Genet. Res., Camb. (2000), **75**, *pp.* 179–198. *With* 13 *figures.* Printed in the United Kingdom

The effect of epistasis on the structure of hybrid zones

N. H. BARTON[1] AND M. SHPAK[2]

[1] *Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK*
[2] *Ecology and Evolutionary Biology Department, Yale University, PO Box 208104, New Haven, CT 06520-8104, USA*

(*Received 24 February 1999 and in revised form 24 June 1999*)

Summary

Within hybrid zones that are maintained by a balance between selection and dispersal, linkage disequilibrium is generated by the mixing of divergent populations. This linkage disequilibrium causes selection on each locus to act on all other loci, thereby steepening clines, and generating a barrier to gene flow. Diffusion models predict simple relations between the strength of linkage disequilibrium and the dispersal rate, σ, and between the barrier to gene flow, B, and the reduction in mean fitness, $\bar{W}$. The aim of this paper is to test the accuracy of these predictions by comparison with an exact deterministic model of unlinked loci ($r = 0{\cdot}5$). Disruptive selection acts on the proportion of alleles from the parental populations (p, q): $W = \exp[-S(4\,pq)^{\beta}]$, such that the least fit genotype has fitness e^{-S}. Where $\beta \ll 1$, fitness is reduced for a wide range of intermediate genotypes; where $\beta \gg 1$, fitness is only reduced for those genotypes close to $p = 0{\cdot}5$. Even with strong epistasis, linkage disequilibria are close to $\sigma^2 p_i' p_j' / r_{ij}$, where p_i', p_j' are the gradients in allele frequency at loci i, j. The barrier to gene flow, which is reflected in the steepening of neutral clines, is given by

$$B = \int_{-\infty}^{\infty} (\bar{W}^{1/\bar{r}} - 1)\,\mathrm{d}x,$$

where $\bar{r}$, the harmonic mean recombination rate between the neural and selected loci, is here 0·5. This is a close approximation for weak selection, but underestimates B for strong selection. The barrier is stronger for small β, because hybrid fitness is then reduced over a wider range of p. The widths of the selected clines are harder to predict: though simple approximations are accurate for $\beta = 1$, they become inaccurate for extreme β because, then, fitness changes sharply with p. Estimates of gene number, made from neutral clines on the assumption that selection acts against heterozygotes, are accurate for weak selection when $\beta = 1$; however, for strong selection, gene number is overestimated. For $\beta > 1$, gene number is systematically overestimated and, conversely, when $\beta < 1$, it is underestimated.

1. Introduction

Hybrid zones (clusters of clines at multiple loci or traits) separate partially reproductively isolated populations. They act as barriers to gene exchange between the diverging populations, and may themselves promote divergence either as sites of reinforcement of prezygotic isolation, or as sources of novel recombinant genotypes ('hybrid speciation'). On the other hand, hybrid zones act as semi-permeable barriers, which allow exchange of genetic material. This gene flow may impede divergence, and may aid adaptation in either population.

As well as giving insight into these processes, hybrid zones allow quantitative study of the genetic basis of reproductive isolation, and its effects on gene flow. However, existing theory is largely based on the

* Corresponding author. e-mail: n.barton@ed.ac.uk

simplest models: heterozygote disadvantage (Bazykin, 1969), multiplicative across loci, or selection favouring different alleles in different places (Haldane, 1948; Fisher, 1950). It also rests mainly on the diffusion approximation (Fisher, 1937; Haldane, 1948; Nagylaki, 1975), which assumes that selection is weak. In fact, reproductive isolation involves strong gene interactions. Here, we examine whether conclusions based on simpler models are robust to various kinds of epistasis, and find how far they hold for strong selection.

In the absence of detailed information on the genetic basis of reproductive isolation, and its effects on fitness, it is hard to know what models of epistasis are appropriate. One way of classifying an epistatic selection scheme is by the number of genes which interact with each other. At one extreme, disruptive selection might act on one additive trait, so that fitness depends on interactions among very many loci, and is lower for all individuals with a mixed genotype. At the other extreme, particular pairs of loci might interact, such that recombinant pairs have lower fitness. In this case, an individual of mixed ancestry might have maximum fitness, if all coadapted pairs remain together, or minimum fitness, if all such pairs are recombined: there is a high variance in fitness among genotypes with the same average ancestry.

Another way of classifying epistatic models is to ask whether the transition between the divergent populations necessarily requires a reduction in mean fitness. With disruptive selection on a single trait, divergence must involve a fitness loss of the same magnitude as the eventual reproductive isolation. With interactions among pairs of genes, divergence is also opposed by natural selection, though the fitness loss may be spread over a number of separate transitions, each only weakly selected.

In less symmetrical models, there may be a path of relatively fit genotypes connecting the parental genotypes, which facilitates divergence by any mechanism (see Bateson, 1909; Dobzhansky, 1937; Muller, 1942; Barton & Charlesworth, 1984; Gavrilets & Gravner, 1997; Orr, 1995, 1997). If incompatibilities tend to be due to recessive alleles, as under the dominance theory of Haldane's Rule (Orr, 1995), then fit intermediate genotypes also become more likely. If fit ancestral paths exist, then they might be recreated across hybrid zones, reducing the fitness loss and the barrier to gene exchange. Such selection of the fitter hybrid genotypes could scatter the clines that make up a hybrid zone, invalidating inferences made assuming concordance. In the extreme, fit hybrid genotypes might escape from the hybrid zone (perhaps with the aid of random drift) to found a recombinant species (Rieseberg, 1995; McCarthy *et al*., 1995; Arnold, 1996).

In this paper, we examine the robustness of inferences to different degrees of epistasis. We concentrate on disruptive selection on a single additive trait, which is just the proportion of genes derived from one of the parental taxa. This model is chosen for two reasons. First, this makes it possible to compare approximations with exact deterministic calculations, avoiding the (considerable) difficulties introduced by random drift in individual-based simulations. Secondly, this model seems most likely to introduce systematic distortions in cline shape: depending on the pattern of epistasis, the marginal selection on each allele may become much weaker or stronger towards the centre of the hybrid zone. We leave aside the important questions of whether speciation involves transition via a chain of fit genotypes, and how new species may descend from fit hybrids. We concentrate on 'endogenous selection', in which selection acts against hybrids regardless of location. A similar analysis can be carried through for 'exogenous selection', where different genotypes are favoured in different places; again, this is an important issue for future work (see Kruuk *et al*., in press).

Our analysis is based on a series of approximations, each of which is compared with exact results. The first, and most accurate, level of approximation is to represent migration between neighbouring demes by diffusion through a continuous habitat, and discrete generations by continuous time. Next, the variance of the trait is divided into a genic contribution, determined by allele frequencies, and a component due to pairwise linkage disequilibria (Bulmer, 1985); the linkage disequilibria are taken to be in a 'quasi-equilibrium' between dispersal and recombination. We then reduce the dynamics of the 2^n haploid genotype frequencies to those of the trait mean and variance. This requires that the n loci are equivalent, and that the distribution of the trait be determined by the mean and variance. Since the trait we consider is bounded by fixation for one or other set of parental alleles, and since linkage disequilibria may be strong, it is not adequate to assume a normal distribution; we take the more general approach of assuming a Beta distribution, or its discrete analogue. Finally, and most difficult, we consider ways of simplifying calculation of the selection response, using various kinds of selection gradient. Thus, as well as examining issues peculiar to hybrid zones, this paper can be seen as a more general examination of alternative approximations to multilocus systems.

We first consider the effect of epistasis on allele frequencies alone. Even in the absence of linkage disequilibria, epistasis is still important, because it influences the marginal effects of each allele, and therefore distorts multilocus clines. We then ask whether linkage disequilibria can be approximated by assuming a short-term 'quasi-linkage equilibrium' between dispersal and recombination, disregarding the associations among loci generated directly by

epistasis. Next, we find whether the barrier to gene flow at a neutral locus, which is determined from the shape of the cline at that locus, can be predicted from the pattern of mean fitness, in the presence of epistasis. Finally, we find how the width of the selected clines, and hence the number of genes inferred from observations on neutral markers, depends on the pattern of epistasis.

2. Methods

In making exact calculations, we assume no linkage between any of the n genes; this ensures that there is a symmetrical solution, in which all loci are equivalent. We assume that selection acts on the viability of diploid individuals, which depends solely on an additive trait, p. This is just the proportion of selected genes derived from one of the parental taxa. Individual fitness is chosen as $W = \exp[-S(4pq)^{\beta}]$ (as in Barton & Gale, 1993), so that the least fit genotype has fitness e^{-S}. Throughout, S is a measure of the total selection acting, while $s \sim S/n$ refers to the selection on each of n loci. The nature of epistasis depends on the parameter β. For $\beta \gg 1$, fitness drops only for genotypes close to intermediate, whereas for $\beta \ll 1$, it is depressed by a slight amount of introgression (Fig. 1).

Exact deterministic calculations can be done if one assumes that all loci are equivalent, and that each carries either '0' or '1' alleles. Then, one need only follow the frequency of haplotypes carrying $k = 0, 1, \ldots n$ '1' alleles. More precisely, one must assume that all genotypes are equally frequent, for a given $p = k/n$ (Kondrashov, 1984; Barton, 1992; Doebeli, 1996; Shpak & Kondrashov, 1998). This requires that there be no linkage, and that selection acts only on p. These conditions are not, however, sufficient: for example, with stabilizing selection in a single population, this symmetrical solution is unstable, and the population fixes for any combination that approaches the optimal phenotype (Wright, 1935). The symmetrical solution is stable for certain forms of disruptive selection, but it is not known how generally it is valid. In a hybrid zone, some kinds of epistasis can cause clines to become staggered (e.g. Hatfield *et al.*, 1992; see Barton & Shpak, 2000).

Suppose that selection acts on diploids. We must follow the number of '1' genes inherited from the mother and father, respectively, $\{i,j\}$. Given the assumption that genotypes are equiprobable, conditional on $\{i,j\}$, we can calculate the expected heterozygosity of an $\{i,j\}$ individual, $B_{h|i,j}$, and hence can allow selection to act both against heterozygotes, against recombinants (with intermediate $i+j$) and in relation to environment. The proportion of gametes carrying k '1' alleles is (Barton, 1992):

$$g_k^* = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{1=\max[0,\,i+j-n]}^{\min[i,j]} \binom{h}{k-1}\left(\frac{1}{2}\right)^h,$$

$$B_{h|i,j}\frac{W_{i,j,h}}{\bar{W}} g_i g_j, \tag{1}$$

where

$$h = i+j-2l,$$

$$B_{h|i,j} = \frac{\binom{i}{1}\binom{n-i}{j-1}}{\binom{n}{j}},$$

$$\bar{W} = \sum_{i=1}^{n}\sum_{j=1}^{n} B[h\,|\,i,j]\, W_{i,j,h}\, g_i g_j.$$

The sum here is over l, which is the number of loci homozygous for '1' alleles; this is related to the number of heterozygous loci by $h = (i+j-2l)$. Note that if the number of '1' alleles in the diploid zygote is even, then so is the number of heterozygous loci, and vice versa. Thus, h changes in steps of 2, as the

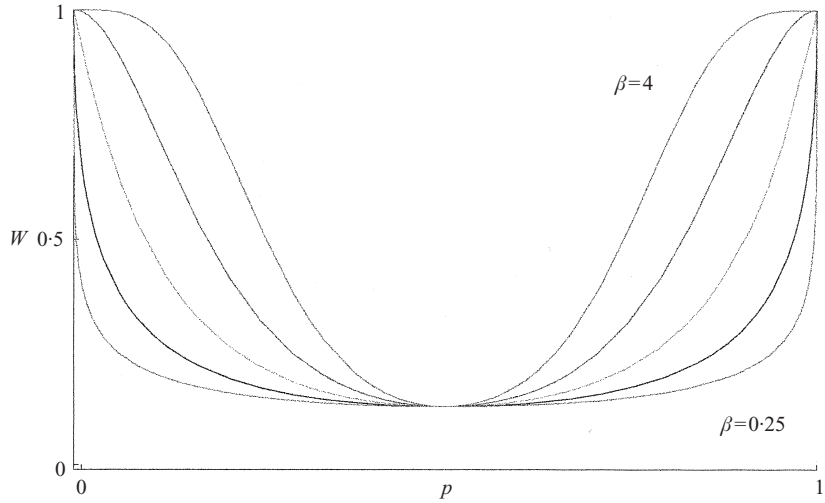

Fig. 1. Individual fitness, $W = \exp[-S(4pq)^{\beta}]$, plotted against the trait, p; $S = 2$, $\beta = 0{\cdot}25$, $0{\cdot}5$, 1, 2, 4 (bottom to top).

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

degree of overlap, l, changes in steps of 1. The simplest interpretation of (1) is that mating is random, so that the frequency of $\{i,j\}$ zygotes is $g_i g_j$, and that diploid viability, $W_{i,j,h}$, depends both on the number of '1' alleles inherited from either parent, $\{i,j\}$, and the number of heterozygous loci, h. However, $W_{i,j,h} g_i g_j$ might more generally represent the relative contribution of $\{i,j\}$ genotypes, including non-random mating among haploids (see Barton & Turelli, 1991).

Equation (1) specifies a symmetrical solution in which all haplotypes with a given number of '1' alleles are equally frequent. However, even when all evolutionary processes act symmetrically, this symmetrical solution may be unstable towards asymmetrical fluctuations. For example, stabilizing selection on a set of unlinked loci causes fixation of any genotype with phenotype at the optimal value; if this is not possible, at most one locus may remain polymorphic, so as to bring the mean close to the optimum (Wright, 1935). The symmetrical solution involves polymorphism at all loci, and so gives a much higher variance than any of the stable equilibria. The stability of the symmetrical solution depends on a $2n \times 2n$ matrix, whose elements are the derivatives of new genotype frequencies with respect to changes in genotype frequency. Barton & Shpak (2000) show how this matrix can be reduced to the product of matrices, each of dimension no more than $(n+1) \times (n+1)$; the method extends to spatially structured models. We have used this method to check that the symmetrical solutions derived below are stable, for $\beta = 0{\cdot}25$, 1, 4 and $S = 0{\cdot}1$, 1. In every case, the leading eigenvalue is close to 1, reflecting neutral shifts of the whole cline to left or right. (The slight deviation from 1 arises because the spatial range is finite.)

The exact model of (1) extends to include an unlinked neutral locus. We now follow u_i, the frequency of a neutral allele at the marker locus, within gametes carrying i '1' alleles at the n selected loci. Since the frequency amongst offspring from a union between gametes in classes i and j is just $(u_i + u_j)/2$, we have:

$$u_k^* g_k^* = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(u_i + u_j)}{2} \sum_{1=\max[0,\, i+j-n]}^{\min[i,j]} \binom{h}{k-1} \left(\frac{1}{2}\right)^h B_{h|i,j} \frac{W_{i,j,h}}{\bar{W}} g_i g_j. \qquad (2)$$

A similar approach was used by Barton (1995) to follow the probability of fixation of a rare selected allele, given selection on n other loci. Equation (2) could readily be extended to find the probability of identity by descent of a pair of neutral genotypes drawn from different genetic backgrounds, $F_{k,k'}$, or the frequency of neutral genotypes at two or more unlinked marker loci. Algorithms for simulating the exact model in a cline are available from http://helios.bto.ed.ac.uk/evolgen/, as a set of *Mathematica* 3.0 packages.

In our analytical approximations, we take space and time to be continuous. The effects of selection and gene flow are added, the latter being represented by a diffusion equation. The equation for the change in allele frequencies or, equivalently, the mean, $\bar{p}$, of the additive trait p, is:

$$\frac{\partial \bar{p}}{\partial t} = 0 = \frac{\sigma^2}{2} \frac{\partial^2 \bar{p}}{\partial x^2} + \Delta \bar{p}_s, \qquad (3)$$

where $\Delta \bar{p}_s$ is the change due to selection. The change in genotype frequency is given by a similar equation, and leads, by a change in variables, to diffusion equations for the linkage disequilibria (see Barton, 1983, 1986). This diffusion approximation can be justified for a variety of discrete or continuous population structures, in the limit where selection is weak (Nagylaki, 1975).

The most difficult task is to approximate the change in mean allele frequency due to selection, $\Delta \bar{p}_s$. This would be trivial if the population were described in terms of the full distribution of p. However, we wish to describe the population in terms of the mean and variance alone, and so must make some assumption about the form of the full distribution. We will show that pairwise linkage disequilibria can be approximated as being in a balance between dispersal and recombination. We then approximate the full distribution of p by choosing some standard distribution, with mean $\bar{p}$ and variance determined by $\bar{p}$ and by the pairwise linkage disequilibria. This distribution can be taken to be continuous, in which case we choose a Beta distribution: this is constrained to the range $0 \leqslant p \leqslant 1$, and converges to the normal when the variance is small:

$$\psi(p) = \frac{\Gamma[\alpha]}{\Gamma[\alpha\bar{p}]\,\Gamma[\alpha\bar{q}]} p^{\alpha\bar{p}-1} q^{\alpha\bar{q}-1} \quad \text{where the variance is } v = \frac{\bar{p}\bar{q}}{\alpha+1}. \qquad (4)$$

The mean fitness can then be calculated by integrating the product of (4) with the individual fitness, $W[p]$. It is more efficient to expand $W = \mathrm{e}^{-S(4pq)^\beta}$ as a Taylor series, and integrate each term. This gives:

$$\begin{aligned} \bar{W} &= \int_0^1 \mathrm{e}^{-S(4pq)^\beta} \psi(p)\, \mathrm{d}p \\ &= \int_0^1 \sum_{k=0}^{\infty} \frac{(-4^\beta S)^j}{j!} \frac{\Gamma[\alpha]}{\Gamma[\alpha\bar{p}]\,\Gamma[\alpha\bar{q}]} p^{\alpha\bar{p}+j\beta-1} q^{\alpha\bar{q}+j\beta-1}\, \mathrm{d}p \\ &= \sum_{j=0}^{\infty} \frac{(-4^\beta S)^j}{j!} \frac{\Gamma[\alpha]\,\Gamma[\alpha\bar{p}+j\beta]\,\Gamma[\alpha\bar{q}+j\beta]}{\Gamma[\alpha\bar{p}]\,\Gamma[\alpha\bar{q}]\,\Gamma[\alpha+2j\beta]}. \end{aligned} \qquad (5)$$

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

Table 1. *Comparison of various approximations to the selection response*

	$\beta = 1/4$			$\beta = 1$			$\beta = 4$		
Deme:	11	12	13	11	12	13	11	12	13
Exact $\Delta\bar{p}_s$	0·0519	0·0434	0·0104	0·0480	0·0371	0·0102	0·0428	0·0289	0·0071
Hypergeometric	0·0403	0·0403	0·0103	0·0459	0·0368	0·0103	0·0431	0·0294	0·0068
Beta	0·0243	0·0275	0·0083	0·0444	0·0356	0·0097	0·0427	0·0301	0·0077
Hypergeometric: L_1	0·0828	0·1000	0·0236	0·0760	0·0572	0·0137	0·0581	0·0282	0·0050
Hypergeometric: L_1, L_2	0·0527	0·0519	0·0118	0·0506	0·0383	0·0103	0·0411	0·0306	0·0074
Beta: L_1	0·0512	0·0930	0·0462	0·0755	0·0575	0·0139	0·0569	0·0275	0·0048
Beta: L_1, L_2	0·0323	0·0441	0·0175	0·0492	0·0379	0·0103	0·0391	0·0296	0·0071

The first row ('Exact $\Delta\bar{p}_s$') shows the change in $\bar{P}$ due to selection in demes 11–13, calculated exactly for a stepping-stone model with $m = 0{\cdot}5$, 20 demes, $S = 1$, and $\beta = 0{\cdot}25$, 1 and 4. The simulation was run for 50 generations, by which time equilibrium was approached, and the change due to selection was balanced by the change due to migration. The 'Hypergeometric' approximation assumes that the distribution of p follows a discrete hypergeometric with observed mean and variance. The 'Beta' approximation assumes a continuous Beta distribution, again with the observed mean and variance. The following rows show approximations which include the selection gradient just on the mean (L_1), or on the mean and the variance (L_1, L_2); these are calculated assuming either a discrete hypergeometric distribution, or a continuous Beta distribution.

For the parameter values used here ($S \leqslant 2$), only the first few terms in (5) need be included.

A more accurate approximation is to take p to follow a discrete distribution, ψ_i, with $i = 0, 1/2n, \ldots, 1$; a natural choice is a hypergeometric distribution, which is obtained by binomial sampling from an underlying Beta distribution:

$$\psi_i = \binom{2n}{i} \frac{(\alpha\bar{p})_i (\alpha\bar{q})_{2n-i}}{(\alpha)_{2n}} \tag{6}$$

where the variance is

$$v = \bar{p}\bar{q}\left(\frac{1}{\alpha+1} + \frac{1}{2n}\right)$$

and

$$(\alpha)_{2n} \equiv \alpha(\alpha+1)\ldots(\alpha+2n-1).$$

With this discrete distribution, and with moderate numbers of genes, it is most efficient to find the mean fitness by summing (6) directly. These choices are arbitrary: there is no guarantee that the form of the actual distribution of p approaches a Beta or hypergeometric distribution in the limit of (say) extreme allele frequency or weak selection. However, these forms are at least mathematically tractable.

By assuming that loci are equivalent, we have reduced the dynamics from those of 2^n haplotype frequencies to the distribution of a single trait, p. The dynamics can be further simplified to those of a single variable, $\bar{p}$, by assuming a form for the distribution of p, and approximating its variance by assuming that pairwise linkage disequilibria are generated primarily by dispersal. The remaining problem, therefore, is to find the change in mean, $\bar{p}$, due to selection. This can, of course, be calculated directly: the distribution of p after selection is equal to the distribution before selection, multiplied by the relative fitness, $W/\bar{W}$. Table 1 compares the exact change in mean due to selection, calculated using the actual equilibrium distribution of p, with that assuming a discrete hypergeometric, or a continuous Beta distribution. Both are accurate for $\beta = 1, 4$. However, for $\beta = 0{\cdot}25$, the continuous Beta approximation becomes inaccurate. This is because fitnesses change sharply near to fixation for $\beta < 1$ (Fig. 1).

It would be useful to find an approximation to this exact selection response which would give a more intuitive understanding, and which would be analytically tractable. For an additive trait with a Gaussian distribution of breeding values, the change in mean is just equal to the product of the genetic variance and the selection gradient on the mean, $\partial_{\bar{p}} \log[\bar{W}]$ (Lande, 1976). However, the use of selection gradients to describe selection on non-Gaussian distributions requires considerable care. In general, the response of a multilocus system to selection is proportional to a set of selection gradients, which are the partial derivatives of log mean fitness with respect to the variables that describe the system. (For a full discussion, see Turelli & Barton (1994).) For example, if we choose to describe the full system in terms of the allele frequencies, p_i, and linkage disequilibria, D_{ij}, $D_{ijk}, \ldots$, the selection response will be proportional to the gradients $\partial_{p_i} \log[\bar{W}]$, $\partial_{D_{ij}} \log[\bar{W}]$, $\partial_{D_{ijk}} \log[\bar{W}], \ldots$. If we consider just the distribution of diploid phenotypes, $p = 0, \frac{1}{2n}, \ldots, 1$, then the description can be reduced to the $2n$ moments of this distribution, $C_k = E[(p-\bar{p})^k]$. Then:

$$\Delta\bar{p} = \sum_{k=1}^{2n} C_{k+1} \frac{\partial \log[\bar{W}]}{\partial C_k}. \tag{7}$$

Note that moments of order higher than $2n$ can be expressed in terms of lower-order moments, and so

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

are redundant. The selection gradient $\partial_{C_k} \log[\bar{W}]$ is the partial derivative of log mean fitness with respect to the kth moment, keeping all other moments constant. For a continuous distribution, this is just proportional to the kth derivative of the fitness, evaluated at the mean:

$$\frac{\partial \log[\bar{W}]}{\partial C_k} = \frac{1}{k!\,\bar{W}} \frac{\partial^k W}{\partial \bar{p}^k}\bigg|_{\bar{p}}. \tag{8}$$

For a distribution which is close to Gaussian, it is more convenient to describe the distribution in terms of the cumulants, K_i, since the higher cumulants are small. The change in mean due to selection is then given exactly by:

$$\Delta\bar{p} = \sum_{k=1}^{2n} K_{k+1} \frac{\partial \log[\bar{W}]}{\partial K_k}. \tag{9}$$

The selection gradient $\partial_{K_k} \log[\bar{W}]$ is the partial derivative of log(mean fitness) with respect to the kth cumulant, keeping all other cumulants constant. For a continuous distribution, this is just proportional to the kth derivative of the mean fitness, with respect to translations of the distribution. The selection gradient on the cumulants can also be expressed as being proportional to the expectation of the kth derivative of the individual fitness:

$$\frac{\partial \log[\bar{W}]}{\partial K_k} = \frac{1}{k!\,\bar{W}} \frac{\partial^k \bar{W}}{\partial \bar{p}^k} = \frac{1}{k!\,\bar{W}} E\left[\frac{\partial^k W}{\partial p^k}\right]. \tag{10}$$

These results can be obtained by expanding W as a Taylor series; see Turelli & Barton (1994).

Considerable difficulties arise in applying selection gradients based on either moments or cumulants to discrete distributions, especially when the fitness functions are not smooth ($\beta < 1$; Fig. 1). For a smooth fitness function, (7)–(10) are almost exact: approximating the selection gradient by (7) or (9) involves approximating the fitness function by a $2n$th-order polynomial. However, this approximation breaks down when there are sharp changes in fitness between adjacent values of p; this is a problem when $\beta < 1$ (Fig. 1). Even with a smooth fitness function ($\beta \geqslant 1$), the sums in (7), (9) do not converge quickly: neglecting selection on the variance and higher moments is therefore inaccurate. Finally, with gradients based on cumulants, the method breaks down altogether, because even for plausible fitness functions, the expectation of the kth derivative of fitness can be infinite: see (10). Although the fitnesses realized by discrete genotypes $p = 0,\ 1/2n, \ldots, 1$ can take reasonable values, the continuous function from which they are calculated may have derivatives that approach infinity at $p = 0,\ 1$.

For distributions which are close to a continuous Beta, or a discrete hypergeometric, selection gradients are best defined with respect to those distributions. The selection gradient with respect to the mean is now defined as the partial derivative of log(mean fitness) with respect to changes in the mean, keeping the variance fixed, and keeping the distribution in the chosen form. The selection gradient with respect to the variance is defined in a similar way. (We do not consider selection gradients on higher moments here: these could be defined by considering deviations of the higher moments away from the specified form.) The change in mean due to selection on the mean and variance is then given by the first two terms of (7) or (9); these are equivalent since the second and third cumulants are the same as the corresponding central moments. This can be seen as an extension of the traditional assumption in quantitative genetics that the breeding value follows a Gaussian distribution; the Beta or hypergeometric forms are appropriate when, as here, the trait is confined to the range $\{0, 1\}$, and when allele frequencies may approach fixation. Table 1 compares the exact selection response with that calculated using the first two selection gradients (L_1, L_2), or just the selection gradient on the mean (L_1). Accounting for selection on the mean and variance is accurate, for both the discrete hypergeometric and continuous Beta approximations. However, accounting for selection on the mean alone is inaccurate. Thus, the response to selection depends on selection on both the mean and variance, and is mediated by the skew as well as by the additive genetic variance (7, 9).

3. Allele frequency clines

In this and the following sections we set out diffusion approximations for cline shape, following Barton (1983, 1986). If selection is weak relative to recombination, then linkage disequilibria can be neglected. Assume linkage equilibrium, and assume that all allele frequencies are equal to $\bar{p}$; the variance of p is then $v = \bar{p}\bar{q}/2n$. Multiply both sides of (3) by $\partial_x \bar{p}$ and integrate:

$$\frac{\sigma^2}{4}\left(\frac{\partial \bar{p}}{\partial x}\right)^2 = -\int_0^{\bar{p}} \Delta\bar{p}_s \, \mathrm{d}\bar{p}'. \tag{11}$$

The form on the right is valid only if fitness does not depend explicitly on position.

At linkage equilibrium, and with large numbers of genes, higher moments are negligible ($C_k \sim n^{1-k}$), and so it may be accurate to consider just the first selection gradient: $\Delta\bar{p}_s = v\partial_{\bar{p}} \log[W[\bar{p}]]$. A more drastic simplification, which applies when n is large enough that the variance around the mean is small, is that the mean fitness is equal to the fitness of an individual at the mean ($\partial_{\bar{p}} \log[\bar{W}] = \partial_{\bar{p}} \log[W[\bar{p}]]$); then, the selection gradient on the mean is $\partial_{\bar{p}} \log[\bar{W}] = \partial_{\bar{p}} W[\bar{p}]/\bar{W}$.

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

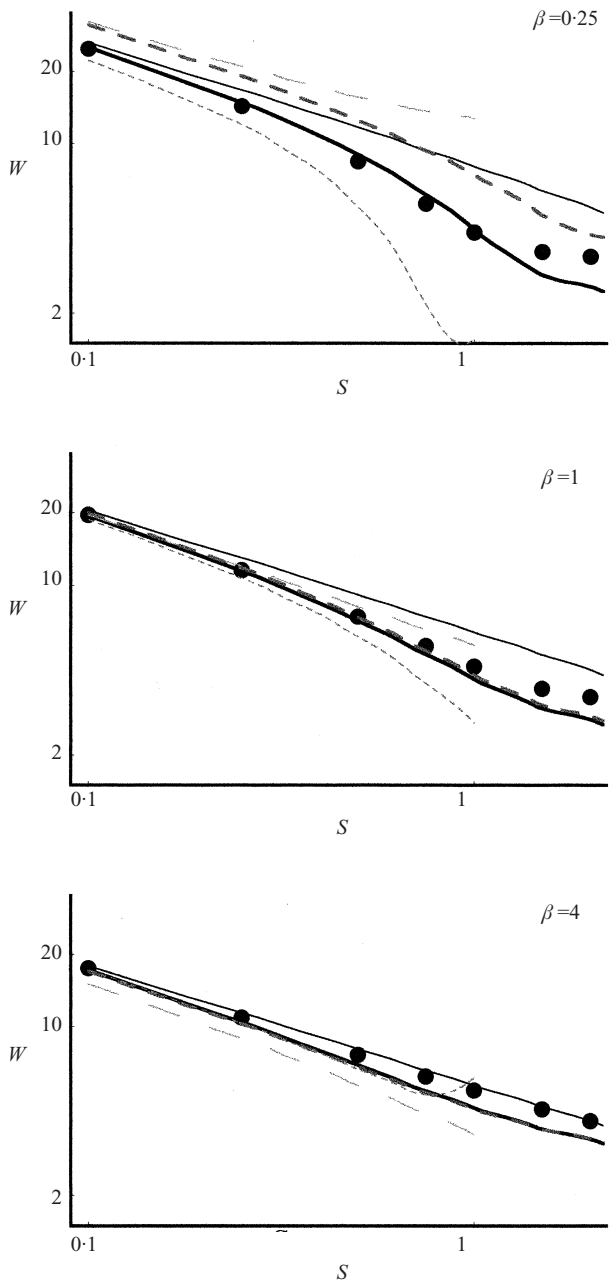

Fig. 2. Cline width, plotted against selection strength S for epistasis $\beta = 0{\cdot}25$, 1, 4. The dots show values from the exact model, with $n = 10$ genes, and $m = 0{\cdot}5$, starting from a sharp step. Width is scaled relative to σ, which here is equal to $\sqrt{m} = \sqrt{0{\cdot}5}$ deme spacings. This model was run for $40/S$ generations for $S < 0{\cdot}8$, and 50 generations for $S > 0{\cdot}8$; there were $\sim 20/\sqrt{S}$ demes for $S < 1$, and 20 demes for $S > 1$. Selection acts on diploids, and is followed by migration of diploid adults; allele frequencies are measured in the gamete pool. The thin continuous line shows the approximation in which the selection response is calculated for a population following a discrete hypergeometric distribution, with linkage equilibrium variance $\bar{p}\bar{q}/n$. The thick continuous curve is calculated in the same way, but accounts for the pairwise linkage disequilibrium generated by dispersal. The dashed lines show various approximations to the selection response; all allow for the increased variance in p due to pairwise linkage disequilibrium. The thick short dashed line approximates p to a continuous Beta distribution. The long dashed line further approximates this continuous distribution by assuming that the mean fitness equals the fitness of an individual at the mean, $\bar{p}$. The thin short dashed line assumes a discrete hypergeometric distribution, but allows only for the selection gradient on the mean, $\partial_p \log[\bar{W}]$.

Substituting $W[\bar{p}] = \exp[-S(4\bar{p}\bar{q})^{\beta}]$ into (11), and assuming that $\bar{W} \sim 1$, gives:

$$\frac{n\sigma^2}{2S}\left(\frac{\partial\bar{p}}{\partial x}\right)^2 = \frac{\beta}{4(\beta+1)}(4\bar{p}\bar{q})^{\beta+1} \qquad (12)$$

We define the width of the cline, w, as the inverse of the maximum slope. In this symmetrical model, the maximum slope is at $\bar{p} = 0{\cdot}5$; thus, the simple approximation of (12) gives

$$w = 2\sigma \quad \left(\frac{n}{2S}\frac{(\beta+1)}{\beta}\right).$$

For large β, where fitness is reduced only for intermediate individuals (upper curves in Fig. 1), the cline width approaches

$$2\sigma \quad \left(\frac{n}{2S}\right).$$

For small β, where fitness is reduced by even slight introgression (lower curves in Fig. 1), the cline becomes wide

$$\left(w \to 2\sigma \quad \frac{n}{2S\beta}\right).$$

This is because for small β there is little selection amongst the genotypes found at the centre, and so the clines there are shallow.

For $\beta = 1$, the simple approximation

$$w = 2\sigma \quad \left(\frac{n}{2S}\frac{(\beta+1)}{\beta}\right) = 2\sigma \quad \frac{n}{S}$$

is very close to that derived from (11), assuming a discrete hypergeometric distribution and linkage equilibrium. This is because the variance around the mean makes little difference to the selection gradients when the fitness varies smoothly with p. However, for $\beta = 0{\cdot}25$, (12) predicts a width 20% greater than (11), and for $\beta = 4$, a width 12% narrower. The error arises because the fitness function $W(p)$ varies sharply with p when β is far from 1 (Fig. 1), and so variation around the mean has a substantial effect on the selection response. In the following, therefore, we concentrate on approximations which take account of variation in p.

Fig. 2 shows how cline width decreases as selection becomes stronger, for three values of β. The thin

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

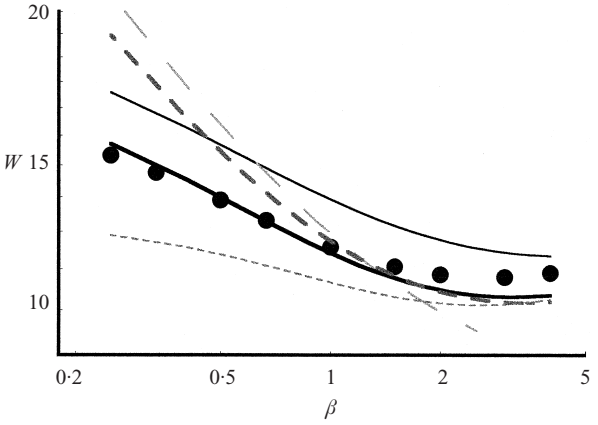

Fig. 3. The width of a cline maintained by epistasis $W = \exp[-S(4pq)^{\beta}]$, plotted against the parameter β. The curves correspond to those in Fig. 2.

continuous curve in Fig. 2 shows the prediction derived from (11), assuming a discrete hypergeometric distribution and linkage equilibrium. This scales almost precisely with $1/\sqrt{S}$, giving a straight line on this log–log plot. In the limit of weak selection, this agrees well with the exact solution for $n = 10$ genes, calculated by iterating (1) from an initial sharp step (dots in Fig. 2). As selection becomes stronger, the clines become narrower than predicted by (11), because linkage disequilibria inflate the variance of p, and hence increase the response to selection. Fig. 3 shows the dependence of cline width on epistasis, β, for selection $S = 0{\cdot}25$. Even for such weak selection, the clines are substantially narrower than expected at linkage equilibrium (compare the continuous line with the dots). We consider the effects of linkage disequilibrium in more detail below.

Since (11) gives the gradient of $\bar{p}$ as a function of $\bar{p}$, it can be integrated to give distance as a function of $\bar{p}$, and hence cline shape. Mean allele frequency is plotted on a logit scale (i.e. $\log(p/q)$) against distance (scaled to σ) in Fig. 4 for weak selection ($S = 0{\cdot}25$), and in Fig. 5 for strong selection ($S = 1$). (The logit scale is chosen because it expands the tails of the cline, and because simple models of single-locus clines give approximately straight lines on this scale.) For all values of β, there is close agreement with the linkage equilibrium prediction for $S = 0{\cdot}25$ (compare the thin continuous line with the dots in Fig. 4). However, for stronger selection ($S = 1$; Fig. 5) the clines are narrowed substantially by linkage disequilibria. For small β, where fitness is reduced by even slight introgression (lower curves in Fig. 1), alleles rapidly approach fixation at the edges. Indeed, under the simplest approximation, where the selection gradient is calculated from the fitness of a mean individual (12), fixation is reached at a *finite* distance. For $\beta = 1$, the cline is close to a straight line on a logit scale,

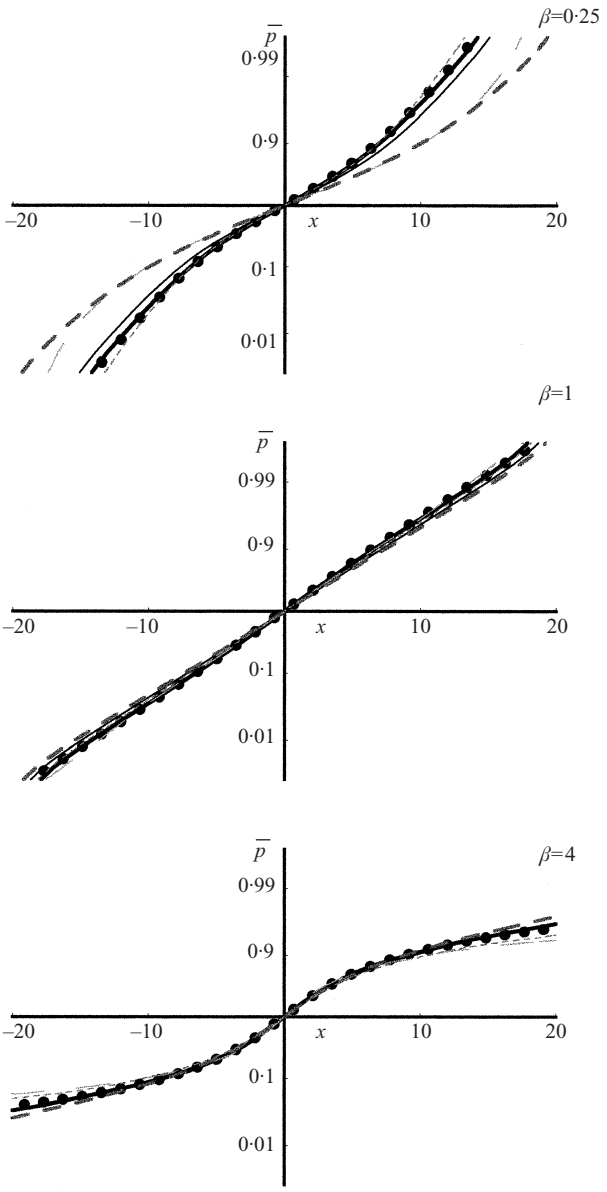

Fig. 4. Allele frequency is plotted against distance, for epistasis $\beta = 0{\cdot}25$, 1, 4; note the logit scale on the vertical axis. Selection is $S = 0{\cdot}25$, and distance is scaled relative to σ. The filled circles show the exact solution, starting with a sharp step and iterating for 800 generations. The curves show various approximations, and are labelled as in Figs 2, 3.

and so has a similar shape to other simple models of clines (Barton & Gale, 1993). For large β, where fitness is reduced only for intermediate individuals (upper curves in Fig. 1), the marginal selection on each allele is weak near the edges of the cline ($\bar{p} \ll 1$), and increases towards the centre. This induces a sharp step in the cline, qualitatively similar to what would be produced by a localized barrier to gene flow, or by linkage disequilibria. This step forms slowly, since introgression out to the tails of the cline is essentially neutral for $\beta > 1$: although the central region

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

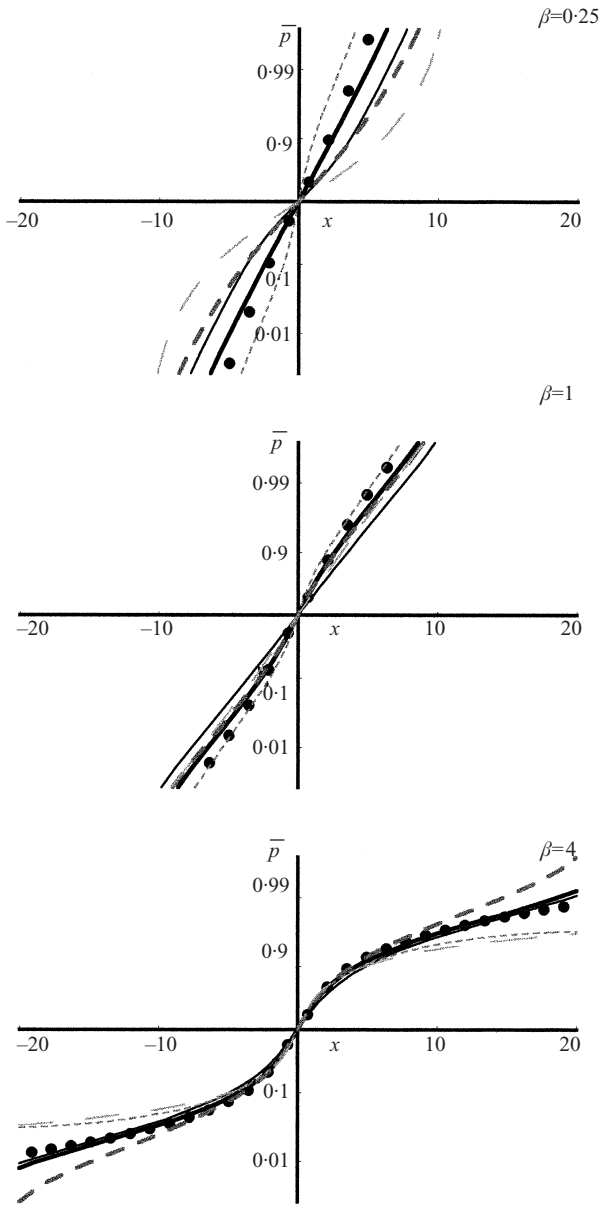

Fig. 5. As for Fig. 4, but with stronger selections ($S = 10$).

approaches equilibrium in approximately 50 generations for $S = 0{\cdot}25$, the tails only equilibrate after approximately 500 generations.

4. QLE approximations for linkage disequilibrium

Now, consider the effect of linkage disequilibria. For the moment, we represent the population in terms of allele frequencies and linkage disequilibria, rather than the distribution of the trait p. Throughout, we refer to the proportion of P alleles in an individual as p, and the mean of that additive trait in the population as $\bar{p}$. We refer to allele frequencies at a particular locus i as p_i, and to linkage disequilibria between loci i and j as D_{ij}. If selection and recombination are both weak, the diffusion approximation gives

$$\frac{\partial D_{ij}}{\partial t} = -r_{ij} D_{ij} + \frac{\sigma^2}{2}\frac{\partial^2 D_{ij}}{\partial x^2} + \sigma^2 \frac{\partial p_i}{\partial x}\frac{\partial p_j}{\partial x} + \sum_k \frac{D_{ijk}}{2}\frac{\partial \log[\bar{W}]}{\partial p_k} + \sum_{k,l} \frac{(D_{ijkl} - D_{ij} D_{kl})}{2}\frac{\partial \log[\bar{W}]}{\partial D_{kl}} + \ldots \quad (13)$$

(Barton, 1986 equation 8b). The first term on the right is due to recombination at a rate r_{ij} between loci i and j. The second term represents the diffusion of linkage disequilibria from place to place. The third term is due to dispersal, which mixes populations with different allele frequencies, and therefore generates associations between loci, D_{ij}, by the Wahlund effect (Li & Nei, 1974; Barton, 1979). The fourth term is due to selection on locus k, which has an indirect effect on the association between i and j through the three-way association D_{ijk}. The sum includes the special cases $k = i$ and $k = j$; with two alleles at each locus, $D_{iij} = -(p_i - q_i) D_{ij}$, and $D_{ijj} = -(p_j - q_j) D_{ij}$. The fifth term represents the effect of epistatic selection on associations between k and l; terms representing selection for higher-order associations have been omitted. Again, this includes special cases such as $k = i$ and $k = j$; for example, with two alleles at each locus, $D_{ijij} = p_i q_i p_j q_j + (p_i - q_i)(p_j - q_j) D_{ij}$. Higher-order terms arise from higher-order disequilibria.

If recombination is much faster than selection, then the population will rapidly approach a 'quasi-equilibrium', that is, a short-term balance between recombination and the forces generating linkage disequilibria (Barton & Turelli, 1991; Nagylaki, 1993). Since the latter are $O(s)$ (where s is a measure of the strength of selection on each locus), linkage disequilibria are $O(s/r) \ll 1$. To leading order in s/r:

$$D_{ij} = \frac{1}{r_{ij}}\left(\sigma^2 \frac{\partial p_i}{\partial x}\frac{\partial p_j}{\partial x} + p_i q_i p_j q_j \frac{\partial \log[\bar{W}]}{\partial D_{ij}}\right) + O\left(\left(\frac{s}{r}\right)^2\right). \quad (14)$$

Here, the selection terms have simplified dramatically, because the leading contribution to associations between loci i and j comes from the pairwise epistasis specifically for that association ($k, l = i, j$ or j, i; Barton, 1986; Hastings, 1986). If one or both of the loci i or j are neutral, or if there is purely directional selection on them (i.e. multiplicative selection), (14) simplifies to

$$D_{ij} = \frac{\sigma^2}{r_{ij}}\frac{\partial p_i}{\partial x}\frac{\partial p_j}{\partial x} + O\left(\left(\frac{s}{r}\right)^2\right). \quad (15)$$

Equations (14), (15) rest on the diffusion approximation, and therefore require that $r \ll 1$. For loose linkage, disequilibria change appreciably

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

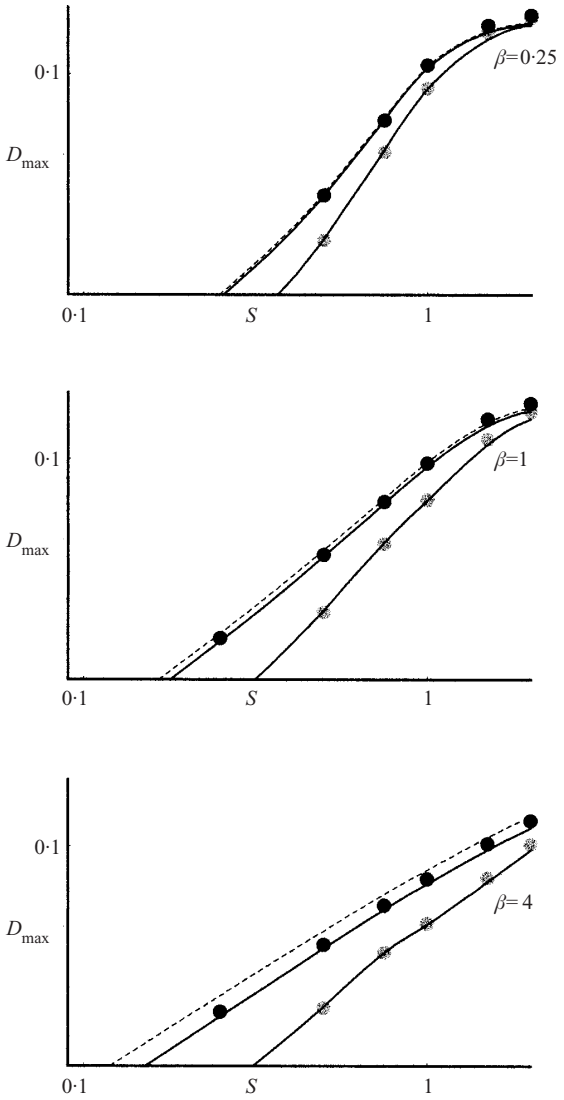

Fig. 6. Dark circles show the linkage disequilibrium, at the centre of the hybrid zone between two selected loci, plotted against selection S for epistasis $\beta = 0{\cdot}25$, 1, 4. Light circles show linkage disequilibrium between a neutral marker and a selected locus. Linkage disequilibria are measured in the gamete pool. These values are based on the exact model of 10 unlinked loci, run as for Fig. 2. Different neutral alleles were held fixed at either end of the range. The continuous curves show the QLE approximation, allowing only for the contribution from dispersal (15). The dashed curves show the full QLE approximation for the disequilibrium between selected loci, which included a small contribution from epistasis (14). These QLE predictions are calculated from the actual mean fitness under the exact model.

through the life cycle; consequently, the approximation to continuous time implied by the diffusion equation may be inaccurate. Changes through the life cycle can be understood as follows. Let D_{ij} be the association between loci i and j in gametes. Immediately after random union, the association between two genes derived from the same gamete is also D_{ij}. The association between genes derived from different gametes is zero; following Barton & Turelli (1991), we denote this cross-genome association by $D_{ij} = 0$. Suppose that migration of diploid individuals increases the association between two genes derived from the same parent by $\Delta_m D_{ij}$. Assuming equal genotype frequencies across the sexes, there is an equal increase in the association between two genes derived from different parents ($\Delta_m D_{i,j} = \Delta_m D_{ij}$). Recombination exchanges associations within genomes for associations across genomes; hence, the association in the next generation of gametes is

$$D^*_{ij} = r_{ij}\Delta_m D_{i,j} + (1-r_{ij})(D_{ij} + \Delta_m D_{ij})$$
$$= (1-r_{ij})D_{ij} + \Delta_m D_{ij}.$$

At equilibrium, $D_{ij} = \Delta_m D_{ij}/r_{ij}$ in the gamete pool. Immediately after dispersal, the association between genes from the same parent is $\Delta_m D_{ij}(1+1/r_{ij})$, and the cross-genome association is $\Delta_m D_{ij}$. We take $\Delta_m D_{ij}$ to be given by the diffusion approximation, in which case (15) is valid for the gametic associations. Kruuk *et al.* (in press) show that, in the absence of selection, the QLE linkage disequilibrium is indeed given by (15), if it is measured in gametes, or before dispersal of diploid adults.

The contribution of epistasis to pairwise linkage disequilibrium, and hence to the variance of the additive trait, p, can be found using the QLE approximation (14). Since $n(n-1)/2$ pairs of genes contribute to the mean pairwise linkage disequilibrium, $\bar{D}$, and since the variance in p is

$$v = \frac{\bar{p}\bar{q} + (n-1)\bar{D}}{2n},$$

the selection gradient on linkage disequilibrium is

$$\frac{\partial \log[\bar{W}]}{\partial D_{ij}} = \frac{2}{n(n-1)}\frac{\partial \log[\bar{W}]}{\partial \bar{D}} = \frac{1}{n^2}\frac{\partial \log[\bar{W}]}{\partial v}. \qquad (16)$$

Fig. 6 shows that (15), which neglects this direct contribution from epistasis, is an excellent approximation, even for loose linkage ($r_{ij} = 0{\cdot}5$) and strong selection, and even for associations between selected loci, where epistasis would be expected to contribute to the disequilibrium (second term in 14). The dashed curves in Fig. 6 were calculated by substituting (16) into (14); they show that epistasis makes a negligible contribution to linkage disequilibrium. Barton (1986) argued that with large numbers of loci, the gradients in allele frequency are maintained to some degree by selection on all the loci, and so increase faster with n than does the contribution from epistasis between any particular pair of loci. This argument assumed that the epistatic coefficient, $\partial_{D_{ij}} \log[W]$, in (14) is of the same order in n as the directional selection coefficient which generates the gradients in allele frequency. That is reasonable if interactions are between small numbers of genes. However, if all genes interact via one trait, as here, pairwise epistasis must decrease as $1/n^2$, whereas directional selection decreases as $1/n$ (16). This is a

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

stronger effect than that identified by Barton (1986). Fig. 6 shows that, at least with the kind of epistasis assumed here, the contribution of epistasis to disequilibrium is negligible even with as few as $n = 10$ genes.

5. The effect of linkage disequilibria on neutral allele frequencies

The positive linkage disequilibria generated by dispersal allow selection on all the loci to act on each locus, and therefore sharpen clines in allele frequency. This sharpening further strengthens linkage disequilibria, leading to a positive feedback which can generate a sharp step at the centre of a set of multilocus clines (Barton, 1983). This process can also be viewed at the level of the trait, p. Dispersal between populations with different means inflates the genetic variance, reflecting increased linkage disequilibrium. This increased variance speeds the response to selection, sharpening clines and thus further inflating the variance. We first examine the effect of linkage disequilibria on a neutral cline (following Barton, 1986), and then turn to the more complicated question of the effect on selected clines.

Consider a neutral locus, i. The diffusion equation for allele frequencies (3) now includes terms due to linkage disequilibrium with all the selected loci, and with all the selected sets of loci:

$$\frac{\partial p_i}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 p_i}{\partial x^2} + \sum_{i \neq j} \frac{D_{ij}}{2}\frac{\partial \log[\bar{W}]}{\partial p_j} + \sum_{i \neq j \neq k} \frac{D_{ijk}}{2}\frac{\partial \log[\bar{W}]}{\partial D_{jk}} \ldots . \quad (17)$$

In the limit of weak selection ($s \ll r \ll 1$), linkage disequilibria are generated primarily by dispersal. Dropping terms due to selection on linkage disequilibria (i.e. epistasis: $\partial_{D_{jk}} \log[\bar{W}], \ldots$), and substituting in the QLE approximation for pairwise associations (15):

$$\frac{\partial p_i}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 p_i}{\partial x^2} + \frac{\sigma^2}{2}\frac{\partial p_i}{\partial x}\sum_{i \neq j}\frac{1}{r_{ij}}\frac{\partial p_j}{\partial x}\frac{\partial \log[\bar{W}]}{\partial p_j}. \quad (18)$$

If we suppose that selected loci are randomly scattered over the genome, independent of the strength of selection on them or the steepness of the associated clines, we can replace the recombination rates r_{ij} by their harmonic mean, $\bar{r}$. Provided that individual fitness does not depend explicitly on location or on genotype frequencies, the spatial gradient in log mean fitness is equal to the sum $\Sigma_j \partial_x p_j \partial_{p_j} \log[\bar{W}]$. (Since linkage disequilibria are assumed to be weak ($\sim s/r$), terms such as $\partial_x D_{jk} \partial_{D_{jk}} \log[\bar{W}]$ are of higher order in (s/r).) Thus:

$$\frac{\partial p_i}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 p_i}{\partial x^2} + \frac{\sigma^2}{2\bar{r}}\frac{\partial p_i}{\partial x}\frac{\partial \log[\bar{W}]}{\partial x}. \quad (19)$$

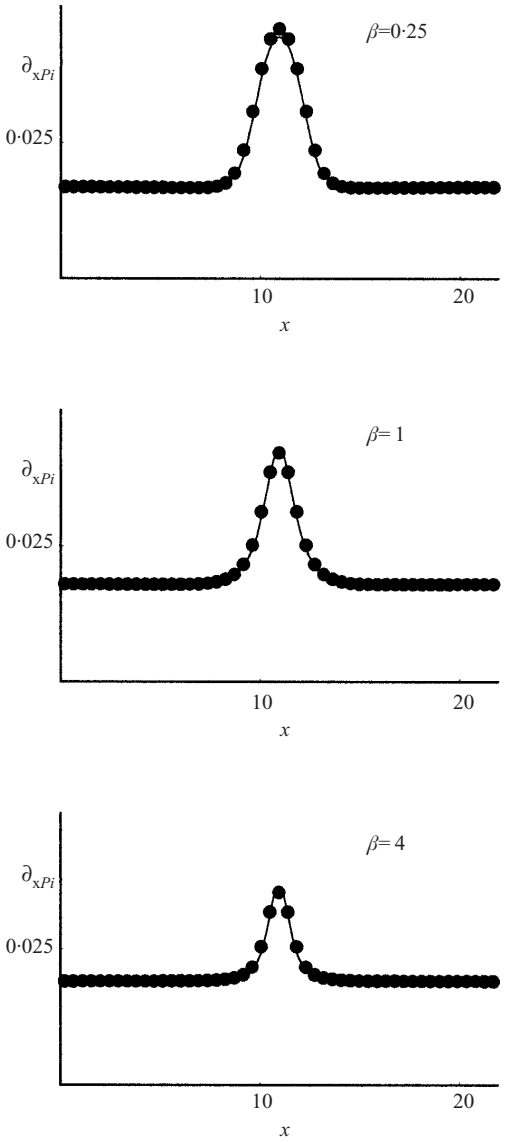

Fig. 7. Filled circles show the gradient in frequency at a neutral locus which is associated with clines at $n = 10$ selected loci. The continuous curve shows the gradient predicted from the mean fitness, $\partial_x p_i \sim \bar{W}^{-1/\bar{r}}$ (20), where $\bar{r} = 0{\cdot}5$ for unlinked loci. Both are plotted against distance, X, scaled relative to $\sigma\sqrt{n}/2S$. The three panels show epistasis $\beta = 0{\cdot}25$, 1, 4. Calculations are as for Fig. 2, with $S = 0{\cdot}5$, but were run for 10 times as long, and over 50% more demes, in order for the neutral clines to equilibrate properly. Gradients are calculated as simple differences between adjacent demes. The continuous curve is a cubic spline interpolation onto the values calculated from mean fitnesses in each deme.

At equilibrium, this integrates to give a simple relation between the mean fitness and the gradient in frequency of the neutral marker:

$$\frac{\partial p_i}{\partial x} = C\bar{W}^{-1/\bar{r}}, \quad (20)$$

where C is a constant determined by boundary conditions.

Fig. 7 shows that, for $S = 0{\cdot}5$, the gradient in

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

neutral allele frequency is close to that predicted from the mean fitness. However, with stronger selection the observed gradient becomes steeper than expected from the mean fitness. One possibility is that this discrepancy arises from the direct effects of epistasis on the neutral allele frequency. These are mediated by third- and higher-order associations such as D_{ijk}, and were neglected in deriving (20). This neglect can be justified on the grounds that epistasis has effects which are of second order in s/r. The QLE approximation for the three-way association, D_{ijk}, is similar to (15), and involves terms such as $\partial_x p_i \partial_x D_{jk}$; since D_{jk} is order s/r, the effect of pairwise epistasis, which is mediated by three-way associations, is of order $(s/r)^2$. Similarly, all higher-order disequilibria make a negligible contribution in the limit $s \ll r$, provided the locus i is neutral or, more generally, is not involved in epistatic interactions with any other loci (Barton, 1986). However, with strong selection, and large numbers of loci, the effect of linkage disequilibria on mean fitness can become substantial. For example, with disruptive selection (Fig. 1) the positive linkage disequilibria generated by dispersal increase the variance in p, and hence increase mean fitness. However, inspection of the higher-order terms shows that these reduce the gradient in neutral allele frequency below that expected from the mean fitness (20). Neglecting these terms therefore cannot account for the breakdown of (20) for $S \geqslant 1$.

Equation (20) implies that in the centre of the hybrid zone, where mean fitness is low, neutral clines will be steep. Overall, the effect is to induce a step in neutral allele frequency, proportional to the gradient at the edge. This is just the same effect as would be induced by a localized barrier to gene flow. The barrier strength is defined as (Nagylaki, 1976):

$$B \equiv \frac{\Delta p_i}{\partial_x p_{i|\pm\infty}}. \tag{21}$$

Here, Δp_i is the step in frequency of the neutral allele frequency due to selection on other loci, and $\partial_x p_{i|+}$ is the gradient in neutral allele frequency just outside the region in which mean fitness is reduced. If gene flow is asymmetrical, such that the gradients on either side of the hybrid zone differ ($\partial_x p_{i|-} \neq \partial_x p_{i|+}$), two different values of B must be defined, corresponding to gene flow in either direction.

The definition of (21) applies to the ideal case where the barrier to gene flow is sharply localized, so that the gradient just outside the barrier can be measured unambiguously. In reality, a physical barrier is spread over some distance in which density or dispersal are reduced, and a genetic barrier is spread over some distance in which mean fitness is reduced. Moreover, the neutral clines may be curved away from the barrier. In the simulations described below, linear clines were maintained artificially, by fixing their endpoints at $p_i = 0, 1$. This ensures a steady influx of neutral alleles, and allows accurate measurement of $\partial_x p_{i|\pm}$ from the difference in allele frequency between n adjacent demes. Accurate empirical measurements of barrier strength (e.g. Szymura & Barton, 1991) require first that, outside the hybrid zone, neutral alleles are changing more slowly than the timescale set by selection within the zone, and secondly that large samples are taken from a sufficiently long transect.

The barrier strength can be related to mean fitness in two ways, both using (20). First, the ratio of gradients at centre and edge depends on the relative fitness at the centre. Defining the cline width as $w = 1/(\partial_x p_{i|x=0})$:

$$B = w\Delta p_i \left(\frac{\bar{W}_0}{\bar{W}_{-\infty}}\right)^{-1/\bar{r}}, \tag{22}$$

where $\bar{W}_{-\infty} = \bar{W}_{\infty}$ is the mean fitness of the parental taxa. (We have assumed that the parental taxa have the same mean fitness; if they differ, the gradients, and hence the rates of introgression, will differ in proportion to $\bar{W}^{-1/\bar{r}}$. The barrier to gene flow will be greater for gene flow into the fitter population.) The step in allele frequency, defined as the *excess* over the allele frequency in the absence of the barrier, can be found by integrating (20) (Barton, 1986):

$$B = \int_{-\infty}^{\infty} \left(\left(\frac{\bar{W}_x}{\bar{W}_{-\infty}}\right)^{-1/\bar{r}} - 1 \right) \mathrm{d}x. \tag{23}$$

Fig. 8 compares the prediction of (23) with exact calculations of the barrier to flow of an unlinked neutral marker; the prediction is based on the actual pattern of mean fitness found from the exact calculations. The barrier is somewhat stronger for small β, because then, fitness is reduced over a wider range of allele frequencies (lower curves in Fig. 1). There is a good fit for weak selection, but with $S \geqslant 1$ the barrier is stronger than expected from the reduction in mean fitness (continuous curve in Fig. 8). Indeed, when selection becomes very strong, the predicted barrier strength *decreases* (right of Fig. 8). This is because linkage disequilibria then greatly inflate the variance at the centre, which generates extreme individuals and hence raises the mean fitness at the centre (upper dashed lines in Fig. 10).

The dashed lines in Fig. 8 show the barrier strength predicted for multiplicative selection against heterozygotes (29). Fitness is defined as $(1-s)^k$, where k is the number of heterozygous loci, and $s = S/n$. Epistasis with $\beta = 0{\cdot}25$ gives a stronger barrier, because mean fitness is reduced over a wider range of allele frequencies; conversely, epistasis with $\beta = 4$ gives a weaker barrier, because fitness is only reduced for a narrow range of genotypes. In comparing these models, S has been defined such that the least fit genotype has fitness $\sim \mathrm{e}^{-S}$. However, the mean fitness

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

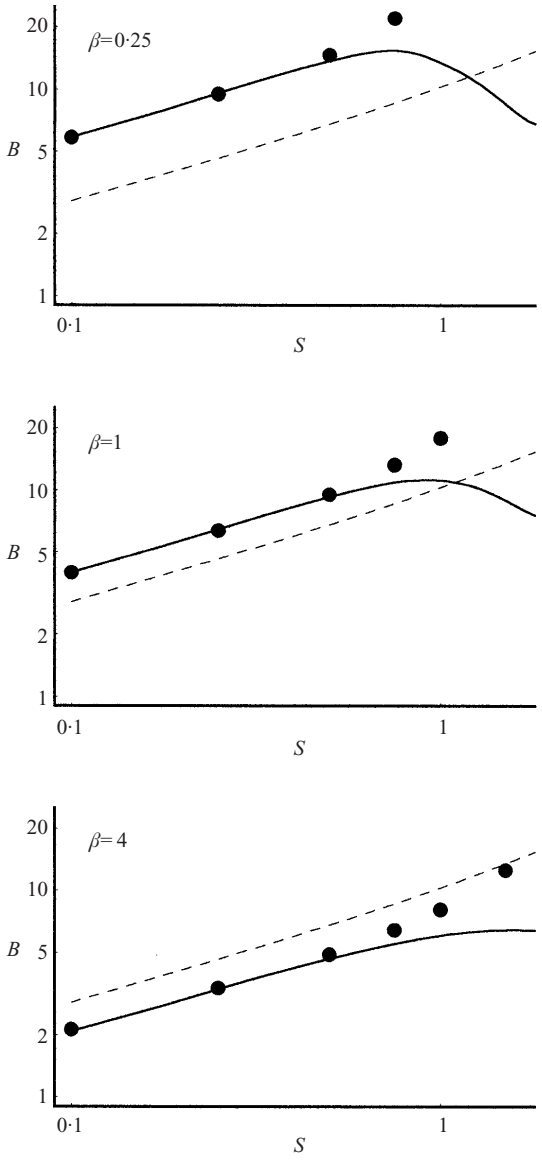

Fig. 8. Barrier strength is plotted against selection, S, for epistasis $\beta = 0{\cdot}25$, 1, 4. Barrier strength is scaled relative to σ. The filled circles show the exact solution, run as for Fig. 7. The continuous curve shows the diffusion approximation of (23). The dashed lines show the barrier strength predicted for selection against heterozygotes (29).

at the centre of the hybrid zone differs between the models. With selection against heterozygotes, the mean fitness is reduced to close to $e^{-S/2}$ at the centre, where half the loci are on average heterozygous. In the epistatic model, mean fitness would approach the minimum e^{-S} at linkage equilibrium, and with large numbers of loci; however, variation around the mean trait value raises the mean fitness substantially. In addition, the width of the region of reduced fitness differs between the models. These factors make it hard to predict how the barrier to gene flow will differ between different models of selection. However, in each case the barrier strength scales with $\sqrt{S}$, as is reflected in the straight lines on the left of Fig. 8.

6. The effect of linkage disequilibria on selected allele frequencies

We now consider how linkage disequilibria between selected loci sharpen each of the clines in allele frequency that makes up a hybrid zone. The response of the mean allele frequency, $\bar{p}$, to selection is given by a sum whose leading term is the product of the additive genetic variance with the selection gradient on the mean (7, 9). This applies even in the presence of linkage disequilibria. However, the variance of p is inflated by pairwise linkage disequilibria

$$\left(v = \frac{\bar{p}\bar{q} + (n-1)\,\bar{D}}{2n}\right),$$

and higher moments by higher-order disequilibria, which increases the response to selection, and hence steepens the clines. As before, we replace the linkage disequilibria by their QLE approximations. Matters are somewhat more complicated than for a neutral marker, because all higher-order terms now contribute to leading order in s/r: while dispersal makes no significant contribution to third- and higher-order disequilibria, epistasis does (see Turelli & Barton, 1994). In order to obtain a simple expression, we include only the disequilibrium generated by dispersal (15), on the grounds that epistasis makes a relatively small contribution to disequilibrium (Fig. 6). Even with this simplification, pairwise epistasis between loci i and j contributes via the association $D_{iij} = -(p_i - q_i)\,D_{ij}$. Thus, we cannot obtain a simple expression in the limit of weak selection ($s \ll r$); further approximations must be made. Terms involving coincident indices, such as D_{iij}, are small for large n, and so are dropped. To find the variance, v, we use (15), replacing the recombination rate r_{ij} by its harmonic mean, $\bar{r}$. Given the variance, and given some assumption about the form of the distribution which determines the higher moments, the selection response can be expressed as a function of $\bar{p}$ and $(\partial_x \bar{p})^2$. This allows (3) to be solved numerically, using any of the various methods for calculating the selection response discussed above (Table 1).

An analytical expression can be obtained if the selection response is approximated by the product of the variance with the selection gradient on the mean. Proceeding as for (18), the allele frequency at locus i is:

$$0 = \frac{\sigma^2}{2}\frac{\partial^2 p_i}{\partial x^2} + \frac{p_i q_i}{2}\frac{\partial \log[\bar{W}]}{\partial p_i} + \frac{1}{2}\sum_{j \neq i} D_{ij}\frac{\partial \log[\bar{W}]}{\partial p_j}. \qquad (24)$$

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

Substituting for D_{ij} from (15), and replacing the recombination rates r_{ij} by their harmonic mean, $\bar{r}$:

$$0 = \frac{\sigma^2}{2}\frac{\partial^2 p_i}{\partial x^2} + \frac{p_i q_i}{2}\frac{\partial \log[\bar{W}]}{\partial p_i} + \frac{\sigma^2}{2\bar{r}}\frac{\partial p_i}{\partial x}\sum_{j\neq i}\frac{\partial p_j}{\partial x}\frac{\partial \log[\bar{W}]}{\partial p_j}. \quad (25)$$

In considering the effect of linkage disequilibria on selected loci, we must separate the effects of selection gradients at other loci from the selection due to the locus itself. Let $\hat{W}$ be the mean fitness excluding the effects of locus i, such that $\partial_x \log[\hat{W}] \equiv \Sigma_{j\neq i}\partial_x p_j \partial_{p_j} \log[\bar{W}]$. If all loci have the same effect, then $\hat{W} = \bar{W}^{1-1/n}$, while if selection varies across loci, $\hat{W} \sim \bar{W}$ for $n \gg 1$. Regarding $(\partial p_i/\partial x)^2$ and $\hat{W}$ as functions of p_i, and integrating:

$$\frac{\sigma^2}{4}\left(\frac{\partial p_i}{\partial x}\right)^2 = -\int_0^{p_i}\frac{\tilde{p}_i\tilde{q}_i}{2}\left(\frac{\hat{W}[\tilde{p}_i]}{\hat{W}[p_i]}\right)^{2/\bar{r}} \frac{\partial \log[\bar{W}]}{\partial \tilde{p}_i}\mathrm{d}\tilde{p}_i. \quad (26)$$

As in (11), (19), this integration involves the assumption that fitness does not explicitly depend on location, and neglects the change in mean fitness due to change in variance and higher moments ($\partial_v \log[\bar{W}]$, etc.), Equation (26) is an extension of (11) which allows for the linkage disequilibria generated by dispersal. If there are such a large number of loci that the variance in p can be ignored, we can set $\bar{W} \sim \hat{W} \sim W = \exp[-S(4pq)^\beta]$. Then:

$$\frac{n\sigma^2}{2S}\left(\frac{\partial \bar{p}}{\partial x}\right)^2 = \frac{\beta}{4}\int_0^{4\bar{p}\bar{q}} h^\beta \exp[2\phi((4\bar{p}\bar{q})^\beta - h^\beta)]\,\mathrm{d}h = \frac{(2\phi)^{-(\beta+1)/\beta}\mathrm{e}^{2\phi(4\bar{p}\bar{q})^\beta}}{4\beta}\Gamma\left[\frac{1}{\beta}, 0, 2\phi(4\bar{p}\bar{q})^\beta\right] - \frac{\bar{p}\bar{q}}{2\phi}, \quad (27)$$

where $h = 4\tilde{p}\tilde{q}$,

$$\Gamma\left[\frac{1}{\beta}, 0, c\right] = \int_0^c t^{\frac{1}{\beta}-1}\mathrm{e}^{-t}\,\mathrm{d}t, \quad \phi = \frac{S}{\bar{r}}\left(1-\frac{1}{n}\right),$$

where $p_i = \bar{p}$ for all i. The cline width is found by setting $4\bar{p}\bar{q} = 1$ in (27).

Figs 2–4 compare various approximations for the cline width and cline shape. The most sophisticated is to calculate the selection response assuming a discrete hypergeometric distribution, with variance derived from (15). This prediction breaks down for strong selection (thick continuous curve at right of Fig. 2), and there are some edge effects in the predictions for cline shape (thick continuous curve in Figs 4, 5). The effect of linkage disequilibrium is also overestimated for large β (Fig. 3). Overall, however, there is good agreement with the exact results (thick continuous line in Figs 2–5). Allowing just for the selection gradient on the mean (dotted curve; 26) gives a good prediction for cline shape for weak selection (Fig. 4), but underestimates cline width for $\beta < 1$ (Fig. 3). Assuming a continuous Beta distribution (which has the advantage that predictions are independent of assumptions as to n) gives good predictions for cline width for $\beta = 1, 4$, but not for small β (short dashed line in Fig. 3); this prediction for cline shape is poor when selection is strong (Fig. 5). Making the further assumption that the variance around the mean is negligible (long dashed line; 27) gives fair accuracy only for $\beta \sim 1$ and weak selection (Fig. 2). The most striking finding is that for a smooth fitness function ($\beta = 1$) all these approximations perform quite well, up to moderately strong selection ($S \sim 1$). However, fitness functions which change sharply, and especially $\beta < 1$, require a more elaborate calculation of selection response.

The cline shape predicted by (27) can be compared with that for multiple clines maintained by heterozygote disadvantage (Barton, 1983). With multiplicative selection s against heterozygotes, at each of n loci, the fitness of an entirely heterozygous individual is $(1-s)^n \sim \mathrm{e}^{-S}$, where $S = ns$. Substituting $\bar{W} = (1-2spq)^n$, $\hat{W} = \bar{W}^{1-1/n}$ into (26) gives:

$$\frac{n\sigma^2}{2S}\left(\frac{\partial \bar{p}}{\partial x}\right)^2 = \frac{1}{4\phi(2\phi+s)}((1-2s\bar{p}\bar{q})^{-2\phi/s} - 4\phi\bar{p}\bar{q} - 1), \quad (28)$$

where

$$\phi = \frac{S}{\bar{r}}\left(1-\frac{1}{n}\right).$$

Evaluating (28) at $\bar{p}\bar{q} = 0{\cdot}25$ shows that the cline width is

$$w = \left(\frac{8\phi^2}{\mathrm{e}^\phi - \phi - 1}\right)\left(\frac{n\sigma^2}{2S}\right).$$

Equation (28) differs from the corresponding formula in Barton (1983),

$$\left(\frac{n\sigma^2}{2S}\right)(\partial_x \bar{p})^2 = (\mathrm{e}^{4\phi\bar{p}\bar{q}} - 4\phi\bar{p}\bar{q} - 1)/(8\phi^2),$$

which made the simplifying assumptions that $(1-2spq)^n \sim \mathrm{e}^{-2nspq}$ and that $\hat{W} \sim \bar{W}$. However, the formulae converge for $s \ll \phi$.

Fig. 9 compares the exact model (dots) with the prediction of (28) (continuous curves), for various numbers of loci. There is good agreement for $n = 20$ loci up to $S = 2$, and for $n = 10$ loci up to $S = 1$. However, for smaller numbers of loci, clines are substantially wider than is predicted by the diffusion

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

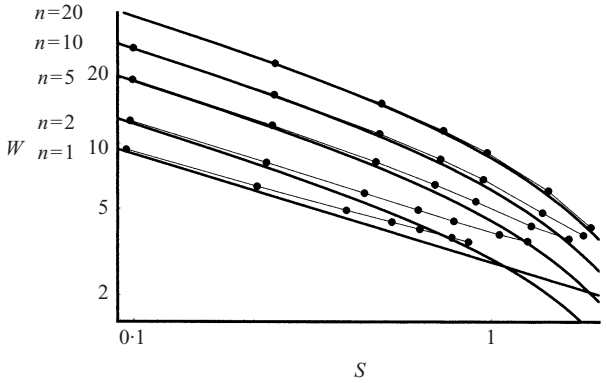

Fig. 9. Cline width, plotted against selection strength S for selection against heterozygotes; fitness is $(1-S)^n$, where k is the number of heterozygous loci, and $s = S/n$. The dots show values from the exact model, with $m = 0{\cdot}5$, starting from a linear gradient. Results are shown for $n =$ 1, 2, 5, 10, 20 loci. The dots are joined for clarity. Width is scaled relative to σ, which here is equal to $\sqrt{m} = \sqrt{0{\cdot}5}$ deme spacings. Parameters were as for Fig. 7. The continuous curves show the prediction of (28), which accounts for the pairwise linkage disequilibrium generated by dispersal; $\phi = (S/\bar{r})(1-1/n)$. For $n = 1$, this reduces to the prediction in the absence of linkage disequilibrium, $4\sigma\sqrt{n}/2S$ (Bazykin, 1969), which shows as a straight line on this log–log plot.

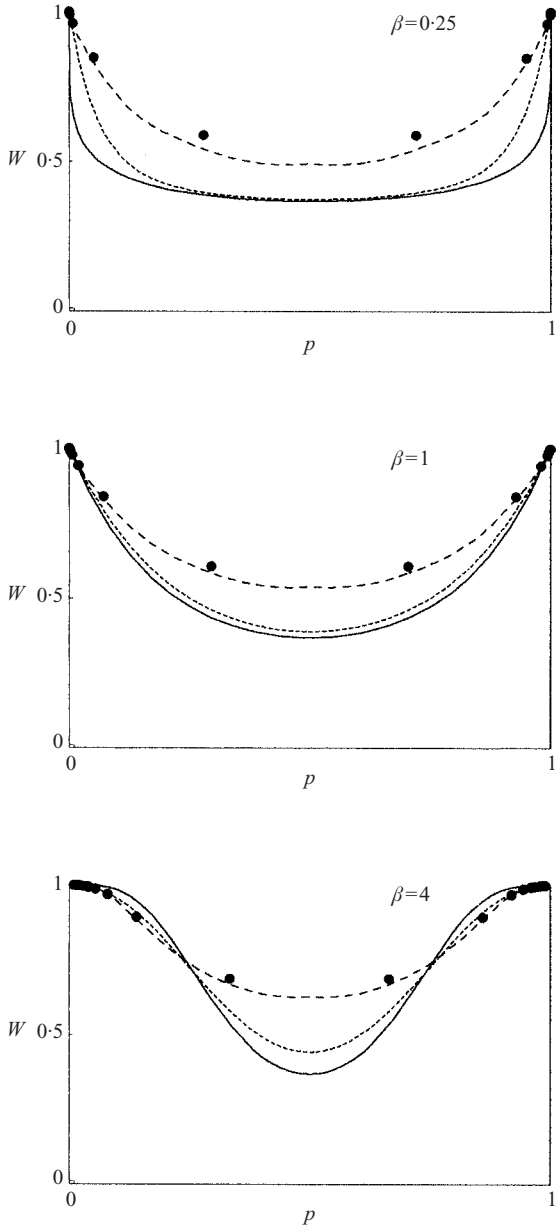

Fig. 10. Fitness plotted against allele frequency at selected loci for $\beta = 0{\cdot}25$, 1, 4; $S = 1$, $n = 10$ loci. In each graph, the lower continuous curve shows individual fitness as a function of p; this would be the mean fitness of a population with zero variance. The dots show the actual mean fitness across the cline, calculated under the exact model, as for Figs 2, 3. The short dashed curves show the approximation in which the mean fitness is calculated for a population following a discrete hypergeometric distribution, with linkage equilibrium variance $\bar{p}\bar{q}/2n$. The upper long dashed curves are calculated in the same way, but account for the pairwise linkage disequilibrium generated by dispersal, using (15).

approximation. One possible source of error is that the mean fitness is higher than e^{-2Spq}, because the number of heterozygous loci varies around the expectation, $2npq$. However, at $S = 1$, with $n = 10$ loci, the mean fitness is only 1·74% higher than expected at the centre, whereas the clines are 13·4% wider than expected. Indeed, the discrepancy does not arise from the effects of linkage disequilibrium, because it is seen for a single selected locus ($n = 1$). The diffusion approximation breaks down when selection on each locus becomes strong, because the clines are then so narrow that the approximation that change is continuous in time and space fails. For $S \sim 2$, linkage disequilibrium is so strong that selection acts as though on a single locus, so that cline width becomes almost independent of the number of loci: the cline widths converge to the same value to the right of Fig. 9.

With disruptive selection, variation around the population mean can substantially increase the mean fitness, thus tending to broaden the clines and reduce the barrier to gene flow. This effect is shown in Fig. 10, in which mean fitness is plotted against allele frequency across the cline, for $S = 1$ and $n = 10$ loci. The lower curve shows the individual fitness as a function of trait value; this would be the mean fitness of a population with zero variance. The actual mean fitness is substantially higher (dots in Fig. 10), because there is considerable variance around the mean, which generates extreme individuals with higher fitness. The genic variance due to heterozygosity at individual loci makes a negligible contribution (lower short dashed line in Fig. 10). Most of the variance is contributed by linkage disequilibrium, which in turn is mostly generated by dispersal into the hybrid zone (Fig. 6). Thus, the mean fitness is accurately predicted from the

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

variance contributed by dispersal (15; upper long dashed curve in Fig. 10). In this example, the net loss of mean fitness is roughly halved by the variance from linkage disequilibrium. It is important to note that this amelioration of selection against hybrids is for the most part not due to selection favouring fitter gene combinations, but rather is a fortuitous consequence of the mixing of populations. If clines were maintained by stabilizing selection towards a changing optimum, one would expect the positive linkage disequilibrium generated by dispersal to *reduce* fitness.

With strong selection, the increase in disequilibrium towards the centre which generates extreme phenotypes could in principle overcompensate for the lower fitness of intermediate genotypes. Indeed, under the QLE approximation for the linkage disequilibrium generated by dispersal, and the discrete hypergeometric approximation to the distribution of genotypes, the mean fitness can *increase* towards the centre. This effect is seen (albeit weakly) for $S = 1$, $\beta = 1, 4$ (upper dashed curves in Fig. 10). However, we have not observed this predicted upturn in exact calculations. First, the clines tend to move so that there is no deme precisely at the centre (dots in Fig. 10). While selection does not maximize mean fitness in the presence of recombination and migration (e.g. Hastings, 1981; Barton & Hewitt, 1989), clines do tend to move so as to prevent the formation of highly unfit populations with $p = 0{\cdot}5$. Secondly, the approximations break down for strong selection and small numbers of genes, and so their predictions fail when selection is strong enough that they predict an upturn in mean fitness. Nevertheless, it is conceivable that for other models of selection the increased variance at the centre could cause a net increase in mean fitness. Even where this is not found, the high variance at the centre does weaken the force of disruptive selection at the centre, and hence broadens the clines.

7. Estimating the number of genes responsible for a barrier to gene flow

Knowledge of the number of genes responsible for reproductive isolation is crucial for distinguishing models of speciation. It has recently become feasible to map such genes directly, just as for other kinds of quantitative trait loci (Wu & Palapoli, 1994; Lynch & Walsh, 1998). However, this approach can only be applied where large numbers of molecular markers can be scored from controlled crosses. An indirect approach is to compare the net loss in mean fitness of a hybrid population with the width of clines at selected loci. Since the former depends on the total selection acting, whereas the latter depends primarily on the selection on each locus, the ratio between them gives an estimate of gene number. The fitness (or at least, some component of fitness) can be measured directly (Barton & Hewitt, 1981), or can be inferred from the distortion induced in clines for neutral markers (Szymura & Barton, 1986, 1991). In this section, we examine the accuracy of the latter method.

The barrier to gene flow at neutral loci, which is reflected in distortion of neutral clines, is determined by the pattern of mean fitness across the hybrid zone (23). This in turn is determined by the shape of the clines at the underlying selected loci. If selection is spread over many loci, then each of these clines will be wide. The barrier will be strong, primarily because selection is spread over a longer distance; in addition, the harmonic mean recombination rate is likely to be lower with larger numbers of genes. Thus, a comparison of the net reduction in mean fitness (inferred from 22) with the net barrier strength (from 23), gives an estimate of the number of genes under selection (Barton & Hewitt, 1981; Szymura & Barton, 1986, 1991). The net barrier strength is proportional to the width of the region over which mean fitness is reduced (23), and therefore to the width of the selected clines. Hence, this method is more sensitive to the pattern of selection than are inferences about rates of dispersal and mean fitness. In most cases, the nature of selection is unknown, and so estimates of gene number must be made assuming an arbitrary model, such as multiplicative selection against heterozygotes. Errors therefore arise not only from the approximation of selection and gene flow by selection gradients and diffusion, but also from differences in the widths of clines maintained in different ways. Here, we investigate the robustness of estimates of gene number to different kinds of selection.

For a cline maintained by multiplicative selection against heterozygotes, the net barrier strength is found by combining (23), (28):

$$B = \phi f_H[\phi, \phi^*] \quad \left(\frac{n\sigma^2}{2S}\right), \tag{29}$$

where

$$f_H[\phi, \phi^*] = \int_0^1 \frac{((1-2s\bar{p}\bar{q})^{-\phi^*/s} - 1)\sqrt{8(1+s/2\phi)}}{\sqrt{[(1-2s\bar{p}\bar{q})^{-2\phi/s} - 4\phi\bar{p}\bar{q} - 1]}} \mathrm{d}\bar{p},$$

where

$$\phi = \frac{S}{\bar{r}}\left(1 - \frac{1}{n}\right), \quad \phi^* = \frac{S}{\bar{r}}.$$

Equation (29) differs from the corresponding formula in Barton (1986), which made the assumption that $(1-2s\bar{p}\bar{q})^{-2\phi/s} \sim \mathrm{e}^{-4\phi\bar{p}\bar{q}}$, and which did not distinguish the coefficient ϕ (which describes the effect of the $n-1$ other loci on one of the n selected loci) from ϕ^* (which describes the effect of the n selected loci on the neutral cline). However, the two formulae converge for $s \ll 1$, $n \gg 1$. As the ratio of selection to recombination, $\phi^* = S/\bar{r}$, increases, the mean fitness, $\bar{W}$, decreases, and

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

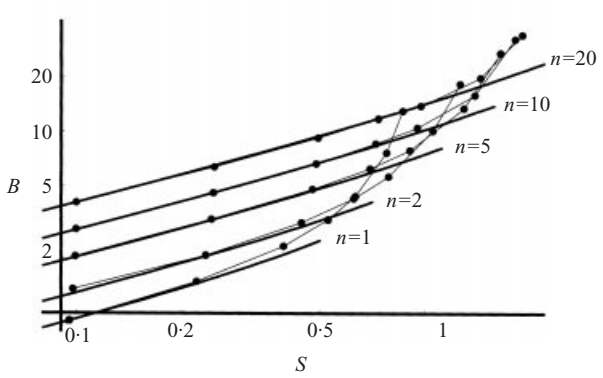

Fig. 11. Barrier strength, plotted against net selection, S, against heterozygotes; fitness is $(1-s)^n$, where k is the number of heterozygous loci, and $s = S/n$. The dots show values from the exact model, with $m = 0{\cdot}5$, starting from a linear gradient. Results are shown for $n = 1, 2, 5, 10, 20$ loci. Barrier strength is scaled relative to σ, which here is equal to $\sqrt{m} = \sqrt{0{\cdot}5}$ deme spacings. Parameters were as for Fig. 7. The thick continuous curves show the prediction of (29), which accounts for the pairwise linkage disequilibrium generated by dispersal. For $n = 1$, this reduces to the prediction in the absence of linkage disequilibrium, $4\sigma\sqrt{n}/2S$ (Bazykin, 1969), which shows as a straight line on this log–log plot.

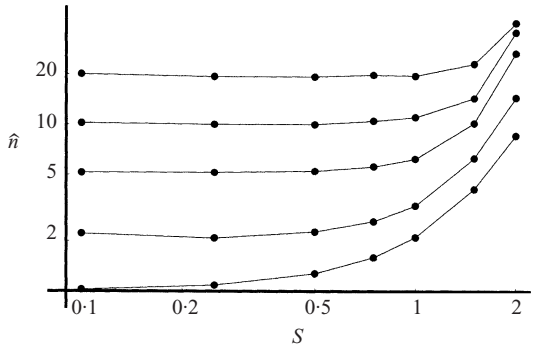

Fig. 12. The number of loci, estimated using (29), on the assumption that selection acts against heterozygotes. Estimates are based on clines in neutral allele frequency, generated with selection against heterozygotes, as in Figs 9, 11. Estimates are plotted against net selection, S, for $n = 1, 2, 5, 10, 20$ loci.

so the term $(\bar{W}^{-1/\bar{r}} - 1) = ((1-2s\bar{p}\bar{q})^{-\phi^*/s} - 1)$ in the numerator increases. However, this is compensated by the narrowing of the clines, which reduces the width of the region over which neutral allele frequencies are steepened. Thus, the function $f_H[\phi, \phi^*]$ lies in a narrow band $(2 \leqslant f_H[\phi, \phi^*] \leqslant 2\sqrt{2}$ for small $s)$. For weak selection, the net barrier strength is close to the simple approximation $B = \frac{\sigma}{\bar{r}}\sqrt{2nS}$. For example, with $n = 10$, $\bar{r} = 1/2$, $S = 1$, there is only a 13% discrepancy between this and (29).

Fig. 11 shows that (29) gives an accurate approximation to the net barrier strength, provided that selection on each locus is weak enough for the power law relation $B \sim \sqrt{S}$ to hold (straight lines to lower left of Fig. 11). When selection is concentrated on a single locus, barrier strength is substantially underestimated for $S > 0{\cdot}5$ ($n = 1$; lower right of Fig. 11). However, when selection is spread across 10 or more loci, the prediction is accurate up to $S \sim 1$. The approximations to both barrier strength (29; Fig. 11) and cline width (28; Fig. 9) break down when selection on each locus is so strong that the hybrid zone is made up of only a few demes.

Since the barrier strength is approximately proportional to $\sqrt{n}$, a comparison with the net reduction in mean fitness ($\sim e^{-S/2}$, inferred from the gradient in neutral clines; 22) and the dispersal rate (σ, inferred from linkage disequilibria; 15) gives an estimate of the number of genes, n. The coupling coefficient, $\phi^* = S/\bar{r}$, can be inferred from the net reduction in mean fitness, provided that some assumption is made about the harmonic mean recombination rate, $\bar{r}$. If selection does in fact act against heterozygotes, errors in estimating gene number arise from two causes. First, the minimum mean fitness is somewhat greater than $e^{S/2}$, because the number of heterozygous loci varies around the expectation of $2pq$, and because the allele frequency in the deme nearest the centre is not equal to 0·5. Secondly, the prediction of (29) is inaccurate for strong selection (right of Fig. 11). Fig. 12 shows estimates made from clines in neutral allele frequency, using (29), where selection acts against heterozygotes at $n = 1, 2, 5, 10, 20$ loci. These estimates are accurate for weak selection, but become much too high when selection is strong and the number of loci small (lower right of Fig. 12). This is because the actual barrier is then much stronger than is predicted by (29) (right of Fig. 11). Nevertheless, when there are more than 10 loci, estimates are reasonably accurate for $S < 1{\cdot}5$, which corresponds to a reduction in mean fitness at the centre of $e^{-S/2} \sim 0{\cdot}47$. This is similar to the value inferred by Szymura & Barton (1991) for the *Bombina* hybrid zone.

In practice, we do not usually know how selection acts to produce reproductive isolation; estimates of gene number must then be made by choosing an arbitrary null model, such as selection against heterozygotes. Fig. 13 shows how estimates of gene number, made assuming this null model, vary with net selection, S, when the barrier is in fact maintained by epistasis amongst $n = 10$ loci. With $\beta = 1$ (middle curve in Fig. 13), the estimate is close to the actual value of $n = 10$ for weak selection. As with underdominance (Fig. 12), numbers are overestimated for strong selection ($S > 0{\cdot}75$), which corresponds to a reduction in mean fitness of $e^{-S} \sim 0{\cdot}47$. With epistasis $\beta = 0{\cdot}25$ (upper curve in Fig. 13), fitness is reduced for a wider range of hybrid genotypes, and so the barrier is stronger; hence, the number of genes involved is overestimated even when selection is weak ($\hat{n} = 20{\cdot}6$ *v.* $n = 10$ for $S = 0{\cdot}1$). Conversely, the gene number is underestimated with epistasis $\beta = 4$ ($\hat{n} = 3{\cdot}38$ *v.* $n = 10$ for $S = 0{\cdot}1$).

https://doi.org/10.1017/S0016672399004334 Published online by Cambridge University Press

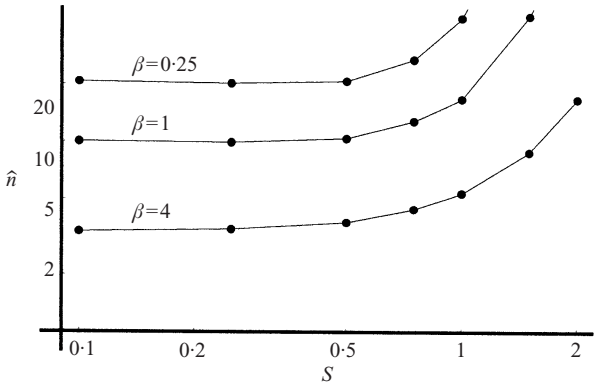

Fig. 13. The number of loci, estimated using (29) on the assumption that selection acts against heterozygotes. Estimates are based on clines in neutral allele frequency, generated under the epistatic model with $\beta = 0{\cdot}25$, 1, 4. Parameters are as in Figs 7, 8, with $n = 10$ loci, and are plotted against the net selection against heterozygotes, S.

With different numbers of genes, the pattern is similar to Fig. 12, though estimates remain independent of S for stronger selection when more genes are involved. In general, there is a systematic error with weak selection when β deviates from 1.

8. Discussion

To a close approximation, there is a quasi-equilibrium between the linkage disequilibrium generated by the diffusion of parental gene combinations into a hybrid zone, and its breakdown by recombination (Fig. 6). This simple relation (15) allows the rate of diffusion, σ^2, to be estimated without any assumptions about what kind of selection is acting. Linkage disequilibrium with selected loci steepens clines at neutral loci, and reflects a barrier to the flow of genes between the hybridizing populations, B. The strength of this barrier depends primarily on the net reduction in mean fitness across the hybrid zone

$$(B = \int_{-\infty}^{\infty} (\bar{W}^{-1/\bar{r}} - 1)\,\mathrm{d}x;\ 23).$$

This approximation is accurate provided that selection is not too strong ($S < 1$, say; Figs. 8, 11), and allows the pattern of mean fitness to be estimated from neutral allele frequencies. Again, there is not need for any assumptions about the nature of selection.

Predicting the shape of the selected clines, and hence mean fitness across the hybrid zone, is more difficult, because it does depend on the genetic basis of fitness differences. In the model of epistasis used here, selection coefficients tend to zero in the centre of the hybrid zone when $\beta \ll 1$; this leads to a cline which is shallow in the centre but approaches fixation rapidly (i.e. faster than exponentially) in the tails. Conversely, if $\beta \gg 1$, selection coefficients tend to zero at the edges, and so there is a sharp central step, flanked by shallow tails. For the intermediate case $\beta = 1$, selection coefficients change smoothly, and cline shape is close to linear on a logistic scale (Figs. 4, 5). This is a similar pattern to that seen for selection against heterozygotes. These distorted cline shapes are unlikely to be observed directly. The mean fitness, which can be observed either directly or through its effects on neutral markers, would not be appreciably distorted even if β were far from 1. The model of epistasis assumed here is also an extreme one, in that fitness depends solely on the net proportion of genes derived from one or other parental population. If hybrid fitness were reduced by the separate effects of several independent sets of interacting genes, the average selection coefficient would not change so drastically across the hybrid zone.

It is important to realize that because the direct effect of epistasis in producing linkage disequilibrium is negligible (Fig. 6), the effects of the parameter β on cline shape and on the barrier to gene flow are not directly due to epistasis components of fitness variation. Rather, they are due to the change in additive selection coefficient with genetic background. The same effects would be seen in the absence of epistasis, if selection coefficients changed as a result of frequency-dependence, or as a result of a changing environment. Thus, observations on genotype frequencies in hybrid zones do not tell us about the epistatic variation segregating within hybrid populations, or about the relative importance of 'endogenous' and 'exogenous' selection.

In order to find analytical approximations for the shape of selected clines, the population must be described in terms of allele frequencies rather than the full set of genotype frequencies. This is straightforward, since pairwise linkage disequilibrium can be closely approximated by a quasi-equilibrium between dispersal and recombination (15). However, predicting the change in allele frequency due to selection is still not simple, because some assumption must be made about the full distribution of genotype frequencies. This is a general difficulty in understanding the response of polygenic traits to selection (Turelli & Barton, 1994). Assuming that the number of '1' alleles follows a discrete hypergeometric distribution gives accurate results (Table 1; Figs. 2–5); this can be seen as an extension of the common assumption that breeding values follow a Gaussian distribution. However, the further approximation that the proportion of '1' alleles follows a continuous Beta distribution is only accurate for $\beta = 1$; for more extreme models of epistasis, there are sharp changes in fitness with genotype, and so approximating to a continuous polygenic trait is inaccurate.

The reduction in mean fitness at the centre of a hybrid zone is a measure of the total selection acting. In contrast, the width of the region of reduced fitness reflects the width of each selected cline, which depends on the selection acting on each locus. Comparison of

the width and depth of the fitness trough (both of which can be inferred from the shape of neutral clines) therefore gives an estimate of the number of genes. If selection is arbitrarily assumed to act against heterozygotes, estimates will be in error by a factor which depends on how selection actually acts (Fig. 13). In addition, the diffusion approximation breaks down when selection on each locus becomes so strong that the hybrid zone spans only a few demes (Figs 12, 13).

The exact model used here assumes that all haploid genotypes with the same number of '1' alleles are equally frequent; it is thus restricted to unlinked loci. With some degree of linkage, the harmonic mean recombination rate will be less than 0·5, and the barrier to gene flow correspondingly stronger. Barton & Gale (1993) and Kruuk *et al*. (in press) investigated stochastic models with linkage, and found that predictions for linkage disequilibrium and barrier strength are accurate up to similar selection strengths to those seen here. However, further investigation of the accuracy of estimates of gene number with linkage is warranted. Note that linkage introduces additional uncertainty in estimating gene number, since the harmonic mean recombination rate depends on the genetic map, and varies randomly with the actual configuration of selected loci.

We have assumed that disruptive selection acts on an additive polygenic trait, namely, the proportion of alleles derived from one or other parental population, p. The comparison of various analytical predictions for the selection response (Table 1) may give a more general insight into methods for approximating quantitative traits. The properties of the cline in p may also apply more generally: because the genetic variance in the trait increases in the centre of the cline, due to both linkage disequilibrium and increased heterozygosity, the selection response increases, steepening the clines into a sharp step. However, the exact model used in this paper cannot describe (say) stabilizing selection towards a changing optimum, because the symmetrical solution is then unstable: some loci approach fixation for '1' alleles, and some for 'O' alleles (Barton, in preparation). Stochastic simulations are necessary for a full understanding of clines in polygenic traits.

We are grateful to Loeske Kruuk and Michael Turelli for their helpful comments on the manuscript. N.B. was supported by grants GR3/11635 from the NERC and GR/L10048 from the EPSRC, and by the Darwin Trust of Edinburgh. M.S. was supported by a graduate student fellowship from the Yale Institute for Biospheric Studies.

References

Arnold, M. (1996). *Natural Hybridization and Introgression*. Princeton: Princeton University Press.

Barton, N. H. (1979). Gene flow past a cline. *Heredity* **43**, 333–339.

Barton, N. H. (1983). Multilocus clines. *Evolution* **37**, 454–471.

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

Barton, N. H. (1992). On the spread of new gene combinations in the third phase of Wright's shifting balance. *Evolution* **46**, 551–557.

Barton, N. H. (1995). Linkage and the limits to natural selection. *Genetics* **140**, 821–841.

Barton, N. H. (1999). Clines in polygenic traits. *Genetical Research*, in press.

Barton, N. H. & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity* **57**, 357–376.

Barton, N. H. & Charlesworth, B. (1984). Genetic revolutions, founder effects, and speciation. *Annual Review of Ecology and Systematics* **15**, 133–164.

Barton, N. H. & Gale, K. S. (1993). Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process* (ed. R. G. Harrison), pp. 13–45. Oxford: Oxford University Press.

Barton, N. H. & Hewitt, G. M. (1981). The genetic basis of hybrid inviability between two chromosomal races of the grasshopper *Podisma pedestris*. *Heredity* **47**, 367–383.

Barton, N. H. & Hewitt, G. M. (1989). Adaptation, speciation and hybrid zones. *Nature* **341**, 497–503.

Barton, N. H. & Shpak, M. (2000). The stability of symmetrical solutions to polygenic models. *Theoretical Population Biology* (submitted).

Barton, N. H. & Turelli, M. (1991). Natural and sexual selection on many loci. *Genetics* **127**, 229–255.

Bateson, W. (1909). Heredity and variation in modern lights. In *Darwin and Modern Science* (ed. A. C. Seward), pp. 85–101. Cambridge: Cambridge University Press.

Bazykin, A. D. (1969). Hypothetical mechanism of speciation. *Evolution* **23**, 685–687.

Bengtsson, B. O. & Christiansen, F. B. (1983). A two-locus mutation–selection model and some of its evolutionary implications. *Theoretical Population Biology* **24**, 59–77.

Bulmer, M. G. (1985). *The Mathematical Theory of Quantitative Genetics*. Oxford: Oxford University Press.

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. New York: Columbia University Press.

Doebeli, M. (1996). A quantitative genetic competition model for sympatric speciation. *Journal of Evolutionary Biology* **9**, 893–910.

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

Fisher, R. A. (1950). Gene frequencies in a cline determined by selection and diffusion. *Biometrics* **6**, 353–361.

Gavrilets, S. & Gravner, J. (1997). Percolation on the fitness hypercube and the evolution of reproductive isolation. *Journal of Theoretical Biology* **184**, 51–64.

Haldane, J. B. S. (1948). The theory of a cline. *Journal of Genetics* **48**, 277–284.

Hastings, A. (1981). Stable cycling in discrete time genetic models. *Proceedings of the National Academy of Sciences of the USA* **78**, 7224–7225.

Hastings, A. (1986). Limits to the relationship among recombination, disequilibrium, and epistasis in two-locus models. *Genetics* **113**, 177–185.

Hatfield, T., Barton, N. H. & Searle, J. (1992). A model of a hybrid zone between two chromosomal races of the common shrew (*Sorex araneus*). *Evolution* **46**, 1129–1145.

Kondrashov, A. S. (1984). On the intensity of selection for reproductive isolation at the beginnings of sympatric speciation. *Genetika* **20**, 408–415.

Kruuk, L. E. B., Baird, S. J. E., Gale, K. S. & Barton, N. H.

(in press). The effect of endogenous and exogenous selection on multilocus clines. *Genetics* (in press).
Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**, 314–334.
Li, W. H. & Nei, M. (1974). Stable linkage disequilibrium without epistasis in subdivided populations. *Theoretical Population Biology* **6**, 173–183.
Lynch, M. & Walsh, J. B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.
McCarthy, E. M., Asmussen, M. A. & Anderson, W. W. (1995). A theoretical assessment of recombinational speciation. *Heredity* **74**, 502–509.
Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biological Symposia* **6**, 71–125.
Nagylaki, T. (1975). Conditions for the existence of clines. *Genetics* **80**, 595–615.
Nagylaki, T. (1976). Clines with variable migration. *Genetics* **83**, 867–886.
Nagylaki, T. (1978). Clines with asymmetric migration. *Genetics* **88**, 813–827.
Nagylaki, T. (1993). The evolution of multilocus systems under weak selection. *Genetics* **134**, 627–647.
Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* **139**, 1805–1813.
Orr, H. A. (1997). Dobzhansky, Bateson and the genetics of speciation. *Genetics* **144**, 1331–1335.
Rieseberg, L. H. (1995). The role of hybridization in evolution: old wine in new skins. *American Journal of Botany* **82**, 944–953.
Shpak, M. & Kondrashov, A. S. (1998). Phenotypic model of a quantitative trait with equal allele frequencies. *Evolution* (in press).
Szymura, J. M. & Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland. *Evolution* **40**, 1141–1159.
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.
Turelli, M. & Barton, N. H. (1994). Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? *Genetics* **138**, 913–941.
Wright, S. (1935). Evolution in populations in approximate equilibrium. *Journal of Genetics* **30**, 257–266.
Wu, C. I. & Palapoli, M. F. (1994). Genetics of postmating reproductive isolation in animals. *Annual Review of Genetics* **27**, 283–308.