The effect of a barrier to gene flow on patterns of geographic variation

N. H. BARTON*

Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, Scotland, UK

(*Received 17 October 2007 and in revised form 29 October 2007*)

Summary

Explicit formulae are given for the effects of a barrier to gene flow on random fluctuations in allele frequency; these formulae can also be seen as generating functions for the distribution of coalescence times. The formulae are derived using a continuous diffusion approximation, which is accurate over all but very small spatial scales. The continuous approximation is confirmed by comparison with the exact solution to the stepping stone model. In both one and two spatial dimensions, the variance of fluctuations in allele frequencies increases near the barrier; when the barrier is very strong, the variance doubles. However, the effect on fluctuations close to the barrier is much greater when the population is spread over two spatial dimensions than when it occupies a linear, one-dimensional habitat: barriers of strength comparable with the dispersal range ($B \approx \sigma$) can have an appreciable effect in two dimensions, whereas only barriers with strength comparable with the characteristic scale ($B \approx L = \sigma/\sqrt{2\mu}$) are significant in one dimension (μ is the rate of mutation or long-range dispersal). Thus, in a two-dimensional population, barriers to gene flow can be detected through their effect on the spatial pattern of genetic marker alleles.

1. Introduction

Natural populations are not distributed uniformly: even a continuous habitat will usually be subdivided by local reductions in density or dispersal. On a larger scale, a biological species may be split into a mosaic of geographic races, separated by narrow zones of hybridization. The flow of genes between these races will be impeded by the reduced fitness of hybrids, or by adaptation to different environments: even if an allele is not itself disadvantageous in the new population, it will be associated with alleles at selected loci, and so will tend to be eliminated with them (Barton & Bengtsson, 1986).

The effects of such barriers to gene flow are of interest from several points of view. Local adaptations tend to be swamped by dispersal from different environments: divergence may be made possible by barriers (Slatkin, 1973; Nagylaki, 1976). Conversely, populations may persist in marginal habitats only if they are sustained by immigration, which directly increases population size (Pulliam, 1988), and also introduces genetic variation. Geographic races often remain distinct, despite the continual exchange of genes across hybrid zones. Thus, although they are members of the same biological species, it is not clear to what extent they evolve independently. The chance that favourable alleles that become established in one race arose as mutations in the other will depend on barrier strength, as will the degree of neutral molecular differentiation.

The strength of a barrier to gene exchange, whether due to a physical obstacle or to genetic incompatibility, can be measured by a quantity B. This has the dimensions of distance, and can be thought of as the length of habitat which would pose the same obstacle to the steady flow of a neutral allele. The effect of the barrier on gene flow depends on the ratio between barrier strength and dispersal rate (B/σ, where σ^2 is the variance in distance between parent and offspring); $(B/\sigma)^2$ has the dimensions of time. A neutral allele will be delayed by $\approx (B/\sigma)^2$ generations, whilst an allele with advantage s will only be delayed significantly if $(B/\sigma)^2$ is much greater than the timescale characteristic of selection, $1/s$ (Barton & Bengtsson, 1986; Pialek & Barton, 1997).

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

The rate of gene exchange between populations can, in principle, be measured directly, from the movements of individuals, and from the fitness of them and their offspring. However, such measurements are laborious, even with the most suitable organisms. Barrier strength can also be measured indirectly, from the shape of clines at selected loci: a barrier will produce a sharp step in allele frequency, whose size is the product of barrier strength and the gradient in gene frequency ($\Delta p = B(\partial p/\partial x)$) (Nagylaki, 1976; Barton & Bengtsson, 1986). However, such clines will only be found at a limited number of loci: a more general method for measuring barrier strength would be to use the pattern of random fluctuations in the frequencies of polymorphic marker alleles that are inevitably generated by random drift. Such measurements could be based on variation detected by enzyme electrophoresis or by DNA sequencing; since the effect of drift and dispersal on additive polygenic characters is formally identical to that on allele frequencies, morphological traits might in principle also be used (Felsenstein, 1986).

The aim of this paper is to develop theoretical predictions for the effect of a local barrier on fluctuations produced by random sampling drift. The results give the covariance between allele frequencies at different spatial positions; however, these can be interpreted in terms of the distribution of pairwise coalescence times, or the pattern of identity by descent.

An absolute barrier to gene flow is equivalent to a range boundary. Wilkins (2004) reviews earlier work on the effects of such absolute boundaries in allele frequencies in two-dimensional populations, and shows how this is related to the distribution of coalescence times. Nagylaki (1988) and Nagylaki & Barcilon (1988) have modelled the effects of a partial barrier on geographic patterns in a linear, one-dimensional population. However, most natural populations are distributed over two spatial dimensions; the effects of drift are then qualitatively different. In one dimension, drift produces smooth fluctuations, so that allele frequencies at points a few dispersal distances apart are strongly correlated. In two dimensions, however, frequencies can differ substantially even over short distances (Wright, 1943; Malecot, 1948; Kimura & Weiss, 1964).

Previous analyses of the effects of barriers to gene flow have assumed that individuals move along a line, in a one-dimensional habitat. However, the effects of random drift are qualitatively different when individuals can move in two dimensions, as is more usual in nature. In this paper, I extend previous results to the two-dimensional case, and show that because fluctuations are localized, barriers have effects which can be detected more readily than in a linear habitat. The existence of local fluctuations

in two dimensions complicates the mathematical formulation, since the diffusion approximation breaks down over small scales (Nagylaki, 1978; Barton *et al.*, 2002). In the first part of the paper, I ignore this difficulty, and derive analytic results which apply over all but the shortest scales. In the Appendix, I derive exact results for this discrete model, and show how these converge to the continuous diffusion approximation.

2. Summary of results for a one-dimensional population

Nagylaki (1988, eq. 14) gives a differential equation for the effects of mutation, drift and dispersal on the probability of identity $F(x, x')$. This is the chance that two genes at positions x and x' in a linear habitat are identical in state; it increases as a result of inbreeding, and decreases as a result of mutation to new alleles. This quantity can also be interpreted as the standardized covariance between fluctuations in allele frequency at x, x' (i.e. $E[(p(x)-\bar{p})(p(x')-\bar{p})]/\bar{p}(1-\bar{p})$, where $E(p) = \bar{p}$ is the expected allele frequency, assumed to be everywhere the same). Nagylaki's equation allows for smooth variations in dispersal and density; he also gives formulae for the effects of discontinuous changes in density or dispersal, and of local regions of reduced density or dispersal (i.e. of local barriers). This paper only considers the simple case where the density of diploid individuals (ρ) and the dispersal rate (σ^2) are uniform everywhere apart from a barrier at $x = 0$. Then, for non-zero x, x':

$$\frac{\partial F}{\partial t} = \frac{\sigma^2}{2}\left[\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial x'^2}\right] - 2\mu F + \frac{(1-F)\delta(x-x')}{2\rho}. \quad (1a)$$

The terms on the right represent dispersal, mutation and drift, respectively. σ^2 is the variance in distance between parent and offspring, and μ is the rate of mutation to new alleles, assuming the infinite alleles model. The equation also describes the effects of long-distance migration at a rate μ from a population with frequency $\bar{p}$; it approximates balancing selection towards an equilibrium at $\bar{p}$, provided that fluctuations are small enough that changes due to selection are roughly linear in $p(\Delta p \approx -\mu(p-\bar{p}))$. The Dirac delta function $\delta(x-x')$ represents the generation of uncorrelated fluctuations by random drift, with variance inversely proportional to the density of genes, $1/2\rho$.

Equation (1a) describes changes in identity across a plane with axes x, x' (Fig. 1a); this represents the relation between pairs of individuals that live in a linear, one-dimensional habitat. The third term, representing sampling drift, increases F on a ridge along the diagonal $x = x'$. As a consequence, F is continuous across this diagonal, but its gradient changes abruptly from
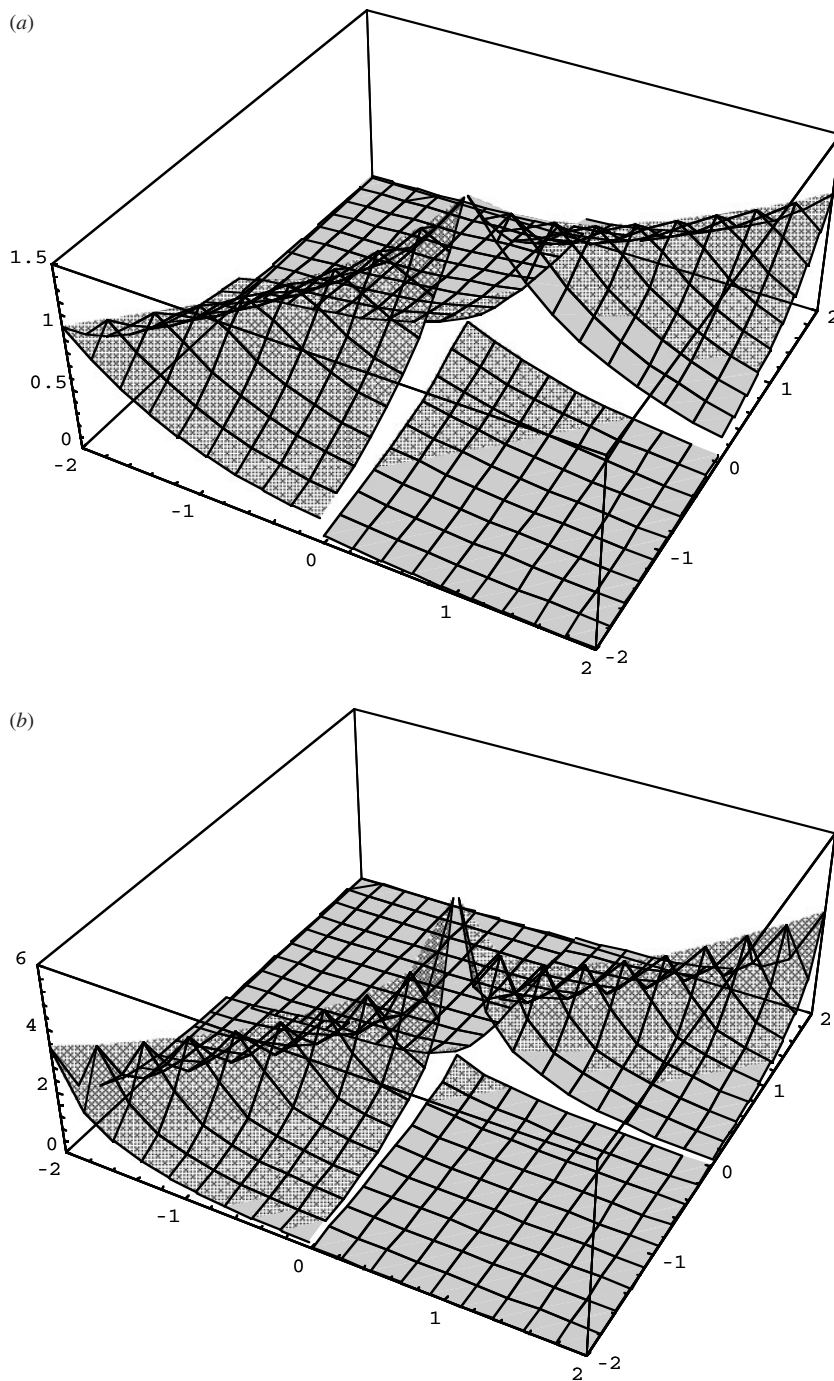
(a)

(b)

Fig. 1. (a) The probability of identity, $F(x, x')$, against x, x', for a population in a one-dimensional habitat. Values are calculated from equation (2). F is standardized relative to its value far from the barrier, $1/(4\rho\sigma\sqrt{2\mu})$, and distance is scaled relative to $L=\sigma/\sqrt{2\mu}$). The barrier strength is twice the characteristic scale ($B=2L=2\sigma/\sqrt{2\mu}$). The height of the ridge along the diagonal ($x=x'$) represents the probability of identity of nearby genes; this is 50% greater near the barrier than far away, as can be seen from the increased height of the ridge towards the centre ($F(0,0)=(1+B/(2L+B))/(4\rho\sigma\sqrt{2\mu})$ vs $F(\infty,\infty)=1/(4\rho\sigma\sqrt{2\mu})$; equation 2a). The probability of identity falls abruptly when the two genes are separated by the barrier; this is shown by the sharp drop in F across the axes $x=0$, $x'=0$. For this example, the probability of identity of two nearby genes, just separated by the barrier, is one-third that of two nearby genes on the same side ($F(0-,0+)=(2L/(2L+B))/(4\rho\sigma\sqrt{2\mu})$ vs $F(0+,0+)=(1+B/(2L+B))/(4\rho\sigma\sqrt{2\mu})$; equation 2b). (b) The analogous graph for a two-dimensional population. Values are calculated from equation (5). F is now standardized relative to the dimensionless parameter $1/(4\pi\rho\sigma^2)$; it is given for points opposite each other (that is, $y=y'$). In a two-dimensional population, the continuous approximation breaks down for nearby genes; the value near the diagonal ($x=x'$, $y=y'$) is calculated with the aid of exact results for the stepping stone model, with $m=0.4$, $\mu=5\times10^{-4}$.

$+(1-F)/(4\rho\sigma^2)$ to $-(1-F)/(4\rho\sigma^2)$. A barrier at zero introduces discontinuities in F along the axes $x=0$ and $x'=0$, representing the decrease in identity between genes that are separated by the barrier. Although the gradient of F is continuous across the barrier, F itself drops by $B(\partial F/\partial x)$ at $x=0$, and by $B(\partial F/\partial x')$ at $x'=0$ (from eq. 54, Nagylaki, 1988). Formally, these conditions give

$$F(x,0+)-F(x,0-)=B\frac{\partial F}{\partial x'} \quad \text{at } x'=0, \qquad (1b)$$

$$F(0+,x')-F(0-,x')=B\frac{\partial F}{\partial x} \quad \text{at } x=0. \qquad (1c)$$

Nagylaki & Barcilon (1988) give equilibrium solutions to this equation for the case of an impenetrable barrier. The exact solution for arbitrary B is a complicated integral (Nagylaki & Barcilon, 1988, eqs. 4.1, 4.4). However, when density and dispersal are large enough that fluctuations are small $(4\rho\sigma\sqrt{2\mu}\gg1)$, the term $(1-F)$ in (1a) can be approximated to 1, and the equation can be solved explicitly:

For x, x' on the same side of the barrier ($x<0$, $x'<0$ or $x>0$, $x'>0$):

$$F(x,x')=\frac{1}{4\rho\sigma\sqrt{2\mu}}\left\{e^{-|x-x'|/L}+\frac{B}{2L+B}e^{-|x+x'|/L}\right\}. \qquad (2a)$$

For x, x' on different sides of the barrier ($x<0$, $x'>0$ or $x>0$, $x'<0$):

$$F(x,x')=\frac{1}{4\rho\sigma\sqrt{2\mu}}\left(\frac{2L}{2L+B}\right)e^{-|x-x'|/L}. \qquad (2b)$$

Here, $L=\sigma/\sqrt{2\mu}$ is the characteristic scale, which is determined by the balance between mutation and dispersal (Slatkin, 1973). When mutation is rare ($\mu\ll1$), then $L\gg\sigma$. This separation between the characteristic scale, L, and the scale of local dispersal, σ, plays a central role in the analysis.

3. The effect of a barrier in two dimensions

Nagylaki (1988) derived his differential equation by writing a recurrence equation for F in a discrete stepping stone model, with deme size N and a rate of exchange $m/2$ between adjacent demes. (Nagylaki *et al.*, 1993, and Ayatia *et al.*, 1999, deal directly with the diffusion limit). If the distance between demes is ε, then this corresponds to a continuous model with $\sigma^2=m\varepsilon^2$, and density $\rho=N/\varepsilon$ in one dimension, N/ε^2 in two dimensions. In his one-dimensional model, Nagylaki (1988) took the limit $\mu\to0$, $N\to\infty$, keeping m and $N\sqrt{\mu}$ fixed; this gives a continuous solution which applies over all scales. Unfortunately, this procedure

fails in two dimensions. This is because the magnitude of random fluctuations then depends on the dimensionless parameter, neighbourhood size, which is the product of density and dispersal: $Nb=4\pi\rho\sigma^2=4\pi Nm$. If m is kept fixed, and $N\to\infty$, then $Nb\to\infty$, and fluctuations vanish. In contrast, fluctuations in one dimension depend on the dimensionless combination $4\rho\sigma\sqrt{2\mu}=4N\sqrt{2m\mu}$. This can remain of order one as $N\to\infty$, and $\mu\to0$ (see Fleming & Su, 1974; Nagylaki, 1978; Barton *et al.*, 2002).

In this section, I ignore these difficulties, and use the obvious two-dimensional analogue of equation (1) to obtain an analytic expression for F, which applies in the limit where fluctuations are small ($F\ll1$). The Appendix shows that though this expression diverges over small scales, it converges to the exact solution of the stepping stone model over scales larger than the deme spacing or dispersal range (ε, σ). Numerical results show that accurate predictions can be made by splicing together analytic results for small and large scales.

In a two-dimensional population, the probability of identity depends on the two coordinates of each of the two genes: $F(x,y,x',y')$. The two-dimensional analogue of equation (1) is

$$\frac{\partial F}{\partial t}=\frac{\sigma^2}{2}\left[\frac{\partial^2 F}{\partial x^2}+\frac{\partial^2 F}{\partial x'^2}+\frac{\partial^2 F}{\partial y^2}+\frac{\partial^2 F}{\partial y'^2}\right]$$
$$-2\mu F+\frac{(1-F)\delta(x-x')\delta(y-y')}{2\rho}, \qquad (3a)$$

$$F(x,y,0+,y)-F(x,y,0-,y)=B\frac{\partial F}{\partial x'} \text{ at } x'=0$$
$$\text{for all } y, \qquad (3b)$$

$$F(0+,y,x',y)-F(0-,y,x',y)=B\frac{\partial F}{\partial x} \text{ at } x=0$$
$$\text{for all } y. \qquad (3c)$$

The barrier is assumed to run along the line $x=0$; dispersal occurs at the same rate σ^2 along the x- and y-axes. In the following, we only consider the case where the probability of identity between nearby genes is small ($F(x,y,x,y)\ll1$), and so neglect the factor $(1-F)$ in (3a).

Equation (3) can be solved by taking the Fourier transform along the y-axis. The method can be understood by thinking of the way fluctuations in allele frequency propagate across the barrier. Fluctuations which do not vary in the y direction will follow the one-dimensional solution given above. Fluctuations which vary sinusoidally in the y direction will propagate across the barrier, in the x direction, according to equation (1); however, as well as decaying at a rate μ as a result of mutation, they will also decay as a result of dispersal in the y direction. The net rate of decay will be $(\mu+\sigma^2 Y^2/2)$, where Y

is the spatial frequency along the *y*-axis. Thus, the covariance of fluctuations with a given frequency can be found simply by modifying equation (2) to include the extra decay due to transverse dispersal. Random drift produces uncorrelated fluctuations, and so generates all frequencies with equal intensity: it produces 'white noise'. Summing over frequencies gives the complete solution.

The Fourier transform of $\partial^2 g / \partial y^2$ is $-Y^2 \tilde{g}$. (The transform of $g(y)$, denoted by $\tilde{g}(Y)$, is defined in the Appendix). Applying the transform to equation (3a) with respect to *y* and *y'*, we have

$$\frac{\partial \tilde{F}}{\partial t} = \frac{\sigma^2}{2} \left[\frac{\partial^2 \tilde{F}}{\partial x^2} + \frac{\partial^2 \tilde{F}}{\partial x'^2} \right] - \frac{\sigma^2}{2} (Y^2 + Y'^2) \tilde{F} - 2\mu \tilde{F}$$
$$+ \frac{\delta(x - x') \delta(Y + Y')}{2\rho}. \quad (4)$$

This has the same form as equation (1), with 2μ being replaced by $2\mu + (\sigma^2/2)(Y^2 + Y'^2)$. The solution will be proportional to $\delta(Y + Y')$, which ensures that *F* depends only the displacement along the *y*-axis, $(y - y')$, and that we need only consider $Y = -Y'$. The solution for $\tilde{F}$ is given by equation (2), with 2μ replaced by $2\mu + \sigma^2 Y^2$, and *L* replaced by $L^* = \sigma / \sqrt{2\mu + \sigma^2 Y^2}$. Taking the inverse transform gives:

For *x*, *x'* on the same side of the barrier ($x < 0, x' < 0$ or $x > 0, x' > 0$):

$$F(x, y, x', y') = \int_{-\infty}^{\infty} \frac{e^{i(y - y')Y} dY}{8\pi\rho\sigma\sqrt{2\mu + \sigma^2 Y^2}} \left\{ e^{-|x - x'|/L^*} \right.$$
$$\left. + \frac{B}{2L^* + B} e^{-|x + x'|/L^*} \right\}. \quad (5a)$$

For *x*, *x'* on different sides of the barrier ($x < 0, x' > 0$ or $x > 0, x' < 0$):

$$F(x, y, x', y') = \int_{-\infty}^{\infty} \frac{e^{i(y - y')Y} dY}{8\pi\rho\sigma\sqrt{2\mu + \sigma^2 Y^2}} \left(\frac{2L^*}{2L^* + B} \right)$$
$$\times e^{-|x - x'|/L^*}. \quad (5b)$$

Though I can find no explicit form for these integrals, they can easily be calculated numerically.

In the case of two coincident genes ($x = x', y = y'$), the contribution of high-frequency components only decreases with $1/\sqrt{1 + Y^2} \approx 1/Y$, so that the integral diverges logarithmically. Thus, in two spatial dimensions, most of the variation in allele frequencies is due to short-lived, local fluctuations. Since the continuous approximation does not describe the local structure of the population, it breaks down when the genes are very close together: *F* is in fact finite. The problem is to combine the local structure with the continuous approximation, so as to get predictions valid over all scales.

In the stepping stone model, with no barrier, *F* is given by an expression of the form

$$F^*(x, y, x', y') = \int_{-\pi/\varepsilon}^{\pi/\varepsilon} \frac{e^{i(y - y')Y} e^{-|x - x'|/L} h(Y) dY}{8\pi\rho\sigma\sqrt{2\mu} + \sigma^2 Y^2}. \quad (6)$$

Since the population is confined to a discrete grid with spacing ε, fluctuations with frequency greater than π/ε do not contribute. High-frequency fluctuations are also damped by the term $h(Y)$, which tends to 1 as $Y\varepsilon$ tends to zero (see Appendix). When the separation between the genes is large compared with the deme spacing ($r \gg \varepsilon$), F^* converges to the continuous approximation, given by equation (5a) with $B = 0$. The probability of identity of nearby genes (and hence the standardized variance of allele frequency) depends on the local structure, and takes the form $\log(L/\kappa)/(4\pi\rho\sigma^2)$ (Slatkin & Barton, 1990; Barton *et al.*, 2002). Here, κ is a small distance, of the same order as the deme spacing, which summarizes the complications of local structure. In the stepping stone model, $\kappa = \varepsilon/\sqrt{32}$ for small *m*, and decreases as *m* approaches its maximum of 1/2. An exact expression for κ can be derived from Kimura & Weiss (1964).

The model with a barrier converges to the same form for high frequencies, so that for genes on the same side of the barrier, *F*, can be calculated as:

$$F(x, y, x', y') = F^*(x, y, x', y') + \int_{-\infty}^{\infty} \frac{e^{i(y - y')Y}}{8\pi\rho\sigma\sqrt{2\mu + \sigma^2 Y^2}}$$
$$\times \left\{ \frac{B}{2L^* + B} e^{-|x + x'|/L^*} \right\} dY. \quad (7a)$$

This formula breaks down when the two genes are close together, and are close to the barrier. Then, the barrier influences local fluctuations. A barrier with strength $B \sim L$ appears impenetrable to localized fluctuations, with a high spatial frequency $Y \gg 1/L$, and we have:

$$F(0+, y, 0+, y) = 2F^*(0, 0, 0, 0) - \frac{2L}{\sqrt{B^2 - 4L^2}}$$
$$\times \arccos\left(\frac{2L}{B}\right) + C[m]. \quad (7b)$$

The function *C*[*m*] is calculated in the Appendix; it is $\pi/2$ for small *m*, ~ -3 for $m = 0.4$, and becomes large and negative as *m* approaches 1/2.

4. Results for two dimensions

The analytic results apply when dispersal and population density are high enough that *F* is small ($F \ll 1$). This is the case of most interest, since natural populations usually show low levels of spatial
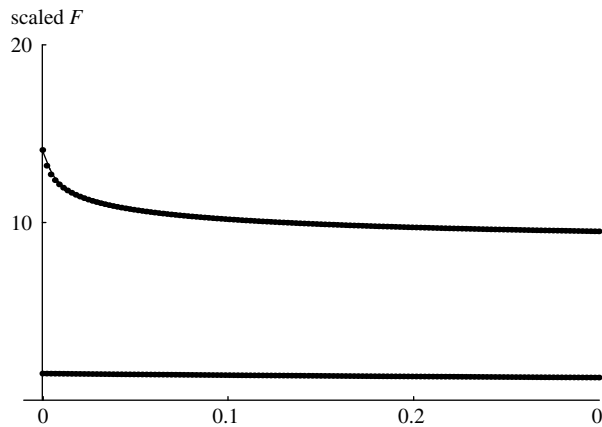
scaled *F*

Fig. 2. The probability of identity between two nearby genes as a function of distance, $X = x/L$, from a barrier of strength $B = 2L$. The lower line gives results for one dimension, and the upper line for two dimensions. As before, F is scaled relative to the dimensionless parameter $1/(4\rho\sigma\sqrt{2\mu})$ in one dimension, and to $1/(4\pi\rho\sigma^2)$ in two dimensions; distance is scaled relative to $L = \sigma/\sqrt{2\mu}$. The dots show exact calculations from the stepping stone model, with $m = 0\cdot4$, $\mu = 10^{-6}$, $L = 447$.

variation, at least for molecular markers ($F_{ST} \approx 0\cdot1$, say; Slatkin, 1987; Rieseberg *et al.*, 2003). (Higher values of F_{ST} may be found in species-wide surveys, but such large-scale patterns are formed over very wide long times. Thus, it is doubtful that there is an equilibrium between drift, mutation and dispersal). I will therefore present the analytic results scaled relative to the dimensionless parameters $4\rho\sigma\sqrt{2\mu}$ in one dimension and $Nb = 4\pi\rho\sigma^2$ in two dimensions (Note that for a given neighbourhood size (proportional to $\rho\sigma$, or $N\sqrt{m}$ in one dimension, $\rho\sigma^2$ or Nm in two), random drift causes a much larger variance in allele frequencies in one dimension than in two, by a factor $\sim 1/\sqrt{\mu}$. This is because fluctuations are dissipated less effectively by dispersal in one dimension than in two. However, throughout this paper results are scaled relative to the variance in the absence of the barrier, which is inversely proportional to $4\rho\sigma\sqrt{2\mu}$ in one spatial dimension, and to $4\pi\rho\sigma^2$ in two).

The overall form of the results is summarized in Fig. 1*a* and 1*b* for one and two dimensions, respectively. Fig. 2 shows how the probability of identity between nearby genes (and hence the variance of allele frequency fluctuations) increases near a barrier ($B = 2L$), where gene flow is less effective. The effect is much more pronounced and more localized in two dimensions than in one: for these parameters, identity increases by 50 % near the barrier in both one and two dimensions, but this increase occurs much closer to the barrier in two dimensions than in one (compare upper curve and lower curves). In this and the following figures, the continuous approximation (curves) agrees closely with exact results for the stepping stone model (dots).

(*a*)

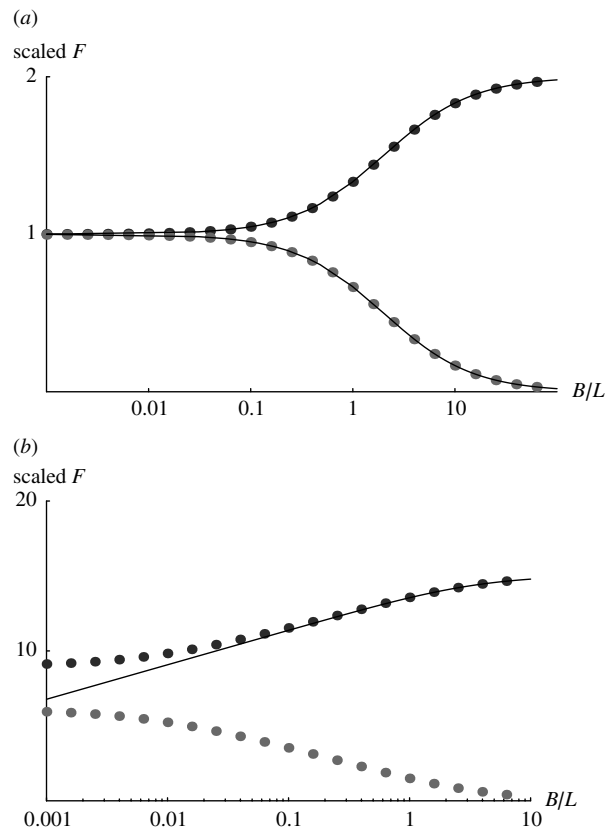scaled *F*

(*b*)

scaled *F*

Fig. 3. The probability of identity between nearby genes on the same side of a barrier, and on different sides, as a function of scaled barrier strength, B/L Scaling is as in Figs 1, 2. (*a*) One-dimensional population. (*b*) Two-dimensional population. Dots show exact results for the stepping stone model with $m = 0\cdot4$, $\mu = 10^{-6}$ and hence $L = 447$.

Fig. 3 shows how barrier strength influences the probability of identity for two genes near the barrier, and either on the same side (upper curves) or on different sides (lower curves). Fig. 3*a* show that in one dimension, only barriers with strength comparable to the characteristic scale, $B \sim L = \sigma/\sqrt{2\mu}$, have a significant effect. In contrast, weak barriers have much more influence in two dimensions (Fig. 3*b*, left). In Fig. 3*b*, the upper curve shows the approximation of equation (7*b*) for the identity between nearby genes that are both at the barrier. This applies only for $B \sim L$; for weaker barriers, this approximation underestimates the effect of the barrier.

We can understand the qualitative difference between one and two dimensions by developing simple analytical formulae for the probability of identity between nearby genes. First, consider one dimension. From equation (2):

$$F(\infty, \infty) = \frac{1}{4\rho\sigma\sqrt{2\mu}}$$

for genes far from the barrier, (8a)

$$F(0+,0+) = \frac{1}{4\rho\sigma\sqrt{2\mu}}\frac{2(L+B)}{(2L+B)}$$

for genes near the barrier, on the same side, (8b)

$$F(0-,0+) = \frac{1}{4\rho\sigma\sqrt{2\mu}}\frac{2L}{(2L+B)}$$

for genes near the barrier, on opposite sides. (8c)

A barrier much stronger than the characteristic scale ($B \gg L$) will increase the probability of identity near the barrier by twofold (equation 8b).

In two dimensions, integration of equation (7a) gives:

$$F(\infty,y,\infty,y) = \frac{\log(L/\kappa)}{4\pi\rho\sigma^2}$$

for genes far from the barrier, (9a)

$$F(0+,y,0+,y) = \frac{1}{4\pi\rho\sigma^2}$$
$$\times \left(2\log(L/\kappa) - \frac{2L}{\sqrt{B^2-4L^2}}\arccos\left(\frac{2L}{B}\right) + C[m]\right)$$

for genes on the same side, (9b)

$$F(0-,y,0+,y) = \frac{1}{4\pi\rho\sigma^2}\frac{2L}{\sqrt{B^2-4L^2}}\arccos\left(\frac{2L}{B}\right)$$

for genes just separated by the barrier. (9c)

When $B < 2L$, $\arccos(2L/B)$ is replaced by $\mathrm{arccosh}$ $(2L/B)$, and $\sqrt{B^2-4L^2}$ by $\sqrt{4L^2-B^2}$. As in the one-dimensional case, the probability of identity approximately doubles near an impenetrable barrier: in the limit of large B, the second term in equation (9b) becomes negligible, and $C[m]$ will be small relative to the first term when mutation rates are low and m is not close to 0·5. (Note that in two dimensions, $m = 0·5$ implies that *all* individuals migrate in each generation – a pathological case.) However, much weaker barriers can cause a substantial increase. When mutation is weak, so that $L \gg \kappa$, the first term in equation (9b) is much larger than the second, even for small B, and so F almost doubles. This can be understood by seeing it as a mixture of contributions from different spatial frequencies (Y). The effect of a barrier on the component with frequency Y depends on its strength relative to the 'effective scale', $L^* = L/\sqrt{1+L^2Y^2}$, which is small for high frequencies. Thus, even a weak barrier can prevent the propagation of high-frequency components. Since these dominate in two dimensions, barriers much weaker than the characteristic scale can have significant effects.

Fig. 4 shows the (scaled) probability of identity between a gene adjacent to the barrier, and a gene at
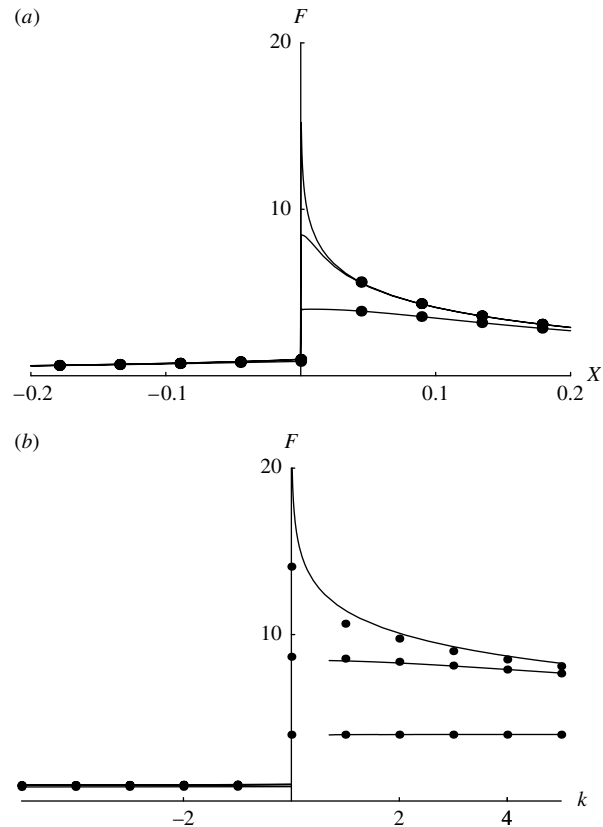

Fig. 4. (*a*) The probability of identity between a gene just next to the barrier, and a gene at $X = x/l$, for a barrier of strength $2L$ for two dimensions. The upper curve is for genes at the same transverse position ($y = y'$); the lower curves are for $y' - y = 0·01L$, $0·1L$. Scaling and simulation results are as in previous figures. (*b*) Detail for genes near the barrier. The upper curve shows the continuum approximation (equation 2a). This is accurate even for adjacent demes (upper dot, $k = 1$), but diverges when the genes coincide (i.e. both next to the barrier; $k = 0$).

x, in two dimensions; there is a barrier of strength $B = 2L$ at $x = 0$. The upper curve is for points opposite each other ($y = y'$), and the lower curves are for displacements $\Delta y = 0·01L$. $0·1L$ F falls away rapidly with distance, reflecting the dominant contribution of local fluctuations. As in Figs 2 and 3, there is good agreement with the stepping stone model.

The basic solution of equation (2) has the interesting property that the decrease in identity between genes on either side of the barrier is equal to the increase in identity between genes on the same side of the barrier, provided that $|x' - x|$ equals $|x + x'|$ in the two cases. This is reflected in the symmetry of the two curves in Fig. 3*a* and 3*b*. Another interesting symmetry is that the identity between two genes separated by a barrier is independent of where the barrier lies on the axis between them. In two dimensions, there is only a weak dependence on the angle of the barrier (Fig. 5).
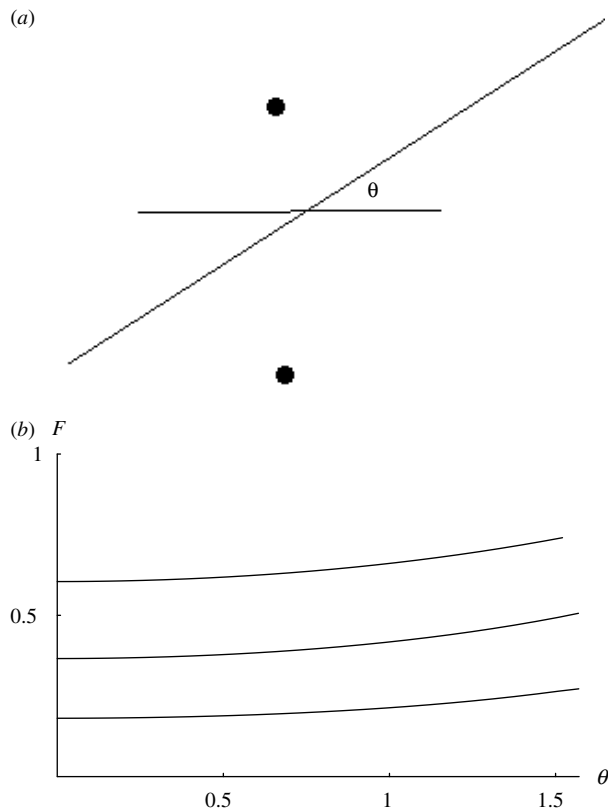
(a)

(b) *F*

Fig. 5. The identity between two genes separated by a barrier is independent of the location of the barrier, and depends only weakly on the orientation of the barrier. The identity does not depend on where the barrier lies along the axis connecting the genes (i.e. in the vertical direction on (a)). (b) In two dimensions, the identity increases somewhat with the angle of the barrier (θ in (a)); this is because when the barrier is close to one or other gene, local fluctuations are amplified. The barrier strength is $B = 2L$; the three curves are for a separation $0.2L$, $0.5L$, L (top to bottom).

5. Discussion

This paper presents simple formulae for the effect of a barrier on random allele frequency fluctuations in one- and two-dimensional populations. A key result is that the effect of a barrier differs qualitatively between one and two dimensions. In one dimension, fluctuations are correlated over the characteristic scale, $L = \sigma/\sqrt{2\mu}$, and are only significantly affected by strong barriers ($B \approx L$). In contrast, local fluctuations dominate spatial patterns in two dimensions and can be impeded by relatively weak barriers ($B \approx \sigma$). This makes it possible to use spatial patterns of allele frequencies to detect weak barriers in two-dimensional populations. Moreover, because the effect of a barrier depends primarily on short-lived, localized fluctuations, inferences are more robust to historical effects and to weak selection than would be the case in a linear habitat.

Since Wright (1931) first emphasized the importance of population structure in evolution, much effort

has been expended in measurements of spatial patterns of genetic variation (Slatkin, 1987). Quantitative analysis of such data has mainly consisted of calculation of some measure of allele frequency variation, such as F_{ST} or Slatkin's (1985) p_0, the frequency of private alleles. This is then used to infer the number of migrants between demes (Nm), or equivalently, the, neighbourhood size in a continuous population ($Nb = 4\pi\rho\sigma^2$). However, spatial patterns contain much more information about population structure than this: in particular, the scale over which relationships decline (Wright, 1943; Malecot, 1948; Epperson, 2003), and the change in relationship across a localized barrier to gene flow.

Barton & Gale (1993, fig. 9) show how the pattern of allele frequencies across a chromosomal hybrid zone in the alpine grasshopper, *Podisma pedestris*, gives an estimate of its strength as a barrier, which agrees with independent estimates based on the chromosomal cline. A maximum-likelihood algorithm for estimating barrier strength is implemented within the package. Analyse (http://www.biology.ed.ac.uk/research/institutes/evolution/software/). Because a barrier mainly affects local fluctuations in two dimensions, its effect equilibrates rapidly. Therefore, allele frequency fluctuations should give a robust estimate of the rate of gene flow in the immediate past, which will not be sensitive to longer-term changes in population structure. This example shows that it is feasible to use readily collected data on molecular polymorphism to make quantitative estimates of the rate of gene flow across physical or genetic barriers. As a rough guide, samples that are large enough to comfortably detect significant variation in allele frequency (as measured by F_{ST}) should also be able to detect the reduced correlation in allele frequency that is caused by a localized barrier. Since the focus is on recent ancestry, samples should be spread over a limited region, spanning the barrier and covering some tens of dispersal ranges.

This paper has followed the classical treatment of genetic variation in structured populations, by working in terms of the identity in allelic state. However, if this identity is regarded as a function of the mutation rate, then it gives a generating function for the distribution of pairwise coalescence times (Slatkin, 1991; Barton & Wilson, 1995). Charlesworth *et al.* (2003) and Wilkins (2004) use this relation to derive the long-term distribution of coalescence times from Maruyama's (1972) results. Wilkins (2004) gives analytical approximations for the distribution of coalescence times in finite two-dimensional populations, and shows how these relate to classical dimensional results on identity in state and allele frequency fluctuations. This paper treats an infinite population, in which the mean coalescence time is infinite. Therefore, it is not helpful to derive the moments of

the distribution of coalescence time by differentiating the generating function near $\mu = 0$. It is, however, possible to find the distribution of coalescence times by numerically inverting the generating function.

It would be interesting to investigate the effects of barriers to gene flow on genealogical structure in more detail. However, it seems unlikely that genealogies at single loci (for example, mtDNA) would give much useful information. This is because in two dimensions, genealogical structure depends on the deep history of the population, and is in any case highly variable (Irwin, 2002; Wilkins, 2004). In contrast, the effects of a localized barrier to gene flow in a two-dimensional population are mainly on recent ancestry, and so can best be detected from variations in the frequency of multiple marker alleles.

Appendix. Convergence of the discrete and continuous solutions

A continuous approximation for two dimensions was derived by splitting fluctuations into a spectrum of frequencies along the y-axis. Fluctuations of different frequencies are uncorrelated, and have a covariance $\tilde{F}(x, Y, x', Y')$ of the same form as the one-dimensional solution for F (equation 4). In this Appendix, I use the same method to derive the exact solution for F in a discrete stepping stone model, and show that it converges to the continuous approximation over all but very small scales.

I begin by deriving F in the absence of a barrier; this is an alternative derivation of the classical results of Malecot (1948) and Kimura & Weiss (1964). First, consider a linear chain of demes, spaced ε apart. After migration and mutation, the probability of identity between demes j and j' is

$$F_{jj'}^* = (1-\mu)^2 \big((m/2)^2 [F_{(j+1)(j'+1)} + F_{(j+1)(j'-1)} + F_{(j-1)(j'+1)} + F_{(j-1)(j'-1)}] + (m/2)(1-m)$$
$$\times [F_{(j+1)j'} + F_{(j-1)j'} + F_{j(j'+1)} + F_{j(j'-1)}] + (1-m)^2 F_{jj'} \big). \tag{A1a}$$

After random sampling of $2N$ genes, F must equal its original equilibrium value:

$$F_{jj'} = \frac{\delta_{jj'}(1-F_{jj}^*)}{2N} + F_{jj'}^*. \tag{A1b}$$

Subtracting $F_{jj}/2N$ from (A1b), and multiplying by $(2N/(2N-1))$ gives

$$F_{jj'} = \frac{(1-F_{jj})\delta_{jj'}}{(2N-1)} + F_{jj'}^*. \tag{A1c}$$

The solution to these equations has the form

$$F_{jj'} = (1-F_{jj})(A_1 \lambda_1^{|j-j'|} + A_2 \lambda_2^{|j-j'|}) \quad (-1 < \lambda, \lambda^* < 1). \tag{A2}$$

λ_1, λ_2 are the eigenvalues of the recurrence equation (A1a):

$$\lambda_1 = z_1 - \sqrt{z_1^2 - 1}$$
$$\text{where } z_1 = \frac{1}{m}\left(\frac{1}{(1-\mu)} - (1-m)\right) = \frac{1}{2}\left(\lambda_1 + \frac{1}{\lambda_1}\right) \tag{A3a}$$

$$\lambda_2 = z_2 + \sqrt{z_2^2 - 1}$$
$$\text{where } z_2 = \frac{1}{m}\left(\frac{-1}{1-\mu} - (1-m)\right) = \frac{1}{2}\left(\lambda_2 + \frac{1}{\lambda_2}\right) \tag{A3b}$$

When $\mu \ll 1$, $\lambda_1 = 1 - \sqrt{2\mu/m} + \mathrm{O}(\mu)$, $\lambda_2 = [2\sqrt{1-m} - (2-m)]/m + \mathrm{O}(\mu)$. The coefficients A_1, A_2 are set by considering the equations for $F_{(j+1)j}$ and F_{jj}. This leads to

$$A_1 = \frac{(1-\mu)}{2(2N-1)\sqrt{\mu(\mu+2m(1-m))}},$$
$$A_2 = \frac{(1-\mu)}{2(2N-1)\sqrt{1+2(1-m)(1-\mu)}},$$
$$F_{jj'} = \frac{A_1 \lambda_1^{|j-j'|} + A_2 \lambda_2^{|j-j'|}}{1+A_1+A_2}. \tag{A4}$$

As μ tends to zero $F_{jj'}$ tends to $\exp(-|x-x'|/L)/(1+2(2N-1)\sqrt{2m\mu})$, where $L = \varepsilon\sqrt{m/2\mu}$, and x is defined as $j\varepsilon$. This tends to the continuous approximation of equation (2a) (with $B = 0$) as N becomes large, and hence F_{jj} small. (Then, the terms in λ_2 decay more rapidly than those in λ_1, and in any case, $A_2/A_1 \approx \sqrt{\mu}$. $\lambda^{|j-j'|}$ tends to $\exp(-|x-x'|/L)$, where $L = \varepsilon\sqrt{m/2\mu}$, and A_1 tends to $1/4N = \sqrt{2m\mu}$.) These results are consistent with Kimura & Weiss (1964).

Now, consider the solution along the x-axis for fluctuations of frequency Y. In continuous space, the Fourier transform of a function $g(y)$ is defined as

$$\tilde{g}(Y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(y)e^{iyY}dy, \tag{A5}$$

where Y is the spatial frequency along the y-axis. The definition of the discrete transform is the same, but with the integral $\int f(y)dy$ being replaced by $\sum \varepsilon f k$, and deme, k being at position $y = k\varepsilon$. $\tilde{g}(Y)$ is defined for $-\pi/\varepsilon < Y < \pi/\varepsilon$: a discrete grid cannot support high frequencies, with $|Y| > \pi/\varepsilon$.

The recursion for $\tilde{F}_{jj'}(Y, Y')$ is

$$\tilde{F}_{jj'}^* = (1-\mu)^2 \times \big((m/2)^2 [\tilde{F}_{(j+1)(j'+1)} + \tilde{F}_{(j+1)(j'-1)} + \tilde{F}_{(j-1)(j'+1)} + \tilde{F}_{(j-1)(j'-1)}] + (m/2)$$
$$\times (1-m-2m\sin^2(\varepsilon Y/2))[\tilde{F}_{(j+1)j'} + \tilde{F}_{(j-1)j'} + \tilde{F}_{j(j'+1)} + \tilde{F}_{j(j'-1)}] + (1-m-2m\sin^2(\varepsilon Y/2))^2 \tilde{F}_{jj'} \big), \tag{A6a}$$

$$\tilde{F}_{jj'} = \frac{(1-F_{jkjk})\varepsilon\delta_{jj'}\delta(Y+Y')}{(2N-1)} + \tilde{F}_{jj'}^{*}. \qquad (A6b)$$

Since the solution must be proportional to $\delta(Y+Y')$, Y' is replaced by Y in (A6a). Equations (A6) have the same form as equation (A1): m is replaced by $m/(1-2m\sin^2(\varepsilon Y/2))$, and $(1-\mu)$ by $(1-\mu)(1-2m\sin^2(\varepsilon Y/2))$. The exact solution to the two-dimensional case is therefore given by making these replacements in equations (A3), (A4), and taking the inverse Fourier transform. This leads to equation (7a).

As in the one-dimensional case, the Fourier transform given by these replacements converges to the continuous approximation (the integrand of (5) with $B=0$) as μ tends to zero, and N becomes large. However, for given μ, the exact and approximate transforms diverge at high frequencies: the discrepancy is of order $1/Y\varepsilon$ for large Y. In two dimensions, these high frequencies do not affect F for genes separated by distances $r \gg \varepsilon$, but do make a substantial contribution to F for nearby genes: for coincident genes, the continuous approximation diverges.

In the presence of a barrier, the one-dimensional solution takes a form analogous to equation (2). The barrier is between demes 0 and 1, and reduces gene flow by a factor γ. This is equivalent to a barrier of strength $B=(1/\gamma-1)\varepsilon$ (Nagylaki, 1976). Because F_{jkjk} now increases near the barrier, rather than being uniform, an explicit solution for small Nm and large F would be complex. (However, one could set bounds on the solution by using a uniform factor $(1-F_{\min})$ or $(1-F_{\max})$, $F_{\min}$ and $F_{\max}$ being the minimum and maximum of F_{jkjk}.) For $F \ll 1$:

$$F_{jj'} = A_1\lambda_1^{|j-j'|} + A_2\lambda_2^{|j-j'|} + (A_1-D_1)\lambda_1^{|j+j'|-1}$$
$$+ (A_2-D_2)\lambda_2^{|j+j'|-1} \quad (j>0, j'>0), \qquad (A7a)$$

$$F_{jj'} = A_1\lambda_1^{|j-j'|} + A_2\lambda_2^{|j-j'|} + (A_1-D_1)\lambda_1^{|j+j'|+1}$$
$$+ (A_2-D_2)\lambda_2^{|j+j'|+1|} \quad (j\leqslant 0,\ j'\leqslant 0), \qquad (A7b)$$

$$F_{jj'} = D_1\lambda_1^{|j-j'|} + D_2\lambda_2^{|j-j'|}$$
$$(j\leqslant 0, j'>0 \text{ or } j>0,\ j'\leqslant 0). \qquad (A7c)$$

The terms $|j+j'|-1$ in (A7a) ensure that $F_{00}=F_{11}$. A_1 and A_2 have the same values as before, while D_1, D_2 are determined by the barrier:

$$D_i = \frac{A_i\varepsilon(1+\lambda_i)}{\varepsilon(1+\lambda_i)+B(1-\lambda_i)}. \qquad (A8)$$

Consider the limit where μ tends to zero and $B/L = (B/\varepsilon)\sqrt{2\mu/m}$ remains finite. Then, $(1-\lambda_1)$ tends to $\sqrt{2\mu/m}$, while $(1-\lambda_2)$ remains between 1 and 1·17 for m between 0 and 0·5 (equation A3). Then, terms involving λ_2 become negligible relative to those

involving λ_1, and the solution converges to that of equation (2).

The solution extends to two dimensions in the same way as in the uniform case. The solution for $\tilde{F}_{jj'}(Y,Y')$ is obtained by replacing m by $m/(1-2m\sin^2(\varepsilon Y/2))$, and $(1-\mu)$ by $(1-\mu)(1-2m\sin^2(\varepsilon Y/2))$ in Eqs A3, A4, A7, A8. Over low frequencies ($Y \ll 1/\varepsilon$), $2m\sin^2(\varepsilon Y/2) \approx m\varepsilon^2 Y^2/2 = \mu(Y/L)^2$. Hence, $\tilde{F}$ converges to the continuous approximation given by the integrand of equation (5a) for $Y \ll 1/\varepsilon$. As Y approaches its upper limit of π/ε, the exact value of $\tilde{F}$ falls below the continuous approximation, ensuring that F_{jkjk} remains finite.

Finally, we must find the function, $C[m]$, which describes the effects of fluctuations near a barrier of strength $B \sim L$. Recall that F^* is the exact solution with no barrier. To justify equation (7b), we must show that the *difference* $(F-2F^*)$ converges to the continuous approximation over all scales with a suitable choice of the constant C. This is assured, since the Fourier transforms $\tilde{F}$ and $\tilde{F}^*$ both decrease as $1/Y$ for large Y. We can therefore find some α such that $(\tilde{F}-\alpha\tilde{F}^*)$ decreases faster than $1/Y$ (in this case, as $1/Y^2$), so that the high-frequency region makes a negligible contribution. This allows us to replace $(F-\alpha F^*)$ by the continuous approximation, and calculate the residual effect of local population structure as a function $C[m]$ that is independent of μ, B. Some tedious integration gives

$$C[m] = \frac{\pi}{2} + \mathrm{Re}\left[\frac{2(1-2m)}{1-m}\Pi\left(\frac{m}{1-m}; \arcsin\left(\frac{1-m}{m}\right);\right.\right.$$
$$\left.\left.\frac{m^2}{(1-m)^2}\right) - \frac{3m}{\sqrt{1-2m}}K\left(-\frac{m^2}{1-2m}\right)\right] \qquad (A9)$$

where K, Π are elliptic integrals.

References

Ayatia, B. P., Dupont, T. F. & Nagylaki, T. (1999). The influence of spatial inhomogeneities on neutral models of geographical variation. IV. Discontinuities in the population density and migration rate. *Theoretical Population Biology* **56**, 337–347.

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

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. & Rouhani, S. (1991). The probability of fixation of a new karyotype in a subdivided population. *Evolution* **45**, 499–517.

Barton, N. H. & Wilson, I. (1995). Genealogies and geography. In *New Uses for Phylogenies* (ed. P. H. Harvey, A. J. Leigh Brown & J. Maynard Smith). Oxford: Oxford University Press.

Barton, N. H., Depaulis, F. & Etheridge, A. M. (2002). Neutral evolution in spatially continuous populations. *Theoretical Population Biology* **61**, 31–48.

Charlesworth, B., Charlesworth, D. & Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology and Systematics* **34**, 99–125.

Epperson, B. K. (2003). *Geographical Genetics*. Princeton, NJ: Princeton University Press.

Felsenstein, J. (1986). Population differences in quantitative characters and gene frequencies: a comment on papers by Lewontin and Rogers. *American Naturalist* **127**, 731–732.

Fleming, W. H. & Su, C. H. (1974). Some one dimensional migration models in population genetic theory. *Theoretical Population Biology* **5**, 431–449.

Irwin, D. E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution* **56**, 2383–2394.

Kimura, M. & Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.

Malecot, G. (1948). *Les Mathematiques de l'Heredite*. Paris: Masson et Cie. (English translation, 1969, W. H. Freeman, San Francisco.)

Maruyama, T. (1972). Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. *Genetics* **70**, 639–651.

Nagylaki, T. (1976). Clines with variable migration. *Genetics* **83**, 867–886.

Nagylaki, T. (1978). A diffusion model for geographically structured populations. *Journal of Mathematical Biology* **6**, 375–382.

Nagylaki, T. (1988). The influence of spatial inhomogeneities on neutral models of geographical variation. I. Formulation. *Theoretical Population Biology* **33**, 291–310.

Nagylaki, T. & Barcilon, V. (1988). The influence of spatial inhomogeneities on neutral models of geographic variation. II. The semi-infinite linear habitat. *Theoretical Population Biology* **33**, 311–343.

Nagylaki, T., Keenan, P. T. & Dupont, T. F. (1993). The influence of spatial inhomogeneities on neutral models of geographical variation. III. Migration across a geographical barrier. *Theoretical Population Biology* **43**, 217–249.

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

Pulliam, R. (1988). Sources, sinks and population regulation. *American Naturalist* **132**, 652–661.

Rieseberg, L. H., Church, S. A. & Morjan, C. L. (2003). Integration of populations and differentiation of species. *New Phytologist* **161**, 59–69.

Slatkin, M. (1973). Gene flow and selection in a cline. *Genetics* **75**, 733–756.

Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution* **39**, 53–65.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science* **236**, 787–792.

Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research* **58**, 167–175.

Slatkin, M. & Barton, N. H. (1990). A comparison of three methods for estimating average levels of gene flow. *Evolution* **43**, 1349–1368.

Wilkins, J. F. (2004). A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168**, 2227–2244.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.