



This is a repository copy of *A modified Wright-Fisher model that incorporates Ne: A variant of the standard model with increased biological realism and reduced computational complexity.*

White Rose Research Online URL for this paper:  
<http://eprints.whiterose.ac.uk/96384/>

Version: Accepted Version

---

**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>

---

Article available under the terms of the CC-BY-NC-ND licence  
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

1

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

5             Lei Zhao<sup>1</sup>, Toni I. Gossmann<sup>2</sup>, and David Waxman<sup>1</sup>

6     <sup>1</sup>Centre for Computational Systems Biology, Fudan University, Shanghai  
7     200433, People's Republic of China

8     <sup>2</sup>Department of Animal and Plant Sciences, University of Sheffield, Sheffield  
9     S10 2TN, United Kingdom

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

12    **Correspondence to:**

13    Professor D. Waxman

14    Centre for Computational Systems Biology, Fudan University, 220 Handan  
15    Road, Shanghai 200433, People's Republic of China.

16    E-mail: davidwaxman@fudan.edu.cn

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

**ABSTRACT**

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

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

## 50 **1 Introduction**

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

63 The first drawback is that the WFM explicitly depends on only one  
64 population size, namely the number of adults present in the population. This  
65 is a quantity we term the *census size*, and denote by  $N$ . Following Wright  
66 and many subsequent authors, it is recognised that biological populations  
67 can behave, in important aspects, as though their size is not the actual  
68 number of adults,  $N$ , but rather a different, typically smaller value,  $N_e$ ,  
69 that is termed the effective population size (Wright, 1931). The effective  
70 size usually arises from a population deviating, in one or more ways, from  
71 being ‘ideal’, such as when individuals do not have a Poisson distributed  
72 number of offspring, or related individuals interbreed, or when populations  
73 show age, stage and spatial structures (Charlesworth 2009). A possible way  
74 to account for a population behaving as if its size is  $N_e$  *appears to be* to

75 simply replace  $N$  by  $N_e$  in, for example, the WFM<sup>1</sup>. However, this does not  
76 directly work, as we show below.

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

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

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

---

<sup>1</sup>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.

99 lation, of census size  $N$ , in an initial state where a single adult carries a focal  
100 allele. The initial frequency of the focal allele is, non-negotiably,  $1/N$ . If  
101 we simply replace the population size,  $N$ , by the effective size,  $N_e$ , then the  
102 smallest non-zero frequency becomes  $1/N_e$ . The effective size is generally  
103 smaller than the census size ( $N_e < N$ ) hence the value of  $1/N_e$  is generally  
104 larger than the smallest non-zero frequency of the actual population ( $1/N$ ),  
105 possibly much larger. For example, if  $N = 1000$  and  $N_e = 100$  then we have  
106  $1/N_e = 10^{-2}$  which is 10 times the value of  $1/N = 10^{-3}$ . Thus, whatever  
107 else that naive replacement of  $N$  by  $N_e$  does, any result for an actual initial  
108 frequency of  $1/N$ , can, at best after the replacement, only be determined  
109 by the smallest non-zero initial frequency of  $1/N_e$  with generally erroneous  
110 results. This problem of mismatch of frequencies in populations of size  $N$   
111 and  $N_e$  is more general than just for the smallest non-zero frequency, and  
112 holds for many other frequencies.

113 The frequency mismatch problem, just described, is evaded under a well-  
114 known approximation of the WFM, namely the diffusion approximation  
115 (Fisher 1922; Wright 1945; Kimura 1955). This is an approximation that  
116 takes both the census size of the population,  $N$ , and the effective population  
117 size,  $N_e$ , into account. The approximation involves a diffusion equation for  
118 the distribution of an allele's frequency (hence the approximation's name),  
119 and has two important features. The first feature is that  $N_e$  takes the place  
120 of  $N$  in the diffusion equation. This means the dynamics of the frequency is  
121 treated as if the population had a census size of  $N_e$ , in accordance with the  
122 general idea behind the effective population size. The diffusion approxima-  
123 tion has a second feature that it treats an allele's frequency as a continuous  
124 quantity. This means that the initial frequency can be chosen to be *any*  
125 *value*. Accordingly, when initially there is, e.g., a single copy of an allele in

126 a population of census size  $N$ , the initial frequency can be chosen to be the  
127 *correct value*, namely  $1/N$ , irrespective of the value of  $N_e$ . In practice the  
128 above two features of the diffusion approximation generally work well to-  
129 gether, to the extent that the diffusion approximation can determine many  
130 properties to good accuracy even for relatively small populations (Ewens  
131 1964).

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

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



## 150 2 Principle

151 The simplest statement of the principle amounts to saying that we should  
152 treat the population as though it has a census (or actual) size of  $N$  when  
153 the copy number of an allele is definitely known, e.g., when a mutation first  
154 appears in a population, but in all subsequent generations, the dynamics  
155 of the allele's frequency behaves as if the actual population size were  $N_e$ .

156 The previous sentence is theoretically equivalent to saying that the popula-  
157 tion size discontinuously changes from  $N$ , in the generation where the copy  
158 number is definitely known, to the size  $N_e$  in the next generation – and all  
159 subsequent generations<sup>2</sup>. This viewpoint, of a discontinuous change of the  
160 population size from  $N$  to  $N_e$  is also a possible interpretation of solutions  
161 of the diffusion equation, where the frequency that is used at an initial time  
162 is correct for a population of size  $N$ , but the subsequent dynamics of the  
163 allele is treated as though the population has a census size of  $N_e$ . As a con-  
164 sequence, the principle we are proposing, to incorporate  $N_e$  into the WFM,  
165 is expected to hold to good accuracy in all of the circumstances where the  
166 diffusion approximation holds to good accuracy.

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

---

<sup>2</sup>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.

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

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

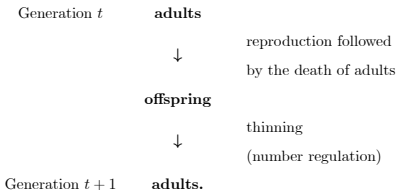
### 185 **3 Test Population**

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

---

<sup>3</sup>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  $t+1, t+2, \dots$  are the local values appropriate to these generations, i.e.,  $N_e(t+1), N_e(t+2), \dots$ .

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



193 We neglect the occurrence of mutations and assume there are reproductive  
 194 differences of carriers of the different alleles, as given in Table 1, which was  
 195 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	$f$	$f^2 \sigma^2$

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

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

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

204

---

205 We interpret the scheme in Table 1 as fertility selection (but see Gillespie  
206 1975), where the  $A$  allele has a selective advantage of  $s$  relative to the  $B$   
207 allele.

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

212 Within the lifecycle, thinning of the population to  $N$  individuals is non-  
213 selectively carried out according to sampling with replacement, i.e., ‘bino-  
214 mial sampling’, as used in the standard WFM.<sup>4</sup>

215 The above specification of a population is, of course, incomplete; a com-  
216 plete description includes the actual distribution of offspring numbers pro-  
217 duced by an adult of the population. While there are many possible distri-  
218 butions that could serve for this purpose, representing different biological  
219 situations, the distribution of offspring numbers we choose is the *negative*  
220 *binomial distribution* (see e.g., Johnson et al. 2005). This is a convenient  
221 and not unreasonable choice. The negative binomial distribution has a vari-  
222 ance that is generally larger than its mean, but has a Poisson distribution

---

<sup>4</sup>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.

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

233 The above constitutes a complete description of the Test Population.

### 234 **3.1 Properties of the Test Population**

235 We note that as the parameter  $\sigma^2$  approaches zero and  $f$  approaches infinity,  
236 the Test Population can be described by a standard WFM where  $N_e = N$   
237 and the  $A$  allele has a selective advantage of  $s$  over the  $B$  allele. However  
238 generally, the Test Population cannot be described by a standard WFM.  
239 Applying the analysis of Gillespie (1974; 1975), suggests that the Test Pop-  
240 ulation is equivalent, under a diffusion approximation, to a population where  
241 the selection coefficient of the  $A$  allele is replaced by an effective value that  
242 may be frequency-dependent, and the census size of the population is re-  
243 placed by an effective size that may also be frequency dependent. We shall  
244 assume that: (i) the  $A$  allele's selection coefficients is small,  $|s| \ll 1$ ; (ii) the  
245 population size is large,  $N \gg 1$ ; (iii) the baseline fertility is large,  $f \gg 1$ ;  
246 (iv) the parameter  $\sigma^2$  is much smaller than  $N$ ,  $\sigma^2 \ll N$ . We will work in  
247 the framework of the reproductive scheme in Table 1, combined with the

248 thinning process of the Test Population. We then find the following (cf.  
249 Gillespie 1974; 1975): (i) apart from small corrections of order  $s\sigma^2/N$ , the  
250 effective selection coefficient is frequency independent and has the value  $s$   
251 (see Table 1); (ii) apart from small corrections of order  $[f(1+\sigma^2)]^{-1}$ , the  
252 effective population size is also frequency independent and given by

$$N_e = \frac{N + \sigma^2}{1 + \sigma^2}. \quad (1)$$

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

256 We can summarise the Test Population by saying it has a census size of  $N$   
257 and, emerging from individual reproduction and thinning of the population,  
258 it has selection of strength  $s$ , and an effective population size given by Eq.  
259 (1).

## 260 **4 Applying the modified Wright-Fisher model to** 261 **the Test Population**

### 262 **4.1 Standard results of a Wright-Fisher model**

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

$$\mathbf{F}(t+1) = \mathbf{WF}(t) \quad (2)$$

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

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

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

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

$$D(x) = x + sx(1-x) \quad (4)$$

279 (see e.g., Ewens 2004).

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

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

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

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

## 289 4.2 Modified Wright-Fisher model for the Test Population

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

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

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

---

<sup>5</sup>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  $\mathbf{u} = (W_{0,1}, W_{0,2}, W_{0,3}, \dots)$ . We shall also write this vector as  $\mathbf{u} = (u_1, u_2, u_3, \dots)$ . 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  $n = 1, 2, \dots$ .



given by

$$\mathbf{F}^{(e)}(1) = \mathbf{W}^{(0)}\mathbf{F}(0) \tag{6}$$

$$\mathbf{F}^{(e)}(t+1) = \mathbf{W}^{(e)}\mathbf{F}^{(e)}(t), \quad t = 1, 2, \dots$$

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

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

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

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

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

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

311 with  $m = 0, 1, \dots, N_e$  and  $n = 0, 1, \dots, N$ , and

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

312 with  $m$  and  $n = 0, 1, \dots, N_e$ .

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

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

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

## 320 5 Illustrative examples involving the Test Popula- 321 tion

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

329 **5.1 Probabilities of fixation and loss**

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

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

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

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

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

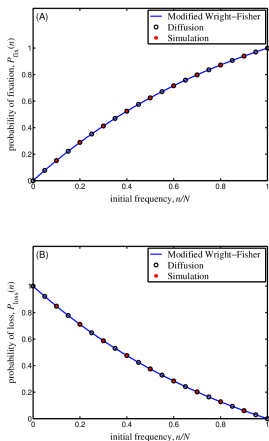
339 and

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

340 (see Appendix A for details).

341 Note that when  $N_e = N$ , and  $\mathbf{G}^{(e)}$  becomes  $\mathbf{G} = (\mathbf{I} - \mathbf{w})^{-1}$ , Eqs. (12)  
 342 and (13) reduce to  $P_{\text{fix}}(n) = [\mathbf{v}\mathbf{G}]_n$  and  $P_{\text{loss}}(n) = [\mathbf{u}\mathbf{G}]_n$  (cf. Waxman  
 343 2009).

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



347 **Figure 1 Caption:** This figure gives results, from three  
 348 different calculational methods, for the probability of ultimate  
 349 fixation of the  $A$  allele as a function of initial frequency (Panel  
 350 A), and the probability of loss of the  $A$  allele as a function of  
 351 initial frequency (Panel B). The three methods are: (i) the mod-  
 352 ified Wright-Fisher model, which was introduced in this work,  
 353 (ii) the diffusion approximation, (Kimura 1962) and (iii) simu-

354 lation. The parameter values adopted were: census population  
 355 size,  $N = 500$ ; selection coefficient of the  $A$  allele relative to the  
 356  $B$  allele,  $s = 0.01$ ; baseline fertility,  $f = 100$ ; value of  $\sigma^2$  (related  
 357 to the variance in offspring number of an adult - see Table 1),  
 358  $\sigma^2 = 9$ . For the simulations we used  $10^5$  replicate populations.  
 359 The approximate value of the effective population size that fol-  
 360 lows from these parameters is  $N_e = 51$ , see Eq. (1).

361

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

## 367 5.2 Mean times to fixation and loss

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

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

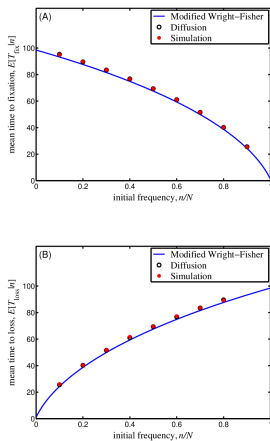
373 and

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

374 (see Appendix A for details).

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



377 **Figure 2 Caption:** This figure gives results, from three dif-  
378 ferent calculational methods, for the mean time to fixation of

379 an  $A$  allele as a function of initial frequency (Panel A), and  
380 the mean time to loss of an  $A$  allele as a function of initial  
381 frequency (Panel B). The three methods are: (i) the modified  
382 Wright-Fisher model, which was introduced in this work, (ii) the  
383 diffusion approximation (Kimura and Ohta 1969) and (iii) sim-  
384 ulation. The parameter values adopted were: census population  
385 size,  $N = 500$ ; selection coefficient of the  $A$  allele relative to the  
386  $B$  allele,  $s = 0.01$ ; baseline fertility,  $f = 100$ ; value of  $\sigma^2$  (re-  
387 lated to the variance in offspring number of an adult - see Table  
388 1),  $\sigma^2 = 9$ . For the simulations,  $10^5$  replicate populations were  
389 used. The approximate value of the effective population size that  
390 follows from these parameters is  $N_e = 51$ , see Eq. (1).

391

---

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

### 396 5.3 Site frequency spectrum

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

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

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

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

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



429 one generation.

### 430 5.3.1 Coarse grained equilibrium site frequency spectrum

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

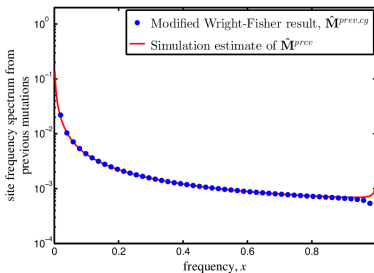
439 While the effective version of  $\hat{\mathbf{M}}^{prev}$ , written as  $\hat{\mathbf{M}}^{prev,e}$ , can be used to  
 440 approximate properties of the exact SFS, the values of  $\hat{\mathbf{M}}^{prev}$  and  $\hat{\mathbf{M}}^{prev,e}$   
 441 are not directly comparable. Each element of the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

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

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

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

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



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

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

475

---

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

#### 480 **5.4 Application to the complex problem of three alleles**

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

487 We assume the three alleles have different selection coefficients but iden-  
488 tical variances in the number of offspring their carriers produce, as shown  
489 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  $\sigma^2$  is a constant.

501

502 Prior to giving any results, we note that apart from the modified WFM  
 503 incorporating the effective population size,  $N_e$ , (unlike the standard WFM),  
 504 the modified WFM for three alleles also has a lower complexity than the  
 505 standard WFM. The complexity of the modified WFM relative to that of  
 506 the standard WFM can be measured by the ratio of the number of elements  
 507 in the transition matrix of the two models. When there are  $\alpha$  distinct alleles,

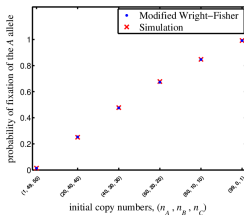
508 the ratio of the number of elements in the transition matrix of the modified  
509 WFM, compared to the number in the standard WFM, is  $(N_e/N)^{2(\alpha-1)}$  (cf.  
510 Waxman 2009). Thus when there are three alleles ( $\alpha = 3$ ) and the census  
511 and effective population sizes are  $N = 100$  and  $N_e = 20$ , respectively, we  
512 have a ratio of  $(N_e/N)^4 = 1/625$  indicating a significantly reduced complex-  
513 ity of the modified WFM<sup>6</sup>.

514 For the three allele problem, random genetic drift, in the absence of mu-  
515 tation and migration, is somewhat different to that of the two allele problem.  
516 With three alleles, loss is not equivalent to fixation: if an allele is lost, the  
517 frequency of the other two alleles can still change; the three allele prob-  
518 lem simply degenerates into a two allele problem. Thus loss is generally  
519 not associated with an absorbing state but fixation is. It follows that in a  
520 three allele problem there are a total of three absorbing states, represent-  
521 ing fixation of each of the three alleles. We shall compare results for the  
522 probability of ultimate fixation, from the modified WFM and simulations.  
523 The expression for the required probability, from the modified WFM, takes  
524 a very similar form to that of Eq. (12), but the matrices that must be used  
525 arise from the ‘higher-dimensional’ three allele problem (Waxman 2009).  
526 Furthermore, the number  $n$  that appears in Eq. (12) must be replaced by

---

<sup>6</sup>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  $\alpha$  alleles, the computational complexity of the modified WFM, relative to the standard WFM, is  $(N_e/N)^{2(\alpha-1)}$  with  $2 < b < 3$ . We generally conclude that the modified WFM leads to a reduced computational complexity, and it may be substantial.

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



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

543 of the three alleles were used in the figure.

544

---

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

## 551 **6 Discussion**

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

564 The modified WFM allows a more efficient capturing of numerical results  
565 than e.g., solving the diffusion equation (which may be hard or impossible

566 to carry out in complex problems). Importantly, we do not just gain com-  
567 putational advantages over a standard WFM: the results obtained apply  
568 under biologically more realistic assumptions. This is of particular interest  
569 for species for which the effective population size,  $N_e$ , differs substantially  
570 from the census size. Such a phenomenon is often observed in animal breed-  
571 ing and conservation biology (Charlesworth, 2009). It may also occur in  
572 species with complex eusocial behaviour, e.g. insects or rodents (Wilson  
573 and Hölldobler 2005; Jarvis 1981) and parasite related differences in sex ra-  
574 tios (Dyson et al. 2002). Differences between  $N_e$  and  $N$  are also observed  
575 in plants species, where the mode of inheritance may differ even between  
576 closely related species. Interestingly, selfing plant species (presumably with  
577 low  $N_e$ ) show a larger geographic range distribution (presumably large  $N$ )  
578 than their outcrossing counterparts (Grossenbacher et al. 2015). These  
579 examples and numerous others illustrate the importance of incorporating  
580  $N$  and  $N_e$  in a biological meaningful framework, when studying important  
581 ecological questions.

582 The human species provides an example of great interest where there  
583 is a dramatic difference between effective and census sizes. However the  
584 census size that is typically reported is not a local quantity, associated with  
585 processes in a particular generation, but rather a harmonic mean, that re-  
586 flects a severe population bottleneck that the population went through, and  
587 from which the effective population size is now recovering (see e.g., Tenesa  
588 et al. 2007). Additionally, the census population size continues to increase,  
589 while the (local) effective size may exhibit a different rate of change, thus the  
590 situation is complex and does not simply warrant the incorporation of an  $N_e$   
591 into a WFM without additional considerations. However, for specific mod-  
592 els/behaviours of the time-dependent *local* effective population size,  $N_e(t)$ ,



593 we can employ the methodology of the present work. This will involve  
594 *rectangular* transition matrices. Suppose, in particular, that the effective  
595 population size changes at generation  $t_0$ , so that  $N_e(t_0 + 1) \neq N_e(t_0)$ . The  
596 relevant transition matrix, connecting generation  $t_0$  to generation  $t_0+1$ , will,  
597 for a haploid population, have the size  $(N_e(t_0 + 1) + 1) \times (N_e(t_0) + 1)$  (Eq.  
598 (8) is an example of a rectangular transition matrix). If there are no further  
599 changes in the effective population size, then all transition matrices after  
600 generation  $t_0+1$  will be square and of size  $(N_e(t_0 + 1) + 1) \times (N_e(t_0 + 1) + 1)$ .  
601 If there is a *pattern* of effective population size changes, with discrete changes  
602 occurring at generations  $t_0, t_1, t_2, \dots$ , then appropriate rectangular transition  
603 matrices need to be introduced into the dynamics at these times.

604 An interesting application of the method of this work is to the site fre-  
605 quency spectrum (SFS). This quantity can be used to obtain information  
606 about the selective effects of mutations segregating in a population (Keight-  
607 ley and Eyre-Walker 2007; Schneider et al. 2011). Even though some meth-  
608 ods consider demographic events when estimating selective effects from the  
609 SFS, very little is known about how differences between effective and cen-  
610 sus population size systematically affect these estimates. For example, the  
611 SFS can be used to infer the amount of adaptive evolution in a McDonald-  
612 Kreitman type of test (McDonald and Kreitman 1991; Eyre-Walker and  
613 Keightley 2009) to deduce whether the (effective) population size is deter-  
614 mining the rate of adaptive evolution across species (Gossmann et al. 2012).  
615 Therefore would the inclusion of both the effective population size,  $N_e$ , and  
616 the census population size,  $N$ , shed further light into this important ongoing  
617 debate (Venton 2012)?

618 A feature of the method of this work, that was explicitly exposed in  
619 the calculations of the site frequency spectrum, is that it leads to ‘coarse

620 grained results'. Distributions and associated quantities are determined at  
621 the frequencies  $n/N_e$  (for a haploid population) with  $n$  an integer. The  
622 splitting of these frequencies, namely  $1/N_e$ , is wider (possibly substantially  
623 wider) than the splitting of the frequencies at which an exact calculation  
624 would yield, namely  $1/N$ . This has the implication that we cannot enquire  
625 into fine features of such distributions that might occur on scales comparable  
626 or smaller than  $1/N_e$ . This does not seem problematic, since if there are  
627 questions about such fine features existing, then calculations based on an  
628 effective population size may, themselves, be questionable without additional  
629 analysis.

## 630 **6.1 Summary**

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

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

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

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

$$\mathbf{F}^{(e)}(t) = \left(\mathbf{W}^{(e)}\right)^{t-1} \mathbf{W}^{(0)} \mathbf{F}(0), \quad t = 1, 2, \dots \quad (\text{A1})$$

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

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

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

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

650 Probabilities of ultimate loss and fixation of the  $A$  allele

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

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

653 where  $\mathbf{G}^{(e)}$  is given in Eq. (11) of the main text and here  $\mathbf{0}$  is a matrix  
 654 of size  $(N_e - 1) \times (N - 1)$  with all elements 0 that occurs because it can  
 655 be argued that all eigenvalues of  $\mathbf{w}^{(e)}$  have magnitude less than unity (see  
 656 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  $n = 0, 1, \dots, N_e$ . This means that we can write equivalently write Eq. (A4) as

$$\begin{aligned} \text{probability of ultimate} & \\ \text{fixation of the } A \text{ allele} & = F_{N_e}^{(e)}(\infty) \\ & = \left( 1, \mathbf{v}^{(0)} + \mathbf{v}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}, 0 \right) \mathbf{F}(0), \\ \\ \text{probability of ultimate} & \\ \text{loss of the } A \text{ allele} & = F_0^{(e)}(\infty) \\ & = \left( 1, \mathbf{u}^{(0)} + \mathbf{u}^{(e)}\mathbf{G}^{(e)}\mathbf{w}^{(0)}, 0 \right) \mathbf{F}(0). \end{aligned} \quad (\text{A5})$$

Assuming there are  $n$  copies of the  $A$  allele present in generation 0, with  $n = 1, 2, \dots, N - 1$ , 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

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

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

(A6)

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

658 Expected times to fixation and loss of the  $A$  allele

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

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

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

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

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

## 673 **Appendix B: Site frequency spectrum**

674 In this appendix we give details of the calculation for the effective site fre-  
 675 quency spectrum (SFS) and a coarse grained SFS using the method intro-  
 676 duced in this work, in the context of a haploid population of census size  $N$   
 677 (number of adults).

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

686 We shall use  $M_n(t)$  to denote the mean number of sites with mutant  
 687 alleles at a frequency of  $n/N$  in generation  $t$  (equivalently,  $M_n(t)$  denotes the  
 688 mean number of sites with  $n$  mutant alleles in generation  $t$ ). The SFS is the  
 689 set of  $M_n(t)$  for  $n = 1, 2, \dots, N-1$ , i.e., it only includes sites where mutations

690 are segregating in the population, and excludes sites where mutations have  
 691 been lost or have not occurred.

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

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

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

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

$$\mathbf{M}(t+1) = \mathbf{wM}(t) + \frac{\theta}{2} \mathbf{i}, \quad (\text{B2})$$

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

$$\hat{\mathbf{M}} = \frac{\theta}{2} \mathbf{G} \mathbf{i} \quad (\text{B3})$$

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

711 In a ‘non ideal’ population, where  $N_e < N$ , the standard results, de-  
712 scribed above, cannot be directly used. We shall use a method associated  
713 with the modified WFM of the present work. As we shall see, this leads to  
714 a ‘coarse grained’ SFS which is defined only at the frequencies  $x_n^{(\epsilon)} = n/N_e$   
715 (with  $n = 1, 2, \dots, N_e - 1$ ) rather than at the exact frequencies  $x_n^{(0)} = n/N$   
716 (with  $n = 1, 2, \dots, N - 1$ ). Because  $N_e < N$  the exact frequencies,  $x_n^{(0)}$ , are  
717 more finely spaced than the  $x_n^{(\epsilon)}$ .

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

733 Consider the part of the SFS associated with mutations which originated  
734 in the previous generation, and which have evolved for at least one generation  
735 in a manner appropriate to a population size of  $N_e$ . Under the approach of  
736 the present work we write this part of the SFS as  $\mathbf{M}^{prev,\epsilon}(t)$ . This is a column



737 vector of length  $N_e - 1$  whose  $n$ 'th element may be approximately viewed  
738 as the mean number of sites corresponding to the mutant allele frequency  
739 lying in an interval of width  $1/N_e$  in the vicinity of the frequency  $n/N_e$  (with  
740  $n = 1, 2, \dots, N_e - 1$ ). Following the viewpoint of the present work, we take  
741 the behaviour of  $\mathbf{M}^{prev,e}(t)$  to be given by

$$\left\{ \begin{array}{l} \mathbf{M}^{prev,e}(1) = \mathbf{w}^{(0)}\mathbf{M}(0), \\ \mathbf{M}^{prev,e}(t+1) = \mathbf{w}^{(e)}\mathbf{M}^{prev,e}(t) + \frac{\theta}{2}\mathbf{w}^{(0)}\mathbf{i}, \quad t = 1, 2, \dots \end{array} \right. \quad (\text{B4})$$

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

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

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

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

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

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

773 and call  $\mathbf{M}^{prev,cg}(t)$  the *coarse grained* SFS. The quantity  $\mathbf{M}^{prev,cg}(t)$  corresponds  
 774 to the frequencies  $x_n^{(e)} = n/N_e$ , which have splittings of  $1/N_e$ ,  
 775 that are larger than the splittings of the exact SFS (which is defined at the  
 776 frequencies  $x_n^{(0)} = n/N$ ) and hence have splittings of  $1/N$ . However, the  
 777 magnitude of  $\mathbf{M}^{prev,cg}(t)$  should closely correspond to the magnitude of the

778 exact SFS, when evaluated at a common frequency. We define the *coarse*  
 779 *grained* equilibrium SFS as

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

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

### 782 **Appendix C: Simulation procedure for the site frequency spec-** 783 **trum**

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

786 To simulate the SFS, we could use the method introduced in the main  
 787 text, for the Test Population, with the added feature that a random number  
 788 of new mutations are introduced in adults at the beginning of each gen-  
 789 eration. However, to determine the equilibrium SFS requires a number of  
 790 generations to ‘forget’ the initial distribution (‘burn in’ time). Furthermore,  
 791 to obtain a smooth result requires averaging the resulting fluctuating spec-  
 792 trum over a very large number of generations (an alternative is carry out  
 793 an average over many replicate populations), and this will cost a large com-  
 794 putation time or require a large computer memory. We adopt the following  
 795 alternative approach.

796 We note that once a mutant occurs at one site, it will evolve according to  
 797 a time homogeneous Markov chain, even though this will not be a standard  
 798 WFM, because  $N_e < N$ . To obtain a simulation result to compare with  
 799 the modified WFM result, and is independent, we use a one step simulation  
 800 (following the simulation described in main text) to estimate the transition

801 matrix of this Markov chain model. This involves using a total of  $R$  trajec-  
 802 tories of the mutants, with the same initial copy number. These trajectories  
 803 are run for a single generation, which includes a random reproductive stage  
 804 (negative binomial random variables are used) and a random thinning stage,  
 805 both of which are described in the Section *Test Population* of the main text.  
 806 Let  $C_j(N, s, n)$  represent the copy number of mutants after one generation of  
 807 trajectory  $j$  ( $j = 1, 2, \dots, R$ ), when the initial copy number of mutants is  $n$ ,  
 808 the census population size is  $N$ , and the selection coefficient of the mutant is  
 809  $s$ . We use the simulated values of the  $C_j(N, s, n)$  to estimate elements of the  
 810 transition matrix of this Markov chain. Writing this estimated transition  
 811 matrix as  $\widetilde{\mathbf{W}}$  we have

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

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

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

$$\widetilde{\mathbf{W}} = \begin{pmatrix} 1 & \widetilde{\mathbf{u}} & 0 \\ \mathbf{0} & \widetilde{\mathbf{w}} & \mathbf{0} \\ 0 & \widetilde{\mathbf{v}} & 1 \end{pmatrix}. \quad (\text{C2})$$

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

820 SFS of

$$\widehat{\mathbf{M}} = \frac{\theta}{2} \widetilde{\mathbf{G}} \mathbf{i}, \quad (\text{C3})$$

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

822 Finally, using the identity  $\widehat{\mathbf{M}} = \widetilde{\mathbf{w}} \widehat{\mathbf{M}} + \frac{\theta}{2} \mathbf{i}$ , and following the definition in

823 Appendix B of the equilibrium form of the SFS from ‘previous’ mutations,

824 which we write as  $\widehat{\mathbf{M}}^{\text{prev}}$ , we obtain  $\widehat{\mathbf{M}}^{\text{prev}} = \widetilde{\mathbf{w}} \widehat{\mathbf{M}}^{\text{prev}}$ . This can be written as

$$\widehat{\mathbf{M}}^{\text{prev}} = \frac{\theta}{2} \widetilde{\mathbf{G}} \widetilde{\mathbf{w}} \mathbf{i}. \quad (\text{C4})$$

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

827 In the Supplementary Material we provide a Matlab function which is a

828 generalised version of the function  $C_j(N, s, n)$  used above.

## 829 **References**

- 830 [Anderson, J. H., E. J. Ward, and S. M. Carlson, 2011 A model for estimat-](#)  
831 [ing the minimum number of offspring to sample in studies of reproductive](#)  
832 [success. J. Hered. 105: 567–576.](#)
- 833 Charlesworth, B., 2009 Fundamental concepts in genetics: effective popu-  
834 lation size and patterns of molecular evolution and variation. Nat. Rev.  
835 Genet. 10: 195–205.
- 836 Dyson, E. A., M. K. Kamath and G. D. D. Hurst, 2002 Wolbachia infec-  
837 tion associated with all-female broods in *Hypolimnas bolina* (Lepidoptera:  
838 Nymphalidae): evidence for horizontal transmission of a butterfly male  
839 killer. Heredity (Edinb.), 88: 166–171.
- 840 [Evans, S. N., Y. Shvets, and M. Slatkin, 2007 Non-equilibrium theory of the](#)  
841 [allele frequency spectrum. Theor. Popul. Biol. 71: 109–119.](#)
- 842 [Ewens, W. J., 1964 The pseudo-transient distribution and its uses in genet-](#)  
843 [ics. J. App. Prob. 1: 141-156.](#)
- 844 [Ewens, W. J., 2004 \*Mathematical population genetics I. Theoretical intro-\*](#)  
845 [duction. 2nd edition. Springer-Verlag, New York.](#)
- 846 Eyre-Walker, A. and P. D. Keightley, 2009 Estimating the rate of adaptive  
847 molecular evolution in the presence of slightly deleterious mutations and  
848 population size change. Mol. Biol. Evol. 26: 2097–2108.
- 849 [Fisher, R. A., 1922 On the dominance ratio, Proc. R. Soc. Edinb. 42:](#)  
850 [321–341.](#)
- 851 [Gillespie, J. H., 1974 Natural selection for within-generation variance in](#)  
852 [offspring number. Genetics 76: 601-606.](#)
- 853 [Gillespie, J. H., 1975 Natural selection for within-generation variance in](#)  
854 [offspring number II. discrete haploid models. Genetics 81: 403-413.](#)

855 [Gossmann, T., P. D. Keightley and A. Eyre-Walker, 2012 The effect of](#)  
856 [variation in the effective population size on the rate of adaptive molecular](#)  
857 [evolution in eukaryotes. \*Genome Biol. Evol.\* 4: 658–667.](#)

858 [Grant, P. R., and B. R. Grant, 2000 Non-random fitness variation in two](#)  
859 [populations of Darwin’s finches. \*Proc. R. Soc. Lond. B. Biol. Sci.\* 267:131–](#)  
860 [138.](#)

861 [Grossenbacher, D., R. B. Runquist, E. E. Goldberg and Y. Brandvain, 2015](#)  
862 [Geographic range size is predicted by plant mating system. \*Ecol Lett.\* 18:](#)  
863 [706–713](#)

864 [Hui, T.-Y. J. and A. Burt 2015 Estimating effective population size from](#)  
865 [temporally spaced samples with a novel, efficient maximum-likelihood algo-](#)  
866 [rithm. \*Genetics\* 200: 285-293.](#)

867 [Keightley, P. D. and A. Eyre-Walker, 2007 Joint inference of the distribution](#)  
868 [of fitness effects of deleterious mutations and population demography based](#)  
869 [on nucleotide polymorphism frequencies. \*Genetics\* 177: 2251–2261.](#)

870 [Jarvis, J. U., 1981 Eusociality in a mammal: cooperative breeding in naked](#)  
871 [mole-rat colonies, \*Science\* 212: 571–573.](#)

872 [Johnson, N. L., S. Kotz and A. W. Kemp 2005 \*Univariate discrete distrib-\*](#)  
873 [utions. 3rd Edition. Wiley Series in Probability and Statistics. John Wiley](#)  
874 [and Sons Ltd, Hoboken.](#)

875 [Kimura, M., 1955 Stochastic processes and distribution of gene frequencies](#)  
876 [under natural selection. \*Cold Spring Harbour Symp. Quant. Biol.\* 20,](#)  
877 [33–53.](#)

878 [Kimura, M., 1962 On the probability of fixation of mutant genes in a pop-](#)  
879 [ulation. \*Genetics\* 47: 713-719.](#)

880 [Kimura, M., 1969 The number of heterozygous nucleotide sites maintained](#)  
881 [in a finite population due to steady flux of mutations. \*Genetics\* 61: 893–903.](#)

882 [Kimura, M., and T. Ohta, 1969 The average number of generations until](#)  
883 [fixation of a mutant gene in a finite population. \*Genetics\* 61: 763-771.](#)

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

886 [Melbourne, B. A., and A. Hastings, 2008 Extinction risk depends strongly](#)  
887 [on factors contributing to stochasticity. \*Nature\* 454: 100-103.](#)

888 Reiss, J. O., 2013 Does selection intensity increase when populations de-  
889 crease? Absolute fitness, relative fitness, and the opportunity for selection.  
890 *Evol. Ecol.* 27: 477-488.

891 Schneider, A., B. Charlesworth, A. Eyre-Walker and P. D. Keightley, 2011  
892 A method for inferring the rate of occurrence and fitness effects of advanta-  
893 geous mutations. *Genetics* 189: 1427-1437.

894 Strassen, V., 1969 Gaussian Elimination is not Optimal. *Numer. Math.* 13:  
895 354-356.

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

899 [Thépot, S., G. Restoux, I. Goldringer, F. Hospital, D. Gouache, I. Mackay,](#)  
900 [and J. Enjalbert, 2015 Efficiently Tracking Selection in a Multiparental Pop-](#)  
901 [ulation: The Case of Earliness in Wheat. \*Genetics\* 199: 609-623.](#)

902 [Venton, D., 2012 Highlight-tracking adaptation's role: do larger populations](#)  
903 [evolve faster?. \*Genome Biol. Evol.\* 4: 668-669.](#)

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

906 [Waxman, D., 2011 Comparison and content of the Wright-Fisher model](#)  
907 [of random genetic drift, the diffusion approximation, and an intermediate](#)  
908 [model. \*J. Theor. Biol.\* 269: 79-87.](#)



909 [Waxman, D., 2012 Population growth enhances the mean fixation time of](#)  
910 [neutral mutations and the persistence of neutral variation. \*Genetics\* 191:](#)  
911 [561–577.](#)

912 [Wilson E. O., and B. Hölldobler 2005 Eusociality: origin and consequences.](#)  
913 [Proc Natl Acad Sci U. S. A, 102: 13367–13371.](#)

914 [Wright, S., 1931 Evolution in Mendelian populations. \*Genetics\* 16: 97–159.](#)  
915 [Wright, S., 1945 Tempo and mode in evolution: a critical review. \*Ecology\*](#)  
916 [26: 415–419.](#)

917 [Zhao, L., X. Y. Yue, and D. Waxman, 2013 Complete numerical solution of](#)  
918 [the diffusion equation of random genetic drift. \*Genetics\* 194: 973–985.](#)

919

## Supplementary Material

920

**A modified Wright-Fisher model that incorporates  $N_e$ :**

921

**A variant of the standard model with increased biological**

922

**realism and reduced computational complexity**

923

Lei Zhao, Toni Gossmann, and David Waxman

924

On the following page we give a Matlab function C.m that generates the

925

number of copies of the mutant allele after a single generation, as described in

926

Appendix C. Repeated use of this function allows estimation of the transition

927

matrix for the Test Population.

928

We use the abbreviation NBD for the negative binomial distribution.

```

930 function n=C(N,n1,s1,s2,sigma1,sigma2)
931 % Simulates copy number of the mutant allele for a haploid population
932 % INPUTS:
933 % N: census population size
934 % n1: copy number of the mutant allele in the current generation
935 % s1 and s2: selection coefficients of mutant and resident alleles, respectively
936 % sigma1 and sigma2: variances in offspring No. of mutant and
937 % resident alleles, respectively
938 % OUTPUT:
939 % n: copy No. of mutant allele one generation after the current generation
940 % 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
941 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=nbirnd(n1*r1,p1); % NBD offspring No.
    n2=nbirnd(n2*r2,p2); % NBD offspring No.
    n=binornd(N,n1./(n1+n2)); % Thinning: copy No. mutant alleles in next generation

```