, and denote by N. Following Wright and many subsequent authors, it is recognised that biological populations can behave, in important aspects, as though their size is not the actual number of adults, N, but rather a different, typically smaller value, N_e, that is termed the effective population size (Wright, 1931). The effective size usually arises from a population deviating, in one or more ways, from being 'ideal', such as when individuals do not have a Poisson distributed number of offspring, or related individuals interbreed, or when populations show age, stage and spatial structures (Charlesworth 2009). A possible way to account for a population behaving as if its size is N_e *appears to be* to simply replace N by N_e in, for example, the WFM[1]. However, this does not directly work, as we show below.

The second drawback of the WFM is that its mathematical description can involve large matrices which, in the simplest problems (such as a single locus with two alleles), involve of the order of N^2 elements. More complicated problems (e.g., one locus with > 2 alleles, or multiple loci, or structured populations, or ...) can lead to matrices involving of the order of N^α elements with $\alpha \geq 4$ (Waxman 2009). Thus even for a modest population sizes, such as $N = 1000$, this can lead to substantial computational issues.

In this work, we provide a method of incorporating the effective population size, N_e, into the WFM. We demonstrate that the method works in a variety of different circumstances, to the extent that we view the method as a useful working principle. The method leads to N being replaced by N_e in the WFM, *but in an appropriate and non trivial way*, and, as we shall see, this resolves the first drawback noted above. Furthermore, if N_e is small compared with N then replacement of N by N_e goes some way to reducing the computational complexity of calculations (with a considerable reduction in computational complexity if $N_e \ll N$), and hence reducing the severity of the second drawback.

The reason we cannot simply replace the census size of the population by the effective size in the WFM is that there is a mismatch between the discrete allele frequencies of a population of size N and the discrete allele frequencies of a population of size N_e. To see this consider a haploid popu-

[1] Throughout this work we assume the effective population size, N_e, takes an integral value. If the effective size is estimated or calculated in some way, then generally it will not be an integer. In the work we present, we shall take N_e to be given by the nearest integer to the estimated/calculated value.

lation, of census size N, in an initial state where a single adult carries a focal allele. The initial frequency of the focal allele is, non-negotiably, $1/N$. If we simply replace the population size, N, by the effective size, N_e, then the smallest non-zero frequency becomes $1/N_e$. The effective size is generally smaller than the census size ($N_e < N$) hence the value of $1/N_e$ is generally larger than the smallest non-zero frequency of the actual population ($1/N$), possibly much larger. For example, if $N = 1000$ and $N_e = 100$ then we have $1/N_e = 10^{-2}$ which is 10 times the value of $1/N = 10^{-3}$. Thus, whatever else that naive replacement of N by N_e does, any result for an actual initial frequency of $1/N$, can, at best after the replacement, only be determined by the smallest non-zero initial frequency of $1/N_e$ with generally erroneous results. This problem of mismatch of frequencies in populations of size N and N_e is more general than just for the smallest non-zero frequency, and holds for many other frequencies.

The frequency mismatch problem, just described, is evaded under a well-known approximation of the WFM, namely the diffusion approximation (Fisher 1922; Wright 1945; Kimura 1955). This is an approximation that takes both the census size of the population, N, and the effective population size, N_e, into account. The approximation involves a diffusion equation for the distribution of an allele's frequency (hence the approximation's name), and has two important features. The first feature is that N_e takes the place of N in the diffusion equation. This means the dynamics of the frequency is treated as if the population had a census size of N_e, in accordance with the general idea behind the effective population size. The diffusion approximation has a second feature that it treats an allele's frequency as a continuous quantity. This means that the initial frequency can be chosen to be *any value*. Accordingly, when initially there is, e.g., a single copy of an allele in a population of census size N, the initial frequency can be chosen to be the *correct value*, namely $1/N$, irrespective of the value of N_e. In practice the above two features of the diffusion approximation generally work well together, to the extent that the diffusion approximation can determine many properties to good accuracy even for relatively small populations (Ewens 1964).

There is, however, a drawback of the diffusion approximation. Except in a rather small subset of problems that can be analytically solved, the diffusion equation, which plays a central role in the approximation, has solutions which can only be found numerically. While numerical procedures exist for the case of one locus with two alleles (see e.g., Zhao et al. 2013) it appears to be very difficult to extend these methods to more complex problems where the dimensionality, associated with allele frequencies, is higher. Alternative ways to proceed are simulations (which have to be sufficiently detailed that they yield an effective population size that differs from the census size) or - the innovation of the present work - a modification of the WFM. Simulations may be time consuming and are subject to statistical errors, however, the WFM, which is formulated in terms of matrices and vectors, is amenable to a computational analysis (in principle, at least, even for complex problems (see Waxman 2009)). In this work we present a *modified* WFM where the effective population size, N_e, is directly incorporated into the WFM, with advantages of both biological realism and computational efficiency.

We now state and explain what we view as a working principle that allows incorporation of the effective population size into the WFM.

2 Principle

The simplest statement of the principle amounts to saying that we should treat the population as though it has a census (or actual) size of N when the copy number of an allele is definitely known, e.g., when a mutation first appears in a population, but in all subsequent generations, the dynamics of the allele's frequency behaves as if the actual population size were N_e. The previous sentence is theoretically equivalent to saying that the population size discontinuously changes from N, in the generation where the copy number is definitely known, to the size N_e in the next generation – and all subsequent generations[2]. This viewpoint, of a discontinuous change of the population size from N to N_e is also a possible interpretation of solutions of the diffusion equation, where the frequency that is used at an initial time is correct for a population of size N, but the subsequent dynamics of the allele is treated as though the population has a census size of N_e. As a consequence, the principle we are proposing, to incorporate N_e into the WFM, is expected to hold to good accuracy in all of the circumstances where the diffusion approximation holds to good accuracy.

We find it helpful to formally state the principle in the simple context of a haploid population with a census size of N and an effective size of N_e, as follows.

[2] If the effective population size changes with time, we shall use the notation $N_e(t)$ to represent the local (in time) effective population size. That is, $N_e(t)$ is a quantity determined from processes occurring in a only *single* generation (Ewens 2004). The quantity $N_e(t)$ is the effective size appearing in the diffusion equation (Waxman 2012), since it is associated with the *instantaneous* rate at which genetic drift increases the genetic variance between different replicate populations. In this work we shall not use averages of the effective population size, such as the harmonic mean, which summarise properties of $N_e(t)$ over multiple generations, and which reflect information about $N_e(t)$ that is non-local in time.

> When a *known number* of n copies of an allele (or mutant) are initially present in a given generation, then in that generation the population size should be taken as the actual census size of the population, N, so the allele's frequency is n/N. However, in all subsequent generations, the population size should be taken as the effective population size, N_e. For time dependent N and N_e this principle is straightforwardly extended[3].

To provide some examples and comparisons that clearly illustrate the working of this principle, we need to have an explicit example of a population whose effective size differs from its census size. There are many different origins of the effective population size, and we shall make use of a specific population which has a well-defined effective population size. We term this population the *Test Population* and introduce it next. We emphasise that the primary interest of the Test Population is to *test* our results; it may or may not be relevant to a real biological scenario of interest.

3 Test Population

We consider a population comprised of haploid individuals. These have a single biallelic locus under selection and we label the two alleles A and B. We assume a constant census size of N, and discrete generations which are labelled by $t = 0, 1, 2, ...$ When n adults carry the A allele, its frequency

[3]Assume both N and N_e depend on time: $N = N(t)$ and $N_e = N_e(t)$. and the copy number of an allele is definitely known in generation t to be n. Then the appropriate population size to use is $N(t)$ and the initial frequency is $n/N(t)$. The relevant effective sizes that should be used in generations $t+1$, $t+2$, ... are the local values appropriate to these generations, i.e., $N_e(t+1)$, $N_e(t+2)$, ...

is n/N. We shall use X_t to denote the frequency of the A allele in adults in generation t, and the corresponding frequency of the B allele is $1 - X_t$. Changes in X_t are assumed to be governed by the following lifecycle.

Generation t	**adults**	
	$\downarrow$	reproduction followed by the death of adults
	offspring	
	$\downarrow$	thinning (number regulation)
Generation $t+1$	**adults.**	

We neglect the occurrence of mutations and assume there are reproductive differences of carriers of the different alleles, as given in Table 1, which was motivated by the work of Gillespie (1974; 1975).

	mean No. of offspring	variance in No. of offspring
carrier of the A allele	$f \times (1+s)$	$f^2\sigma^2$
carrier of the B allele	f	$f^2\sigma^2$

Table 1 Title: Basic statistics of reproductive outputs in the Test Population

Table 1 Caption: This table shows basic statistics of the reproductive outputs of different allele-carriers in the Test Population. The quantity f represents a baseline fertility, while s is the selection coefficient of an A allele relative to a B allele.

Both alleles have the same variance in offspring number, which is taken to be $f^2\sigma^2$ where σ^2 is a constant.

We interpret the scheme in Table 1 as fertility selection (but see Gillespie 1975), where the A allele has a selective advantage of s relative to the B allele.

The quantity f represents a baseline fertility. Its presence in both the mean and the variance in Table 1 leads to a coefficient of variation (= standard deviation/mean) of the number of offspring that is independent of f and results in a simple form of the effective population size (see below).

Within the lifecycle, thinning of the population to N individuals is non-selectively carried out according to sampling with replacement, i.e., 'binomial sampling', as used in the standard WFM.[4]

The above specification of a population is, of course, incomplete; a complete description includes the actual distribution of offspring numbers produced by an adult of the population. While there are many possible distributions that could serve for this purpose, representing different biological situations, the distribution of offspring numbers we choose is the *negative binomial distribution* (see e.g., Johnson et al. 2005). This is a convenient and not unreasonable choice. The negative binomial distribution has a variance that is generally larger than its mean, but has a Poisson distribution

[4]The Test Population involves independent reproduction of each individual, followed by population thinning that ensures the census size is N. Generally, all calculations for the Test Population should be conditioned on the number of offspring equalling/exceeding N, since it is possible that after reproduction, the total number of offspring is smaller than N, and thinning cannot be carried out. For the parameters we later adopt in this work for simulations, this conditioning is not required, because population non-replacement is extremely improbable, and was never observed in the simulations.

(which is often adopted for offspring numbers) as a limiting case. A negative binomial distribution is controlled by two parameters, and specification of its mean and variance fully determine these parameters and hence the distribution. This conveniently means there is no need to introduce additional parameters beyond those of Table 1. Additionally, there is evidence in the literature that reproductive success in some species is reasonably approximated by a negative binomial distribution (Grant and Grant 2000; Anderson, Ward and Carlson 2011), and some studies have described models where the lifecycle involves randomness associated with a negative binomial distribution (Melbourne and Hastings 2008; Reiss 2013).

The above constitutes a complete description of the Test Population.

3.1 Properties of the Test Population

We note that as the parameter σ^2 approaches zero and f approaches infinity, the Test Population can be described by a standard WFM where $N_e = N$ and the A allele has a selective advantage of s over the B allele. However generally, the Test Population cannot be described by a standard WFM. Applying the analysis of Gillespie (1974; 1975), suggests that the Test Population is equivalent, under a diffusion approximation, to a population where the selection coefficient of the A allele is replaced by an effective value that may be frequency-dependent, and the census size of the population is replaced by an effective size that may also be frequency dependent. We shall assume that: (i) the A allele's selection coefficients is small, $|s| \ll 1$; (ii) the population size is large, $N \gg 1$; (iii) the baseline fertility is large, $f \gg 1$; (iv) the parameter σ^2 is much smaller than N, $\sigma^2 \ll N$. We will work in the framework of the reproductive scheme in Table 1, combined with the thinning process of the Test Population. We then find the following (cf. Gillespie 1974; 1975): (i) apart from small corrections of order $s\sigma^2/N$, the effective selection coefficient is frequency independent and has the value s (see Table 1); (ii) apart from small corrections of order $\left[f\left(1+\sigma^2\right)\right]^{-1}$, the effective population size is also frequency independent and given by

$$N_e = \frac{N+\sigma^2}{1+\sigma^2}. \tag{1}$$

In general, the value of N_e following from this equation is not an integer. As stated in the Introduction, the N_e that we shall use in calculations will be the closest integer to the result in Eq. (1).

We can summarise the Test Population by saying it has a census size of N and, emerging from individual reproduction and thinning of the population, it has selection of strength s, and an effective population size given by Eq. (1).

4 Applying the modified Wright-Fisher model to the Test Population

4.1 Standard results of a Wright-Fisher model

We shall make use of some results of a standard WFM for a haploid population of finite census size N, with discrete generations, where individuals have a single locus with two alleles, labelled A and B. The population is assumed have an effective population size that coincides with the census size. The behaviour of the distribution for this population can be written as

$$\mathbf{F}(t+1) = \mathbf{W}\mathbf{F}(t) \tag{2}$$

where $\mathbf{F}(t)$ is a column vector with $N+1$ elements, corresponding to probabilities of the different frequency states of a population of size N, and $\mathbf{W}$ is square matrix of size $(N+1)\times(N+1)$ - the transition matrix - which contains probabilities of transitions between frequency states of the population.

If the A allele has a small selective advantage of s over the B allele, and there is no mutation and migration, then it is well known that the transition matrix for the finite population is given by

$$W_{m,n} = \binom{N}{m} \left[D\left(x_n^{(0)}\right)\right]^m \left[1 - D\left(x_n^{(0)}\right)\right]^{N-m} \tag{3}$$

where n and m take the values $0, 1, ..., N$, the quantity $\binom{N}{m} = \frac{N!}{m!(N-m)!}$ is a binomial coefficient,while $x_n^{(0)} = n/N$ are the possible frequencies of an allele in a haploid population of size N and, with small corrections of order s^2,

$$D(x) = x + sx(1-x) \tag{4}$$

(see e.g., Ewens 2004).

In the calculations we shall present, it is useful to write the transition matrix in a 'block' form (see e.g., Waxman 2011). For the transition matrix of Eq. (3) we write

$$\mathbf{W} = \begin{pmatrix} W_{0,0} & W_{0,1} & \cdots & \\ W_{1,0} & W_{1,1} & \cdots & \\ \vdots & \vdots & \ddots & \\ & & & W_{N,N} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{u} & 0 \\ \mathbf{0} & \mathbf{w} & \mathbf{0} \\ 0 & \mathbf{v} & 1 \end{pmatrix} \tag{5}$$

where $\mathbf{0}$, is a column vector of length $N-1$ where all elements are 0, while $\mathbf{u}$ and $\mathbf{v}$ are row vectors of length $N-1$, and $\mathbf{w}$ is square matrix of size

$(N-1) \times (N-1)$. The elements of $\mathbf{u}$ and $\mathbf{v}$ are probabilities of transition from states of the population where the A allele is segregating to states where this allele is lost or fixed[5]. The elements of the matrix $\mathbf{w}$ are transition probabilities between pairs of states where the A allele is segregating.

4.2 Modified Wright-Fisher model for the Test Population

The modified WFM of the present work, that incorporates N_e, can be directly applied to the Test Population. In doing so, we will make use of the standard WFM results given in Eqs. (2) and (3), along with Eq. (4).

The modified WFM follows from assuming that in a given generation, say generation 0, the population size is N and the distribution is known. The method then assumes that incorporating the effective population size into the dynamics is equivalent to the population size changing from N to N_e at the end of generation 0, and then remaining at the value N_e.

For generation 0 we write the distribution of the A allele's frequency as $\mathbf{F}(0)$. This distribution describes a population of size N and is a column vector with $N+1$ elements, corresponding to probabilities of the different frequency states of the population.

After generation 0 we describe the population by an *effective distribution* that is appropriate to a population with N_e individuals. We write the effective distribution for generation t (with $t \geq 1$) as $\mathbf{F}^{(e)}(t)$. This is a column vector with $N_e + 1$ elements. The behaviour of the effective distribution is

[5] Let us point out the labelling convention we use for elements of matrices with a block structure like that of $\mathbf{W}$ (Eq. (5)). The elements of the row vector $\mathbf{u}$ in Eq. (5) correspond to $\mathbf{u} = (W_{0,1}, W_{0,2}, W_{0,3}, ...)$. We shall also write this vector as $\mathbf{u} = (u_1, u_2, u_3, ...)$. In other words, for *row vectors* such as $\mathbf{u}$ (and $\mathbf{v}$), their elements have labels that start at 1 and not 0. For such row vectors, we shall sometimes use the notation $[\mathbf{u}]_n$ to denote the n'th element, i.e., to denote u_n, with $n = 1, 2, ...$.

given by

$$\mathbf{F}^{(e)}(1) = \mathbf{W}^{(0)}\mathbf{F}(0)$$
$$\mathbf{F}^{(e)}(t+1) = \mathbf{W}^{(e)}\mathbf{F}^{(e)}(t), \qquad t = 1, 2, \ldots \,. \tag{6}$$

Here $\mathbf{W}^{(0)}$ is a *rectangular* transition matrix of size $(N_e+1)\times(N+1)$ that takes into account the genetic drift and selection of the Test Population that occurs in going from generation 0 (where the population size is N), to generation 1 (where the population size is treated as N_e), while $\mathbf{W}^{(e)}$ is an effective transition matrix of size $(N_e+1)\times(N_e+1)$ that is defined in complete analogy to a standard WFM, but for a population of census size N_e.

We write the possible allele frequencies in populations of size N and N_e as $x_n^{(0)}$ and $x_n^{(e)}$, respectively, with

$$x_n^{(0)} = n/N \text{ with } n = 0, 1, 2, ..., N$$
$$x_n^{(e)} = n/N_e \text{ with } n = 0, 1, 2, ..., N_e. \tag{7}$$

The two transition matrices can be written in terms of the function $D(x)$ of Eq. (4) as

$$W_{m,n}^{(0)} = \binom{N_e}{m} \left[D\left(x_n^{(0)}\right)\right]^m \left[1 - D\left(x_n^{(0)}\right)\right]^{N_e - m} \tag{8}$$

with $m = 0, 1, .., N_e$ and $n = 0, 1, .., N$, and

$$W_{m,n}^{(e)} = \binom{N_e}{m} \left[D\left(x_n^{(e)}\right) \right]^m \left[1 - D\left(x_n^{(e)}\right) \right]^{N_e - m} \tag{9}$$

with m and $n = 0, 1, .., N_e$.

A 'block' form of the transition matrices $\mathbf{W}^{(e)}$ and $\mathbf{W}^{(0)}$ of Eqs. (8) and (9) that is similar to that of a standard WFM (Eq. (5)), turns out to be useful in the calculations that follow. These take the form

$$\mathbf{W}^{(e)} = \begin{pmatrix} 1 & \mathbf{u}^{(e)} & 0 \\ \mathbf{0}^{(e)} & \mathbf{w}^{(e)} & \mathbf{0}^{(e)} \\ 0 & \mathbf{v}^{(e)} & 1 \end{pmatrix}, \qquad \mathbf{W}^{(0)} = \begin{pmatrix} 1 & \mathbf{u}^{(0)} & 0 \\ \mathbf{0}^{(0)} & \mathbf{w}^{(0)} & \mathbf{0}^{(0)} \\ 0 & \mathbf{v}^{(0)} & 1 \end{pmatrix}. \tag{10}$$

Here: $\mathbf{0}^{(e)}$ and $\mathbf{0}^{(0)}$ are column vectors of length $N_e - 1$ with all elements 0; $\mathbf{u}^{(e)}$ and $\mathbf{v}^{(e)}$ are row vectors of length $N_e - 1$; $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$ are row vectors of length $N - 1$; $\mathbf{w}^{(e)}$ is a square matrix of size $(N_e - 1) \times (N_e - 1)$; $\mathbf{w}^{(0)}$ is a rectangular matrix of size $(N_e - 1) \times (N - 1)$.

5 Illustrative examples involving the Test Population

We now consider some illustrative examples involving the Test Population, which we note is one possible way an effective population size, N_e, can arise. We shall apply our modified WFM, that incorporates N_e, using Eq. (1). We can then make the comparison with the diffusion approximation (when results are available). As an independent test, we shall also carry out simulations (which do not assume validity of the diffusion approximation), and which are also based on the Test Population.

5.1 Probabilities of fixation and loss

We use arguments, based on Eq. (6), that are very similar to those used in the standard WFM to determine the probabilities of ultimate fixation and loss of the A allele. These results involve 'blocks' from the matrices $\mathbf{W}^{(0)}$ and $\mathbf{W}^{(e)}$ given in Eq. (10). The results are concisely expressed in terms of a matrix $\mathbf{G}^{(e)}$ defined by

$$\mathbf{G}^{(e)} = \left(\mathbf{I}^{(e)} - \mathbf{w}^{(e)}\right)^{-1} \tag{11}$$

where $\mathbf{I}^{(e)}$ is an identity matrix that is the same size as $\mathbf{w}^{(e)}$.

We find that when n copies of the A allele are initially present in the population ($n = 1, 2, .., N - 1$), so the A allele is initially at a frequency of n/N, the probabilities of fixation and loss of the A allele are

$$P_{\text{fix}}(n) = \left[\mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n \tag{12}$$

and

$$P_{\text{loss}}(n) = \left[\mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n \tag{13}$$

(see Appendix A for details).

Note that when $N_e = N$, and $\mathbf{G}^{(e)}$ becomes $\mathbf{G} = (\mathbf{I} - \mathbf{w})^{-1}$, Eqs. (12) and (13) reduce to $P_{\text{fix}}(n) = [\mathbf{vG}]_n$ and $P_{\text{loss}}(n) = [\mathbf{uG}]_n$ (cf. Waxman 2009).

In Figure 1 we illustrate how the results in Eqs. (12) and (13), from the modified WFM, compare with results from the diffusion approximation and simulations.

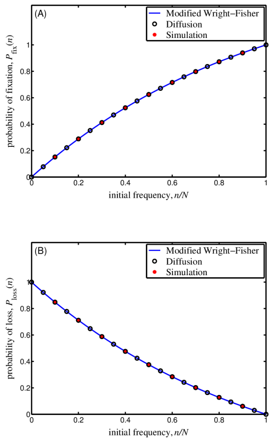

Figure 1 **Caption:** This figure gives results, from three different calculational methods, for the probability of ultimate fixation of the A allele as a function of initial frequency (Panel A), and the probability of loss of the A allele as a function of initial frequency (Panel B). The three methods are: (i) the modified Wright-Fisher model, which was introduced in this work, (ii) the diffusion approximation, (Kimura 1962) and (iii) simulation. The parameter values adopted were: census population size, $N = 500$; selection coefficient of the A allele relative to the B allele, $s = 0.01$; baseline fertility, $f = 100$; value of σ^2 (related to the variance in offspring number of an adult - see Table 1), $\sigma^2 = 9$. For the simulations we used 10^5 replicate populations. The approximate value of the effective population size that follows from these parameters is $N_e = 51$, see Eq. (1).

It is evident from Figure 1 that the results from all three methods of calculation used (modified WFM, diffusion approximation and simulation) are extremely close to each other, despite there being a very substantial difference between the census size ($N = 500$) and the effective population size ($N_e = 51$).

5.2 Mean times to fixation and loss

The mean times to fixation or loss of the A allele are conditional on fixation or loss of this allele ultimately occurring. When there are n copies of the A allele initially present in the population in generation 0 ($n = 1, 2, .., N - 1$), so the A allele is at a frequency of n/N, we write these mean times as $E[T_{\text{fix}}|n]$ and $E[T_{\text{loss}}|n]$, respectively. We find

$$E[T_{\text{fix}}|n] = 1 + \frac{\left[\mathbf{v}^{(e)}\left(\mathbf{G}^{(e)}\right)^2\mathbf{w}^{(0)}\right]_n}{\left[\mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n} \tag{14}$$

and

$$E[T_{\text{loss}}|n] = 1 + \frac{\left[\mathbf{u}^{(e)}\left(\mathbf{G}^{(e)}\right)^2\mathbf{w}^{(0)}\right]_n}{\left[\mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n} \tag{15}$$

(see Appendix A for details).

Note that when $N_e = N$ Eqs. (14) and (15) reduce to $E[T_{\text{fix}}|n] = \left[\mathbf{vG}^2\right]_n / \left[\mathbf{vG}\right]_n$ and $E[T_{\text{loss}}|n] = \left[\mathbf{uG}^2\right]_n / \left[\mathbf{uG}\right]_n$ (cf. Waxman 2009).

In Figure 2 we illustrate how the results in Eqs. (14) and (15), from the modified WFM compare with results from the diffusion approximation and simulations.

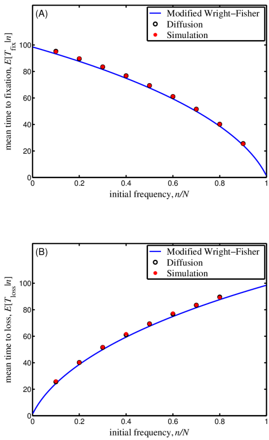

Figure 2 **Caption:** This figure gives results, from three different calculational methods, for the mean time to fixation of an A allele as a function of initial frequency (Panel A), and the mean time to loss of an A allele as a function of initial frequency (Panel B). The three methods are: (i) the modified Wright-Fisher model, which was introduced in this work, (ii) the diffusion approximation (Kimura and Ohta 1969) and (iii) simulation. The parameter values adopted were: census population size, $N = 500$; selection coefficient of the A allele relative to the B allele, $s = 0.01$; baseline fertility, $f = 100$; value of σ^2 (related to the variance in offspring number of an adult - see Table 1), $\sigma^2 = 9$. For the simulations, 10^5 replicate populations were used. The approximate value of the effective population size that follows from these parameters is $N_e = 51$, see Eq. (1).

It is evident from Figure 2 that the results from all three methods of calculation used (i.e., modified WFM, diffusion approximation, and simulation) are close to each other, despite there being a very substantial difference between the census size ($N = 500$) and the effective population size ($N_e = 51$).

5.3 Site frequency spectrum

We shall incorporate the effective population size into results for the site frequency spectrum (SFS), assuming an effectively infinite number of independent (unlinked) sites (see e.g., Evans, Shvets and Slatkin 2007). Mutations are assumed to occur in adults at the beginning of a generation, and once mutations have arisen, each site is described by the dynamics of a Test Population, where no additional mutations occur.

With μ denoting the expected number of new mutations in an adult each generation, and with θ the scaled mutation rate, defined by $\theta = 2N\mu$, the expected number of mutations entering the adult population in a generation is $\theta/2$.

In the main text we only consider the equilibrium SFS, which we write as a column vector $\hat{\mathbf{M}}$ (dynamics of the SFS is considered in Appendix B). The elements of $\hat{\mathbf{M}}$, written $\hat{M}_n$, with $n = 1, 2, ..., N-1$, represent the mean number of sites with mutants at a frequency of n/N. We only include those sites in the SFS where mutant alleles are segregating in the population, and exclude contributions from sites where mutations have become lost or have not occurred or have become fixed.

A consequence of the assumption of an effectively infinite number sites is that each site can, at most, suffer only one mutation; double mutations of a site happen with negligible probability.

The equilibrium SFS represents a steady state situation, where the single copies of new mutations represent an input that balances mutations that are removed by fixation and loss. From dynamical considerations, we can view the value of $\hat{\mathbf{M}}$, in any generation, as arising from two sources: (i) from sites where new mutations originated in adults at the beginning of the generation, written as $\hat{\mathbf{M}}^{new}$, and (ii) from sites associated with mutations which originated in the previous generation, or yet earlier generations, written as $\hat{\mathbf{M}}^{prev}$. We thus have $\hat{\mathbf{M}} = \hat{\mathbf{M}}^{new} + \hat{\mathbf{M}}^{prev}$. The form of $\hat{\mathbf{M}}^{new}$ is explicitly known; it is a column vector where only the first element is non-zero and has the value $\theta/2$. Following the approach of the present work, we can obtain an approximation for $\hat{\mathbf{M}}^{prev}$, which, by assumption, corresponds to sites which have evolved in a manner appropriate to a population size of N_e for at least one generation.

5.3.1 Coarse grained equilibrium site frequency spectrum

The exact SFS is defined at the frequencies $x_n^{(0)} = n/N$ with $n = 1, 2, .., N-1$. By contrast, the effective SFS that we determine is defined at the frequencies $x_n^{(e)} = n/N_e$ with $n = 1, 2, .., N_e - 1$. The $x_n^{(e)}$ represent a coarser grid than the $x_n^{(0)}$, with the spacing between adjacent $x_n^{(e)}$ (i.e., $1/N_e$) being larger than the spacing between adjacent $x_n^{(0)}$ (i.e., $1/N$). For example, if $N_e = N/10$ then for 10 adjacent frequencies where the exact SFS is defined (the $x_n^{(0)}$), there corresponds just one frequency where the effective SFS is defined ($x_n^{(e)}$).

While the effective version of $\hat{\mathbf{M}}^{prev}$, written as $\hat{\mathbf{M}}^{prev,e}$, can be used to approximate properties of the exact SFS, the values of $\hat{\mathbf{M}}^{prev}$ and $\hat{\mathbf{M}}^{prev,e}$ are not directly comparable. Each element of the effective result, $\hat{\mathbf{M}}^{prev,e}$ represents approximately N/N_e elements of the exact result, $\hat{\mathbf{M}}^{prev}$. For the example used above, where $N_e = N/10$, each element of $\hat{\mathbf{M}}^{prev,e}$ represents approximately 10 elements of $\hat{\mathbf{M}}^{prev}$. However, two quantities which are directly comparable are $\hat{\mathbf{M}}^{prev}$ and $(N_e/N) \times \hat{\mathbf{M}}^{prev,e}$. We therefore define the 'coarse grained' approximation of $\hat{\mathbf{M}}^{prev}$ as $\hat{\mathbf{M}}^{prev,cg} = (N_e/N) \times \hat{\mathbf{M}}^{prev,e}$. Thus $\hat{\mathbf{M}}^{prev,cg}$ is defined on a coarse frequency grid given by the $x_n^{(e)}$ however, the magnitude of $\hat{\mathbf{M}}^{prev,cg}$ should be closely comparable with the exact result, $\hat{\mathbf{M}}^{prev}$, when both $\hat{\mathbf{M}}^{prev,cg}$ and $\hat{\mathbf{M}}^{prev}$ are evaluated at a common (or near common) frequency.

In Appendix B we give details of the calculation leading to the coarse-grained form for $\hat{\mathbf{M}}^{prev}$, namely $\hat{\mathbf{M}}^{prev,cg}$. The result is

$$\hat{\mathbf{M}}^{prev,cg} = \frac{\theta_e}{2} \mathbf{G}^{(e)} \mathbf{w}^{(0)} \mathbf{i} \tag{16}$$

where in this equation: $\theta_e = N_e\theta/N = 2N_e\mu$, while the matrices $\mathbf{G}^{(e)}$ and $\mathbf{w}^{(0)}$ appear in Eqs. (10) and (11), and $\mathbf{i}$ is a column vector with $N_e - 1$ elements, all of which are zero except the first, which is unity.

In Figure 3 we plot the equilibrium coarse grained SFS, $\hat{\mathbf{M}}^{prev,cg}$, at the frequencies $x_n^{(e)} = n/N_e$, which is calculated from the modified WFM. We also plot an estimate of the exact form for $\hat{\mathbf{M}}^{prev}$ that is based on simulations of the Test Population, with details of the simulations given in Appendix C.

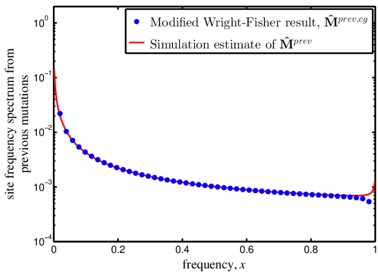

Figure 3 Caption: This figure illustrates the equilibrium SFS that arises from existing mutations, showing the coarse grained result $\hat{\mathbf{M}}^{prev,cg}$ from Eq. (16) (blue dots), and an estimate of $\hat{\mathbf{M}}^{prev}$ from simulations (red line). The parameter values adopted for the figure were: scaled mutation rate, $\theta/2 = 1$; census population size, $N = 500$; selection coefficient of the A allele

relative to the B allele, $s = 0.01$; baseline fertility, $f = 100$; value of σ^2 (related to the variance in offspring number of an adult - see Table 1), $\sigma^2 = 9$. The approximate value of the effective population size that follows from these parameters is $N_e = 51$, see Eq. (1). Note that the equilibrium SFS is proportional to θ, so for a different value of θ, the results in Figure 3 simply become multiplied by $\theta/2$. The simulation procedure used for this figure was different to that used in Figures 1 and 2: see Appendix C for details.

It is evident from Figure 3 that the coarse grained equilibrium SFS and the simulation results are, where the SFS is appreciable, very close to each other. This applies despite the very substantial difference between the census size ($N = 500$) and the effective population size ($N_e = 51$).

5.4 Application to the complex problem of three alleles

We shall apply the modified WFM to an extension of the Test Population to three alleles and shall determine some results for the probability of fixation, when $N_e \neq N$. The diffusion equation (of the diffusion approximation) is very hard or impossible to solve with more than two alleles. Thus prior to the present work, the only viable approach that could incorporate a nontrivial N_e was simulations.

We assume the three alleles have different selection coefficients but identical variances in the number of offspring their carriers produce, as shown in Table 2.

	mean No. of offspring	variance in No. of offspring
carrier of the A allele	$f \times (1 + s_A)$	$f^2\sigma^2$
carrier of the B allele	$f \times (1 + s_B)$	$f^2\sigma^2$
carrier of the C allele	$f \times (1 + 0)$	$f^2\sigma^2$

Table 2 Title: Basic statistics of reproductive outputs in a population with three alleles

Table 2 Caption: This table shows basic statistics of the reproductive outputs of different allele-carriers in the a population with three alleles. This population is a direct generalisation of the Test Population, to three alleles. The quantity f represents a baseline fertility, while s_A and s_B are, respectively, the selection coefficients of the A allele and the B allele, relative to the C allele, which has a vanishing selection coefficient ($s_C = 0$). All three alleles have the same variance in offspring number, which is taken to be $f^2\sigma^2$, where σ^2 is a constant.

Prior to giving any results, we note that apart from the modified WFM incorporating the effective population size, N_e, (unlike the standard WFM), the modified WFM for three alleles also has a lower complexity than the standard WFM. The complexity of the modified WFM relative to that of the standard WFM can be measured by the ratio of the number of elements in the transition matrix of the two models. When there are α distinct alleles, the ratio of the number of elements in the transition matrix of the modified WFM, compared to the number in the standard WFM, is $(N_e/N)^{2(\alpha-1)}$ (cf. Waxman 2009). Thus when there are three alleles ($\alpha = 3$) and the census and effective population sizes are $N = 100$ and $N_e = 20$, respectively, we have a ratio of $(N_e/N)^4 = 1/625$ indicating a significantly reduced complexity of the modified WFM[6].

For the three allele problem, random genetic drift, in the absence of mutation and migration, is somewhat different to that of the two allele problem. With three alleles, loss is not equivalent to fixation: if an allele is lost, the frequency of the other two alleles can still change; the three allele problem simply degenerates into a two allele problem. Thus loss is generally not associated with an absorbing state but fixation is. It follows that in a three allele problem there are a total of three absorbing states, representing fixation of each of the three alleles. We shall compare results for the probability of ultimate fixation, from the modified WFM and simulations. The expression for the required probability, from the modified WFM, takes a very similar form to that of Eq. (12), but the matrices that must be used arise from the 'higher-dimensional' three allele problem (Waxman 2009). Furthermore, the number n that appears in Eq. (12) must be replaced by

[6] This reduced complexity indicates there is a qualitative reduction in *computational* complexity of the modified WFM (as measured by number of elementary operations, or mean time of running of a program). A quantitative measure of the reduced computational complexity of the modified WFM depends on the quantity calculated. Restricting ourselves to quantities which just require matrix multiplication, the multiplication of two matrices of size n has a running time which scales as n^b with exponent $2 < b < 3$. For example, a fast multiplication algorithm leads to a running time which scales as $n^{2.807}$ (Strassen 1969) and more recent algorithms have yet smaller exponents. For the problem with α alleles, the computational complexity of the modified WFM, relative to the standard WFM, is $(N_e/N)^{b(\alpha-1)}$ with $2 < b < 3$. We generally conclude that the modified WFM leads to a reduced computational complexity, and it may be substantial.

an appropriate scalar index that corresponds to the initial numbers of all three alleles (Waxman 2009). In Figure 4 we illustrate results associated with the probability of ultimate fixation of the A allele.

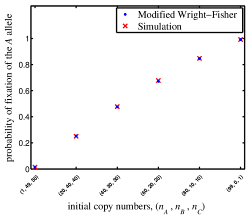

Figure 4 **Caption:** We extended the Test Population to three alleles that we have labelled A, B and C. The three alleles have different selection coefficients but the same variance in the number of offspring. The figure illustrates the probability of ultimate fixation of the A allele. For the figure, the census population size was $N = 100$, while other parameters, as described in Table 2, have the values: baseline fertility, $f = 100$; value of σ^2 for all three alleles, $\sigma^2 = 4$; selection coefficients of the three alleles, $s_A = 0.01$, $s_B = -0.01$ and $s_C = 0$. The effective population size that follows from Eq. (1), which was derived for the Test Population but also applies for the three allele model, is $N_e = 21$. We have written the initial copy numbers of the three alleles as (n_A, n_B, n_C); six different sets of initial copy numbers of the three alleles were used in the figure.

It is evident from Figure 4 that for the parameter set and initial copy-numbers adopted, the results following from the modified WFM are very close to the simulation results. As a quantitative illustration of this, we looked at the difference between the calculated and simulated values of the six fixation probabilities plotted in Figure 4. The maximum difference was found to be smaller than 2%.

6 Discussion

In this work we have presented a method of incorporating the effective population size into a Wright-Fisher model (WFM), but in a manner that also contains information on the census size, which also plays a key role, and cannot be ignored. We have called the resulting model a modified WFM, and have explicitly illustrated the method on 'non-ideal' haploid populations, where the effective and census population sizes are very different. However, as already pointed out, the closeness of the logic we employ, to that of the diffusion approximation, suggests that in all situations where the diffusion approximation works well, the modified WFM will also work well. Thus the modified WFM should have broad applicability and apply, for example, to diploid populations, as well as accommodating multiple alleles, multiple loci, structured populations etc.

The modified WFM allows a more efficient capturing of numerical results than e.g., solving the diffusion equation (which may be hard or impossible to carry out in complex problems). Importantly, we do not just gain computational advantages over a standard WFM: the results obtained apply under biologically more realistic assumptions. This is of particular interest for species for which the effective population size, N_e, differs substantially from the census size. Such a phenomenon is often observed in animal breeding and conservation biology (Charlesworth, 2009). It may also occur in species with complex eusocial behaviour, e.g. insects or rodents (Wilson and Hölldobler 2005; Jarvis 1981) and parasite related differences in sex ratios (Dyson et al. 2002). Differences between N_e and N are also observed in plants species, where the mode of inheritance may differ even between closely related species. Interestingly, selfing plant species (presumably with low N_e) show a larger geographic range distribution (presumably large N) than their outcrossing counterparts (Grossenbacher et al. 2015). These examples and numerous others illustrate the importance of incorporating N and N_e in a biological meaningful framework, when studying important ecological questions.

The human species provides an example of great interest where there is a dramatic difference between effective and census sizes. However the census size that is typically reported is not a local quantity, associated with processes in a particular generation, but rather a harmonic mean, that reflects a severe population bottleneck that the population went through, and from which the effective population size is now recovering (see e.g., Tenesa et al. 2007). Additionally, the census population size continues to increase, while the (local) effective size may exhibit a different rate of change, thus the situation is complex and does not simply warrant the incorporation of an N_e into a WFM without additional considerations. However, for specific models/behaviours of the time-dependent *local* effective population size, $N_e(t)$), we can employ the methodology of the present work. This will involve *rectangular* transition matrices. Suppose, in particular, that the effective population size changes at generation t_0, so that $N_e(t_0 + 1) \neq N_e(t_0)$. The relevant transition matrix, connecting generation t_0 to generation t_0+1, will, for a haploid population, have the size $(N_e(t_0 + 1) + 1) \times (N_e(t_0) + 1)$ (Eq. (8) is an example of a rectangular transition matrix). If there are no further changes in the effective population size, then all transition matrices after generation t_0+1 will be square and of size $(N_e(t_0 + 1) + 1)\times(N_e(t_0 + 1) + 1)$. If there is a *pattern* of effective population size changes, with discrete changes occurring at generations t_0, t_1, t_2, ..., then appropriate rectangular transition matrices need to be introduced into the dynamics at these times.

An interesting application of the method of this work is to the site frequency spectrum (SFS). This quantity can be used to obtain information about the selective effects of mutations segregating in a population (Keightley and Eyre-Walker 2007; Schneider et al. 2011). Even though some methods consider demographic events when estimating selective effects from the SFS, very little is known about how differences between effective and census population size systematically affect these estimates. For example, the SFS can be used to infer the amount of adaptive evolution in a McDonald-Kreitman type of test (McDonald and Kreitman 1991; Eyre-Walker and Keightley 2009) to deduce whether the (effective) population size is determining the rate of adaptive evolution across species (Gossmann et al. 2012). Therefore would the inclusion of both the effective population size, N_e, and the census population size, N, shed further light into this important ongoing debate (Venton 2012)?

A feature of the method of this work, that was explicitly exposed in the calculations of the site frequency spectrum, is that it leads to ‘coarse grained results'. Distributions and associated quantities are determined at the frequencies n/N_e (for a haploid population) with n an integer. The splitting of these frequencies, namely $1/N_e$, is wider (possibly substantially wider) than the splitting of the frequencies at which an exact calculation would yield, namely $1/N$. This has the implication that we cannot enquire into fine features of such distributions that might occur on scales comparable or smaller than $1/N_e$. This does not seem problematic, since if there are questions about such fine features existing, then calculations based on an effective population size may, themselves, be questionable without additional analysis.

6.1 Summary

In summary, we have provided a method that incorporates the effective population size into the Wright-Fisher model. This increases the biological realism of this model, and, importantly, is a viable way of obtaining numerical results. We have thus provided a tool that will allow new analyses to be systematically carried out, without the need of detailed simulations, or numerical solution of the diffusion equation.

Appendices

Appendix A: Calculating quantities in the modified Wright-Fisher model for the Test Population

In this appendix we use the modified Wright-Fisher model (WFM) to calculate the probabilities of ultimate fixation and loss, and the mean times to fixation and loss, for the Test Population.

To begin, we determine the effective distribution in different generations. From Eq. (6) of the main text we can show that the solution for $\mathbf{F}^{(e)}(t)$ is given by

$$\mathbf{F}^{(e)}(t) = \left(\mathbf{W}^{(e)}\right)^{t-1}\mathbf{W}^{(0)}\mathbf{F}(0), \qquad t = 1, 2, ... \tag{A1}$$

Using the block form for $\mathbf{W}^{(e)}$ (Eq. (10) of the main text) we obtain

$$\left(\mathbf{W}^{(e)}\right)^{t-1} = \begin{pmatrix} 1 & \mathbf{u}^{(e)} \sum_{k=0}^{t-2} \left(\mathbf{w}^{(e)}\right)^{k} & 0 \\ \mathbf{0}^{(e)} & \left(\mathbf{w}^{(e)}\right)^{t-1} & \mathbf{0}^{(e)} \\ 0 & \mathbf{v}^{(e)} \sum_{k=0}^{t-2} \left(\mathbf{w}^{(e)}\right)^{k} & 1 \end{pmatrix} \tag{A2}$$

with the understanding that $\sum_{k=0}^{t-2} \left(\mathbf{w}^{(e)}\right)^{k} = 0$ for $t = 1$. Combining this with the block form of $\mathbf{W}^{(0)}$ (Eq. (10) of the main text) leads to

$$\mathbf{F}^{(e)}(t) = \begin{pmatrix} 1 & \mathbf{u}^{(0)} + \mathbf{u}^{(e)} \sum_{k=0}^{t-2} \left(\mathbf{w}^{(e)}\right)^{k} \mathbf{w}^{(0)} & 0 \\ \mathbf{0}^{(e)} & \left(\mathbf{w}^{(e)}\right)^{t-1} \mathbf{w}^{(0)} & \mathbf{0}^{(e)} \\ 0 & \mathbf{v}^{(0)} + \mathbf{v}^{(e)} \sum_{k=0}^{t-2} \left(\mathbf{w}^{(e)}\right)^{k} \mathbf{w}^{(0)} & 1 \end{pmatrix} \mathbf{F}(0). \tag{A3}$$

Probabilities of ultimate loss and fixation of the A allele

For long time properties, we determine the $t \to \infty$ limit of the above equation with the result

$$\mathbf{F}^{(e)}(\infty) = \begin{pmatrix} 1 & \mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)} & 0 \\ \mathbf{0}^{(e)} & \mathbf{0} & \mathbf{0}^{(e)} \\ 0 & \mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)} & 1 \end{pmatrix} \mathbf{F}(0) \tag{A4}$$

where $\mathbf{G}^{(e)}$ is given in Eq. (11) of the main text and here $\mathbf{0}$ is a matrix of size $(N_e - 1) \times (N - 1)$ with all elements 0 that occurs because it can be argued that all eigenvalues of $\mathbf{w}^{(e)}$ have magnitude less than unity (see Appendix C of Waxman 2009).

The n'th element of $\mathbf{F}^{(e)}(t)$, namely $F_n^{(e)}(t)$, has the interpretation as the probability of occurrence of the frequency n/N_e at time t, with $n = 0, 1, ..., N_e$. This means that we can write equivalently write Eq. (A4) as

$$\begin{aligned} \begin{array}{l}\text{probability of ultimate} \\ \text{fixation of the } A \text{ allele}\end{array} &= F_{N_e}^{(e)}(\infty) \\ &= \begin{pmatrix} 1, & \mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}, & 0 \end{pmatrix} \mathbf{F}(0), \\ \begin{array}{l}\text{probability of ultimate} \\ \text{loss of the } A \text{ allele}\end{array} &= F_{0}^{(e)}(\infty) \\ &= \begin{pmatrix} 1, & \mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}, & 0 \end{pmatrix} \mathbf{F}(0). \end{aligned} \tag{A5}$$

Assuming there are n copies of the A allele present in generation 0, with $n = 1, 2, ..., N-1$, the vector $\mathbf{F}(0)$ has only one non-zero element, namely $F_n(0)$, which equals 1. Using the notation $[\mathbf{a}]_n$ to denote the n'th element of the row vector $\mathbf{a}$, we have, for

$$\begin{array}{l}\text{probability of ultimate fixation}\\ \text{of the } A \text{ allele when } n \text{ copies}\\ \text{are initially present}\end{array} = \left[\mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n,$$

$$\begin{array}{l}\text{probability of ultimate loss}\\ \text{of the } A \text{ allele when } n \text{ copies}\\ \text{are initially present}\end{array} = \left[\mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n. \tag{A6}$$

This pair of equations corresponds to Eqs. (12) and (13) of the main text.

Expected times to fixation and loss of the A allele

We shall focus just on the expected time to fixation, since the corresponding quantity for loss has a form which can be simply inferred.

Assuming there are n copies of the A allele present in generation 0, with $n = 1, 2, ..., N-1$, the vector $\mathbf{F}(0)$ has only one non-zero element, namely $F_n(0)$, which equals 1. The interpretation of the fixation part of Eq. (A3) is that $\mathbf{v}^{(0)}\mathbf{F}(0) = \left[\mathbf{v}^{(0)}\right]_n$ is the probability of fixation occurring precisely in generation 1, and similarly $\mathbf{v}^{(e)}\left(\mathbf{w}^{(e)}\right)^{t-2}\mathbf{w}^{(0)}\mathbf{F}(0) = \left[\mathbf{v}^{(e)}\left(\mathbf{w}^{(e)}\right)^{t-2}\mathbf{w}^{(0)}\right]_n$ is the probability of fixation precisely occurring in generation t for $t \geq 2$.

From these results, the mean time to fixation, conditional on fixation ultimately occurring, is written $E[T_{\text{fix}}|n]$ and given by

$$E[T_{\text{fix}}|n] = \frac{\left[\mathbf{v}^{(0)}\right]_n + \sum_{t=2}^{\infty} t \left[\mathbf{v}^{(e)} \left(\mathbf{w}^{(e)}\right)^{t-2} \mathbf{w}^{(0)}\right]_n}{\left[\mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\right]_n}. \tag{A7}$$

Evaluating the sum and simplifying the result quickly yields Eq. (14) of the main text. Replacing $\mathbf{v}$'s by $\mathbf{u}$'s in the result leads to the expected time to loss, conditional on loss ultimately occurring, and yields Eq. (15) of the main text.

Appendix B: Site frequency spectrum

In this appendix we give details of the calculation for the effective site frequency spectrum (SFS) and a coarse grained SFS using the method introduced in this work, in the context of a haploid population of census size N (number of adults).

In the lifecycle given in the main text, mutation has been assumed neglectable, because only a single locus was under consideration. This is not the case for the SFS, where the mutational target is an extended part of the genome. We thus include mutations which we take to occur in adults at the beginning of a generation. We use μ to denote the expected number of new mutations each generation. In terms of the scaled mutation rate $\theta = 2N\mu$ there are an expected number of $\theta/2$ mutations entering the adult population each generation.

We shall use $M_n(t)$ to denote the mean number of sites with mutant alleles at a frequency of n/N in generation t (equivalently, $M_n(t)$ denotes the mean number of sites with n mutant alleles in generation t). The SFS is the set of $M_n(t)$ for $n = 1, 2, ..., N-1$, i.e., it only includes sites where mutations are segregating in the population, and excludes sites where mutations have been lost or have not occurred.

The model of the SFS we consider is based on the assumption that there are an effectively infinite number of independent (unlinked) sites where mutations can occur, that is to say, the infinite sites model (Kimura 1969). A consequence of this assumption is that each site can, at most, suffer only one mutation; double mutations of a site are considered to happen with negligible probability.

When $N_e = N$, a standard Wright-Fisher model for a haploid population with census size N can be applied. The SFS obeys

$$M_n(t+1) = \sum_{m=1}^{N-1} w_{n,m} M_m(t) + \frac{\theta}{2}\delta_{n,1}, \tag{B1}$$

where the $w_{n,m}$ are elements of a submatrix $\mathbf{w}$ of the Wright-Fisher transition matrix which takes into account transitions between segregating states of the population (see Eq. (5) of the main text), and $\delta_{a,b}$ is a Kronecker delta ($\delta_{a,b}$ is 1 if $a = b$ and is 0 if $a \neq b$). The presence of the term $\frac{\theta}{2}\delta_{n,1}$ in Eq. (B1) reflects the assumption that new mutants originate as single copies in the population, at a rate of $\theta/2$ per generation. Equation (B1) can be written as the matrix equation

$$\mathbf{M}(t+1) = \mathbf{w}\mathbf{M}(t) + \frac{\theta}{2}\mathbf{i}, \tag{B2}$$

where both $\mathbf{M}(t)$ and $\mathbf{i}$ are column vectors of length $N-1$. The first element of $\mathbf{i}$ is 1 with all others being 0. From Eq. (B2) the equilibrium SFS, written $\hat{\mathbf{M}}$, is found to be

$$\hat{\mathbf{M}} = \frac{\theta}{2}\mathbf{G}\mathbf{i} \tag{B3}$$

where $\mathbf{G} = (\mathbf{I} - \mathbf{w})^{-1}$ and $\mathbf{I}$ is an identity matrix (the same size as $\mathbf{w}$).

In a 'non ideal' population, where $N_e < N$, the standard results, described above, cannot be directly used. We shall use a method associated with the modified WFM of the present work. As we shall see, this leads to a 'coarse grained' SFS which is defined only at the frequencies $x_n^{(e)} = n/N_e$ (with $n = 1, 2, .., N_e - 1$) rather than at the exact frequencies $x_n^{(0)} = n/N$ (with $n = 1, 2, ..., N - 1$). Because $N_e < N$ the exact frequencies, $x_n^{(0)}$, are more finely spaced than the $x_n^{(e)}$.

Note that the initial SFS $\mathbf{M}(0)$ (assumed known) and the contribution from new mutations, to the SFS, $\frac{\theta}{2}\mathbf{i}$, are both defined for the exact frequencies, $x_n^{(0)} = n/N$. Thus $\mathbf{M}(0)$ and $\frac{\theta}{2}\mathbf{i}$ are both column vectors of length $N - 1$. However, the SFS that is associated with a modified WFM, where the effective population size is N_e is, under the approach of this work, described as a column vector of length of $N_e - 1$. The difference in the lengths of the vectors of the SFS, of the actual model and the model following from the modified WFM, make it impossible to *directly* evolve the SFS, according to Eq. (B1). To overcome this, we decompose the value of $\mathbf{M}(t)$ in generation t ($t \geq 1$) into two contributions: (i) from sites where new mutations originated at the beginning of generation t, and (ii) from sites associated with mutations which originated in the previous or earlier generations. We write this decomposition as $\mathbf{M}(t) = \mathbf{M}^{new} + \mathbf{M}^{prev}(t)$. The form of $\mathbf{M}^{new}$ is known; it originates purely from new mutations and is a column vector of length $N - 1$ where only the first element is non-zero and has the value $\theta/2$.

Consider the part of the SFS associated with mutations which originated in the previous generation, and which have evolved for at least one generation in a manner appropriate to a population size of N_e. Under the approach of the present work we write this part of the SFS as $\mathbf{M}^{prev,e}(t)$. This is a column vector of length $N_e - 1$ whose n'th element may be approximately viewed as the mean number of sites corresponding to the mutant allele frequency lying in an interval of width $1/N_e$ in the vicinity of the frequency n/N_e (with $n = 1, 2, ..., N_e - 1$). Following the viewpoint of the present work, we take the behaviour of $\mathbf{M}^{prev,e}(t)$ to be given by

$$\begin{cases} \mathbf{M}^{prev,e}(1) = \mathbf{w}^{(0)}\mathbf{M}(0), \\ \\ \mathbf{M}^{prev,e}(t+1) = \mathbf{w}^{(e)}\mathbf{M}^{prev,e}(t) + \dfrac{\theta}{2}\mathbf{w}^{(0)}\mathbf{i}, \qquad t = 1, 2, ... \end{cases} \tag{B4}$$

An explanation of the various terms in Eq. (B4) is as follows. The quantity $\mathbf{w}^{(0)}$ is a rectangular matrix that 'converts' the segregating part of a definitely known distribution in a generation where the population size is N, to the corresponding *effective distribution* in the next generation (see Eq. (10) of the main text), where the population size is treated as N_e. Thus $\mathbf{w}^{(0)}\mathbf{M}(0)$ reflects the conversion of the known quantity $\mathbf{M}(0)$, where the population size is N, to the next generation, where the population size is treated as N_e. The quantity $\mathbf{w}^{(e)}$ is a square matrix that takes the segregating part of the distribution of a population in any generation where the population size is treated as N_e, and yields the corresponding distribution in the next generation, where the population size is also treated as N_e (see Eq. (10) of the main text). Thus $\mathbf{w}^{(e)}\mathbf{M}^{prev,e}(t)$ represents the part of $\mathbf{M}^{prev,e}(t+1)$ that was contributed by mutations prior to generation t, while $\frac{\theta}{2}\mathbf{w}^{(0)}\mathbf{i}$ represents new mutations that occurred at the beginning of generation t, whose contribution is 'converted' to generation $t + 1$.

The equilibrium form of $\mathbf{M}^{prev,e}(t)$ that follows from Eq. (B4) is written as $\hat{\mathbf{M}}^{prev,e}$ and given by

$$\hat{\mathbf{M}}^{prev,e} = \frac{\theta}{2}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\mathbf{i}, \tag{B5}$$

where $\mathbf{G}^{(e)}$ is given in Eq. (11) of the main text.

Now consider the mean number of sites in a frequency range δx around a frequency x. We assume δx is small in value ($\ll 1$) but still large compared with $1/N$ and $1/N_e$, so it covers many frequency states. Furthermore, assume that we can approximately write x as either n/N or m/N_e where n and m are integers. Then the mean number of sites whose mutant frequency lies in the frequency range δx around the frequency x is given (approximately) by either adding $\frac{\delta x}{1/N} = N\delta x$ adjacent elements of the exact SFS M_n^{prev}, or adding (the smaller number of) $\frac{\delta x}{1/N_e} = N_e \delta x$ adjacent elements of the effective spectrum $M_m^{prev,e}$. That is, we approximately have $N\delta x M_n^{prev} = N_e \delta x M_m^{prev,e}$. This tells us that M_n^{prev} and $M_m^{prev,e}$ are not of the same magnitude, but are related by $M_n^{prev} = (N_e/N) \times M_m^{prev,e}$. To obtain an approximate quantity that should be directly comparable with the exact spectrum we shall generally define

$$\mathbf{M}^{prev,cg}(t) = \frac{N_e}{N}\mathbf{M}^{prev,e}(t) \tag{B6}$$

and call $\mathbf{M}^{prev,cg}(t)$ the *coarse grained* SFS. The quantity $\mathbf{M}^{prev,cg}(t)$ corresponds to the frequencies $x_n^{(e)} = n/N_e$, which have splittings of $1/N_e$, that are larger than the splittings of the exact SFS (which is defined at the frequencies $x_n^{(0)} = n/N$) and hence have splittings of $1/N$. However, the magnitude of $\mathbf{M}^{prev,cg}(t)$ should closely correspond to the magnitude of the exact SFS, when evaluated at a common frequency. We define the *coarse grained* equilibrium SFS as

$$\hat{\mathbf{M}}^{prev,cg} = \frac{N_e}{N}\hat{\mathbf{M}}^{prev,e} = \frac{\theta_e}{2}\mathbf{G}^{(e)}\mathbf{w}^{(0)}\mathbf{i}, \tag{B7}$$

where $\theta_e = \frac{N_e}{N}\theta = 2N_e\mu$ and this appears to work well (see Figure 4 of the main text).

Appendix C: Simulation procedure for the site frequency spectrum

In this appendix we give details of the simulation procedure adopted to estimate the SFS.

To simulate the SFS, we could use the method introduced in the main text, for the Test Population, with the added feature that a random number of new mutations are introduced in adults at the beginning of each generation. However, to determine the equilibrium SFS requires a number of generations to 'forget' the initial distribution ('burn in' time). Furthermore, to obtain a smooth result requires averaging the resulting fluctuating spectrum over a very large number of generations (an alternative is carry out an average over many replicate populations), and this will cost a large computation time or require a large computer memory. We adopt the following alternative approach.

We note that once a mutant occurs at one site, it will evolve according to a time homogeneous Markov chain, even though this will not be a standard WFM, because $N_e < N$. To obtain a simulation result to compare with the modified WFM result, and is independent, we use a one step simulation (following the simulation described in main text) to estimate the transition matrix of this Markov chain model. This involves using a total of R trajectories of the mutants, with the same initial copy number. These trajectories are run for a single generation, which includes a random reproductive stage (negative binomial random variables are used) and a random thinning stage, both of which are described in the Section *Test Population* of the main text. Let $C_j(N, s, n)$ represent the copy number of mutants after one generation of trajectory j ($j = 1, 2, ..., R$), when the initial copy number of mutants is n, the census population size is N, and the selection coefficient of the mutant is s. We use the simulated values of the $C_j(N, s, n)$ to estimate elements of the transition matrix of this Markov chain. Writing this estimated transition matrix as $\widetilde{\mathbf{W}}$ we have

$$\widetilde{W}_{m,n} = \frac{\sum_{j=1}^{R} \delta_{m,C_j(N,s,n)}}{R}. \tag{C1}$$

This result is determined for $m = 0, 1, 2, ..., N$ and for $n = 1, 2, ..., N - 1$, while we determine the remaining elements using $\widetilde{W}_{m,0} = \delta_{m,0}$ and $\widetilde{W}_{m,1} = \delta_{m,1}$.

The matrix $\widetilde{\mathbf{W}}$ has the same general structure as the matrix $\mathbf{W}$ of Eq. (5) of the main text, namely,

$$\widetilde{\mathbf{W}} = \begin{pmatrix} 1 & \widetilde{\mathbf{u}} & 0 \\ \mathbf{0} & \widetilde{\mathbf{w}} & \mathbf{0} \\ 0 & \widetilde{\mathbf{v}} & 1 \end{pmatrix}. \tag{C2}$$

It also has a size of $(N + 1) \times (N + 1)$, irrespective of the value of N_e. In a standard WFM, the equilibrium SFS is given by Eq. (B3), and analogously, the result of the above procedure leads to an estimate of the equilibrium

SFS of

$$\widehat{\widetilde{\mathbf{M}}} = \frac{\theta}{2}\widetilde{\mathbf{G}}\mathbf{i}, \tag{C3}$$

where $\widetilde{\mathbf{G}} = \left(\widetilde{\mathbf{I}} - \widetilde{\mathbf{w}}\right)^{-1}$ and $\widetilde{\mathbf{I}}$ is an identity matrix (the same size as $\widetilde{\mathbf{w}}$).

Finally, using the identity $\widehat{\widetilde{\mathbf{M}}} = \widetilde{\mathbf{w}}\widehat{\widetilde{\mathbf{M}}} + \frac{\theta}{2}\mathbf{i}$, and following the definition in Appendix B of the equilibrium form of the SFS from 'previous' mutations, which we write as $\widehat{\widetilde{\mathbf{M}}}^{prev}$, we obtain $\widehat{\widetilde{\mathbf{M}}}^{prev} = \widetilde{\mathbf{w}}\widehat{\widetilde{\mathbf{M}}}$. This can be written as

$$\widehat{\widetilde{\mathbf{M}}}^{prev} = \frac{\theta}{2}\widetilde{\mathbf{G}}\widetilde{\mathbf{w}}\mathbf{i}. \tag{C4}$$

and is the result we use in Figure (3) of the main text, as our estimate from simulations. The value of R used for the figure was $R = 10^5$.

In the Supplementary Material we provide a Matlab function which is a generalised version of the function $C_j(N, s, n)$ used above.

References

Anderson, J. H., E. J. Ward, and S. M. Carlson, 2011 A model for estimating the minimum number of offspring to sample in studies of reproductive success. J. Hered. 105: 567–576.

Charlesworth, B., 2009 Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10: 195–205.

Dyson, E. A., M. K. Kamath and G. D. D. Hurst, 2002 Wolbachia infection associated with all-female broods in Hypolimnas bolina (Lepidoptera: Nymphalidae): evidence for horizontal transmission of a butterfly male killer. Heredity (Edinb.), 88: 166–171.

Evans, S. N., Y. Shvets, and M. Slatkin, 2007 Non-equilibrium theory of the allele frequency spectrum. Theor. Popul. Biol. 71: 109–119.

Ewens, W. J., 1964 The pseudo-transient distribution and its uses in genetics. J. App. Prob. 1: 141-156.

Ewens, W. J., 2004 *Mathematical population genetics I. Theoretical introduction.* 2nd edition. Springer-Verlag, New York.

Eyre-Walker, A. and P. D. Keightley, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108.

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

Gillespie, J. H., 1974 Natural selection for within-generation variance in offspring number. Genetics 76: 601-606.

Gillespie, J. H., 1975 Natural selection for within-generation variance in offspring number II. discrete haploid models. Genetics 81: 403-413.

Gossmann, T., P. D. Keightley and A. Eyre-Walker, 2012 The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. 4: 658–667.

Grant, P. R., and B. R. Grant, 2000 Non-random fitness variation in two populations of Darwin's finches. Proc. R. Soc. Lond. B. Biol. Sci. 267:131–138.

Grossenbacher, D., R. B. Runquist, E. E. Goldberg and Y. Brandvain, 2015 Geographic range size is predicted by plant mating system. Ecol Lett. 18: 706–713

Hui, T.-Y. J. and A. Burt 2015 Estimating effective population size from temporally spaced samples with a novel, efficient maximum-likelihood algorithm. Genetics 200: 285-293.

Keightley, P. D. and A. Eyre-Walker, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177: 2251–2261.

Jarvis, J. U., 1981 Eusociality in a mammal: cooperative breeding in naked mole-rat colonies, Science 212: 571–573.

Johnson, N. L., S. Kotz and A. W. Kemp 2005 *Univariate discrete distributions*. 3rd Edition. Wiley Series in Probability and Statistics. John Wiley and Sons Ltd, Hoboken.

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

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713-719.

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

Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics 61: 763-771.

McDonald J. H. and M. Kreitman, 1991 Adaptive evolution at the Adh locus in Drosophila. Nature 351: 652–654

Melbourne, B. A., and A. Hastings, 2008 Extinction risk depends strongly on factors contributing to stochasticity. Nature 454: 100-103.

Reiss, J. O., 2013 Does selection intensity increase when populations decrease? Absolute fitness, relative fitness, and the opportunity for selection. Evol. Ecol. 27: 477–488.

Schneider, A., B. Charlesworth, A. Eyre-Walker and P. D. Keightley, 2011 A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189: 1427–1437.

Strassen, V., 1969 Gaussian Elimination is not Optimal. Numer. Math. 13: 354-356.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard and P. M. Visscher 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Thépot, S., G. Restoux, I. Goldringer, F. Hospital, D. Gouache, I. Mackay, and J. Enjalbert, 2015 Efficiently Tracking Selection in a Multiparental Population: The Case of Earliness in Wheat. Genetics 199: 609-623.

Venton, D., 2012 Highlight–tracking adaptation's role: do larger populations evolve faster?. Genome Biol. Evol. 4: 668–669.

Waxman, D., 2009 Fixation at a Locus with Multiple Alleles: Structure and Solution of the Wright Fisher Model. J. Theor. Biol. 257: 245-251.

Waxman, D., 2011 Comparison and content of the Wright-Fisher model of random genetic drift, the diffusion approximation, and an intermediate model. J. Theor. Biol. 269: 79-87.

Waxman, D., 2012 Population growth enhances the mean fixation time of neutral mutations and the persistence of neutral variation. Genetics 191: 561–577.

Wilson E. O., and B. Hölldobler 2005 Eusociality: origin and consequences. Proc Natl Acad Sci U. S. A, 102: 13367–13371.

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

Wright, S., 1945 Tempo and mode in evolution: a critical review. Ecology 26: 415-419.

Zhao, L., X. Y. Yue, and D. Waxman, 2013 Complete numerical solution of the diffusion equation of random genetic drift. Genetics 194: 973–985.

Supplementary Material

A modified Wright-Fisher model that incorporates N_e: A variant of the standard model with increased biological realism and reduced computational complexity

Lei Zhao, Toni Gossmann, and David Waxman

On the following page we give a Matlab function C.m that generates the number of copies of the mutant allele after a single generation, as described in Appendix C. Repeated use of this function allows estimation of the transition matrix for the Test Population.

We use the abbreviation NBD for the negative binomial distribution.

C.m

```
function n=C(N,n1,s1,s2,sigma1,sigma2)
% Simulates copy number of the mutant allele for a haploid population
% INPUTS:
% N: census population size
% n1: copy number of the mutant allele in the current generation
% s1 and s2: selection coefficients of mutant and resident alleles, respectively
% sigma1 and sigma2: variances in offspring No. of mutant and
% resident alleles, respectively
% OUTPUT:
% n: copy No. of mutant allele one generation after the current generation
% CALCULATION
f=100;                          % baseline fertility, taken as a constant
n2=N-n1;                        % copy No. of resident alleles in current generation
m1 = f*(1+s1);                  % mean No. offspring of a carrier of the mutant allele
m2=f*(1+s2);                    % mean No. offspring of a carrier of the resident allele
v1=f^2*sigma1;                  % variance offspring No. of a carrier of the mutant allele
v2=f^2*sigma2;                  % variance offspring No. of a carrier of the resident allele
p1=m1/v1;                       % parameter p of NBD for mutant alleles
p2=m2/v2;                       % parameter p of NBD for resident alleles
r1=p1/(1-p1)*m1;                % parameter r of NBD for mutant alleles
r2=p2/(1-p2)*m2;                % parameter r of NBD for resident alleles
n1=nbinrnd(n1*r1,p1);           % NBD offspring No.
n2=nbinrnd(n2*r2,p2);           % NBD offspring No.
n=binornd(N,n1./(n1+n2));       % Thinning: copy No. mutant alleles in next generation
```