Ancestry in adapting, spatially-extended populations

Daniel B. Weissman
Dept. of Physics, Emory University, Atlanta, GA 30322

Abstract

Selective sweeps affect neutral genetic diversity through hitchhiking. While this effect is limited to the local genomic region of the sweep in panmictic populations, we find that in spatially-extended populations the combined effects of many unlinked sweeps can affect patterns of ancestry (and therefore neutral genetic diversity) across the whole genome. Even low rates of sweeps can be enough to skew the spatial locations of ancestors such that neutral mutations that occur in an individual living outside a small region in the center of the range have virtually no chance of fixing in the population.

Introduction

In large populations even a fairly low rate of selective sweeps is sufficient to reduce diversity across most of the genome via hitchhiking [1, 2]. Most modeling of the effects of hitchhiking on diversity has considered well-mixed populations. However, the effects are potentially quite different in spatially-extended populations, because instead of quickly fixing through logistic growth, sweeps must spread out in a spatial wave of advance over the whole range [3]. [4] recently showed that this increase in the time to sweep tends to reduce the size of the genomic region over which diversity is depressed by a sweep.

While the effect of sweeps on genetic diversity at linked loci is therefore reduced by spatial structure, we show here that collective effect of sweeps on the diversity at *unlinked* loci can be much stronger than in panmictic populations. Surprisingly, this effect is dependent on the geometry of the range – it only appears for realistic range shapes with relatively well-defined central regions, not for the perfectly symmetric idealizations of ring-shaped and toroidal ranges often used in theoretical models. In particular, we find that probability of fixation of an allele can be strongly position-dependent, with alleles near the center of the range orders of magnitude more likely to fix than those at typical locations. This can produce a false signal of a population expansion (in number, not space) even at loci on chromosomes where no adaptation is taking place. The basic mechanism driving this effect is that sweeps tend to arrive at each location from the direction of the center of the range, and so bias the ancestry back towards the center.

Methods

We wish to find the expected number of copies that an allele found in an individual at spatial position $\mathbf{x}$ will leave far in the future, i.e., its reproductive value [5], which we denote $\phi(\mathbf{x})$. Equivalently, $\phi(\mathbf{x})\rho(\mathbf{x})$, where ρ is the population density, is the probability density of a present-day individual's ancestor being at location $\mathbf{x}$ at some time in the distant past. [6] showed that in the absence of selection, $\phi(\mathbf{x}) \equiv 1$ regardless of the details of the population structure, as long as dispersal does not change expected allele frequencies. Here we show that this result does not extend to populations undergoing selection. Populations living in perfectly symmetric ranges (circles in one dimension, tori in two) necessarily have $\phi(\mathbf{x}) \equiv 1$, but when this symmetry is broken, recurrent sweeps can create a substantial imbalance, making ϕ much higher in a small region in the center of the range and decreasing it everywhere else.

Model

We consider a population with uniform, constant density ρ distributed over a d-dimensional range with radius L, with uniform local dispersal with diffusion constant D. We assume that selective sweeps with advantage s occur in the population at a rate Λ per generation, originating at points uniformly distributed over time and space. As long as the density is sufficiently high ($\rho \gg (s/D)^{d/2}/s$, [7, 4]), they will spread roughly deterministically in waves with speed $c \approx 2\sqrt{Ds}$ with characteristic wavefront width $l \approx \sqrt{D/s}$ [3], which we take to be much smaller than the range size, $l \ll L$. We assume that Λ is low enough compared to the frequency of outcrossing, f, and the average number of crossovers per outcrossing, K, that the waves do not interfere with each other. The definitions of symbols are collected in Table 1.

One and two dimensions

We consider both one-dimensional ranges (lines with length $2L$) and two-dimensional ranges. In two dimensions, the shape of the range will have some effect on many of our results; however, as long as the shape is fairly "nice", with a clear center and single characteristic length scale L, this effect will be modest. We will therefore ignore it for simplicity. For our purposes, the main difference between one and two dimensions will be in the density of individuals a distance x from the center, $\rho(x)$. Since we are assuming a uniform spatial density, in one dimension this is just ρ, a constant. In two dimensions, however, we must account for the fact that there is more area at larger x, and thus $\rho(x) \approx 2\pi x\rho$. (Obviously, $\rho(x > L) = 0$ in both one and two dimensions.)

Results

A sweep at a tightly-linked locus a genetic map distance $r \ll s$ away pulls a lineage a distance that is approximately exponentially distributed with mean c/r, going backwards in time. This is only approximate, because actually there is an upper cutoff at the distance to the origin of the sweep [4]. For sweeps at loosely-linked loci with $r \gg s$, the lineage is only pulled an expected distance $c/2r$ (see Appendix A). However, since the average displacement still only falls off like $1/r$, and since there is an upper cutoff on the effect of sweeps as $r \to 0$, the total average displacement for a typical locus can be dominated by the effect of the many unlinked sweeps rather than the few linked ones, assuming that sweeps are uniformly distributed over the genome, and that the genome is long. We will make the approximation that the expected displacement is solely due to unlinked sweeps, with $r = f/2$; we will consider the additional effects of the rare, tightly-linked sweeps below.

For a lineage a distance x from the center, there is an excess of approximately $\sim \Lambda x/L$ sweeps per generation pulling it back toward the center, each of which pulls it an expected distance c/f. (Note that the effect of the upper cutoff on the displacement from these sweeps is negligible as long as $L \gg c/f$.) The expected distance from the center therefore decays exponentially (backwards in time) like

$$\overline{x} \approx x_0 \exp\left(-\frac{\Lambda c}{Lf}t\right). \tag{1}$$

This implies that there is a time t_{con} before which individuals' ancestors are unlikely to be found outside the center of the range, with

$$t_{\text{con}} = \frac{Lf}{\Lambda c}. \tag{2}$$

This deterministic move back to the center is opposed by dispersal, and also by the effect of occasional tightly-linked sweeps which pull the lineage a distance $\sim L$, effectively randomizing its position. The balances between these forces means that the ancestry of the population is not completely concentrated at the center of the range, but is instead distributed around it in some region of size $\sim x_c$.

Balance with dispersal

If tightly-linked sweeps are relatively rare, either because the overall rate of sweeps is low or because the focal locus lies in a region of the genome that is not undergoing much adaptation, the main balance will be

between the diffusive effect of dispersal and the pull of unlinked sweeps. In this case, the position of the ancestry is an Ornstein-Uhlenbeck process. The stationary distribution is therefore normal and concentrated in the center of the range according to:

$$\phi(x) \propto \exp\left(-\frac{x^2}{2x_c^2}\right), \text{ with } \frac{x_c}{L} = \sqrt{\frac{fl}{2\Lambda L}}. \tag{3}$$

If $x_c \ll L$, then the reproductive value of an individual at the center of the range can be orders of magnitude higher than than one at a typical distance $\sim L/2$ from the center (Fig. 1).

From Eq. (3), we see that the ancestral range will be substantially reduced by selection if the rate of sweeps per sexual generation is greater than the ratio of the cline width to the range size: $\Lambda/f > l/L$. It is unclear what ranges these ratios take in natural populations. $\Lambda/(fK)$ is unlikely to be much more than $\mathcal{O}(1)$ [2], but in organisms with many chromosomes (large K), Λ/f may be substantial. Looking at the right-hand side of the inequality, modeling sweeping alleles by waves spreading across the range necessarily requires $l/L \ll 1$, so even small values of Λ/f may be enough to distort the distribution of ancestry. Surprisingly little is known about typical values of l for the waves of advance of sweeping alleles in nature, but it seems plausible that for many species it should be much smaller than the total species range [3]. For the spread of insecticide resistance in *Culex pipiens* in southern France, the width of the wave of advance was ~ 20 km [8], much smaller than the global scale of the species range, but the dynamics were more complex than a simple selective sweep [9]. Much more is known about the width of stable clines and hybrid zones, which are frequently much smaller than species ranges [10]. To the extent that the selection maintaining them is comparable in strength to the selection driving sweeps, these should have roughly the same width as the wavefronts.

Balance with tightly-linked sweeps

Finding the balance between concentrating effect of unlinked sweeps and the randomizing effect of tightly-linked sweeps is slightly trickier. Indeed, finding an exact expression for $\phi(x)$ is intractable. However, we can find an approximate expression by using the fact that the mean squared displacement of the ancestral lineage due to linked sweeps is dominated by rare very tightly-linked sweeps rather than the many loosely-linked ones [4]. This suggests that for large x, the probability that an individual's ancestor was farther than x from the center at time t_0 in the distant past is roughly just the probability that a single very tightly-linked sweep pulled it there at some time within $\sim t_{\text{con}}$ generations of t_0. Since the distance that a sweep at recombination fraction r pulls the lineage goes like $1/r$, the rate of sweeps close enough on the genome to pull the ancestry a distance of at least x falls off like $1/x$. Therefore, the probability of finding the ancestry at a distance of at least x should also fall off like $1/x$; the probability density of being exactly at x, $\phi(x)\rho(x)$, should then fall off like $1/x^2$.

In the appendix, we calculate this more formally, and find

$$\phi(x)\rho(x) \approx \frac{2L\left(1-(x/L)^d\right)}{Kx^2} \text{ for } x \gg x_c = 2L/K. \tag{4}$$

The factor $1-(x/L)^d$ (where $d = 1$ or 2 is the dimension of the habitat) reflects the fact that for very large x, $x \sim L$, most sweeps start at distances less than x and cannot pull the lineage that far from the center. For $x \ll x_c = 2L/K$, lineages will tend to experience many sweeps pulling them distances greater than x in time $\sim t_{\text{con}}$, so the approximation used to derive Eq. (4) breaks down; for these small values of x, the randomizing effects of moderately-linked sweeps smooth out $\phi(x)$ and make it roughly constant.

Barton et al. describe the randomizing effect of tightly-linked sweeps by "D_{eff}," an effective dispersal rate, with $D_{\text{eff}} \approx \frac{16L\Lambda}{3lKf}D$ (Eq. (9) of [4]). Comparing Eqs. (3) and (4), however, we see that their effect cannot simply be described as an increase in the dispersal rate, since they create a much longer tail in the spatial distribution of ancestry. Because of this, it is possible that while the bulk of the distribution of ancestry is determined by a balance between unlinked sweeps and dispersal, with linked sweeps too rare to make a difference, linked sweeps make the dominant contribution to the tails of the ancestry distribution (Fig. 3).

Combining dispersal and tightly-linked sweeps

Combining Eqs. (3) and (4), we see that unlinked sweeps reduce the effective size of the ancestral range by a factor x_c/L:

$$\frac{x_c}{L} \approx \max\left\{\sqrt{\frac{fl}{2\Lambda L}}, \frac{2}{K}\right\}. \tag{5}$$

For typical numbers of chromosomes K, it would seem that ancestry could be concentrated by about an order of magnitude. However, the result $2/K$ was derived under the assumption that sweeps are distributed uniformly across the genome. If, on the other hand, adaptation is mostly occurring in just a few genes, the rest of the genome will not experience any tightly-linked sweeps, and ordinary dispersal will be the only force counteracting the concentration, meaning that the effect could potentially be much stronger. This has the surprising implication that selection can have a stronger effect on some features of the spatial distribution of ancestry at far-away loci than at those nearby.

Effect on diversity

While the effect of recurrent sweeps on neutral diversity can be quite large, detecting the effect in data from real populations may be tricky. It might seem to be indistinguishable from a range expansion in the absence of time-series data, but there is a simple way to tell them apart: under recurrent sweeps, there is no serial founder effect reducing diversity away from the center. One way to see this is by looking at isolation by distance. The probability $\psi(x)$ that two individuals separated by a distance x are genetically identical can be written in terms of the neutral mutation rate μ and their coalescence time T as

$$\psi(x) = E\left[e^{-2\mu T} \mid x\right]. \tag{6}$$

For x large compared to the size of a single deme (i.e., the spatial scale over which individuals interact within a generation) and loci far on the genome from any recent sweeps, there are two simple regimes for Eq. (6). If $x \ll x_c$, then we expect that the pull due to sweeps should be unimportant, and $\psi(x)$ is just given by the neutral value, $\psi(x) \propto x^{(1-d)/2} e^{-\sqrt{\mu/D}x}$ [11], which says that the probability of identity falls off rapidly with distance. On the other hand, larger values of x are quickly collapsed by the pull of sweeps in time $\sim \log(x)/t_{\text{con}}$, so we expect that ψ should be of the form $\psi(x) \propto x^{-2\mu t_{\text{con}}}$. A detailed calculation in Appendix C confirms that this is true for $x \gg x_c\sqrt{2 + 4\mu t_{\text{con}}}$. The probability of identity thus has a long tail in distance – individuals at opposite sides of the range (separated by $\approx 2L$) are nearly as related as individuals separated by, say, $L/2$. Notice that ψ does not depend on from where in the range we sampled the pair of individuals. This implies that, while reproductive value is concentrated in the center of the range, genetic diversity is more evenly spread, distinguishing this scenario from a range expansion.

Above, we have ignored loci that are close to recent sweeps. If we are considering large enough loci so that $\mu t_{\text{con}} \gg 1$, then usually only these recently swept regions will be identical between individuals from different parts of the range. In this case, because each sweep causes coalescence between individuals separated by a large distance x over a region of genome with length $r \propto 1/x$ [4], ψ should still have a long tail, but with an exponent that is independent of the population parameters, $\psi \propto 1/x$ (see Appendix C.1). This characteristic exponent is another effect of rare, tightly-linked sweeps that cannot be accounted for by any effective dispersal rate D_{eff}.

Discussion

Because selection and demography are often difficult or impossible to measure directly in natural populations, both are typically inferred from patterns of genetic diversity. This inference can be difficult, because the two processes can produce similar signals. For instance, both purifying selection and population expansion tend to produce site frequency spectra with a relative excess of rare alleles. In order to tease apart the two factors, demography is often first inferred using data from loci that are thought to be neutral, and then the answer is used to infer the pattern of selection at the remaining loci. However, in order for the demography to be inferred correctly, this method requires that the first set of loci be not just neutral, but also unaffected by

selection at linked loci. Typically, this is done by using loci that are far from sites where selection is thought to have been important (e.g., [12]). Our results suggest that this may be problematic in spatially-structured populations – even diversity at these loci may be strongly affected by unlinked sweeps.

Geometry, not topology, of range is important

Our results might seem to show that the genetic diversity in a population depends sensitively on the topology of the range and can therefore change drastically as the result of small perturbations to the environment. For example, a circular range (which has no concentration of ancestry since it is perfectly symmetric) can be transformed into a linear one (with very concentrated ancestry) by removing a single point. However, this is a misleading interpretation. In fact, a "circular" range is an annulus with radius large compared to its thickness (Fig. 4a). A small perturbation that slightly reduces the population in one part of the range will only have a correspondingly small effect on the distribution of ancestry (Fig. 4b), and the bias of the ancestry ancestry increases smoothly as the perturbation grows (Fig. 4c), until the annulus is completely pinched off (Fig. 4d). More generally, the common-sense intuition that the pattern of diversity should not depend on the details of the shape of the range is correct. All that matters is the extent to which there is a "central region" that sweeps tend to pass through on their way to dominating the population.

Extensions

We have focused on a very simple population model. Here we consider several possible modifications. First, we have assumed that the density ρ is constant in time. If density fluctuations typically occur on timescales longer than t_{con}, this approximation should be accurate, and if they are rapid compared to the sweep time L/c they should average out, but it is unclear how fluctuations on moderate timescales should interact with dynamics discussed here.

We have also neglected the possibility of rare long-range dispersal. Tightly-linked sweeps already effectively produce occasional long-range jumps in the ancestry of neutral sites, so adding long-range dispersal might not have a large direct effect, but it is likely to have dramatic effects on how sweeps spread [13], and therefore a large indirect effect on the hitchhiking dynamics. It is not at all clear what this effect should be – on the one hand, the sweeps will spread faster, increasing their pull, but on the other hand, the direction of that pull may be less reliably towards the center.

We have also neglected the possibility that many sweeps may be "soft", starting from multiple alleles [14]. If these alleles typically descend from a recent single ancestor, i.e, are concentrated in a small region at the time when they begin to sweep, then the results should be essentially unchanged, with the possible exception of the coalescent effects of tightly-linked sweeps (Appendix C.1). The same should be true if sweeps are "firm", i.e., multiple mutant lineages contribute to each sweep, but the most successful one typically colonizes most of the population. But sweeps in which many widely-spread mutations contribute equally would likely not consistently concentrate ancestry in space.

We have focused on the effect of sweeps on neutral variation, but they will of course also affect selected alleles. Most obviously, if recombination is limited they will interfere with each other [15]. They will interfere even more strongly with weakly-selected variants. We are currently preparing a manuscript addressing these issues. It is also important to consider how the concentration of reproductive value interacts with spatially-varying selection. It seems plausible that it would reduce the potential for local adaptation and thereby limit population ranges.

Acknowledgements

This work would not have been possible without Nick Barton's generous assistance at all stages.

Tables

Table 1: **Symbol definitions**

Symbol	**Definition**
ρ	Density of individuals
L	Radius of range
D	Dispersal constant
f	Frequency of outcrossing
K	Expected number of crossovers per outcrossing
s	Selective advantage of adaptive alleles
Λ	Rate of adaptive substitutions
$c \approx 2\sqrt{Ds}$	Expected rate of advance of a sweeping beneficial allele
$l \approx \sqrt{D/s}$	Characteristic width of the wave of advance of a sweeping beneficial allele

The definitions of the main symbols used in the text.

Figures

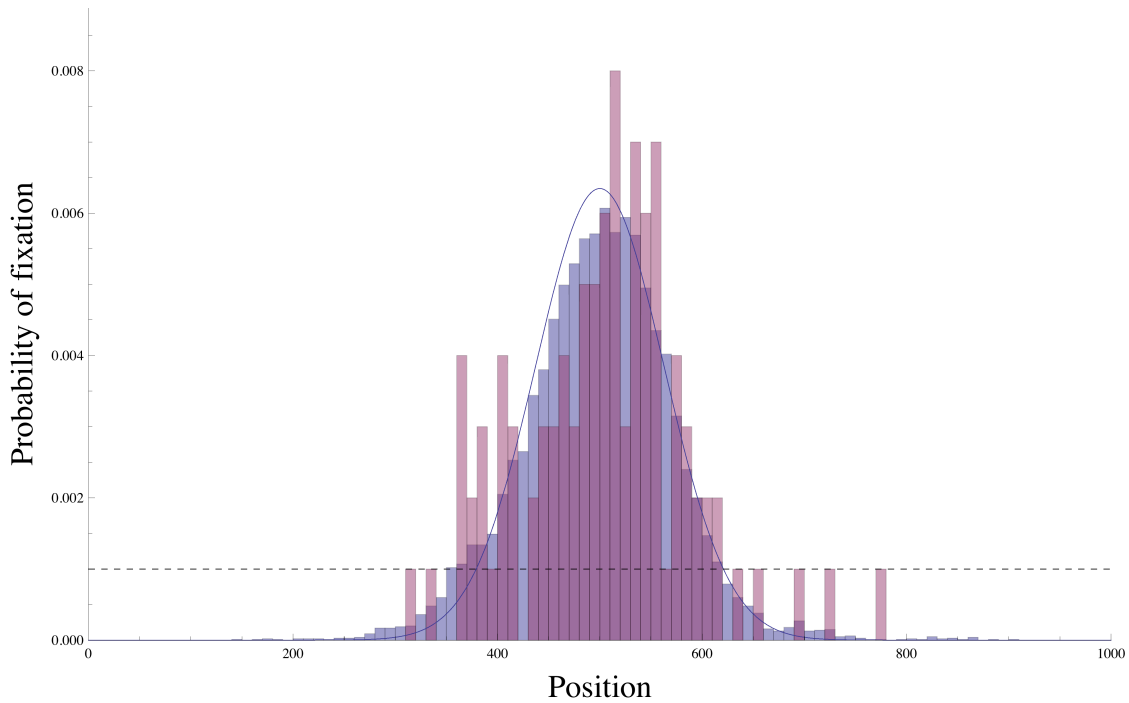

Figure 1: Unlinked sweeps concentrate ancestry in the center of the range. The plot shows the long-term probability of finding the ancestor of a neutral allele at a given deme in a linear range. The purple histogram shows full-population forward simulations, the blue histogram shows approximate backwards simulations, and the blue curve shows the predicted distribution, Eq. (3). The dotted black line shows a uniform distribution. Parameters are $L = 500, \rho = 100, s = 0.05, D = 0.125, \Lambda = 0.1, f = 1, K = 100$; see Appendix D for simulation methods.

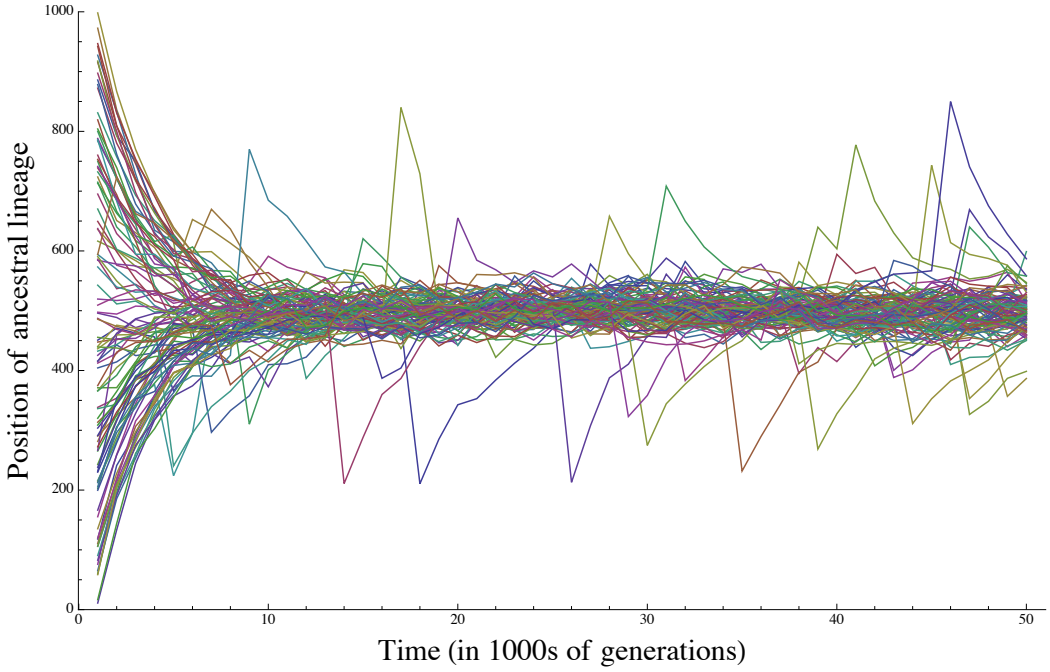

Figure 2: Tracing the ancestry of an individual back through time. Individuals throughout the range rapidly trace their ancestry at a neutral locus back to a small region in the center of the range. The width of this region is determined by strength of dispersal. The occasional excursions out of the center are due to hitchhiking on tightly-linked sweeps. Parameters are $L = 500, s = 0.05, D = 0.125, \Lambda = 1, f = 1, K = 300$. Each curve is an independent simulation.

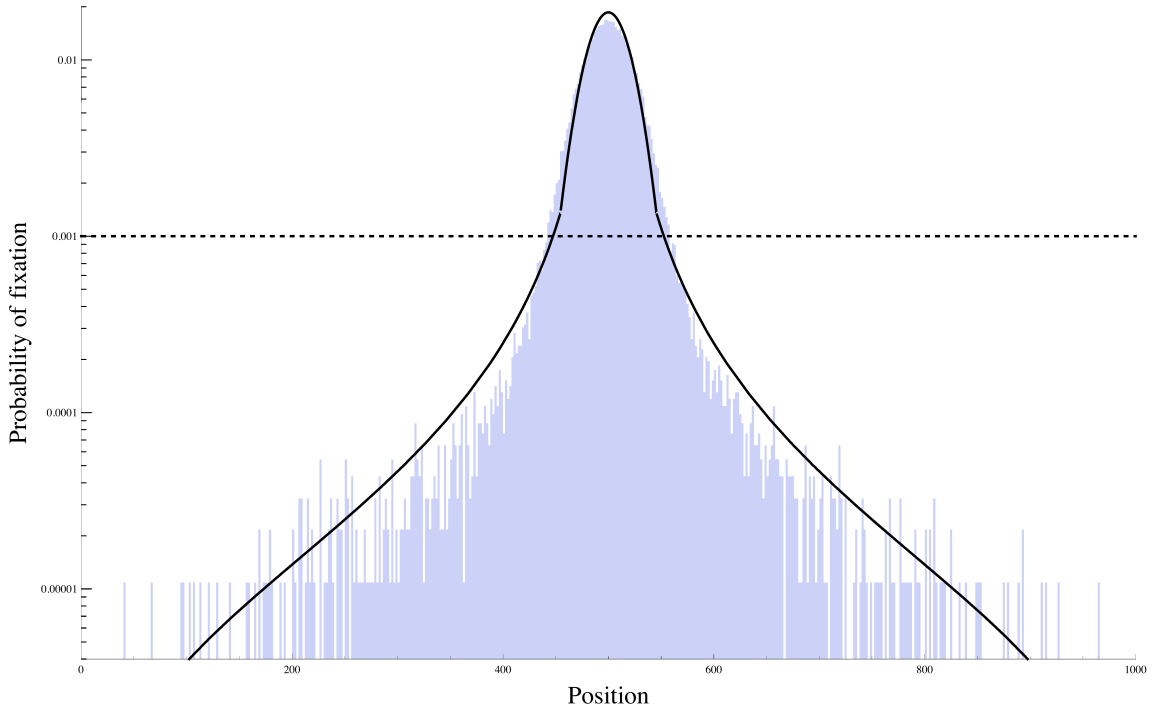

Figure 3: The long-term probability of finding the ancestor of a neutral allele at a given deme in a linear range. The blue histogram shows approximate backwards simulations, the dotted black line shows a uniform distribution, and the blue curve shows the predicted distribution, Eqs. (3) and (4). Typical loci are in the center of the range, in a balance between the pull of unlinked sweeps and dispersal (Eq. (3)), but the (small) probability that the ancestor is found outside the center of the range is dominated by the probability that there has recently been a tightly-linked sweep (Eq. (4)). Parameters are as in Fig. 2.

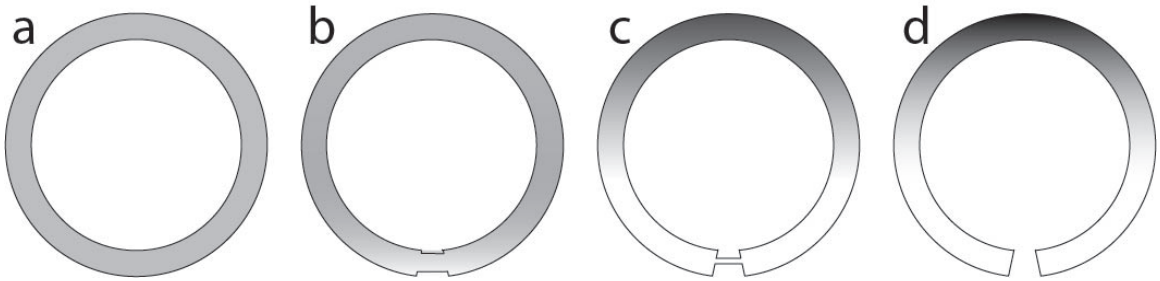

Figure 4: The distribution of ancestry depends smoothly on the shape of the range as it deforms from a perfectly symmetric circle (a) to a curve with endpoints (d). Shading represents the reproductive value of each location, from high (dark) to low (light). In (a), the ancestry is necessarily evenly distributed. Slight asymmetries in the shape of the range introduce slight differences in the distribution of ancestry (b). When the range has a well-defined middle (d), the ancestry is concentrated there; including a weak connection between the end of the range (c) does not change this much.

References

[1] Gillespie JH (2000) Genetic drift in an infinite population: the pseudohitchhiking model. Genetics 155: 909–919.

[2] Weissman DB, Barton NH (2012) Limits to the rate of adaptive substitution in sexual populations. PLoS Genetics 8: e1002740.

[3] Fisher RA (1937) The wave of advance of advantageous genes. Annals of Eugenics 7: 355-369.

[4] Barton NH, Etheridge AM, Kelleher J, Véber A (2013) Genetic hitchhiking in spatially extended populations. Theoretical Population Biology 87: 75–89.

[5] Barton NH, Etheridge AM (2011) The relation between reproductive value and genetic contribution. Genetics 188: 953-973.

[6] Maruyama T (1970) On the fixation probability of mutant genes in a subdivided population. Genet Res 15: 221-225.

[7] Nagylaki T (1978) Random genetic drift in a cline. Proceedings of the National Academy of Sciences 75: 423–426.

[8] Lenormand T, Bourguet D, Guillemaud T, Raymond M (1999) Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. Nature 400: 861–864.

[9] Labbé P, Berticat C, Berthomieu A, Unal S, Bernard C, et al. (2007) Forty years of erratic insecticide resistance evolution in the mosquito *Culex pipiens*. PLOS Genetics 3: e205.

[10] Barton NH, Hewitt GM (1985) Analysis of hybrid zones. Annual review of Ecology and Systematics 16: 113-148.

[11] Barton NH, Depaulis F, Etheridge AM (2002) Neutral evolution in spatially continuous populations. Theoretical Population Biology 61: 31–48.

[12] Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLoS Genetics 7: e1001302.

[13] Hallatschek O, Fisher DS (2014) Acceleration of evolutionary spread by long-range dispersal. Proceedings of the National Academy of Sciences 111: E4911-4919.

[14] Hermisson J, Pennings PS (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335–2352.

[15] Martens EA, Hallatschek O (2011) Interfering waves of adaptation promote spatial mixing. Genetics 189: 1045–1060.

[16] Pham H (2009) Continuous-Time Stochastic Control and Optimization with Financial Applications. Stochastic modelling and applied probability. Springer-Verlag.

[17] Wolfram Research (2016). Hermite function. URL `http://functions.wolfram.com/07.01.02.0001.01`. Online; accessed 30-Oct-2016.

[18] Kim Y, Stephan W (2003) Selective sweeps in the presence of interference among partially linked loci. Genetics 164: 389–98.

A Calculating the "pull" of an loosely-linked sweep

We would like to find the expected spatial displacement of a lineage caused by an loosely-linked sweep, tracing backwards in time. To do so, suppose that we sample an allele in a present-day individual in the middle of a very large one-dimensional range, and that a long time ago a selective sweep occurred at a locus a recombination fraction r away from the focal allele, starting a very long distance away from our sample. We wish to find the expected location of the ancestor of the sampled allele before the sweep began. Let $p(x,\tau)$ be the probability density for finding the ancestor at location x τ generations in the past, with $x=0$ corresponding to the present location. We want to find

$$\begin{aligned}\lim_{\tau\to\infty}\langle X\rangle &\equiv \lim_{\tau\to\infty}\int dx\, x p(x,\tau)\\ &= \int d\tau \int dx\, x\partial_\tau p(x,\tau).\end{aligned} \tag{7}$$

To find $\partial_\tau p$, first define $p_i(x,\tau)$ as the probability density that the ancestor was at location x and in genetic background i, where $i=0$ is the ancestral genetic background, and $i=1$ is the background with the allele that swept. (Note $p=p_0+p_1$.) If we define $u(x,\tau)\equiv u_1(x,\tau)$ and $u_0(x,\tau)$ to be the frequencies of the sweeping allele and the background allele, respectively, with $u_1+u_0=1$, p_i satisfies the partial differential equation

$$\partial_\tau p_i = r(u_i p_{1-i} - u_{1-i}p_i) + D\partial_x(\partial_x p_i - 2p_i\partial_x \log u_i). \tag{8}$$

Technically, in models with discrete generations, Eq. (8) only applies when the recombination rate per generation is small, but we will use it for unlinked loci anyway.

The equivalent of linkage disequilibrium in this system is $\Delta\equiv u_0p_1-u_1p_0$; we expect it to be small for large r. Using Δ to change variables back to p, Eq. (8) becomes

$$\partial_\tau p = D\partial_x\left(\partial_x p - 2\frac{\partial_x u}{u(1-u)}\Delta\right) \tag{9}$$

$$\partial_\tau\Delta = -r\Delta + D\partial_x^2\Delta - (\partial_\tau u + D\partial_x^2 u)p + 2(2u-1)D\partial_x\left(\frac{\partial_x u}{u(1-u)}\Delta\right) + 2D\frac{\partial_x u}{u(1-u)}\Delta \tag{10}$$

Plugging Eq. (9) into Eq. (7), we have

$$\begin{aligned}\lim_{\tau\to\infty}\langle X\rangle &= D\int d\tau\int dx\, x\partial_x\left(\partial_x p - 2\frac{\partial_x u}{u(1-u)}\Delta\right)\\ &= 2D\int d\tau\int dx\frac{\partial_x u}{u(1-u)}\Delta,\end{aligned} \tag{11}$$

where we have used integration by parts and the fact that $p(\pm\infty,\tau)=0$. It now remains to find an expression for Δ. Eq. (10) is quite complicated, but for large r we will have $\Delta\ll p$ and the dominant balance will be between the first and third terms on the right-hand side, giving

$$\Delta \approx -\frac{1}{r}(\partial_\tau u + D\partial_x^2 u)p_{\text{neut}}, \tag{12}$$

where p_{neut} is the value of p ignoring the perturbation caused by the sweep, i.e., $p_{\text{neut}} = \frac{1}{\sqrt{4\pi D\tau}}\exp\left(-\frac{x^2}{4D\tau}\right)$. We can simplify this further by noting that u solves the differential equation $\partial_\tau u + D\partial_x^2 u = -su(1-u)$. (Recall that τ is backwards time.) Using this relation and substituting Eq. (12) into Eq. (11), we have

$$\lim_{\tau\to\infty}\langle X\rangle = \frac{2Ds}{r}\int d\tau\int dx\,\partial_x u\, p_{\text{neut}}. \tag{13}$$

Recall that we are interested in the effect of a long-past sweep. Let τ_0 be the time at which the wave of advance passed the point where we sampled the allele; we will take τ_0 to be extremely large. At time τ_0, p_{neut} has width $\sim\sqrt{D\tau_0}$, so the wave crosses the region where the ancestor might have lived in a

time $\sim \sqrt{D\tau_0}/c \ll \tau_0$, and the integral in Eq. (13) is dominated by times τ in the approximate range $|\tau - \tau_0| \lesssim \sqrt{D\tau_0}/c$. Since τ does not vary by much (proportionately) in this interval, $p_{\text{neut}}(x,\tau) \approx p_{\text{neut}}(x,\tau_0)$ is approximately constant in τ. Using this approximation in Eq. (13) yields

$$\begin{aligned}\lim_{\tau\to\infty} \langle X \rangle \approx & \frac{2Ds}{r} \int dx\, p_{\text{neut}}(x,\tau_0) \int d\tau\, \partial_x u \\ = & \frac{2Ds}{r}(1)\left(-\frac{1}{c}\right) \\ = & -\frac{c}{2r}. \end{aligned} \tag{14}$$

Note that this result did not depend on the form of p_{neut}, only that it was approximately constant in time; in particular, it also holds if the ancestry settles down to a stationary distribution, as in Eq. (3).

A.1 Other kinds of loosely-linked sweep

Above, we have assumed that the sweeping allele spread according to the FKPP equation, $\partial_\tau u + D\partial_x^2 u = -su(1-u)$, which describes an allele with a constant selective advantage s. However, the allele may have a varying selective advantage if, for instance, dominance or frequency-dependent effects are important, or if there is environmental variation. More generally, the changing allele frequency is described by

$$\partial_\tau u + D\partial_x^2 u = -sf(u,x,\tau)u(1-u) \tag{15}$$

for some function f.

Otherwise, the derivation of the expected displacement is the same as above, and we have

$$\lim_{\tau\to\infty} \langle X \rangle \approx \frac{2Ds}{r} \int dx\, p_{\text{neut}}(x,\tau_0) \int d\tau\, f(u,x,\tau)\partial_x u. \tag{16}$$

Assuming that f is such that $u(x,\tau)$ is still a traveling wave moving at some speed c, we can change variables in the second integral to obtain:

$$\lim_{\tau\to\infty} \langle X \rangle \approx -\frac{2Ds}{rc} \int dx\, p_{\text{neut}}(x,\tau_0) \int_0^1 du\;\; f(u,x,\tau(x,u)). \tag{17}$$

B Effect of tightly-linked sweeps

We wish to calculate $\phi(x)$ for large x, including the effect of occasional tightly-linked sweeps. It is easiest to consider $\int_x^L dy\, \rho(y)\phi(y)$, which we can think of as the probability that at some time t_0 in the distant past, the ancestor of a present-day individual was at a distance greater than x from the center. For large x, we expect that this is dominated by the probability that it was pulled there by a 'recent' tightly-linked sweep t generations 'before' t_0 (i.e., t generations closer to the present), with t not too large. This sweep must have pulled the lineage out to a distance of at least $xe^{t/t_{\text{con}}}$ for it still to be at a distance of at least x t generations 'later', and therefore the sweep must have originated a distance $z > xe^{t/t_{\text{con}}}$ from the center. Given that it did, the probability that it pulled the lineage out far enough is $\exp\left[-\frac{r}{c}xe^{t/t_{\text{con}}}\right]$. Putting this all together, and using that the density of sweeps per generation per unit map length per distance (or area in two dimensions) at distance z from the center and genetic map distance r from the focal locus is $2\Lambda/(fKL)$ (or $4\Lambda z/(fKL^2)$ in two dimensions), the expected number of sweeps that would have left the lineage more than x from the center at time t_0 is

$$\begin{aligned}\int_x^L dy\, \rho(y)\phi(y) \approx & \frac{2\Lambda}{fKL^d} \int_0^{t_{\text{con}}\log(L/x)} dt \int_{xe^{t/t_{\text{con}}}}^L dz\, (2z)^{d-1} \int dr\, \exp\left[-\frac{rx}{c}e^{t/t_{\text{con}}}\right] \\ = & \frac{2L}{Kx} \times \begin{cases} 1-(1+\log(L/x))x/L & \text{for } d=1 \\ (L-x)^2/L^2 & \text{for } d=2. \end{cases} \end{aligned} \tag{18}$$

Taking the derivative of both sides of Eq. (18) with respect to x gives the probability density, Eq. (4).

Note that Eq. (18) approximates the probability that there is at least one tightly-linked sweep by the expected number of such sweeps, so it is only valid when the right-hand side is small, $x \gg 2L/K$. It also obviously typically breaks down as x approaches L and the particular geometry of the habitat begins to matter.

C Isolation by distance

We wish to find the probability $\psi(x)$ that a pair of lineages a distance x apart will be identical at a neutral locus. Let us assume that the locus is far from any recent sweeps. (We relax this assumption below.) Then tracing the ancestry back in time, the separation X_τ between them can be approximated by a Brownian motion, with diffusion constant $2D$ (since it combines the motion of both lineages), and with the lineages moving together at a mean rate of $\approx -\Lambda c X/fL = -X/t_{\text{con}}$ from (unlinked) sweeps that start in between them. In other words, we can approximate the motion by

$$dY_\tau = -t_{\text{con}}^{-1} Y_\tau d\tau + 2\sqrt{D} dB_\tau, \tag{19}$$

where B is a Brownian motion. We write Y to emphasize that this is not quite the same as the real path of the lineages X. In particular, unlike X, Y does not include coalescence. (In two dimensions, Y fails to approximate X even when the lineages are just very close together, but since most of the coalescence time will be spent at some distance away, it is still a useful approximation.)

We would like to find an explicit form for Eq. (6). To do this, we can rewrite in terms of the behavior of Y. First, note that the rate of coalescence for the two lineages when they are in the same place is $1/\rho$, and therefore the probability density of coalescence at time τ is $\approx \frac{\delta(Y_\tau)}{\rho} \exp\left(-\int_0^\tau d\tau' \frac{\delta(Y_{\tau'})}{\rho}\right)$, where δ is the Dirac delta. (The exponential factor accounts for the possibility that the two lineages have already coalesced.) Plugging this into Eq. (6) gives:

$$\begin{aligned}\psi(x) =& E_X\left[e^{-2\mu T} \mid X_0 = x\right] \\ \approx& E_Y\left[\int_0^\infty d\tau \frac{\delta(Y_\tau)}{\rho} \exp\left(-2\mu\tau - \int_0^\tau d\tau' \frac{\delta(Y_{\tau'})}{\rho}\right) \mid Y_0 = x\right].\end{aligned} \tag{20}$$

We can use the Feynman-Kac formula ([16], p25) to rewrite Eq. (20) as an ordinary differential equation:

$$0 = 2D\psi'' + \left(2D\frac{d-1}{x} - \frac{x}{t_{\text{con}}}\right)\psi' - 2\mu\psi + \frac{1}{\rho}\delta(x)(1-\psi), \tag{21}$$

where δ is the Dirac delta. Eq. (21) breaks down for $x \to 0$ in $d = 2$ dimensions; in this case, some kind of small-scale cutoff is needed, but this does not change the shape of $\psi(x)$ at larger scales [11]. The last term in Eq. (21) is just a boundary condition that sets the overall normalization of ψ.

The solution to Eq. (21) can be written exactly in terms of special functions. (For $d = 1$, Eq. (21) is the Hermite equation, with solution $\psi(x) \propto H_{-2\mu t_{\text{con}}}\left(\frac{x}{2\sqrt{Dt_{\text{con}}}}\right)$, where $H_\nu(z)$ is a Hermite function [17].) However, approximate asymptotic solutions are more useful. For $x \gg x_c\sqrt{2 + 4\mu t_{\text{con}}}$, the dominant balance is $\psi' \approx -2\mu t_{\text{con}}\psi/x$, with solution $\psi \propto x^{-2\mu t_{\text{con}}}$. For $x \ll x_c$, the pull of unlinked sweeps is negligible, and the solutions are close to the neutral solutions in [11]. At intermediate distances, there is a crossover regime where the form of the dependence of ψ on x is independent of the mutation rate.

C.1 Tightly-linked sweeps

Above, we have focused on regions of the genome far from any recent sweeps. Ideally, however, we would like to be able to extend our analysis to include recently-swept regions. As a first approximation, we can say that the main effect of tightly-linked sweeps is that they can cause two widely-separated lineages to rapidly coalesce. The probability that a sweep recombining at rate r with the focal neutral locus will cause coalescence between two lineages separated by x is $\approx \exp(-rx/c)/(1 + 2r\Upsilon)$, where Υ is mean coalescence

time for two lineages inside the wavefront of the sweep [4]. We can therefore account for the effect of sweeps uniformly distributed over the genome by changing the coalescence kernel in Eq. (20) from $\delta(x)/\rho$ to

$$\begin{aligned} p_{\text{coal}}(x) &\approx \delta(x)/\rho + \frac{2\Lambda}{fK}\int_0^\infty dr \frac{e^{-rx/c}}{1+2r\Upsilon} \\ &\approx \frac{2\Lambda}{fK}\frac{c}{x} \text{ for } x \gg c\Upsilon. \end{aligned}$$

For $x \gg c\Upsilon$, Eq. (21) then becomes

$$0 = 2D\psi'' + \left(2D\frac{d-1}{x} - \frac{x}{t_{\text{con}}}\right)\psi' - 2\mu\psi + \frac{2\Lambda}{fK}\frac{c}{x}(1-\psi).$$

For large x, there are two possible tail behaviors for the solution. If $2\mu t_{\text{con}} < 1$, then the pull of unlinked sweeps is strong enough that it is likely to bring lineages close together before they mutate, and $\psi \propto x^{-2\mu t_{\text{con}}}$ as above. For $2\mu t_{\text{con}} > 1$, only recently-swept loci share recent enough ancestry to be likely to be identical in distant individuals, and $\psi \propto x^{-1}$.

D Simulation methods

Forward-time simulations (purple histogram in Fig. 1) were conducted using the algorithm from [2] (which draws on that of [18]), modified so that population was subdivided into a line of L demes of ρ individuals each, with random dispersal between adjacent demes. Because these simulations were extremely computationally demanding, we also conducted approximate backwards-time simulations to get better statistics and investigate rare events (blue histogram in Fig. 1, and Figs. 2 and 3). These simulations followed a single lineage back in time at one neutral locus as it diffused through a continuous one-dimensional space. Sweeps were treated as instantaneous events arising uniformly at random in space and time and across the genome. Sweeps occurring at a recombination fraction r from the focal locus pulled the lineage an exponentially-distributed distance with mean c/r or $c/(2r)$ (for $r < s$ and $r > s$, respectively), truncated at the origin of the sweep. For both sets of simulations, the focal locus was at the center of a linear genome with map length K Morgans.