

# Beyond clines: lineages and haplotype blocks in hybrid zones

Alisa Sedghifar<sup>\*</sup>, Yaniv Brandvain<sup>§</sup>, and Peter Ralph<sup>†</sup>

<sup>\*</sup>Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology,  
Princeton University, Princeton, New Jersey, 08544

<sup>§</sup>Department of Plant Biology, University of Minnesota, St. Paul, Minnesota, 55108

<sup>†</sup>Department of Molecular and Computational Biology, University of Southern California, Los Angeles,  
California, 90089

## Abstract

Hybrid zones formed between recently diverged populations offer an opportunity to study the mechanisms underlying reproductive isolation and the process of speciation. Here, we use a combination of analytical theory and explicit forward simulations to describe how selection against hybrid genotypes impacts patterns of introgression across genomic and geographic space. By understanding how lineages move across the hybrid zone, in a model without coalescence, we describe how clines form and how parental haplotypes are broken up during introgression. Clines form in about  $1/s$  generations, where  $s$  is the strength of selection against hybrids, and linked clines persist over a genomic scale of  $1/T$ , where  $T$  is the age, in generations, of the hybrid zone. *This is in May1975? Should it be removed?* Locally disadvantageous alleles tend to exist as small families, whose lineages trace back to the side from which they originated at speed  $\sqrt{s}$  dispersal distances per generation. The lengths of continuous tracts of ancestry provide an additional source of information: blocks of ancestry surrounding single-locus incompatibilities can be substantially longer than the genome-wide average block length at the same spatial location, an observation that might be used to characterize the age of hybrid zones and identify candidate targets of selection.

**Keywords:** hybridization, tension zone, introgression, haplotypes, diffusion methods.

## Introduction

The process of speciation commonly involves populations diverging with relatively little gene flow (?). However, when formerly isolated populations come into contact before reproductive isolation is complete, some gene flow is possible. Interbreeding and migration between such populations creates a gradient of alleles derived from the two populations across geographic space, centered on their point of contact (reviewed by ?). If the populations are sufficiently diverged, this process leaves a distinct pattern of variation across the genome, in which long tracts of divergent haplotypes from each ancestral population are broken up by historical recombination events, forming chromosomal junctions in ancestry (???). These patterns of correlated ancestry in admixed populations have been previously used to infer histories of hybridization and admixture (e.g. ???).

Our work here builds on the large body of theory describing tension zones *and aims to describe transient dynamics with a focus on haplotype patterns*. To do this, we take the reverse-time perspective, understanding the temporal dynamics of the system by describing properties of the lineages tracing back from the zone.

The arrangement of segments of genome inherited from particular ancestral populations is shaped by the number of recombination events that have taken place between ancestral genotypes. The length distribution of such segments therefore contains information about the age of the hybrid zone and the relative strength of selection affecting subsets of the genome.

*We show here that heterogeneous patterns of the ancestry length distribution across the genome may help identify putative targets of selection in hybridizing populations, providing another source of information in addition to gradients in allele frequencies used in current approaches (e.g., ??). Specifically, we find that selection against hybrid incompatibility loci changes the extent of correlated ancestry — that is, because selection rapidly removes incompatible alleles from heterospecific genomes, hybrid incompatibilities will be surrounded by disproportionately long ancestry blocks.*

Using a combination of theory and simulated hybrid zones, we describe genome-wide patterns of coancestry as they relate to hybrid zone age, genetic distance from selected locus and geographic distance from hybrid zone center. *We extend the framework presented by ? to describe how lineages move across a zone of secondary contact. We consider the influence of selection against heterozygotes at a single locus on the movement of lineages across space, conditioning on the frequencies of the selected allele and allowing lineages to recombine between selected background* (analogous to the structured coalescent applied to selection, as in ?).

If hybrid ancestry at a given locus is disfavored, migrant haplotypes containing the selected allele will be removed rapidly from the population, preventing introgression of surrounding genomic regions. We therefore expect a deficit of short blocks of foreign ancestry surrounding the selected locus, with this effect becoming more pronounced further away from the center of a hybrid zone. As a corollary to this, conditional on having the ancestry that is at lower frequency (that is, being on the ‘wrong’ side of the hybrid zone), the length of unbroken ancestry surrounding the selected locus is expected to be relatively long when far away from the zone center. This is because an unfit haplotype is more likely to have been recently inherited from the other side of the hybrid zone, and therefore will not have as many ancestors of the locally common type as do neutral haplotypes. *This reasoning is applicable to instances of both tension zones generated by intrinsic genetic incompatibilities, and ecotone models of extrinsic ecological selection (e.g., local adaptation).*

*Because of the well-known problems with spatial coalescence (??), incorporating genetic drift is not straightforward, and we therefore neglect genetic drift in our analytical results.*

This approach is common to nearly all theoretical work, that study the reaction-diffusion equations that govern the deterministic, high-density limit (as in ?), or the discrete analogues (?). There is substantial work on how clines are affected by genetic drift (??????), and quantifying the barrier to gene flow caused by clines (????). *(why are references to 'barrier to gene flow' in the paragraph on drift? perhaps move to discussion?)* While we do not attempt to account for finite population size in our analytic results, our simulation results (below) suggest that although genetic drift strongly affects local noisiness of the system, it does not substantially influence our conclusions at moderate population densities (hundreds within one dispersal distance), perhaps because of the short time-scales that we examine. *(maybe remove this next sentence? but want to say something about the very low densities we don't examine.)* At much lower population densities, local extinction and recolonization are expected to have a stronger effect.

*(nice wrap-up sentence shouldn't be at the end of paragraph on drift:) Is it weird to end intro on above paragraph? Could this wrap-up sentence hold up on its own? Our results look beyond the traditional signature of selection observed in the width of genetic clines, and present a possible new approach for identifying genomic regions experiencing selection in hybrid zones.*

## Methods

### The Model

We consider two isolated populations — labeled species  $A$  and species  $B$  — that came into contact  $T$  generations in the past. We will say that species  $A$  was initially on the “left” of the zone of contact, which corresponds to spatial positions  $x < 0$ . After contact, the populations live across continuous geographical space, with random, local dispersal that we take to be Gaussian (although this should not affect the conclusions if the true dispersal distribution is not too fat-tailed).

We model selection through a single, underdominant locus: at this locus, there are two alleles, one corresponding to each of the ancestral populations. Individuals who are heterozygous at this locus produce on average  $1 - s$  times as many offspring than either homozygote. Although most selection in hybrid zones is likely more complex than this simple situation, previous work has shown that such models generate clines that are very similar to models of ecological selection or epistatic systems (??), *because selection on an allele depends on its marginal effect on fitness.* Exceptionally strong epistatic selection which acts more

quickly than free recombination may therefore, yield results that differ slightly from ours. Likewise, while we do not model ecological selection, theoretical and simulation studies show that clines formed by underdominant selection and extrinsic selection are effectively indistinguishable without additional information (??), especially over short time scales (?), with their shape determined mostly by the mean fitness of the heterozygote (?).

As in ?, we consider where the ancestors of modern-day individuals fall across geography as we look further back towards the time of first contact. We say that a locus in a sampled individual is of *ancestry A* if it has been inherited from an ancestor of species *A*, i.e., if its lineage traces back to a spatial position  $x < 0$  at the time of secondary contact. A block of genome is of ancestry *A* if every locus in it does the same; this occurs if there is no recombination in this block, or if all lineages generated by recombination events in this block trace back to the *A* side. This process, in which recombination events cause such lineages to *branch*, is illustrated in Figure ?. We say that a block of genome is on the *A background* if it is physically linked to an allele of ancestry *A* at the selected locus; if a block of ancestry *A* includes the selected site then it is necessarily on the *A background*. Because the identity of the selected allele determines how selection acts on the haplotype, and because linked alleles can only move between the *A* and the *B* background in heterozygotes, a key factor in these models is the density and fecundity of heterozygotes.

## Analysis

In this section we aim to give both heuristics for how lineages move, that are helpful in providing intuition and for establishing order-of-magnitude estimates, and analytics, mostly in terms of partial differential equations needing numerical solution. *As noted above, we ignore the effects of drift and the pedigree, so that the underlying allele frequency dynamics are deterministic. This should provide a good model for stochastic motion of a single lineage (our main focus), but does not model correlations between lineages.*

**Establishment of the cline at the selected locus** Deterministic theory predicts that after secondary contact, the two alleles at the selected locus will form a stable cline, affecting neighboring loci as well (??). First, we turn our attention to how the cline at the selected locus itself is formed. Write  $p(x, t)$  for the frequency of the *A* allele at spatial location  $x$  and time  $t$ . Suppose that secondary contact occurred at time  $t = 0$ , and the *A* allele is initially fixed on the left, so  $p(x, 0) = 1$  if  $x < 0$  and  $p(x, 0) = 0$  if  $x > 0$ . We assume that dispersal is local and unbiased; and that the mean squared displacement between parent and offspring in the direction perpendicular to the cline is  $\sigma^2$ . Assuming that: (i) alleles locally assort into diploids randomly, (ii) habitat and dispersal are homogeneous, and (iii)

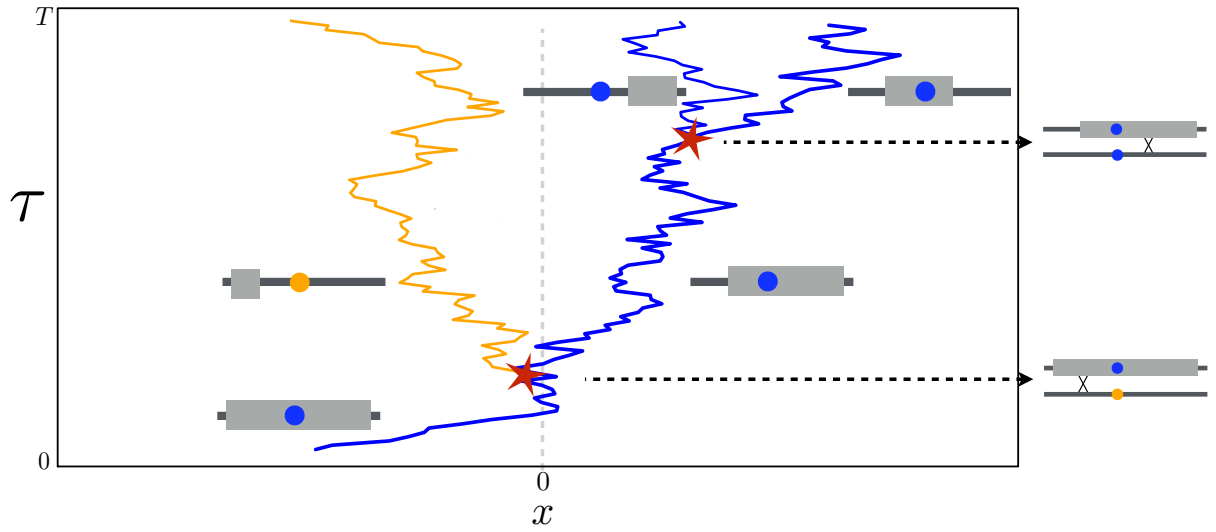


Figure 1: **A schematic figure of the lineages along which a segment of genome has been inherited**, showing the location (horizontal axis) and time (vertical axis) at which the ancestors of a hypothetical sampled haplotype lived. At the time of sampling ( $\tau = 0$ ), the segment contains the  $B$  allele at the selected locus. Paths traced by lineages across space are depicted by blue and yellow lines. Branching events occur when there was a recombination event within the sampled segment. The *rightmost* blue line represents the path of the selected locus. The colors of the other paths indicate the identity at the linked selected locus (blue: linked to allele  $B$ ; yellow: linked to allele  $A$ ). The center of the hybrid zone is at geographic position  $x = 0$ ; in the region  $x < 0$ , ancestry  $B$  is less common, and because of selection, the selected lineage spends little time there. Based on position of the lineages at  $\tau = 100$  generations ago, from left to right the segments are of ancestry  $(A, B, B)$ .

population density is large, then as in ?, the commonly-used equation that  $p$  approximately solves is

$$\frac{d}{dt}p = \frac{\sigma^2}{2} \frac{d^2}{dx^2}p + 2p(1-p) \times s\left(\frac{p}{2} - 1\right), \quad (1)$$

where  $\frac{d}{dt}p$  is the local rate of change of the frequency of  $A$ . The first term,  $\frac{\sigma^2}{2} \frac{d^2}{dx^2}p$ , describes the net rate at which  $A$  alleles arrive by local dispersal. The second term describes the impact of selection on local allele frequency as the product of local genotypic variance,  $2p(1-p)$ , and selection on allele  $A$ ,  $s(p/2 - 1)$ . *I don't think 'genotypic variance' is the right interpretation? Let me try again: The second term is the product of the local frequency of heterozygotes,  $2p(1-p)$ , and the strength of selection on allele  $A$ ,  $s(p/2 - 1)$ .* The term  $s(p/2 - 1)$  describes underdominance –  $A$  is favored when common ( $p > 1/2$ ) and disfavored when rare ( $p < 1/2$ ). The equation can be derived as in ?; the assumption of locally random mating was shown to have only a small effect ?.

The equation is approximate because it omits terms of order  $s^2$  (so deviations from the prediction occur over  $1/s^2$  generations), and relies on the Central Limit Theorem to approximate generation-to-generation dispersal by the Gaussian (so if dispersal is non-Gaussian, will fit best over longer time scales, say, tens of generations). As we show below, the solutions provide a good approximation across the range of realistic values of  $s$ . Although there is no known exact analytic solution to this equation, the steady-state solution is  $\lim_{t \rightarrow \infty} p(x, t) = (1 + \tanh(-2x\sqrt{s}/\sigma))/2$  (?). This relies on several approximations, but the general conclusions should be quite robust: the stable cline has width of order  $\sigma/\sqrt{s}$ , and decays exponentially.

As noted by others (e.g. ??), rescaling space and time by  $\sigma/\sqrt{s}$  and  $1/s$  respectively in equation (??) results in a dimensionless equation, implying that the cline establishes over a timescale of the order  $1/s$ . Since this is how long it takes diffusion at rate  $\sigma^2$  to smooth across a region of width  $\sigma/\sqrt{s}$ , this means we can think of the cline as established by diffusion, despite being slowed down somewhat by selection against heterozygotes.

For simplicity we mostly work with one-dimensional clines, but the mathematics applies as well to two-dimensional systems with some modifications: for instance, if the hybrid zone is a straight line, the description above applies to motion of a lineage transverse to the hybrid zone. In this case,  $d^2/dx^2$  should be replaced by the Laplacian operator, and  $d/dx$  by the gradient. In fact, if the landscape is heterogeneous or dispersal is non-Gaussian, the theory still holds replacing the Laplacian by the appropriate operator. Note, however, that with low population densities, drift can have a strong effect that differs between one and two dimensional systems (???).

**Lineages at the selected locus** Now suppose that the frequency profile of the selected allele,  $p(x, t)$ , is known as a function of time and space. We seek to first understand how lineages of sampled individuals behave at the selected locus, and next extend the analysis to lineages at loci linked to the selected locus.

Consider the collection of  $A$  alleles found at geographic location  $x$ . The expected number, among these alleles, that had a parent at location  $y$  in the previous generation is proportional to the number of  $A$  alleles that were at  $y$  multiplied by the per-generation probability of dispersing from  $y$  to  $x$ . In other words, lineages move as a random walk determined by the dispersal kernel, but biased towards locations where the selected allele they carry is at higher frequency.

Making the same assumptions as for equation (??), the derivation in Appendix ?? shows that the lineage of an  $A$  allele at position  $x$  moves as Brownian motion driven by dispersal with mean displacement proportional to the derivative of  $\log p$ . This has the effect of pulling  $A$  lineages towards regions where they are more common, i.e., the left side of the range. (the analogous physical system is a particle that moves at speed  $\sigma^2$  in the potential  $-\log p$ ). To see why this is true, note that the probability that the parent of an  $A$  allele found at  $x$  lived at position  $x - r$  is proportional to  $\mathbb{P}\{R = r\}p(x - r)$ , where  $R$  is the random dispersal distance; and so the mean displacement from offspring to parent is  $\mathbb{E}[Rp(x - R)]/\mathbb{E}[p(x - R)]$ . Using the fact that  $\mathbb{E}[R] = 0$  and  $\mathbb{E}[R^2] = \sigma^2$  and expanding  $p(\cdot)$  to first order about  $x$  shows that this mean displacement is approximately  $\sigma^2 p'(x)/p(x)$ , which is  $\sigma^2$  multiplied by the gradient of  $\log p(x)$ . This description holds even when the frequency profile of the  $A$  allele changes with time (replacing  $p(x)$  by  $p(x, t)$ ).

At equilibrium, the  $A$  allele is nearly fixed at a geographic position far to the left ( $p \approx 1$ ), and is rare far to the right, with mean frequency at distance  $x$  proportional to  $\exp(-x\sqrt{s}/\sigma)$ . Therefore, roughly, lineages on “their own” side wander randomly, while lineages on “the wrong” side are pushed at constant speed  $\sigma\sqrt{s}$  back towards the side where they are more common (since here  $\frac{d}{dx} \log p(x) \approx \sqrt{s}/\sigma$  and the speed is  $\sigma^2$ ). Since an  $A$  allele must, by definition, have been inherited from the  $A$  side of the barrier at the time of secondary contact, this push must get more intense the closer it is to the time of secondary contact.

This description gives more information than the steady-state cline, which depends only on  $\sigma/\sqrt{s}$ . Here we see that lineages with  $\sigma = 10$  and  $s = .16$  move much faster than lineages with  $\sigma = 1$  and  $s = .0016$ , reflecting strong differences in selection against heterozygotes, even though the stable clines have the same form, and therefore have additional information about the hybrid zone. Even though a rescaling of space and time as described above can make the models equivalent, this difference in lineage speed can be seen through the action of recombination, which we explore next.

**Lineages at linked loci** The behavior of a lineage at a linked locus is similar to a selected locus. However, there is one important difference – in heterozygotes the lineage linked to the selected locus may recombine onto the other selected background. Therefore, if we follow back through time a lineage at a locus linked to an  $A$  allele, it will first tend to be inherited from ancestors to the left (as  $A$  lineages drift to the left). However, with sufficient time in the hybrid zone, recombination allows this linked locus to have been inherited from a  $B$ -carrying individual, whose ancestors will tend to be more from the right.

*Suppose we sample an allele today  $r$  Morgans from the selected site, and follow its lineage back through time, using  $\tau$  to denote “generations ago” (reserving  $t$  for time measured in the usual direction). If  $X_\tau$  is the geographic location of its ancestor  $\tau$  generations ago, then we say that  $X$  moves as a diffusion pushed by either  $\log(p)$  or  $\log(1 - p)$  (as described for a selected allele in equation (??)). Following this lineage back in time, the identity of the selected allele that ancestor carried at time  $\tau$ ,  $Z_\tau$ , jumps between  $A$  and  $B$  with recombination events between the neutral and selected locus (with frequency  $r$ ), that result in a change in the selected background. Thus, by the assumption of locally random assortment of alleles,  $Z$  shifts from  $A$  to  $B$  at rate  $r(1 - p)$ , and  $B$  to  $A$  at rate  $rp$  (see Appendix ?? for a more precise description).*

We can describe this process formally in Itô notation: with  $B_\tau$  a standard Brownian motion,  $T_B(\tau)$  the most recent time before  $\tau$  that  $Z_\tau = B$ , and likewise for  $T_A$ ,

$$dX_\tau = \sigma dB_\tau + \begin{cases} \sigma^2 \frac{d}{dx} \log(p(X_\tau, \tau)) d\tau & \text{if } Z_\tau = A \\ \sigma^2 \frac{d}{dx} \log(1 - p(X_\tau, \tau)) d\tau & \text{if } Z_\tau = B \end{cases} \quad (2)$$

$$\mathbb{P}\{T_B(\tau) > \tau + u \mid Z_\tau = A\} = \exp\left(-r \int_\tau^{\tau+u} (1 - p(X_s, s)) ds\right)$$

$$\mathbb{P}\{T_A(\tau) > \tau + u \mid Z_\tau = B\} = \exp\left(-r \int_\tau^{\tau+u} p(X_s, s) ds\right).$$

In the first expression, giving the distribution of how the lineage location changes, the first term ( $\sigma dB_\tau$ ) is Brownian noise driven by dispersal, and the second is the mean displacement, which moves the lineage “downhill” towards its selected allele’s ancestral range on either  $-\log p$  or  $-\log(1 - p)$ , depending on which selected allele the lineage is linked to. (In two dimensions,  $d/dx$  is replaced by the gradient.) *The second expression gives the probability that a lineage on the  $A$  background at time  $\tau$  does not switch to the  $B$  background across  $u$  generations, i.e., the probability that there was no recombination between the lineage and the selected site in those  $u$  generations. Since the rate of recombination between the two is  $r$ , and the probability the lineage is in a heterozygote at time  $s$  is equal to the local proportion of  $B$  alleles,  $1 - p(X_s, s)$ , the mean number of recombinations in heterozygotes across that*



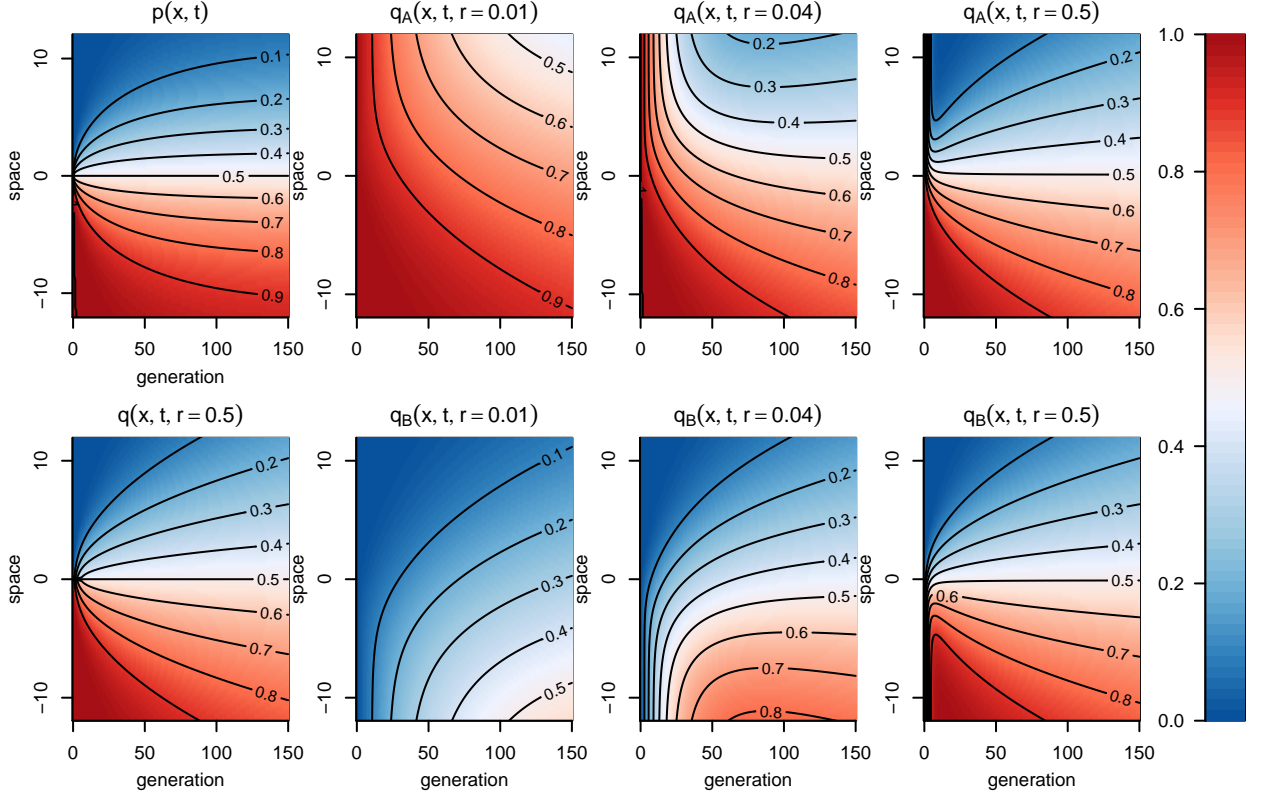


Figure 2: **Probabilities of A ancestry**, across space (vertical axis, in units of  $\sigma$ ) and time (horizontal axis, in generations). In each plot, color corresponds to the expected frequency of A ancestry at a particular location in time and space. The selection coefficient is  $s = .02$ . **Top left:** at the selected site, showing establishment and stabilization of the cline on a time scale of  $1/s = 50$  generations. **Bottom left:** at an unlinked site, with cline flattening continuing with  $\sqrt{t}$ . **Remaining figures** show frequencies of A ancestry *conditional* on the ancestry at the selected site, at different distances from the selected site ( $r = .01, .04$ , and  $0.5$  Morgans), as described in the text (see definition of  $q_z(x, t, r)$ ). See figure ?? for the same figure over a longer period of time.

time period is  $r \int_{\tau}^{\tau+u} (1 - p(X_s, s)) ds$ ; assuming recombinations are Poisson, the probability of no such recombinations is the expression given. The third expression is the same, except for switching B to A.

**Linked clines** We can use this diffusion model for lineages to find expected clines in ancestry, i.e., the expected proportion of individuals who inherit from species A, as a function of space, time, and position on the genome. Precisely, we need the probability that an allele sampled  $t$  generations after secondary contact at location  $x$ , at recombination distance  $r$  to a selected allele of type  $z$ , is inherited from an individual of ancestry A, where  $z$  can be either A or B. We denote this probability  $q_z(x, t, r)$ . In the notation above,

$$q_z(x, t, r) = \mathbb{P}^x\{Z_t = A \mid X_0 = x, Z_0 = z\}. \quad (3)$$

(Recall that  $Z$  depends implicitly on  $r$ .) The description above implies that  $q_z$  solves the following Kolmogorov backward equation:

$$\begin{aligned}\frac{d}{dt}q_A(x, t, r) &= \frac{\sigma^2}{2} \frac{d^2}{dx^2} q_A(x, t, r) + \sigma^2 \frac{d}{dx} \log(p(x, t)) \cdot \frac{d}{dx} q_A(x, t, r) \\ &\quad + r(1 - p(x, t))(q_B(x, t, r) - q_A(x, t, r)) \\ \frac{d}{dt}q_B(x, t, r) &= \frac{\sigma^2}{2} \frac{d^2}{dx^2} q_B(x, t, r) + \sigma^2 \frac{d}{dx} \log(1 - p(x, t)) \cdot \frac{d}{dx} q_B(x, t, r) \\ &\quad + rp(x, t)(q_A(x, t, r) - q_B(x, t, r)),\end{aligned}\tag{4}$$

with boundary conditions  $q_A(x, 0, r) = 1$  and  $q_B(x, 0, r) = 0$ . The three terms in these equations come from Brownian movement of a lineages (i.e., the smoothing action of local dispersal), the tendency of a lineage to inherit from regions where its type is more common (i.e., the net flux induced by reduced fitness in the hybrid zone), and recombination between selected backgrounds, respectively. *I think the first and second terms may have been switched?? We note that because recombination occurs within a deme, the recombination terms in equation (??) are quite standard (?)*.

We will have more use for the differential operators on the right-hand sides of these equations, so define these as  $G^A = \sigma^2 \frac{d}{dx} \log(p(x, t)) \cdot \frac{d}{dx} + (\sigma^2/2) \frac{d^2}{dx^2}$  and  $G^B = \sigma^2 \frac{d}{dx} \log(1 - p(x, t)) \cdot \frac{d}{dx} + (\sigma^2/2) \frac{d^2}{dx^2}$ , so that equation (??) can be written more compactly as

$$\begin{aligned}\frac{d}{dt}q_A &= G^A q_A + r(1 - p)(q_B - q_A) \\ \frac{d}{dt}q_B &= G^B q_B + rp(q_A - q_B).\end{aligned}$$

In technical terms,  $G^A$  and  $G^B$  are the generators of the diffusions of lineages of selected alleles of ancestry  $A$  and  $B$ , respectively; informally, they encode the stochastic motion of lineages at the selected loci.

**Numerical computation** To determine how the cline at a linked locus is expected to relax, we can solve the partial differential equations (??) and (??) numerically. For instance, Figure ?? shows a heatmap of  $p(x, t, r)$ , the expected frequency of ancestry  $A$  at location  $x$  and time  $t$  at a site at recombination distance  $r$  from the selected site, which is computed as the frequency of ancestry  $A$  on each selected background weighted by the frequencies of each background:  $q(x, t, r) = p(x, t)q_A(x, t, r) + (1 - p(x, t))q_B(x, t, r)$ . The equations are solved numerically in R, using the ReacTran package (?).

**Haplotype lengths** We now develop our model further to find the frequency at which *entire* blocks of genome (haplotypes) of a single ancestry are found, at a given location and time. Suppose we sample an individual at spatial location  $x$  and time  $t$  after the initiation

of gene flow, and genotype them on the genomic segment between positions  $a$  and  $b$ , relative to the selected site. We are interested in the probability  $g_z(x, t; a, b)$  of finding an entire segment  $(a, b)$  of ancestry  $A$  given that the individual has selected allele of type  $z$ . For instance,  $g_B(0, 10; 0, .2)$  is the probability that an individual carrying a selected  $B$  allele sampled at the center of the zone at  $t = 10$  has a block of  $A$  ancestry for 0.2 Morgans to the right of the selected site.

As in ?, a given block of genome is inherited along a single lineage ever since the most recent recombination event that fell within that block. Prior to this, there are two lineages to follow (see Figure ??), and so lineages behave as labeled, branching diffusions, where the total branching rate is conserved. We do not consider subsequent coalescence. The general description of the process, again looking backwards in time, is as follows: The lineage of a segment of genome moves as a linked locus described in equations (??), with recombination distance  $r$  equal to the rate at which the segment recombines away from the selected site. (If the selected site lies inside the segment,  $r = 0$ .) Additionally, at rate equal to the genetic length of the segment, recombination occurs, at which point the segment splits in two at a uniformly chosen location between  $a$  and  $b$ , each of which proceeds as before, independently. In this description,  $g_z(x, T; a, b)$  is then the probability that all branches are found on the  $A$  side of the hybrid zone at the time of secondary contact  $T$  units of time ago. We will write  $r(a, b)$  for the distance from the segment to the selected site: always taking  $a \leq b$ ,  $r(a, b) = 0$  if  $a \leq 0 \leq b$  and  $r(a, b) = \min(|a|, |b|)$  otherwise.

The resulting equation is similar to (??) with a term added for branching, which is written (omitting the  $(x, t)$  for conciseness):

$$\begin{aligned} \frac{d}{dt} g_A(a, b) = & G^A g_A(a, b) + r(a, b)(1 - p)(g_A(a, b) - g_B(a, b)) \\ & + (1 - p) \left( \int_a^{\min(0, b)} g_B(a, \theta) g_A(\theta, b) d\theta + \int_{\max(0, a)}^b g_A(a, \theta) g_B(\theta, b) d\theta \right) \\ & + p \int_a^b g_A(a, \theta) g_A(\theta, b) d\theta - (b - a) g_A(a, b), \end{aligned} \quad (5)$$

Here the first term ( $G^A g_A$ ) represents spatial mixing, and the second term results from recombinations *between* the block and the selected site in heterozygotes, which switch the identity of the linked, selected locus *without splitting the block* (the factor  $(1 - p)$  is the probability that the block, initially linked to an  $A$  allele, encounters a  $B$  allele under the assumption of locally random mating). In the integrals,  $\theta$  is the genomic position of the recombination event occurring *inside* the block. The third term accounts for such recombinations in a heterozygote: the portion of the block nearest the selected site remains linked to an  $A$  allele, and the remaining portion becomes linked to a  $B$  allele (as is the case in

the most recent recombination event shown in Fig. ??). To account for cases where the selected locus lies outside the interval  $[a, b]$ , we say that the integral  $\int_a^{\min(0, b)}$  is zero if  $a > 0$  and likewise  $\int_{\max(0, a)}^b$  is zero if  $b < 0$ . The last integral results from recombination inside the block in homozygotes for  $A$  (as is the case for the second recombination event in Fig. ??), and the final term balances the outflux due to all recombinations. The equation for  $g_B(x, t; a, b)$  is identical after exchanging  $A \leftrightarrow B$  and  $p \leftrightarrow (1 - p)$ . The boundary conditions are  $g_A(x, 0; a, b) = 1$  and  $g_B(x, 0; a, b) = 0$ .

**Numerical solutions** Notice that equations (??) are hierarchical in  $(a, b)$ : the equation for haplotype identity probabilities on a segment  $(a, b)$  depends only on those probabilities for segments contained in  $(a, b)$ . This allows for numerical solutions, described in more detail in Appendix ??.

**Correlations in ancestry** To compute correlations in local ancestry (i.e., “ancestry disequilibrium”, as in ??), we need only follow lineages at two sites, instead of an entire region. Doing so only requires computing correlations in ancestry between markers, which can be done directly using our numerical code; see Appendix ?? for more detail.

## Simulations

We implemented forwards-time simulations of a one-dimensional grid of demes with non-overlapping generations and fixed population sizes (a Wright-Fisher model). Individuals are diploid, with haploid number  $n = 2$  chromosomes, each of length 1 Morgan. One chromosome pair harbors, at position 0.5M, a single locus that reduces fitness by  $s$  in heterozygotes, while the other contains no sites under direct selection.

Each deme has exactly  $N_d$  diploid hermaphroditic individuals at the start of each generation. Then, every individual disperses to a (possibly) new deme by choosing a Gaussian displacement with mean zero and variance  $\sigma^2$ , then dispersing to the nearest deme. (The mean displacement is zero, and when comparing to theory we compute  $\sigma$  as the standard deviation of this distribution.) Migrants past either end of the range remain at the terminal demes. Then, fitnesses are computed (heterozygotes at the selected locus have fitness  $1 - s$ ; homozygotes have fitness 1), and in each deme  $N_d$  pairs of parents are chosen, with replacement, with probability proportional to their fitness and selfing allowed. Since migration is not conservative, demes may have no available parents; in this case, parents are chosen from other demes with probability proportional to fitness multiplied by  $\exp(-x^2/\sigma^2)$ , where  $x$  is the distance to the other deme. The next generation is formed by carrying out meiosis in each parent and combining the resulting gametes such that each pair of parents leaves

one descendant. Meiosis results in alternating blocks of the gamete's chromosome being inherited from the two parental chromosomes, with the blocks separated by a Poisson(1) number of uniformly chosen recombination points along the chromosome, and the order of the parental chromosomes chosen randomly. The simulation software works by recording, for each chromosome, a list of ancestry breakpoints, and the index of the ancestor from which the chromosome inherited that genomic region. We then assigned ancestry at individual loci by looking up which side of the zone the ancestor lived on.

The simulations were executed in R, with scripts available at <https://github.com/petrelharp/clinal-lineages>.

## Measures of introgression

The above theory and simulations generate predictions about patterns of ancestry surrounding selected loci. In reality, however, such loci are not usually known, so it is useful to have per-site statistics that may allow for detection of candidate targets of selection. The most straightforward measure is  $l_B(m, x)$ , the mean length of all contiguous segments of ancestry  $B$  sampled at position  $x$  that contain genomic position  $m$ . Likewise,  $l_A(m, x)$  is the mean length of segments of ancestry  $A$ . We compute similarly the unconditioned mean block length  $l(m, x)$  by averaging all segment lengths without regard to ancestral identity.

We also look at the mean length of the two chunks,  $m-$  and  $m+$ , that flank, to the left and right respectively, the block of unbroken ancestry containing  $m$ . As described below, these blocks tend to be *shorter* than average when  $m$  is the selected site, motivating us to define the statistic

$$C(m, x) = \frac{2 \sum_i l_B(m_i, x)}{\sum_i l_A(m_i+, x) + l_A(m_i-, x)} \quad (6)$$

where  $m_i$  is the block length in individual  $i$ , and the sum is over individuals at location  $x$ . The statistic  $C$  is the mean block length around  $m$  in the population, divided by the mean lengths of the two blocks directly flanking the block containing  $m$ .

*Below we use  $l_B$  to mean the normalized version also. I am defining consistent notation,  $\bar{l}_B$ .*

To identify regions of the genome with abnormal distributions of block lengths, we compute  $l_B(m, x)$  and  $C(m, x)$  at a grid of positions across geography and across the genome. From these, we compute normalized versions  $\bar{l}_B(m, x)$  and  $\bar{C}(m, x)$  by dividing by the empirical mean across the genome for each geographic location  $x$ : for instance,  $\bar{l}_B(m, x) = n_x l_B(m, x) / \sum_i l_B(m_i, x)$ , where the sum is over the  $n_x$  individuals at location  $x$ . While there is useful information contained in the geographic distribution of block length

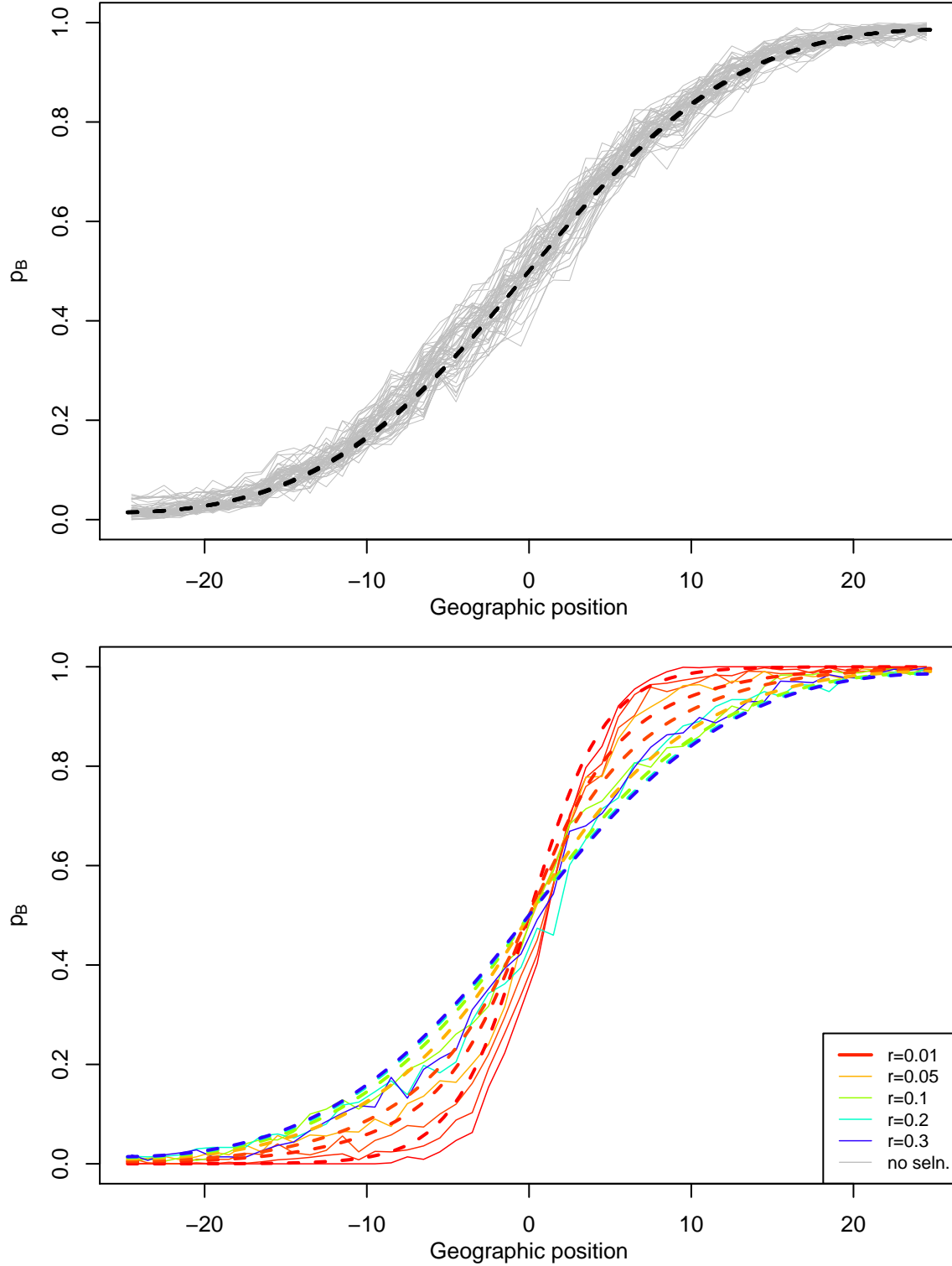


Figure 3: **Comparison of simulated to theoretical cline:** As in Figure ??, frequency of ancestry  $B$ , across geography and at several physical positions on the genome, at  $T = 100$  generations after secondary contact, and at  $s = 0.1$ . Dotted lines are expected frequencies of  $B$  ancestry  $1 - q(x, t, r)$  (unconditional on the ancestry of the linked allele), computed numerically. Each line represents a locus some distance  $r$  away from the true target of selection, and  $r = 0$  represents the locus that is under selection. Colors correspond to different values of  $r$ , transitioning from red (tight linkage to selected site) to blue (distantly linked).

patterns, obtaining good spatial sampling can often be difficult, and this allows us to search for patterns if block lengths are only known within relatively few populations.

We are able to partially trace the genealogy of haplotypes in our simulations. In particular, we wish to know the relative number of ancestors from  $T$  generations ago that are represented in present day populations at a given locus, given the frequency of a particular ancestry. This is a reflection of the average size of a haplotype's family, and is potentially a source of additional information about selection. Within each deme  $x$ , we calculated  $F_B(m, x)$ , which is the number of independent ancestors from time  $T$  contributing to the pool of alleles of ancestry  $B$  at site  $m$ , divided by the number of individuals of ancestry  $B$  at site  $m$ .

## Results

### Single locus clines

As expected, clines in simulations form and begin to flatten, with ancestry frequencies at the selected site maintaining a stable shape, and tightly linked sites flattening more slowly than distant sites. This is seen in Figure ??, and supplementary figures ??, ?? and ??. Clines in ancestry frequency matched expected values found from deterministic theory up to stochasticity due to genetic drift, which is more pronounced at lower population densities. Figure ?? shows that expected clines computed from equation (??) match simulated clines in Figure ?? quite well. Figures ??, ??, and ?? show this comparison with a lower population density than in Figure ??, and separating linked clines based on the identity of the linked selected allele. The agreement is good over hundreds of generations, although as the clines flatten they are seen to wobble (which happens at rate proportional to  $\sqrt{T}$ ; ?).

Loci not under direct selection can in principle spread across the cline unimpeded. In practice, however, it can take quite some time for even unlinked neutral loci to homogenize, due to the decreased fitness of heterozygotes (?) and the relative slowness of diffusive movement. We display the spread of ancestry across space and time in Figure ??. The cline in ancestry at a locus  $r$  Morgans from the selected site will have flattened out to distance  $x$  (say, on the  $B$  side) if there is a good chance that the corresponding lineage that begins linked to a  $B$  allele traces back to an  $A$  allele on the opposite side of the hybrid zone. Since lineages linked to  $B$  alleles move nearly as unbiased Brownian motion on the  $B$  side, this is only possible if Brownian motion has had enough time to travel distance  $x$ , i.e., if  $T > \sqrt{x}$ . This square-root flattening is seen in Figure ?? (and is discussed for environmentally determined clines by ?; but see ?). A linked lineage must also spend at least  $1/r$  generations in heterozygotes to have a good chance of recombining, so clines with

$r < 1/T$  will still resemble the selected cline, which can also be seen in Figure ??.

In principle, the genomic window about the selected site in which clines remain narrow could be quite a bit wider, since the only way to move linked lineages between selected backgrounds is via recombination in a heterozygote, and heterozygotes for the selected allele are only found at high frequency in the cline. The majority of lineages are generally pushed away from the cline but have no bias far away, so the amount of time a lineage spends in heterozygotes should grow as  $\sqrt{T}$  for large  $T$ , and so the width of the genomic region showing clines about the selected locus could be substantially larger than  $1/T$ . However, this distinction appears hard to observe for realistic parameter values.

## Blocks of ancestry

The distribution of contiguous ancestry block lengths contains more information than allele frequency alone. We are specifically interested in how the tracts of ancestry surrounding the selected locus compares to the rest of the genome. Ideal information – true ancestry assignments for a few simulated individuals sampled from across space – are shown in Figure ?? ( $T = 100$  generations) and Figure ?? ( $T = 1000$  generations). For the more recent hybrid zone ( $T = 100$ ), the selected cline has established, but linked clines are still flattening. After a longer period ( $T = 1000$ ), clines over much of the chromosome are flat (since the width of the entire population is less than  $1/\sqrt{T} = 31.6\sigma$ ), but a distinct enrichment of each ancestry is observed around the selected site.

We expect that, in the absence of selection, blocks of  $A$  ancestry across the genome will tend to be shorter the further one goes into the  $B$  side of the cline, because they have had more opportunities to recombine with  $B$  haplotypes. This is seen in Figure ?. However, we expect that stretches of  $A$  ancestry *containing a selected site* will be longer than those that do not contain the selected site at the same spatial location, because lineages containing the selected site have usually been inherited from the  $A$  side of the cline recently. As discussed above, we expect these lineages move at speed roughly  $\sigma\sqrt{s}$ , so (selected)  $A$  alleles at distance  $x$  from the cline center have last had an ancestor on the  $A$  side of the cline around  $x/(\sqrt{s}\sigma)$  generations ago (compared to  $x^2/\sigma^2$  for a neutral allele). This implies the scale on which  $A$  haplotypes are found surrounding the locus should be no longer than about  $\sigma\sqrt{s}/x$ .

**Identifying selected loci** The statistic  $C(m, x)$  shows promise for identifying selected loci under some circumstances. As expected, regions surrounding a locus under selection are more resistant to introgression, as seen in Figures ?? and ?. When present, we expect haplotypes that contain the locally less common allele to be longer than the genome-wide average. Indeed, as shown in Figures ??, ?? and ??, the mean length of such haplotype



blocks is up to three times longer than the average for that geographic location, peaking quite sharply around the location of the selected site.

*To provide some intuition about the physical scale of this signal, we note that* genomic regions are inherited from ancestors  $T$  generations ago in blocks of size roughly  $1/T$ . There will be a long segment of  $A$  ancestry if many such adjacent blocks are all inherited from the  $A$  side. Ancestries of neighboring blocks are correlated, due to the branching process described above. But, if we assume they are independent, then since the block surrounding location  $r$  in an individual at geographic location  $x$  is of ancestry  $A$  with probability  $q_A(x, T, r)$ , we'd expect to see, roughly,  $q_A(x, T, r)(1 + 2q_A(x, T, r))/(1 - q_A(x, T, r))$  consecutive blocks of ancestry  $A$  about a given site unlinked to the selected site. (This assumes, crudely, that the number of  $A$ -blocks on either side of the enclosing block has a Geometric distribution with parameter  $q_A(x, T, r)$ .) This implies the mean length  $l(x, r)$  would be  $q_A(x, T, r)(1 + 2q_A(x, T, r))/((1 - q_A(x, T, r))T)$ , and the mean conditional length  $l_A(x, r)$  would be  $(1 + 2q_A(x, T, r))/(T(1 - q_A(x, T, r)))$ . Furthermore, we know that if  $x = 0$  or if  $T$  is large and  $r$  is not too small, that  $q_A$  is close to  $1/2$  (so  $l_A(x, r) \approx 4/T$ ); and for  $r < 1/T$  that  $q_A$  looks like the selected cline. Also, we know from the discussion above that the lineage of a *selected*  $A$  allele, if it is in the region where  $A$  is rare, moves at speed roughly  $\sigma\sqrt{s}$  back towards the  $A$  side of the zone, returning to the region where  $A$  is common in about  $x/(\sigma\sqrt{s})$  generations. Therefore, a selected  $A$  allele found on the  $B$  side of the zone should carry with it a haplotype of average length  $\sigma\sqrt{s}/x$  that looks like haplotypes from the center of the zone. This analysis suggests that  $A$  haplotypes in the center of the zone should be of average length  $4/T$ ; this is indeed what is seen at distant sites, for instance, in Figure ???. Haplotypes at the selected site are expected to be longer, but still of a length proportional to  $1/T$ , suggesting that the normalization in the statistic  $C(x, m)$  is appropriate, as shown in Figure ??, although a numerical prediction of the value of  $C(x, m)$  is elusive.

The mean haplotype length found without conditioning on ancestry,  $l(x, m)$ , shows a smaller increase near the selected locus, because most blocks will be of the locally common type, and so do not trace back to regions of different block lengths (Figures ??, ?? and ??).

Power will be optimal at intermediate values of  $s$ . If selection is too strong, it may be difficult to observe this signal due to a lack of introgressed selected sites, while if selection is weak, the selected lineage does not move very fast, and so the strength of signal from elevated  $l_B(m)$  is diminished (*Fig. ??*). Indeed, if  $s$  is less than  $x/(\sigma T)$ , then only one block of ancestry  $A$  is expected to be seen about a selected  $A$  allele, and  $l_A(x, m)$  is expected to be  $2/T$ . For similar reasons, power and resolution are best at intermediate  $T$ .

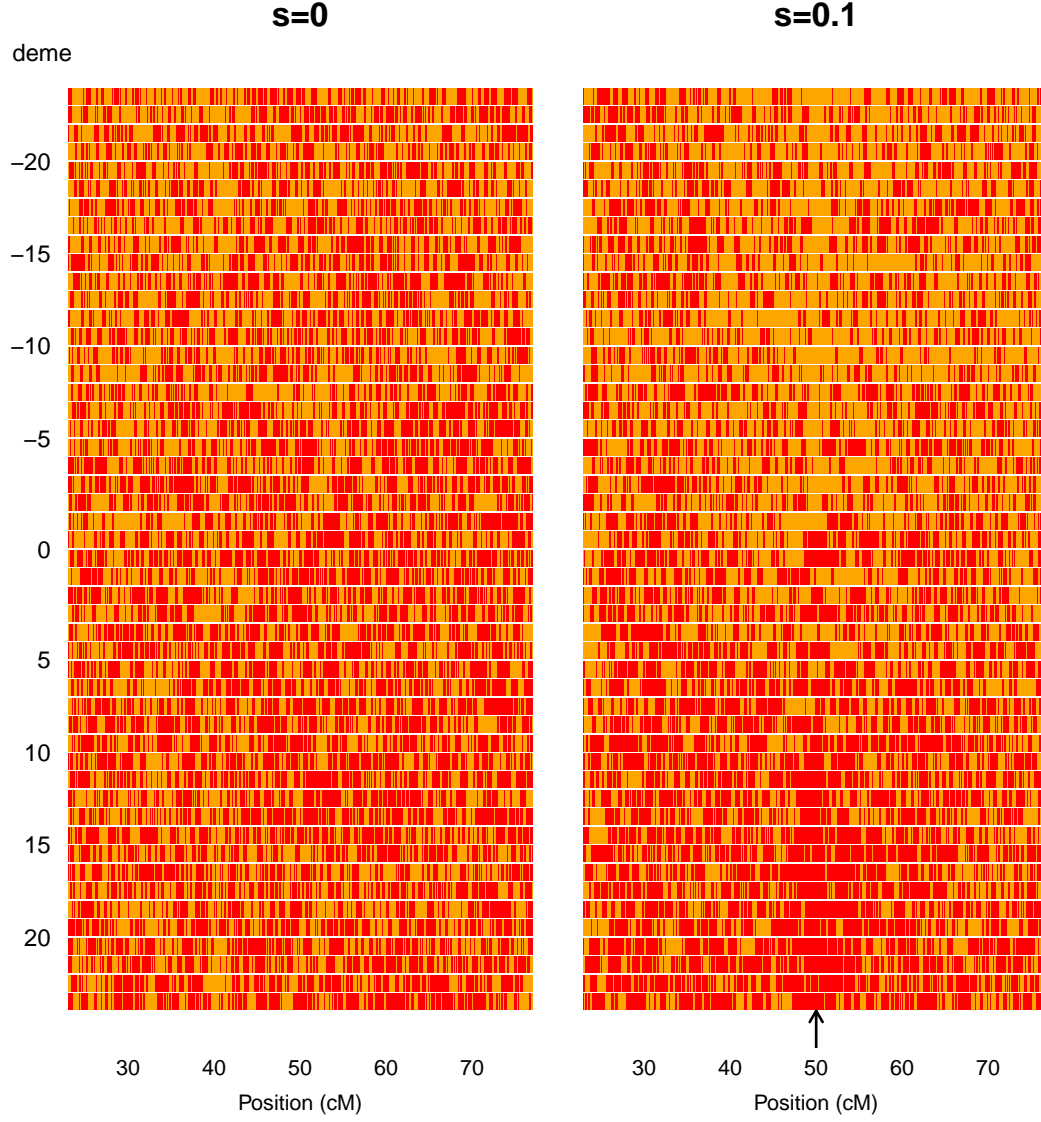


Figure 4: **Ancestry blocks for randomly sampled chromosomes** across a hybrid zone of age  $T = 1000$ . Here we compare chromosomes of length 1M from a neutral zone to a zone that has a single under-dominant locus with  $s = 0.1$  in the middle of the chromosome (indicated by black arrow). Red blocks along the chromosome denote ancestry  $B$ , and orange blocks are ancestry  $A$ . The simulation was performed in a population with 50 demes, each with 500 diploids, and  $\sigma = 1$ . An analogous figure at  $T = 100$  is shown in Figure ??.

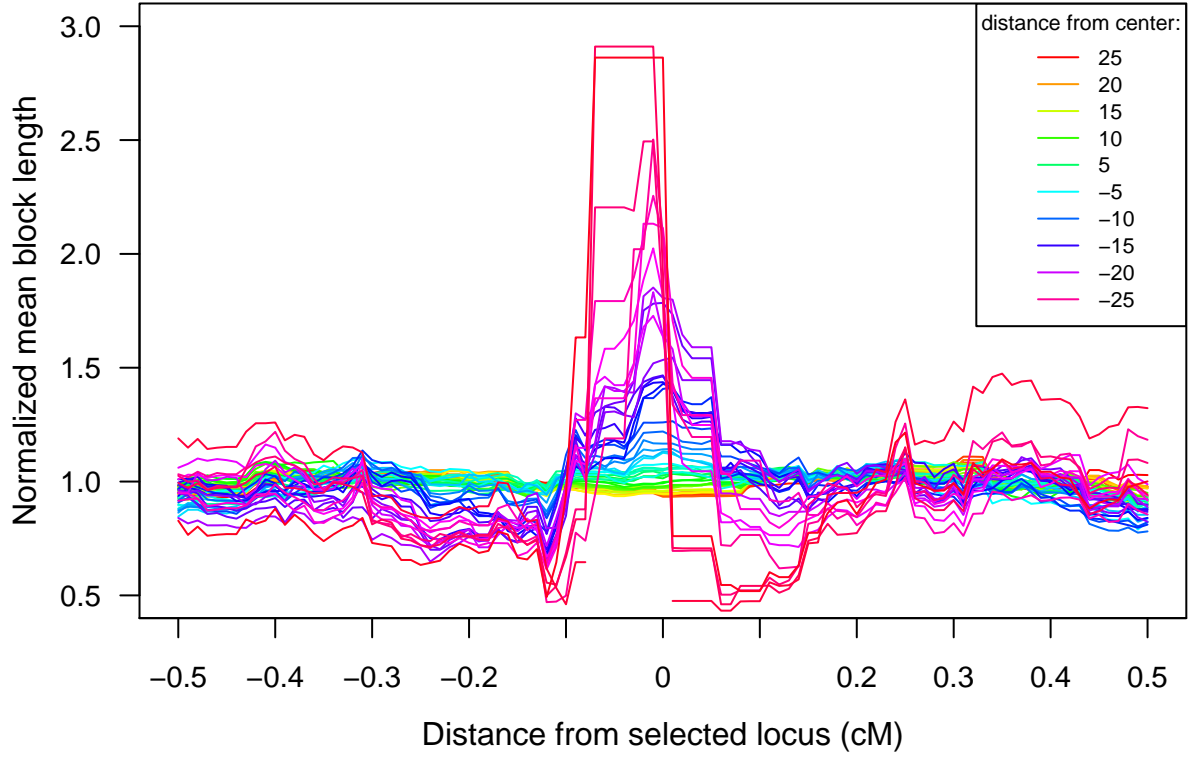


Figure 5: **Normalized mean enclosing block length**,  $\bar{l}_B(m, x)$ , along a simulated chromosome of length  $1M$  with selected locus of  $s = 0.01$  at position  $0.5M$  in a hybrid zone of  $T = 1000$ . Here each line shows  $\bar{l}_B(m, x)$ , in a given deme some distance from the zone center. Chromosome were sampled across the hybrid zone, which consists of 50 demes, each containing 500 diploid individuals. The same simulation and statistic are shown in Fig. ??, on coarser genomic scale.

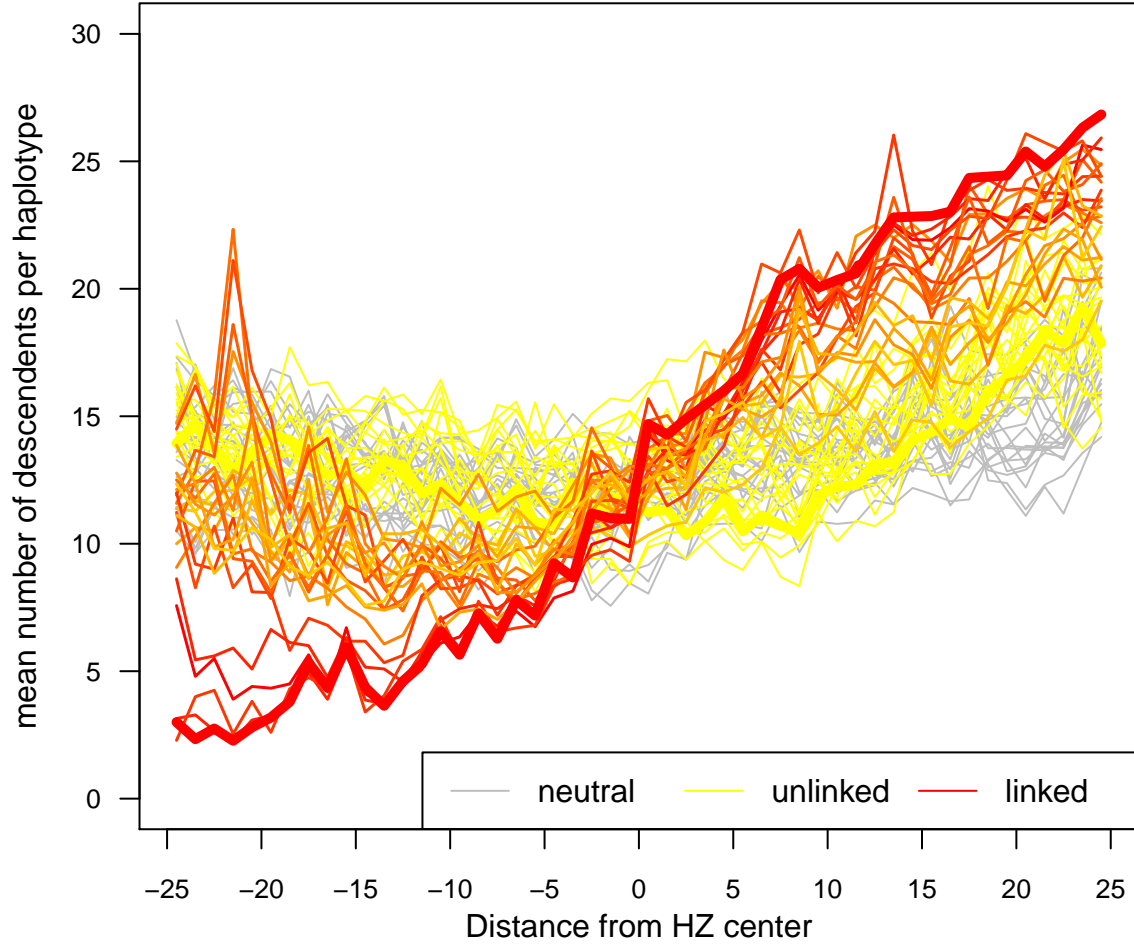


Figure 6: **Mean family size of haplotypes.** The number of individuals of ancestry  $B$  per number of ancestors ( $F_B(m, x)$ ) from secondary contact occurring  $T = 1000$  generations ago, represented across geographic space. The hybrid zone is 50 demes, each containing 500 diploid individuals. Each red or orange line represents a site some distance away (ranging from 0 - 0.5cM) from the selected site (here  $s = 0.01$  and  $\sigma = 1$ ). Yellow lines are corresponding positions on an unlinked chromosome with no selected loci, and grey lines are corresponding positions in a simulation with no selection. Bold lines depict the target of selection when present, and corresponding position on chromosomes not harboring any selected sites.

## The size of migrant families

Within-ancestry haplotype diversity, i.e., the number of ancestral haplotypes of each type, could provide additional information about whether introgression is through relatively few, successful migrants, or through many migrants that each contribute relatively little. In our simulations, the average local family size of a selected  $B$  allele ( $F_B(m, x)$ ) decreases with distance into the  $A$  side of the hybrid zone, and is lowest far away from the zone center, where ancestry  $B$  is at low frequency (Figure ??). Unlinked loci have local family sizes similar to neutral simulations, and loci linked to the selected locus have intermediate sizes. This pattern is consistent with the prediction that unfit lineages tend to be recent migrants, which will have smaller families.

## Discussion

Using a combination of theory and simulations, we present a description of the process of cline formation and haplotype structure in a relatively young (i.e., non-equilibrium) hybrid zone. We show why clines establish over time  $1/s$ , and why lineages of selected loci tend to move back towards their ‘ancestral home’ when in a geographic region where they are unfit. This occurs at speed  $\sigma\sqrt{s}$ . Based on this we predict, and observe in simulations, that blocks of ancestry surrounding these selected loci are longer, especially when distant from the center of the cline, than those surrounding neutral loci. This extends previous theoretical work on hybrid zones, which has primarily focused on stable clines in allele frequency. Additionally, our work suggests that the ancestry block length distribution can help detect targets of purifying selection in hybrid zones. The resolution of this approach is expected to scale with  $1/T$ , as this is the physical scale over which linked clines persist.

## Genomic signals associated with targets of selection

Popular approaches to identifying loci under selection in hybrid zones involve identifying alleles that are exceptional in terms of frequency across space, or genome-wide admixture proportion (??). The availability of genomic data has made it possible to use local ancestry as an additional source of information. In particular, ancestry deconvolution facilitated by programs such as hapMIX (?), LAMP (?) and fineSTRUCTURE (?) can inform the demographic history of hybridization/admixture from present day samples (e.g. ?). We described how selection against hybrid incompatibilities results in long contiguous blocks of ancestry around these loci. This is because regions surrounding selected loci do not readily introgress, and so introgressed alleles have been inherited from “their” side of the zone recently. We get quantitative understanding from the fact that lineages of loci under selection move as Brownian motion pushed at speed  $\sigma\sqrt{s}$  towards its ancestral side. In other words, disadvantageous alleles that have encroached deep into the other side of the hybrid zone have done so by chance and, since they do not persist for long on the “wrong” side, are likely to have done so relatively recently. Haplotypes surrounding the selected locus therefore have had relatively few opportunities for recombination with different ancestries and this is reflected in longer blocks of contiguous ancestry.

Through our simulations we find also that blocks of ancestry that have crossed the hybrid zone and are closely linked to the selected site without overlapping it (i.e., blocks that are adjacent to the block containing the selected site) are shorter than average for the spatial location (Figure ??). An intuitive explanation for this pattern is that loci physically linked to a selected site have recently come from the center of the zone or beyond. Compared to

other chromosomes in their new geographic location, these migrants will have on average longer  $B$  haplotypes at the selected locus, and shorter  $A$  haplotypes nearby.

Our results suggest that the statistic  $l(x, m)$  – the mean genetic length of ancestry surrounding a site of a given lineage – could help identify selected loci in hybrid zones. The ratio of mean adjacent block lengths,  $C(x, m)$ , also shows a very sharp peak at the selected site, suggesting that, despite the fact that haplotypes surrounding incompatibilities might be quite long, genome scans may have the power to extract fine-grained regions near selected loci from the large chunks of ancestry that will often flank these regions.

Our results focus on the length distribution of ancestry blocks — a statistic that can only be obtained in the few systems for which dense genetic markers and phase information is known. However, as the cost of sequencing continues to drop and long-phased reads become more common, the patterns described here could further aid in the identification of selected sites in hybrid zones in many non-model taxa. In the meantime, patterns of elevated pairwise LD, which is easily computed from readily available low-density and unphased genomic data, could offer an alternative path forward for empirical work. Importantly, the numerical solutions derived above readily and efficiently predict correlations in ancestry, and so our results can be applied immediately to genotype data for any taxa with markers placed on a genetic map.

Ancestry assignments have the additional benefit that relatively unbiased estimates are possible even with markers of problematic ascertainment such as those on SNP arrays. The power and resolution of these approaches depends strongly on the strength of selection, the time since secondary contact, and the strength of genetic drift: in our simulations, we found good power and resolution at  $s = .01$ ,  $T = 1000$  generations after secondary contact, and with hundreds of individuals per dispersal distance.

## Assumptions

Here we review the assumptions we made and their likely impact.

**The nature of selection** Although our assumed scenario of selection against heterozygotes at a single locus is uncommon in nature, we believe that this is unlikely to drastically influence our findings. As discussed in the Introduction, previous studies have demonstrated that models of underdominant selection share similar properties with more realistic models, e.g., selection on an ecological gradient or multilocus, epistatic selection (???).

Our assumption of a single selected locus may have greater consequences. Our model is most relevant to scenarios with few targets of selection scattered throughout the genome, and therefore, our predictions may differ significantly from situations in which the density of

selected sites is higher. Having numerous selected sites within one ancestry block increases the strength of selection (?), and the relatively short map distances between linked incompatibilities generates a longer unit which is not readily broken up by recombination (?). Both factors are expected to result in regions that are surrounded by even longer segments of unbroken ancestry, and will impact expectations of genome-wide patterns of clines and block lengths, as well as the resolution to which one could detect targets of selection (??).

To speed calculations and simulations, we assumed a homogeneous, one-dimensional geographical range. Our analytical results further assumed a large population density, effectively working with a deterministic model that ignores coalescence and associated stochasticity. In contrast, our simulations model regularly spaced demes of finite size. In reality, populations may be patchily connected, especially at the edges of species ranges where hybrid zones may occur. The degree to which inhomogeneous geography would affect the predictions depends on how patchy the zone is; the differential equations provide a way to evaluate this in specific circumstances.

Extending our analytical results to capture stochasticity arising from coalescence/pedigree structure represents an important future direction. Indeed, correlated fluctuations visible in simulations (e.g., Figure ??) are likely due to coupling due to demographic stochasticity; and simulations at lower density show larger fluctuations than those at higher density. Furthermore, ignoring pedigree structure can result in an underestimate of covariance in ancestry, as we have ignored additional sharing of ancestry through shared genealogy (?); but there is nonetheless good agreement between analytic predictions and simulations (which include an explicit pedigree structure).

## Theory and simulation

We have taken two complementary approaches, using both simulation and theory, and comparing the two. As usual, simulations make fewer biological simplifications, while theory provides more generalizable conclusions. To do this, we have described the branching diffusion process that approximates the lineages along which haplotypes are inherited. Since the expected motion of a lineage depends on the local frequencies of the selected alleles, these diffusions are time-inhomogeneous.

The diffusion model for lineages predicts that quantities of interest solve sets of coupled partial differential equations (PDE), which we have written down. As there are no known analytical solutions to these PDE, we have constructed numerical solutions (and provide the source code for doing this). A main role of these solutions in our work has been to verify that theory based on the diffusion model of lineages matches realistic, individual-based models. These solutions easily and quickly provide predictions of joint frequencies at small numbers of

loci: about 1 second to compute predicted clines, as opposed to hours for the full simulation. However, due to the high dimensionality of the haplotype problem (spatial position  $\times$  time  $\times$  endpoints of the haplotype), numerical solutions for mean haplotype lengths along the genome can be as computationally intensive as simulations (although are substantially less noisy). More work could be done to develop more efficient methods of solution, but it may be better to perform more biologically realistic forwards-time simulations that includes coalescence and drift to better characterize the block length distribution. If however the correlation in ancestry, rather than the block-length distribution, is of interest the PDE approach may be preferable because it is easily modified to provide predictions for spatial and temporally inhomogeneous systems – for instance, across maps of real landscapes.

## Patterns of divergence

A number of studies have described heterogeneous patterns of genetic divergence across the genome. Work on these “islands of divergence” (??) and related patterns have been largely descriptive (??). Our study here contributes to a model-based understanding of how migration and selection may influence such patterns across the genome of hybridizing populations. Overall, focusing on lengths of ancestry blocks across the genome brings focus to the processes of migration and selection rather than high-level summaries that are somewhat abstracted from the evolutionary process.

## Adaptive introgression

While our focus has been on hybrid incompatibilities, unconditionally adaptive loci are expected to easily introgress across hybrid zones (????). Future work could take a similar approach to understand how positive selection shapes ancestry block lengths, and predict signatures of adaptive introgression in hybridizing populations using similar statistics presented here. These could eventually be combined to gain a fuller understanding of the forces shaping patterns of introgression in hybrid zones. In particular, beneficial alleles tightly linked to incompatibilities cannot introgress until recombination separates them; our model *provides some rough expectation on how quickly this should happen.*

## Acknowledgements

We thank Graham Coop, Nancy Chen, Emily Josephs, Kristin Lee, Nick Barton and two anonymous reviewers for helpful feedback and suggestions. PR was supported in part by NSF grant DBI-1262645 and a Sloan Research Fellowship. *More info here*



## References

- Arnold ML (2004) Transfer and origin of adaptations through natural hybridization: Were Anderson and Stebbins right? *The Plant Cell*, **16**, 562–570.
- Baird S, Barton N, Etheridge A (2003) The distribution of surviving blocks of an ancestral genome. *Theoretical Population Biology*, **64**, 451–471.
- Barton N (1986) The effects of linkage and density-dependent regulation on gene flow. *Heredity*, **57**, 415–426.
- Barton N, Bengtsson B (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **56**, 357–376.
- Barton N, Partridge L (2000) Limits to natural selection. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **22**, 1075–84.
- Barton NH (1979a) Gene flow past a cline. *Heredity*, **43**, 333–339.
- Barton NH (1979b) The dynamics of hybrid zones. *Heredity*, **43**, 341–359.
- Barton NH (1983) Multilocus Clines. *Evolution*, **37**, 454–471.
- Barton NH (2008) The effect of a barrier to gene flow on patterns of geographic variation. *Genetics Research*, **90**, 139–49.
- Barton NH, Depaulis F, Etheridge AM (2002) Neutral evolution in spatially continuous populations. *Theoretical Population Biology*, **61**, 31–48.
- Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In *Hybrid Zones and the Evolutionary Process* (herausgegeben von R Harrison), S. 13–45. Oxford University Press.
- Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.
- Barton NH, Hewitt GM (1989) Adaptation, speciation and hybrid zones. *Nature*, **341**, 497–503.
- Barton NH, Kelleher J, Etheridge AM (2010) A new model for extinction and recolonization in two dimensions: Quantifying phylogeography. *Evolution*.
- Barton NH, Shpak M (2000) The effect of epistasis on the structure of hybrid zones. *Genetical Research*, **75**, 179–198.
- Bazykin AD (1969) Hypothetical mechanism of speciation. *Evolution*, **23**, 685–687.

- Chapman NH, Thompson EA (2002) The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, **162**, 449–458.
- Christiansen FB, Andreassen V, Poulsen ET (1995) Genotypic proportions in hybrid zones. *Journal of Mathematical Biology*, **33**, 225–249.
- Cox JT, Durrett R (1995) Hybrid zones and voter model interfaces. *Bernoulli*, **1**, 343–370.
- Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates.
- Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.
- Durrett R, Zähle I (2007) On the width of hybrid zones. *Stochastic Process. Appl.*, **117**, 1751–1763.
- Eddelbuettel D, Francois R (2011) Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18.
- Felsenstein J (1975a) Genetic drift in clines which are maintained by migration and natural selection. *Genetics*, **81**, 191–207.
- Felsenstein J (1975b) A pain in the torus: Some difficulties with models of isolation by distance. *The American Naturalist*, **109**, 359–368.
- Fisher R (1954) A fuller theory of “junctions” in inbreeding. *Heredity*, **8**, 187–197.
- Gompert Z, Parchman TL, Buerkle CA (2012) Genomics of isolation in hybrids. *Philosophical transactions of the Royal Society of London B Biological Sciences*, **367**, 439–50.
- Gravel S (2012) Population genetics models of local ancestry. *Genetics*, **191**, 607–19.
- Hanson WD (1966) Effects of partial isolation (distance), migration, and different fitness requirements among environmental pockets upon steady state gene frequencies. *Biometrics*, **22**, pp. 453–468.
- Hartl DL, Clark AC (1989) *Principles of Population Genetics*. Sinauer, Sunderland, MA.
- Hellenthal G, Busby GBJ, Band G, *et al.* (2014) A genetic atlas of human admixture history. *Science*, **343**, 747–51.
- Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

- Kruuk LE, Baird SJ, Gale KS, Barton NH (1999) A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–71.
- Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS genetics*, **8**, e1002453.
- Liang M, Nielsen R (2014) The lengths of admixture tracts. *Genetics*, **197**, 953–967.
- Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.
- May R, Endler J, McMurtrie RE (1975) Gene frequency clines in the presence of selection opposed by gene flow. *American Naturalist*, **109**, 659–676.
- Nagylaki T (1975) Conditions for the existence of clines. *Genetics*, **80**, 595–615.
- Nagylaki T (1978) Random genetic drift in a cline. *Proceedings of the National Academy of Sciences of the United States of America*, **75**, 423–6.
- Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.
- Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.
- Polechová J, Barton N (2011) Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, **189**, 227–235.
- Pool JE (2015) The mosaic ancestry of the Drosophila genetic reference panel and the *D. melanogaster* reference genome reveals a network of epistatic fitness interactions. *Molecular Biology and Evolution*, **32**, 3236–3251.
- Porter AH, Wenger R, Geiger H, Scholl A, Shapiro AM (1997) The pontia daplidice-edusa hybrid zone in northwestern italy. *Evolution*, **51**, 1561–1573.
- Price AL, Tandon A, Patterson N, *et al.* (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, **5**, e1000519.
- Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, **82**, 290 – 303.
- Schumer M, Brandvain Y (2016) Determining epistatic selection in admixed populations. *Molecular Ecology*, S. ??–??

- Sedghifar A, Brandvain Y, Ralph P, Coop G (2015) The spatial mixing of genomes in secondary contact zones. *Genetics*, **201**, 243–61.
- Slatkin M (1973) Gene flow and selection in a cline. *Genetics*, **75**, 733–756.
- Slatkin M (1975) Gene flow and selection in a two-locus system. *Genetics*, **81**, 787–802.
- Slatkin M, Maruyama T (1975) Genetic drift in a cline. *Genetics*, **81**, 209–22.
- Soetaert K, Meysman F (2012) Reactive transport in aquatic ecosystems: Rapid model prototyping in the open source software R. *Environmental Modelling & Software*, **32**, 49–60.
- Soetaert K, Petzoldt T, Setzer RW (2010) Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, **33**, 1–25.
- Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.

## Data Accessibility

All scripts used in production of this paper are available at <https://github.com/petrelharp/clinal-lineages> under the GPLv3 software license, and on Data Dryad <http://datadryad.org/XXXX>.

## A Rescaling a discrete model to obtain the lineage motion

For concreteness, here we describe a discrete model that rescales to the continuous model we consider.

In this discrete model, the total number of individuals at location  $x$  is  $N(x)$ , and we count “individuals” as *haploid*, so each individual is of type either A or B at the selected locus, and the proportion of individuals of type A and location  $x$  and time  $t$  is  $p(x, t)$  (but, we often neglect the  $t$ ). Suppose that type A individuals at location  $x$  reproduce at rate  $s_A(x)$ , and likewise type B at rate  $s_B(x)$ . Assuming locally random mating, we will then have that  $s_A(x) = 1 - s(1 - p(x))$  and  $s_B(x) = 1 - sp(x)$ .

At reproduction, individuals recombine with others in the same location, with recombination occurring between the locus we follow and the selected locus with probability  $r$ , and the offspring choose a new location  $y$  with probability  $m(x, y)$ . The population dynamics

are random, but suppose that  $N(x)$  is sufficiently large that these do not vary substantially with time. Suppose this is a Moran model. There are four things that can happen:

$x \xrightarrow{AA} y$  One type A individual at location  $x$  reproduces, either does not recombine or recombines with another type A, and sends the offspring to  $y$ .

$x \xrightarrow{AB} y$  An individual at location  $x$  reproduces, recombines with the other type, and sends to  $y$  an offspring who inherits at the selected locus from the type  $B$  parent and the at the neutral locus from the type A parent.

$x \xrightarrow{BA} y$  An individual at location  $x$  reproduces, recombines with the other type, and sends to  $y$  an offspring who inherits at the selected locus from the type A parent and the at the neutral locus from the type  $B$  parent.

$x \xrightarrow{BB} y$  One type  $B$  individual at location  $x$  reproduces, either does not recombine or recombines with another type  $B$ , and sends the offspring to  $y$ .

These four things happen at rates:

$$x \xrightarrow{AA} y \quad w_{AA}(x, y) = p(x)s_A(x)(1 - r(1 - p(x)))N(x)m(x, y) \quad (7)$$

$$x \xrightarrow{AB} y \quad w_{AB}(x, y) = rp(x)(1 - p(x))\frac{s_A(x) + s_B(x)}{2}N(x)m(x, y) \quad (8)$$

$$x \xrightarrow{BA} y \quad w_{BA}(x, y) = rp(x)(1 - p(x))\frac{s_A(x) + s_B(x)}{2}N(x)m(x, y) \quad (9)$$

$$x \xrightarrow{BB} y \quad w_{BB}(x, y) = (1 - p(x))s_B(x)(1 - rp(x))N(x)m(x, y) \quad (10)$$

For instance, there are  $p(x)N(x)$  type A individuals at  $x$ , that reproduce at rate  $s_A(x)$ ; the chance that each reproduction is with a type  $B$  is  $(1 - p(x))$ , and the chance there is a recombination that gives the offspring genotype  $AB$  is  $r/2$ ; the offspring has probability  $m(x, y)$  to disperse to  $y$ ; the total rate at which individuals at  $x$  with the  $A$  allele produce offspring at  $y$  with the  $B$  allele is the product of these terms,  $p(x)N(x)s_A(x)(1 - p(x))r/2$ . The second and third rates (for  $x \xrightarrow{AB} y$  and  $x \xrightarrow{BA} y$ ) are the same; they are this value plus the corresponding term for individuals carrying the  $B$  allele producing offspring that carry the  $A$  allele.

Note that at equilibrium, we require  $N$  to solve

$$0 = \sum_y N(y)m(y, x) \{p(y)s_A(y)(1 - p(x)) - (1 - p(y))s_B(y)p(x)\}.$$

## Lineage movement

These rates tell us the rates at which a lineage will move, backwards in time. For instance, the rate at which a lineage at the selected locus currently in a type A individual at location  $x$  jumps to another type A individual at location  $y$  is equal to the rate of influx of migrants

756 from  $y$  divided by the number of A alleles at  $x$ , or

$$\begin{aligned} r_A(x, y) &= \frac{w_{AA}(y, x)}{p(x)N(x)} \\ &= N(y)p(y)s_A(y)m(y, x) \frac{1}{N(x)p(x)}. \end{aligned}$$

757 Let  $X_t$  denote the position of the lineage of a selected locus of time  $A$  at time  $t$  in the  
758 past, and let  $f$  be test function with  $f(\rho) = 0$ . Then,

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[f(X_t) | X_0 = x] &= \sum_y r_A(y, x)(f(y) - f(x)) \quad (11) \\ &= \frac{1}{N(x)p(x)} \sum_y N(y)p(y)s_A(y)m(y, x)(f(y) - f(x)). \quad (12) \end{aligned}$$

## 759 Diffusion limit

760 Now suppose that  $m(x, y)$  is symmetric, and depends on a parameter  $\sigma$  so that as  $\sigma \rightarrow 0$ ,  
761 the associated random walk converges to Brownian motion, so that for an arbitrary smooth  
762 function  $f$ ,

$$\lim_{\sigma \rightarrow 0} \sum_y \frac{m(x, y)(f(y) - f(x))}{\sigma^2} = \frac{1}{2} \frac{d^2}{dx^2} f(x).$$

763 Write  $f'(x) = \frac{d}{dx} f(x)$ , and note that

$$\begin{aligned} \frac{1}{\sigma^2} \sum_y g(y)m(y, x)(f(y) - f(x)) &= \frac{1}{\sigma^2} \sum_y m(y, x) (g(y)f(y) - g(x)f(x) + (g(x) - g(y))f(x)) \\ &= \frac{1}{\sigma^2} \sum_y m(y, x) (g(y)f(y) - g(x)f(x)) \\ &\quad - f(x) \frac{1}{\sigma^2} \sum_y m(y, x)(g(y) - g(x)) \\ &\xrightarrow{\sigma \rightarrow 0} \frac{1}{2} \frac{d^2}{dx^2} (g(x)f(x)) - \frac{1}{2} f(x) \frac{d^2}{dx^2} g(x) \\ &= \frac{1}{2} (g(x)f'(x) + 2g'(x)f'(x) + f(x)g''(x) - f(x)g''(x)), \end{aligned}$$

764 which tells us the differential operator that best approximates the discrete sum:

$$\frac{1}{\sigma^2} \sum_y g(y)m(y, x)(f(y) - f(x)) \xrightarrow{\sigma \rightarrow 0} \frac{1}{2} g(x)f''(x) + g'(x)f'(x). \quad (13)$$

Under these assumptions, combining (??) and (??),

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[f(X_{t/\sigma^2}) | X_0 = x] &\xrightarrow{\sigma \rightarrow 0} \frac{1}{2} s_A(x) \frac{d^2}{dx^2} f(x) + \frac{1}{N(x)p(x)} \frac{d}{dx} \{N(x)p(x)s_A(x)\} \frac{d}{dx} f(x) \\ &= \frac{1}{2} s_A(x) f''(x) + (s'_A(x) + \log(N(x)p(x))' s_A(x)) f'(x), \end{aligned}$$

i.e.  $X_{t/\sigma^2}$  converges to a diffusion with mean displacement (“drift” in diffusion terminology)  $\frac{1}{N(x)p(x)} \frac{d}{dx} (N(x)p(x)s_A(x))$  and killed at rate  $\rho k(x)$ .

In our case, since  $s_A(x) = 1 - s(1 - p(x))$ , the drift is  $p'(x)/p(x)$  to first order in  $s$ ; the time scaling by  $\sigma^2$  implies that the Brownian noise and the mean displacement should both be scaled by  $\sigma^2$ .

## B Numerical calculation of haplotype probabilities

In this section, we give details for how we found numerical solutions to the partial differential equations (PDE) of the text, which are all of reaction-diffusion type. The R code, with worked examples, is available in our git repository. Spatial grids were usually chosen to have at least four grid sites per dispersal distance, but using substantially finer grids did not affect the results.

The forwards-time evolution of the selected alleles, equation (??), presents no difficulty; we use the ReacTran package (?) to compute discrete approximations to the diffusion term, and the deSolve package (?) to solve the equation.

The equations (??) describing probabilities that a lineage descends from  $A$  ancestry, conditional on the linked selected allele, required some more attention. First,  $s > 0.1$  the system of equations can be *stiff* (as is commonly observed for reaction-diffusion equations), and hence slow to solve, because of the extremely steep slope of the selected allele frequency  $p(x, t)$ . In practice we used  $s < 0.1$ . Second, the ReacTran function `tran.1D` that converts the diffusion portion of the PDE into a system of ODE contains a term like

$$\frac{1}{A(x)} \frac{d}{dx} A(x) f(x) = \frac{d}{dx} f(x) + f(x) \frac{d}{dx} \log A(x),$$

where  $A(x)$  is the interface area between grid cells, and  $f(x)$  is the flux. (See ? for more details.) Discrete approximations of the left-hand side, as implemented in ReacTran, run into numerical difficulties if  $A(x)$  is small, and in our case,  $A(x)$  is equal to  $p$ , the local frequency of the selected  $A$  allele. To avoid this, we made minor modifications to the `tran.1D` to provide a discrete approximation to the right-hand side.

Haplotype probabilities are obtained from equations (??), which is a coupled system of integro-differential equations in three variables plus time. One method for solution would

be via a Wild sum over the number of recombination events, as we did in ?. Here, we solved the equations numerically, again discretizing the equations and using the ode1D function of the deSolve package. The reaction-diffusion part is the same as for equations (??). The functions  $g_A(x, t; a, b)$  and  $g_B(x, t; a, b)$  are functions of space ( $x$ ), time ( $t$ ), and the endpoints of the block in question ( $a$  and  $b$ , with  $a < b$ ). The second integral in (??) is

$$\int_a^b g_A(a, \theta) g_A(\theta, b) d\theta. \quad (14)$$

Conceptually, this is an integral transformation  $g(a, b) \mapsto \int_a^b g(a, \theta) g(\theta, b) d\theta$ ; the reason it appears here is that for the entire segment  $(a, b)$  to be of ancestry  $A$ , if there was a recombination at  $\theta$ , the two segments  $(a, \theta)$  and  $(\theta, b)$  must both be of ancestry  $A$  (and, ignoring coalescence, these probabilities are independent). Suppose we have divided the segment of chromosome into a regular grid, say,  $r_1 < \dots < r_n$ . The natural discretization approximates this transformation by a sum, and is equivalent to keeping track of only a finite number of loci. Writing  $g_A(i, j)$  for  $g_A(x, t; r_i, r_j)$ , the discrete transformation we use corresponding to (??) is  $g_A(i, j) \mapsto$  the sum

$$\sum_{k=i}^{j-1} (r_{k+1} - r_k) g_A(i, k) g_A(k+1, j).$$

This is the correct term for the process only tracking loci on the grid, because for all the alleles at  $r_i, r_{i+1}, \dots, r_j$  to be of ancestry  $A$ , when a recombination occurs between  $r_k$  and  $r_{k+1}$ , the sequences of alleles at  $r_i, \dots, r_k$  and at  $r_{k+1}, \dots, r_j$  must each be of ancestry  $A$ . The first integral term in (??) is

$$\int_a^0 g_B(a, \theta) g_A(\theta, b) d\theta + \int_0^b g_A(a, \theta) g_B(\theta, b) d\theta \quad (15)$$

We now require that one of the grid points along the chromosome is exactly at the selected site; say this is  $r_\ell = 0$ . The discrete term corresponding to (??) is

$$\sum_{k=i}^{\ell-1} (r_{k+1} - r_k) g_B(i, k) g_A(k+1, j) + \sum_{k=\ell}^{j-1} (r_{k+1} - r_k) g_A(i, k) g_B(k+1, j).$$

Since these are not easily vectorizable in R, for efficiency we implemented the discrete transformations in C, using the Rcpp package (?). In implementing these, we kept track of  $g_A(i, j)$  in a vector in the order that the upper triangular elements of a matrix are encountered when traversing the matrix column-wise, allowing for efficient computation of the sum.



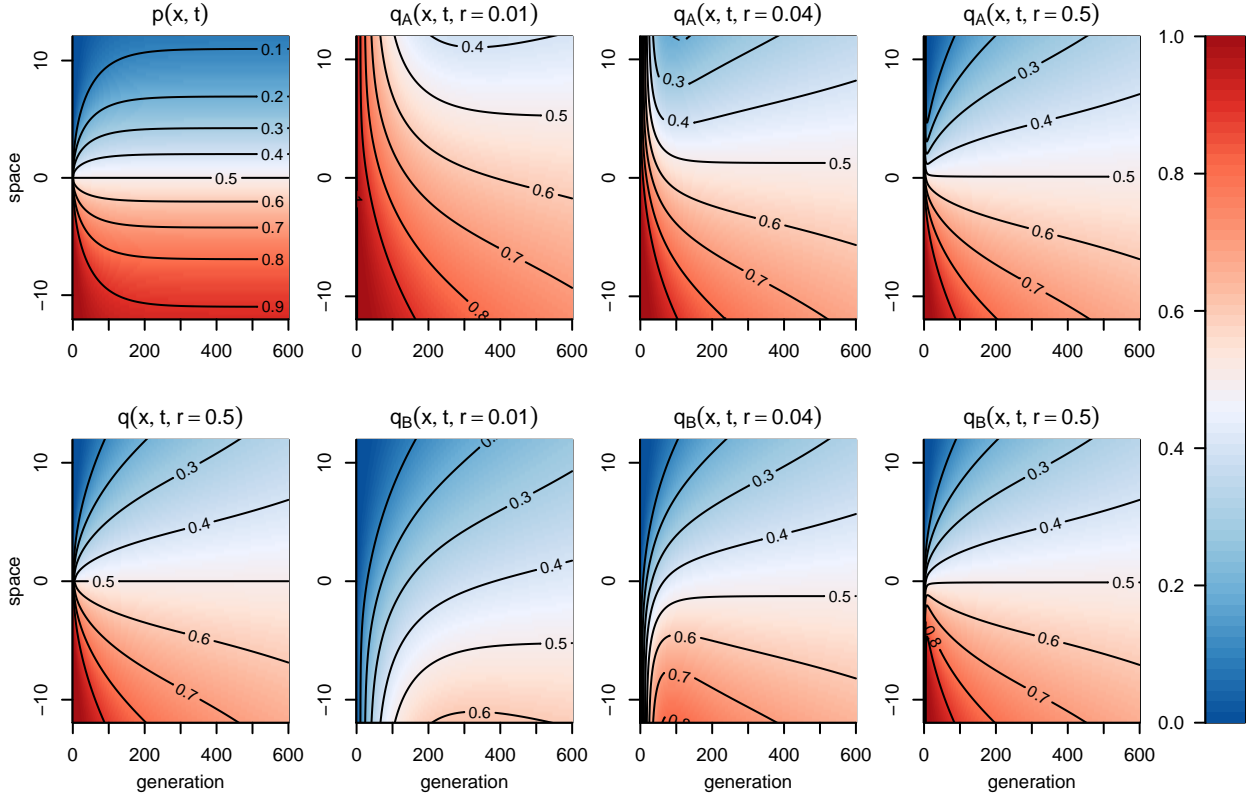


Figure S1: **Probabilities of A ancestry**, across space (vertical axis, in units of  $\sigma$ ) and time (horizontal axis, in generations). In each plot, color corresponds to the expected frequency of A ancestry at a particular location in time and space. The selection coefficient is  $s = .02$ . **Top left:** at the selected site, showing establishment and stabilization of the cline on a time scale of  $1/s = 50$  generations. **Bottom left:** at an unlinked site, with cline flattening continuing with  $\sqrt{t}$ . Remaining figures show frequencies of A ancestry *conditional* on the ancestry at the selected site, at different distances from the selected site ( $r = .01, .04$ , and  $0.5$  Morgans), as described in the text (see definition of  $q_z(x, t, r)$ ). See figure ?? for the same figure over a shorter period of time.

**A note on symmetry:** The equations we present have the symmetry that they are invariant after exchanging  $A$  and  $B$  and reversing space. For instance,  $q_A(x, t, r) = 1 - q_B(-x, t, r)$ . Using this fact would speed up the code by a factor of two, at the cost of generality: as written, it would be easy to modify the code to allow space to be inhomogeneous (which would break this symmetry).

## C Supplementary Figures

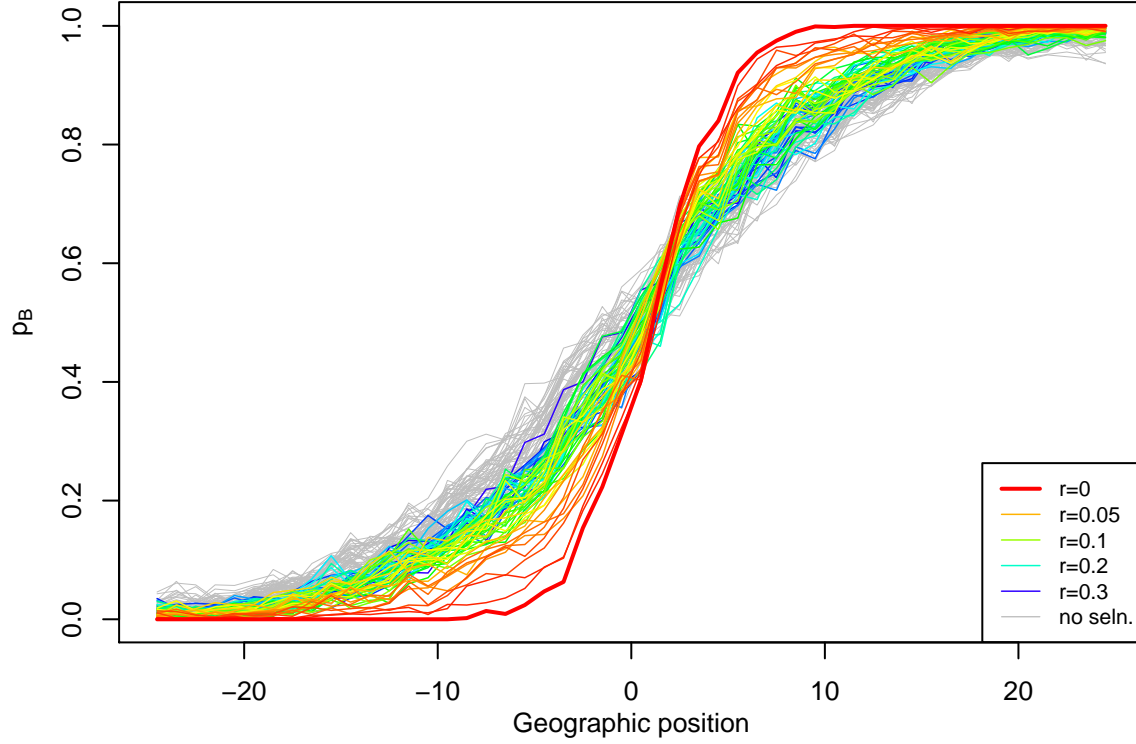


Figure S2: Frequency of ancestry  $B$ , across geography at different physical positions on the genome, simulated for a hybrid zone  $T = 100$  generations after secondary contact, with  $s = 0.1$ , using 50 demes, each with 500 diploid individuals and  $\sigma = 1$ . Each line represents a locus some distance  $r$  away from the true target of selection with colors corresponding to different values of  $r$ , transitioning from red (tight linkage to selected site) to blue (distantly linked). Grey lines represent the same positions from a simulation with identical parameters except that  $s = 0$ . Corresponding theoretical quantities are shown juxtaposed in Figure ??; the same plot is shown with weaker selection in Figure ?? and at a longer time in Figure ??.

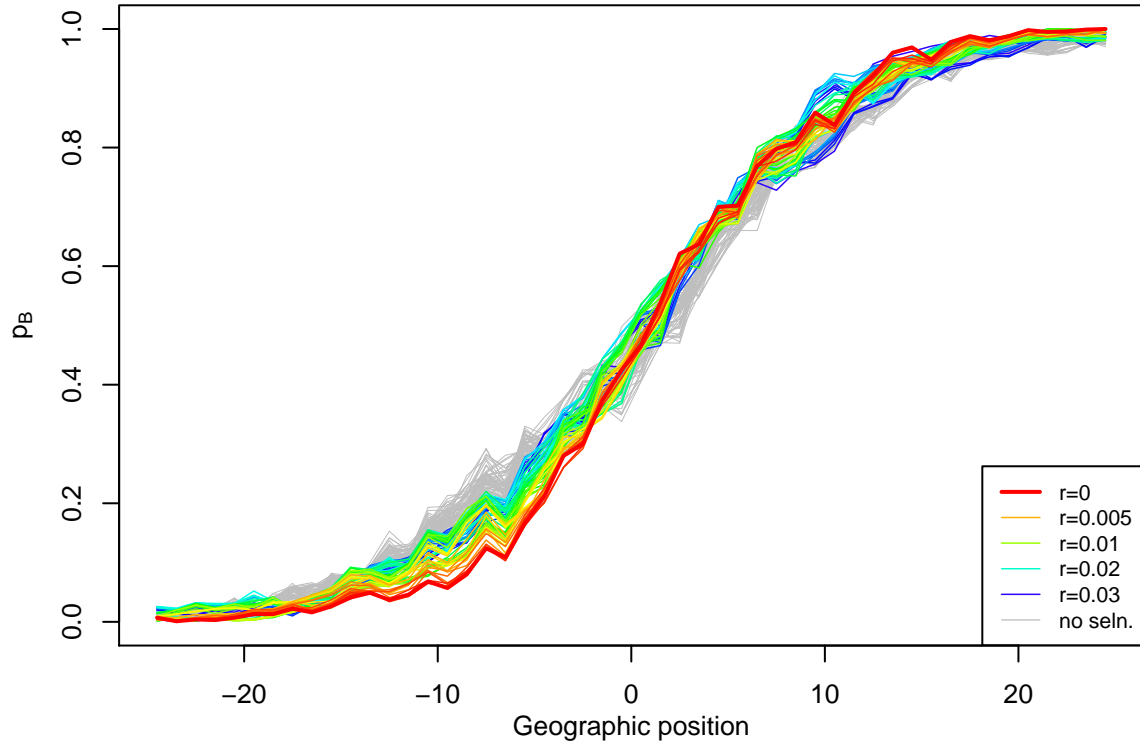


Figure S3: Frequency of ancestry  $B$ , across geography and at several physical positions on the genome, simulated for a hybrid zone  $T = 100$  generations after secondary contact, and with  $s = 0.01$ . 50 demes, each with a population size of 500 diploid individuals. Each line represents a locus some distance  $r$  away from the true target of selection, and  $r = 0$  represents the locus that is under selection. Grey lines represent the same positions from a simulation with identical parameters except that  $s = 0$ .

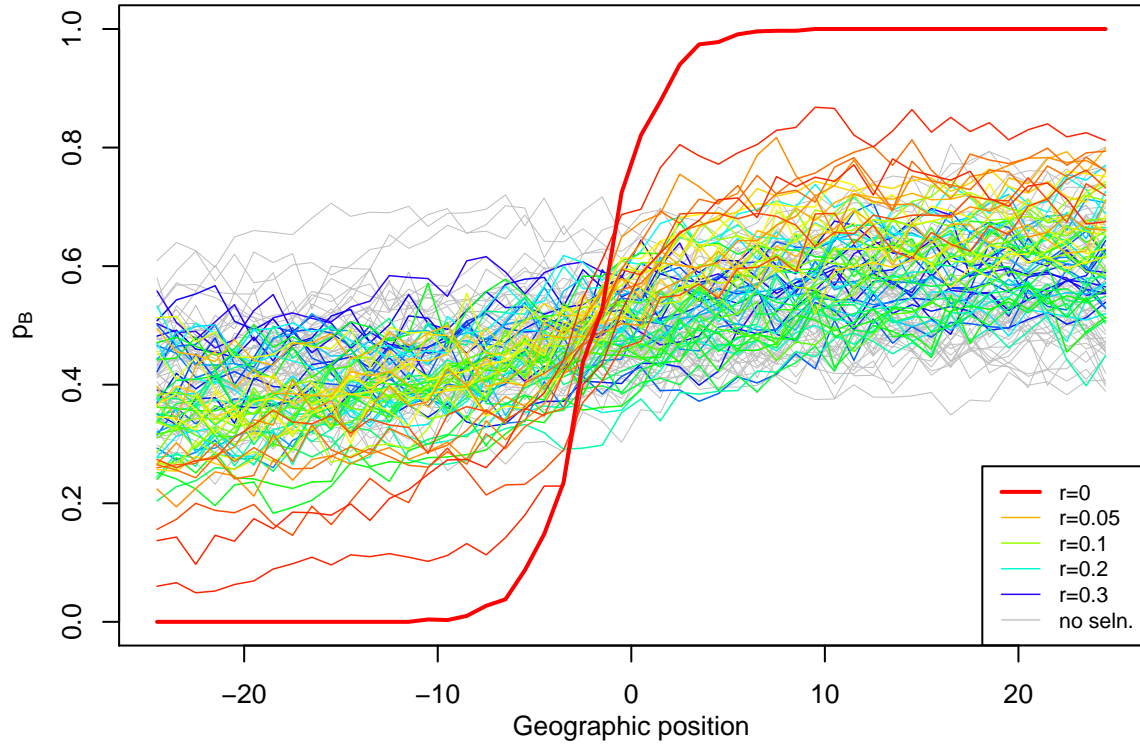


Figure S4: Frequency of ancestry  $B$ , across geography and at several physical positions on the genome, simulated for a hybrid zone  $T = 1000$  generations after secondary contact, and with  $s = 0.1$ . 50 demes, each with a population size of 500 diploid individuals. Each line represents a locus some distance  $r$  away from the true target of selection, and  $r = 0$  represents the locus that is under selection. Grey lines represent the same positions from a simulation with identical parameters except that  $s = 0$ .

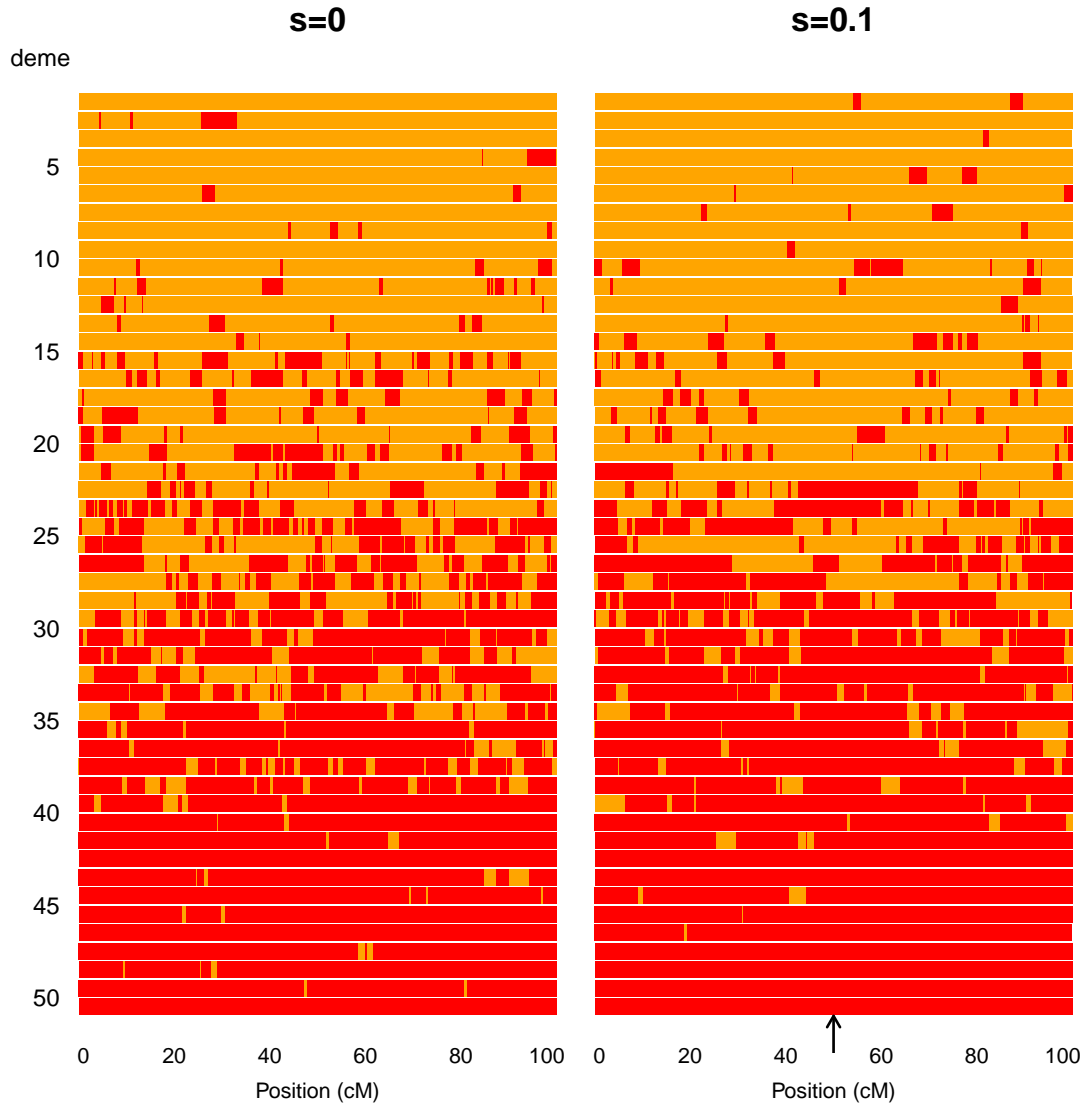


Figure S5: Randomly sampled chromosomes across a hybrid zone of age  $T = 100$ . Here we compare chromosomes of length 1M from a neutral zone to one that has a single under-dominant locus ( $s = 0.1$ ) in the middle of the chromosome (indicated by black arrow). Red blocks along the chromosome denote ancestry  $B$ , and orange blocks are ancestry  $A$ .

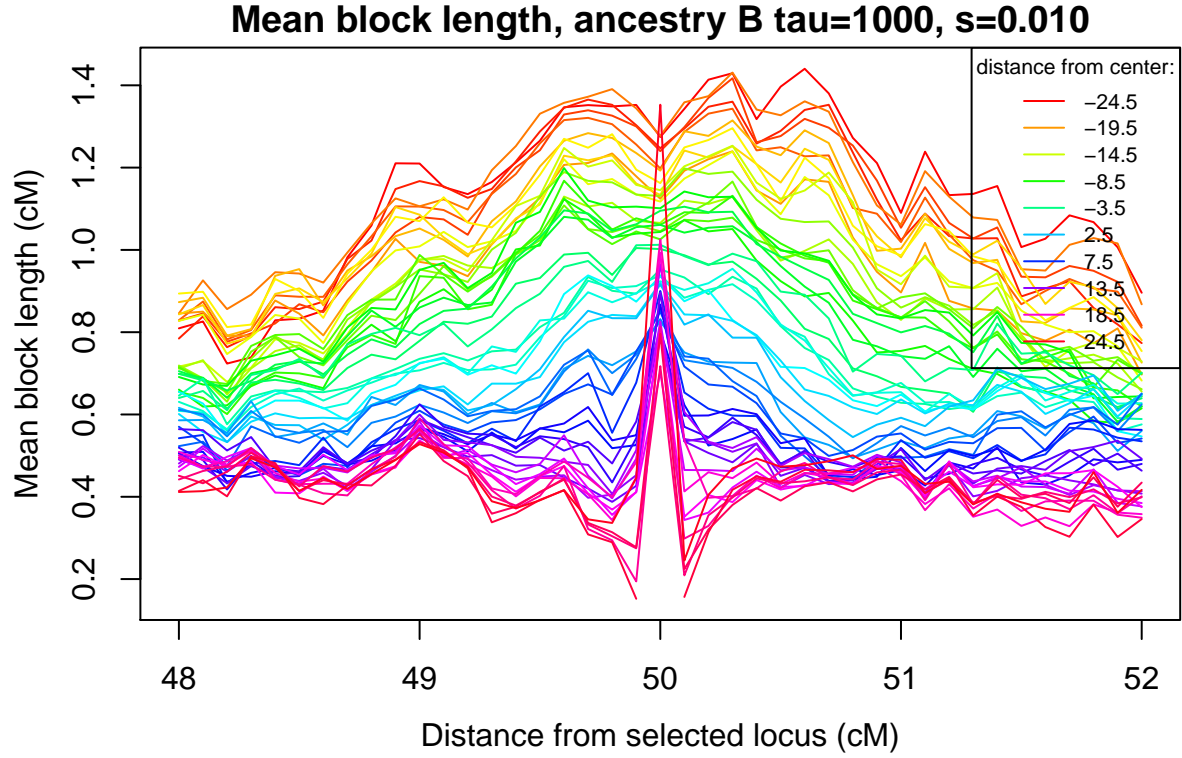


Figure S6: **Mean haplotype lengths of  $B$  haplotypes**,  $l_B(x, m)$ , across the genome (horizontal axis) and at different spatial locations (colored lines), from a simulation with 50 demes having 500 individuals each,  $s = 0.01$ ,  $\sigma = 1$ , and after  $T = 1000$  generations.

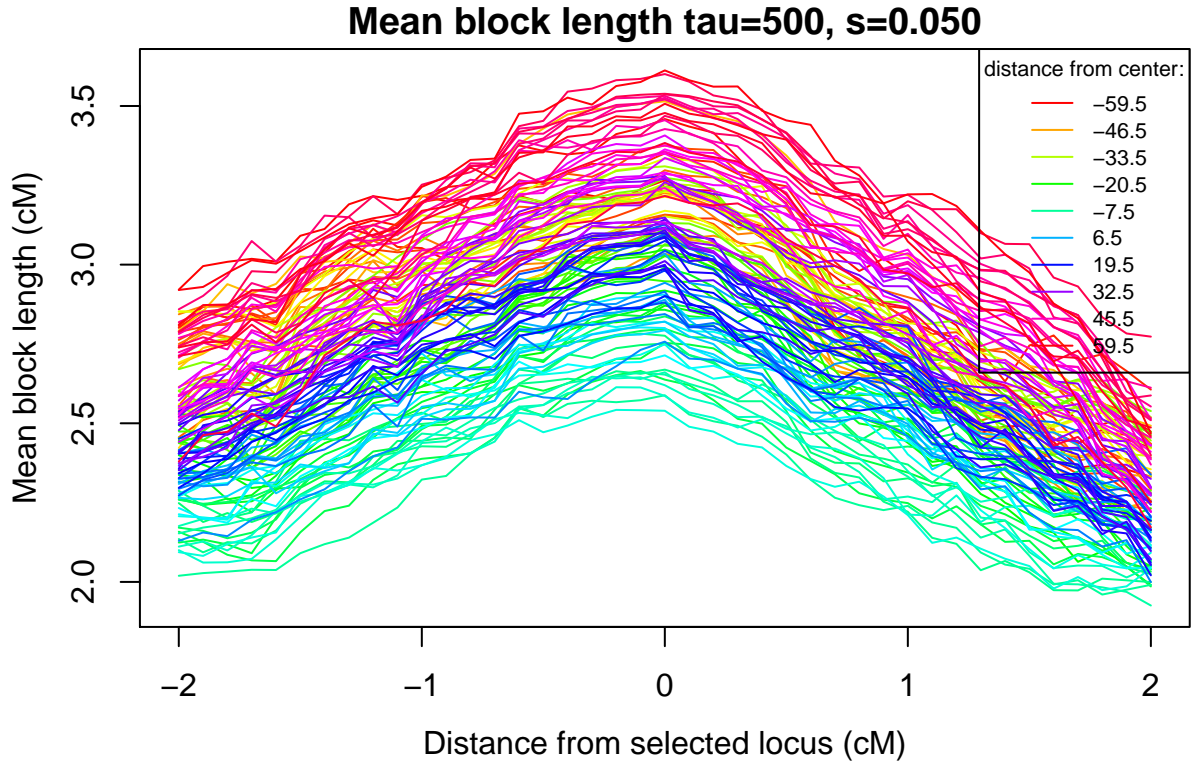


Figure S7: **Mean enclosing block length  $l(m, x)$** , across the genome (horizontal axis) and at different geographic positions (different colored lines). Results are from a simulation with 120 demes of 200 diploids each, selection  $s = .05$ , dispersal  $\sigma = 3$ , and after  $T = 500$  generations.

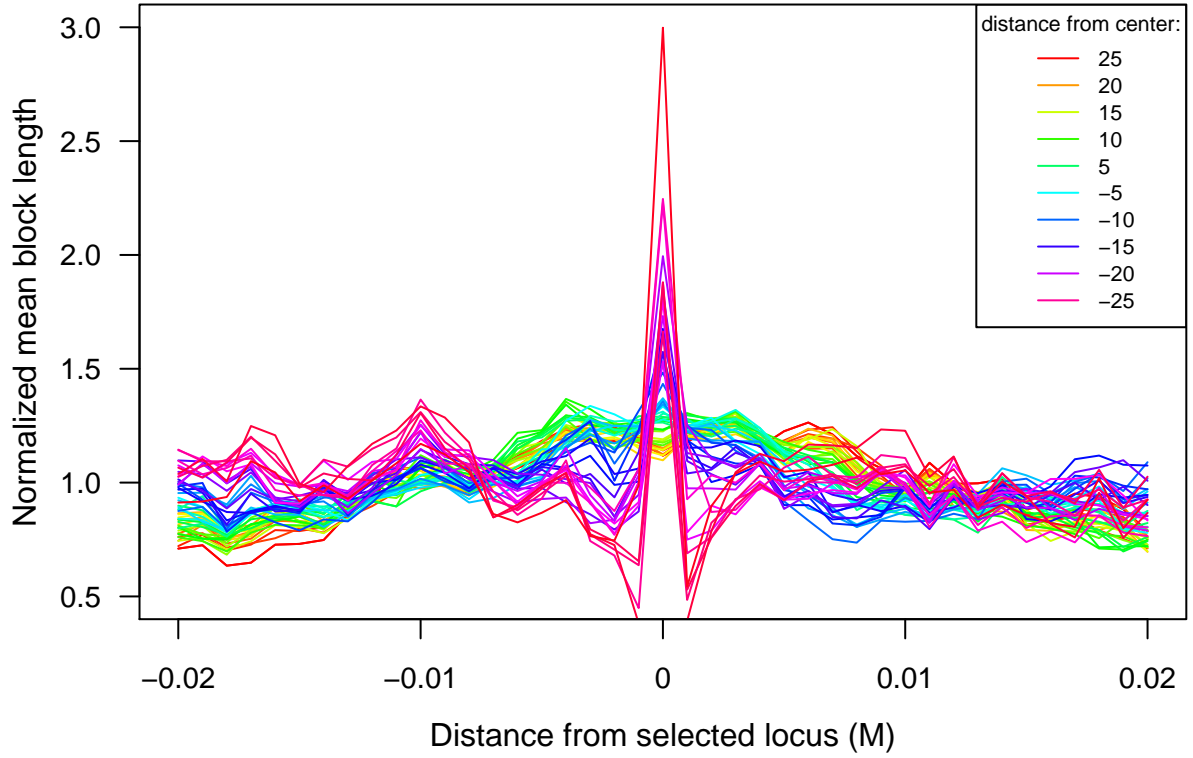


Figure S8: **Normalized mean enclosing block length**,  $\bar{l}_B(m, x)$ , after  $T = 1000$  generations, against position relative to the selected locus (horizontal axis) located in the center of a 1M chromosome. Each line shows the mean block length at that spatial and genomic position divided by the mean over the chromosome at that location; the simulation was run with  $s = 1$  and  $\sigma = 1$ , 50 demes, each containing 500 diploid individuals.

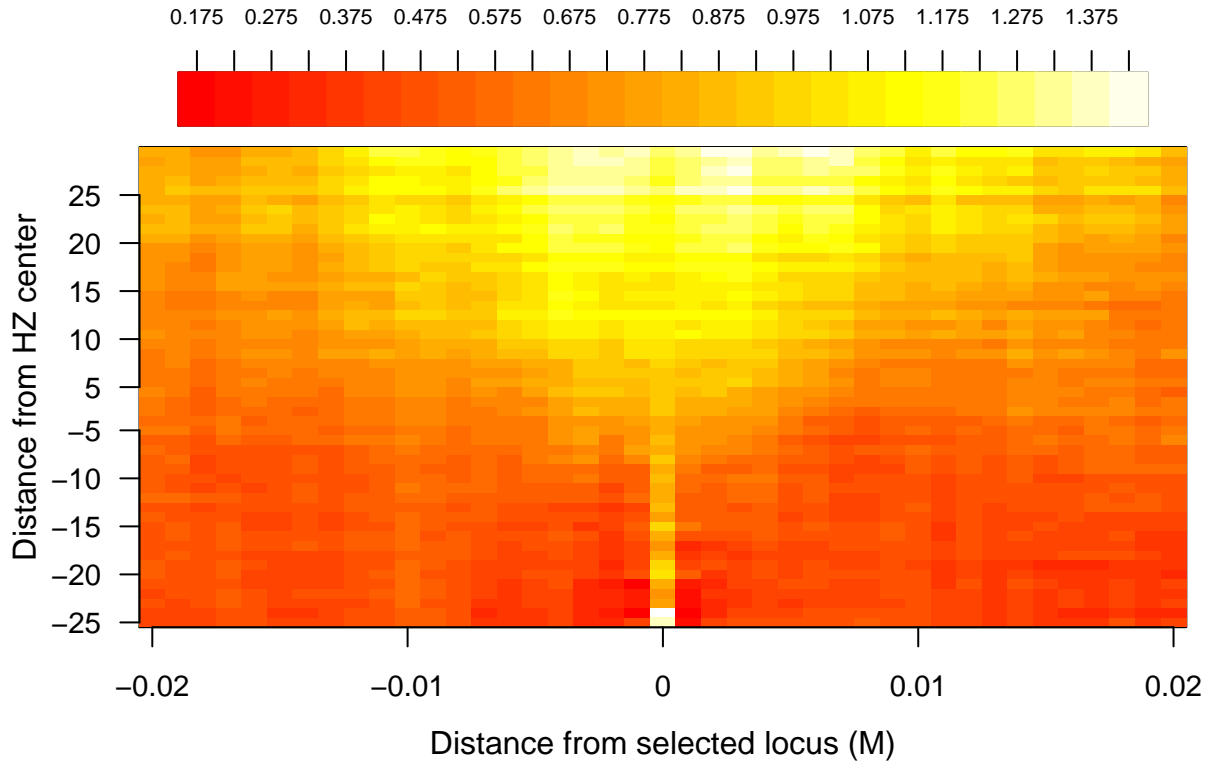


Figure S9: Heatmap of mean block length  $l_B(m)$  along a simulated chromosome.

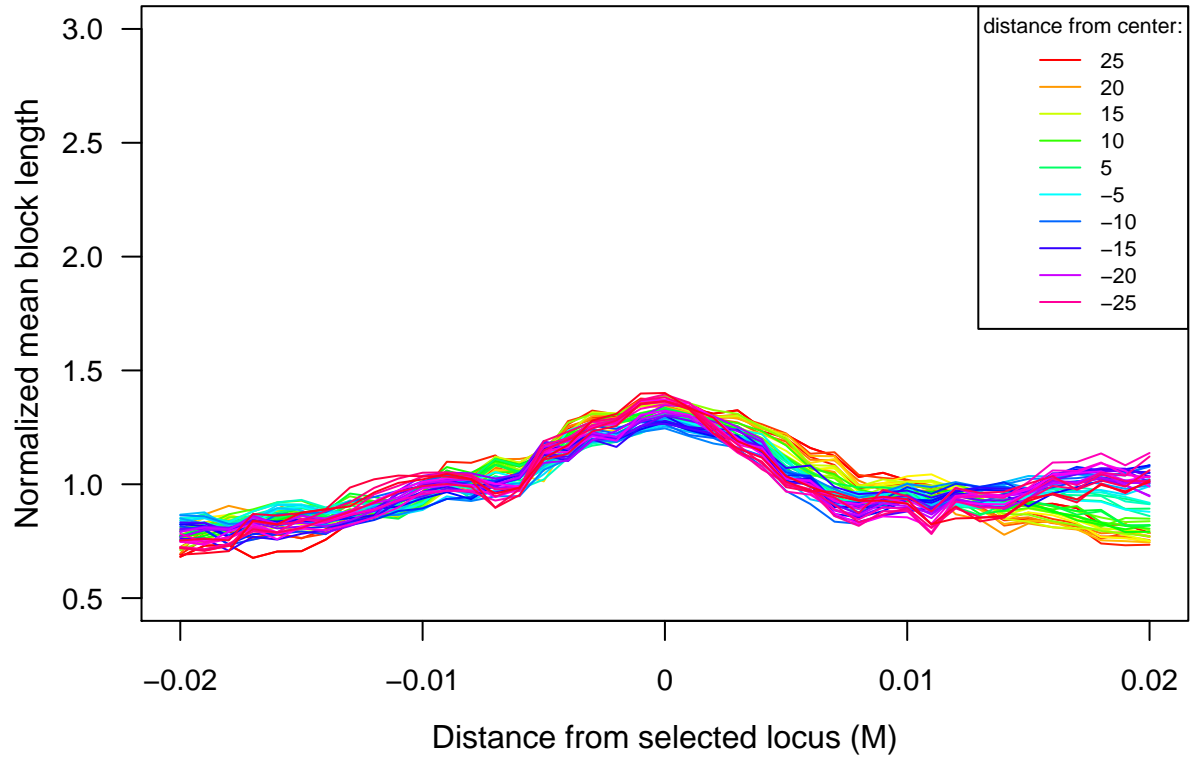


Figure S10: Mean block length  $l(m)$  surrounding a given position along the genome with a single underdominant site ( $s = 0.01, T = 1000$ ).

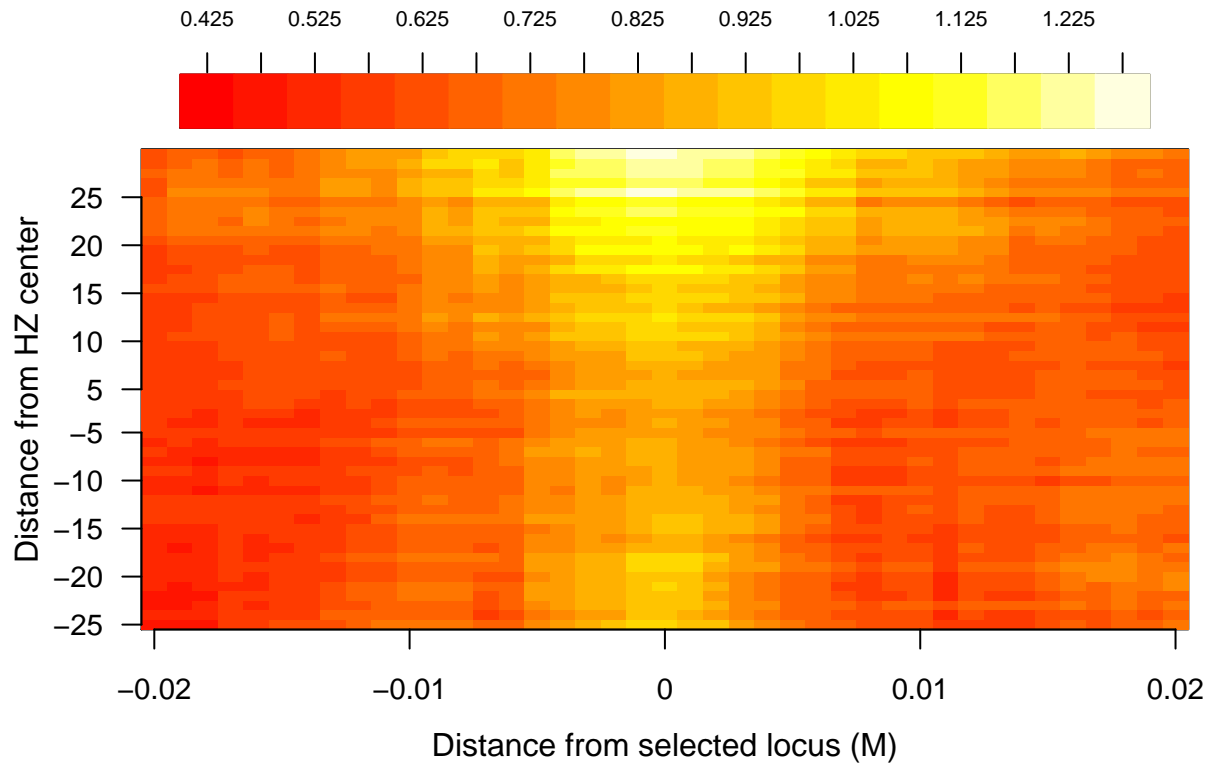


Figure S11: Heatmap of mean block length  $l(m)$  along a simulated chromosome with a single underdominant site ( $s = 0.01, T = 1000$ )



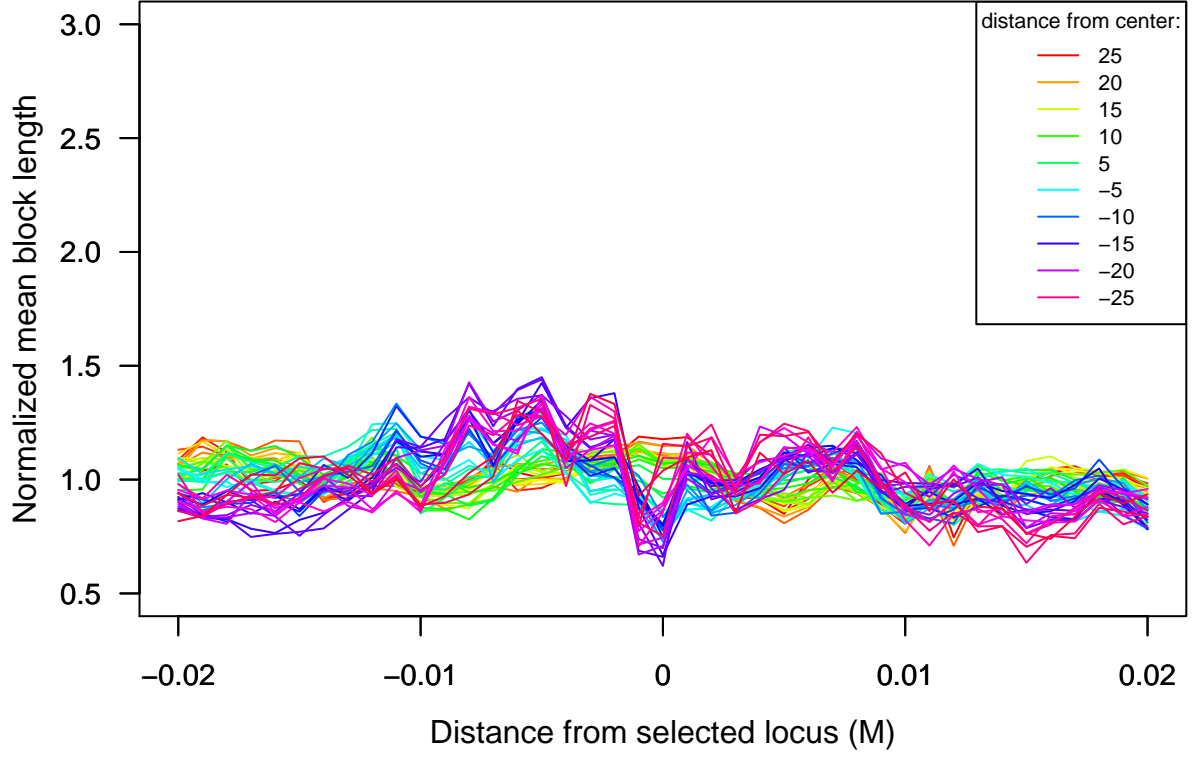


Figure S12: Mean block length of  $l_A(m_{\pm})$  across chromosome with single under dominant site, conditioning on ancestry  $B$  at the selected locus. ( $s = 0.01, T = 1000$ )

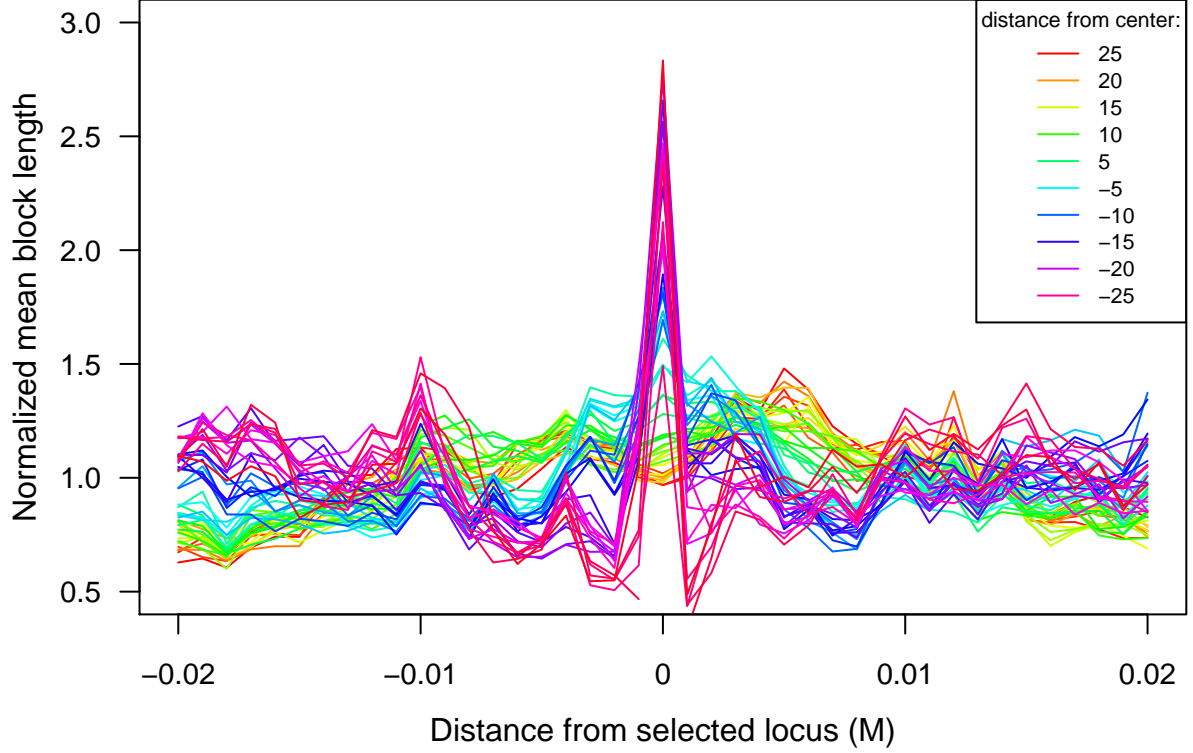


Figure S13: The statistic  $\bar{C}(m, x)$ : the ratio  $\frac{2\sum l_B(m_i)}{\sum l_A(m_i-) + l_A(m_i+)}$  of mean block length and mean adjacent block lengths across a simulated chromosome with a single underdominant site and conditioning on ancestry  $B$  at the selected site ( $s = 0.01, T = 1000$ ). Each line represents a deme and is normalized by mean block length across the chromosome in the deme.

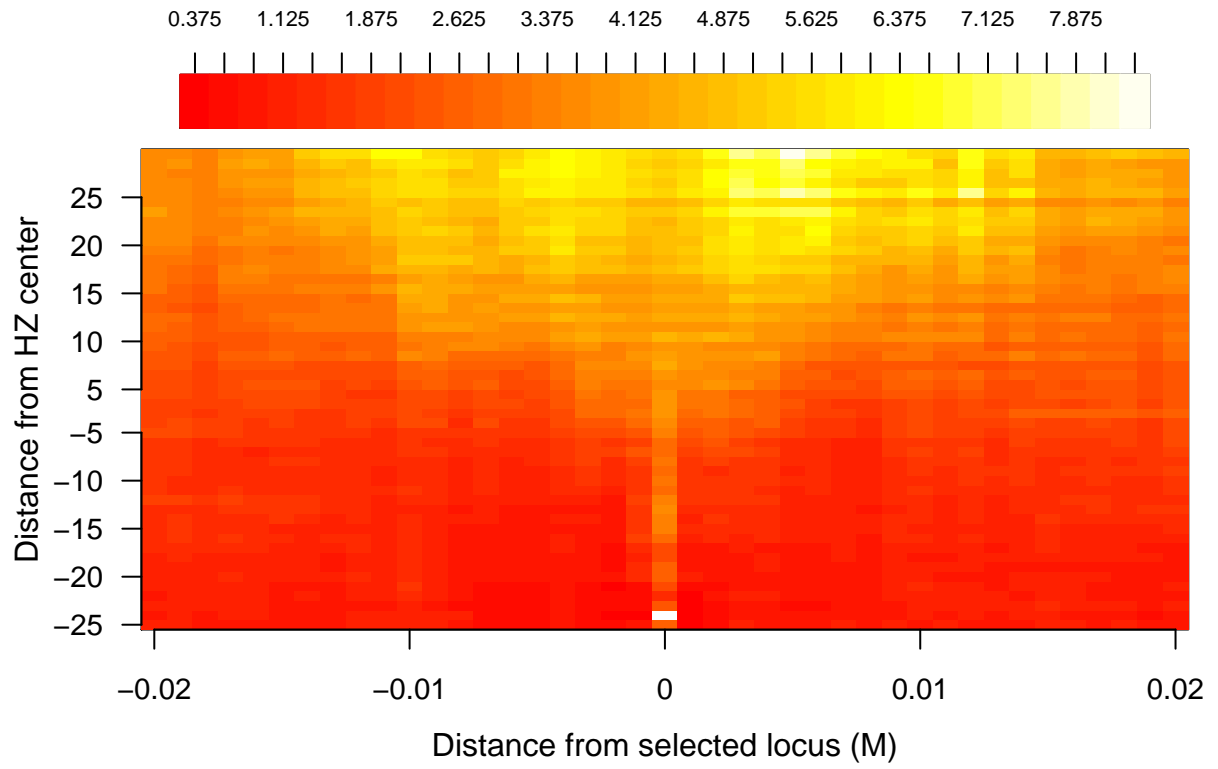


Figure S14: Heatmap of  $C(m, x) = \frac{2 \sum l_B(m_i)}{\sum l_A(m_i-) + I_A(m_i+)}$  across a simulated chromosome with a single underdominant site and conditioning on ancestry  $B$  at the selected site ( $s = 0.01, T = 1000$ ).

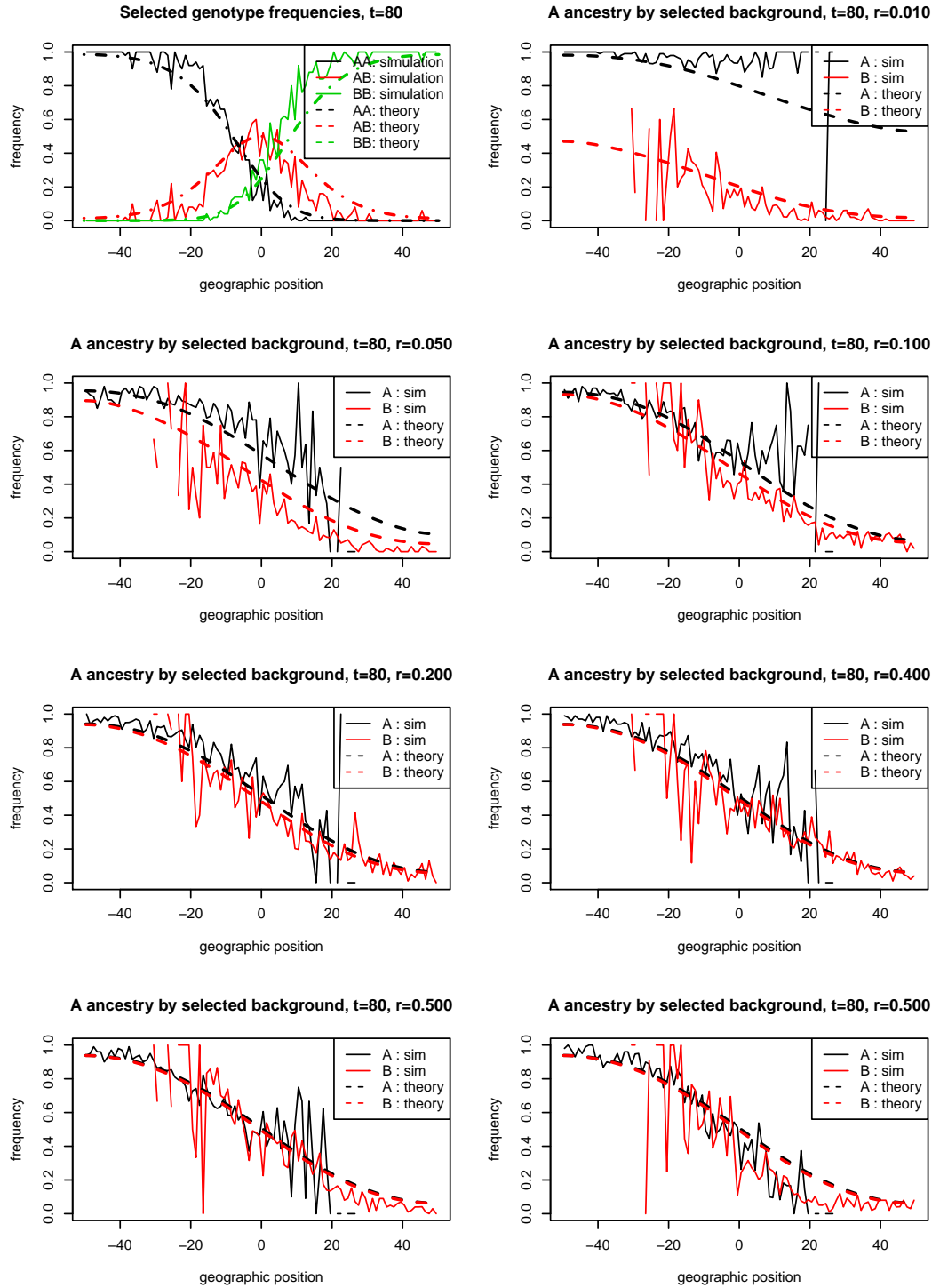


Figure S15: **Conditional frequencies of ancestry  $A$  at  $\tau = 80$** , comparing simulation and theory, with  $\sigma = 3$  deme spacings,  $s = 0.05$ , and 50 individuals per deme. The top left figure shows observed and expected genotype frequencies for the two homozygotes and the heterozygote at the selected locus; expected genotype counts were obtained assuming random mating, and by solving equation (??) numerically. The remaining figures show observed and expected frequencies of  $A$  ancestry, separately conditioned on the identity of the linked allele at the selected site. Observed frequencies become much noisier where the linked allele becomes rarer.

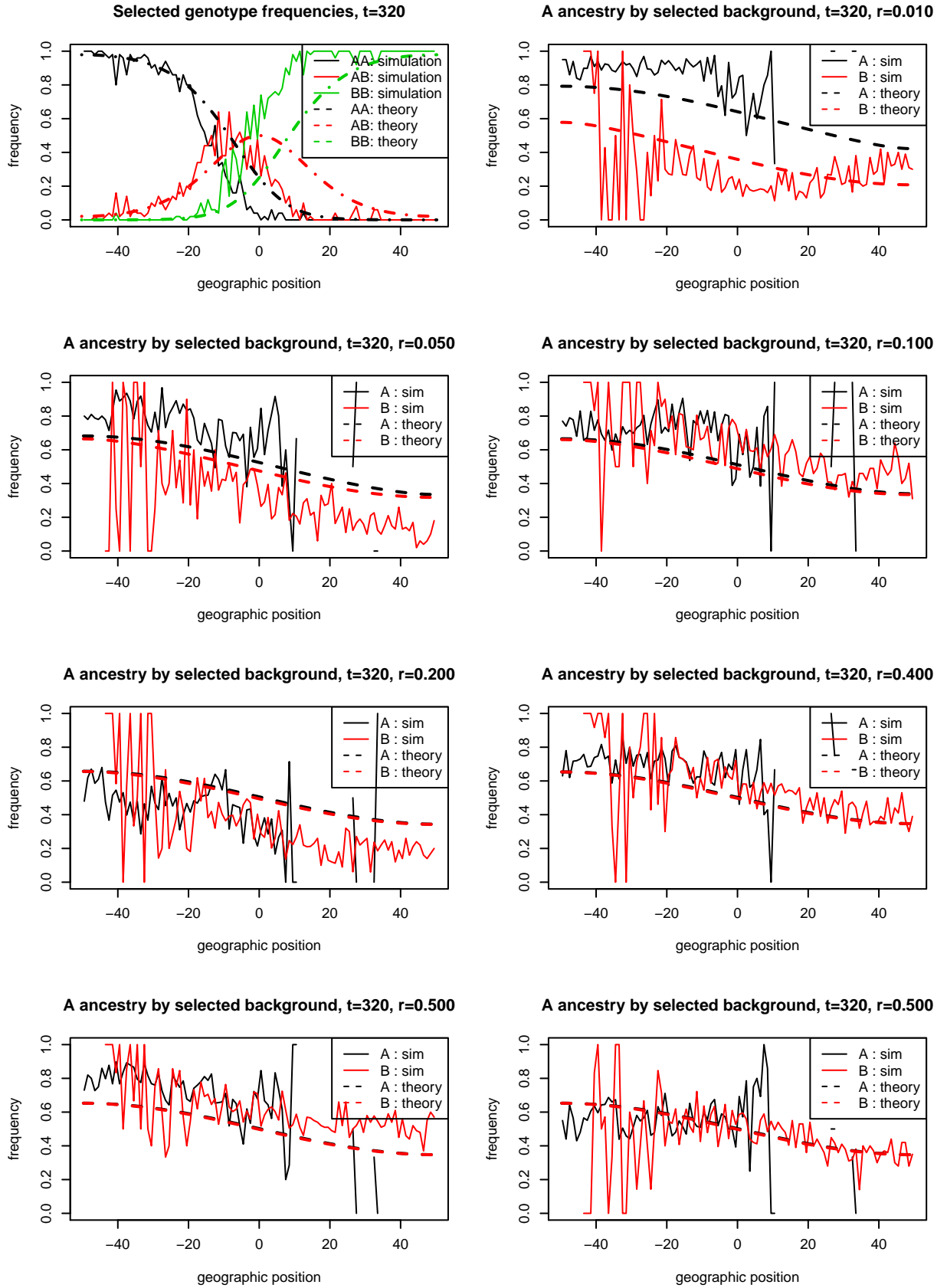


Figure S16: **Conditional frequencies of ancestry  $A$  at  $\tau = 320$** , as in figure ???. Deviations are larger than at  $\tau = 80$ , due to genetic drift.

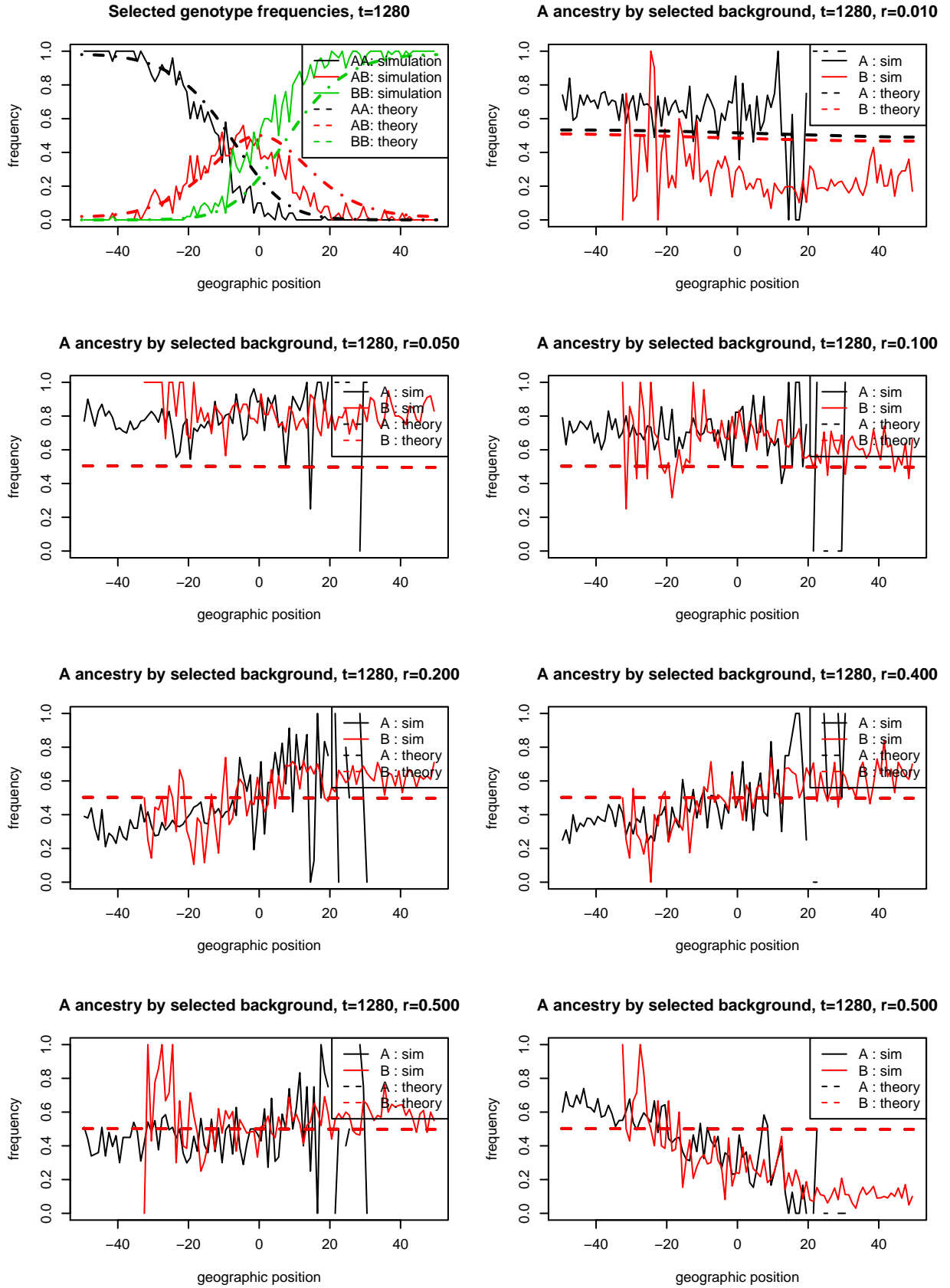


Figure S17: **Conditional frequencies of ancestry  $A$  at  $\tau = 1280$** , as in figure ???. Deviations are larger still than at  $\tau = 320$ , due to genetic drift.

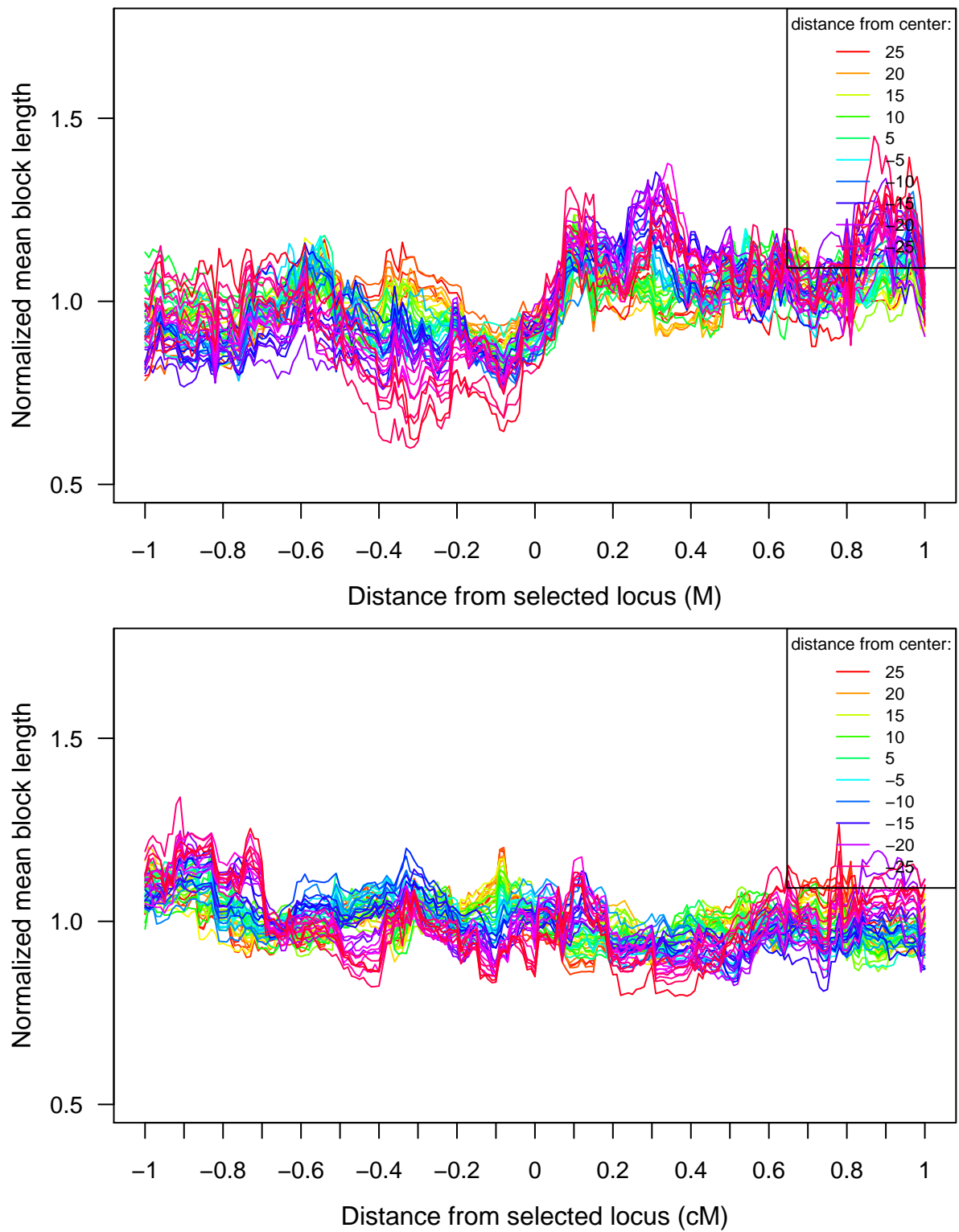


Figure S18: Block lengths along chromosome under  $s = 0.001$ , conditioning on ancestry B (top) and without conditioning (bottom)

Resubmission Cover Letter  
*Molecular Ecology*

Lewis-Sigler institute for integrative ge-  
nomics  
Princeton University  
Yaniv's address  
and Molecular and Computational Biol-  
ogy  
University of Southern California  
April 26, 2016

**To the Editor(s) –**

We did all the things, thanks.

*these aren't points because they are all points below in Reviewer 1 & 2's sections; we'll just say here that we did 'em, see below*

Reviewer 1 would like to see development of the finding that ancestral blocks are longer around a selected locus into a practical test. I think that this would be a substantial project for the future, and that it is valuable to first explain and understand the theoretical prediction, as is done here. Those analysing actual data can of course look for this pattern without needing an already implemented test.

Rev. 1 is also concerned that only under dominance is simulated. However, it is clear that any form of selection that gives the same marginal selection coefficient will give essentially the same effect: a large body of work on clines shows that different selective mechanisms are indistinguishable, and it is reasonable for the authors to rely on that. The other main concern was about the novelty of Eqs 4, 5; these are indeed standard (which I think the authors intended by referring to them as the Kolmogorov backward equations). I don't know of a reference where they have been written in exactly this form, but the recombination terms and the diffusion terms in Eq. 4 both go back a long way.

Reviewer 2 is mainly concerned about the fit between theory and simulations. I am not worried about this, since the deterministic mathematics stands independently, and the numerical computations are straightforward. The stochastic simulations are useful mainly as illustrations of the process. Theory and simulations are compared in the supplementary material; it should be stated more clearly how closely these match. (A precise fit is not expected anyway because there is substantial drift). Rev. 2 also worries that the simulations are over a rather narrow range; I suspect that the authors should have results that address sensitivity to the boundaries. Regarding specific point 1, I don't know of a previous derivation of exactly these equations, but similar equations have been used widely, and I don't think that the authors suggest otherwise.

Sincerely,

Alisa Sedghifar, Yaniv Brandvain, and  
Peter Ralph

**Reviewer AE:**

This is a nice analysis of how neutral genome introgresses past a cline maintained by selection against heterozygotes. There has been a good deal of work on related problems, but mostly assuming equilibrium, and mostly on the forwards evolution of allele frequencies, rather than tracing back the ancestry of blocks of genome.

The theory presented here is therefore new, and directly relevant to interpreting genomes sampled from hybrid zones.

---

**(AE.1) Fig 1 caption** “allele”

**Reply:** Fixed.

---

**(AE.2) (p. ??, l. ??)** “ $s$  and  $\sigma$  are small” - but relative to what? This is not at all obvious.

**Reply:** We just meant that in the rescaling argument for a discrete model this is required formally, without giving it enough thought. We’ve added more thoughtful discussion.

---

**(AE.3) (p. ??, l. ??)** Surely if time is rescaled as well as distance and genome length, everything should look the same? All the results, and to a good approximation the simulations, are in the continuous regime where we can rescale.

**Reply:** Good point; we in fact outlined this rescaling at (p. ??, l. ??). It still seems fair to say there is “more information” here, because we can pick up the difference in time scaling using haplotypes and linkage; we’ve edited this paragraph to make the point better.

---

**(AE.4)** A comment on two dimensions would be helpful. Ancestral lineages simply diffuse transversely at a rate  $\sigma$ , independent of their motion parallel to the cline.

**Reply:** We’ve added a short note on this (with more references) (p. ??, l. ??).

---

**(AE.5) (p. ??, l. ??)** This is a discrete deme model, so dispersal can’t follow a continuous Gaussian. I guess what is happening is that  $\sigma$  is chosen larger relative to deme spacing. However, there is some approximation here, and the approach will fail if  $\sigma$  is small relative to deme spacing.

**Reply:** There might have been some confusion about our use of the floor function? We now describe the dispersal distribution in a more straightforward way (p. ??, l. ??). It’s a good point more generally that if  $\sigma$  is smaller than the deme spacing this scheme won’t get dispersal with  $\sigma$  close to the  $\sigma$  that we plug into the Gaussian; it’s for this reason that we compute numerically what  $\sigma$  actually is when comparing simulation to theory.

---

**(AE.6)** Missing semicolon?

**Reply:** Fixed.

---

**(AE.7)** “are have”

**Reply:** Fixed.

---

**(AE.8) (p. ??, l. ??)** “reality” means “simulation” here??

**Reply:** Well, we hoped to mean “reality”, but perhaps “verify that theory matches reality” is too strong; we’ve adjusted this to say “realistic, individual-based models.”.

---

**(AE.9)** Some appendices might be transferred to supplementary material – the derivations are fairly standard.

**Reply:** We’re not strongly against moving the appendices to supplementary material, but we’d prefer to keep them as appendices (and thus distributed with the document) because although the results are fairly standard, they might not be familiar to our readership. We think it is also useful to have a more detailed description of a concrete model that falls in the domain of attraction that we discuss here.



## Reviewer 1:

This manuscript presents a discussion of patterns of linkage and allele frequency clines in a hybrid zone between two populations, under a scenario of mild hybrid incompatibility. It presents deterministic equations for the distribution of ancestral haplotypes along the genome and argues that the lengths of ancestry segments can be used to infer selection.

The idea of using the lengths of ancestry blocks to detect selection in hybrid zones is interesting, and I had not encountered it before. This article will likely be of most interest to theorists, however, as the discussion of the practical confounders of the selection test is limited.

---

**(1.1) (p. ??, l. ??)** *A serious limitation of the test is that it only detects selection against heterozygotes. They claim that this would not affect results, [previously at] 1.453-456: “Although our assumed scenario of selection against heterozygotes at a single locus is uncommon in nature, we believe that this is unlikely to drastically influence our findings, as previous studies have demonstrated that underdominant loci share similar properties with more realistic models” is unconvincing.*

**Reply:** We had failed to cite the substantial literature on the matter (referred to by the AE); we have now added pointers to this in the introduction (p. ??, l. ??) and to the referenced section (p. ??, l. ??). The point is made quite well by ? : the details of the shape are determined mostly by the mean fitness of heterozygotes, a parameter that collapses many models (approximately) onto each other, at least after drift and environmental heterogeneity is considered.

---

**(1.2)** *From a theoretical perspective, the article makes modest but valid contributions, and Equation 4 and 5 are potentially interesting extensions to equation (1). The practical benefit of using these equations rather than forward simulations appears limited.*

**Reply:** We also pondered this point, and feel that the equations contribute substantially to building intuitive understanding for how to address haplotype lengths, and aim to present analytical expressions and simulations so that they each support the conclusions of the other. We believe that analytical results can allow for a deeper understanding than simulation output, as for instance the intuitive discussion at (p. ??, l. ??), which mostly derives from the analytical model. To aid our readers in developing this understanding, we have gone through considerable pains to map terms in our equations to the biological processes of migration, selection and recombination. We believe that our efforts will help build a better mechanistic understanding in our readers in addition to making them aware of our results.

Furthermore, as noted at (p. ??, l. ??), the equations can be used to generate expectations for LD much more efficiently than forward-time simulation.

---

**(1.3)** *While reading the manuscript, I found it difficult to distinguish the most important observations from more tangential observations. The manuscript reads like a long list of observations, and would benefit from more focus on the main findings and more structure.*

**Reply:** *We agree that a long set of theoretical results can be tedious. We have therefore reshaped the introduction to highlight our major question “How does selection against incompatibilities in a hybrid zone influence the length distribution of ancestry tracts?”*

---

**(1.4)** *Since the recombination-based terms in Eqs 4 and 5 do not depend on the spatially explicit part of the model or on the selection term, I would be surprised if they had not been derived somewhere.*

**Reply:** We agree that this representation has likely been derived many times. We turned to Hartl and Clark and found a similar derivation, which is now cited in this point in the text.

---

**(1.5) Eq.??** *If  $a < b < \theta < 0$ , then I interpret  $g(\theta, b) = 0$ . If correct, this should be specified. Alternately, the bounds of the integral could be modified to account for this situation.*

**Reply:** Thanks for catching the mistake; we modified the bounds of the integral.

---

**(1.6)** *As an example of a section that felt unstructured, consider the discussion starting at (p. ??, l. ??). This long paragraph starts with a long list of observations, without giving a hint as to where we are going. Beginning the paragraph with the main idea for the paragraph would be useful.*

**Reply:** *(comment is more general than that particular location; say something about this)*  
We added an introductory sentence at (p. ??, l. ??) to provide some structure. We have also gone through to clarify other sections that felt unclear.

---

**(1.7) (p. ??, l. ??)** *“In the notation above,  $q_z(x, t, r) = P^x Z_t = A$ .” I did not see the notation  $P^x$  above, and the RHS does not depend on small  $z$ .*

**Reply:** Good point; we’ve used a more common notation to specify that  $X_0 = x$  (and  $Z_0 = z$  now).

---

**(1.8)** *“As we show below the solutions provide an good approximation”*

**Reply:** Fixed.

---

**(1.9) (p. ??, l. ??)** *It seems like the  $x+r$  and  $x-r$  should have the same sign.*

**Reply:** That’s correct; we’ve fixed the sign issues.

---

**(1.10) (p. ??, l. ??)**  *$g_z(x, T, a, b)$  is defined as: “We are interested in the probability  $g_z(x, t; a, b)$  of finding an entire segment  $(a, b)$  of ancestry  $A$  in an individual having selected allele of type  $z$ , ...” This definition does not make it clear that the probability is conditional on  $z$ ;  $\mathbb{P}(A|z)$  vs  $\mathbb{P}(A, z)$ .*

**Reply:** This was indeed confusing; it’s rewritten to be hopefully less confusing.

---

**(1.11) 1.606** *Reference issues*

**Reply:** *This is a reference that is expected to appear in the same issue, so we do not have page numbers. The hope is that this can be filled in during final proofs for the issue*

---

## Reviewer 2:

In this paper, the authors extend previous work on the diffusion of neutral alleles in a secondary-contact zone to include selection against hybrids. Building on the classic diffusion model for the dynamics of a cline at a selected locus, they formalize the dynamics of linked neutral loci and investigate the distribution of ancestry block-lengths around the selected locus.

The authors obtain numerical results for the flattening of the cline at a neutral locus linked to an under-dominant selected locus in the absence of genetic drift. They also outline scaling arguments for the rates of movement and length of the introgressing blocks around the selected locus. Simulations are set up to test the theoretical predictions and illustrate other interesting patterns such as the distribution of haplotype (ancestry block?) lengths.

---

**(2.1)** *Unfortunately, the fit of the theoretical predictions to simulations is largely not shown, and those comparisons in the manuscript are limited to the frequency dynamics of the neutral alleles (Figure S3). The predictive theoretical results are based on scaling arguments - but it is not clear which of these results are uniquely derived in this paper.*

**Reply:** *Our hope was that the simulations and theory would provide qualitative support for one another. Some predictions made by the theory e.g. time and spatial scales do seem to correspond. We will try to do a better job of signposting our unique contributions??*

The rescaling arguments we provide are of a standard sort in the literature; we've inserted comments to this effect, at (p. ??, l. ??) and (p. ??, l. ??).

---

**(2.2)** The authors also aim to characterize the movement of the genomic block that contains a selected locus through a hybrid zone. Yet, the size of the simulated habitat barely covers the cline width, and the simulation results may be sensitive to the boundaries.

Overall, this MS presents some interesting insights, and the formalization of the block dynamics can be useful. However, I think that more work is needed to explain and test (by simulations) the novel predictions. Currently, the theory and the matching simulations are largely presented separately.

**Reply:** *Should we be running more sims? Nick seems to think we have done this, but I don't think we have. run sims with longer range also run sims in part of param range where nothing's supposed to happen?*

---

**(2.3) Figure 1** Plotting the haplotypes to the side of the plot would help.

**Reply:** *We have remade Figure 1, with additional information that we hope will help with interpretation. We have also related this figure to the terms in Equation 5.*

---

**(2.4) (p. ??, l. ??)** As the theory is set up in the absence of genetic drift, and, correspondingly, the local population size is large, it is unclear how a deme with 500 individuals could become empty in a Wright-Fisher model. The (narrow) width of the habitat also needs justification.

**Reply:** Although most of the simulations shown in the main text have high population densities, we have run simulations with considerably lower densities, as shown for instance in figures ??, ??, and ??. Demes are never empty for very large populations, but we did encounter empty demes in some simulations with smaller population sizes. *add reference to more low pop density figures if added*

---

**(2.5) Figure 3:** What are the replicate lines for the same  $r$ ? Why re the lines for  $r = 0$  all slightly displaced to the right? Also, it would appear that the color scheme in associated Figure S2 (bottom) is mismatched with the legend, shifted to the red part of the spectrum. After fixing this, I think that the Figure S2 should be in the main text instead of Fig. 3.

**Reply:** *We have switched figures S2 and 3 and fixed the colors. The colored lines are not all represented in the legend which seems to have caused some confusion and we have tried to address this in the figure legends. We don't interpret the mismatch as a displacement to the right, as translating the lines horizontally will not improve the overall fit, and attribute this to drift in the simulations.*

---

**(2.6) (p. ??, l. ??)** Total population size is irrelevant when predicting the effects of genetic drift (besides, some simulations only have 50 demes across, not 100). In a one-dimensional habitat, one uses  $N\sigma\sqrt{s}$  to judge whether the effect of genetic drift is likely to be negligible.

**Reply:** *We meant that the allele frequencies in the entire population had shifted, which is determined by the total (effective) population size. However, we decided to simply state the our observed wobble of the center of the tension zone is consistent with previous theory. However, more broadly throughout the manuscript we are not sure about relying on  $N\sigma\sqrt{s}$  as we are concerned with loci linked to the selected site, and we are not aware of results extending this rule to an unlinked locus. (Do you mean "linked locus"?)*

---

**(2.7) Scaling arguments, p. 14, 15:** Please give a citation with the known scaling, and take more care with the novel ones.

**Reply:** As noted above, we have provided some citations to similar arguments in the literature (although such arguments are often used without citation; see, for instance, ?).

---

**(2.8) (p. ??, l. ??)** *I think the sentence about the maximum length of the enclosing haplotype,  $\sigma\sqrt{s}/x$  is mis-phrased: presumably this should be a statement on the distribution (expected size?) of the introgressing block length at distance  $x$ , as a lucky block can travel many sigmas in one generation?*

**Reply:** Good point, it was meant to be a statement about an order of magnitude bound to a stochastic quantity rather than a precise bound; it's been rephrased.

---

**(2.9) Discussion:** *It is important to make the reader aware that even neutral clines do not just keep flattening – genetic drift leads to a finite cline width even for neutral clines. Despite the occasional mention of genetic drift, the MS does not clearly state that the parameter space is chosen such that genetic drift is weak and does not indicate when this is a reasonable assumption (see comment above [about (p. ??, l. ??)]).*

**Reply:** We now have much more discussion about drift, in particular at (p. ??, l. ??).

*I don't understand this: We should add a line or two in the discussion about drift. However, IBD is perhaps not relevant over the timescales we are considering and is not necessarily ancestry informative, and since we are not interested in the equilibrium state, it is not clear that we need to worry about this limit.*

---

**(2.10) rates at (p. ??, l. ??), e.g. Eq. (7,8):** *I find the explanation of the rates a bit unclear. I understand that the recombinant AB can arise either by A reproducing at rate  $s_A$  and recombining with B, or B reproducing at rate  $s_B$  and recombining with A – such that the neutral locus of the type A is kept. Or alternatively, the BA recombinant arises, where the neutral locus at B is kept. Which one of the two happens has a probability of 1/2.*

**Reply:** Right; we've added more discussion of this.

---

**(2.11)** *The “Data Accessibility section” is missing. The folder at appears to be a working folder with lots of other files and the r-scripts hidden under a ‘sim’ directory. The README file appears insufficient to allow orientation in the scripts, and worryingly, it has an ‘oh, wrong recombination distance’ note to it.*

**Reply:** *We have created a new repository with minimally required scripts and a descriptive README file*

---

**(2.12)** *The Acknowledgements section is missing.*

**Reply:** This has been added. Thanks.