US
UNIVERSITY OF SUSSEX

Sussex Research

Mutation and selection in a large population

Joel Peck, David Waxman, A. Cruikshank

Publication date
01-01-2004

Licence
This work is made available under the Copyright not evaluated licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

Citation for this work (American Psychological Association 7th edition)
Peck, J., Waxman, D., & Cruikshank, A. (2004). *Mutation and selection in a large population* (Version 1). University of Sussex. https://hdl.handle.net/10779/uos.23311103.v1

Published in
BioSystems

Link to external publisher version
https://doi.org/10.1016/j.biosystems.2003.12.004

Copyright and reuse:
This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at sro@sussex.ac.uk. Discover more of the University's research at https://sussex.figshare.com/

Mutation and Selection in a Large Population

J. R. Peck, D. Waxman and A. Cruikshank

Centre for the Study of Evolution,

School of Life Sciences

University of Sussex,

Falmer BN1 9QG, Sussex UK

Running Head: Mutation and Selection in a Large Population

Key Words: one locus, many alleles, diploid population, genetic drift, large population size

Correspondence to:

Dr. J. R. Peck, School of Life Sciences, University of Sussex, Brighton BN1 9QG, Sussex UK.

E-mail: J. R. Peck@sussex.ac.uk, Phone: +44 (0)1273 678843

Abstract

In this paper we study a large, but finite population, in which mutation and selection occur at a single genetic locus in a diploid organism. We provide theoretical results for the equilibrium allele frequencies, their variances and covariances and their equilibrium distribution, when the population size is larger than the reciprocal of the mean mutation rate. [[We are also able to infer that the equilibrium distribution of allele frequencies takes the form of a constrained multivariate Gaussian distribution.]] Our results provide a rapid way of obtaining useful information in the case of complex mutation and selection schemes when the population size is large. We present numerical simulations to test the applicability of our theoretical formulations. The results of these simulations are in very reasonable agreement with the theoretical predictions.

1. Introduction

Biological evolution depends on changes in allele frequencies and these changes can occur because of various evolutionary "forces" that include selection, mutation, and genetic drift. Understanding how these evolutionary forces combine to produce distributions of allele frequencies is, generally, a complex task. Most progress has been made in the case of infinite populations (Crow and Kimura, 1970), however for the more realistic case of finite populations, there has been less progress.

In this paper we focus on the case a finite population in which mutation and selection occur at a single genetic locus in a diploid organism with non-overlapping generations. Our main objective is to provide results that can help in the analysis of situations that are either difficult to approach with purely analytic methods or are highly time-consuming when simulated on a computer. The results found can, in particular, provide useful information in the case of a complex selection scheme where the population is too large to allow a complete study using only computer simulations.

The primary restrictions on the applicability of our approach are that the number of alleles is finite, hence continuum of alleles models are not included, and that the reciprocal of the population size is small compared with the mean mutation rate. We clarify the origin of this restriction in Subsection 6.1.

The genetic locus under consideration has n possible alleles and we describe these by the column vector $\mathbf{p}(t) \equiv (p_1(t), p_2(t), ..., p_n(t))^T$ (where T denotes transpose), and the i'th element of $\mathbf{p}(t)$ is the frequency of allele i at generation (i.e. time) t $(= 0, 1, 2, 3...)$. At the time of census, population size is fixed at N. Thus the frequency of any allele can only be one of the values

given by

$$\text{allowed allele frequencies} = \frac{0}{2N}, \frac{1}{2N}, \frac{2}{2N}, \frac{3}{2N}, ..., \frac{2N}{2N} \tag{1.1}$$

and there are a total of $2N+1$ possible values for each element of $\mathbf{p}(t)$. Because a value of $\mathbf{p}(t)$ is simply a specification of the value of each of the n elements of $\mathbf{p}(t)$, there exists a finite number of possible values of $\mathbf{p}(t)$. This number would have the value $(2N+1)^n$ if the elements of $\mathbf{p}(t)$ were independent, however they are constrained to sum to unity. This results in the number of possible values of $\mathbf{p}(t)$ being generally a much smaller number and given by $\frac{(2N+n-1)!}{(2N)!(n-1)!}$.

In general, evolutionary biologists are most interested in the long-term outcome of evolution. Therefore, we will concentrate on characterising $\mathbf{p}(t)$ for large values of t. The analysis we present applies for the class of models where the value of $\mathbf{p}(t)$, in an infinite population, approaches a *unique* equilibrium value at long times.

We shall focus on the calculation of the mean and variance of the various allele frequencies, along with the covariances, over time, between allele frequencies. Our calculations hold for the long term, once no systematic trends are exhibited by the population and only the effects of stochastic drift are present. We shall loosely refer to this state of the population as "equilibrium" but emphasise that there may be considerable stochasticity present. Once this equilibrium regime is achieved, we can interpret the results of calculations that involve genetic drift in two different ways, both of which are valid. The first way views the results for summary statistics as being derived from an average, over a large number of replicate populations that differ from each other due to their different stochastic histories. The second way takes the view that there is a *single* population and the distributions or summary statistics arising from the calculations describe a time average over this single population. In this work we shall generally adopt the single

population viewpoint.

As we shall see, the quantities we calculate allow the determination of evolutionarily important quantities that include the level of genetic variance, the level of heterozygosity, the mean fitness and also the loss of fitness due to genetic drift (the drift load). An advantage of our work is that dependence on population size, N, is explicitly present in the results, so comparing results for different population sizes requires no additional calculation.

Previous theoretical studies on mutation and selection in finite populations have generally assumed a particular pattern of fitnesses and mutations. Summary statistics, such as mean fitness, genetic variance and the level of heterozygosity, have been found by computer simulations and analytic approximations. Our work allows for calculation of these quantities and can readily deal with general schemes of mutation and selection in a single framework.

We begin the presentation of the analysis with a study of the infinite-population case. This is, essentially, a re-formulation of previous work (Crow and Kimura, 1970). We then use the results from the infinite-population case as the basis for investigating the case of finite-populations.

2. The model

Consider a diploid organism in which generations are discrete. During each generation, the population undergoes four phases:

1. The adults produce gametes which, by random union, form zygotes. These mature into juveniles. We assume that a very large (effectively infinite) number of zygotes are produced.

2. All of the adults die, leaving only the juveniles alive.

3. Viability selection occurs. The probability that a particular juvenile will survive viability selection depends only on their genotype.

4. We assume that the resources present in the environment are only sufficient to support N adults, thus a non-selective thinning process occurs, where N juveniles are selected at random. These juveniles become the adults of the next generation, while the remainder die.

The n possible alleles at the one locus under selection are numbered $1, 2, ..., n$ and the i'th allele is denoted by A_i. An individual who inherited A_i from one parent and A_j from the other will be referred to as an individual of type (i, j). The probability that a juvenile of type (i, j) will survive viability selection is given by $W_{ij} \equiv W_{ji}$. [[Note that we do not assume any relation between W_{ii}, W_{ij} and W_{jj}, thus no particular dominance relation is assumed between any pair alleles.]]

It is convenient to be able to work in terms of *relative* fitnesses, rather than absolute fitnesses Therefore, we define the *relative fitness* of type (i, j) juveniles, denoted w_{ij}, as:

$$w_{ij} = \frac{W_{ij}}{W_{nn}} \tag{2.1}$$

Thus, type (n, n) individuals are arbitrarily chosen as the reference genotype, and have a realtive fitness of unity. It is also convenient to define the selection coefficient associated with genotype (i, j) by

$$s_{ij} = w_{ij} - 1. \tag{2.2}$$

Finally, we assume that mutations occur during the production of gametes. In any given gamete containing allele j, the probability of mutation to an allele of type i is given by μ_{ij}.

We shall analyse the above model when the reciprocal of the population size, N^{-1}, is a small quantity in the model and allows an expansion in N^{-1}. In particular, this means N^{-1} should be much smaller than the mean mutation rate.

3. The infinite population limit

To begin the analysis, we consider the limit as population size, N, goes to infinity. In this case the allowed allele frequencies, given in (1.1), become continuous. The equilibrium of the population is described by the vector $\mathbf{\Lambda} = (\Lambda_1, \Lambda_2, \ldots, \Lambda_n)^T$ and this is assumed to be *unique*. Thus, many generations after an arbitrary starting point, the frequency of allele A_i $(i = 1, 2, ..., n)$ has a value given by the i'th component of $\mathbf{\Lambda}$, namely Λ_i. Furthermore, in the infinite population limit, the values of the frequencies are known with certainty and have no fluctuations about their values.

We have assumed models which, when $N \longrightarrow \infty$, the long time limit of $\mathbf{p}$, i.e. $\mathbf{p}(\infty)$, always achieves the same value, namely $\mathbf{p}(\infty) = \mathbf{\Lambda}$, corresponding to the existence of a unique equilibrium. While this is the relevant case in many situations, it is possible to choose the values of μ_{ij} and w_{ij} such that there may be multiple equilibria possible. Other possibilities are that allele frequencies may exhibit chaotic or other complex behaviours. In this paper, we will not consider models with these properties [[although we shall briefly comment on multiple equilibria in SubSection 5.1]].

In special cases it is possible to write analytic expressions for the equilibrium allele frequencies, $\mathbf{\Lambda}$, in terms of the values of w_{ij} and μ_{ij} and there is a large literature on this topic, starting in

the early days of theoretical population genetics (Felsenstein, 1981). However we are not aware of any general expressions for $\mathbf{\Lambda}$. Nevertheless it is straightforward to numerically calculate $\mathbf{\Lambda}$ to a high degree of accuracy. One simply specifies an initial set of frequencies, $\mathbf{p}(0)$, and iterates, to convergence, the equation that determines the gene frequencies in subsequent generations. This equation is

$$\mathbf{p}(t+1) = \mathbf{p}(t) + \mathbf{\Omega}(\mathbf{p}(t)), \tag{3.1}$$

where $\Omega(\mathbf{p})$ is an n component column vector with elements

$$\Omega_i(\mathbf{p}) = \frac{p_i \left[\sum_j w_{ij} p_j - \sum_{jk} w_{jk} p_j p_k\right] + \sum_{jk} \left[\mu_{ij} w_{jk} p_j p_k - \mu_{ji} w_{ik} p_i p_k\right]}{\bar{w}(\mathbf{p})} \tag{3.2}$$

and

$$\bar{w}(\mathbf{p}) = \sum_{jk} w_{jk} p_j p_k. \tag{3.3}$$

4. Finite populations

What is the outcome of evolution when the population is finite in size? No general answer to this question exists, however, a great deal can be said if we restrict ourselves to the situation where the population size, N, is sufficiently large that the allele frequencies of (1.1) can be treated as continuous variables lying in the interval $[0, 1]$. In this case, we can incorporate the most important effects of finite population size by adding the random genetic drift term $\boldsymbol{\xi}(t) = (\xi_1(t), \xi_2(t), \ldots, \xi_n(t))^T$ on the right hand side of (3.1):

$$\mathbf{p}(t+1) = \mathbf{p}(t) + \mathbf{\Omega}(\mathbf{p}(t)) + \boldsymbol{\xi}(t). \tag{4.1}$$

With E denoting the expectation operator and $\delta_{i,j}$ the Kronecker delta ($\delta_{i,j} = 1$ if $i = j$ and is zero otherwise), the $\xi_i(t)'$s satisfy the standard conditional expectations

$$E[\xi_i(t) \quad | \quad \mathbf{p}(t] = 0, \quad E[\xi_i(t)p_k(t) \mid \mathbf{p}(t] = 0$$

$$E[\xi_i(t_1)\xi_j(t_2) \quad | \quad \mathbf{p}(t)] = \delta_{t_1,t_2}\frac{p_i(t_1)\delta_{i,j} - p_i(t_1)p_j(t_2)}{2N} \tag{4.2}$$

[[where the last result follows from a multinomial distribution.]]

The fundamental quantities we are interested in are the equilibrium allele frequencies along with their variances and the covariances between different allele frequencies. We can use (4.1) and (4.2) to derive approximate equation for these quantities when N is suitably large.

We note that Barlett (1978) has presented calculations for the leading effects of finite population size on a one locus, two allele model. His work exploits the fact, as does this work, that N^{-1} may be used as an expansion parameter in the calculations.

4.1. Equations that determine the mean allele frequencies and their variances and covariances

To determine the approximate means, variances and covariances, we first take the unconditional expectation of (4.1). In equilibrium (where t arguments are omitted) we obtain

$$E\left[\Omega_i(\mathbf{p})\right] = 0. \tag{4.3}$$

Denoting the mean value of $\mathbf{p}$ in equilibrium by $\bar{\mathbf{p}}$:

$$E\left[\mathbf{p}\right] = \bar{\mathbf{p}} \tag{4.4}$$

we subtract $\bar{\mathbf{p}}$ from (4.1) yielding $p_i(t+1) - \bar{p}_i = p_i(t) - \bar{p}_i + \Omega_i(\mathbf{p}(t)) + \xi_i(t)$. We combine this equation with the corresponding equation where i is replaced by j by multiplying the i and j equations together and take expectation values to obtain

$$\begin{aligned} & E\left[\left(p_i(t+1) - \bar{p}_i\right)\left(p_j(t+1) - \bar{p}_j\right)\right] \\ = \; & E\left[\left(p_i(t) - \bar{p}_i\right)\left(p_j(t) - \bar{p}_j\right)\right] + E\left[\left(p_i(t) - \bar{p}_i\right)\Omega_j(\mathbf{p}(t))\right] \\ & + E\left[\Omega_i(\mathbf{p}(t))\left(p_j(t) - \bar{p}_j\right)\right] + E\left[\xi_i(t)\xi_j(t)\right]. \end{aligned} \tag{4.5}$$

In equilibrium this reduces to

$$E\left[\left(p_i - \bar{p}_i\right)\Omega_j(\mathbf{p}) + \Omega_i(\mathbf{p})\left(p_j - \bar{p}_j\right)\right] = -E\left[\xi_i(t)\xi_j(t)\right]. \tag{4.6}$$

Using (4.3) and (4.4), we can write (4.6) as

$$\begin{aligned} & E\left[\left(p_i - \bar{p}_i\right)\left(\Omega_j(\mathbf{p}) - \Omega_j(\bar{\mathbf{p}})\right) + \left(\Omega_i(\mathbf{p}) - \Omega_i(\bar{\mathbf{p}})\right)\left(p_j - \bar{p}_j\right)\right] \\ = \; & -E\left[\xi_i(t)\xi_j(t)\right]. \end{aligned} \tag{4.7}$$

Equations (4.3) and (4.7) are, as they stand within our model, exact. Let us now use them to obtain approximations for the allele frequency means along with their variances and covariances.

4.2. Approximation

In order to derive useful approximations, we must make certain plausible assumptions (assumptions 1-3 below). We will test the accuracy of these assumptions shortly. Note that assumptions 1-3 are consistent with Eqs. (4.3) and (4.7), in the limit of very large N.

The assumptions are:

1. The mean allele frequencies, $\bar{\mathbf{p}}$, consist of $\mathbf{\Lambda}$ (the $N = \infty$ deterministic equilibrium result) plus a correction whose leading term is of order N^{-1}.

2. The variances and covariances of the various allele frequencies, $E\left[\left(p_i - \bar{p}_i\right)\left(p_j - \bar{p}_j\right)\right]$ are of order N^{-1}.

3. Higher order correlations such as $E\left[\left(p_i - \bar{p}_i\right)\left(p_j - \bar{p}_j\right)\left(p_k - \bar{p}_k\right)\right]$ are of order N^{-2} or higher order in N^{-1}.

We determine $\bar{\mathbf{p}}$ and $E\left[\left(p_i - \bar{p}_i\right)\left(p_j - \bar{p}_j\right)\right]$ up to and including terms of order N^{-1}. To proceed, let us introduce the quantities B_i and C_{ij} which are defined via

$$E\left[p_i\right] \equiv \bar{p}_i = \Lambda_i + B_i/N + O\left(1/N^2\right), \tag{4.8}$$

$$E\left[\left(p_i - \bar{p}_i\right)\left(p_j - \bar{p}_j\right)\right] = C_{ij}/N + O\left(1/N^2\right) \tag{4.9}$$

thus B is an n component column vector and C is an $n \times n$ matrix and both are independent of N.

It is natural to first determine C_{ij} and we do this by expanding $\mathbf{\Omega}(\bar{\mathbf{p}})$ in (4.7) about $\mathbf{p} = \bar{\mathbf{p}}$ to *first* order in $(\mathbf{p} - \bar{\mathbf{p}})$. Thus $E\left[(p_i - \bar{p}_i)\left(\Omega_j(\mathbf{p}) - \Omega_j(\bar{\mathbf{p}})\right)\right]$ in (4.7) yields

$$
\begin{aligned}
& E\left[(p_i - \bar{p}_i)\left(\Omega_j(\mathbf{p}) - \Omega_j(\bar{\mathbf{p}})\right)\right] \\
= & \sum\nolimits_k E\left[(p_i - \bar{p}_i)(p_k - \bar{p}_k)\right] \left.\frac{\partial \Omega_j(\mathbf{p})}{\partial p_k}\right|_{\mathbf{p}=\bar{\mathbf{p}}} + O(1/N^2) \\
= & \sum\nolimits_k \frac{C_{ik}}{N} \left.\frac{\partial \Omega_j(\mathbf{p})}{\partial p_k}\right|_{\mathbf{p}=\mathbf{\Lambda}} + O(1/N^2) \qquad (4.10)
\end{aligned}
$$

the last equality following from the assumption of (4.8), that $\bar{\mathbf{p}}$ is, to leading order in N^{-1}, equal to $\mathbf{\Lambda}$. Additionally, the right-hand-side of (4.7) has

$$
\begin{aligned}
\frac{E\left(p_i \delta_{ij} - p_i p_j\right)}{2N} &= \frac{\bar{p}_i \delta_{ij} - \bar{p}_i \bar{p}_j - E\left[(p_i - \bar{p}_i)(p_j - \bar{p}_j)\right]}{2N} \\
&= \frac{\Lambda_i \delta_{ij} - \Lambda_i \Lambda_j}{2N} + O(N^{-2}) \qquad (4.11)
\end{aligned}
$$

the last equality using (4.8) and (4.9). Thus the introduction of [[N independent matrices Γ and A given by]]

$$
\Gamma_{ij} \stackrel{\text{def}}{\equiv} \frac{\Lambda_i \delta_{ij} - \Lambda_i \Lambda_j}{2} \qquad (4.12)
$$

and

$$A_{jk} \stackrel{\text{def}}{\equiv} - \left.\frac{\partial \Omega_j(\mathbf{p})}{\partial p_k}\right|_{\mathbf{p}=\mathbf{\Lambda}} \tag{4.13}$$

(4.10) leads to the matrix equation that determines C:

$$AC + CA^T = \Gamma. \tag{4.14}$$

Once C is known, we can determine B by similarly expanding the left-hand-side of (4.3) to *second* order in $\mathbf{p} - \bar{\mathbf{p}}$. This yields the equation

$$\sum\nolimits_j \left.\frac{\partial \Omega_i(\mathbf{p})}{\partial p_j}\right|_{\mathbf{p}=\mathbf{\Lambda}} B_j + \frac{1}{2}\sum\nolimits_{j,k} \left.\frac{\partial^2 \Omega_i(\mathbf{p})}{\partial p_j \partial p_k}\right|_{\mathbf{p}=\mathbf{\Lambda}} C_{jk} = 0. \tag{4.15}$$

4.3. Calculation of B and C

Here we give a *prescription* by which B and C can be calculated. The rationale underlying this is given in Appendix A.

With $\mathbf{\Lambda}$ assumed known from numerical or analytic methods, explicit calculations require the form of $A_{ij} = -\left.\partial \Omega_i(\mathbf{p})/\partial p_j\right|_{\mathbf{p}=\mathbf{\Lambda}}$ and also $\left.\partial^2 \Omega_j(\mathbf{p})/\partial p_k \partial p_l\right|_{\mathbf{p}=\mathbf{\Lambda}}$. For completeness, we state the results in the case of frequency-independent selection:

$$\begin{aligned} A_{jk} \;=\; & -\frac{1}{\overline{w}(\Lambda)} \Big\{ \delta_{j,k} \left(\textstyle\sum_r w_{jr}\Lambda_r - \overline{w}(\Lambda) - \sum_{r,s} \mu_{rj} w_{rs} \Lambda_s \right) \\ & + \Lambda_j \left(w_{jk} - 2\textstyle\sum_r w_{kr}\Lambda_r - \sum_r \mu_{rj} w_{jk} \right) \\ & + \textstyle\sum_r \mu_{jr}\Lambda_r w_{rk} + \mu_{jk} \sum_r w_{kr}\Lambda_r \Big\}, \end{aligned} \tag{4.16}$$

$$\left.\frac{\partial^2\Omega_i(\mathbf{p})}{\partial p_r \partial p_s}\right|_{\mathbf{p}=\Lambda} = \frac{1}{\overline{w}(\Lambda)}\Big\{2A_{is}\sum_j w_{rj}\Lambda_j + 2A_{ir}\sum_j w_{sj}\Lambda_j$$

$$+\delta_{i,s}\left(w_{ir} - 2\sum_j w_{rj}\Lambda_j - \sum_j \mu_{ji}w_{ir}\right)$$

$$\delta_{i,r}\left(w_{is} - 2\sum_j w_{sj}\Lambda_j - \sum_j \mu_{ji}w_{is}\right) \tag{4.17}$$

$$+\mu_{ir}w_{rs} + \mu_{is}w_{rs} - 2\Lambda_i w_{rs}\Big\}.$$

The solutions for B and C are written in terms of ψ_i and χ_i^T, which are the right and left eigenvectors of the matrix A associated with eigenvalue λ_i, $i = 1, 2, ..., n$. These are selected to obey

$$A\psi_i = \lambda_i\psi_i, \qquad \chi_i^T A = \lambda_i\chi_i^T, \qquad \chi_i^T\psi_j = \delta_{i,j}. \tag{4.18}$$

Then the matrix C can be written as

$$C = \sum_{i,j=1}^{n} \frac{\psi_i\chi_i^T\Gamma\chi_j\psi_j^T}{\lambda_i + \lambda_j}. \tag{4.19}$$

For the vector B it is simpler to write out the components rather than give an expression for the entire vector. The i'th component of B is given by

$$B_i = \frac{1}{2}\sum_{j,k,l=1}^{n} \left(A^{-1}\right)_{ij} \left.\frac{\partial^2\Omega_j(\mathbf{p})}{\partial p_k \partial p_l}\right|_{\mathbf{p}=\Lambda} C_{kl}. \tag{4.20}$$

With these expressions, we have, via (4.8) and (4.9) the means and variances or covariances of allele frequencies to order N^{-1}.

5. Expressions for some biologically relevant quantities

Various quantities of biological interest may be expressed in terms of the mean allele frequencies and their covariances. Using (4.8) and (4.9), these may, if desired, be expressed in terms of B and C.

5.1. Probability distribution

Perhaps the most fundamental quantity we can approximately determine is the stationary probability density, $\Phi(\mathbf{p})$, which has the interpretation that $\Phi(\mathbf{p})dp_1dp_2...dp_n$ is the probability that p_1 lies in the range $(p_1, p_1 + dp_1)$, p_2 lies in the range $(p_2, p_2 + dp_2)\ldots$. It can be shown that the following distribution yields mean allele frequencies and covariances that are, to order N^{-1}, identical to the results (4.8) and (4.9):

$$\Phi(\mathbf{p}) = Z\delta(F^T\mathbf{p} - 1)\exp\left[-\frac{N}{2}(\mathbf{p} - \bar{\mathbf{p}})^T[C]^{-1}(\mathbf{p} - \bar{\mathbf{p}})\right] \tag{5.1}$$

In (5.1), Z is a constant that ensures the integral of $\Phi(\mathbf{p})$ over all allele frequencies is unity, $\int dp_1dp_2...dp_n\Phi(\mathbf{p}) = 1$, as is required of a probability density. The quantity $\delta(\bullet)$ denotes a Dirac delta function (which satisfies $\int_{-\infty}^{\infty}\delta(x-a)g(x)dx = g(a)$ for $g(x)$ an arbitrary function). The quantity F is an n component column vector with all elements equal to 1: $F = (1, 1, 1, \ldots)^T$ and $[C]^{-1}$ denotes the *pseudo-inverse* of the matrix C. [[We note that from Eq. (4.9), C contains all information, to $O(N^{-1})$, about all variances and covariances of allele frequencies.]]

The form of (5.1) can be understood as follows. The factor $\delta(F^T\mathbf{p} - 1) \equiv \delta(\sum_{i=1}^{n} p_i - 1)$ ensures that $\Phi(\mathbf{p})$ is only non-zero at frequencies that sum to unity. The remaining factor is a multivariate Gaussian corresponding to a mean lying at $\mathbf{p} = \bar{\mathbf{p}}$, and the Gaussian is characterised

by fluctuations about the mean, i.e. variances and covariances, that are of order N^{-1}. In the limit $N \to \infty$, $\Phi(\mathbf{p})$ in (5.1) collapses to $\delta(\mathbf{p} - \mathbf{\Lambda})$, which corresponds to a distribution with sharply defined allele frequencies given by the components of $\mathbf{\Lambda}$.

[[It seems very plausible that in the event of multiple, well-separated, equilbria, Eq. (5.1) describes a population that is trapped in the vicinity of the particular equilibrium located at $\mathbf{p} = \bar{\mathbf{p}}$. One can also envisage a population that makes drift induced transitions between *nearby* equlibria, or other movement between equilibria, however the analysis of these lies well beyond the present work.]]

5.2. Mean heterozygosity

The mean heterozygosity is the average proportion of individuals that are heterozygous. A particular population, with allele frequencies given by the elements of $\mathbf{p}$, has the fraction of heterozygotic individuals given by $\sum_{i,j\,(i\neq j)} p_i p_j = 1 - \sum_i p_i^2 \equiv 1 - \mathbf{p}^T\mathbf{p}$. Time averaging this quantity yields the expected mean heterozygosity, H:

$$\begin{aligned} H &= E\left[1 - \mathbf{p}^T\mathbf{p}\right] = 1 - \bar{\mathbf{p}}^T\bar{\mathbf{p}} + \mathrm{Tr}\left[C\right]/N \\ &= 1 - \Lambda^T\Lambda + 2\Lambda^T B/N + \mathrm{Tr}\left[C\right]/N \end{aligned} \tag{5.2}$$

where $\mathrm{Tr}\left[C\right] \equiv \sum_i C_{ii}$.

5.3. Genetic variance

Let the column vector $x = (x_1, x_2, ..., x_L)^T$ contain the effects of the different alleles. The variance of allelic effects of a population whose allele frequencies are $\mathbf{p}$, at a particular time, is $2\left[\sum_i p_i x_i^2 - \left(\sum_i p_i x_i\right)^2\right]$. The *expected* genetic variance is the time average of this quantity:

$$\begin{aligned} V_g &= 2E\left[\sum\nolimits_i p_i x_i^2 - \sum\nolimits_{i,j} p_i p_j x_i x_j\right] \\ &= 2\left[\sum\nolimits_i \bar{p}_i x_i^2 - \sum\nolimits_{i,j} \left(\bar{p}_i \bar{p}_j + C_{ij}/N\right) x_i x_j\right] \\ &= 2\left[\sum\nolimits_i \Lambda_i x_i^2 - \sum\nolimits_{i,j} \Lambda_i \Lambda_j x_i x_j\right] \\ &\quad + \frac{1}{N}\left[\sum\nolimits_i B_i x_i^2 - 2\sum\nolimits_{i,j} B_i B_j x_i x_j + \sum\nolimits_{i,j} C_{ij} x_i x_j\right]. \end{aligned} \tag{5.3}$$

5.4. Drift load

The drift load is the fraction of the population that die each generation due to genetic drift causing some individuals to have a fitness that is less than the optimum. With $\bar{w}(\mathbf{p})$ defined in (3.3), $E\left[\bar{w}(\mathbf{p})\right]$ is the expected (i.e. time averaged) mean fitness of the population in equilibrium. Furthermore, in an infinite equilibrium population, the allelic frequencies are precisely given by Λ (with no deviations about this value), thus $\overline{w}(\Lambda)$ is the mean equilibrium fitness of an infinite population. Therefore the expected drift load is given by

$$L_{drift} = \frac{\bar{w}(\Lambda) - E\left[\bar{w}(\mathbf{p})\right]}{\bar{w}(\Lambda)}. \tag{5.4}$$

Using (4.8) and (4.9) we find $\bar{w} = \sum_{j,k} w_{jk}\left(\bar{p}_j\bar{p}_k + \frac{C_{jk}}{N}\right) =$ $\sum_{j,k} w_{jk}\left(\Lambda_j\Lambda_k + \frac{\Lambda_j B_k + B_j\Lambda_k + C_{jk}}{N}\right)$ hence

$$L_{drift} = \frac{1}{N}\left(\frac{-\sum_{j,k} w_{jk}\left(2\Lambda_j B_k + C_{jk}\right)}{\sum_{j,k} w_{jk}\Lambda_j\Lambda_k}\right). \tag{5.5}$$

6. Comparison with results for 2 alleles

Having derived estimates for the mean allele frequencies and their covariances from a large N approximation of diffusion analysis, we now compare these with a diffusion analysis results for the case of 2 alleles. This serves to make clear the domain of validity of our approximate results.

Following Ewens (1969) we use the notation

$$\begin{aligned} \mu_{21} &= u, \qquad \mu_{12} = v \\ w_{11} &= 1 + s_1, \qquad w_{12} = 1 + s_2, \qquad w_{22} = 1 \end{aligned} \tag{6.1}$$

with $s_1, s_2, u, v \ll 1$ [[but no particular relation between s_1, 1 and s_2, so allelic effects are, in general, neither additive nor multiplicative.]] Then diffusion analysis, (Ewens, 1969), gives

$f(x)dx$ as the probability that the frequency of allele A_1 will lie in the range $(x, x+dx)$, where

$$f(x) = \frac{x^{4Nv-1}(1-x)^{4Nu-1}\exp\left[4Ns_2x + 2N(s_1-2s_2)x^2\right]}{\int_0^1 dy\, y^{4Nv-1}(1-y)^{4Nu-1}\exp\left[4Ns_2y + 2N(s_1-2s_2)y^2\right]}. \tag{6.2}$$

The mean frequency of allele A_1 is thus given, in the diffusion approximation, by

$$\bar{p}_1 = \int_0^1 dx\, x\, f(x) \tag{6.3}$$

and the mean frequency of allele A_2 is $\bar{p}_2 = 1 - \bar{p}_1$. The covariance of the frequencies of A_1 and A_2 is, in the diffusion approximation,

$$\begin{aligned} \mathrm{cov}(p_1, p_2) &\equiv E(p_1 - \bar{p}_1, p_2 - \bar{p}_2) = \int_0^1 dx\, x(1-x)\, f(x) - \bar{p}_1(1-\bar{p}_1) \\ &= -\left(\int_0^1 dx\, x^2 f(x) - \bar{p}_1^2\right) \end{aligned} \tag{6.4}$$

6.1. Selectively neutral case

In the case where both s_1 and s_2 are zero, $\bar{p}_1$ and $\mathrm{cov}(p_1, p_2)$, as given by (6.3) and (6.4) may be evaluated in closed form. They are

$$\left.\begin{aligned} \bar{p}_1 &= \frac{v}{u+v} \\ \mathrm{cov}(p_1, p_2) &= -\frac{1}{4N}\frac{uv}{(v+u)^2\left(v+u+\frac{1}{4N}\right)} \end{aligned}\right\} \text{results of standard diffusion analysis.} \tag{6.5}$$

If we specialise the results given for the calculations of B and C in (4.19) and (4.20) to the $n = 2$ case, we obtain, after some work,

$$\left.\begin{aligned} \bar{p}_1 &= \frac{v}{u+v} \\ \mathrm{cov}(p_1, p_2) &= -\frac{1}{4N}\frac{uv}{(v+u)^3} \end{aligned}\right\} \begin{array}{l} \text{results of large } N \\ \\ \text{analysis of this work.} \end{array} \tag{6.6}$$

A comparison of $\mathrm{cov}(p_1, p_2)$ from (6.5) and (6.6) indicates that the two results are approximately equal only if $N^{-1} \ll 4\,(u+v)$. This appears to be the typical limitation of our approach and we shall conservatively take this to mean that N^{-1} must be much smaller than the mean allelic mutation rate. This is not a strict criterion. If we consider the 2 allele case, it is evident that the probability density will only be similar to a Gaussian (i.e. will be a unimodal distribution) when the factor $x^{4Nv-1}(1-x)^{4Nu-1}$ in (6.2) does *not* result in sharp peaks at $x = 0$ and $x = 1$, corresponding to quasi-fixation of alleles in the vicinity of their boundary-value frequencies. A unimodal distribution will be obtained when $4Nv - 1 > 0$ and $4Nu - 1 > 0$. We infer that the large N results of the present work are applicable when, apart from N being sufficiently large, the pattern of mutation probabilities, μ_{ij}, is such that the population cannot get irreversibly "trapped" at some alleles. To make stronger theoretical statements concerning this seems to be formidably hard. Let us therefore discuss the numerical work and simulations we have performed.

6.2. More general 2 allele case

We have carried out numerical comparisons of the predictions of diffusion analysis given in (6.3) and (6.4) and the results of this work summarised in (4.19) and (4.20). We have restricted

selection coefficients to be small to allow the use of diffusion results. We find that when N^{-1} is reasonably smaller than the allelic mutations rates, the agreement is extremely good, as Table 1 illustrates.

Table 1

7. Standardised selection/mutation scheme

As a further application of our results, we consider a single set of mutation rates and two different choices for the fitnesses. We refer to these as *Standard Set* 1 and 2 and compare the results with *numerical simulations*. We take, for both Standard Set 1 and 2, a population size of 2000 with 10 alleles segregating at the locus in question, thus

$$N = 2000, \qquad n = 10. \tag{7.1}$$

7.1. Results for Standard Set 1

Mutation rates and fitnesses μ_{std} and w_{std} are given in Appendix B. The maximum mutation rates were of chosen to be of order 10^{-3}. This very large value was chosen to speed up the approach to equilibrium of the numerical simulations. The fitnesses of Standard Set 1 correspond to relatively small selection coefficients.

We present the results of the approximation of this work ("large N approximation") and numerical simulations of the life-cycle of the one-locus randomly mating diploid organism considered in this work for the standard set of fitnesses and mutation rates.

7.1.1. Mean allele frequencies

Using (4.8),and (4.20), the approximation of this work yields, for the mean allele frequencies,

$$\underbrace{\bar{\mathbf{p}} = 10^{-1} \times (0.810, 1.182, 1.244, 0.902, 1.026, 1.099, 1.124, 0.804, 0.690, 1.118)^T}_{\text{Result of large N approximation}} \tag{7.2}$$

while the numerical simulations produced

$$\underbrace{\bar{\mathbf{p}} = 10^{-1} \times (0.810, 1.182, 1.244, 0.902, 1.025, 1.099, 1.126, 0.804, 0.691, 1.119)^T}_{\text{Result of numerical simulation}} \tag{7.3}$$

7.1.2. Covariances

From (4.9) and (4.19), the matrix of covariances is given by C/N where C is a symmetric matrix, whose independent elements are

$$
\begin{aligned}
& C \\
= \; & 10^{-1} \times \\
& \underbrace{\begin{pmatrix}
3.604 & & & & & & & & & \\
-1.181 & 7.694 & & & & & & & & \\
-0.013 & -2.846 & 7.099 & & & & & & & \\
-0.127 & -0.884 & -0.149 & 3.187 & & & & & & \\
-0.220 & -0.842 & -0.227 & -0.533 & 5.405 & & & & & \\
-0.798 & -0.489 & -1.007 & -0.480 & -1.038 & 5.129 & & & & \\
-0.501 & -0.738 & -0.719 & -0.353 & -1.052 & -0.445 & 4.679 & & & \\
-0.139 & -0.146 & -0.611 & -0.254 & -0.354 & -0.366 & -0.511 & 3.168 & & \\
-0.328 & 0.105 & -1.014 & 0.083 & -0.370 & -0.340 & -0.016 & -0.234 & 2.585 & \\
-0.299 & -0.673 & -0.514 & -0.490 & -0.769 & -0.166 & -0.343 & -0.552 & -0.471 & 4.277
\end{pmatrix}}_{\text{Result of large N approximation}}
\end{aligned}
\tag{7.4}
$$

The quantity that C can be directly compared with from the numerical simulations is

$$
\begin{aligned}
& N \times \text{matrix of covariances} \\
= \; & 10^{-1} \times \\
& \underbrace{\begin{pmatrix}
3.672 & & & & & & & & & \\
-1.213 & 7.674 & & & & & & & & \\
-0.008 & -2.811 & 7.346 & & & & & & & \\
-0.145 & -0.809 & -0.270 & 3.328 & & & & & & \\
-0.243 & -0.901 & -0.173 & -0.556 & 5.562 & & & & & \\
-0.767 & -0.540 & -1.166 & -0.425 & -1.018 & 5.210 & & & & \\
-0.551 & -0.777 & -0.695 & -0.399 & -1.109 & -0.379 & 4.818 & & & \\
-0.137 & -0.012 & -0.591 & -0.267 & -0.358 & -0.373 & -0.530 & 3.176 & & \\
-0.295 & 0.047 & -1.082 & 0.086 & -0.390 & -0.332 & -0.0032 & -0.239 & 2.618 & \\
-0.313 & -0.658 & -0.550 & -0.545 & -0.813 & -0.211 & -0.345 & -0.668 & -0.382 & 4.484
\end{pmatrix}}_{\text{Result of numerical simulation}}
\end{aligned}
\tag{7.5}
$$

7.1.3. Comparison

There is very good agreement between (7.2) and (7.3). This is not surprising since the result is primarily the $N = \infty$ result,

$$
\mathbf{\Lambda} = 10^{-1} \times (0.809, 1.183, 1.241, 0.901, 1.026, 1.099, 1.126, 0.806, 0.691, 1.119)^T . \tag{7.6}
$$

We have previously expressed $\overline{\mathbf{p}}$ as $\mathbf{\Lambda}$ plus an $O(N^{-1})$ correction term, (4.8). However for the population sizes considered, the $O(N^{-1})$ correction term is hard to extract and compare with theory because statistical errors in $\bar{\mathbf{p}}$ cannot be disentangled from this term.

By contrast, the leading term in the covariances is not $O(N^0)$ but the $O(N^{-1})$ term C/N and consequently this term is far more readily observable than the $O(N^{-1})$ terms in $\bar{\mathbf{p}}$. We have presented C in (7.4) and $N\times$ (the matrix of covariances from the simulations) in (7.5). There is a reasonably good agreement between the approximation of this work and the result of the simulations, thereby suggesting that the covariances do scale as N^{-1} for large N when selection is weak.

It should be noted that once $\mathbf{\Lambda}$ is known, the results of this work for $\bar{\mathbf{p}}$ and $\mathrm{cov}(p_i, p_j)$ were calculated, for any N, w and μ on a standard PC in seconds, while the numerical simulation results took an appreciable amount of computer time.

7.2. Results for Standard Set 2

Mutation rates for this set were identical to those of Standard Set 1. The fitnesses w_{std} for this set are given in Appendix C and correspond to quite large selection coefficients.

7.2.1. Mean allele frequencies

Using (4.8),and (4.20), the approximation of this work yields, for the mean allele frequencies,

$$\underbrace{\bar{\mathbf{p}} = 10^{-1} \times (0.060, 6.468, 0.046, 0.100, 0.110, 2.056, 0.268, 0.506, 0.173, 0.213)^T}_{\text{Result of large N approximation}} \tag{7.7}$$

while the numerical simulations produced

$$\underbrace{\bar{\mathbf{p}} = 10^{-1} \times (0.060, 6.468, 0.046, 0.100, 0.110, 2.057, 0.268, 0.505, 0.172, 0.213)^T}_{\text{Result of numerical simulation}} \tag{7.8}$$

7.2.2. Covariances

From (4.9) and (4.19), the matrix of covariances is given by C/N where C is a symmetric matrix, whose independent elements are

$$\begin{aligned} & C \\ = \; & 10^{-1} \times \\ & \underbrace{\begin{pmatrix} 0.021 & & & & & & & & & \\ 0.012 & 8.226 & & & & & & & & \\ 0.000 & -0.132 & 0.0253 & & & & & & & \\ 0.000 & -0.103 & 0.002 & 0.047 & & & & & & \\ 0.002 & 0.238 & -0.004 & -0.004 & 0.216 & & & & & \\ -0.040 & -7.906 & 0.118 & 0.066 & -0.487 & 9.288 & & & & \\ 0.000 & -0.232 & 0.001 & -0.001 & -0.010 & -0.025 & 0.273 & & & \\ 0.006 & 0.166 & -0.011 & -0.007 & 0.059 & -1.104 & -0.010 & 0.948 & & \\ 0.000 & -0.059 & 0.000 & 0.001 & 0.000 & -0.029 & 0.002 & -0.011 & 0.099 & \\ 0.000 & -0.210 & 0.003 & 0.001 & -0.010 & 0.119 & 0.001 & -0.036 & -0.003 & 0.136 \end{pmatrix}}_{\text{Result of large N approximation}} \end{aligned} \tag{7.9}$$

The quantity that C can be directly compared with from the numerical simulations is

$$
\begin{aligned}
& N \times \text{matrix of covariances} \\
= \; & 10^{-1} \times \\
& \underbrace{\begin{pmatrix}
0.033 & & & & & & & & & \\
0.007 & 8.736 & & & & & & & & \\
-0.001 & -0.136 & 0.034 & & & & & & & \\
-0.001 & -0.109 & 0.001 & 0.064 & & & & & & \\
0.003 & 0.237 & -0.005 & -0.003 & 0.254 & & & & & \\
-0.046 & -8.183 & 0.116 & 0.055 & -0.521 & 9.629 & & & & \\
-0.002 & -0.268 & 0.002 & 0.000 & -0.013 & -0.027 & 0.321 & & & \\
0.007 & 0.037 & -0.013 & -0.008 & 0.066 & -1.110 & -0.014 & 1.083 & & \\
0.000 & -0.066 & 0.000 & 0.000 & -0.004 & -0.046 & -0.001 & -0.010 & 0.129 & \\
0.000 & -0.256 & 0.002 & 0.000 & -0.013 & 0.132 & 0.002 & -0.038 & -0.003 & 0.174
\end{pmatrix}}_{\text{Result of numerical simulation}}
\end{aligned}
\tag{7.10}
$$

7.2.3. Comparison

Again there is very good agreement between (7.7) and (7.8) and for completeness we state the $N = \infty$ result:

$$
\mathbf{\Lambda} = 10^{-1} \times (0.060, 6.461, 0.046, 0.100, 0.109, 2.064, 0.268, 0.506, 0.173, 0.213)^T . \tag{7.11}
$$

The leading term in the covariances, for large N, is given by C/N and a comparison of C

in (7.9) and $N\times$ (the matrix of covariances from the simulations) in (7.10) indicates that the covariances do depend on N^{-1} to a reasonable approximation when selection is strong. [[Beyond this factor of N^{-1} that is present in the covariances, and arises as the leading term in N^{-1}, from multinomial sampling of the population, the evidence is that there is also good agreement in the general *pattern* of covariances predicted by the methods of this work.]]

8. Summary

In this work we have investigated some of the equilibrium properties of a finite population in which selection and mutation occur at a single genetic locus of a diploid organism. The theoretical results presented are an approximation that allows the rapid determination of allele frequencies along with covariances between them and are able to determine this information for complex mutation and selection schemes.

Our results show that covariances between allele frequencies can be quite substantial, even when mutation rates are low and population size is quite large. [[It is important to recognise that in an infinite population the equilibrium covariances between allele frequencies would all be zero. The finding of non-zero covariances under the regime that we have studied therefore represents a qualitative difference from the infinite-population case.]]

[[For large population sizes the amount of load generated by genetic drift is quite small *for any given locus*, considered in isolation. However, eukaryotic organisms typically have many thousands of genes, and each gene can have many stretches of nucleotide sequence that are maintained by selection. Thus, over the entire genome it may be possible to generate very substantial amounts of drift load, even when the population is large (Kondrashov, 1995; Peck et

al. 1997).]]

We note that the methods presented here can only be used when population size is large in comparison to the inverse of the allelic mutation rate. Thus in sexual DNA-based organisms, where allelic mutation rates tend to be small, the methods presented here will be of most interest for the calculation of quantities whose leading term is of order N^{-1}, such as the covariances between allele frequencies. However in RNA-based organisms, where mutation rates are much higher, the methods presented here can be useful for calculating a variety of different statistics. The same is true for asexual organisms, where the entire genome can be treated as a single locus where the relevant mutation rate tends to be substantial.

Appendices

A. Solutions of the B and C equations

In this appendix, we indicate how (4.14),

$$AC + CA^T = \Gamma \tag{A.1}$$

and (4.15),

$$\sum\nolimits_j \frac{\partial \Omega_i(\mathbf{p})}{\partial p_j}\bigg|_{\mathbf{p}=\mathbf{\Lambda}} B_j + \frac{1}{2}\sum\nolimits_{j,k} \frac{\partial^2 \Omega_i(\mathbf{p})}{\partial p_j \partial p_k}\bigg|_{\mathbf{p}=\mathbf{\Lambda}} C_{jk} = 0 \tag{A.2}$$

may be solved for C and B.

We begin using the properties of left and right eigenvectors of A

$$A\psi_i = \lambda_i \psi_i, \qquad \chi_i^T A = \lambda_i \chi_i^T \tag{A.3}$$

$$\chi_i^T \psi_j = \delta_{ij}, \qquad \sum_i \psi_i \, \chi_i^T = I \text{ (unit } L \times L \text{ matrix)} \tag{A.4}$$

Operating on (A.1) with χ_i^T from the left and χ_j from the right and using the eigenvalue equation, (A.3), it follows that $\chi_i^T C \chi_j = \chi_i^T \Gamma \chi_j / (\lambda_i + \lambda_j)$. Then using (A.4) yields an explicit solution to (A.1):

$$C = \sum\nolimits_{i,j} \frac{\psi_i \chi_i^T \Gamma \chi_j \psi_j^T}{\lambda_i + \lambda_j}. \tag{A.5}$$

(4.15) is then solved by

$$B_i = \frac{1}{2} \sum\nolimits_{j,k,l} \left(A^{-1}\right)_{ij} \left.\frac{\partial^2 \Omega_j(\mathbf{p})}{\partial p_k \partial p_l}\right|_{\mathbf{p}=\mathbf{\Lambda}} C_{kl}. \tag{A.6}$$

Thus combining (4.19), (4.20) leads to explicit predictions for the mean allele frequencies along with their covariances in the limit of large N, for an arbitrary number of alleles.

B. Standard set 1

In this Appendix, we give a set of mutation rates and fitnesses that were generated randomly. Results for the mean allele frequencies and the matrix of covariances are calculated from these and in the main text, the results are compared with numerical simulations.

We take

$$\text{Number of alleles } n = 10 \tag{B.1}$$

$$\text{Population size } N = 2000 \tag{B.2}$$

and

$$\mu_{std} = 10^{-3} \times \begin{pmatrix} 0 & 5 & 5 & 3 & 1 & 2 & 5 & 5 & 6 & 8 \\ 0 & 0 & 1 & 6 & 6 & 5 & 3 & 9 & 7 & 4 \\ 7 & 0 & 0 & 8 & 9 & 9 & 1 & 1 & 7 & 8 \\ 7 & 4 & 4 & 0 & 3 & 9 & 9 & 8 & 10 & 3 \\ 9 & 1 & 7 & 4 & 0 & 1 & 1 & 8 & 9 & 4 \\ 4 & 4 & 9 & 2 & 8 & 0 & 5 & 8 & 2 & 5 \\ 5 & 7 & 8 & 10 & 5 & 5 & 0 & 1 & 3 & 5 \\ 8 & 6 & 3 & 7 & 2 & 5 & 3 & 0 & 4 & 3 \\ 0 & 9 & 0 & 8 & 3 & 3 & 9 & 7 & 0 & 2 \\ 1 & 8 & 7 & 7 & 4 & 10 & 5 & 9 & 6 & 0 \end{pmatrix} \tag{B.3}$$

$$w_{std} = 10^{-3} \times \begin{pmatrix} 943 & 918 & 963 & 954 & 955 & 921 & 918 & 954 & 945 & 946 \\ 918 & 984 & 942 & 939 & 986 & 989 & 965 & 987 & 945 & 951 \\ 963 & 942 & 976 & 984 & 956 & 923 & 960 & 986 & 920 & 949 \\ 954 & 939 & 984 & 941 & 945 & 939 & 905 & 933 & 980 & 929 \\ 955 & 986 & 956 & 945 & 951 & 937 & 931 & 971 & 947 & 943 \\ 921 & 989 & 923 & 939 & 937 & 975 & 970 & 963 & 966 & 964 \\ 918 & 965 & 960 & 905 & 931 & 970 & 911 & 973 & 967 & 961 \\ 954 & 987 & 986 & 933 & 971 & 963 & 973 & 919 & 923 & 918 \\ 945 & 945 & 920 & 980 & 947 & 966 & 967 & 923 & 918 & 914 \\ 946 & 951 & 949 & 929 & 943 & 964 & 961 & 918 & 914 & 915 \end{pmatrix} \tag{B.4}$$

C. Standard Set 2

In this Appendix, we give a second set of mutation rates and fitnesses that correspond to strong selection. Results for the mean allele frequencies and the matrix of covariances are calculated from these and in the main text, the results are compared with numerical simulations.

We take the same number of alleles population size and mutation rates as used in Standard Set No. 1, i.e. as given by (B.1), (B.2) and (B.3).

The matrix of fitnesses is now given by

$$w_{std} = 10^{-2} \times \begin{pmatrix} 43 & 18 & 63 & 54 & 55 & 21 & 18 & 54 & 45 & 46 \\ 18 & 84 & 42 & 39 & 86 & 89 & 65 & 87 & 45 & 51 \\ 63 & 42 & 76 & 84 & 56 & 23 & 60 & 86 & 20 & 49 \\ 54 & 39 & 84 & 41 & 45 & 39 & 5 & 33 & 80 & 29 \\ 55 & 86 & 56 & 45 & 51 & 37 & 31 & 71 & 47 & 43 \\ 21 & 89 & 23 & 39 & 37 & 75 & 70 & 63 & 66 & 64 \\ 18 & 65 & 60 & 5 & 31 & 70 & 11 & 73 & 67 & 61 \\ 54 & 87 & 86 & 33 & 71 & 63 & 73 & 19 & 23 & 18 \\ 45 & 45 & 20 & 80 & 47 & 66 & 67 & 23 & 18 & 14 \\ 46 & 51 & 49 & 29 & 43 & 64 & 61 & 18 & 14 & 15 \end{pmatrix} \tag{C.1}$$

References

Crow, J. F. and Kimura, M. (1970): *An Introduction to Population genetics Theory.* New York: Harper and Row

Felsenstein, J. (1981): *Bibliography of Theoretical Population Genetics.* Dowden

Turelli, M., (1984) Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theor. Popul.Biol. **25**: 138–193.

Bartlett, M. S.(1978): *An Introduction to Stochastic Processes*, Cambridge: CUP

Ewens, W. (1969): *Population Genetics.* London: Methuen

Kondrashov, A. S. (1995): Contamination of the genome by very slightly deleterious mutations - Why have we not died 100 times over? Journal of Theoretical Biology 175:583-594.

Peck, J. R., G. Barreau, and S. C. Heath. (1997): Imperfect genes, Fisherian mutation and the evolution of sex. Genetics **145**, 1171-1199.

Table 1

N	s_1	s_2	u	v	diffusion result $\bar{p}_1$	large N approx $\bar{p}_1$	diffusion result $\mathrm{cov}(p_1, p_2)$	large N approx $\mathrm{cov}(p_1, p_2)$
10^4	0.000	0.000	0.00040	0.00080	0.6667	0.6667	0.0045	0.0046
10^4	0.001	0.000	0.00040	0.00080	0.7759	0.7757	0.0030	0.0030
10^4	0.001	0.002	0.00040	0.00080	0.6683	0.6683	0.0030	0.0030
10^5	0.001	0.002	0.00004	0.00008	0.6676	0.6676	0.0007	0.0007
10^5	0.010	-0.010	0.00004	0.00008	0.9980	0.9980	3×10^{-7}	3×10^{-7}

Table 1 Caption

A set of results comparing the standard diffusion results and the large N approximate results of this work, for the case of a locus with two alleles.