Genetic hitchhiking in spatially extended populations

N. H. Barton[a,1], A. M. Etheridge[b,2], J. Kelleher[c,3], A. Véber[d,4]

[a] Institute of Science and Technology
Am Campus I
A-3400 Klosterneuberg
Austria
[b] Department of Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
UK
[c] Institute of Evolutionary Biology
University of Edinburgh
Kings Buildings
West Mains Road
Edinburgh EH9 3JT
UK
[d] Centre de Mathématiques Appliquées
École Polytechnique
Route de Saclay
91128 Palaiseau Cedex
France

Abstract

When a mutation with selective advantage s spreads through a panmictic population, it may cause two lineages at a linked locus to coalesce; the probability of coalescence is $\exp(-2rT)$, where $T \sim \log(2Ns)/s$ is the time to fixation, N is the number of haploid individuals, and r is the recombination rate. Population structure delays fixation, and so weakens the effect of a selective sweep. However, favourable alleles spread through a spa-

Email addresses: n.barton@ist.ac.at (N. H. Barton), etheridg@stats.ox.ac.uk (A. M. Etheridge), jerome.kelleher@ed.ac.uk (J. Kelleher), amandine.veber@cmap.polytechnique.fr (A. Véber)

[1] Work supported by European Research Council grant 250152

[2] Work supported in part by EPSRC grant EP/I01361X/1

[3] Work supported by EPSRC grant EP/I013091/1

[4] Work supported by the chaire Modélisation Mathématique et Biodiversité of Veolia Environment - École Polytechnique - Museum National d'Histoire Naturelle - Fondation X and by the ANR project MANEGE (ANR-09-BLAN-0215)

tially continuous population behind a narrow wavefront; ancestral lineages are confined at the tip of this front, and so coalesce rapidly. In extremely dense populations, coalescence is dominated by rare fluctuations ahead of the front. However, we show that for moderate densities, a simple quasi-deterministic approximation applies: the rate of coalescence within the front is $\lambda \sim 2g(\eta)/(\rho\ell)$, where ρ is the population density and $\ell = \sigma\sqrt{2/s}$ is the characteristic scale of the wavefront; $g(\eta)$ depends only on the strength of random drift, $\eta = \rho\sigma\sqrt{s/2}$. The net effect of a sweep on coalescence also depends crucially on whether two lineages are ever both within the wavefront at the same time: even in the extreme case when coalescence within the front is instantaneous, the net rate of coalescence may be lower than in a single panmictic population. Sweeps can also have a substantial impact on the rate of gene flow. A single lineage will jump to a new location when it is hit by a sweep, with mean square displacement $\sigma_{\text{eff}}^2/\sigma^2 = (8/3)(L/\ell)(\Lambda/R)$; this can be substantial if the species' range, L, is large, even if the species-wide rate of sweeps per map length, Λ/R, is small. This effect is half as strong in two dimensions. In contrast, the rate of coalescence between lineages, at random locations in space and on the genetic map, is proportional to $(c/L)(\Lambda/R)$, where c is the wavespeed: thus, on average, one-dimensional structure is likely to reduce coalescence due to sweeps, relative to panmixis. In two dimensions, genes must move along the front before they can coalesce; this process is rapid, being dominated by rare fluctuations. This leads to a dramatically higher rate of coalescence within the wavefront than if lineages simply diffused along the front. Nevertheless, the net rate of coalescence due to a sweep through a two-dimensional population is likely to be lower than it would be with panmixis.

Keywords: spatial structure, hitchhiking

1. Introduction

For many years, population geneticists have been trying to distinguish the signature of natural selection from other forces of evolution. If a single mutation is fixed by selection in a hard sweep, then diversity will be eliminated at the target of selection, but can also be substantially reduced at sufficiently tightly linked loci, as the alleles that were fortunate enough to be carried by the chromosome on which the favourable mutation first arose 'hitch a lift' to achieve higher frequency in the population. The first attempt to quantify this genetic hitchhiking was due to Maynard Smith and Haigh (1974), and their work has been refined and extended by many authors since (see e.g. Stephan et al. (1992); Barton (1998); Durrett and Schweinsberg (2004); Schweinsberg and Durrett (2005); Etheridge et al. (2006)). Essentially all this work deals with panmictic populations. However, most populations have spatial structure, with many evolving in two dimensional spatial continua. In this setting, if a favourable mutation fixes, it will sweep through the population in an expanding wave. Very little is known about the degree of hitchhiking as such a sweep passes through the population and our intuition from the panmictic case may be of little help. Slatkin and Wiehe (1998) conisder hitchhiking in a subdivided population under the assumption that migration rates are small. Kim and Maruki (2011) consider higher migration rates, but their analysis is restricted to the case when the population is subdivided into just two demes. Barton (2000) derives the increase of a linked neutral allele due to a sweep through a continuous two-dimensional habitat, using a deterministic analysis forwards in time. Here, we focus on the effect of a sweep on coalescence, backwards in time, and in a spatially continuous habitat.

It is not at all obvious whether the net effect of a sweep on the genetic diversity at linked neutral loci will be stronger or weaker in a spatially structured population. On the one hand fixation will take much longer, extending the timescale over which recombination can 'free' the hitchhiking allele, but on the other hand, local founder events at the wavefront may greatly increase genetic drift (Klopfstein et al., 2006; Excoffier et al., 2009; Hallatschek et al., 2007; Hallatschek and Nelson, 2008, 2010).

In order to quantify the strength of hitchhiking in a spatially structured population, we focus on the net rate of coalescence of neutral genes due to sweeps at linked loci passing through the population. We leave other questions, such as "What is the spatial signature of genetic hitchhiking?" for future work.

Our analysis rests on understanding the wave of advance of a favoured allele. Fisher (1937) described this through a deterministic diffusion equation (1) which exhibits travelling wave solutions. However, the speed of the Fisher wave is determined by its behaviour in regions where the frequency of the favoured allele is extremely low and, consequently, it is very sensitive to stochastic fluctuations. In recent years there has been considerable progress in understanding the coupling between such fluctuations and the progress of the 'bulk' of the wave (Brunet and Derrida (1997, 2001); van Saarloos (2003); Brunet et al. (2006); Hallatschek and Nelson (2008); Mueller et al. (2011); Berestycki et al. (2012)). Much of this work is concerned with the spread of a new species into an empty habitat, but the mathematical models apply equally to the spread of a selectively favoured allele through a stable population.

Analytic results for models that mimic noisy Fisher waves are restricted to one spatial dimension and they deal almost exclusively with asymptotic behaviour as the population density tends to infinity. Three ramifications of the presence of small amounts of genetic drift have been explored: first, drift slows down the rate of advance of the wave by a factor proportional to the dimensionless quantity $1/(\log(\rho\sigma\sqrt{s/2}))^2$ where ρ is population density, σ^2 is the rate of diffusion and s is the selection coefficient (Brunet et al., 2006; Mueller et al., 2011); second, the resulting cline shape (that is the shape of the wavefront) is well-approximated by a truncated Fisher wave (described in more detail in §2.1), Brunet and Derrida (1997); Mueller et al. (2011); and third, the genealogy of a sample from the wavefront will be dominated by 'founder effects' (resulting from fluctuations in the wavefront) and in an appropriate timescale is approximated not by a Kingman coalescent, but instead by the so-called Bolthausen-Sznitman coalescent in which, in particular, multiple lineages merge in a single event (Brunet et al., 2006; Berestycki et al., 2012).

Our goal is to apply this body of theory to understand the effect of a selective sweep on the genealogy of a sample from a spatially distributed population. Key to this is to investigate the accuracy of the (asymptotic) predictions of the theory for biologically realistic population densities. What we shall see in one dimension is that although the prediction for the reduction in the wavespeed is remarkably robust, by contrast, for realistic population densities, fluctuations are much less important for the genealogy of a sample. Instead a simple deterministic approximation (which nonetheless draws on the asymptotic theory) provides a surprisingly good approximation: a substantial fraction of coalescence occurs within the 'average' wave rather than in extreme events.

Armed with an understanding of the genealogy of a sample from the selected locus, we can investigate the strength of hitchhiking. Lineages can escape the sweep through recombination into the less fit genetic background. The rate of 'successful' recombination events will depend on the availability of individuals of the unfavoured type with whom to recombine. Since lineages ancestral to the selected locus typically lie near the 'front' of the wave, and, at least if the local population density is not *too* small, the favoured allele comprises a small proportion of the population there, we suppose that all recombinations will be successful. (For very low population densities, this approximation is less accurate and instead we should consider an 'effective' recombination rate which takes into account the allele frequencies in a neighbourhood of a typical ancestral lineage.) Thus, we can find the chance that two genes sampled from behind a sweep will coalesce within it, rather than recombining away. In a finite range, we can also find the chance that lineages will trace all the way back to the original favourable mutation. For one dimension, we summarise these arguments by finding the net rate of coalescence between randomly sampled genes, and averaging over a random rate of sweeps. We compare this net rate of coalescence to that in the absence of sweeps, and to that due to classic genetic sweeps within a panmictic population.

In two spatial dimensions there is essentially no rigorous theory on which to draw. It is still the case that at the selected locus, genes sampled from behind the advancing wavefront trace back to ancestors that lived at the front of the wave. Even if the population density is high, such ancestors were in effect confined to an extremely small population. This greatly increases the rate of random genetic drift. In models of species expanding into empty habitat, this has been shown to lead to 'surfing', in which neutral, and even deleterious, alleles that are lucky enough to arise at the tip of the wave can be carried to high frequency in the population (Klopfstein et al., 2006; Excoffier et al., 2009). This effect will be reinforced by fluctuations at the front: occasionally an individual moves far ahead of the rest of the population, where the only competition that it experiences is from its own family. Such individuals leave an exceptionally large number of descendants. In particular, descendants of this single individual can comprise a significant piece of the wavefront at a later time.

In two dimensions, we consider only the simplest case of a linear wavefront. In that setting, it seems natural to assume that ancestral lineages become trapped in a narrow (effectively one-dimensional) front, along which they diffuse. Our one-dimensional arguments are then easily adapted to give explicit approximations for the rate of coalescence. However, we show (by simulation) that this gives a very poor approximation. In contrast to one dimension, even when neighbourhood size is small, coalescence is dominated by the effect of fluctuations in the wavefront. This leads to a drastically stronger effect of hitchhiking in two dimensions, and makes it much more likely that genetic diversity is shaped by 'genetic draft' in spatially extended populations.

For ease of reference, definitions of the important parameters in what follows are collected together in Table 1.

Table 1: Frequently used notation

s	selection coefficient
σ^2	dispersal rate
r	recombination rate
ρ	population density
p	allele frequency
x	distance in direction of wave movement
y	distance transverse to the wave
$\eta = \sigma\rho\sqrt{\frac{s}{2}}$	dimensionless parameter
$c_\infty = \sigma\sqrt{2s}$	wave speed at infinite density
$c_\eta < c_\infty$	wave speed at finite density
$\ell = \sigma\sqrt{\frac{2}{s}}$	characteristic spatial scale
$w = \int 4p(1-p)dx$	width of wave
$\lambda = \frac{2}{\rho\ell}g(\eta)$	rate of coalescence within the wavefront
m	migration rate between demes
N	# of haploids in each deme
$f(x)$	distribution of ancestral lineages (Eq. (7))
$g(\eta)$	scaled rate of coalescence in the front (Eq. (B.2)
$h\left(\frac{L\lambda}{c}\right)$	$\sim$ net rate of coalescence (Eq. C.2)
$\phi(t,x)$	density of the time taken for biased Brownian motion from x to hit 0 (Eq. 10))
$\Psi(t,z)$	density of time to coalescence of lineages z apart (Eq. (14))
L	length of species' range
Λ	rate of sweeps per genome, per generation
R	map length of genome

2. Previous work

2.1. One dimensional waves of advance

The classical approach to modelling the spread of alleles through a spatially distributed population dates back to Fisher (1937) and Kolmogorov, Petrovsky & Piscounov (1937). Fisher considers a population evolving in $\mathbb{R}$. Denoting the frequency of the favoured allele at point x and time t by $p(t,x)$, he found

$$\frac{\partial p}{\partial t}(t,x) = \frac{\sigma^2}{2}\frac{\partial^2 p}{\partial x^2}(t,x) + sp(t,x)\left(1 - p(t,x)\right). \tag{1}$$

We shall refer to this equation as the Fisher-KPP equation. It has a whole family of travelling wave solutions. However, if we start from a non-negative initial condition which tends to 1 at $-\infty$ and 0 at $+\infty$ and decays quickly enough in space (for example one in which favoured alleles are initially confined to the negative half line), then the solution to (1) converges to the non-negative travelling wave of the smallest possible velocity, $c_\infty = \sigma\sqrt{2s}$, (Bramson, 1983). If we write the corresponding travelling wave solution as $p(t,x) = p_{c_\infty}(x - c_\infty t)$, then, as Fisher showed, when $p_{c_\infty}(z)$ is small it can be approximated by $\exp(-c_\infty z/\sigma^2)$.

The first obstruction that we must overcome is that equation (1) tacitly assumes an infinite population density at every point in space (hence our notation c_∞) and so we see no coalescence in this picture. This problem is traditionally circumvented by supposing the population to be subdivided into discrete demes and assuming that allele frequencies can be modelled by a stepping stone model with selection and, say, nearest neighbour migration. We assume that the population is subdivided into demes, each containing N haploid individuals. Reproduction within each deme is according to the Wright-Fisher model (with selection). Thus, if the proportion of favoured alleles in deme i at generation t is $p_i(t)$, then each offspring, independently, is of the favoured type with probability $(1+s)p_i(t)/(1+sp_i(t))$. The reproduction step is followed by a migration step in which a proportion m of offspring are exchanged with neighbouring sites. Our simulations will all be based upon this model.

If selection is weak, and the population size within each deme is reasonably large, then it is often mathematically convenient to replace this discrete (individual and generation) model, by a diffusion approximation (Kimura, 1953). In one dimension this takes the form

$$dp_i(t) = \frac{m}{2}\left(p_{i+1}(t) + p_{i-1}(t) - 2p_i(t)\right)dt + sp_i(t)\left(1 - p_i(t)\right)dt + \sqrt{\frac{1}{\rho}p_i(t)\left(1-p_i(t)\right)}dW_i(t), \quad i \in \mathbb{Z}, \tag{2}$$

where $p_i(t)$ once again denotes the frequency of the favoured allele in deme i at time t and $\{W_i(\cdot)\}_{i\in\mathbb{Z}}$ are independent Brownian motions. The quantity ρ represents population density and will correspond to N in our simulations. In one spatial dimension one can perform a diffusive rescaling to obtain a continuum analogue of this system encoded by the stochastic partial differential equation

$$dp(t,x) = \frac{\sigma^2}{2}\frac{\partial^2}{\partial x^2}p(t,x)dt + sp(t,x)\left(1-p(t,x)\right)dt \\ + \sqrt{\frac{1}{\rho}p(t,x)\left(1-p(t,x)\right)}W(dt,dx) \quad (3)$$

where $W(dt,dx)$ is a space-time white noise (Shiga, 1988). The scaling is made explicit (in the neutral case) in Lemma 2.7 of Barton et al. (2002); this spatial diffusion approximates a wide range of models, both discrete and continuous, provided that the distribution of dispersal distance declines fast enough that the motion of a single ancestral lineage rescales to Brownian motion. To mimic the behaviour of (3) we take $N=\rho$ and $m=\sigma^2$ in our simulations.

There is now a considerable body of work exploring the stochastic pde (3) in the case of small noise (corresponding to weak selection and very high local population density). Most of this work deals with the scaled equation

$$d\tilde{p}(t,x) = \frac{\partial^2}{\partial x^2}\tilde{p}(t,x)dt + \tilde{p}(t,x)\left(1-\tilde{p}(t,x)\right)dt \\ + \sqrt{\frac{1}{\eta}\tilde{p}(t,x)\left(1-\tilde{p}(t,x)\right)}W(dt,dx). \quad (4)$$

These results can be translated into the setting of (3) through the transformation

$$p(t,x) = \tilde{p}\left(st, \frac{\sqrt{2s}}{\sigma}x\right)$$

which is valid on setting

$$\eta = \sigma\rho\sqrt{\frac{s}{2}}.$$

The dimensionless quantity η, introduced in Nagylaki (1978), will be at the heart of our analysis: it measures the strength of random drift, relative to selection and dispersal.

There is a travelling wave solution to (3) (Mueller and Sowers, 1995), for which, in contrast to the classical Fisher wave, the region in which $p \notin \{0,1\}$ is bounded. Indeed, even in the case without selection, if we start from an initial condition in which the 'interface' between the two types is bounded, then the region in which $p(t,y) \notin \{0,1\}$ remains bounded for all time (Mueller and Tribe, 1997).

The Fisher wave is what is known as a pulled wave. It is propagating from a stable (saturated) state into an unstable state, and its asymptotic speed is determined through linearisation about the unstable state (where $p = 0$). As a result, the fluctuations (of order $\sqrt{p/\rho}$) induced by the noise in (3) have a surprisingly big effect on the speed of the wave, slowing it down by an amount proportional to $1/(\log \eta)^2$ for large η. (Here, and throughout, we use log to denote the natural logarithm.) This is sometimes referred to as Brunet-Derrida behaviour and has now been established rigorously for a whole raft of related models (see Berestycki et al. (2012); Bérard and Gouéré (2011) for recent reviews) and it is reasonable to expect that the reduction of wavespeed in (2) compared to that in the limit as $\eta \to \infty$ will (to leading order) also be proportional to $1/(\log \eta)^2$. The wavespeed in (3) is just that in (4) multiplied by $\sigma\sqrt{s/2}$. Thus writing c_η for the wavespeed (in any of our models), the theory predicts that

$$c_\eta \approx c_\infty \left(1 - \frac{A}{(\log(\eta))^2}\right) \tag{5}$$

for some constant A. In the case of (3) it has been proved rigorously that $A = \pi^2/2$ (Mueller et al., 2011).

The heuristic arguments of Brunet and Derrida, and indeed the rigorous proof of Brunet-Derrida behaviour of (3), rest on comparing the solution to the travelling wave solution of the deterministic Fisher-KPP equation (1) corresponding to the wavespeed c_η. Since $c_\eta < c_\infty$, this deterministic solution will not remain positive and so we truncate it at the smallest x at which it is zero. We denote the resulting wavefront by $p_{c_\eta}(\cdot)$. We discuss the form of the solution in a little more detail in Appendix A.

The rigorous mathematical results described above apply only in the limit as the population density tends to infinity. However, in spatially distributed populations one might expect neighbourhood size to be small. There are essentially no results available for the resulting 'strong noise' scenario, although Hallatschek and Korolev (2009) suggest that in the strong noise limit, the exponential decay in the front of the wave will be replaced by a power law decay. Here we shall restrict our attention to large, but biologically realistic, deme sizes ($N = 10, \ldots, 10^6$, $m = 0.25$, $s = 0.01, 0.05, 0.1$ in our individual based stepping stone model, corresponding to values of η ranging from approximately 10^{-1} to 10^5). Note that since we assume the diffusion scaling, the value of m in itself should not affect the results; we choose $m = 0.25$ to make simulations most efficient.

2.2. Location of a single ancestral lineage

Suppose now that we sample an individual of the favoured type from our population. We assume that the sweep has passed through as a travelling wave (with a stationary form). The ancestral lineage of our sampled individual will initially move around according to a Brownian motion (or random walk in the stepping stone setting), but (tracing backwards in time) at some point it will be 'caught' by the wave and will start to experience a drift (in the mathematical sense) away from the wavefront and back towards the origin of the sweep. Only then does it have any chance of recombining with an individual of the unfavoured type, and thus escaping the sweep. The 'escape' of lineages from the sweep is illustrated in Fig. 1.

We are interested in ascertaining the movement of an ancestral lineage within the wavefront. If the shape of the wave is deterministic, the drift is easily calculated to be $\sigma^2 p'_{c_\eta}(x)/p_{c_\eta}(x)$ where $p'_{c_\eta}(x) = dp_{c_\eta}/dx$. (To see why this is true, recall that in a stepping stone model on $\mathbb{Z}$ in which the population size of the deme at k is N_k, and with symmetric nearest neighbour migration with rate m, an ancestral lineage migrates from k to $k-1$ at rate $mN_{k-1}/(2N_k)$ and to $k+1$ at rate $mN_{k+1}/(2N_k)$. Applying the diffusive rescaling, and assuming that the population size scales to p_{c_η} one obtains a Brownian motion with the drift given above.) In other words, writing the position of the lineage (relative to the tip of the travelling wave) at time t as X_t and c_η for the speed of the wave,

$$dX_t = -\left(\frac{\sigma^2 p'_{c_\eta}(X_t)}{p_{c_\eta}(X_t)} + c_\eta\right) dt + \sigma dB_t, \tag{6}$$

where $\{B_t\}_{t\geq 0}$ is a one-dimensional Brownian motion. If $\{X_t\}_{t\geq 0}$ has a stationary distribution, with density f say, then

$$f(x) \propto p^2_{c_\eta}(x) \exp\left(2\frac{c_\eta x}{\sigma^2}\right) \tag{7}$$

which, if the right hand side is integrable (and so can be normalised to be a probability density function), provides an explicit expression for f (see, for example, Karlin and Taylor (1981) for the theory of one-dimensional diffusions). Equation (7) is also derived in Hallatschek and Nelson (2008). If the wavefront is given by a deterministic Fisher wave and the lineage is far out in the front, where the frequency of favoured alleles is small, then the drift away from the wavefront reduces to $-\sigma\sqrt{2s}$, exactly sufficient to compensate for the speed of the wave. If a lineage escapes too far from the tip of the wave, then $\sigma^2 p'_{c_\eta}(x)/p_{c_\eta}(x)$ decreases in magnitude, is no

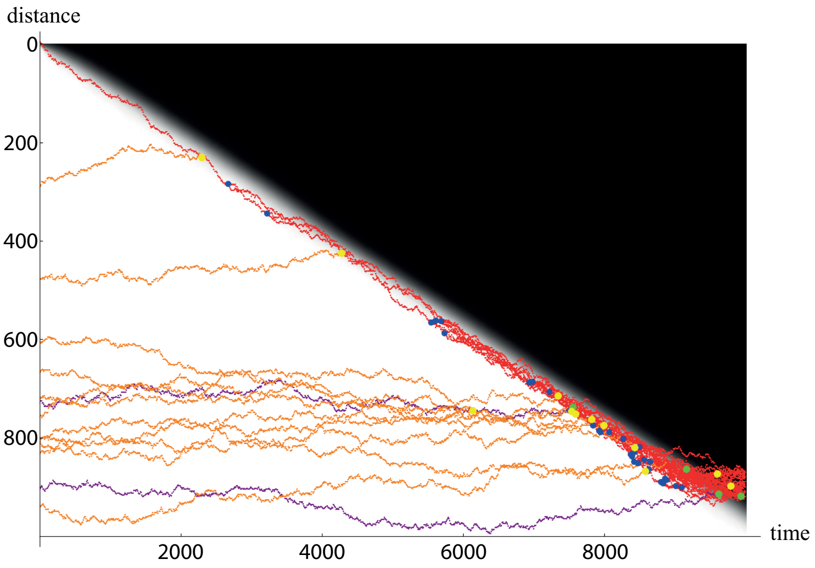

Figure 1: The effect of a sweep through a one-dimensional population on genealogies at linked loci. The picture is the result of a forwards simulation of a stepping stone model on $\mathbb{Z}$ (see the Supplementary Material for details). Initially, an allele with advantage $s = 0.01$ is fixed for $x < 0$, but absent elsewhere (top left). There is nearest-neighbour migration, with a fraction $m = 0.5$ moving per generation; demes contain $N = 10^5$ haploid individuals. The advance rapidly settles to a wave with speed slightly slower than that of the corresponding deterministic wave which is $\sqrt{2ms} = 0.1$ demes per generation, and width $\sim \sqrt{2m/s} = 10$; the grey scale is proportional to $\sqrt{p}$, to make the low frequencies out at the front more visible. At time $t = 10^4$, 5 genomes are sampled from each of 6 locations just behind the wavefront, and 10 demes apart (bottom right); their ancestry at the selected locus is shown in red, and coalescence events are shown as blue dots. The MRCA of the 30 lineages at the selected locus is 7338 generations back (blue dot at upper left). The ancestry of ten marker alleles, spaced $0.01cM$ apart on one side of the selected locus was also traced, backwards through time. Purple dots show lineages where the closest marker has recombined onto the less fit background; orange lines show lineages where the second closest marker, $0.02cM$ away, has recombined onto the ancestral background. Green dots show recombination events that separate the closest marker from the favoured allele, and yellow dots show recombinations that separate ancestral material at the first and second markers. Note that almost all coalescence events are on the selected genealogy, and occur in the wavefront: there are 33 coalescence events on the ancestral selection graph, of which 29 are responsible for the coalescence of the 30 lineages at the selected locus. The remaining four are between various recombinant lineages, but still occur within the wavefront. Similarly, almost all recombinations lead to escape of neutral lineages. There are four recombinations that separate the proximal marker from the selected locus (green dots) of which two allow escape onto the ancestral background (purple lines) and the remaining two (which occur further back from the front) are betweeen parents carrying the fitter allele. Moving along the genome, there are 12 events that separate the two closest marker loci (yellow dots) of which all occur within the front, leading to escape of the second marker (orange lineages). However, one event separates these two markers along a lineage that has already escaped (yellow dot at $t = 5726$).

longer sufficient to compensate the wavespeed, and the wave starts to catch up, driving the lineage back into the front. As $\eta \to \infty$, the width of the region in which the allele frequencies are between any two assigned values is proportional to $\sigma\sqrt{2/s}$ (Fisher, 1937) and so for large s/σ^2, this suggests that ancestral lineages can be thought of as trapped within a narrow front. We shall write

$$\ell = \sigma\sqrt{2/s}$$

for the characteristic lengthscale of the travelling wavefront.

In fact, for the Fisher wave, $\sigma^2 p'_{c_\infty}(x)/p_{c_\infty}(x)$ only becomes close to the asymptotic value $-\sigma\sqrt{2s}$ when $p_{c_\infty}(x)$ is extremely small, and so for a large (but finite) value of η the mechanism described above drives lineages into regions where the classical Fisher wave is no longer a valid approximation for allele frequencies and the correction achieved by replacing p_{c_∞} by p_{c_η} becomes important.

2.3. Genealogies in one dimension

We now turn to the genealogy of a sample from the population. In the discussion above, we have considered only the 'bulk' of the travelling wave and we have ignored the fluctuations in the position of the front. The analyses of Brunet and Derrida (2001); Brunet et al. (2006); Berestycki et al. (2012) of models which are believed to mirror the behaviour of solutions to (3) suggest that, at least for large η, for most of the time the deterministic approximation $p_{c_\eta}(\cdot)$ provides a good approximation to the shape of the wavefront, but at time intervals of order $\mathcal{O}((\log\eta)^3)$ there are appreciable fluctuations. Roughly, these occur because an individual in the population manages to get significantly ahead of the bulk of the wave. There it reproduces without competition (other than from its own descendants) until the rest of the population catches up ($\mathcal{O}((\log\eta)^2)$ units of time later). By that time, it has reproduced so successfully that a significant portion of the individuals in the wavefront are descendants of this single individual. This will have significant impact on the genealogy of a sample from the population.

If we could ignore the stochastic fluctuations in the wavefront, then the genealogy of a sample of individuals from the population would be determined by a system of Brownian motions with drift (as in (6)) which can coalesce upon meeting at a rate which is inversely proportional to $\rho p_{c_\eta}(x)$, where $p_{c_\eta}(x)$ is the frequency of the favoured allele at the point where they meet. However, the work of Brunet et al. (2006); Berestycki et al. (2012) suggests that in fact it is coalescence due to fluctuations in the wavefront that dominate. More precisely, if we take a sample of lineages from the wavefront, the expected time until a pair of lineages coalesces is $\mathcal{O}((\log \eta)^3)$, which determines the appropriate timescale on which to view coalescence. (The exact spatial locations at which we sample the lineages is not important as long as they are within the front, where the favoured allele is not fixed.) On this timescale, asymptotically (as $\eta \to \infty$), the genealogy of the sample is given by the so-called Bolthausen-Sznitman coalescent, introduced (in a very different context) in Bolthausen and Sznitman (1998). In particular, we see *multiple* mergers, in which more than two ancestral lineages coalesce. Each coalescence event corresponds to the coalescence of all lineages descended from an individual that escaped the rest of the population during an extreme fluctuation. At least as $\eta \to \infty$, it is the coalescences driven by these fluctuations that dominate the genealogy. One of our goals is to establish whether this asymptotic regime is representative of the moderate population densities that we can expect in natural populations.

2.4. Two dimensions

Although Fisher only considered one spatial dimension, equation (1) can be applied in dimensions $d \geq 2$ on replacing $\partial^2/\partial x^2$ by the d-dimensional Laplacian. The difficulties begin when we try to incorporate random genetic drift. The analogue of the stochastic partial differential equation (3) has no solution in dimensions two or higher and so the theory that we have appealed to above has no counterpart.

Ralph and Coop (2010) examine the patterns of diversity resulting from competing selective sweeps arising through parallel adaptation. The randomness in their model is associated with the time and spatial location at which favoured alleles arise rather than genetic drift. Martens and Hallatschek (2011) investigate clonal interference in a two-dimensional setting. Other work on stochastic models in two dimensions has largely been motivated not by the study of spatial selective sweeps, but instead by the desire to understand the patterns of genetic variation resulting from range expansion (Excoffier et al., 2009). The 'surfing' to high frequency of neutral alleles that arise at the frontier of an expanding species, observed in simulations of the stepping stone model, mirrors the rare appearances of highly successful individuals at the wavefront of a population evolving according to (3) described in §2.3. Hallatschek et al. (2007); Hallatschek and Nelson (2010) consider expanding microbial colonies and through both in vitro experiments and numerical simulation of individual based models shows how two competing, but equally fit, strains, expanding into new territory in a circular wave, naturally subdivide into sectors of the two types. There are many other observations of sectoring in expanding colonies (e.g Yin (1993);

Wei and Krone (2005); Krone et al. (2007)). This provides some information on the structure of genealogies in what corresponds to a strong noise and strong selection setting, although there is not yet any detailed theoretical work on how the sectors are created.

We recently introduced a model for the evolution of allele frequencies in populations evolving in spatial continua (Barton et al., 2010a,b) which captures (for different parameter values) both weak and strong (and indeed intermediate) noise. It is readily extended to incorporate selection and recombination, but the (challenging) mathematical analysis is deferred to later work.

3. Results

Our results are based on simulations of the stepping stone model. They are mostly concerned with the case of one spatial dimension where we explore the accuracy of the asymptotic approximations described above for moderate population densities and their implications for the genealogy of a sample from a spatially distributed population after the passage of a selective sweep.

3.1. Outline of the argument

Our first goal is to estimate the net rate of coalescence due to sweeps passing through the population. We begin by focussing on genes at the selected locus; this argument is easily extended to a linked neutral locus. The simplest case is when each sweep is an isolated wave spreading through a one-dimensional population which is dispersed over an infinite range. Suppose that we sample two individuals at separation Δx at some time after the sweep has passed. What is the probability that when we trace backwards in time, the coalescence of their two ancestral lineages is caused by the passage of the sweep? Initially the lineages will follow independent random walks which coalesce on meeting at a rate which is inversely proportional to the local population density. Once caught by the wave, each lineage will be carried back in a narrow wavefront whose shape we approximate by the (truncated) deterministic travelling wave $p_{c_\eta}(\cdot)$. Once both lineages are within the front they evolve according to a random walk with drift whose stationary distribution can be approximated by that for the corresponding Brownian motion with drift which has density f determined by (7). The proportion of time for which the two lineages are both at the location x is then approximately $f(x)^2$, assuming that mixing within this small region is sufficiently fast. When in the same location they coalesce at rate $1/(\rho p_{c_\eta}(x))$ and so we can estimate the net rate of coalescence between two genes Δx

apart in a one-dimensional population given some *rate* of sweeps. We investigate this approximation for a variety of parameter values in §3.3.1.

So far we have considered just the genealogy at the selected locus, for which coalescence is inevitable. Our second goal is to understand the effect of the sweep on neutral variation at linked loci. For this we need to find the rate at which lineages recombine out into the unfavoured background. If the population density is high and ancestors are far out in the wave front, where the favoured allele forms a small proportion of the population, then this is just determined by the recombination rate. If local population density is low, this need not be the case, because lineages may find themselves in regions where the favoured allele is at appreciable frequency. Then, the instantaneous rate of recombination is given by the recombination rate multiplied by $1 - p_{c_\eta}(X_t)$ where $p_{c_\eta}(X_t)$ is the frequency of the favoured allele at the current location of the lineage. Once again we can approximate this from the deterministic frequencies. In fact we can (and do) omit this (small) correction for the range of population densities simulated here.

In two spatial dimensions things are significantly more difficult. We restrict ourselves to the simplest setting of a linear wavefront. Orthogonal to the wavefront we can expect lineages to follow a random walk with drift towards the origin of the sweep. The question is: what is the behaviour of lineages transverse to the wave? One might guess that, at least for small values of $\ell = \sigma\sqrt{2/s}$, the lineages would be trapped in an effectively one-dimensional wavefront along which they diffuse independently until coalescing on meeting at a rate inversely proportional to an effective population density in the wavefront. The distribution of their coalescence times would then be determined by the one-dimensional Wright-Malécot formula (D.1). We test this intuition (and find it badly wanting) in §4.3.

3.2. Speed and shape of the wave

3.2.1. Average behaviour

We measure the position of the wavefront by its centre of mass and we use twice the total number of heterozygotes, that is $w(t) = \int 4p(t,x)(1 - p(t,x))dx$, as a proxy for its width. At any time, the expected rate of advance, $c = d/dt \int p(t,x)dx$, is given by $sw(t)/4$. In Fig. 2, we illustrate the way in which the speed and the 'width' of the wave fluctuate together.

In Fig. 3 we investigate the effect of noise on the average speed of the advancing wave of favoured alleles. Even though our simulations are based on the stepping stone model, the prediction of Brunet and Derrida (1997) that the correction should be $\mathcal{O}(1/(\log\eta)^2)$ remains accurate. However, for the range of values of η considered here, we are not close enough to

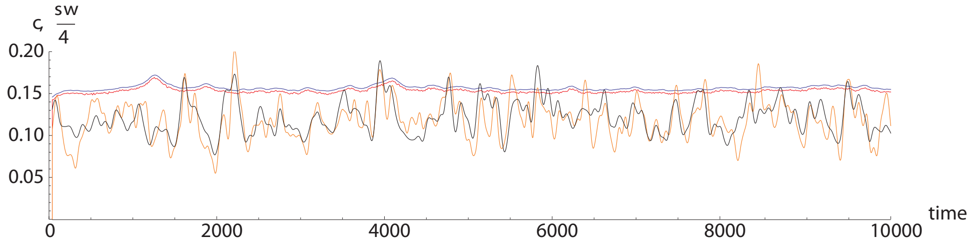

Figure 2: Under the weak selection approximation (2), the expected speed of advance, speed being defined as $c = d/dt \int p(t,x)dx$, is necessarily proportional to the expected 'width' defined as $w(t) = \int 4p(t,x)(1-p(t,x))dx$. The graph shows this relationship (for our simulations, detailed in the Supplementary Material, of the individual based stepping stone model which (2) approximates) for $N = 100$ (lower pair of lines; black: $sw/4$, orange: c) and $N = 10^6$ (upper pair of lines; blue: $sw/4$, red: c). Speed is defined as the change in total allele frequency between generations; both speed and width are smoothed for clarity, using a Gaussian kernel with standard deviation 5. The slight discrepancy is due to the inaccuracy of the weak selection approximation; here, $s = 0.05$.

the asymptotic regime for the constant A in equation (5) to be $\pi^2/2$. We therefore use the fitted formula from Fig. 3 (rather than the asymptotic prediction) to approximate c_η for the remainder of our analysis.

In Fig. 4 we investigate the shape of the wave. This is achieved by calculating average allele frequencies (at each distance from the centre of mass) over 10^5 generations. The expected shape of the wave relative to its centre of mass is close to the approximation provided by the deterministic Fisher wave. This does converge to an exponential distribution far away from the centre of the wave, but extremely slowly, and before we achieve this regime the expected wavefront turns down, in agreement with the approximation p_{c_η}. At the extreme tip of the wave, the expected allele frequency is substantially higher than p_{c_η}. We attribute this to occasional large fluctuations.

3.2.2. Fluctuations in the wavespeed

Although the bulk of the wave is close to the deterministic solution, there are rare large fluctuations in the front (as predicted by the theory). These propagate back to cause modest fluctuations in the 'width' and speed of the wave. We define the 'mass in the front' to be the total mass more than a specified distance ahead of the centre of mass of the wave. Then we observe a delay: an increase in mass in the front is followed by an increase of the speed of the wave which then relaxes down again. See the Supplementary Material for simulations. If we simulated over a long enough time, we would expect to generate a long-tailed distribution of such excursions.

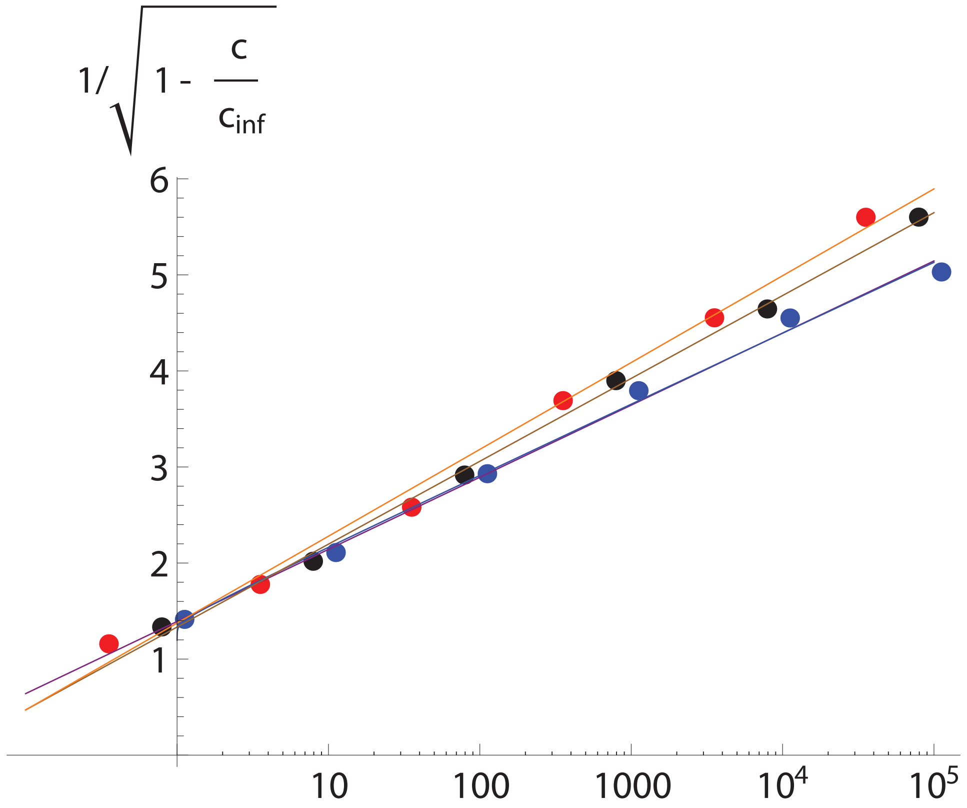

Figure 3: The results of Brunet and Derrida (1997) suggest that the speed of the wave should be reduced in the presence of noise by a factor of the form $(1 - A/(\log(B\eta))^2)$ for some constants A and B. Thus, $1/\sqrt{1 - c/c_\infty}$ should be a linear function of $\log \eta$, where $c_\infty = \sqrt{2ms}$ is the speed of the deterministic wave corresponding to $N = \infty$. The graph is based on the mean speed over 10^5 generations; $m = 0.25$, $s = 0.01, 0.05, 0.10$ (red, black, blue). The lines show least-squares fits for these three selection strengths. The best fit, combining $s = 0.01, 0.05$ gives $A = 6.83$ and $B = 24.82$ so that $c_\eta/c_\infty \sim 1 - 6.83/(\log(35.10\eta)^2$. (See the Supplementary Material for details of the simulations.)

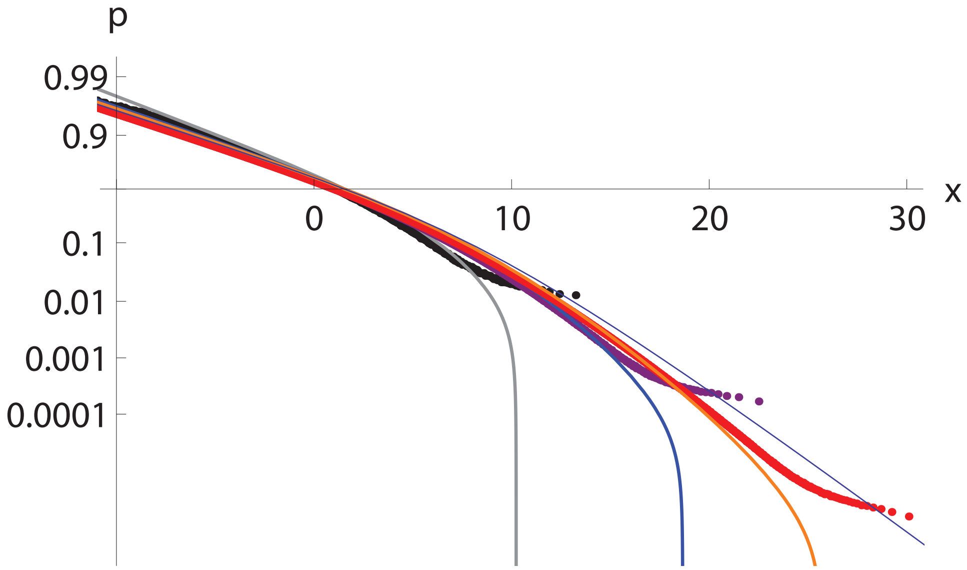

Figure 4: Cline shape for $N = 100$, 10^4, 10^6 (black, purple, red dots); $s = 0.05$, $m = 0.25$, 10^5 generations. For each generation, the position, x, of each deme relative to the centre of mass of the wave, is plotted against its allele frequency, p; each dot shows the means of 5000 such (x, p) values. The upper black curve shows the solution to the deterministic Fisher-KPP equation, which has speed $\sqrt{2ms}$; the grey, blue and orange curves show the deterministic solutions assuming that speed is reduced by a factor $1 - 6.83/(\log(24.82N\sqrt{sm}))^2$ (as estimated from Fig. 3). The actual shape fits closely to this prediction, except at very low frequency, where the average allele frequency is higher. We attribute this to occasional fluctuations that take the front of the wave well ahead of the centre of mass. Details of the simulations are in the Supplementary Material.

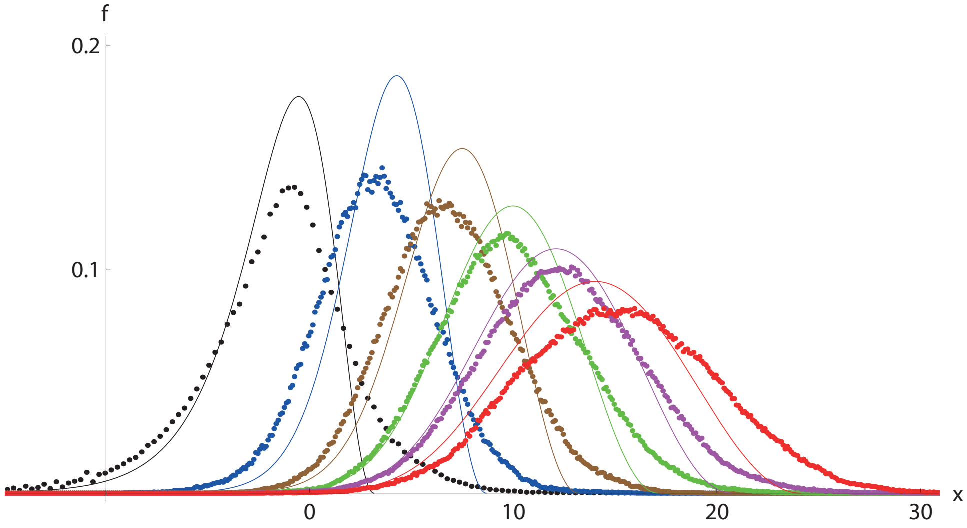

Figure 5: The dots show the distribution of locations of ancestral lineages, relative to the centre of mass, for $N = 10, 100, \ldots, 10^6$ (left to right); $s = 0.05$, $m = 0.25$. The curves show the predicted locations, $f \propto p_{c_\eta}^2 e^{2c_\eta x/\sigma^2}$, where the allele frequency is calculated using the deterministic Fisher-KPP equation, with speed estimated from the formula in Fig. 3. For each N, four replicate lineages were propagated back through 10^5 generations, using a single realisation of the forwards process. Ancestors tend to be ahead of the deterministic prediction, which may be because of the upturn in allele frequency which we attribute to random fluctuations. Details of the simulations are in the Supplementary Material.

3.3. Sampling from the wavefront

3.3.1. A single ancestral lineage within the wavefront

In Fig. 5 we plot the position of ancestral lineages relative to the centre of mass of the wave. We superpose on this the approximation for the stationary distribution of the position of the ancestral lineage predicted by (7) with $p_{c_\eta}(\cdot)$ determined by the (truncated) deterministic Fisher wave and c_η predicted by the formula fitted in Fig. 3. The fit is quite good, illustrating the fact that the large fluctuations in the wave front are rather rare and hardly distort the picture at all.

One might hope to obtain a better approximation by substituting the actual cline shape of the simulation (rather than the deterministic wave with speed c_η) in (7). However, this approach fails. The fluctuations in the wavefront are such that we can no longer normalise the distribution (7) to obtain a probability measure.

3.3.2. Coalescence within the wavefront

The theory summarised in §2.1-§2.3 predicts that, at least for very high population densities, the genealogy of a sample taken from the propagating wavefront will be dominated by a series of 'founder events' at the tip of the wave, caused by (relatively) rare events in which an individual moves substantially ahead of the bulk of the wave and then multiplies unimpeded by competition for long enough that its descendants form a significant proportion of the population in the wave at all future times.

For modest population densities, at least in one spatial dimension, these fluctuations can be expected to be less effective since the competition from the individual's own family rapidly becomes significant. In Fig. 6 we compare the rate of coalescence in the simulated front to that predicted in the absence of fluctuations, $f^2/(\rho p_{c_\eta})$, with f given by (7) and p_{c_η} the approximation to the cline given by the deterministic Fisher-KPP equation. As before, c_η is approximated using the formula in Fig. 3. Although, as we expect, for large values of η this approximation underestimates the coalescence rate, we see in Fig. 6 that it provides a surprisingly good approximation for modest values of η. This suggests that for these moderate values the fluctuations are much less significant. For small population densities, the ancestral lineages are no longer trapped within the front and coalescence outwith the wavefront becomes significant (so that the actual coalescence rate is slower than that predicted by the approximation). In Fig. 7 we look at the locations of coalescence events within the front. Once again, for very large values of η fluctuations become more important.

4. Analysis

We now turn to some analytic predictions. Having presented the basis for the analysis, in §4.2.2, §4.2.3 we use a simple approximation to compare the effect of sweeps in a one-dimensional population to two baselines: a neutral structured population and a sweep in a panmictic population. In two spatial dimensions our approximations are no longer valid and there is essentially no existing theory to draw upon. Instead, in §4.3 we present a preliminary qualitative analysis of the new effects that we see in simulations.

4.1. Assumptions underlying the analysis

For the rest of our analysis, we shall ignore fluctuations in the wavespeed and assume that it takes the constant value c. (We shall drop the subscript η in our notation, but it should be implicitly understood that $c = c_\eta$.) We shall also suppose that the species range, L, satisfies $L \gg \ell$. Under

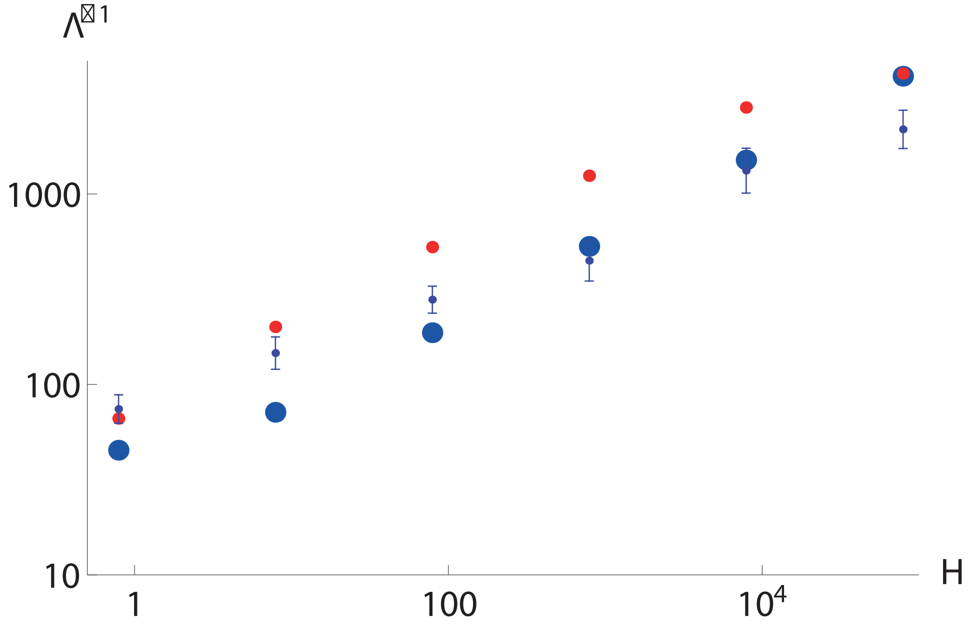

Figure 6: The inverse rate of coalescence within the front is plotted against *eta*. The observed values are indicated by bars, showing twice the standard error; this is based on 10 replicates, each of 400 pairs, started at 0, $10^4, \dots, 9 \times 10^4$ generations; $s = 0.05$, $m = 0.25$. Large blue dots show the *a priori* predictions, based on the Fisher-KPP equation with speed given as in Fig. 3. Small red dots show the predicted rate $f^2/(Np_{c_\eta})$, when one takes the observed distribution of ancestors for f. These two predictions correspond to the upper and middle curves in the distributions of locations of coalescences in Fig. 7. The deterministic prediction fits $\lambda^{-1} \sim (1.4/s)(\rho\sigma\sqrt{s})^{0.44}$ (blue dots), whereas the simulations (detailed in the Supplementary Material) fit $\lambda^{-1} \sim (3.7/s)(\rho\sigma\sqrt{s})^{0.30}$ (error bars).

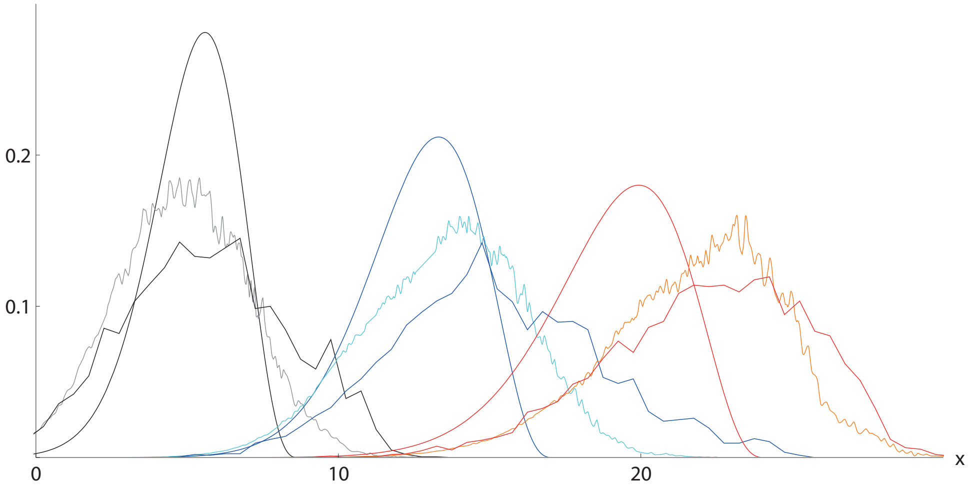

Figure 7: The distribution of coalescence events, plotted against the position x relative to the centre of mass of the cline. Three sets of curves are shown for $N = 100$, 10^4, 10^6 (black, blue, red). Each set is based on a single forwards sweep of 10^5 generations,with $s = 0.5$, $m = 0.25$. The smooth curve is the prediction $f^2/(Np_{c_\eta}) \propto p_{c_\eta}^3 e^{2c_\eta x/\sigma^2}/N$, calculated as in Fig. 3. The middle jagged curve, drawn in a lighter shade, is the prediction f^2/Np_{c_η} obtained by using the actual distribution of the location of ancestral lineages in the simulation for f. The broadest curve in each set is the observed distribution of coalescence events. Each is based on sampling one pair of lineages, replicated 400 times from each of 10 time points at $t = 0, 10^4, \ldots, 9 \times 10^4$ (counting backwards). The distribution of coalescence events is close to the prediction based on the actual locations of ancestral locations, though somewhat further out to the front. The prediction based on the deterministic Fisher-KPP equation is sharper and lies further back - especially for the largest deme size, $N = 10^6$ (red curves at the right). Details of the simulations are in the Supplementary Material.

this assumption, as we trace backwards in time, a lineage sampled at a distance x behind the travelling wave will be 'caught' by the wavefront at a time with mean x/c whose distribution we estimate in (10) below. The assumption $L \gg \ell$ also allows us to assume that from its inception the travelling wavefront of favoured alleles is in its stationary form. In particular, we can then make the approximation $\lambda = \int_0^\infty f^2(x)/(\rho p_{c_\eta}(x))dx$ for the coalescence rate within the front. In Appendix B we show that λ can be written as the product of $2/(\rho\ell)$ and a function, g, of the single dimensionless parameter η, that is $\lambda = 2g(\eta)/(\rho\ell)$ (although we have no explicit expression for g).

4.2. Sampling from outside the front

We have seen that, once two lineages enter the wavefront, they coalesce at approximately the rate λ. We now build on this to find the effect of sweeps on the ancestry of genes sampled at arbitrary locations in space and on the genetic map, and at arbitrary times. To do this, we must allow for three factors: the diffusion of ancestral lineages as we trace backwards in time, before they are caught by the wavefront; their rate of escape from the front through recombination onto the unfavoured background; and their chance of tracing back to coalesce at the original favourable mutation. Figure 8 shows these mechanisms in action in a simulation. These calculations allow us to compare the rate of coalescence at a neutral locus due to sweeps arising at rate Λ and crossing a species range, to the rate of coalescence in the absence of sweeps. Finally, we compare the average rate of coalescence due to successive sweeps with the rate due to those same sweeps in a panmictic population.

4.2.1. Motion of a single lineage

Before considering the rate of pairwise coalescence, we consider the effect of hitchhiking on the spatial motion of a single lineage. Each passing sweep can cause the lineage to experience a large jump, which may in the long run be more significant for the dynamics of the lineages than simple diffusion at rate σ^2. This effect can be seen clearly for the lineages in Fig. 8. To quantify this, note that once within a wavefront a neutral gene at map distance r from the selected locus will escape from the front at a rate r. (As explained in §1, we will assume that the favoured allele is sufficiently rare in the vicinity of the ancestral lineage that all recombination events are with an individual of the unfavoured type and so result in escape from the sweep.) If the wave is travelling at speed c and the selected mutation arose at a distance x away (in space) from the current location of the lineage, then the

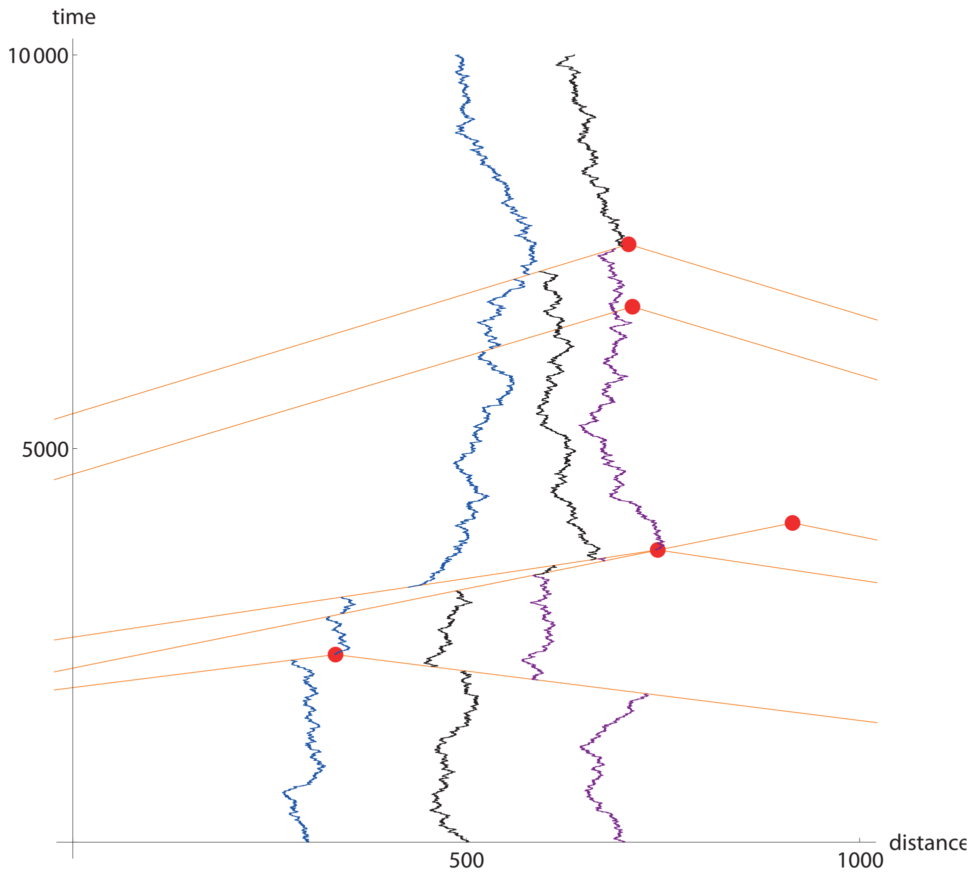

Figure 8: The effect of 5 successive selective sweeps on the ancestry of three genes, sampled at $x = 300$, 500, 700; genes diffuse at a rate $\sigma^2 = 1$. The origin of five selective sweeps is shown by the red dots, and the advancing wavefronts by the orange lines. The origin, speed and position on the genetic map were drawn from a uniform distribution; in this example, the map location, relative to the focal locus, is −0.90cM, +0.80cM, −0.77cM, +0.83cM, −0.19cM. When a lineage hits a wavefront, it is carried back towards the favourable mutation (red dot) but may escape by recombination. The population is assumed very dense, so that coalescence only occurs within the wavefront. In this example, the purple and black lineages lie within the same wavefront ($x \sim 600 - 750$, $t \sim 3500$), but escape from it without coalescing. The only coalescence event occurs at the origin of the oldest selective sweep ($x \sim 700$, $t \sim 6800$), when both the purple and black lineages are carried back to coalesce on the genome that carried the favourable mutation. See Supplementary Material for details.

expected length of time between the lineage being caught by the sweep and the origin of the favourable mutation is x/c (see equation (10) below for the approximate distribution of this time). In particular, the probability that the lineage will trace back to that mutation, and so return to the ancestral background without recombination, is approximately $\exp(-rx/c)$. More generally, as a result of a sweep which is travelling at rate c, the lineage will move an exponentially distributed distance (with parameter c/r), truncated at the location of the origin of the mutation. This represents a mean square displacement of a lineage due to the sweep of

$$2\left(\frac{c}{r}\right)^2\left(1-\left(1+\frac{rx}{c}\right)e^{-rx/c}\right), \tag{8}$$

see Appendix C.1.

We assume that sweeps are uniformly distributed over a genetic map of length R and that they originate at points that are uniformly distributed over a one-dimensional range of length L. In Appendix C.1 we show that if the overall rate of sweeps is Λ across the entire range and genome, and $RL/c \gg 1$, then the mean square displacement per unit time of a lineage sampled uniformly from the species range due to hitchhiking, relative to that without hitchhiking, is

$$\frac{\sigma_{\text{eff}}^2}{\sigma^2} \approx \frac{8}{3}\frac{L}{\ell}\frac{\Lambda}{R}, \tag{9}$$

where, as before, $\ell = \sigma\sqrt{2/s}$ is the characteristic length scale of the wavefront when $\eta = \infty$. There will be some bias in the direction of this displacement, because unless a lineage is sampled from exactly the centre of the range, there will always be more sweeps arising on one side of it than the other.

4.2.2. Comparison to a neutral one-dimensional population

The long-term effective size of a structured population is determined by the rate of coalescence between two genes, sampled uniformly at random from the whole population (Charlesworth et al., 2003). In this subsection we compare this quantity in the presence of sweeps to that in a spatially structured neutral population.

Suppose first that we sample a lineage at distance x behind the current position of the wavefront. Relative to the front, it follows a random walk with drift (which as usual we approximate by its Brownian counterpart). If $x/\ell \gg 1$, then we can suppose that the time until a lineage is ‘caught by the wave' is approximately the time for a Brownian motion with drift $-c$ to hit zero for the first time, given that it started at x. This time has density

$$\phi(t,x) = \frac{x}{\sigma\sqrt{2\pi t^3}} \exp\left(-\frac{(x-ct)^2}{2\sigma^2 t}\right), \tag{10}$$

see e.g. Cox and Miller (1977) p.221, Equation 74. Notice that, at least if $x \gg \sigma$, this is essentially a Gaussian density, centred on $t = x/c$, the time since the wave was at the current location of the lineage, with standard deviation σ/c.

Now suppose that we sample two lineages from behind the wavefront. For all but small population densities, we can neglect the chance that the lineages will coalesce before they are both captured by the wavefront and so until that time they evolve independently of one another and equation (10) determines the distribution of that time. Once a lineage enters the front, it can recombine out onto the less fit background at a rate r. (We ignore the chance that it recombines in again, since the fitter background is typically rare, and is receding as we trace backwards in time). Once both lineages are within the wavefront, each can escape at rate r or the two lineages can coalesce at rate λ. Thus, if the favourable mutation is sweeping across an infinite range, then once in the front the chance of coalescence before either lineage escapes through recombination is just $\lambda/(\lambda+2r)$. Putting all this together, if the favourable mutation were sweeping across an infinite range, the probability of coalescence of two lineages sampled at distances x_1, x_2 behind the wavefront would be approximately

$$P\left(x_1, x_2\right) = \frac{\lambda}{\lambda+2r}\int_0^\infty \int_0^\infty \phi\left(t_1, x_1\right)\phi\left(t_2, x_2\right) e^{-r|t_2-t_1|} dt_1 dt_2. \tag{11}$$

(The integrand is the probability that the lineages enter the front at times t_1 and t_2 respectively, multiplied by the probability that the lineage which is captured first does not escape through recombination before the second lineage is captured.) In a finite range we will have to correct this expression to take into account the possibility that the lineages trace back to the origin of the sweep before recombining away. We perform this correction in Appendix C.1. From equation (11) we see that (provided the sweep is not of too recent origin) the chance that both lineages will meet in the wavefront (i.e., the double integral in (11)) depends on the scaled distances from the front, x_i/ℓ, where, as before, $\ell = \sigma\sqrt{2/s}$, and on the rate of recombination relative to selection, r/s. Figure 9 shows that (in an infinite range) provided $r \ll s$, if we sample two lineages from the same location, even if that location is a

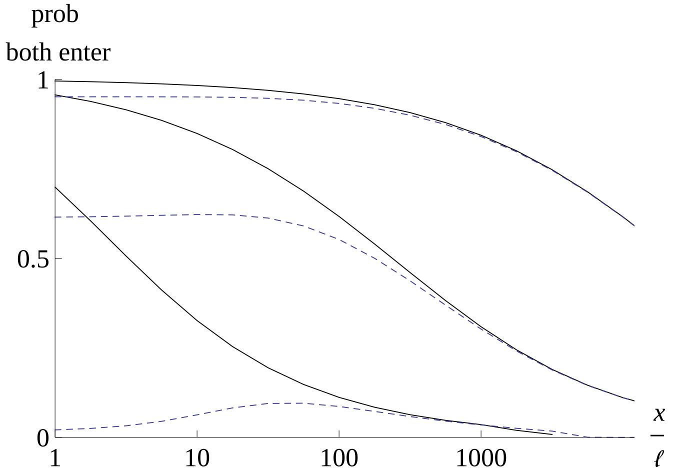

Figure 9: The probability that two lineages will meet in the wavefront, given that they are sampled at distances x, $x + \Delta x$ behind the front. The three pairs of curves are for $r/s = 0.01$, 0.1, 1 (top to bottom); solid curves are for $\Delta x = 0$, dashed curves for $\Delta x = 10$. The probability of coalescence equals this probability, multiplied by $\lambda/(\lambda + 2r)$ (equation (11)).

long way from the wavefront, there is a high probability that as we trace backwards in time the ancestral lineages will meet in the wavefront. The explanation is that the lineages will not diffuse far apart before being caught by the wave and so the time during which there is exactly one lineage in the front is short. In particular, the chance of us seeing the recombination event required for the lineage to 'escape' during this period is small. A separation Δx between genes sampled from just behind the front makes little difference to the chance of the two lineages meeting in the front if linkage is tight (upper dashed curve, $r/s = 0.01$ in Fig. 9) but greatly reduces their chances if linkage is loose (lower dashed curve, $r/s = 1$). This suggests that, as one expects, the portion of the rate of coalescence between two randomly chosen lineages that can be attributed to hitchhiking is almost entirely due to sweeps at loci that are linked tightly to the selected locus. In Appendix C.1, we show that the mean rate of coalescence, averaged over random locations of genes and sweep, and over a genetic map of length $R \gg 1$ is

$$2\frac{\Lambda}{R}\frac{c}{L}\, h\left(\frac{L\lambda}{c}\right),$$

for a function $h(\theta)$ which tends to 1 for small θ, and to

$$2\left(\log\left(\frac{\theta}{2}\right)+\gamma-1\right)$$

for large θ, where γ=0.577... is Euler's gamma. Since $h(\theta)$ increases only logarithmically with θ, the net rate of coalescence is insensitive to the rate of coalescence within the front: the limiting factor is the rate at which tightly linked sweeps occur.

The average rate of coalescence per sweep per map length is inversely proportional to L/c, which is the time a sweep takes to cross the species' range. Since this is typically large, and Λ/R is expected to be small, the rate of coalescence due to hitchhiking will be very small. However, in a sufficiently dense population, genetic sampling drift will be negligible, so that hitchhiking may still be the main cause of coalescence (Maynard Smith and Haigh, 1974; Gillespie, 2000). The effective size of a neutral haploid one-dimensional population is $\rho L + \frac{1}{12}\left(\frac{L}{\sigma}\right)^2$, where ρL is the total number of haploid individuals (Charlesworth et al., 2003). The contribution of hitchhiking will be larger than this if

$$2\frac{\Lambda}{R}\frac{c}{L}\,h\left(\frac{L\lambda}{c}\right) > \frac{1}{\rho L}$$

or if $2(\Lambda/R)c\rho h(L\lambda/c) > 1$. The factor $c\rho$ is proportional to $sw\rho$, where w is the total number of heterozygotes, which we introduced as a proxy for the width of the wave in §3.2.1. We expect w to be proportional to $\ell = \sigma\sqrt{2/s}$ and so $sw\rho$ will be proportional to the quantity that we have called η. Thus, in one dimension, if the product of the rate of sweeps per map length and the parameter η is large, hitchhiking will be the main cause of coalescence.

4.2.3. Comparison to panmixis

We can also compare the probability of coalescence of two neutral lineages (sampled uniformly from the population) due to the sweep with the same probability for a sweep in a panmictic population. If we suppose that the species range is big enough that we can ignore the probability of individuals tracing back to the original mutation (that is $Lr \gg c$) then, as we saw in the last section, the chance that two lineages sampled at separation Δx meet in the front and coalesce is approximately

$$\frac{\lambda}{\lambda+2r}e^{-r\Delta x/c}. \tag{12}$$

Assuming that we sample uniformly from the species' range, the probability of coalescence of a randomly chosen pair of lineages is then approximately

$$\frac{\lambda}{\lambda + 2r} \frac{2c}{rL}, \tag{13}$$

see Appendix C.1. As we recall in Appendix D, in a panmictic population of N haploid individuals, the probability that a sweep will cause two lineages to coalesce is approximately $(2Ns)^{-2r/s}$ (Maynard Smith and Haigh, 1974). This is of a quite different form to (13): the effect of a sweep on coalescence falls away much faster on the genetic map in a panmictic population than in one spatial dimension ($\sim e^{-r/s}$ vs. $\sim \frac{1}{r}$), but (still assuming $r \ll s$) is less sensitive to population size ($\sim N^{-r/s}$ vs. $\frac{1}{L}$).

Now suppose that we sample two individuals at separation Δx from behind the wavefront. Until caught by the wave lineages have no chance to escape the sweep through recombination. The effectiveness of hitchhiking is determined by the length of the time period during which exactly one ancestral lineage is in the wavefront. (After this period, the probability of escape is $2r/(\lambda + 2r)$.) This period is of duration roughly $\Delta x/c$ which should be compared to the duration of a sweep through a panmictic population (throughout about half the course of which favoured alleles are rare and so both lineages have a chance to escape through recombination). Thus the effectiveness of hitchhiking in a spatially structured population will be greater than in a panmictic popluation for individuals sampled at a separation Δx with $\Delta x/c < \log(2Ns)/s$ which we can write $\Delta x/\ell < \log(2Ns)$ where, as usual, $\ell = \sigma\sqrt{2/s}$.

4.3. Two spatial dimensions

Our analysis so far has been entirely concerned with populations distributed across a one-dimensional range. In higher dimensions there is essentially no analytic theory on which to draw. We shall show that understanding the effect of selective sweeps on the movement and coalescence of genes in two dimensions will not be a straightforward extrapolation of the one-dimensional results.

We focus on the simplest case of a linear wavefront. Let us try to mimic what we did in one dimension and examine the probability that the ancestral lineages of two neutral genes sampled from behind the wavefront coalesce as a result of a sweep. We distinguish their positions transverse to the wave, which we denote by y_1, y_2, from those orthogonal to the wave, x_1, x_2. Tracing backwards in time, just as in one dimension, if we can neglect the fluctuations in the wavefront, the orthogonal distances between the ancestral lineages and the wavefront will follow (approximately) Brownian motions with drift and so the distribution of the times t_1 and t_2 when they are caught by the wave is determined by (10).

Since the favoured allele is rare in the wavefront, we assume, as before, that all recombination events to affect a lineage in the wavefront result in escape. The chance that the first lineage to be caught escapes the sweep before the second one is caught is then $e^{-r|t_2-t_1|}$. If that does not happen, then both lineages will be caught in the front. We assume that until they are both caught by the wavefront, their separation transverse to the wave is governed by a diffusion of rate $2\sigma^2$, started from $y = y_1 - y_2$. Once within the front, we consider the lineages to be trapped. Let us write $\Psi(t, z)$ for the probability density of the time until the ancestral lineages of two favoured alleles, sampled at separation z from within the wavefront, coalesce. The chance of coalescence due to the sweep of our two neutral lineages sampled from behind the wave can then be written

$$\int_0^\infty \int_0^\infty \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{4\pi\sigma^2|t_2-t_1|}} e^{-(y-z)^2/(4\sigma^2|t_2-t_1|)} \\ \phi(t_1, x_1)\phi(t_2, x_2)e^{-r|t_2-t_1|}e^{-2rt}\Psi(t, z)dzdtdt_2dt_1. \quad (14)$$

The size of this probability will hinge on the nature of the relative motion of the two lineages transverse to the wavefront (which, combined with the 'effective population density' in the wavefront will determine $\Psi(t, z)$).

A natural conjecture is that the motion of a single lineage along the wavefront is governed by a one-dimensional Brownian motion with variance σ^2. If lineages evolve independently until meeting, their separation will also be governed by a Brownian motion, but with twice the variance. Just as in one dimension, essentially all recombination events that take place in the wavefront lead to a lineage escaping the sweep. Provided we work on an infinite range, the probability of recombination before coalescence, which in the notation above is

$$\int_0^\infty e^{-2rt}\Psi(t, z)dt, \quad (15)$$

can then be calculated from the classical one-dimensional Wright-Malécot formula (D.1), with recombination playing the rôle of mutation and with an effective population density $\rho_e = 1/\lambda$.

In Fig. 10, we plot the resulting probability of coalescence multiplied by $(1 + 2\rho_e\sigma\sqrt{2r})$. This constant is a common factor in the Wright-Malécot equation, corresponding to the probability that two lineages coalesce before

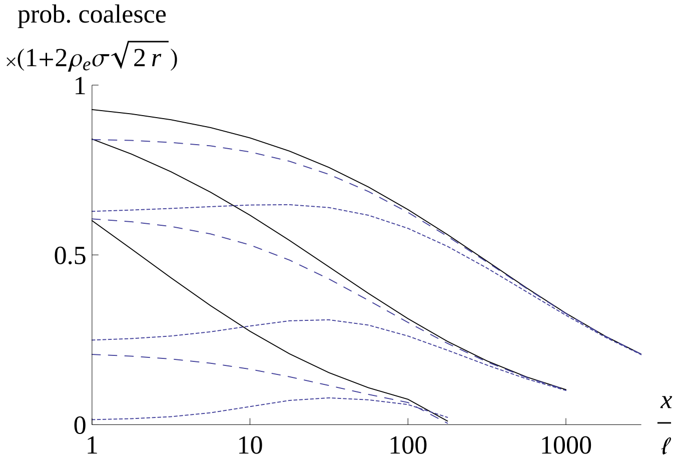

Figure 10: The probability that two lineages will coalesce within the wavefront, given that they are sampled at distances x, $x+\Delta x$ behind the front, and separated by Δy transverse to it. The plotted values must be divided by $1+2\rho_e\sigma\sqrt{2r}$ to give the actual probability of coalescence. The three pairs of curves are for $r/s = 0.001$, 0.01, 0.1 (top to bottom); solid curves are for $\Delta x, \Delta y = 0$; long dashed curves are for $\Delta x = 10\ell$, $\Delta y = 0$; and dotted curves are for $\Delta x = 0$, $\Delta y = 10\ell$. Note that linkage is ten times tighter here than in the corresponding Fig. 9 for one dimension, and that separation in the transverse direction reduces coalescence by much more than separation in the direction of movement (i.e., dotted curves ($\Delta x = 0$, $\Delta y = 10\ell$) are lower than dashed curves ($\Delta x = 10\ell$, $\Delta y = 0$).

recombination if they start in the same location (the analogue in one dimension is the factor $\lambda/(\lambda + 2r)$ in equation (11)). Since we have no way to estimate it, we prefer to scale it out rather than assign it an arbitrary value. The resulting quantity is the probability that two lineages meet (at separation zero) in the wavefront before either escapes the sweep. We can compare this to the one-dimensional results in Fig. 9. We notice that if this analysis applies, then in two dimensions recombination is much more effective at enabling lineages to escape the sweep than in one dimension.

It is also clear that displacement transverse to the wave reduces the chance of coalescence by much more than separation in the direction of the wave movement. This is because motion in the direction of the wave is much faster than transverse diffusion. In Appendix C.2 we outline the analogue of our analysis of §4.2.3 and show that if motion of lineages along the wavefront really were governed by independent diffusions, genetic hitchhiking would be remarkably ineffective in two spatial dimensions.

If the analysis above were correct, then we would see essentially no coalescence within the wavefront, with lineages instead tracing back to the origin of the selective sweep. To investigate this, we simulated a sweep on a cylindrical habitat, with circumference 200 demes, driven by $s = 0.05$ and with $m = 0.25$; demes were extremely large, with $N = 10^6$. The sweep quickly settled to advance at 0.155 demes per generation, very slightly slower than the deterministic prediction $\sigma\sqrt{2s} = 0.158$ demes per generation. Tracing backwards through time, just as in one dimension, lineages become trapped within the front. Following one lineage at a time, transverse to the front, they diffuse at a rate close to the predicted $\sigma^2 = m = 0.25$ (variance increases as $0.24t$ for 50 replicate lineages, started at each of 20 points spaced evenly around the circumference). However, coalescence events are strongly clustered, along ~ 3 paths. These correspond to occasional fluctuations, when a few individuals well ahead of the wavefront have extremely large numbers of offspring. Thus, while the gross rate of movement of single lineages is close to a simple diffusion, this conceals a clustering that greatly increases the rate of coalescence (Fig. 11a). Hallatschek et al. (2007) suggest that the motion of lineages within the front is superdiffusive (with variance increasing like $t^{4/3}$). Although our simulations appear to show mean square displacement increasing linearly with time, there is some noise and it would require more work to reject superdiffusivity.

Simulations with moderate deme size show a yet more extreme pattern. In Fig. 11(b), $N = 100$ and $m = 0.25$, and so 'neighbourhood size' is $4\pi Nm \sim 314$ - large enough that the population is well-mixed in the absence of selection. Nevertheless, once again, the ancestry of lineages within the front is dominated by rare events, in which a few individuals that happen to be far ahead of the wave have very many offspring. Tracing back, lineages from a wide section of the front derive from these few ancestors, and so are very likely to coalesce there. Such events will very greatly increase the rates of coalescence, far above that predicted from the diffusion analysis, by pulling lineages close together as we trace back in time. This effect has the potential to make sweeps through a two dimensional population a powerful source of local coalescence. Moreover, a single sweep would lead to multiple local founder events, producing substantial spatial heterogeneity ('sectoring'), which should be readily detectable.

Nevertheless, even if fluctuations at the tip of the wave are frequent enough that lineages coalesce rapidly once they both enter the wavefront, it is unlikely that the net rate of coalescence for individuals sampled uniformly across the whole population can be higher than in panmixis (see (C.5) in Appendix C.2). Indeed, just as in one dimension, the probability of coalescence of two lineages sampled at separation Δx is limited by the interval of length $\Delta x/c$ between the times when they hit the front; the probability of coalescence is bounded above by the expression in (12). Thus the net rate of coalescence is bounded by that in one-dimension which, we have seen, is unlikely to be larger than in panmixis. Understanding the local and global effects of hitchhiking rigorously presents a considerable mathematical challenge.

5. Discussion

A favourable allele sweeps through a population behind a wave which is assumed to be much narrower than the species' range. Most ancestry derives from a few individuals at the very tip; although the bulk of the wave advances almost deterministically, its speed is limited by stochastic fluctuations at the tip. In the limit where $\log \eta = \log(\rho\sigma\sqrt{s/2})$ becomes very large, these fluctuations dominate (Berestycki et al., 2012). However, we show that for moderately dense one-dimensional populations, the rate of coalescence within the wavefront is reasonably well approximated by a simple deterministic calculation, that treats the ancestral population as well-mixed and homogeneous.

We use this deterministic approximation to the configuration of the ancestral population to find the effect of sweeps on arbitrarily placed lineages. When a sweep of a favourable mutation at map distance r, travelling at speed c, hits a single lineage, it causes a sudden jump in location with mean square displacement $2(c/r)^2$. From equation (9) we see that, averaging over the genome, sweeps will cause more gene flow than simple diffusion if $(\Lambda/R)(L/\ell)$ is large. The quantity Λ/R is the rate of sweeps per map length, and L/ℓ is the size of the range relative to the width, ℓ, of the wavefront which is taken to be $\ell = \sigma\sqrt{2/s}$. (Recall that the width of the region in which the allele frequencies in the classical Fisher wave are between two prescribed values is proportional to $\sigma\sqrt{2/s}$.) Thus, even if the rate of sweeps per map distance is low, if the species' range is long, then in the long run hitchhiking can be more important than diffusion for the motion of ancestral lineages. The rate of adaptive sweeps per map length (Λ/R) is poorly known; the best estimates are from divergence between *Drosophila melanogaster* and *D. simulans*, where perhaps $\Lambda/R \sim 0.02$; it may be much higher in populations under strong artificial selection, but lower in organisms with longer genetic maps (see the discussion in Weissman and Barton (2012)). Thus, hitchhiking could substantially inflate gene flow if the species range is much wider than the wavefront ($L/\ell > 50$, say).

To put this another way, since $c = c_\eta \approx \sigma\sqrt{2s}(1 - A/(\log\eta)^2)$, hitchhiking will be a significant source of gene flow if the product of the selection

.

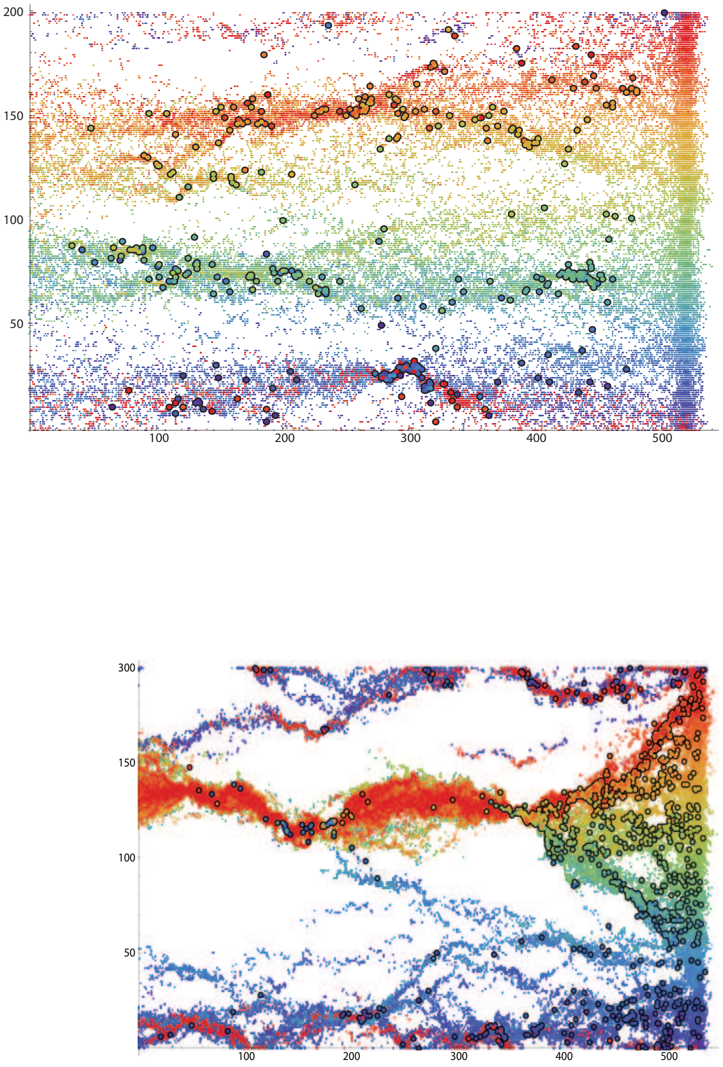

Figure 11: Coalescence events in two dimensions cluster onto three paths, where most ancestors are located. The habitat is a cylinder with circumference 200 demes (vertical axis); time runs from left to right, moving ~ 520 demes in 4000 generations; $s = 0.05$, $m = 0.25$, a) $N = 10^6$, b) $N = 100$. Coalescence was simulated in two ways. First, 100 pairs of genes were started at each of 20 locations $y = 0, 10 \ldots, 190$ at right; coloured blue…green…red); large dots show where these pairs coalesce. By 4000 generations, $461/2000 = 23\%$ had coalesced. Second, 50 single ancestral lineages were propagated back from $y = 0, 10 \ldots, 190$ (coloured in the same way); small dots show where any pair that started from the same location coincide in the same deme. This second simulation is much faster, and indicates the distribution of potential coalescence events, given this particular selective sweep. The lower panel shows the same for $N = 100$. In that case 85% had coalesced by 4000 generations. Details of the simulations can be found in the Supplementary Material.

coefficient and the time taken for a sweep to cross the species range is large: $sL/c \approx L/\ell \gg 1$. However, the spatial motion of lineages due to hitchhiking will be quite different in nature to the diffusion of lineages in a neutral population which is uniformly distributed across its range. Hitchhiking acts through very rare events in which a favourable mutation arises in tight linkage, and carries the neutral lineage across a substantial fraction of the range.

The rate of pairwise coalescence, and hence the long-term effective size of the population, may be much more weakly influenced than gene flow, simply because both genes must be trapped within the front if they are to coalesce there. In both one and two dimensions, the net probability of coalescence, averaged over the whole genome, depends primarily on the number of crossovers during the time it takes for the sweep to cross the whole range (i.e. $\sim 1/(RL/c)$). This parallels the result of Maynard Smith and Haigh (1974) for panmixis, where the probability of coalescence of two neutral genes also depends on the amount of recombination over the whole time-course of the sweep. Since fixation must be substantially slower with spatial structure, we expect the net effect of a sweep to be lower, despite the possibility of coalescence within the front. In two dimensions, we saw that lineages coalesce much faster than expected from independent diffusions; fluctuations dominate, and so greatly increase the rate of coalescence within the front. Nevertheless, it is unlikely that lineages can be brought together fast enough to make coalescence faster than in a well-mixed population. Under panmixis, a single mutation will fix in $\sim (1/s)\log(4N_e s)$ generations, and it is hard for any process to bring together two lineages in different parts of the broad species' range faster than this.

Hitchhiking in space may be significant nevertheless. First, the rate of local sweeps (in which an allele spreads through a limited region) could be far higher than the rate of species-wide sweeps, which is limited by the observed rate of substitutions in coding and functional non-coding sequences that is seen in comparisons between species. Moreover, sweeps from standing variation will cause 'surfing' in much the same way as 'hard sweeps' that start from single mutations: all that matters is that a few individuals at the front of the wave dominate reproduction. Yet, such sweeps would not be detected in surveys that count regions of low sequence diversity. It is not clear how we could estimate this local rate of sweeps: is the rate of sweeps that pass any given point much higher than the rate of sweeps that fix across the whole population? Second, hitchhiking in space leaves a distinct signature, which may be detected even if it makes a weak contribution to the total species-wide rate of coalescence. It will be interesting to find the pattern of summary statistics such as F_{st}. However, in any particlar case there will be a limited number of ‘founder events’ at the wavefront, and it may be more sensible to locate these, rather than to attempt to estimate any long-term expectation. This could best be done through intensive studies of variation around loci that are known to have recently spread across a broad area (e.g. Hoekstra et al. (2006); Karasov et al. (2010)).

Acknowledgements

We thank two referees for their valuable suggestions and Chrisine Krebs for her assistance with the figures.

References

Barton, N.H., 1998. The effect of hitch-hiking on neutral genealogies. Gen. Res. 72, 123–133.

Barton, N.H., 2000. Genetic hitchhiking. Phil. Trans. Roy. Soc. (Lond.) B 355, 1553–1562.

Barton, N.H., Depaulis, F., Etheridge, A.M., 2002. Neutral evolution in spatially continuous populations. Theor. Pop. Biol. 61, 31–48.

Barton, N.H., Etheridge, A.M., Véber, A., 2010a. A new model for evolution in a spatial continuum. Electron. J. Probab. 15, 162–216.

Barton, N.H., Kelleher, J., Etheridge, A.M., 2010b. Genetic hitchhiking. Evolution 64, 2701–2715.

Bérard, J., Gouéré, J.B., 2011. Survival probability of the branching random walk killed below a linear boundary. Electron. J. Probab. 16, 396–418.

Berestycki, J., Berestycki, N., Schweinsberg, J., 2012. The genealogy of branching brownian motion with absorption. Ann. Probab. To appear.

Bolthausen, E., Sznitman, A.S., 1998. On ruelle’s probability cascades and an abstract cavity method. Comm. Math. Phys. 197, 247–276.

Bramson, M., 1983. Convergence of solutions of the Kolmogorov equation to travelling waves. Mem. Amer. Math. Soc. 44.

Brunet, E., Derrida, B., 1997. Shift in the velocity of a front due to a cutoff. Phys. Rev. E 56, 2597–2604.

Brunet, E., Derrida, B., 2001. Effect of microscopic noise on front propagation. J Statist. Phys. 103, 269–282.

Brunet, E., Derrida, B., Mueller, A.H., Munier, S., 2006. Noisy travelling waves: effect of selection on genealogies. Europhys. Lett. 76, 1–7.

Charlesworth, B., Charlesworth, D., Barton, N.H., 2003. the effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. syst. 34, 99–125.

Cox, D., Miller, H., 1977. The theory of stochastic processes. Chapman & Hall/CRC.

Durrett, R., Schweinsberg, J., 2004. Approximating selective sweeps. Theor. Pop. Biol. 66, 129–138.

Etheridge, A., Pfaffelhuber, P., Wakolbinger, A., 2006. An approximate sampling formula under genetic hitchhiking. Ann. Appl. Probab. 16, 685–729.

Excoffier, L., Foll, M., Petit, R.J., 2009. Genetic consequences of range expansions. Annual Review of Ecology, Evolution and Systematics 40, 481–501.

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

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

Hallatschek, O., Hersen, P., Ramanathan, S., Nelson, D.R., 2007. Genetic drift at expanding frontiers promotes gene segregation. Proc. Natl. Acad. Sci. (USA) 104, 19926–19930.

Hallatschek, O., Korolev, K., 2009. Fisher waves in the strong noise limit. Phys Rev Lett 103, 108103.

Hallatschek, O., Nelson, D., 2008. Gene surfing in expanding populations. Theor. Pop. Biol. 73, 158–170.

Hallatschek, O., Nelson, D., 2010. Life at the front of an expanding population. Evolution 64, 193–206.

Hoekstra, H.E., Hirschmann, R.J., Bundey, R.A., Insel, P.A., Crossland, J.P., 2006. A single amino acid mutation contributes to adaptive color pattern in beach mice. Science 313, 101–104.

Karasov, T., Messer, P.W., Petrov, D.A., 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. PloS Genetics 6, e1002740. doi:10.1371/journal.pgen.1000924.

Karlin, S., Taylor, H.M., 1981. A second course in stochastic processes. Academic Press.

Kim, Y., Maruki, T., 2011. Hitchhiking effect of a beneficial mutation spreading in a subdivided population. Genetics 189, 213–226.

Kimura, M., 1953. Stepping stone model of population. Ann. Rep. Nat. Inst. Genetics Japan 3, 62–63.

Klopfstein, S., Currat, M., Excoffier, L., 2006. The fate of mutations surfing on the wave of a range expansion. Mol. Biol. Evol. 23, 482–490.

Kolomogorov, A., Petrovsky, I., Piscounov, N., 1937. Étude de l'equation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. Moscow Univ. Math. Bull. 1, 1–25.

Krone, S.M., Lu, R., Fox, R., Suzuki, H., Top, E.M., 2007. Modelling the spatial dynamics of plasmid transfer and persistence. Microbiology 153, 2803–2816.

Malécot, G., 1948. Les Mathématiques de l'hérédité. Masson et Cie, Paris.

Martens, E.A., Hallatschek, O., 2011. Interfering waves of adaptation promote spatial mixing. Genetics 189, 1045–1060.

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

Mueller, C., Mytnik, L., Quastel, J., 2011. Effect of noise on front propagation in reaction-diffusion equations of KPP type. Inv. Math. 184, 405–453.

Mueller, C., Sowers, R., 1995. Travelling waves for the KPP equation with noise. J. Functional Anal. 128, 439–498.

Mueller, C., Tribe, R., 1997. Finite width for a random stationary interface. Elect. J. Prob. 2, paper 7, 1–27.

Nagylaki, T., 1978. Random genetic drift in a cline. Proc. Nat. Acad. Sci. USA 75, 423–426.

Ralph, P., Coop, G., 2010. Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 186, 647–668.

Schweinsberg, J., Durrett, R., 2005. Random partitions approximating the coalescence of lineages during a selective sweep. Ann. Appl. Probab. 15, 1591–1651.

Shiga, T., 1988. Stepping stone models in population genetics and population dynamics, in: et al, S.A. (Ed.), Stochastic processes in physics and engineering, D Reidel Publishing Company.

Slatkin, M., Wiehe, T., 1998. Genetic hitch-hiking in a subdivided populations. Genetical Research 71, 155–160.

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

van Saarloos, W., 2003. Front propagation into unstable states. Phys. Rep. 386, 29–222.

Wei, W., Krone, S.M., 2005. Spatial invasion by a mutant pathogen. J. Theor. Biol. 236, 335–348.

Weissman, D.B., Barton, N.H., 2012. Limits to the rate of adaptation in sexual populations. PloS Genetics 8, e1002740. doi:10.1371/journal.pgen.1002740.

Yin, J., 1993. Evolution of bacteriophage T7 in a growing plaque. J. Bacteriol. 175, 1272–1277.

Appendix A. The truncated Fisher wave

Travelling wave solutions to (1) are found by writing $p(t,x) = p(x - ct)$ where p satisfies

$$\frac{\sigma^2}{2}\frac{d^2p}{dx^2} + c\frac{dp}{dx} + sp(1-p) = 0. \tag{A.1}$$

In order to understand the shape of such solutions in the 'front' of the wave, where the favoured allele is rare, we look for solutions of the form

$p(x) = \exp(-\gamma x)$. The parameter γ is related to the wavespeed c through

$$\gamma c(\gamma) = \frac{\sigma^2}{2}\gamma^2 + s. \tag{A.2}$$

The smallest value of c for which this yields a real-valued γ is $c_\infty = \sigma\sqrt{2s}$, corresponding to $\gamma_\infty = \sqrt{2s}/\sigma = 2/\ell$. Noise has the effect of slowing the wave down, so the corresponding γ will be complex-valued. To leading order, if we assume only a small perturbation (corresponding to very large values of ρ in (3)), the real part will be γ_∞, but there will be a complex correction so that, after a shift of the x-coordinate, the solution becomes

$$p(x) \approx \text{Const.}\frac{W}{\pi}\sin\left(\frac{\pi x}{W}\right)\exp\left(-\gamma_\infty x\right). \tag{A.3}$$

We suppose that the solution is truncated at $x = W$, to ensure that it is always non-negative. The value of W is such that $e^{-\gamma_\infty W} \sim 1/\eta$ (as then the fluctuations will outweigh the deterministic solution), that is $W \sim (\log\eta)/\gamma_\infty$. Writing $p(x)$ in equation (A.3) as $e^{-\gamma_\eta x}$ we see that

$$\gamma_\eta \approx \gamma_\infty + \frac{i\pi}{W},$$

and so using (A.2) we obtain

$$c_\eta = c(\gamma_\eta) \approx c(\gamma_\infty) - \frac{1}{2}\left(\frac{\pi}{W}\right)^2 c''(\gamma_\infty).$$

Now

$$c''(\gamma_\infty) = \frac{2s}{\gamma_\infty^3} = \frac{2s\sigma^3}{(2s)^{3/2}} = \frac{\sigma^3}{\sqrt{2s}} = \frac{c_\infty}{\gamma_\infty^2},$$

and so

$$c_\eta \approx c_\infty\left(1 - \frac{\pi^2}{2(\log\eta)^2}\right).$$

This argument is not rigorous, but for (3), writing $p_{c_\eta}(z)$ for the allele frequency in the stationary wavefront, at least where allele frequencies are small we have (Brunet et al., 2006; Mueller et al., 2011; Berestycki et al., 2012)

$$p_{c_\eta}(z) \propto \frac{W}{\pi}\sin\left(\frac{\pi z}{W}\right)e^{-c_\infty z/\sigma^2}, \tag{A.4}$$

where in these coordinates the 'front' of the wave is at $z = W = (\log\eta + 3\log\log\eta)\ell/2$.

Note that η has to be extremely large for this approximation to be accurate. Asymptotically, the minimum speed non-negative solution to the deterministic Fisher-KPP equation (1) tends to $e^{-\gamma_\infty x}$ with $\gamma_\infty = 2/\ell$. However, choosing coefficients such that $\gamma_\infty = 1$ and solving numerically, we see that even when $p(x) = 0.01$ the gradient of $\log p(x)$ is only 0.776 and when $p(x) = 0.0001$ it is only 0.895. Consequently, although the form of the logarithmic correction to the wavespeed in (5) is accurate, the constants in the asymptotic form (A.4) are quite misleading.

Appendix B. The coalescence rate within the front

We should like to understand the dependence of our approximation

$$\lambda = \int_0^\infty \frac{1}{\rho p_{c_\eta}(x)} f^2(x) dx, \tag{B.1}$$

where f is determined by (7), for the rate of coalescence within the front, on the parameters η and ℓ.

Let $\tilde{p}_{c_0}(x)$ solve (A.1) with $\sigma^2 = 2$, $s = 1$ and $c = c_0 = 2(1 - A/(\log \eta)^2)$. Then setting

$$p(x) = \tilde{p}_{c_0}\left(\frac{\sqrt{2s}}{\sigma} x\right),$$

we find that p solves (A.1) with $c = \sigma c_0 \sqrt{s/2}$. This tells us, in particular, that, in an obvious notation, the stationary distributions determined by (7) for these parameter values are related through $f(x) = (\sqrt{2s}/\sigma)\tilde{f}(x\sqrt{2s}/\sigma)$. From this we see that

$$\lambda = \int_0^\infty \frac{1}{\rho p_{c_\eta}(x)} f^2(x) dx = \int_0^\infty \frac{1}{\rho \tilde{p}_{c_0}(x\frac{\sqrt{2s}}{\sigma})} \frac{2s}{\sigma^2} \tilde{f}^2\left(x \frac{\sqrt{2s}}{\sigma}\right) dx$$
$$= \frac{\sqrt{2s}}{\rho\sigma} \int_0^\infty \frac{1}{\tilde{p}_{c_o}(x)} \tilde{f}^2(x) dx. \tag{B.2}$$

The integral on the right hand side is a function of η alone which we denote by $g(\eta)$ and then the coalescence rate becomes

$$\lambda = \frac{\sqrt{2s}}{\rho\sigma} g(\eta) = \frac{2}{\rho\ell} g(\eta).$$

Appendix C. Averaging over randomly distributed lineages

Appendix C.1. One dimension

We first derive equation (8) which gives the mean square displacement of a lineage at x due to a mutation that occurs at 0. The displacement, which we denote by X, is exponentially distributed with parameter r/c, but truncated at x. Thus

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^x \frac{r}{c} y^2 e^{-ry/c} dy + x^2 e^{-rx/c} \\
&= \left[-y^2 e^{-ry/c} \right]_0^x + \int_0^x 2y e^{-ry/c} dy + x^2 e^{-rx/c} \\
&= \left[-2y \frac{c}{r} e^{-ry/c} \right]_0^x + \int_0^x 2\frac{c}{r} e^{-ry/c} dy \\
&= -2\frac{xc}{r} e^{-rx/c} + 2\frac{c^2}{r^2} \left(1 - e^{-rx/c} \right) \\
&= 2\left(\frac{c}{r}\right)^2 \left(1 - e^{-rx/c} \left(1 + \frac{xr}{c} \right) \right).
\end{aligned}$$

We now assume that sweeps are uniformly disributed over a genetic map of length R and that they originate at points that are uniformly distributed over a one-dimensional range of length L. The mean square displacement per unit time of a lineage due to hitchhiking can be obtained by averaging equation (8) over $r \in [0, R]$ and $x \in [0, L]$. Averaging first over $x \in [0, L]$ we obtain

$$\frac{2L^2}{\theta^4} \left(6 - 4\theta + \theta^2 - 2\ e^{-\theta}(3 + \theta) \right),$$

where $Lr/c = \theta$. Multiplying by the rate of sweeps, Λ, and averaging over a genetic map, length $R \gg 1$ gives

$$\sigma^2_{\text{eff}} = \frac{4}{3} cL \frac{\Lambda}{R}.$$

Dividing by the rate of gene flow with no sweeps, σ^2, and substituting $c/2\sigma^2 = 1/\ell$, we have

$$\frac{\sigma^2_{\text{eff}}}{\sigma^2} = \frac{8}{3} \frac{L}{\ell} \frac{\Lambda}{R}.$$

Now, consider a sweep which originates with a favourable mutation arising at z, and two lineages sampled at x_1, x_2. If these are on opposite sides of the sweep ($x_1 < z < x_2$ or vice-versa), then they can coalesce only at the origin, which will happen with probability $e^{-r|x_2-x_1|/c}$. If they are on the same side ($x_1 < x_2 < z$, say), then they coalesce with probability (approximately)

$$\frac{\lambda}{\lambda+2r}\left(1-e^{-(\lambda+2r)\,(z-x_2)/c}\right)e^{-r(x_2-x_1)/c}.$$

Integrating over a uniform distribution of x_1, x_2, z on $[0, L]$, the probability of coalescence is:

$$\begin{aligned}
P = \frac{1}{L^3}\int_0^L \Big(2\int_0^z\int_z^L & e^{-r\,(x_2-x_1)/c}dx_2dx_1 \\
+4\int_0^z\int_0^{x_2} & \frac{\lambda}{\lambda+2r}\left(1-e^{-(\lambda+2r)\,(z-x_2)/c}\right)e^{-r\,(x_2-x_1)/c}dx_1dx_2\Big)\,dz \\
&= \frac{2}{\theta^3}\Big(\frac{2e^{-\theta(2\omega+1)}}{(1+\omega)(1+2\omega)^3}+\frac{e^{-\theta\omega}(2+\theta(1+\omega))}{\omega^2(1+\omega)} \\
&+\frac{4\theta\;\omega^3(2+\theta)-\;(2+\theta)+4\omega^2\left(\theta^2-4\right)+\omega\;(\theta(\theta-4)-10)}{\omega^2(1+2\omega)^3}\Big) \qquad \text{(C.1)}
\end{aligned}$$

where $\omega = r/\lambda$, $\theta = L/c$ and where the factors of 2, 4 in the integral arise from summing over the distinct orders of x_1, x_2, z. Integrating over a long map of length R, the mean rate of coalescence due to sweeps

$$\frac{2\Lambda}{R}\int_0^\infty P dr = \frac{2c\Lambda}{LR}\,h\Big(\frac{L\lambda}{c}\Big),$$

where

$$\begin{aligned}
h(\theta) = \frac{2}{\theta^2}\Big\{&\left(1-\theta^2\right)-(1+\theta)e^{-\theta} \\
+&\Big(2e^{\theta}(\mathrm{Ei}[-2\theta]-\mathrm{Ei}[-\theta])-\left(2+2\theta+\theta^2\right)\Big(\mathrm{Ei}[-\theta]-\gamma-\log\Big(\frac{\theta}{2}\Big)\Big)\Big)\Big\}, \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(C.2)}
\end{aligned}$$

$\mathrm{Ei}[z]$ is the exponential integral, and γ is Euler's γ. The function $h(\theta)$ tends to 1 for small θ, and to $2\,(\log(\theta/2)+\gamma-1)$ for large θ.

Now we derive the expression (13) for the probability of coalescence of a randomly chosen pair of neutral genes due to the passage of a single sweep. It is obtained by averaging the probability of coalescence of a pair of genes sampled at points x_1 and x_2, which we obtained in §4.2.3, with respect to a uniform sampling distribution. Ignoring the constant factor $\lambda/(\lambda+2r)$, we have

$$\begin{aligned}
\frac{1}{L^2}\int_0^L\int_0^L e^{-r|x_1-x_2|/c}dx_1dx_2 &= \frac{2}{L^2}\int_0^L (L-y)e^{-ry/c}dy \\
&= \frac{2}{L^2}\int_0^L\int_y^L e^{-ry/c}dxdy \\
&= \frac{2}{L^2}\int_0^L\int_0^x e^{-ry/c}dydx \\
&= \frac{2}{L^2}\int_0^L \frac{c}{r}\left(1-e^{-rx/c}\right)dx \\
&= \frac{2}{L}\frac{c}{r}-\frac{2}{L^2}\frac{c^2}{r^2}\left(1-e^{-rL/c}\right) \qquad \text{(C.3)} \\
&\approx \frac{2}{L}\frac{c}{r},
\end{aligned}$$

since $rL \gg c$ by assumption. The result now follows on multiplying by $\lambda/(\lambda+2r)$.

Appendix C.2. Two dimensions

Let the distribution of the time to coalescence of two favoured alleles sampled at separation $|y|$ from within the wavefront be $\Psi(t,|y|)$. Just as in our one-dimensional calculations, we suppose that the time until a lineage sampled at a distance x orthogonal to (and behind) the front is caught by the wave is approximately x/c and let us suppose that at the moment when they are first caught, their separation transverse to the wave is $|y|$. (In Fig. 10 we assumed that up to that time they each followed independent diffusions with rate σ^2 transverse to the wave as in (14), but this assumption is not important for what follows.) Then the probability of coalescence of two neutral genes sampled at distances x_1 and x_2 orthogonal to the front and with a separation transverse to the front of $|y|$ is

$$\int_0^\infty \Psi(t,|y|)e^{-2rt}e^{-r|x_1-x_2|/c}dt.$$

Averaging over random locations x_1 and x_2 and using (C.3) we obtain

$$\begin{aligned}
&\frac{1}{L^2}\int_0^L\int_0^L\int_0^\infty \Psi(t,|y|)e^{-2rt}e^{-r|x_1-x_2|/c}dtdx_1dx_2 \\
&\qquad\qquad = \frac{2c}{L^2r^2}\left(ce^{-rL/c}-c+rL\right)\int_0^\infty \Psi(t,|y|)e^{-2rt}dt.
\end{aligned}$$

Integrating over a linear genome we obtain

$$\frac{2}{R}\int_0^R\int_0^\infty \Psi(t,|y|)e^{-2rt}\frac{2c}{L^2r^2}\left(ce^{-rL/c}-c+rL\right)dtdr$$
$$\approx \frac{4c}{Lr}\int_0^\infty \Psi(t,|y|)\left(\left(1+\frac{2ct}{L}\right)\log\left(1+\frac{L}{2ct}\right)-1\right)dt. \quad \text{(C.4)}$$

Let us write $\tau(|y|)$ for the coalescence time of two favoured lineages sampled at separation $|y|$ from within the wavefront, then (C.4) can be written as an expectation:

$$\frac{4}{RT}\mathbb{E}\left[\left(1+\frac{2\tau(|y|)}{T}\right)\log\left(1+\frac{T}{2\tau(|y|)}\right)-1\right], \quad \text{(C.5)}$$

where $T = L/c$ is, approximately, the time that it takes for a sweep to cross the whole range. The comparable quantity for a classic sweep through a panmictic population is

$$\frac{2}{R}\int_0^R e^{-2rT^*}dr \approx \frac{1}{RT^*},$$

where $T^* = (\log(2Ns))/s$. Even if genes that hit the front come together and coalesce quickly ($\tau(|y|) \ll T$), it is hard for the expression in (C.5) to exceed $1/(RT^*)$. Thus hitchhiking will be much less effective in a population distributed across two spatial dimensions than in a panmictic population.

Appendix D. Some classical results

For ease of reference, we record two classical results.

Appendix D.1. The Wright-Malécot formula in one dimension

We calculate the Laplace transform of the time to coalescence of individuals sampled at separation x from an infinite one-dimensional range. This is equivalent to assuming an infinitely many alleles mutation model and asking for the probability, $\phi(x)$ that two individuals sampled at separation x are identical in state. In the absence of sweeps this is given by the classical Wright-Malécot formula (Malécot, 1948). In one spatial dimension this takes the form

$$\phi(x) = \frac{1}{1+4\rho\sigma\sqrt{2\mu}}\exp\left(-\frac{|x|}{l}\right) \quad \text{(D.1)}$$

where ρ is the local population density and $l = \sigma/\sqrt{2\mu}$.

Appendix D.2. Sweeps in panmictic populations

In the classical model of a sweep through a panmictic population, we suppose that an allele starts in a single copy and increases to fixation. We consider a haploid population of N genes and suppose that $Ns \gg 1$. Allele frequencies are assumed to follow the Wright-Fisher diffusion with selection (corresponding to the non-spatial version of (2)). The time for the allele to reach high frequency ($p = 0.5$ say), is $t \sim \frac{1}{s}\log(2Ns)$. (Note that this includes a stochastic acceleration by a factor $1/(2s)$ due to conditioning on fixation.) For most of this time, the favoured allele is at very low frequency. The additional time to fixation is dominated by the period when the allele frequency is close to one. Suppose that we take a sample of size two from the population and ask for the probability that they coalesce before one of them escapes the sweep through recombination. This is only going to be possible during the period when allele frequencies are low. Assuming that the lineages coalesce only at the time of origin of the sweep, the chance that neither escapes the sweep through recombination is just $\exp(-2rt) \sim (2Ns)^{2r/s}$. This approximation is accurate as long as $Ns \gg 1$.