Genet. Res., Camb. (1994), **64**, *pp.* 199–208 *With 4 text-figures*

The reduction in fixation probability caused by substitutions at linked loci

N. H. BARTON
Institute of Cell, Animal and Population Biology, University of Edinburgh, Kings' Buildings, Edinburgh EH9 3JT UK

(*Received 13 September 1994*)

Summary

The probability of fixation of a mutation with selective advantage s will be reduced by substitutions at other loci. The effect of a single substitution, with selective advantage $S \gg s$, can be approximated as a sudden reduction in the frequency of the favourable allele, by a fraction $w = 1-(s/S)^{r/S}$ (where r is the recombination rate). An expression for the effect of a given sequence of such catastrophes is derived. This also applies to the ecological problem of finding the probability that a small population will survive, despite occasional disasters. It is shown that if substitutions occur at a rate Λ, and are scattered randomly over a genetic map of length R, then an allele is unlikely to be fixed if its advantage is less than a critical value, $s_{\text{crit}} = (\pi^2/6)(2\Lambda S/(R\log(S/s)))$. This threshold depends primarily on the variance in fitness per unit map length due to substitutions, $\text{var}(W)/R = 2\Lambda S/R$. With no recombination, the fixation probability can be calculated for a finite population. If $\Lambda > s$, it is of the same order as for a neutral allele ($\approx \Lambda/(2N(\Lambda-s))$), whilst if $\Lambda \ll s$, fixation probability is much higher than for a neutral allele, but much lower than in the absence of hitch-hiking ($1/2N \ll 2s/(4Ns)^{\Lambda/s} \ll 2s$). These results suggest that hitch-hiking may substantially impede the accumulation of weakly favoured adaptations.

1. Introduction

Recent surveys have shown that DNA sequence variation may be influenced by selection at linked loci. Variation may be reduced by substitutions ('hitch-hiking'; Maynard Smith & Haigh, 1974), reduced by deleterious mutations ('background selection'; Charlesworth *et al.* 1990; Charlesworth, 1994), or increased by balancing selection ('associative over-dominance'; Ohta & Kimura, 1970). For example, there is greater nucleotide diversity in the region adjacent to the F/S polymorphism at the *Adh* locus of *Drosophila melanogaster* (Hudson *et al.* 1987), giving evidence that this polymorphism is maintained by balancing selection. Conversely, nucleotide diversity is lower in regions of the *Drosophila* genome with reduced crossing over (Aquadro & Begun, 1993). This may be explained by either adaptive substitutions (Kaplan *et al.* 1989) or deleterious mutations (Charlesworth *et al.* 1993; Charlesworth, 1994). As well as reducing neutral diversity, hitch-hiking also makes it less likely that favourable alleles will become established (Fisher, 1930; Muller, 1932; Hill & Robertson, 1966). This interference between selection at different loci impedes adaptation, and gives a long-term advantage to sex and recombination (Felsenstein, 1974, 1988).

Barton (1994) sets out a method for finding the probability of fixation of favourable alleles in a large population which is subdivided into a variety of genetic backgrounds. In particular, the favourable allele might be associated with an advantageous allele at another locus, increasing its chance of fixation, or it might be associated with the deleterious allele, decreasing its fixation probability. Overall, the fixation probability is reduced. The effect is greatest when the background substitution is driven by much stronger selection than that favouring the rare allele whose survival is in question. Then, hitch-hiking is equivalent to a sudden catastrophe that reduces the frequency of the rare allele by some factor which depends on the relative rates of recombination and selection. Any one event may have a small effect; however, a weakly favoured allele is likely to be vulnerable to extinction for many generations, and so may suffer from many hitch-hiking events. (An allele with advantage s will be vulnerable for $\approx 1/s$ generations; Barton, 1994). Thus, to find the net effect of a sequence of substitu-

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

tions at linked loci, the net effect of a sequence of catastrophes must be calculated. This is also relevant to ecological problems, where small populations may be in greatest danger from occasional catastrophes, rather than from demographic fluctuations (Mangel & Tier, 1993). In this paper, analytic and numerical results combine to show that the fixation probability declines linearly with the rate of substitutions. There is a threshold beyond which fixation becomes extremely unlikely in a large population. The case of complete linkage is studied in more detail, and yields results for large but finite populations.

2. Assumptions and general strategy

Suppose that a favourable allele with selective advantage s enters a very large population at $t = 0$, in a single copy ($n_0 = 1$). It then segregates at low frequency for some time, before either being lost, or increasing exponentially to fixation. Substitutions occur at other loci and involve alleles with advantage S. The population is assumed to be extremely large ($Ns \gg 1$), so that by the time these substitutions have much effect on the allele in question, they are present at appreciable frequency, and so increase deterministically: we need not consider the stochastic fluctuations which they themselves survived before beginning their deterministic increase. Barton (1994) sets out numerical results for the reduction in fixation probability caused by substitutions at other loci. The net reduction is substantial only if these are fixed by relatively strong selection ($S \gg s$), and are moderately tightly linked ($r < S$). In this case, substitutions can be approximated by instantaneous jumps, which reduce the numbers of the allele of interest from n to wn ($0 < w < 1$). The factor w is given by:

$w = 1-(s/S)^{r/S}$ (1; eqn 7 of Barton, 1994).

Suppose that the favourable allele arises a time t before the substitution at the other locus. (Time here is counted from when that substitution reaches its midpoint at equal allele frequencies). Its probability of fixation is (from eqn A 1 of Barton, 1994):

$$P \approx \frac{2sw}{w+(1-w)\exp(st)} \quad \text{for} \quad St \ll 0, s \ll S. \qquad (2)$$

The allele is vulnerable to interference from hitch-hiking for a time $t \approx 1/s$, which may be very long if the allele has a slight advantage.

The above formula is an average over cases when the new favourable allele arises in coupling or in repulsion with the existing substitution. There is a small chance that the weakly favoured allele will occur in coupling with a highly favoured allele while the latter is still at low frequency. This will greatly increase its chance of fixation. However, this is always outweighed by the greater chance that the alleles arise in repulsion. Overall, hitch-hiking always reduces the chance that a favourable allele will be fixed.

On the approximation of eqns 1, 2, the net effect of a sequence of hitch-hiking events at times $t_1, \ldots, t_k$ reduces to the net effect of a sequence of catastrophes at those times. The probability of fixation $u(n_0)$ can be calculated as follows:

$$u(n_0) = \int_0^{2N} \psi(n_0, n_1, t_1) \int_0^{2N} \psi(w_1 n_1, n_2, t_2 - t_1) \ldots \times \int_0^{2N} \psi(w_k n_k, 2N, \infty) \ldots \mathrm{d}n_k \ldots \mathrm{d}n_2 \mathrm{d}n_1. \qquad (3)$$

Here, $\psi(n_0, n_1, t)$ is the chance that the allele is not lost, and changes under selection and drift from n_0 to n_1 in time t. By using the diffusion approximation, it can be calculated explicitly:

$$\psi(n_0, n_1, t) = \frac{z}{2n_1} \exp(-s(n_1+n_0)\coth(st/2)+s(n_1-n_0))\, I_1(z), \qquad (4)$$

where

$$z = \left(\frac{2s\sqrt{(n_0 n_1)}}{\sinh(st/2)}\right).$$

3. The effects of a series of hitch-hiking events

The expression for the transition probability $\psi(n_0, n_1, t)$ given in eqn 4 can be substituted into eqn 3, to give the probability of fixation in the presence of substitutions at other loci. For simplicity, I first consider a single substitution, and then go on to find an explicit formula for the effect of a series of substitutions which occur at known times (t_i), and have known effects (w_i). The final step is to average over the distribution of times and effects.

(i) *The effect of a single substitution*

Suppose that the weakly advantageous allele is introduced in n_0 copies; after t generations, a rapid substitution occurs at another locus, which reduces the allele from n_1 copies to wn_1 copies. Since the probability of ultimate fixation immediately after this sequence of events is $(1-\exp(-2swn_1))$, the net probability is:

$$u(n_0) = \int_0^{2N} \psi(n_0, n_1, t_1)(1-\exp(-2swn_1))\,\mathrm{d}n_1. \qquad (5a)$$

Substituting from eqn 4, assuming that $2N \gg 1$, and using the result in Abramowitz & Stegun (1965, 11.4.31):

$$u(n_0) = 1-\exp\left(\frac{-2wn_0 s}{w+(1-w)\exp(-st)}\right) \qquad (5b)$$

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

This equation is exact, given the assumption that Ns is large. If a single copy is introduced ($n_0 = 1$), then since selection is assumed to be weak ($s \ll 1$), eqn 5*b* reduces to eqn 2, which was derived by a different route in eqn A 1 of Barton (1994). If the catastrophe occurs just after the introduction of the allele, $\exp(-st) \approx 1$, and $u(n_0) = (1-\exp(-2swn_0))$; this is the same value as if it had been introduced in wn_0 copies. If it occurs long after the introduction, $\exp(-st) \approx 0$, and the probability of fixation is not affected: $u(n_0) = (1-\exp(-2sn_0))$.

(ii) *The effect of a sequence of substitutions*

As shown below, with a given set of one or more hitch-hiking events, the fixation probability must have the form $u(n) = (1-\exp(-2sn_0/\theta)) = 1-[\exp(-2s/\theta)]^{n_0}$. Alleles introduced at low frequency are lost independently of each other; thus, the probability that n_0 are all lost is the product of the probabilities that each one is lost: $1-u(n_0) = (1-u(1))^{n_0}$. The factor θ can be thought of as the factor by which the effective selection pressure favouring the allele is reduced; eqn 5*b* shows that a single event changes θ from 1 to

$$[1+(1/w-1)\exp(-st))] = [(1-\exp(-st))+\exp(-st)/w].$$

Because the fixation probability $u(n_0)$ retains the same form for any number of hitch-hiking events, it can be integrated repeatedly to give the effect of an arbitrary series of substitutions. By induction:

$$\theta = \left((1-y_1)+\frac{y_1}{w_1}\left((1-y_2)+\frac{y_2}{w_2}\left((1-y_3)+\frac{y_3}{w_3}(\ldots)\right)\right)\right) \quad (6)$$

(where $\exp(-st_i) = y_i$). If there are n events in all, the series is terminated by setting y_{n+1} to 0 (corresponding to the $n+1$th event occurring at $t = \infty$). For example, suppose that one substitution occurs soon after the weakly selected mutant arises ($y_1 = \exp(-st_1) = 0{\cdot}8$), but only reduces the mutant's frequency by $w_1 = 0{\cdot}8$. A second substitution occurs much later ($y_2 = \exp(-st_2) = 0{\cdot}8$), but has a larger effect ($w_2 = 0{\cdot}5$). The probability of fixation is reduced by a factor $\theta = (0{\cdot}2+(0{\cdot}8/0{\cdot}8)(0{\cdot}9+0{\cdot}1/0{\cdot}5)) = 0{\cdot}2+1{\cdot}1 = 1{\cdot}3$. Thus, the first substitution is twice as important as the second in raising θ above 1.

(iii) *The distribution of times and effects*

The next step is to average over the distribution of possible hitch-hiking events (i.e. over their times and effects, t and w). The simple form $u(n_0) = (1-\exp(-2sn_0/\theta))$ will now be lost: this is because the chance that one allele will be lost becomes correlated with the chance that others are lost: all are affected by the same events. This correlation will reduce the fixation probability of alleles present in large numbers, because these may all be lost if a strong substitution occurs close by.

We assume that substitutions occur randomly and independently; since substitutions may be clustered in time (for example, around periods of environmental change), this may underestimate the effects of hitch-hiking. However, because only that small fraction of substitutions which happen to be tightly linked ($r < S$) to the locus of interest will significantly interfere with it, this may not introduce much error. With this assumption, the distribution of times between events is $\Lambda\exp(-\Lambda t)$, where Λ is the total rate of substitutions anywhere in the genome. It will be convenient to change variables from t to $y = \exp(-st)$; the distribution of y is $\tilde{\Lambda}y^{\tilde{\Lambda}-1}$, where $\tilde{\Lambda} = \Lambda/s$ is the expected number of events during the time $1/s$ when the new allele is vulnerable to loss.

The effect of a strongly selected substitution is equivalent to a catastrophe $w = 1-(s/S)^{-r/S}$ (eqn 1). If there is a single short chromosome of map length R, with the locus of interest at its centre, then r will be uniformly distributed between 0 and $R/2$; w is therefore distributed between 0 and $1-\epsilon$, with density $\phi/(1-w)$ (where $\phi = 2S/(R\log_e(S/s))$, $\epsilon = \exp(-1/\phi)$). This model of the genetic map is quite unrealistic, since it neglects the cumulative effect of unlinked loci, and so will underestimate the effect of hitch-hiking. However, because the calculations below show that only closely linked loci have a significant effect, the configuration of distant loci makes little difference.

The net fixation probability is given by a multiple integral over $y_i = \exp(-st_i)$ and w_i. This is intractable. However, explicit results can be obtained for the probability of fixation of a single mutant in two extreme cases: where substitutions are rare ($\tilde{\Lambda} \ll 1$), and where the selection driving each substitution is weak relative to the map length ($\phi \ll 1$).

$$u(1) = 2s\left[1-\tilde{\Lambda}\phi\left\{\frac{\pi^2}{6}-Li_2(e^{-1/\phi})\right\}\right] \quad (7a;\quad \tilde{\Lambda} \ll 1)$$

$$u(1) = 2s\left[1-\tilde{\Lambda}\phi\left(\frac{\pi^2}{6}\right)\right] \quad (7b;\quad \phi \ll 1)$$

These are derived in the Appendix; Li_2 is the second polylogarithm function, which decreases monotonically. $Li_2(e^{-1/\phi})$ tends to zero as ϕ tends to zero, showing that the two results are consistent. There might be many substitutions at other loci while a weakly selected allele is climbing to high frequency, so that $\tilde{\Lambda}$ may be large. However, the selection associated with adaptive substitutions is likely to be small, compared with the map length of sexually reproducing higher organisms ($\phi \ll 1$). Thus, the most relevant formula is the simplest one, eqn 7*b*.

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

(iv) *Numerical results for large* $\tilde{\Lambda}\phi$

To check the derivation of eqn 7, and to investigate cases where hitch-hiking has large effects ($\tilde{\Lambda}\phi \approx 1$), the fixation probability was calculated directly by Monte Carlo integration. A series of catastrophes was simulated, by drawing random values of t and w from the appropriate distributions. These values were put into eqn 6, giving the fixation probability for that sequence of events. This procedure was repeated to give the expected fixation probability. The results agree well with the simple prediction from (7*b*), up to $\approx \tilde{\Lambda}\phi = 0{\cdot}5$ (Fig. 1). For small $\tilde{\Lambda}\phi$, the discrepancy is most noticeable when $\phi = 0{\cdot}4$ (Fig. 1*b*). It is accounted for by the second term in (7*a*). The two straight lines in Fig. 1*b* show the predictions of eqns 7*a*, 7*b*: the upper line, which corresponds to the prediction appropriate for large ϕ (eqn 7*b*), fits best.

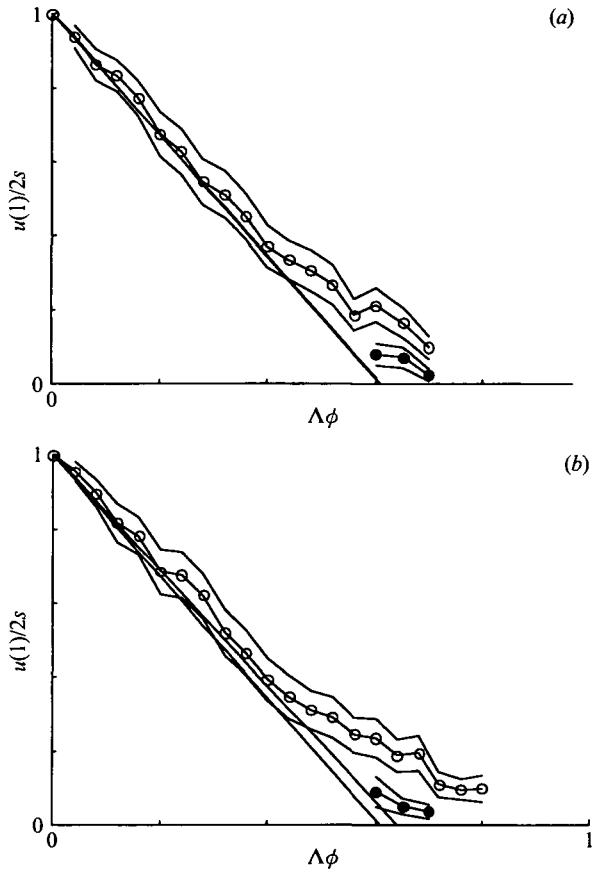

Fig. 1. The reduction in fixation probability due to hitch-hiking events at other loci. The graphs show the ratio by which fixation probability is reduced ($u(1)/2s$), as a function of the product $\tilde{\Lambda}\phi = (\Lambda/s)(2S/(R\log_e(S/s)))$. Each point shows the mean for 100 random sequences; the flanking lines show 95% confidence intervals. Each sequence consisted of k randomly generated hitch-hiking events. For the main series, k was the larger of four or $10\tilde{\Lambda}$; to give greater accuracy, $k = 40\tilde{\Lambda}$ was used for the lower series of three points near to the threshold. Each graph also shows two straight lines. The lower is the prediction from eqn 7*b*, and the upper is the prediction from eqn 7*a*. These are indistinguishable for $\phi = 0{\cdot}2$ (Fig. 1*a*), but differ slightly for $\phi = 0{\cdot}4$ (Fig. 1*b*).

The numerical results fit closely with the linear predictions of eqn 7, even though these were derived by asymptotic arguments valid only for small Λ or ϕ. This suggests that the fixation probability declines to zero if hitch-hiking events occur more frequently than a critical value close to $\tilde{\Lambda}\phi = 6/\pi^2 \approx 0{\cdot}61$. For large $\tilde{\Lambda}\phi$, numerical estimates of the fixation probability are inaccurate, because only a finite sequence was simulated. However, if the number of catastrophes used in the simulations is increased fourfold (from $10\tilde{\Lambda}$ to $40\tilde{\Lambda}$), the fixation probability decreases towards the prediction (sequence of three points at lower right of Fig. 1).

(v) *The threshold*

This threshold behaviour can be understood by considering the increase of the favourable allele in the long term. Suppose that the favourable allele is present in large enough numbers that it increases deterministically ($\approx \exp(st)$), but is still rare ($1/s \ll n \ll 2N$). After some long period T, the number of hitch-hiking events will converge to $k = \Lambda T$, and numbers will have increased by a factor

$$\exp\left(sT + \sum_{i=1}^{k} \log(w_i)\right).$$

By the central limit theorem, the sum will approach a normal distribution around the expectation $kE(\log(w))$, and with variance $\mathrm{var}(\log(w))/k$. The long-term rate of increase will therefore converge to $\exp(T(s+\Lambda E(\log(w))))$. Averaging over the distribution of w gives:

$$(s+\Lambda E(\log(w)) = s - \Lambda\phi\left\{\frac{\pi^2}{6} - Li_2(e^{-1/\phi})\right\}. \qquad (8)$$

The two terms in eqn 8 correspond to the rate of increase due to the selective advantage, s, and the rate of decrease caused by a random sequence of hitch-hiking events. If the selective advantage is less than some critical value, s_{crit}, the number of copies of the allele are expected to decrease, and fixation becomes impossible. This argument confirms the threshold suggested by eqn 7*b*, and by the numerical results. One can also show by a similar argument that if $s < s_{crit}$, the denominator of eqn 6 tends to infinity.

These arguments show that eqn 7*b* is correct for small $\tilde{\Lambda}$ and for small ϕ, and that it also correctly predicts a threshold beyond which fixation is impossible. The numerical results suggest that eqn 7*b* is in fact correct for all $\tilde{\Lambda}$, ϕ, and is always equal to twice its long-term expected rate of increase, as given by eqn 8. It is tempting to justify this by a simple branching-process argument. However, this would not be valid, because alleles are not lost independently of each other: catastrophes eliminate many together.

The case of most interest is where the genetic map is long ($\phi \ll 1$), but substitutions are frequent relative

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

to the selection on the allele of interest ($\tilde{\Lambda} \gg 1$). Then, eqn. 7*a* is accurate, and the threshold value is:

$$s_{\text{crit}} = \frac{\pi^2}{6}\Lambda\phi = \frac{\pi^2}{3}\frac{\Lambda S}{R\log_e(S/s)}. \quad (9a)$$

The net variance in fitness associated with adaptive substitutions is $\text{var}(W) = 2\Lambda S$ (Crow, 1970), and so:

$$s_{\text{crit}} = \frac{\pi^2}{6}\frac{\text{var}(W)}{R\log_e(S/s)}. \quad (9b)$$

This critical value only depends only logarithmically on the relative selection coefficients on substitutions and on the allele in question. It therefore depends primarily on $(\text{var}(W)/R)$, the variance in fitness associated with the substitution of new mutants, per unit map length.

4. No recombination

The results derived above suggest that with complete linkage ($\phi \to \infty$), the probability of fixation should become very small. In this section, I derive an expression for this probability which applies to large but finite populations, and which is exact, given the diffusion approximation and the assumption that Ns is large. As well as being relevant to strictly asexual organisms, this will give some insight into the way the fixation probability tends to zero as population size tends to infinity when $s < s_{\text{crit}}$.

Much of the previous work on the effects of hitch-hiking on the rate of adaptation has dealt with the limit of no recombination. However, most authors have considered the mutual interference between substitutions at different loci, with each having a similar selective advantage: this is the most obvious representation of Fisher's (1930) and Muller's (1932) argument. Hill & Robertson (1966) gave a wide range of simulation results for two loci; they explained these using some heuristic arguments for the limit of no recombination. Maynard Smith (1971) compares the rates of evolution of sexual and asexual populations, assuming that favourable alleles all have the same advantage. Felsenstein (1974) simulated substitutions at many loci, and extended previous work by Crow & Kimura (1965) to find a rough analytic approximation to the fixation probability. More recently, he has given a simple formula for the case where two mutants arise in the same generation (Felsenstein, 1988). Keightley (1991) gives analytic and simulation results using a model where the selective effects of new mutations are chosen at random. Here, I consider the influence of strongly selected substitutions on a weakly selected allele. This greatly simplifies the analysis, since one need only consider the influence of strong substitutions on the locus of interest, thus avoiding consideration of interactions among potentially large numbers of segregating genes. It is also biologically reasonable: only weakly selected alleles are substantially affected by hitch-hiking, and only strongly selected alleles substantially contribute to hitch-hiking.

Suppose that the first substitution occurs t_1 generations after the weakly favoured mutant arises: that mutation has by then reached a frequency $p_1 = n_1/2N$. If the strongly selected mutation occurs on a chromosome carrying the previous mutant, then it will carry that mutant to fixation; otherwise, it will displace it. Since the probability that the new mutation arises in coupling with the old is p_1, the probability of fixation, $u(1)$, is just the expectation of p_1. If substitutions are common ($\Lambda \gg s$), this will be close to its initial value of $1/2N$ when the first strong substitution occurs; $u(1)$ will therefore be close to that for a neutral variant. If substitutions are rare ($\Lambda \ll s$), then by that time, the weakly favoured allele will either have been lost already, or will have approached 1 with probability $2s$. The net fixation probability will thus be $2s$, and hitch-hiking will have a negligible effect.

During its early life, the weakly selected mutant is likely to be rare, and so selection will act as a linear force ($\Delta p = sp$). Since drift does not change the expected frequency, the net expectation (including cases of both fixation and loss) is just $\exp(st)/2N$. The fixation probability is the expectation of $\exp(st_1)/2N$, taken over the distribution of times, t_1, of the first hitch-hiking event:

$$u(1) = \frac{1}{2N}\int_0^\infty \Lambda\, e^{-\Lambda t} e^{st}\, dt = \frac{1}{2N}\frac{\Lambda}{\Lambda - s} = \frac{1}{2N}\frac{\tilde{\Lambda}}{\tilde{\Lambda}-1}. \quad (10)$$

As hitch-hiking events become very frequent ($\Lambda/s = \tilde{\Lambda} \gg 1$), $u(1)$ tends to the neutral value, $1/2N$.

The approximation that $E(p_1) = \exp(st_1)/2N$ breaks down completely when $\tilde{\Lambda} < 1$, since then, it is likely that the allele will have risen to high frequency before the first hitch-hiking event occurs. This is true even if Ns is extremely large. The expected frequency for arbitrarily long times can be found in the same way as above, with the difference that instead of approximating by an exponential, the logistic increase of the favourable allele is followed. Deterministic increase from an initial frequency of $(1/2N)$ gives $p = 1/(1+2N\exp(-st))$ after t generations. However, random sampling drift during the establishment of the allele accelerates or delays the eventual increase ($p = 1/(1+2N\exp(-s(t+\tau)))$). Given that an allele has increased to high frequency despite drift, the increase tends to be accelerated: $\exp(s\tau)$ is exponentially distributed with mean $1/2s$ (from eqn 4). Taking into account those cases where the allele is lost while rare, the expected frequency after t generations is the expectation of $2s/(1+2N\exp(-s(t+\tau)))$:

$$E\left(\frac{2s}{(1+2N\exp(-s(t+\tau)))}\right) = \int_{\tau=-\infty}^{\infty} \frac{2s}{(1+2N\exp(-s(t+\tau)))} \times \exp(-2s\exp(s\tau))\, d(2s\exp(s\tau)).$$

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

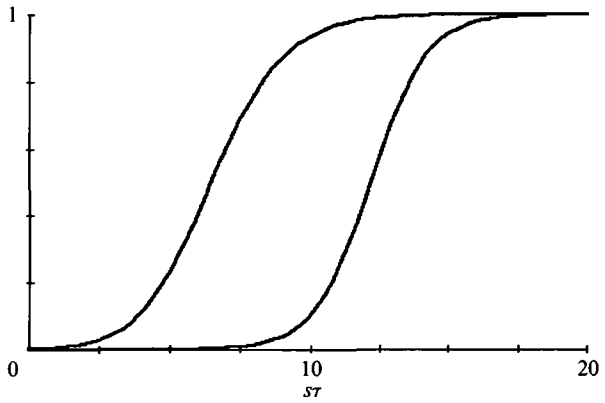

Fig. 2. The left-hand curve shows the expected frequency of an advantageous allele that is destined to be fixed (eqn 11). This is compared with the deterministic prediction, shown on the right. In this example, $N = 100000$ and $s = 0{\cdot}001$.

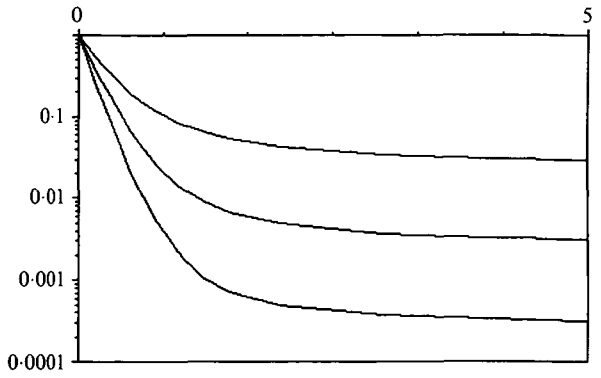

Fig. 3. The reduction in fixation probability when there is no recombination. The ratio $u(1)/2s$ is plotted on a log scale, against $\tilde{\Lambda} = \Lambda/s$, for $Ns = 10$, 100, and 1000 (top to bottom). The asymptote at the right is $(1/2N)/2s = 1/4Ns$.

Substituting $z = 2s\exp(s\tau), C = 4Nse^{-st}$:

$$= \int_0^\infty \frac{2sz\,e^{-z}}{z+C}dz = 2s[1 - Ce^C E_1(C)] \qquad (11)$$

(where $E_1(x)$ is the exponential integral, $\int_x^\infty e^{-z}/z\,dz$). This expected frequency is plotted as the leftmost curve in Fig. 2, for $Ns = 100$. It has almost the same form as the deterministic curve to the right, but is accelerated by $\tau \approx \log(1/2s)/s$. When t is small (so that $C \gg 1$), it reduces to $e^{st}/2N$, which agrees with the value derived above. Equation 11 therefore applies for all t.

This expectation must now be averaged over the distribution of $t_1, \Lambda e^{-\Lambda t_1}$. It is convenient to change variables from t_1 to $C = 4Nse^{-st_1}$, whose distribution is $(4Ns)^{-\tilde{\Lambda}}\tilde{\Lambda}C^{\tilde{\Lambda}-1}$. Integrating eqn 11 across this distribution gives:

$$u(1) = \frac{2s}{(4Ns)^{\tilde{\Lambda}}}\int_0^{4Ns}\int_0^\infty \frac{z\,e^{-z}\,\tilde{\Lambda}\,C^{\tilde{\Lambda}-1}}{(z+C)}dz\,dC. \qquad (12)$$

This can be written as the difference between the integral from $C = 0$ to ∞, and the integral from $C = 4Ns$ to ∞. The first can be evaluated explicitly, whilst the second gives an asymptotic expansion which will be accurate for large Ns:

$$u(1) = \frac{2s}{(4Ns)^{\tilde{\Lambda}}}\left[\int_0^\infty\int_0^\infty \frac{z\,e^{-z}\,\tilde{\Lambda}\,C^{\tilde{\Lambda}-1}}{(z+C)}dz\,dC - \int_{4Ns}^\infty\int_0^\infty \frac{z\,e^{-z}\,\tilde{\Lambda}\,C^{\tilde{\Lambda}-1}}{(z+C)}dz\,dC\right]$$

$$= \frac{2s[\tilde{\Lambda}\Gamma(\tilde{\Lambda})]^2\,\Gamma(1-\tilde{\Lambda})}{(4Ns)^{\tilde{\Lambda}}} - \sum_{k=1}^{\infty}\frac{2sk!\tilde{\Lambda}}{(4Ns)^k(k-\tilde{\Lambda})}. \qquad (13)$$

Note that this is an asymptotic series: the sum does not converge, but for a fixed number of terms, it gives an increasingly accurate approximation as Ns becomes large. When $\tilde{\Lambda} < 1$, the first term dominates, whilst when $\tilde{\Lambda} > 1$, the first term in the sum dominates; this is the same as the formula derived by considering only rare alleles (eqn 10). Thus:

$$u(1) = \frac{2s[\tilde{\Lambda}\Gamma(\tilde{\Lambda})]^2\,\Gamma(1-\tilde{\Lambda})}{(4Ns)^{\tilde{\Lambda}}} + O\left(\frac{1}{4Ns}\right) \quad (\tilde{\Lambda} < 1) \qquad (14a)$$

$$= \frac{2s}{(4Ns)^{\tilde{\Lambda}}} + O(\tilde{\Lambda}) + O\left(\frac{1}{4Ns}\right) \quad (\tilde{\Lambda} \ll 1), \qquad (14b)$$

$$u(1) = \frac{\tilde{\Lambda}}{2N(\tilde{\Lambda}-1)} + O\left(\frac{1}{(4Ns)^{\tilde{\Lambda}}}\right) + O\left(\frac{1}{(4Ns)^2}\right) \quad (\tilde{\Lambda} > 1). \qquad (14c)$$

Figure 3 shows the proportion by which the probability of fixation is reduced ($u(1)/2s$), as a function of the rate of hitch-hiking events, $\Lambda/s = \tilde{\Lambda}$. These curves were calculated using two terms from the sum in eqn 13, which gives essentially the same values as those found by numerical integration of eqn 11. However, the approximations of eqn 14*a*, *c* are very accurate, except near to $\tilde{\Lambda} = 1$.

This analysis shows that when Ns is large, there are two qualitatively different regimes. When substitutions are common, relative to the rate of increase of the new allele ($\Lambda > s$, or $\tilde{\Lambda} > 1$), the probability of fixation is equal to the neutral value ($1/2N$), multiplied by a factor which depends only on $\tilde{\Lambda}$, and is of order 1. Thus, the probability of fixation becomes extremely small in a large population. When substitutions are relatively rare ($\Lambda < s$, or $\tilde{\Lambda} < 1$), the probability of fixation is very much larger than the neutral value, but very much smaller than in the absence of hitch-hiking ($(1/2N) \ll 2s/(4Ns)^{\tilde{\Lambda}} \ll 2s$). These relations are illustrated in Fig. 4, which shows the probability of fixation of an allele with advantage s, in a population of 10^8. If other loci were not evolving, or if there were free recombination, alleles with extremely small advantages ($\approx 10^{-8}$) could accumulate (albeit slowly). However, if substitutions occur at a rate Λ, alleles with $s < \Lambda$ will be effectively neutral. If s is larger than Λ (more precisely, larger than $\Lambda\log(4Ns)$), then hitch-hiking has little effect (eqn 14*c*, Fig. 4).

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

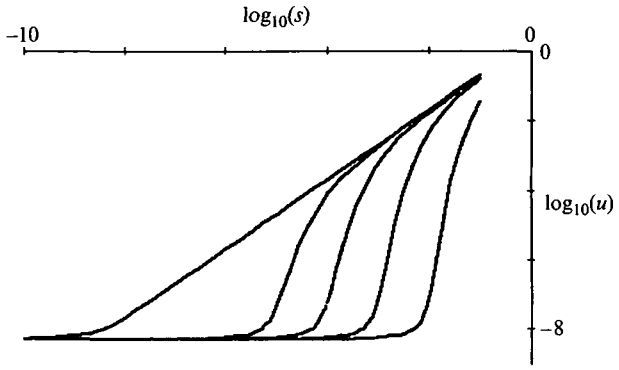

Fig. 4. The fixation probability, $u(1)$, in the absence of recombination, as a function of selective advantage, s. Both are plotted on a $\log_{10}$ scale, for a population size of $N = 10^8$. The leftmost curve applies when there is no hitch-hiking; the subsequent curves are for $\Lambda = 10^{-5}$, 10^{-4}, 10^{-3}, 10^{-2}. These were calculated from the series in eqn 13, using two terms in the sum.

5. Discussion

The effect of sporadic substitutions on the probability of fixation of a favourable mutation depends primarily on the variance in fitness which they produce, per unit map length. If the advantage of a new mutation is smaller than this, then its chance of fixation is substantially reduced, and tends to zero as population size increases (eqn 9*b*). If there is no recombination at all, the fixation probability depends critically on whether a substitution is likely to occur at another locus before the new allele has reached high frequency: if it is, the probability is reduced to not much more than that for a neutral allele ($\approx (1/2N)(\Lambda/(\Lambda - s))$ if $\Lambda > s$; eqns 10, 14*c*).

The above analysis can be applied to ecological as well as genetical questions: it gives the probability that a small population will become established (that is, will rise above some threshold size) despite the twin hazards of demographic fluctuations and ecological catastrophes. Mangel & Tier (1993) review the considerable literature on this topic, and set out a general algorithm for finding the expected time to extinction, and the probability of establishment. However, this algorithm assumes that the *number* of individuals which die in a catastrophe approaches a definite limit when the population is small. It therefore does not apply to the case analysed here, where catastrophes eliminate a fixed *proportion* of individuals. It is this assumption which leads to a threshold growth rate, s_{crit}, below which extinction is certain. This is the growth rate which just balances the proportion lost through catastrophes. If a constant number were culled, the proportion culled would decrease as the population grew, so that establishment would become almost certain if the population were sufficiently large.

The key approximation in the analysis of genetic hitch-hiking was that only the influence of substitutions of large effect on those of small effect need be considered. In fact, there must be a continuous range of selection coefficients. However, this is not restrictive. First, the actual rate of substitution, Λ, has been taken as given; this may have itself been reduced by interference between loci, but that does not affect calculations of the effects of those substitutions that do occur. Secondly, the analysis in Barton (1994) applies to alleles of arbitrary selective effect, but showed that there is only a substantial net effect when the selection on the vulnerable allele is much weaker than on the substitution which interferes with it. In that case, hitch-hiking can be approximated by a sudden catastrophe that culls a certain fraction of individuals. Thirdly, the effect of a substitution depends on its contribution to the variance in fitness: alleles with advantages weak enough to be influenced by hitch-hiking will make a negligible contribution to this variance, and so will not themselves have a significant effect on other loci. Thus, there will be a group of alleles which are selected strongly enough for them to be little affected by other loci, but not so strongly that they themselves cause hitch-hiking effects. These moderately selected alleles separate the two classes of loci which are considered in this paper.

Whether hitch-hiking of the sort analysed here significantly impedes adaptation depends primarily on the variance in relative fitness per unit map length caused by substitutions. There is almost no direct evidence on the net rate of adaptive substitutions, or their selective effect. However, Haldane (1957) suggested that the 'substitution load' sets an upper limit of one substitution every 30 generations for mammals ($\Lambda < 1/30$). A similar value is obtained if one assumes that most amino-acid substitutions are adaptive (Gillespie, 1992, p. 41). If S averages 5%, the variance in fitness is $\mathrm{var}(W) = 2\lambda S = 0{\cdot}0033$; spread over a map of length 10 Morgans, eqn 9 gives $s_{crit} = 0{\cdot}85 \times 10^{-4}$. While such weakly selected alleles cannot be responsible for the recently evolved adaptations that distinguish related species, they may contribute to long-term molecular adaptations such as codon usage bias. This issue is discussed in more detail in Barton (1994).

The effect of hitch-hiking on the probability of fixation differs qualitatively from the effect of neutral heterozygosity, and hence could not be derived by defining a 'variance-effective' population size (Crow & Kimura, 1970). The variance-effective size is only useful for calculating the variance in allele frequency, and statistics such as the average heterozygosity which depend on that variance. It is misleading when applied to find other quantities. Using the diffusion approximation, one would calculate the probability of fixation in a large population as $2s(N_e/N)$, where N_e is defined as the population size that would produce the same variance in allele frequency as does hitch-hiking. Hitch-hiking can reduce the probability of fixation of a favourable allele to indefinitely small values, and yet does not cause an indefinite increase in the rate of drift

of neutral alleles, or an indefinite reduction in neutral heterozygosity (Maynard Smith & Haigh, 1974; Birky & Walsh, 1988; Kaplan *et al.* 1989; Stephan *et al.* 1992). Moreover, the effect on fixation probability depends on the ratio of selection coefficients (s/S), whereas the effect on neutral heterozygosity depends on $2NS$ (Stephan *et al.* 1992). The effects differ because hitch-hiking produces occasional catastrophes, which cannot be adequately modelled using the diffusion approximation alone.

DNA sequence variation gives good evidence of hitch-hiking (e.g. Hudson *et al.* 1987; Aquadro & Begun, 1993; Kaplan *et al.* 1989; Charlesworth *et al.* 1993). However, hitch-hiking has different effects on neutral diversity and on fixation probabilities. Stephan *et al.* (1992, eqn 17) show that a single substitution reduces neutral heterozygosity by a factor $(2r/S)(2NS)^{-2r/S}\Gamma(-2r/S, 1/2NS)$, where

$$\Gamma(a,x) = \int_x^\infty \exp(-t)\, t^{a-1}\, \mathrm{d}t$$

is the incomplete gamma function. If the population is large ($NS \gg 1$), this approximates to $1-(2NS)^{-2r/S}$. In contrast, hitch-hiking reduces fixation probability by a factor $1-(S/s)^{-r/S}$. Since we consider alleles with a substantial advantage ($s \gg 1/2N$), and since the exponents differ ($2r/S$ *vs.* r/S), fixation probability must always be reduced more than neutral heterozygosity. For example, consider a substitution favoured by $S = 10\%$, which occurs $r = 1$ cM from the locus in question; the population size is $N = 10^6$. Hitch-hiking reduces neutral heterozygosity by a factor $\approx 1-(2NS)^{-2r/S} = 0{\cdot}91$, but reduces the probability of fixation by $\approx 1-(S/s)^{-r/S} = 0{\cdot}50$. Thus, the observation that hitch-hiking has a substantial effect on neutral diversity implies a substantially greater effect on the accumulation of weakly favoured alleles.

Successive substitutions make it very unlikely that alleles with selective advantage below some threshold will be fixed. (More precisely, as the population size increases, the fixation probability tends to zero for alleles with $s < s_{\mathrm{crit}}$). The critical selective advantage below which hitch-hiking overwhelms natural selection is proportional to the additive genetic variance in fitness due to substitutions, per unit map length ($\mathrm{var}(W)/R$; eqn 25). The scanty evidence on fitness variation is discussed in more detail in Barton (1994). However, arguments based either on the substitution load (Haldane, 1957), or on the total number of amino-acid substitutions (Gillespie, 1992, p. 41) suggest the number of adaptive substitutions per generation must be low ($\lambda < 1/30$ say). Assuming (arbitrarily) an average selective advantage $S = 0{\cdot}05$, then the variance in fitness would be $2\lambda S = 0{\cdot}0033$. Over a map of length 10 Morgans, eqn 25 gives the critical selection pressure $s_{\mathrm{crit}} = 0{\cdot}85 \times 10^{-4}$. Thus, hitch-hiking is unlikely to thwart those alleles which confer a moderate advantage in outcrossing populations. However, it might well impede adaptations which involve detailed molecular adjustments (for example, bias in codon usage), and so sets a limit on the power of natural selection.

This work was supported by the SERC (GR/E/08507) and by the Darwin Trust. Thanks are due to W. G. Hill, K. S. Gale, M. Slatkin, J. Maynard Smith and M. Turelli for their helpful comments.

Appendix

The probability of fixation of a single mutation is $u(1) = 2s/\theta$. For any particular sequence of events, θ is given by eqn 6; the problem is to find the expectation of $2s/\theta$, taken over the distribution of $y_i = \exp(-st_i)$ and w_i. (These were derived above, and are, respectively $\tilde{\Lambda} y^{\tilde{\Lambda}-1}$ $(0 < y < 1)$, and $\phi/(1-w)$ $(0 < w < 1-\epsilon$; $\epsilon = \exp(-1/\phi))$). Here, the expectation is derived in two limits: where few substitutions occur during the lifetime of the weakly selected allele ($\tilde{\Lambda} = \Lambda/s \to 0$), and where the effect of each substitution is small ($\phi = 2S/(R\log(S/s)) \to 0$).

(i) *Infrequent substitutions*: $\tilde{\Lambda} \to 0$

For a specific sequence of events, the fixation probability is reduced by a factor:

$$\frac{u(1)}{2s} = \frac{1}{\theta}$$
$$= 1/1 + y_1\left(-1+\frac{1}{w_1}\left(1+y_2\left(-1+\frac{1}{w_2}\right.\right.\right.$$
$$\left.\left.\left.\times\left(1+y_3\left(-1+\frac{1}{w_3}(1+\ldots)\right)\right)\right)\right)\right). \quad \text{(A 1)}$$

Write θ as $1/(1+y_1 C_1)$, and integrate over the time of the first event y_1:

$$E\left(\frac{1}{\theta}\right) = \int_0^1 \frac{\tilde{\Lambda}\, y_1^{\tilde{\Lambda}-1}\, \mathrm{d}y_1}{(1+C_1 y_1)}. \quad \text{(A 2)}$$

In the limit $\tilde{\Lambda} \to 0$, this reduces to:

$$1 - \tilde{\Lambda}\log(1+C_1) + O(\tilde{\Lambda}^2). \quad \text{(A 3)}$$

The next step is to average over the distribution of effects of the first event, w_1. $(1+C_1)$ can be rewritten as $w_1/(1+y_2 C_2)$. The expectation over w_1 is thus:

$$1 - \tilde{\Lambda}\phi \int_0^{1-\epsilon} \frac{[\log(w_1) - \log(1+y_2 C_2)]\, \mathrm{d}w_1}{(1-w_1)} + O(\tilde{\Lambda}^2)$$
$$= 1 - \tilde{\Lambda}\phi\{Li_2(1) - Li_2(\epsilon)\} - \tilde{\Lambda}\log(1+y_2 C_2), \quad \text{(A 4)}$$

$Li_2(\epsilon)$ is the dilogarithm function, defined by

$$\int_{1-\epsilon}^1 \frac{\log(1/z)}{(1-z)}\, \mathrm{d}z.$$

The formula still includes the effects of the second and subsequent events through $y_2 C_2$; however, when

https://doi.org/10.1017/S0016672300032857 Published online by Cambridge University Press

hitch-hiking events are rare ($\tilde{\Lambda} \to 0$), the distribution of $y = \exp(-st)$ clusters around zero, and the last term in eqn A 4 can be neglected. (Formally, the expectation of $\tilde{\Lambda}\log(1+y_2 C_2)$ is of order $\tilde{\Lambda}^2$, and so can be ignored). Since $Li_2(1) = \pi^2/6$, the expectation over the whole sequence is given by eqn 7*a*. As one might expect, this is the same formula as would be obtained by considering only a single event: when $\tilde{\Lambda}$ is small, it is unlikely that two hitch-hiking events will occur while the weakly selected allele is on its way to fixation.

(ii) *Substitutions with small effects*: $\phi \to 0$

If $\tilde{\Lambda}$ is large, an allele is likely to be influenced by many events, and so the calculation is somewhat more complicated. However, a simple formula can be obtained when the expected effect of any one event is small. Equation A 1 can be rewritten:

$$\frac{u(1)}{2s} = \frac{1}{\theta} = \frac{1}{\alpha_1 + \beta_1/w_1}, \tag{A 5}$$

where

$$\alpha_1 = 1-y_1, \beta_1 = y_1\left(1-y_2+\frac{y_2}{w_2}\left(1-y_3+\frac{y_3}{w_3}(\ldots)\right)\right).$$

First, average over the effect of the first event, w_1:

$$\frac{1}{\theta} = \frac{1}{(\alpha_1+\beta_1)} - \frac{\phi\beta_1}{\alpha_1(\alpha_1+\beta_1)}\log\left(1+\frac{\alpha_1(1-\epsilon)}{\beta_1}\right) \tag{A 6}$$

$(\alpha_1+\beta_1)$ can be rewritten as $\alpha_2+\beta_2/w_2$, where $\alpha_i = 1-(y_1 y_2 \ldots y_i)$, and

$$\beta_i = (1-\alpha_i)\left(1-y_{i+1}+\frac{y_{i+1}}{w_{i+1}}\left(1-y_{i+2}+\frac{y_{i+2}}{w_{i+2}}(\ldots)\right)\right).$$

The first term can therefore be integrated as before. The process can be repeated: each time the first term is integrated, it produces another term proportional to ϕ. Because the product $(y_1 y_2 \ldots y_k)$ tends to zero for large k, $\alpha_k+\beta_k$ tends to 1 as $k \to \infty$. Hence we obtain:

$$1-\phi\sum_{k=1}^{\infty}\frac{\beta_k}{\alpha_k(\alpha_k+\beta_k)}\log\left(1+\frac{\alpha_k(1-\epsilon)}{\beta_k}\right). \tag{A 7}$$

So far, the equation is exact: it still depends on the w_k and y_k through α_k and β_k. However, because the second term is proportional to ϕ, we can evaluate the sum by taking the expectation over w in the limit $\phi \to 0$. This is equivalent to setting $w_k = 1$ throughout; then, $\beta_k = (y_1 y_2 \ldots y_k)$, and $\alpha_k = 1-\beta_k$. ϵ also tends to zero in this limit, and so:

$$1-\phi\sum_{k=1}^{\infty}\frac{\beta_k}{(1-\beta_k)}\log\left(\frac{1}{\beta_k}\right)+O(\phi^2). \tag{A 8}$$

It now remains to average over the times of the events. Now, $\beta_k = (y_1 y_2 \ldots y_k) = \exp(-s(t_1+t_2\ldots+t_k))$. The sum of k exponentially distributed variables has a gamma distribution; thus, β_k can be shown to have the distribution:

$$\psi(\beta_k) = \frac{\tilde{\Lambda}^k \log(1/\beta_k)^k \beta_k^{\tilde{\Lambda}-1}}{(k-1)!}. \tag{A 9a}$$

This can be rewritten as the kth differential of a generating function:

$$\psi(\beta_k) = \frac{-(-\tilde{\Lambda})^k}{\beta_k(k-1)!}\frac{\partial^{k-1}\{\beta_k^{\tilde{\Lambda}}\}}{\partial\tilde{\Lambda}^{k-1}}. \tag{A 9b}$$

The expectation of eqn A 8 is therefore:

$$1-\phi\sum_{k=1}-\frac{(-\tilde{\Lambda})^k}{(k-1)!}\frac{\partial^{k-1}}{\partial\tilde{\Lambda}^{k-1}}\left[\int_0^1\frac{\beta_k^{\tilde{\Lambda}}\log(1/\beta_k)}{(1-\beta_k)}\right]+O(\phi^2). \tag{A 10}$$

This is a Taylor's series. If the term in square brackets is denoted $f(\tilde{\Lambda})$, then the whole sum is $f(\tilde{\Lambda}-\tilde{\Lambda}) = f(0) = \pi^2/6$. Thus, the fixation probability is given by eqn 7*b* in this limit. It is remarkable that when substitutions have small average effects ($\phi \to 0$), the change in fixation probability is directly proportional to the rate of substitution, $\tilde{\Lambda}$.

References

Abramowitz, M. & Stegun, I. A. (1965). *Handbook of Mathematical Functions*. New York: Dover.

Aquadro, C. F. & Begun, D. J. (1993). Evidence for and implications of genetic hitch-hiking in the *Drosophila* genome. In *Mechanics of Molecular Evolution* (ed. N. Takahata and A. G. Clark), pp. 159–178. Sunderland, MA: Sinauer Press.

Barton, N. H. (1994). Linkage and the limits to natural selection. *Genetics* (in the press).

Birky, C. W. & Walsh, J. B. (1988). Effects of linkage on rates of molecular evolution. *Proceedings of the National Academy of Science (USA)* **85**, 6414–6418.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research* **63**, 213–228.

Crow, J. F. (1970). Genetic loads and the cost of natural selection. In *Mathematical Topics in Population Genetics* (ed. K. I. Kojima), pp. 128–177. Berlin: Springer-Verlag.

Crow, J. F. & Kimura, M. (1965). Evolution in sexual and asexual populations. *American Naturalist* **99**, 439–450.

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

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics* **78**, 737–756.

Felsenstein, J. (1988). Sex and the evolution of recombination. In *The Evolution of Sex* (ed. R. E. Michod and B. R. Levin), pp. 74–86. Sunderland, Massachussetts: Sinauer Press.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.

Gillespie, J. H. (1992). *The Causes of Molecular Evolution*. Oxford: Oxford University Press.

Haldane, J. B. S. (1957). The cost of natural selection. *Journal of Genetics* **55**, 511–524.

Hill, W. G. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hudson, R. R., Kreitman, M. & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **114**, 93–110.

Kaplan, N. L., Hudson, N. L. & Langley, C. H. (1989). The hitch-hiking effect revisited. *Genetics* **123**, 887–899.

Keightley, P. D. (1991). Genetic variance and fixation probabilities at quantitative trait loci in mutation-selection balance. *Genetical Research* **58**, 139–144.

Mangel, M.& Tier, C. (1993). Dynamics of metapopulations with demographic stochasticity and environmental catastrophes. *Theoretical Population Biology* **44**, 1–31.

Maynard Smith, J. (1971). What use is sex? *Journal of Theoretical Biology* **30**, 319–335.

Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

Muller, H. J. (1932). Some genetic aspects of sex. *American Naturalist* **66**, 118–138.

Ohta, T. & Kimura, M. (1970). Development of associative overdominance through linkage disequilibrium in finite populations. *Genetical Research* **16**, 165–177.

Stephan, W., Wiehe, T. H. & Lenz, M. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology* **41**, 237–254.