PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Genetic hitchhiking

N. H. Barton

Phil. Trans. R. Soc. Lond. B 2000 **355**, 1553-1562
doi: 10.1098/rstb.2000.0716

THE ROYAL SOCIETY

doi 10.1098/rstb.2000.0716

Genetic hitchhiking

N. H. Barton

Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK
(n.barton@ed.ac.uk)

Selection on one or more genes inevitably perturbs other genes, even when those genes have no direct effect on fitness. This article reviews the theory of such genetic hitchhiking, concentrating on effects on neutral loci. Maynard Smith and Haigh introduced the classical case where the perturbation is due to a single favourable mutation. This is contrasted with the apparently distinct effects of inherited variation in fitness due to loosely linked loci. A model of fluctuating selection is analysed which bridges these alternative treatments. When alleles sweep between extreme frequencies at a rate λ, the rate of drift is increased by a factor $(1 + E[1/pq]\lambda/(2(2\lambda + r)))$, where the recombination rate r is much smaller than the strength of selection. In spatially structured populations, the effects of any one substitution are weaker, and only cause a local increase in the frequency of a neutral allele. This increase depends primarily on the rate of recombination relative to selection (r/s), and more weakly, on the neighbourhood size, $Nb = 4\pi\rho\sigma^2$. Spatial subdivision may allow local selective sweeps to occur more frequently than is indicated by the overall rate of molecular evolution. However, it seems unlikely that such sweeps can be sufficiently frequent to increase significantly the drift of neutral alleles.

Keywords: genealogy; linkage disequilibrium; genetic diversity

1. INTRODUCTION

When a favourable mutation arises, and increases to fixation, it gives a fortuitous advantage to all the genes with which it was originally associated. Maynard Smith & Haigh (1974) termed this process 'hitchhiking' and showed that, in large populations, it could reduce neutral diversity much more than random genetic drift. In an asexual population, hitchhiking can be seen directly through the phenomenon of 'periodic selection' (Dykhuizen 1990): the steady increase in diversity at a marker locus caused by neutral mutation is punctuated by an abrupt loss of variation whenever a favourable substitution occurs. Even in a sexual population, any variants associated with a favourable mutation will increase until separated from it by recombination.

This article reviews the theory of genetic hitchhiking in its broadest sense. The term was originally coined to describe the effects of the substitution of a favourable mutation on linked loci and is sometimes still used in this restricted sense. However, I will use it to refer to the indirect effects of selection at one or more loci on the rest of the genome. Selection may be for alleles which are unconditionally favourable; it may act to eliminate deleterious mutations; or it may fluctuate in space or time. The surrounding loci may be neutral; they may modify the genetic system; or they may themselves be directly selected. The same processes underlie all these various cases, and so it is illuminating to consider them together. Indeed, the idea of 'hitchhiking' brings together apparently diverse aspects of population genetics.

Hitchhiking can be understood in many ways. It is caused by linkage disequilibria: that is, by statistical associations between the states of different genes. In the classical case, hitchhiking is caused by the complete association between the new mutation and the genome in which it happens to arise. This association is eroded by recombination, but while it exists, selection for the favourable mutation also increases the frequency of all the genes originally associated with it. Here, linkage disequilibria are generated by the random sampling of the one genome in which the mutation arises: in an infinite population, with all mutations recurring many times, there would on average be no linkage disequilibrium and no hitchhiking (Maynard Smith 1978, p. 112).

In any finite population, genetic drift generates random associations between polymorphic loci and hence causes selection on one locus to spill over onto others. Any one locus experiences random perturbations, which on average interfere with selection (Robertson 1961; Hill & Robertson 1966). This kind of hitchhiking can thus be understood as causing an amplification of random sampling drift, and a reduction in effective population size. In this article, I concentrate on linkage disequilibria generated by random genetic drift. However, associations generated by migration or by epistasis can interfere with selection in a similar way.

Hitchhiking is important for many reasons. It was first proposed as an explanation of why very abundant species do not show correspondingly high levels of genetic diversity (Maynard Smith & Haigh 1974). Conversely, the pattern of marker variation can reveal the action of selection in the surrounding genome. The idea that hitchhiking effects provide a way of inferring the overall amount and nature of selection is long standing (e.g. Clegg *et al.* 1976), but has been greatly stimulated by the present abundance of DNA sequence data. In particular, the low genetic diversity seen in regions of low recombination in *Drosophila* and other organisms (Aguade *et al.* 1989; Aquadro *et al.* 1994) must be due to some kind of selection. Sequence data have also stimulated the re-framing of population-genetic theory in terms of genealogies rather than allele frequencies, giving a more complete understanding of the effects of selection on neutral genes. In a separate development, studies of the evolution of genetic systems (and especially, the evolution of sex and recombination) have centred on understanding indirect selection on modifier loci. Moreover, because hitchhiking limits the ability of populations to respond to selection, it also generates selection for increased recombination rates.

In this article, I will bring together different kinds of hitchhiking, in order to show how they are related. Although the population genetics of linkage disequilibria among multiple loci can be complicated, some plausible approximations yield simple and general results. I begin by summarizing the classical analysis of hitchhiking by Maynard Smith & Haigh (1974) and then contrast this with an argument based on inherited variation in fitness, introduced by Robertson (1961). I show that these two ways of looking at hitchhiking can be seen as extremes, which can be connected by a simple model of fluctuating selection. I then extend Maynard Smith & Haigh's (1974) analysis to a spatially extended population, to show how genetic and spatial structure interact. Throughout, I restrict attention to hitchhiking effects on neutral genes. There is a considerable literature on the way hitchhiking interferes with selection (especially by reducing the chance of fixation of weakly favoured alleles) and a still larger literature on the evolution of genetic systems through indirect selection on modifier alleles. Though there is not space here to cover these areas, much of the theory for neutral loci carries over directly.

2. NEUTRAL GENES IN A SINGLE POPULATION

(a) *Maynard Smith and Haigh's deterministic argument*

In outline, Maynard Smith & Haigh's (1974) argument is as follows. Suppose that a single copy of a new favourable allele, P, arises by mutation at a locus previously fixed for allele Q. The new mutation increases heterozygous fitness by a factor $(1+s)$; $s \ll 1$. Its frequency, p, is assumed to increase deterministically from $p_0 = 1/2N$ to fixation at $p_\infty = 1$. Selection is assumed to be additive, so that PP homozygotes have relative fitness $1+2s$. (The fitness of the homozygotes makes little difference, because most of the hitchhiking effect occurs while P is at low frequency.) Approximating by a continuous time model, $(p/q) = (p_0/q_0)\exp(st)$.

Alleles U, V segregate at a linked neutral locus, at frequencies u, v; and the recombination rate is r. It is simplest to follow the frequencies of allele U within the two alternative genetic backgrounds defined by the selected alleles, u_P, u_Q. The coefficient of linkage disequilibrium is then $D = pq(u_P - u_Q)$, and the overall frequency of allele U is $u = u_P p + u_Q q$. If P originates with U, then initially $u_{P0} = 1$ and $u_{Q,0} = u_0$. Now, selection does not alter the proportions of the neutral alleles within each of the selected backgrounds. A proportion q of P backgrounds pair with Q backgrounds, and a fraction r of these undergo recombination. Thus, u_P changes at a rate $rq(u_Q - u_P)$, and conversely, u_Q changes at a rate $rp(u_P - u_Q)$. The difference $(u_P - u_Q)$ decreases by a factor $(1-r)$ in every generation, and so at time t, $(u_P - u_Q) = (u_{P0} - u_{Q,0})\exp(-rt)$. The final step is to calculate the total change in the frequency of U over the whole course of the substitution. The frequency $u = u_P p + u_Q q$ changes at a rate $(u_P - u_Q)\mathrm{d}p/\mathrm{d}t$ as a result of the change in p due to selection. That is, the net change in neutral allele frequency is

$$\Delta u = \int_0^\infty (u_P - u_Q)\frac{\mathrm{d}p}{\mathrm{d}t}\mathrm{d}t = \int_{p_0}^1 (u_P - u_Q)\mathrm{d}p,$$

$$= \int_{p_0}^1 (u_{P,0} - u_{Q,0})\mathrm{e}^{-rt}\mathrm{d}p = \int_{p_0}^1 (1-u_0)\left(\frac{p_0 q}{q_0 p}\right)^{r/s}\mathrm{d}p,$$

$$= (1-u_0){p_0}^{r/s} = (1-u_0)(2N)^{-r/s}$$

$$\text{for } r < s,\ p_0 = \frac{1}{2N} \ll 1. \qquad (1)$$

Maynard Smith & Haigh's (1974) equation 14 is the first-order approximation to equation (1) for $r \ll s$. Barton (1998, equation 3) gives a slightly more accurate approximation.

Equation (1) has a simple interpretation. A new mutation takes approximately $\log(2N)/s$ generations to increase from $p_0 = 1/2N$ to high frequency. During this time, its association with the neutral locus is dissipating at a rate r. The neutral marker only increases appreciably when the selected locus is itself increasing (say, from 10 to 90%), and the residual association by this time is just $\exp(-r\log(2N)/s) = (2N)^{-r/s}$. Thus, the hitchhiking effect decreases with population size, because it takes longer for a new mutation to reach a high frequency, by which time the genes it was originally associated with have become separated by recombination. However, the dependence on population size is only logarithmic.

There is a chance u_0 that the new mutation arises with allele U, and increases by approximately $v_0(2N)^{-r/s}$, as assumed above. There is a chance v_0 that P arises with V, in which case u decreases by $u_0(2N)^{-r/s}$. On average, there is no change in neutral allele frequency. However, the variance in allele frequency is $u_0 v_0 (2N)^{-2r/s}$, and on average the heterozygosity, $2u(1-u)$, decreases by a factor $(1-(2N)^{-2r/s})$. If substitutions occur at random locations on a genome of map length R, and at a rate Λ per generation, then the rate of loss of heterozygosity is approximately $s\Lambda/(R\log(2N))$. If the population size is sufficiently large, this will be greater than the rate of loss due to random genetic drift, $1/2N$.

While this calculation gives essentially the correct result, it ignores the initial random fluctuations in the frequency of the selected allele, and in the associations of the neutral marker with that allele. In particular, a favourable allele that is destined to fix is likely to increase faster than the deterministic expectation (Maynard Smith & Haigh 1974, p. 25). The expected frequency conditional on ultimate fixation is increased by a factor $1/2s$, and hence the expected increase in the neutral allele is increased by a factor $(2s)^{-r/s}$. This stochastic acceleration increases the variance in allele frequency by a factor $(2s)^{-2r/s}$, to $u_0v_0(4Ns)^{-2r/s}$. This and other stochastic complications are analysed by Stephan *et al.* (1992), Kaplan *et al.* (1989) and Barton (1998). Their full effect is to further increase the hitchhiking effect. However, for most purposes the deterministic argument of Maynard Smith & Haigh (1974) is sufficiently accurate. This is because the effect of a selected substitution depends primarily on the time between the origin of the mutation and its ultimate fixation. This time is approximately $\log(2N)$ generations, regardless of the initial random phase.

(b) *Genealogies*

A complete description of the effect of a selected substitution on a sample of neutral genes requires that we find the distribution of their genealogical relationships. In principle, this is a straightforward application of the 'structured coalescent' (Hudson 1990; Notohara 1990; Hey 1991; Kaplan *et al.* 1991; Gillespie 2000). As one traces lineages back in time, two processes occur: pairs of genes that are in the same genetic background coalesce at a rate equal to the inverse of the number of genes in that background, and genes move between backgrounds by recombination or by mutation at the loci that define the backgrounds. The fixation of a single mutation, considered above, is not strictly described by the structured version of Kingman's (1982) coalescent, which assumes that the number of lineages is much smaller than the total number of genes—in the early stages, the number of genes in the rare genetic background is small, and of the same order as the number of lineages. However, the process is exactly described by discrete-time recursions similar to those for the continuous-time coalescent.

Hudson & Kaplan (1988) derived recursions for changes in the number of lineages within each background, averaging over random changes in background frequencies. They allowed for the possibility that Ns is small, in which case drift is important even when the selected alleles are common. Kaplan *et al.* (1989) used this framework to find the distribution of the total length of a genealogy, given multiple substitutions, and hence found the distribution of the number of segregating sites in a sample of DNA sequences. These results allow tests that distinguish the effects of substitutions from those of population bottlenecks, based on the joint distribution of the number of segregating sites and the nucleotide diversity (Tajima 1989; Braverman *et al.* 1995).

The coalescent approach is closely related to the analysis based on allele frequency, via the classical concept of identity by descent. To see this relationship, consider a favourable mutation that swept rapidly to fixation at some time of order $2N$ generations in the past. Tracing a sample of genes back in time, lineages coalesce in time according to the usual neutral process, at a rate $1/2N$ per generation, until the time when allele P was fixed. Next, trace back a further approximately $\log(2N)/s$ generations to a time when allele P was rare ($p^* \ll 1$), but still in large numbers ($Np^* \gg 1$). The chance of a coalescence during this brief interval is negligible if $Ns \gg 1$: the only significant process is the exchange of lineages between backgrounds P and Q by recombination. The probability that a lineage traces back into background P, when that background was at frequency p^*, is just $(p^*)^{r/s}$, as in Maynard Smith & Haigh's (1974) deterministic calculation. Going back further, there is now an appreciable chance of coalescence within background P, the rate being $1/(2Np)$ per generation. Eventually, all lineages must either coalesce, or escape into background Q to coalesce in the distant past ($t \sim 2N$).

The probability that two genes present just after the stochastic phase, at time t^*, coalesce within background P is equal to the probability f^*_{PP} that those genes are identical by descent relative to the population immediately before the mutation. The probability that two genes immediately after fixation coalesce as a result of the substitution is $f_{PP} = f^*_{PP}(p^*)^{2r/s}$, since the lineages trace back from fixation to t^* without interacting with each other. The probability of identity by descent is directly proportional to the variance in neutral allele frequency, through the relationship $\text{var}(\Delta u) = V = u_0v_0f_{PP}$. Thus, the distribution of pairwise coalescence times, and of pairwise identity by descent, can be reconstructed from the variance in allele frequency. Similarly, the genealogical relationship among sets of n genes can be reconstructed from the nth moments of the distribution of neutral allele frequencies.

The net effect of a rapid substitution is to cause the sudden coalescence of j lineages into a set of k families, containing $n = \{n_1, n_2, \ldots\}$ descendant lineages. For large Ns, a complete description of the effect on a genealogy is given by the distribution of n. Simulations show that the distribution of family sizes is qualitatively different from that generated by a population bottleneck: one large family which traces back to the ancestral mutation tends to dominate (Barton 1998). Moreover, the distribution of family sizes depends on both population size ($2N$) and the relative rates of recombination and selection (r/s): if the population is very large, and linkage is tight, there tend to be fewer, larger families for a given pairwise diversity than if linkage is loose and the population size smaller. However, very large samples of genes, and well-resolved genealogies, would be needed to detect these effects in practice. Two more promising methods for distinguishing bottlenecks from selective sweeps have been proposed. First, population bottlenecks (which must affect all loci) can be distinguished from the hitchhiking effect of successful mutations (which affect only closely linked sites) by testing whether reductions in diversity at different loci occur simultaneously (Galtier *et al.* 2000). Second, positive hitchhiking raises derived variants to high frequency. If the direction of evolution can be determined by comparison with an outgroup, this gives a sensitive assay for the effects of selective sweeps (Fay & Wu 2000).

(c) *Random drift caused by natural selection*

A selected substitution has a similar effect on linked neutral loci to random sampling drift (Medina & Petit 1979; Gillespie 2000). The variance in allele frequency, $\text{var}(u)$, increases in proportion to allele frequencies, u_0v_0, and the rate of coalescence of ancestral lineages increases. The distribution of allele frequencies, and the pattern of coalescence, differ from that expected under random sampling drift, but the qualitative effect is nevertheless a dispersion of allele frequencies and an increase in identity by descent.

Robertson (1961) first pointed out that selection causes random drift at other loci (see also Nei & Murata 1966). In the long term, variation in allele frequency must be due to variation in the fitness of genes: some genes must leave more descendants than others. When fitness variation is not correlated across generations, the variance in allele frequency increases at a rate

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\left(\frac{1}{2}+\frac{V_{\mathrm{f}}}{4}\right), \qquad (2)$$

where V_{f} is the variance in family size (Wright 1939). The first component is due to random segregation of heterozygotes, while the second is due to variation in the fitness of pairs of diploid individuals—if family size is Poisson distributed with mean and variance 2, we recover the rate $uv/2N$. Inherited variation in fitness causes larger fluctuations in allele frequency because fluctuations persist over generations. Additive fitness variation due to a locus r recombination units away cause perturbations which decay at a rate $(1-r)$ and so have a cumulative effect $1+(1-r)+\ldots = 1/r$. Hence, the influence of additive genetic variance in fitness, V_{a}, on the long-term drift of a neutral locus is inflated by a factor $1/r^2$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\left(\frac{1}{2}+\frac{V_{\mathrm{f}}}{4}+\frac{V_{\mathrm{a}}}{2r^2}\right), \qquad (3)$$

where V_{f} is the non-inherited component of family size. (Note that V_{a} is defined as the additive genetic variance associated with a diploid individual.)

Robertson's (1961) argument does not take account of the reduction in genetic variance caused by selection, which significantly reduces hitchhiking effects when selection is strong. This has been taken into account by Wray & Thompson (1990), Santiago & Caballero (1998) and Woolliams *et al.* (1999) for unlinked loci. Santiago & Caballero (1998) give a general formula that applies to any kind of selection and that allows for non-random mating. It is based on the infinitesimal model, which assumes that fitness depends on very many unlinked loci, each of small effect.

(d) *Reconciling Maynard Smith and Haigh with Robertson: fluctuating selection*

Theoretical results on hitchhiking effects fall into two apparently distinct classes: on the one hand, the effects of inherited variation in fitness, which scale with $1/2N$, and can be thought of as reducing the effective population size, N_{e}, by some factor; and on the other, the effect of a single mutation, which is not equivalent to a simple change in N_{e}, does not scale with $1/2N$, and so can be significant even in very large populations. The reasons for this distinction are that

(i) random drift occurs when the mutation is present in very few copies, so that total numbers, $2N$, have little influence; and

(ii) the effect of this drift is greatly amplified by the rise of the new allele to high frequency.

Santiago & Caballero's (1998) results depend on the infinitesimal model, in which changes in allele frequency are negligible. As we will see below, their approach only extends to explicit genetic models with a finite number of linked loci if linkage is loose ($r \gg s$) so that allele frequencies do not change significantly during the time for which linkage disequilibria persist. Note also that even when Santiago & Caballero's (1998) formula holds, the effects of selection are qualitatively different from those of simple drift, since fluctuations are correlated over time.

To understand the relationship between Robertson's (1961) argument, in which inherited variance in fitness inflates random drift, and the classical hitchhiking process analysed by Maynard Smith & Haigh (1974), it is helpful to analyse the effects of fluctuating selection. This bridges the two regimes: when allele frequencies vary little, or linkage is loose, the effect is an inflation of the rate of drift in proportion to the fitness variance, whereas when alleles sweep from low frequency to high, results are close to those of Maynard Smith & Haigh (1974). If balancing selection is widespread, and if polymorphisms fluctuate in frequency, then fluctuating selection could be the main cause of hitchhiking. This is because the hitchhiking effect of substitutions is bounded by the observed slow rate of molecular evolution, and the effect of inherited variance in fitness is bounded by the total rate of recombination and by the genetic component of fitness variance. Fluctuating selection could in principle have a stronger effect than either of these two extreme cases. (See Gillespie (1997) for a simulation study of various kinds of fluctuating selection, and a discussion of their likely importance as causes of hitchhiking.)

In Appendix A, I derive the hitchhiking effect of an arbitrary pattern of selection at a linked locus. When linkage is loose relative to selection, this reduces to equation (3). In the opposite regime, allele frequencies sweep back and forth rapidly relative to recombination ($r \ll s$):

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\left(1+\frac{\lambda}{2(2\lambda+r)}E\left[\frac{1}{pq}\right]\right) \text{ for } r \ll s, \qquad (4)$$

where λ is the rate of sweeps ($\lambda \ll s$). Note that equation (4) does not involve the selection coefficient, and depends primarily on the rate of sweeps relative to recombination (λ/r), and on the expected rate at which drift generates linkage disequilibria, through $E[1/p_{\tau'}q_{\tau'}]$. The key point is that the hitchhiking effect of a polymorphism that passes between extreme allele frequencies can be much greater than would be suggested by the additive variance in fitness.

Averaging over a genome of length R map units gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\left(1+\frac{\Lambda}{R}\log\left[1+\frac{R}{2\Lambda}\right]E\left[\frac{1}{pq}\right]\right), \qquad (5)$$

where Λ is the rate of sweeps over the whole genome. This formula is valid if $\Lambda \ll s$, r. However, it should be possible to extend it to sweeps at multiple loci, provided that the selected loci become polymorphic at different times ($\Lambda \ll s$).

(e) *Deleterious mutations*

When deleterious mutations are eliminated by selection, genetic diversity at linked loci is also lost. This kind of hitchhiking is the converse of that due to fixation of favourable mutations. Although any one deleterious mutation has little effect, the number of such mutations is much greater than the number of favourable mutations. Thus, the cumulative effect of such 'background selection' may be large (Charlesworth *et al*. 1993). The hitchhiking effect of deleterious mutations can be derived in the same way as for favourable mutations and fluctuating polymorphisms. Using this approach, the rate of random drift of a neutral allele is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\left(1+\frac{\mu s}{(s+r)^2}\right), \tag{6}$$

where μ is the rate of deleterious mutations at a linked locus; each mutation reduces the fitness of heterozygotes by s. Averaging over a genome of length R map units ($R \gg 1$), and assuming multiplicative effects across loci,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\exp\left(\frac{U}{R}\right), \tag{7}$$

where $U = 2\Sigma\mu$ is the total mutation rate over the diploid genome. This remarkable result was obtained by Hudson (Hudson & Kaplan 1995) using a coalescent-based argument and by Nordborg *et al.* (1996) using diffusion equations. As for extreme fluctuations in allele frequency (equations (4) and (5)), the rate of drift is independent of selection pressure.

(f) *Balancing selection*

If selection maintains polymorphism for very long periods, then the separate genetic backgrounds can diverge. This increases overall diversity at closely linked loci. Thus, balancing selection can be inferred from local increases in neutral diversity—variation in neutral mutation rates can be corrected for by comparison with an out-group (Hudson *et al*. 1987).

The reduction in the long-term rate of drift due to balancing selection seems at first to conflict with the increased rate of drift due to genetic structure discussed above. To understand the apparent discrepancy, consider complete linkage. Then, random drift occurs independently in the two pools, u_P, u_Q. At time t, u_P has variance $u_0v_0(1-\exp(-(t/2N)E[1/p]))$, and similarly for u_Q. Averaging over the final values of p:

$$\begin{aligned}\mathrm{var}[\Delta u] = u_0v_0\Bigg((E[p]^2+\mathrm{var}(p))\left(1-\exp\left(-\frac{t}{2N}E\left[\frac{1}{p}\right]\right)\right)\\ +(E[q]^2+\mathrm{var}(p))\left(1-\exp\left(-\frac{t}{2N}E\left[\frac{1}{q}\right]\right)\right)\Bigg).\end{aligned} \tag{8}$$

To leading order in $1/N$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2N}\Bigg((E[p]^2+\mathrm{var}(p))E\left[\frac{1}{p}\right]\\ +(E[q]^2+\mathrm{var}(p))E\left[\frac{1}{q}\right]\Bigg).\end{aligned} \tag{9}$$

Thus, if allele frequencies are held constant, there is no effect on the rate of drift to order $1/N$. However, over longer time-scales ($t \sim 2N$), the rate of drift is reduced, asymptotically, to $(uv/2N)\ (E[p]^2+E[q]^2+2\mathrm{var}(p))$.

Because this effect of balancing selection is only manifest over long times, it is appreciable only for very tight linkage ($r \sim 1/N$). It is this extremely local effect that allows the precise target of balancing selection to be located, in the first study of this kind to my knowledge, to a particular amino-acid polymorphism in the alcohol dehydrogenase gene of *Drosophila melanogaster* (Hudson *et al*. 1987). However, it is important to realize that slight fluctuations in a balanced polymorphism reduce diversity over a wide range, and increase it only within a very narrow range (Sved 1983). If a pair of balanced polymorphisms are held in strong linkage disequilibrium by epistatic selection, then diversity can be increased throughout the intervening region (Kelly & Wade 2000). The net effect of fluctuating and epistatic selection at a set of polymorphic loci is as yet unknown.

3. POPULATION STRUCTURE

The evolution of neutral alleles associated with diverse genetic backgrounds is analogous to their evolution in a population that is subdivided into discrete demes. Genes can move between genetic backgrounds by recombination and by mutation at the selected loci, just as they can move between demes by migration. Most of the theory of spatial subdivision deals with discrete demes which are maintained at constant size (Nagylaki 1986). The rate of drift of the whole population is then reduced by subdivision, because diversity is preserved within local isolates—just as diversity is increased at sites closely linked to polymorphisms held at constant frequency. However, if deme sizes fluctuate, genetic variation may be greatly reduced, just as with fluctuating selection. This effect is due to variation in reproductive success between individuals, which arises from the varying fortunes of the demes in which they live. Because fluctuations may be correlated across generations, they can greatly amplify random drift. Overall, population subdivision is likely to reduce genetic diversity (Whitlock & Barton 1997). Similarly, the net effect of balancing selection may also be to reduce diversity if allele frequencies change over time.

(a) *The island model*

For the island model, in which demes exchange genes with a common migrant pool, the long-term rate of drift can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{uv}{2n}(1+\mathrm{var}(v))E\left[\frac{(1-F_{\mathrm{ST}})}{2N_{\mathrm{d}}}\right], \tag{10}$$

where n is the number of demes, v is the eventual contribution of individuals in a deme to the whole population (scaled such that $\Sigma_i v_i = n$), N_{d} is the local deme size, and the expectation is over demes, weighted by v^2 (Whitlock & Barton 1997). Equation (10) is similar to equation (3), but allows for the reduction in rate of drift due to subdivision through the factor $(1-F_{\mathrm{ST}})$—this is the fraction of genetic variation held within demes. However, the average over $1/N_{\mathrm{d}}$, and the variance in long-term contribution both increase the rate of drift. While equation (10) is quite general, it is unhelpful, since both v and F_{ST} are complicated functions of the population dynamics and migration rates. Moreover, the long-term contribution v

cannot easily be related to observable quantities. This approach is similar to Santiago & Caballero's (1998)—there is a general relationship between the rate of drift and variance in reproductive success, but this depends on the long-term accumulation of fluctuations and hence on details of the models.

(b) *Substitution of new mutations*

The most interesting questions arise from the combined effects of genetic and spatial structure. First, consider a favourable mutation that sweeps through a spatially extended population. The net hitchhiking effect is expected to be much smaller than for a single population because it takes much longer for the allele to spread, giving more time for associations to dissipate. The ultimate result is a local increase in the frequency of linked neutral alleles and hence the generation of spatial differentiation. However, unless linkage is very tight, this increase will be restricted to the immediate neighbourhood of the birthplace of the new allele. Slatkin & Wiehe (1998) consider the effects of such spatial hitchhiking for models with discrete demes, which exchange few migrants ($Nm \ll 1$). Fixation in one deme is completed before spread begins in the next and so the process can be modelled by following the number of demes fixed for one or other allele. However, most natural populations show moderate geographical differentiation, implying large Nm. Spread therefore begins in many demes before it is completed in the first and is largely deterministic.

To understand this regime, in which random drift is confined to the early increase of a single mutation, consider spread through a two-dimensional continuum, in which genes diffuse at a rate σ^2. (This is the mean square distance between parent and offspring along some axis.) While it is rare, a favourable allele increases in a branching process independently of spatial structure. Allowing for the stochastic acceleration of an allele that is destined to be fixed, the expected number of copies at time t is $\exp(st)/2s$. These genes are distributed in a Gaussian distribution with variance $\sigma^2 t$ (ignoring random variation in this distribution). Since the initial frequency, integrated over the whole area, is $\int p_0(x)\mathrm{d}x = 1/2\rho$, where ρ is the population density, we have

$$p = \frac{\exp\left(st - \dfrac{r^2}{2\sigma^2 t}\right)}{2sNbt}, \tag{11}$$

where r is the distance from the origin and $Nb = 4\pi\rho\sigma^2$ is Wright's 'neighbourhood size'. Suppose that P is initially associated with the neutral allele U. During this time, the frequency of PU gametes increases in a similar way, but with a rate r of dilution by recombination with QV gametes. Since $u_Q \sim u$ initially, we have

$$u_P = u + v\,\mathrm{e}^{-rt}. \tag{12}$$

At a time $st \sim \log(2Nb)$, the selected allele becomes common. It rises rapidly to local fixation, and then spreads in a wave of advance with speed $c = \sqrt{2\sigma^2 s}$ (Fisher 1937). The time taken to spread throughout a species containing $2N$ genes, occupying a circular area of N/ρ, is $st \sim \sqrt{2Ns/Nb}$. If $(r/s)\sqrt{2Ns/Nb} \gg 1$, then linkage disequilibrium will have dissipated before fixation throughout the species, and the net effect will be a localized increase in neutral allele frequency.

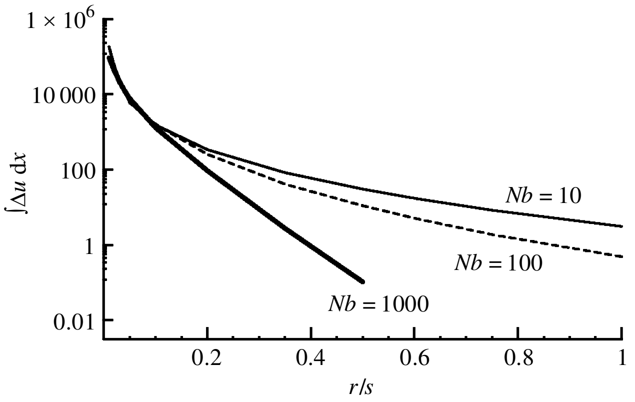

Figure 1. The net increase in neutral allele frequency, $\int \Delta u\,\mathrm{d}x$, caused by the spread of a new mutation, which was initially at low frequency. Distance is scaled relative to $\sqrt{\sigma^2/2s}$, so that the values on the vertical axis must be multiplied by $\sigma^2/2s$ to give the effective area occupied by additional copies of the neutral allele. Values are plotted against the ratio between recombination and selection (r/s) for $Nb = 4\pi\rho\sigma^2 = 10, 100, 1000$.

A quantitative solution can be found using the diffusion equations for the spread of two linked loci (Slatkin 1975). Figure 1 shows how this net increase in neutral allele frequency depends on (r/s), for neighbourhood sizes $Nb = 10$, 100, 1000. There are two regimes. For tight linkage (left of figure 1), the association between P and U remains almost complete when the wave of advance is established; the subsequent increase $\int \Delta u\,\mathrm{d}x$ is therefore independent of neighbourhood size. Examination of the solutions shows that as the wave advances, it raises neutral allele frequency by an amount proportional to the gradient, to $\Delta u = B(\mathrm{d}u/\mathrm{d}x)$. The quantity B has the dimensions of a distance, and has the form $B = b[r/s](\sigma^2/2s)$, with $b \sim 1.73(s/r)$ After the wave has passed, the solution follows $u \sim \exp(-r/B)$, and the net increase is $\int \Delta u\,\mathrm{d}x = 2\pi B^2$. The second regime applies when linkage is loose enough that recombination dissipates linkage disequilibria before the selected allele is locally fixed (right of figure 1). Then, the net increase in the neutral allele decreases with neighbourhood size, because the mutation starts at lower frequency in a denser population. The net increase decreases approximately exponentially with (r/s) (as shown by the straight lines in the linear-log plot of figure 1), as for hitchhiking in a single population.

The magnitude of the increase in neutral allele frequency can be large. For $r \ll s$, it tends to approximately $2\pi B^2$, which is equivalent to fixation within a radius $\sqrt{2}B \sim 1.73(s/r)\sqrt{2\sigma^2/s}$. For example, with $r = 0.01$, $s = 0.1$, the hitch is equivalent to fixation within a radius *ca.* 100σ. However, the overall effect on the whole species may nevertheless be small. A net increase of $\int \Delta u\,\mathrm{d}x$ raises allele frequency overall by $\int \Delta u\,\mathrm{d}x/A$. Overall, the rate of drift of the whole population is therefore $(\Lambda/A^2)E[(\int \Delta u\,\mathrm{d}x)^2]$, where Λ is the total rate of substitution, and the expectation is over the recombination rates and neighbourhood sizes. If Λ is independent of area, hitchhiking becomes negligible relative to simple drift (*ca.*

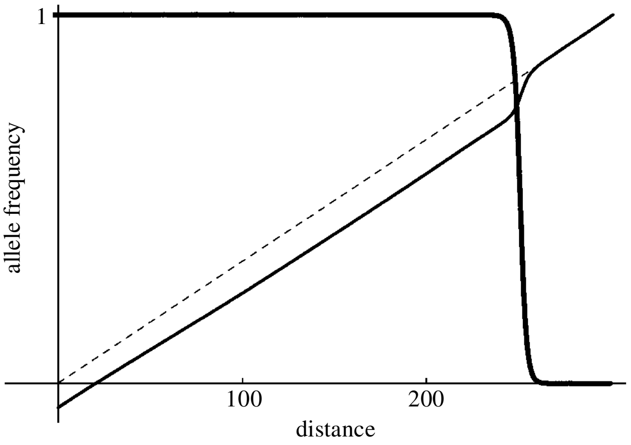

Figure 2. The effect of a wave of advance on standing variation. The heavy line shows the frequency of a favourable allele, moving from left to right. For $(r/s) = 0.05$, this causes a drop in the frequency of a neutral allele of 13.3 $(\mathrm{d}u/\mathrm{d}x)$. Distance (horizontal axis) is scaled relative to $\sqrt{\sigma^2/2s}$.

$1/(2\rho A))$ as the species' range (A) becomes large. It is possible that the rate of substitution scales relative to area. Then, hitchhiking is significant relative to drift only if $(\Lambda/A)E[(\int \Delta u\,\mathrm{d}x)^2] \sim 1/2\rho$.

Even after the initial association has dissipated, the advance of a new mutation has a further effect on the standing pattern of geographical variation. The wave front causes an increase in neutral allele frequency proportional to any pre-existing gradient, by $B(\mathrm{d}u/\mathrm{d}x)$ (figure 2). This is equivalent to shifting the whole pattern sideways by a distance B, without altering the magnitude of variation. Asymmetrical introgression across hybrid zones is often interpreted as being due to past movement, as modelled here (e.g. Marchant *et al.* 1988), though asymmetrical selection could also be responsible.

(c) *Local selective sweeps*

Seen from a different perspective, spatial subdivision may allow for greater hitchhiking effects. The observed slow rate of divergence between species limits the number of substitutions that can occur in the whole species. As we saw above, fluctuations in the frequency of balanced polymorphisms can also be a powerful source of drift if $E[1/pq]$ is small, even if species-wide substitutions are rare. However, it is hard to see that polymorphisms could approach extreme frequencies throughout a subdivided population. More likely, alleles sweep through individual demes as conditions change or as they replace alleles previously lost by drift. Using the approximation discussed in equation (2), the net rate of drift over the whole population is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{var}(\Delta u) = \frac{\Lambda}{n}uv(4Ns)^{-2r/s} \text{ for } r \ll s, \qquad (13)$$

where there are n demes, and Λ is the rate of substitutions per deme. Now, the rate of local substitutions might be very much greater than the global rate, but is still constrained by the cost of natural selection (Haldane 1957). Thus, local substitutions can only contribute substantially if genes interact in such a way as to allow a high rate of substitution with a tolerable mean fitness (Maynard Smith 1968; Sved 1968). Moreover, the net rate of drift scales inversely with the number of demes and so (unlike a single population) hitchhiking will not necessarily dominate over drift in a very large population.

Stephan (1994) and Begun & Aquadro (1993) have invoked local 'selective sweeps' to explain the presence of fixed differences between *Drosophila* populations in regions of reduced crossing over. However, Charlesworth (1998) has pointed out that such a pattern can also be produced by a process such as 'background selection', which reduces within-population diversity. While samples of sequences from multiple populations might allow detection of localized substitutions, theoretical interpretation is problematic. Moreover, if substitutions are frequent enough to be significant, observations are needed on a fine temporal and geographical scale.

(d) *Static barriers*

If selection maintains stable genetic differentiation, neutral genes must recombine onto a new genetic background in order to move to a new location. Thus, genetic structure generates an additional barrier to gene flow. This form of hitchhiking is distinct from those considered so far because linkage disequilibrium is generated by dispersal rather than drift (Nei & Li 1973)—genes moving into a new population carry with them many selected genes characteristic of their native population and tend to be eliminated by their association with locally deleterious genes. Conversely, if immigrants are locally favoured, for example by heterosis, then gene flow is increased (Ingvarsson & Whitlock 2000). The net effect of spatial barriers can be much stronger than classical hitchhiking—in the limit, selection against hybrids maintains separate biological species. For continuous clines, the barrier to gene flow is measured by the ratio between the change in neutral allele frequency, Δu, across the cline, and the gradient on either side $(B = \Delta u/(\mathrm{d}u/\mathrm{d}x)$, just as for a favourable allele, discussed in § 3(b).

To a good approximation, $B = w\overline{W}^{-1/r}$, where w is the cline width, $\overline{W}$ is the reduction in mean fitness at the centre and r the harmonic mean recombination rate between the selected loci and the neutral marker (Barton 1986; Kruuk *et al.* 1999). Such barriers can be detected through their effects on marker frequency and give an indirect method of measuring mean fitness (e.g. Szymura & Barton 1991). Gene flow may be substantially reduced but genetic barriers are nevertheless unlikely to impede the flow of a favourable allele much unless almost complete (Pialek & Barton 1997).

Physical barriers inflate neutral diversity by reducing the effective rate of gene flow (Nagylaki 1988), and genetic barriers act in the same way (Petry 1983). Nordborg (1997) has analysed the effects of local selection on genealogies and showed that if migration, selection and recombination act faster than drift, the effects on coalescence time can be summarized by a reduction in the effective rate of gene flow. Such effects might be detected through localized peaks in between-population divergence surrounding loci under spatially varying selection. However, such effects extend over a map distance of the same order as the selection coefficient and so would be hard to detect from studies of individual genes (Charlesworth *et al.* 1997).

4. DISCUSSION

Maynard Smith & Haigh (1974) introduced the term 'hitchhiking' to describe the effect of the substitution of a new mutation on a linked neutral locus. However, it is fruitful to see this as one instance of a general phenomenon—the perturbation of one locus by selection on other loci that are associated with it. When the selected loci are sufficiently loosely linked that their allele frequencies do not shift significantly during the time for which associations persist, the rate of drift is inflated in proportion to the additive variance in fitness (Robertson 1961; Santiago & Caballero 1998). These two apparently distinct kinds of hitchhiking are bridged by a simple model of fluctuating selection. This shows that when allele frequencies sweep between extreme values, hitchhiking effects can be much stronger than indicated by the variance in fitness, and depend primarily on the rate of sweeps relative to the rate of recombination. An analysis of hitchhiking in a spatially extended population shows that the effect of a substitution is weaker because it takes longer for an allele to fix throughout the population. However, tightly linked loci can increase over a large area, generating geographical differentiation as well as enhancing random drift of the species as a whole.

How significant are hitchhiking effects likely to be in nature? The theoretical results depend on relatively simple parameters, such as the net rate of substitutions, the genomic rate of deleterious mutations and the inherited variance in fitness. Unfortunately, we are still largely ignorant of the magnitudes of these key parameters. The rate of substitution would seem to be bounded by the observed rate of amino-acid divergence between species and by Haldane's (1957) 'cost of natural selection'. However, substitutions within local demes may be much more frequent than in the whole species, weak selection on very many non-coding sites could be significant in aggregate (Kondrashov 1995; McVean & Charlesworth 2000), and epistasis of the right form allows large numbers of substitutions for a given mean fitness (Maynard Smith 1968; Sved 1968). Our best current estimates are not so different from the original calculations of Maynard Smith & Haigh (1974), which suggested that hitchhiking due to favourable mutations is potentially the dominant process limiting neutral variation in large populations.

Indirect arguments based on relative rates of synonymous and non-synonymous divergence suggest a high genomic rate of deleterious mutation, U, in the human lineage (Eyre-Walker & Keightley 1999). Accumulation of mutations on balancer chromosomes in *Drosophila* suggests similarly high rates, but there are ambiguities in these experiments (Keightley 1998; Keightley & Eyre-Walker 1999) and deleterious mutation rates in some other model organisms seem too low to cause significant hitchhiking (Keightley & Caballero 1997; Lynch *et al.* 1999). The additive variance in fitness is notoriously hard to measure—there is indirect evidence that it is high, but little direct information (Burt 1995). Moreover, the theory presented here shows that the additive variance in fitness does not by itself determine the degree of hitchhiking—effects can be much stronger if selection causes wide variations in allele frequency.

The best evidence for the importance of hitchhiking comes from the lower nucleotide diversity seen in regions of reduced recombination of *Drosophila* (Stephan & Langley 1989; Aquadro *et al.* 1994), plants (Stephan & Langley 1998), mice (Nachman 1997) and humans (Nachman *et al.* 1998), on Y chromosomes (Charlesworth 1996; Filatov *et al.* 2000) and in selfing plants (Liu *et al.* 1998). Stronger effects yet are seen in bacterial populations, which reproduce predominantly asexually (Maynard Smith 1990). However, this does not directly tell us whether hitchhiking has a significant effect on variation in regions of high recombination or what kind of selection is responsible. Resolution of this issue may come from a better theoretical understanding of how hitchhiking shapes genetic diversity. That will require a much better understanding of the statistical properties of samples of sequences under different kinds of selection and of the effects of population subdivision.

I am grateful to B. Charlesworth and M. Slatkin for their helpful comments. This work was supported by the Biotechnology and Biological Sciences Research Council, the Natural Environment Research Council, and the Darwin Trust of Edinburgh.

APPENDIX A. THE HITCHHIKING EFFECT OF FLUCTUATING SELECTION

Consider a polymorphism with two alleles, at frequencies p, q. Assume Nsp, $Nsq \gg 1$, so that this frequency changes deterministically. The difference in neutral allele frequency between genetic backgrounds, $(u_P - u_Q)$ decays exponentially at a rate r, and is perturbed by independent sampling within backgrounds p, q:

$$(u_{P,t+1} - u_{Q,t+1}) = (1-r)(u_{P,t} - u_{Q,t}) + (\delta u_{P,t} - \delta u_{Q,t}), \quad \text{(A1)}$$

where

$$\mathrm{var}(\delta u_P)\frac{u_P v_P}{2Np}\mathrm{var}(\delta u_Q) = \frac{u_Q v_Q}{2Nq}\mathrm{cov}(\delta u_P, \delta u_Q) = 0. \quad \text{(A2)}$$

To leading order in $1/N$, $u_P \sim u_Q \sim u$, and $\mathrm{var}(\delta u_{P,t} - \delta u_{Q,t}) = uv/(2Npq)$ where $u = pu_P + qu_Q$. The value of $(u_{P,t} - u_{Q,t})$ is the cumulative effect of successive fluctuations at previous times τ. Approximating to continuous time

$$(u_{P,t} - u_{Q,t}) = \int_0^t (\delta u_{P,\tau} - \delta u_{Q,\tau})\mathrm{e}^{-r(t-\tau)}\mathrm{d}\tau.$$

Now, the overall allele frequency changes as a result of random sampling and of changes in p due to selection

$$\frac{\mathrm{d}u_t}{\mathrm{d}t} = \delta u_t + (u_{P,t} - u_{Q,t})\frac{\mathrm{d}p_t}{\mathrm{d}t}, \ \mathrm{var}(\delta u) = \frac{uv}{2N}. \quad \text{(A3)}$$

To leading order in $1/N$, these two components are uncorrelated. The net change in neutral allele frequency is

$$\Delta u = \int_0^t \delta u_\tau e^{-r(t-\tau)} d\tau + \int_0^t \frac{dp_\tau}{d\tau} \int_0^\tau (\delta u_{P,\tau'} - \delta u_{Q,\tau'}) e^{-r(\tau-\tau')} d\tau' d\tau. \tag{A4}$$

The variance is obtained by taking the expectation of the square of equation (A4). Note that the $\delta u_{P,\tau'}$ are uncorrelated between generations, and the covariance between the two terms is negligible. Thus

$$\mathrm{var}(\Delta u) = \int_0^t E[\delta u_\tau^2] d\tau.$$

$$+ \int_0^t \int_0^t \int_0^{\min(\tau_1,\tau_2)} \frac{dp_{\tau_1}}{d\tau_1} \frac{dp_{\tau_2}}{d\tau_2} e^{-r(\tau_1+\tau_2-2\tau')}$$

$$\times E[(\delta u_{P,\tau'} - \delta u_{Q,\tau'})^2] d\tau' d\tau_1 d\tau_2, \tag{A5}$$

$$= \frac{uv}{2N} \left(t + \int_0^t \int_0^t \int_0^{\min(\tau_1,\tau_2)} \frac{dp_{\tau_1}}{d\tau_1} \frac{dp_{\tau_2}}{d\tau_2} \right.$$

$$\left. \times \frac{e^{-r(\tau_1+\tau_2-2\tau')}}{p\tau' q\tau'} d\tau' d\tau_1 d\tau_2 \right).$$

The second integral gives the cumulative effects of random linkage disequilibria generated at time τ', acting through subsequent selection at times τ_1, τ_2. The rate of increase of variance in Δu can be found as the long-term effect of associations generated at τ':

$$\frac{d}{dt}\mathrm{var}(\Delta u) = \frac{uv}{2N} \left(1 + E\left[\frac{1}{p_{\tau'} q_{\tau'}} \left(\int_0^\infty \frac{dp_\tau}{d\tau} e^{-r(\tau-\tau')} d\tau \right)^2 \right] \right), \tag{A6}$$

where the expectation is over the time-course of p_τ.

Equation (A6) is closely related to Maynard Smith & Haigh's (1974) calculation, in which fluctuations are amplified by the integral

$$\int_0^\infty (dp\tau/d\tau) e^{-r(\tau-\tau')} d\tau.$$

However, if the substitution begins from a single mutation, then the derivation given above fails, because $p \sim O(1/2N)$. The correct result can be obtained by allowing for the reduction in $E[u_P v_P]$ due to drift (see Barton 1998, equation 5).

If recombination is faster than selection ($r \gg s$), then equation (A6) reduces to that given by Robertson's (1961) argument (equation (3) above). The opposite regime is where allele frequencies sweep back and forth rapidly relative to recombination ($r \ll s$)—for simplicity, assume that sweeps are between states of near fixation, so that

$$\int_0^\infty (dp\tau/d\tau) e^{-r(\tau-\tau')} d\tau$$

is approximately $\pm e^{-r(\tau-\tau')}$, the sign depending on the direction of the sweep. Assume that sweeps occur at exponentially distributed intervals, at a rate λ. Further, assume that drift is mainly generated between sweeps, during periods of near fixation, so that a factor $E[1/p_\tau q_{\tau'}]$ can be separated out. Then, the rate of drift depends on the expectation $E[(e^{-rt_1} - e^{-r(t_1+t_2)} + \ \ldots)^2]$, where the t_i are independent exponentially distributed variables. This is just $\lambda/(2(2\lambda + r))$ and leads to equation (4) above.

REFERENCES

Aguade, M., Miyashita, N. & Langley, C. H. 1989 Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. *Genetics* **122**, 607–615.

Aquadro, C. F., Begun, D. J. & Kindahl, E. C. 1994 Selection, recombination and DNA polymorphism in *Drosophila*. In *Non-neutral evolution* (ed. B. Golding), pp. 46–56. London: Chapman & Hall.

Barton, N. H. 1986 The effects of linkage and density-dependent regulation on gene flow. *Heredity* **57**, 415–426.

Barton, N. H. 1998 The effect of hitch-hiking on neutral genealogies. *Genet. Res.* (*Camb.*) **72**, 123–133.

Begun, D. J. & Aquadro, C. F. 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* **353**, 548–549.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Burt, A. 1995 The evolution of fitness. *Evolution* **49**, 1–8.

Charlesworth, B. 1996 The evolution of chromosomal sex determination and dosage compensation. *Curr. Biol.* **6**, 149–162.

Charlesworth, B. 1998 Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15**, 538–543.

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

Charlesworth, B., Nordborg, M. & Charlesworth, D. 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.* (*Camb.*) **70**, 155–174.

Clegg, M. T., Kidwell, J. F., Kidwell, M. G. & Daniel, N. J. 1976 Dynamics of correlated genetic systems. I. Selection in the region of the glued locus of *Drosophila melanogaster*. *Genetics* **83**, 793–810.

Dykhuizen, D. E. 1990 Experimental studies of natural selection in bacteria. *A. Rev. Ecol. Syst.* **21**, 373–398.

Eyre-Walker, A. & Keightley, P. D. 1999 High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347.

Fay, J. C. & Wu, C. I. 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Filatov, D. A., Moneger, F., Negrutiu, I. & Charlesworth, D. 2000 Low variability in a Y-linked plant gene and its implications for Y-chromosome evolution. *Nature* **404**, 388–390.

Fisher, R. A. 1937 The wave of advance of advantageous genes. *Annls Eugenics* **7**, 355–369.

Galtier, N., Depaulis, F. & Barton, N. H. 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**, 981–987.

Gillespie, J. H. 1997 Junk ain't what junk does: neutral alleles in a selected context. *Gene* **205**, 291–299.

Gillespie, J. H. 2000 Genetic drift in an infinite population: the pseudo-hitchhiking model. *Genetics* **155**, 909–919.

Haldane, J. B. S. 1957 The cost of natural selection. *J. Genet.* **55**, 511–524.

Hey, J. 1991 A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theor. Popul. Biol.* **39**, 30–48.

Hill, W. G. & Robertson, A. 1966 The effect of linkage on limits to artificial selection. *Genet. Res.* (*Camb.*) **8**, 269–294.

Hudson, R. 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Hudson, R. B. & Kaplan, N. L. 1988 The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R. R. & Kaplan, N. L. 1995 Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.

Hudson, R. R., Kreitman, M. & Aguade, M. 1987 A test for neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Ingvarsson, P. K. & Whitlock, M. C. 2000 Heterosis increases the effective migration rate. *Proc. R. Soc. Lond.* B **267**, 1321–1326.

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

Kaplan, N., Hudson, R. R. & Iizuka, M. 1991 The coalescent process in models with selection, recombination and geographic subdivision. *Genet. Res.* **57**, 83–91.

Keightley, P. D. 1998 Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: a simulation study. *Genetics* **150**, 1283–1293.

Keightley, P. D. & Caballero, A. 1997 Genomic mutation rates for lifetime reproductive output and lifespan in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **94**, 3823–3827.

Keightley, P. D. & Eyre-Walker, A. 1999 Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**, 515–523.

Kelly, J. K. & Wade, M. J. 2000 Molecular evolution near a two-locus balanced polymorphism. *J. Theor. Biol.* **204**, 83–102.

Kingman, J. F. C. 1982 The coalescent. *Stochastic Proc. Appl.* **13**, 235–248.

Kondrashov, A. S. 1995 Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *J. Theor. Biol.* **175**, 583–594.

Kruuk, L. E. B., Baird, S. J. E., Gale, K. S. & Barton, N. H. 1999 A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics* **153**, 1959–1971.

Liu, F., Zhang, L. & Charlesworth, D. 1998 Genetic diversity in *Leavenworthia* populations with different inbreeding levels. *Proc. R. Soc. Lond.* B **265**, 293–301.

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassilieva, L. & Willis, J. 1999 Perspective: spontaneous deleterious mutation. *Evolution* **53**, 645–663.

McVean, G. A. T. & Charlesworth, B. 2000 The effects of Hill–Robertson interference between weakly selected sites on patterns of molecular evolution and variation. *Genetics.* (In the press.)

Marchant, A. D., Arnold, M. L. & Wilkinson, P. 1988 Gene flow across a chromosomal tension zone. I. Relicts of ancient hybridisation. *Heredity* **61**, 321–328.

Maynard Smith, J. 1968 'Haldane's dilemma' and the rate of evolution. *Nature* **219**, 1114–1116.

Maynard Smith, J. 1978 *The evolution of sex*. Cambridge University Press.

Maynard Smith, J. 1990 The evolution of prokaryotes—does sex matter? *A. Rev. Ecol. Syst.* **21**, 1–12.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35.

Medina, J. R. & Petit, C. 1979 The hitch-hiking effect as a dispersive process. *J. Theor. Biol.* **81**, 235–246.

Nachman, M. W. 1997 Patterns of DNA variability at X-linked loci in *Mus domesticus*. *Genetics* **147**, 1303–1316.

Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. 1998 DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141.

Nagylaki, T. 1986 Neutral models of geographic variation. In *Stochastic spatial processes* (ed. P. Tautu), pp. 216–237. Berlin: Springer.

Nagylaki, T. 1988 The influence of spatial inhomogeneities on neutral models of geographical variation. I. Formulation. *Theor. Popul. Biol.* **33**, 291–310.

Nei, M. & Li, W. H. 1973 Linkage disequilibrium in subdivided populations. *Genetics* **75**, 213–219.

Nei, M. & Murata, M. 1966 Effective population size when fertility is inherited. *Genet. Res.* **8**, 257–260.

Nordborg, M. 1997 Structured coalescent processes on different timescales. *Genetics* **146**, 1501–1514.

Nordborg, M., Charlesworth, B. & Charlesworth, D. 1996 The effect of recombination on background selection. *Genet. Res.* (*Camb.*) **67**, 159–174.

Notohara, M. 1990 The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.* **29**, 59–75.

Petry, D. 1983 The effect on neutral gene flow of selection at a linked locus. *Theor. Popul. Biol.* **23**, 300–313.

Pialek, J. & Barton, N. H. 1997 The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes. *Genetics* **145**, 493–504.

Robertson, A. 1961 Inbreeding in artificial selection programmes. *Genet. Res.* **2**, 189–194.

Santiago, E. & Caballero, A. 1998 Effective size and polymorphism of linked neutral loci in populations under directional selection. *Genetics* **149**, 2105–2117.

Slatkin, M. 1975 Gene flow and selection in a two-locus system. *Genetics* **81**, 209–222.

Slatkin, M. & Wiehe, T. 1998 Genetic hitch-hiking in a subdivided population. *Genet. Res.* (*Camb.*) **71**, 155–160.

Stephan, W. 1994 Effects of genetic recombination and population subdivision on nucleotide sequence variation in *Drosophila ananassae*. In *Non-neutral evolution* (ed. B. Golding), pp. 57–66. London: Chapman & Hall.

Stephan, W. & Langley, C. H. 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the vermilion and forked loci. *Genetics* **121**, 89–99.

Stephan, W. & Langley, C. H. 1998 DNA polymorphism in *Lycopersicon* and crossing-over per physical length. *Genetics* **150**, 1585–1593.

Stephan, W., Wiehe, T. H. & Lenz, M. 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theor. Popul. Biol.* **41**, 237–254.

Sved, J. A. 1968 Possible rates of gene substitution in evolution. *Am. Nat.* **102**, 283–293.

Sved, J. A. 1983 Does natural selection increase or decrease variability at linked loci? *Genetics* **105**, 239–240.

Szymura, J. M. & Barton, N. H. 1991 The genetic structure of the hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*: comparisons between transects and between loci. *Evolution* **45**, 237–261.

Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Whitlock, M. C. & Barton, N. H. 1997 The effective size of a subdivided population. *Genetics* **146**, 427–441.

Woolliams, J. A., Bijma, P. & Villanueva, B. 1999 Expected genetic contributions and their impact on gene flow and genetic gain. *Genetics* **153**, 1009–1020.

Wray, N. R. & Thompson, R. 1990 Prediction of rates of inbreeding in selected populations. *Genet. Res.* **55**, 41–54.

Wright, S. 1939 *Statistical genetics in relation to evolution.* Actualites scientifiques et industrielles 802. Exposes de biometrie et de la statistique biologique XIII. Paris: Hermann et Cie.