Beyond clines: lineages and haplotype blocks in hybrid zones

Alisa Sedghifar*, Yaniv Brandvain§, and Peter Ralph†

*Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, 08544
§Department of Plant Biology, University of Minnesota, St. Paul, Minnesota, 55108
†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 show 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 physical scale of $1/T$, where T is the age, in generations, of the hybrid zone. 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

Much of the process of speciation commonly involves populations diverging with relatively little gene flow (Coyne & Orr 2004). 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 in Barton & Hewitt 1985). 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 (Fisher 1954; Chapman & Thompson 2002; Baird *et al.* 2003). These patterns of correlated ancestry in admixed populations have been previously used to infer histories of hybridization and admixture (e.g. Gravel 2012; Hellenthal *et al.* 2014; Sedghifar *et al.* 2015).

While for much of the genome, gene flow will eventually erase both geographic clines in ancestry and strong correlations in ancestry across linked markers, subsets of the genome can maintain stable clines in ancestry if selection against migrant, or hybrid, genotypes impedes gene flow across the hybrid zone (Barton 1979). Studies of hybrid zones have primarily been based on single-site statistics such as clines, but today, the availability of dense genotype data in combination with linkage information makes it now possible to use patterns of the block length distribution to better understand the processes operating in hybrid zones.

There is substantial theoretical work on genetic clines. We study those maintained only by selection against hybrids – *tension zones* – as in Bazykin (1969), reviewed in Barton & Hewitt (1985). Clines may also be maintained by environmental gradients (Haldane 1948; Fisher 1950); although these are in principle different, Slatkin (1973) pointed out, in a far-reaching general analysis, that while the environment may determine the location of a cline, the fitness of the heterozygote will likely determine the shape of the cline, and Barton & Bengtsson (1986a) show that the shapes of several sorts of stable clines are similar. There is substantial work on how clines are affected by genetic drift (Slatkin & Maruyama 1975; Felsenstein 1975a; Barton 2008; Polechová & Barton 2011), and quantifying the barrier to gene flow caused by clines (Barton 1979; Barton & Hewitt 1989; Barton & Gale 1993; Barton & Partridge 2000). In this work, we focus on *transient* patterns after the initial formation of a tension zone, in particular describing haplotype patterns. Although the selected cline establishes quickly, transient patterns can persist for quite some time, and haplotypes can potentially carry substantial information about the formation of recent zones. 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. This is particularly suited to understanding transient patterns, although because of the well-known problems with spatial coalescence (Felsenstein 1975b; Barton *et al.* 2002), incorporating genetic drift is not straightforward. Common to nearly all studies is the use of the reaction-diffusion equations that govern the deterministic, high-density limit (as in Nagylaki 1975), or the discrete analogues (Hanson 1966).

The arrangement of segments of genome inherited from particular ancestral populations provides information about the number of recombination events that have taken place between ancestral genotypes, and therefore contains clues concerning the age of the hybrid zone and the relative strength of selection affecting subsets of the genome. In addition to interpreting non-equilibrium patterns of diversity in hybrid zones, this could be especially important for understanding the impact of selection on the mixing of ancestral genomes. In the extreme case of zero fitness in first-generation hybrids, for example, parental genomes will never mix and the stable hybrid zone will resemble a recently formed one. If hybrids

have non-zero fitness, some amount of introgression can occur, and the distribution of ancestry tract length in such populations has be used to infer the extent of hybridization and timing of secondary contact (Price *et al.* 2009; Hellenthal *et al.* 2014). In addition to facilitating demographic inference, heterogeneous patterns of ancestry across the genome can help identify putative targets of selection in hybridizing populations (Porter *et al.* 1997; Gompert *et al.* 2012).

If hybrid ancestry at a given locus is disfavored, migrant haplotypes containing the selected locus 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.

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. In Sedghifar *et al.* (2015) we described how lineages move across a neutral zone of secondary contact, obtaining predictions for distributions of lengths of blocks of ancestry, which we fit to several datasets. In this work, we extend the same framework to the case of selection against heterozygotes at a single locus, by describing how lineages move conditional 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 Hudson & Kaplan (1988)).

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 (Kruuk *et al.* 1999; Barton & Shpak 2000).

As in Sedghifar *et al.* (2015), we consider where the ancestors of modern-day individuals lived, i.e., where their lineages 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 individual of species A at the time of secondary contact, i.e., if its lineage traces back to a spatial position $x < 0$. 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 1. 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 establishing order-of-magnitude estimates, and analytics, mostly in terms of partial differential equations needing numerical solution.

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 (Barton 1979; Barton & Bengtsson 1986b). First, we turn our attention to how the cline at the selected locus itself is formed. Let $p(x, t)$ denote the frequency of the A allele at spatial location x and time t. Suppose that secondary contact occurred at time $t = 0$, and that $p(x, 0) = 1$ if $x < 0$ and $p(x, 0) = 0$ if $x > 0$. Let σ^2 be the mean squared distance between parent and offspring (i.e., the variance of the distribution of dispersal distance). Assuming that: (i) alleles locally assort into diploids randomly, (ii) habitat and dispersal are homogeneous, (iii) s and σ are small, *relative to N and τ respectively?* and (iv) population density is large, then the commonly-used equation

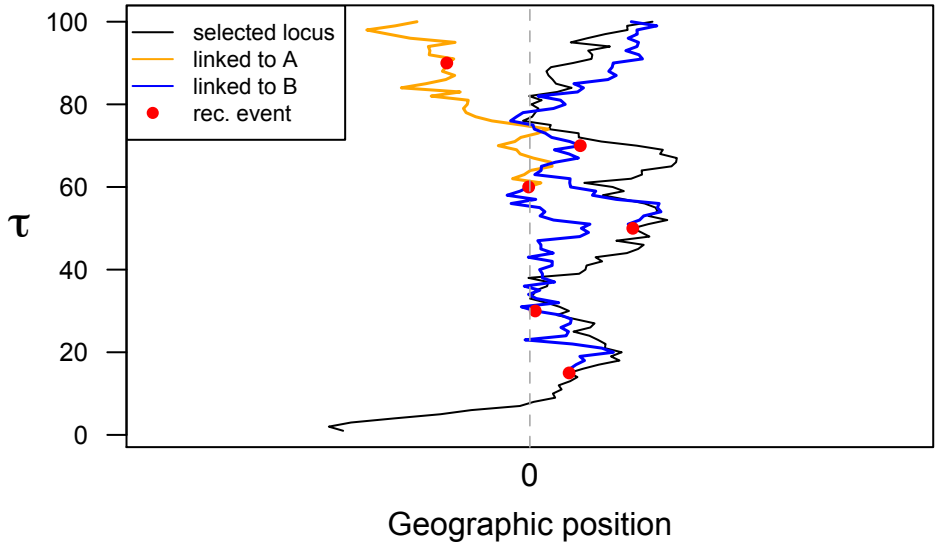

Figure 1: **A schematic figure of the lineages along which a segment of genome has been inherited,** showing the location (horizontal axis) and time (vertical axis) at which the ancestors of a hypothetical sampled haplotype lived. At the time of sampling ($\tau = 0$), the segment contains the B allele at the selected locus. Paths of lineages in space, backwards in time, are depicted by blue and yellow lines. Splitting events represent recombination events along the sampled segment. The thick 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 negative 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).

that p solves is

$$\dot{p} = \frac{\sigma^2}{2}\Delta p + sp(1-p)(2p-1), \tag{1}$$

where $\dot{p}(t,x)$ is the time derivative and Δp is the Laplacian, as in Bazykin (1969); this can be derived as in Nagylaki (1975); for discussion of diploidy see Christiansen *et al.* (1995). As we show below the solutions provide an 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 $\lim_{t\to\infty} p(x,t) = (1+\tanh(-2x\sqrt{s}/\sigma))/2$ (Bazykin 1969). 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. Note that rescaling space and time by $\sigma/\sqrt{s}$ and $1/s$ respectively in equation (1) results in a dimensionless equation (Slatkin 1973), implying that the cline establishes over a timescale of the order $1/s$. In other words, the cline is established through diffusion, despite being slowed down somewhat by selection against heterozygotes.

Lineages at the selected locus Now suppose that the frequency profile of the selected allele, $p(x,t)$, is known. 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. (Lineages are also biased towards regions with higher fecundity, and thus away from the hybrid zone, but we assume that s is small and ignore this.)

Making the same assumptions as for equation (1), the derivation in Appendix A shows that the lineage of an A allele moves as Brownian motion with speed σ with bias $\sigma^2\nabla \log p(x,t)$, where ∇ is the gradient (i.e., it moves at speed σ^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 $p()$ to first order about x shows that this mean displacement is approximately $\sigma^2 p'(x)/p(x)$, which is σ^2 multiplied by the gradient of $\log p(x)$.

At a geographic position far from the center, $p=1$ to the left, and is proportional

to $\exp(-x\sqrt{s}/\sigma)$ to the right. 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 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.

Note that this description gives more information than the steady-state cline $p(x)$, which depends only on $\sigma/\sqrt{s}$. Here we see that lineages with $\sigma = 10$ and $s = .16$ move quite differently to lineages with $\sigma = 1$ and $s = .0016$, reflecting strong differences in selection against heterozygotes, even though the stable clines have the same form.

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 move between backgrounds. 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. We will follow its lineage back through time, using τ to denote "generations ago" to avoid confusion. If X_τ is the geographic location of its ancestor τ generations ago, and Z_τ is the identity of the selected allele that ancestor carried, then we say that X moves as a diffusion pushed by either $\log(p)$ or $\log(1-p)$ (as described for a selected allele), and that Z jumps between A and B at rate either $r(1-p)$ (from A to B) or rp (from B to A), by the assumption of locally random assortment of alleles (see Appendix A).

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

$$\begin{aligned} dX_\tau &= \begin{cases} \sigma^2 \nabla \log(p(X_\tau, \tau))d\tau + \sigma dB_\tau & \text{if } Z_\tau = A \\ \sigma^2 \nabla \log(1 - p(X_\tau, \tau))d\tau + \sigma dB_\tau & \text{if } Z_\tau = B \end{cases} \\ \mathbb{P}\{T_B > \tau + u \,|\, Z_\tau = A\} &= \exp\left(-r \int_\tau^{\tau+u} (1 - p(X_s, s))ds\right) \\ \mathbb{P}\{T_A > \tau + u \,|\, Z_\tau = B\} &= \exp\left(-r \int_\tau^{\tau+u} p(X_s, s)ds\right). \end{aligned} \tag{2}$$

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

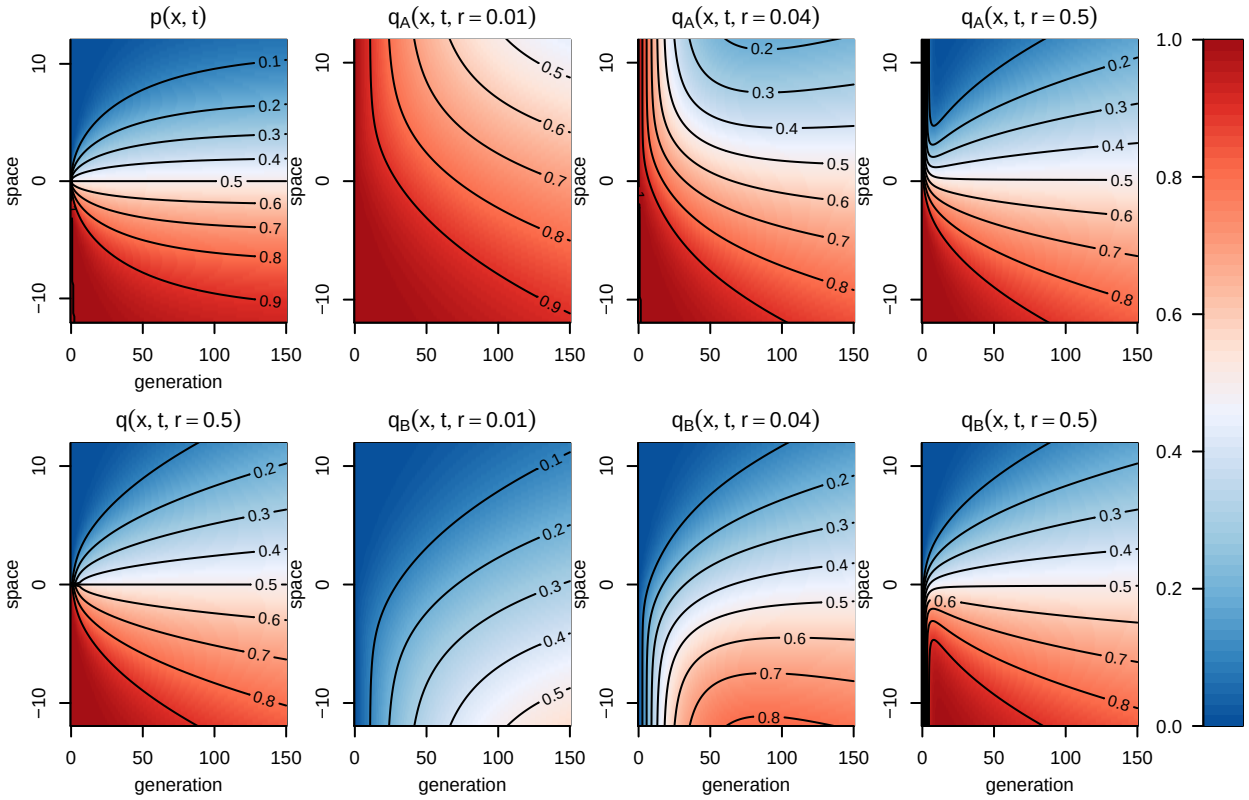

Figure 2: **Probabilities of A ancestry,** across space (vertical axis, in units of σ) and time (horizontal axis, in generations). In each plot, color corresponds to the expected frequency of A ancestry at a particular location in time and space. The selection coefficient is $s = .02$. **Top left:** at the selected site, showing establishment and stabilization of the cline on a time scale of $1/s = 50$ generations. **Bottom left:** at an unliked 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 S1 for the same figure over a longer period of time.

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,

$$q_z(x,t,r) = \mathbb{P}^x\{Z_t = A\}. \tag{3}$$

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

$$\begin{aligned}
\dot{q}_A(x,t,r) &= \sigma^2 \nabla \log(p(x,t)) \cdot \nabla q_A(x,t,r) + \frac{\sigma^2}{2} \Delta q_A(x,t,r) \\
&\quad + r(1-p(x,t))(q_B(x,t,r) - q_A(x,t,r)) \\
\dot{q}_B(x,t,r) &= \sigma^2 \nabla \log(1-p(x,t)) \cdot \nabla q_B(x,t,r) + \frac{\sigma^2}{2} \Delta q_B(x,t,r) \\
&\quad + rp(x,t)(q_A(x,t,r) - q_B(x,t,r)),
\end{aligned} \tag{4}$$

with boundary conditions $q_A(x,0,r) = 1$ and $q_B(x,0,r) = 0$, and where ∇ is the gradient, and Δ is the Laplacian.

We will have more use for the differential operators on the right-hand sides of these equations, so define these as $G^A = \sigma^2 \nabla \log(p(x,t)) \cdot \nabla + (\sigma^2/2)\Delta$ and $G^B = \sigma^2 \nabla \log(1 - p(x,t)) \cdot \nabla + (\sigma^2/2)\Delta$, so that equation (4) can be written more compactly as

$$\begin{aligned}
\dot{q}_A &= G^A q_A + r(1-p)(q_B - q_A) \\
\dot{q}_B &= G^B q_B + rp(q_A - q_B).
\end{aligned}$$

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.

Numerical computation To determine how the cline at a linked locus is expected to relax, we can solve the partial differential equations (1) and (4) numerically. For instance, Figure 2 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 $q(x,t,r) = p(x,t)q_A(x,t,r) + (1-p(x,t))q_B(x,t,r)$. The equations are solved numerically in R, using the ReacTran package (Soetaert & Meysman 2012).

Haplotype lengths We now find the frequency at which *entire* blocks of genome (haplotype) of a single ancestry are found, at a given location and time. Consider a segment of the genome 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 in an individual

having selected allele of type z, sampled at spatial location x at time t after the initiation of gene flow.

As in Sedghifar *et al.* (2015), 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 1), and so lineages act like a particular class of 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 (2), 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.

The resulting equation is similar to (4) with a term added for branching, which is written (omitting the (x, t) for conciseness):

$$\begin{aligned}\dot{g}_A(a,b) = G^A g_A(a,b) + r(a,b)(1-p)(g_A(a,b) - g_B(a,b)) \\ + (1-p)\left(\int_a^0 g_B(a,\theta)g_A(\theta,b)d\theta + \int_0^b g_A(a,\theta)g_B(\theta,b)d\theta\right), \\ + p\int_a^b g_A(a,\theta)g_A(\theta,b)d\theta - (b-a)g_A(a,b)\end{aligned} \tag{5}$$

where θ is the genomic position of a recombination event, the integral $\int_a^0$ is zero if $a > 0$ and likewise $\int_0^b$ is zero if $b < 0$; and $r(a, b)$ is the distance from the segment to the selected site: $r(a, b) = 0$ if $ab < 0$ and $r(a, b) = \min(|a|, |b|)$ otherwise. 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 (5) 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 B.

Correlations in ancestry To compute correlations in local ancestry (i.e., "ancestry disequilibrium", as in Pool (2015); Schumer & Brandvain (2016)), 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 B 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 can migrate to a new deme offset by $\lfloor Z + 1/2 \rfloor$ from their original position, where Z is a Gaussian random variable with mean zero and variance σ^2. (The mean displacement is zero, and when comparing to theory we compute σ 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

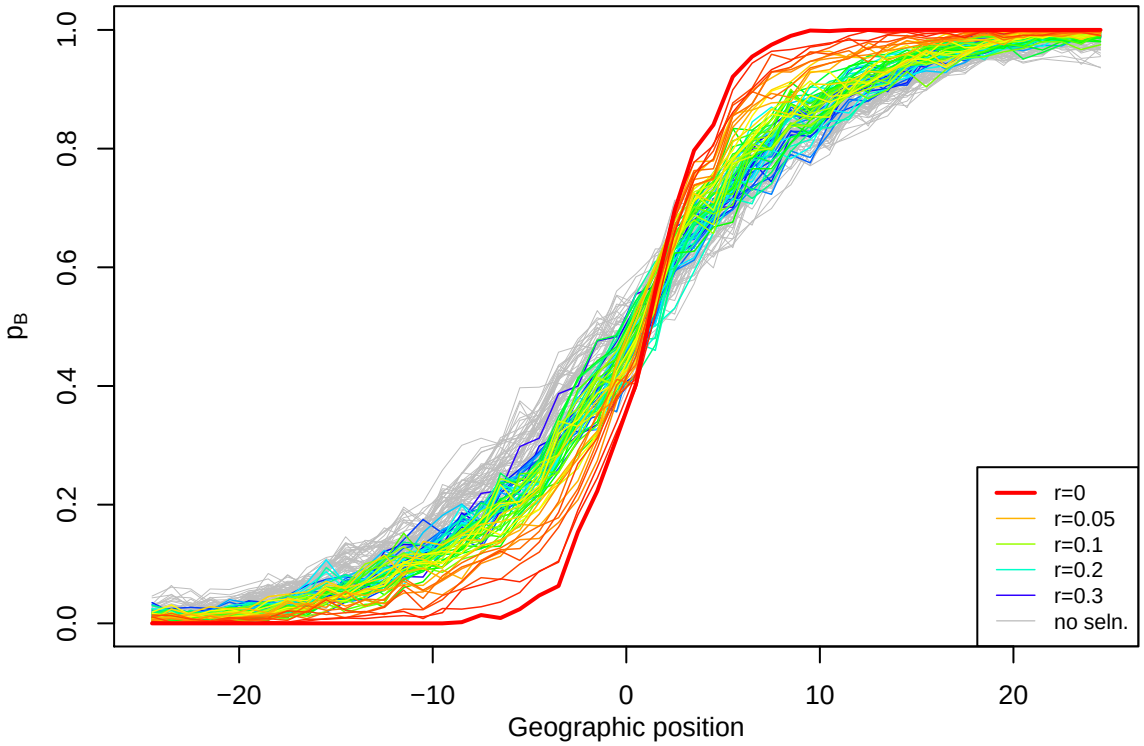

Figure 3: Frequency of ancestry B, across geography at different physical positions on the genome, simulated for a hybrid zone $T = 100$ generations after secondary contact, with $s = 0.1$, using 50 demes, each with 500 diploid individuals and $\sigma = 1$. Each line represents a locus some distance r away from the true target of selection. Grey lines represent the same positions from a simulation with identical parameters except that $s = 0$. Corresponding theoretical quantities are shown juxtaposed in Figure S2; the same plot is shown with weaker selection in Figure S3 and at a longer time in Figure S4.

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, motivating us to define the statistic

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

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.

We are interested in identifying regions of the genome with abnormal distributions of block lengths, and at each spatial location x, compute $l_B(m,x)$ and $C(m,x)$ at a grid of positions across the genome. Where indicated, we normalize $l_B(m,x)$ and $C(m,x)$ by dividing by the empirical mean of l_B and C across the genome for location x. While there is useful information contained in the geographic distribution of block length 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.

Although our analytical model does not account for coalescence, 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 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-loci: 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, for instance, in Figures 3, S3 and S4. 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 S2 shows

that expected clines computed from equation (4) match simulated clines in Figure 3 quite well. Figures S15, S16, and S17 show this comparison with a lower population density than in Figure 3, and separating linked clines based on the identity of the linked selected allele. The agreement is good over hundreds of generations, but by $T = 1280$ generations, stochasticity has substantially moved allele frequencies range-wide (which is unsurprising, as the total population size is only 5,000 diploids: Polechová & Barton 2011).

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 (Barton & Bengtsson 1986a) and the relative slowness of diffusive movement. We display the spread of ancestry across space and time in Figure 2. 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 2. 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 2.

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 Figures S5 ($T = 100$ generations) and 4 ($T = 1000$ 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 S7. 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/σ^2 for a neutral allele). This implies the enclosing A haplotype should be no longer than $\sigma\sqrt{s}/x$.

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 4 and S5. 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 5, S8 and S9, 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.

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 S6. 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 S12, 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. This is shown in Figures S7, S11 and S10.

Power will be optimal at intermediate values of s. If selection is too strong, it may be difficult to observe the 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. 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.

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 6). 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 that clines establish over time $1/s$, and that 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

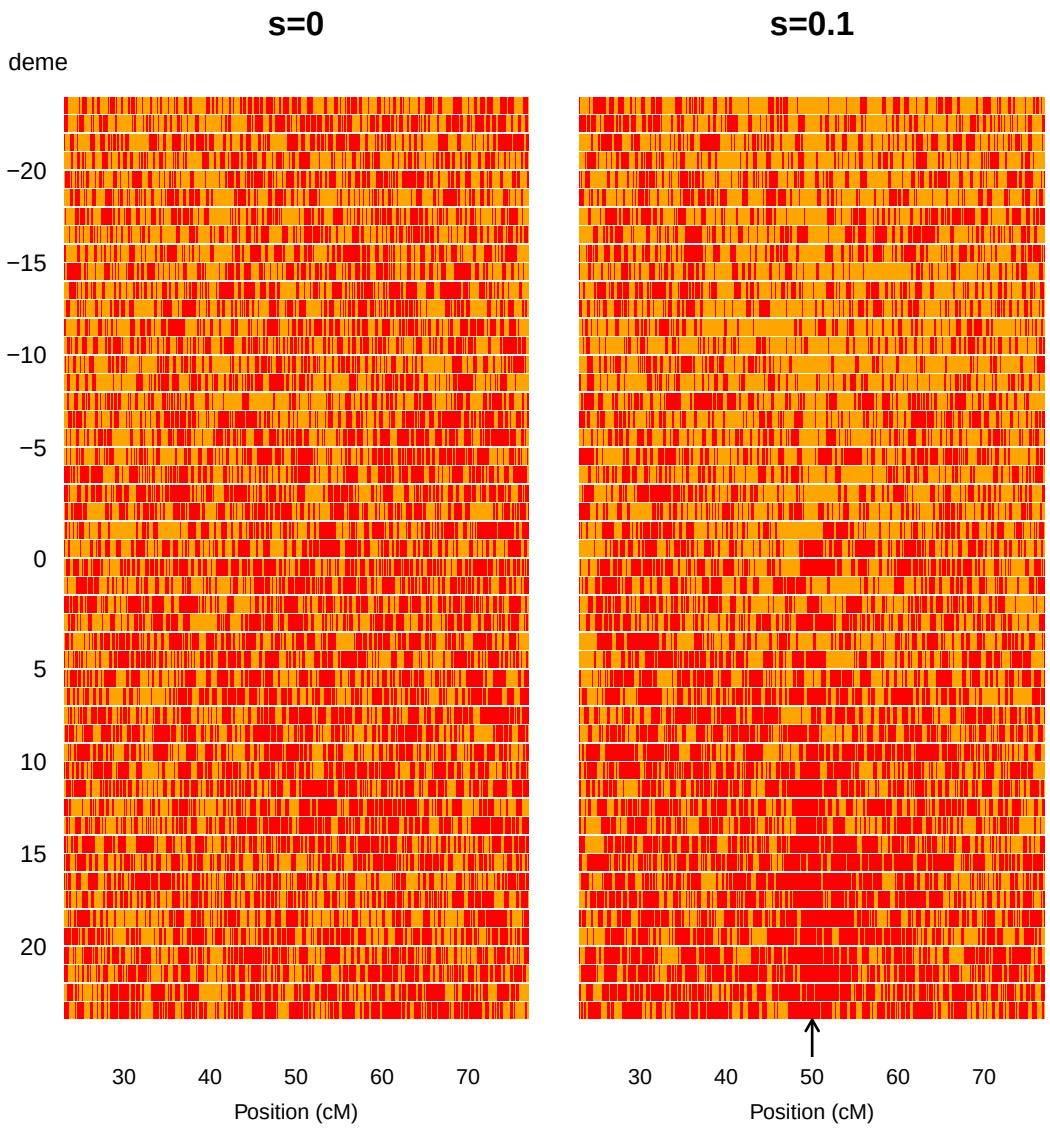

Figure 4: **Ancestry blocks for randomly sampled chromosomes** across a hybrid zone of age $T = 1000$. Here we compare chromosomes of length 1M from a neutral zone to a zone that has a single under-dominant locus with $s = 0.1$ in the middle of the chromosome (indicated by black arrow). Red blocks along 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 S5.

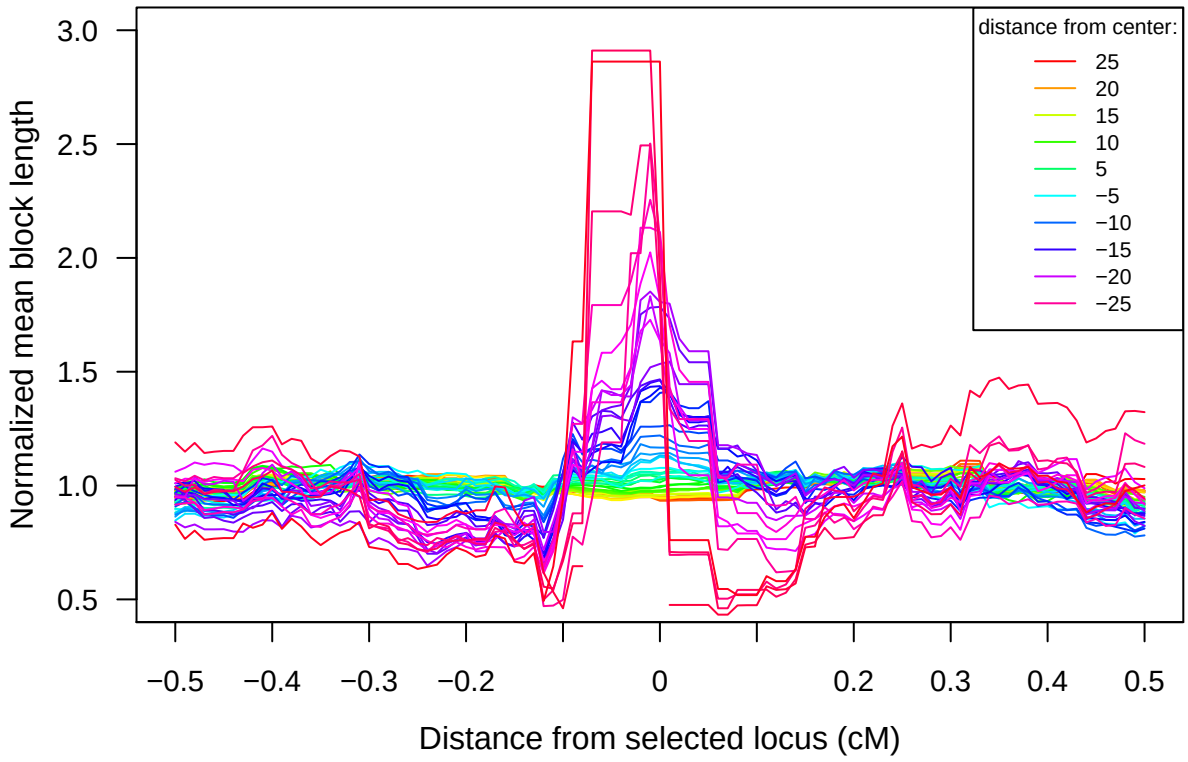

Figure 5: **Normalized mean enclosing block length**, $l_B(m)$, along a simulated chromosome of length $1M$ with selected locus of $s = 0.01$ at position $0.5M$ in a hybrid zone of $T = 1000$. Here each line represents $l_B(m)$, in a given deme, and is normalized by dividing by mean $l_B(m)$ along the chromosome within the deme. Chromosome were sampled across the hybrid zone, which consists of 50 demes, each containing 500 diploid individuals. This figure represents the same simulation and statistic as Fig. S8, on a finer genomic scale.

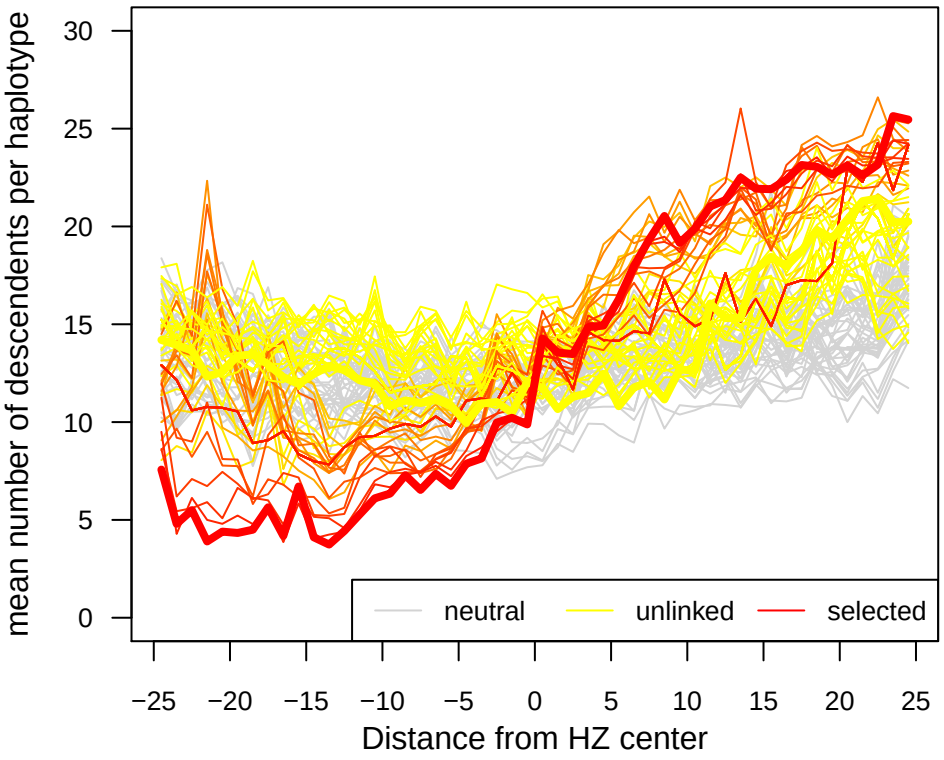

Figure 6: **Mean family size of haplotypes.** The number of individuals of ancestry B per number of ancestors ($F_B(m, x)$) from secondary contact occurring $T = 1000$ generations ago, represented across geographic space. The hybrid zone is 50 demes, each containing 500 diploid individuals. Each red or orange line represents a site some distance away (ranging from 0 - 0.01) 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.

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 (Porter *et al.* 1997; Gompert *et al.* 2012). 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 (Price *et al.* 2009), LAMP (Sankararaman *et al.* 2008) and fineSTRUCTURE (Lawson *et al.* 2012) can inform the demographic history of hybridization/admixture from present day samples (e.g. Hellenthal *et al.* 2014). 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 S12). 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 a few natural statistical summaries could help identify selected loci in hybrid zones. Specifically, the statistic $l_z(k)$ – the genetic length of ancestry surrounding a site of a given lineage – and similarly, and similarly $C(k)$, could provide a simple and informative summary. The ratio of mean adjacent block lengths $C(x, m)$ shows a very sharp peak at the selected site, suggesting that, despite the fact that haplotypes surround-

ing 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 previous studies have demonstrated that underdominant loci share similar properties with more realistic models (e.g., ecological selection (Barton & Hewitt 1989; Barton & Gale 1993)).

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 (Barton & Bengtsson 1986a), and the relatively short map distances between linked incompatibilities generates a longer unit which is not readily broken up by recombination (Barton 1986). 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 pattens of clines and block lengths, as well as the resolution to which one could detect targets of selection (Slatkin 1975; Barton 1983).

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 S8) 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 (Liang & Nielsen 2014); 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 reality. 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" (Turner *et al.* 2005; Nosil *et al.* 2009) and related patterns have been largely descriptive (Cruickshank & Hahn 2014; Noor & Bennett 2009). 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 (Barton 1979; Barton & Bengtsson 1986a; Martinsen *et al.* 2001; Arnold 2004). 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 describes at least roughly how quickly this should happen.

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 (1986a) The barrier to genetic exchange between hybridising populations. *Heredity*, **56**, 357–376.

Barton N, Bengtsson B (1986b) 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 (1979) Gene flow past a cline. *Heredity*, **43**, 333–339.

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, Shpak M (2000) The effect of epistasis on the structure of hybrid zones. *Genetical Research*, **75**, 179–198.

Bazykin AD (1969) Hypothetical mechanism of speciaton. *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, Andreasen V, Poulsen ET (1995) Genotypic proportions in hybrid zones. *Journal of Mathematical Biology*, **33**, 225–249.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates.

560 Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation
561 are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

562 Eddelbuettel D, Francois R (2011) Rcpp: Seamless R and C++ integration. *Journal of*
563 *Statistical Software*, **40**, 1–18.

564 Felsenstein J (1975a) Genetic drift in clines which are maintained by migration and natural
565 selection. *Genetics*, **81**, 191–207.

566 Felsenstein J (1975b) A pain in the torus: Some difficulties with models of isolation by
567 distance. *The American Naturalist*, **109**, 359–368.

568 Fisher R (1954) A fuller theory of "junctions" in inbreeding. *Heredity*, **8**, 187–197.

569 Fisher RA (1950) Gene frequencies in a cline determined by selection and diffusion. *Bio-*
570 *metrics*, **6**, pp. 353–361.

571 Gompert Z, Parchman TL, Buerkle CA (2012) Genomics of isolation in hybrids. *Philosoph-*
572 *ical transactions of the Royal Society of London B Biological Sciences*, **367**, 439–50.

573 Gravel S (2012) Population genetics models of local ancestry. *Genetics*, **191**, 607–19.

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

575 Hanson WD (1966) Effects of partial isolation (distance), migration, and different fitness re-
576 quirements among environmental pockets upon steady state gene frequencies. *Biometrics*,
577 **22**, pp. 453–468.

578 Hellenthal G, Busby GBJ, Band G, *et al.* (2014) A genetic atlas of human admixture history.
579 *Science*, **343**, 747–51.

580 Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recom-
581 bination. *Genetics*, **120**, 831–840.

582 Kruuk LE, Baird SJ, Gale KS, Barton NH (1999) A comparison of multilocus clines main-
583 tained by environmental adaptation or by selection against hybrids. *Genetics*, **153**, 1959–
584 71.

585 Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using
586 dense haplotype data. *PLoS genetics*, **8**, e1002453.

587 Liang M, Nielsen R (2014) The lengths of admixture tracts. *Genetics*, **197**, 953–967.

588 Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter
589 gene introgression between species. *Evolution*, **55**, 1325–1335.

590 Nagylaki T (1975) Conditions for the existence of clines. *Genetics*, **80**, 595–615.

591 Noor MAF, Bennett SM (2009) Islands of speciation or mirages in the desert? examining
592 the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

593 Nosil P, Funk DJ, Oritz-Barrientos D (2009) Divergent selection and heterogeneous genomic
594 divergence. *Molecular Ecology*, **18**, 375–402.

595 Polechová J, Barton N (2011) Genetic drift widens the expected cline but narrows the
596 expected cline width. *Genetics*, **189**, 227–235.

597 Pool JE (2015) The mosaic ancestry of the Drosophila genetic reference panel and the *D.*
598 *melanogaster* reference genome reveals a network of epistatic fitness interactions. *Molec-*
599 *ular Biology and Evolution*, **32**, 3236–3251.

600 Porter AH, Wenger R, Geiger H, Scholl A, Shapiro AM (1997) The pontia daplidice-edusa
601 hybrid zone in northwestern italy. *Evolution*, **51**, 1561–1573.

602 Price AL, Tandon A, Patterson N, *et al.* (2009) Sensitive detection of chromosomal segments
603 of distinct ancestry in admixed populations. *PLoS genetics*, **5**, e1000519.

604 Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in
605 admixed populations. *The American Journal of Human Genetics*, **82**, 290 – 303.

606 Schumer M, Brandvain Y (2016) Determining epistatic selection in admixed populations.
607 *Molecular Ecology*, S. ??–??

608 Sedghifar A, Brandvain Y, Ralph P, Coop G (2015) The spatial mixing of genomes in
609 secondary contact zones. *Genetics*, **201**, 243–61.

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

611 Slatkin M (1975) Gene flow and selection in a two-locus system. *Genetics*, **81**, 787–802.

612 Slatkin M, Maruyama T (1975) Genetic drift in a cline. *Genetics*, **81**, 209–22.

613 Soetaert K, Meysman F (2012) Reactive transport in aquatic ecosystems: Rapid model
614 prototyping in the open source software R. *Environmental Modelling & Software*, **32**,
615 49–60.

616 Soetaert K, Petzoldt T, Setzer RW (2010) Solving differential equations in R: Package deS-
617 olve. *Journal of Statistical Software*, **33**, 1–25.

618 Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles*
619 *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 occuring 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:

$$x \xrightarrow{AA} y \qquad w_{AA}(x,y) = p(x)s_A(x)\left(1 - r(1-p(x))\right)N(x)m(x,y) \tag{6}$$

$$x \xrightarrow{AB} y \qquad w_{AB}(x,y) = rp(x)(1-p(x))\frac{s_A(x)+s_B(x)}{2}N(x)m(x,y) \tag{7}$$

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

$$x \xrightarrow{BB} y \qquad w_{BB}(x,y) = (1-p(x))s_B(x)\left(1 - rp(x)\right)N(x)m(x,y) \tag{9}$$

Note that at equilibrium, we require N to solve

$$0 = \sum_y N(y)m(y,x)\left\{p(y)s_A(y)(1-p(x)) - (1-p(y))s_B(y)p(x)\right\}. \tag{10}$$

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

$$r_A(x,y) = \frac{w_{AA}(y,x)}{p(x)N(x)} \tag{11}$$

$$= N(y)p(y)s_A(y)m(y,x)\frac{1}{N(x)p(x)}. \tag{12}$$

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

$$\frac{d}{dt}\mathbb{E}[f(X_t) \mid X_0 = x] = \sum_y r_A(y,x)(f(y) - f(x)) \tag{13}$$

$$= \frac{1}{N(x)p(x)}\sum_y N(y)p(y)s_A(y)m(y,x)(f(y) - f(x)). \tag{14}$$

Diffusion limit

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

$$\lim_{\sigma\to 0}\sum_y \frac{m(x,y)(f(y) - f(x))}{\sigma^2} = \frac{1}{2}\frac{d^2}{dx^2}f(x). \tag{15}$$

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

$$\frac{1}{\sigma^2}\sum_y g(y)m(y,x)(f(y)-f(x)) = \frac{1}{\sigma^2}\sum_y m(y,x)\left(g(y)f(y)-g(x)f(x)+(g(x)-g(y))f(x)\right) \tag{16}$$

$$= \frac{1}{\sigma^2}\sum_y m(y,x)\left(g(y)f(y)-g(x)f(x)\right) \tag{17}$$

$$- f(x)\frac{1}{\sigma^2}\sum_y m(y,x)(g(y)-g(x)) \tag{18}$$

$$\xrightarrow{\sigma\to 0} \frac{1}{2}\frac{d^2}{dx^2}\left(g(x)f(x)\right) - \frac{1}{2}f(x)\frac{d^2}{dx^2}g(x) \tag{19}$$

$$= \frac{1}{2}\left(g(x)f'(x)+2g'(x)f'(x)+f(x)g''(x)-f(x)g''(x)\right) \tag{20}$$

$$= \frac{1}{2}g(x)f''(x)+g'(x)f'(x). \tag{21}$$

Under these assumptions, combining (14) and (21),

$$\frac{d}{dt}\mathbb{E}[f(X_{t/\sigma^2})\,|\,X_0=x] \xrightarrow{\sigma\to 0} \frac{1}{2}s_A(x)\frac{d^2}{dx^2}f(x) + \frac{1}{N(x)p(x)}\frac{d}{dx}\left\{N(x)p(x)s_A(x)\right\}\frac{d}{dx}f(x) \tag{22}$$

$$= \frac{1}{2}s_A(x)f''(x) + \left(s_A'(x)+\log(N(x)p(x))'s_A(x)\right)f'(x), \tag{23}$$

i.e. X_{t/σ^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 σ^2 implies that the Brownian noise and the mean displacement should both be scaled by σ^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 (1), presents no difficulty; we use the ReacTran package (Soetaert & Meysman 2012) to compute discrete approximations to the diffusion term, and the deSolve package (Soetaert *et al.* 2010) to solve the equation.

The equations (4) describing probabilities that a lineage descends from A ancestry, con-

ditional 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 Soetaert & Meysman (2012) 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 (5), 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 Sedghifar *et al.* (2015). Here, we solved the equations numerically, again discretizing the equations and using the ode.1D function of the deSolve package. The reaction-diffusion part is the same as for equations (4). 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 (5) is

$$\int_a^b g_A(a,\theta)g_A(\theta,b)d\theta. \tag{24}$$

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 θ, the two segments (a,θ) and (θ,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, $r_1 < \ldots < r_n$. 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 (24) 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 $r_i, r_{i+1}, \ldots, r_j$ to be of ancestry A, when a recombination occurrs between r_k and

r_{k+1}, the sequences of alleles at $r_i, \ldots, r_k$ and at $r_{k+1}, \ldots, r_j$ must each be of ancestry A. The first integral term in (5) is

$$\int_a^0 g_B(a,\theta) g_A(\theta, b) d\theta + \int_0^b g_A(a,\theta) g_B(\theta, b) d\theta \tag{25}$$

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 (25) 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 (Eddelbuettel & Francois 2011). 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.

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

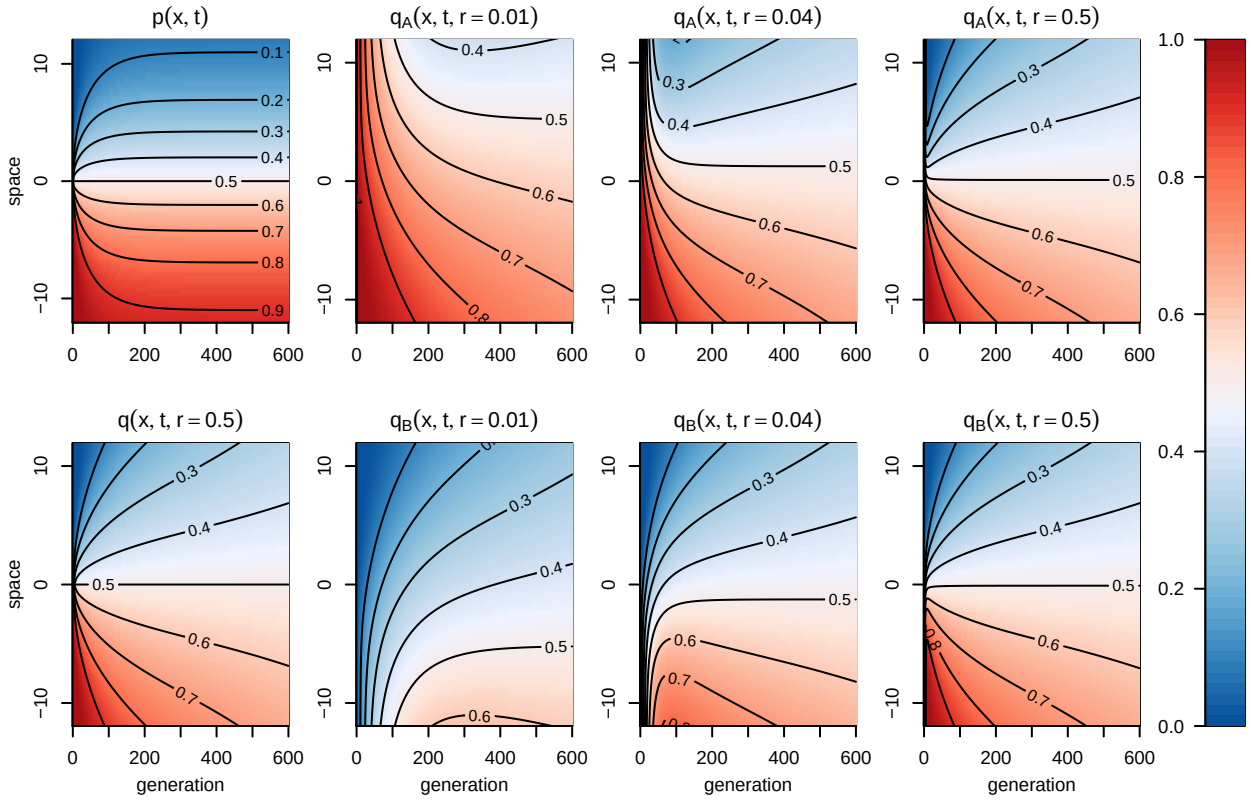

Figure S1: **Probabilities of A ancestry,** across space (vertical axis, in units of σ) and time (horizontal axis, in generations). In each plot, color corresponds to the expected frequency of A ancestry at a particular location in time and space. The selection coefficient is $s = .02$. **Top left:** at the selected site, showing establishment and stabilization of the cline on a time scale of $1/s = 50$ generations. **Bottom left:** at an unliked 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 2 for the same figure over a shorter period of time.

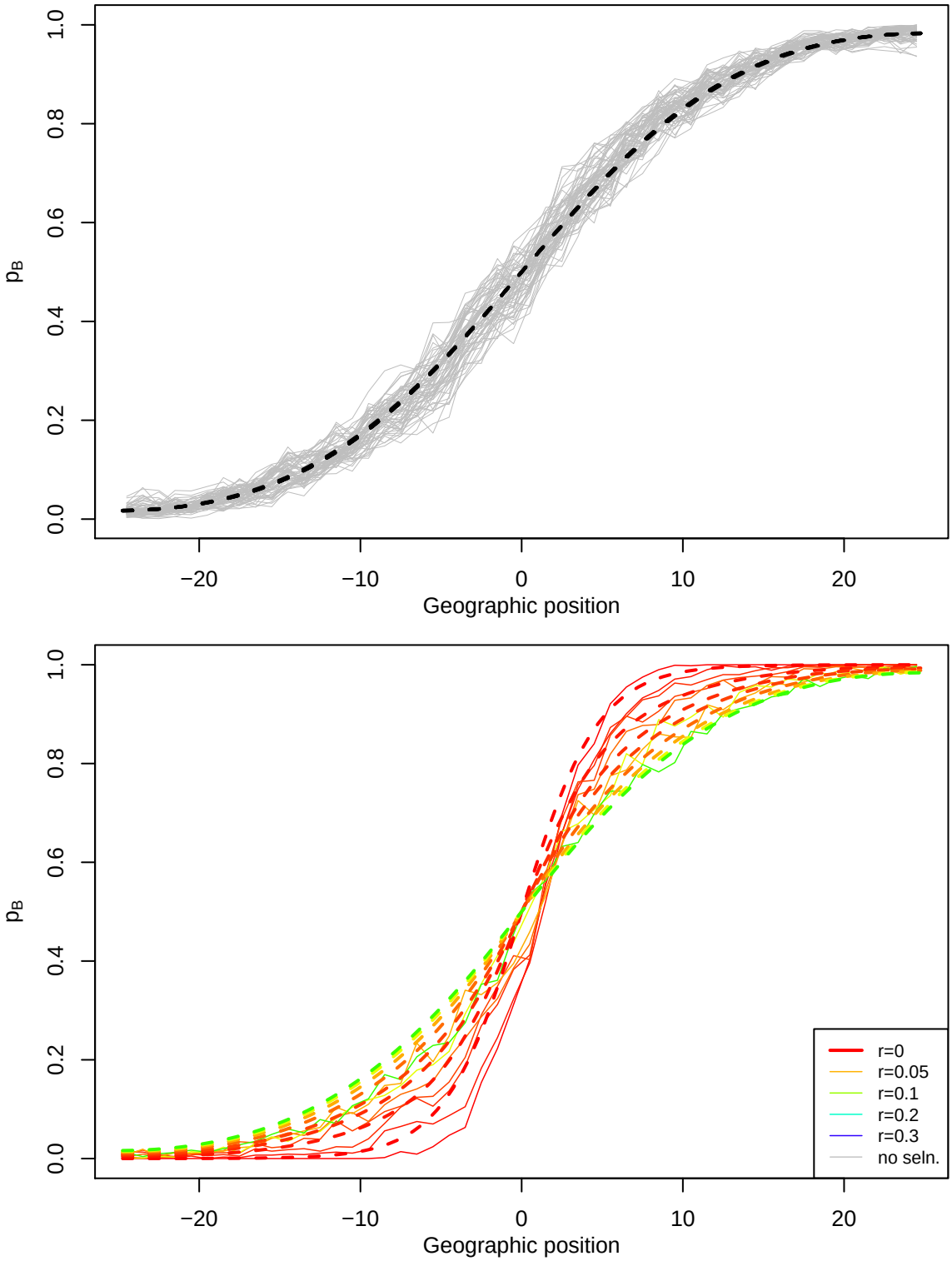

Figure S2: **Comparison of simulated to theoretical clines:** As in Figure 3, frequency of ancestry B, across geography and at several physical positions on the genome, at $T = 100$ generations after secondary contact, and at $s = 0.1$. Dotted lines are expected frequencies of B ancestry $1-q(x,t,r)$ (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.

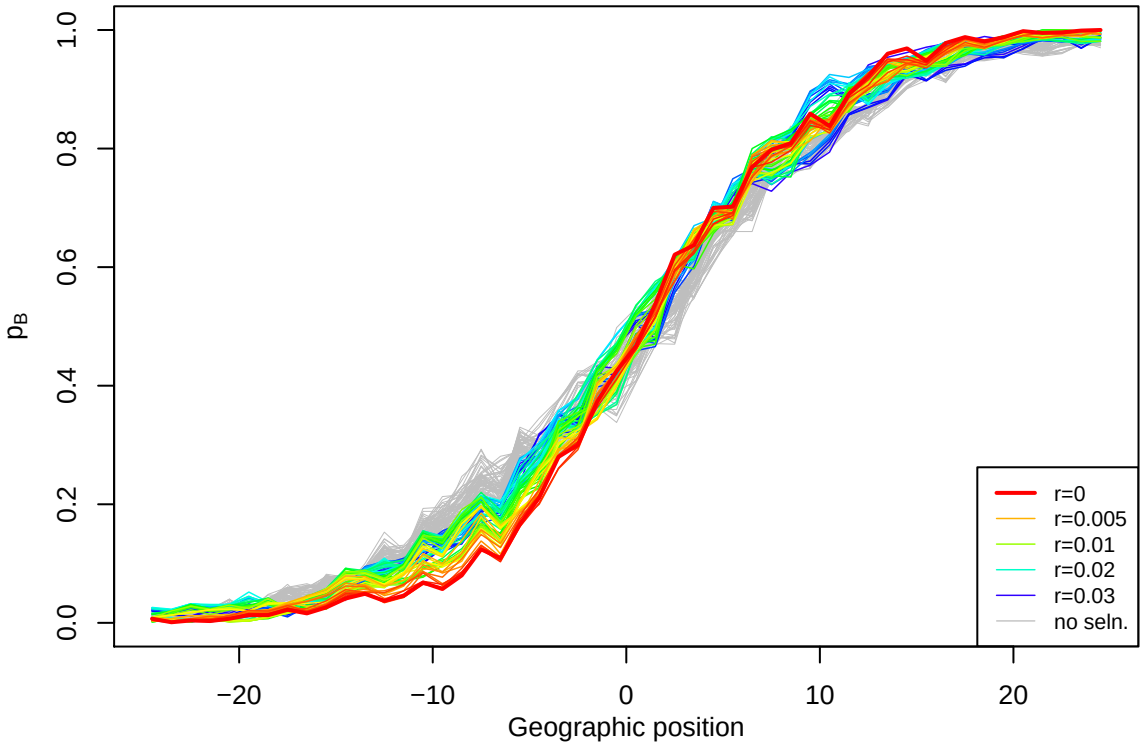

Figure S3: Frequency of ancestry B, across geography and at several physical positions on the genome, simulated for a hybrid zone $T = 100$ generations after secondary contact, and with $s = 0.01$. 50 demes, each with a population size of 500 diploid individuals. Each line represents a locus some distance 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 $s = 0$.

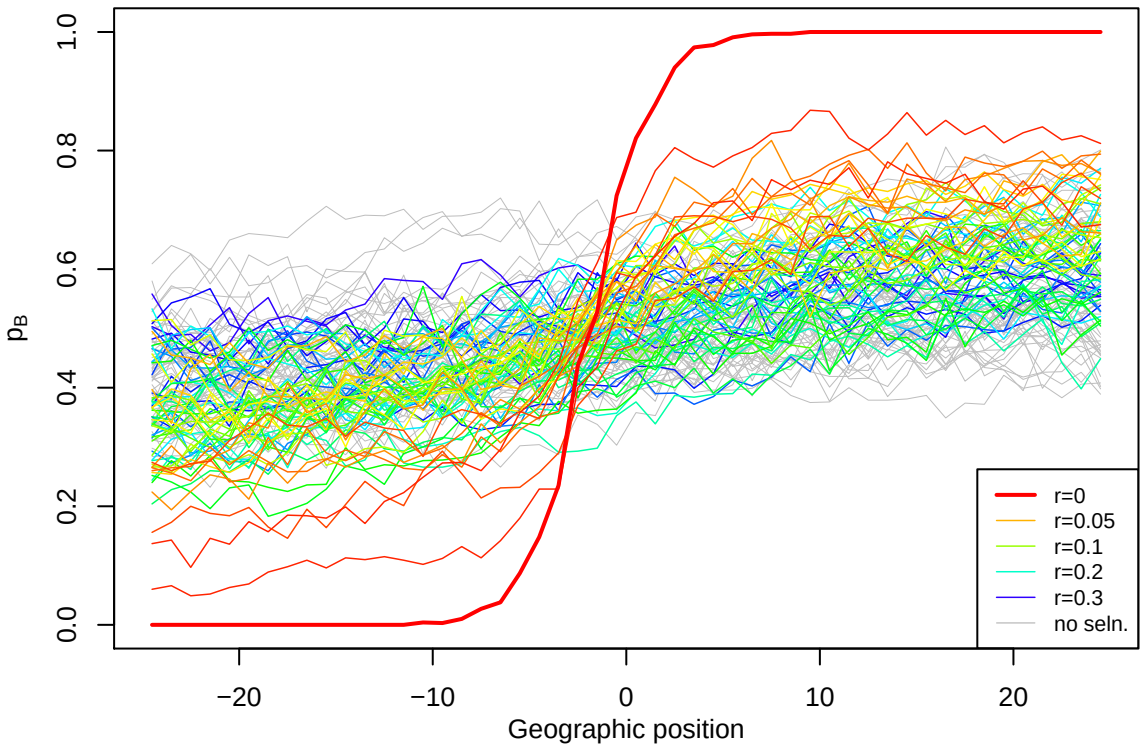

Figure S4: Frequency of ancestry B, across geography and at several physical positions on the genome, simulated for a hybrid zone $T = 1000$ generations after secondary contact, and with $s = 0.1$. 50 demes, each with a population size of 500 diploid individuals. Each line represents a locus some distance 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 $s = 0$.

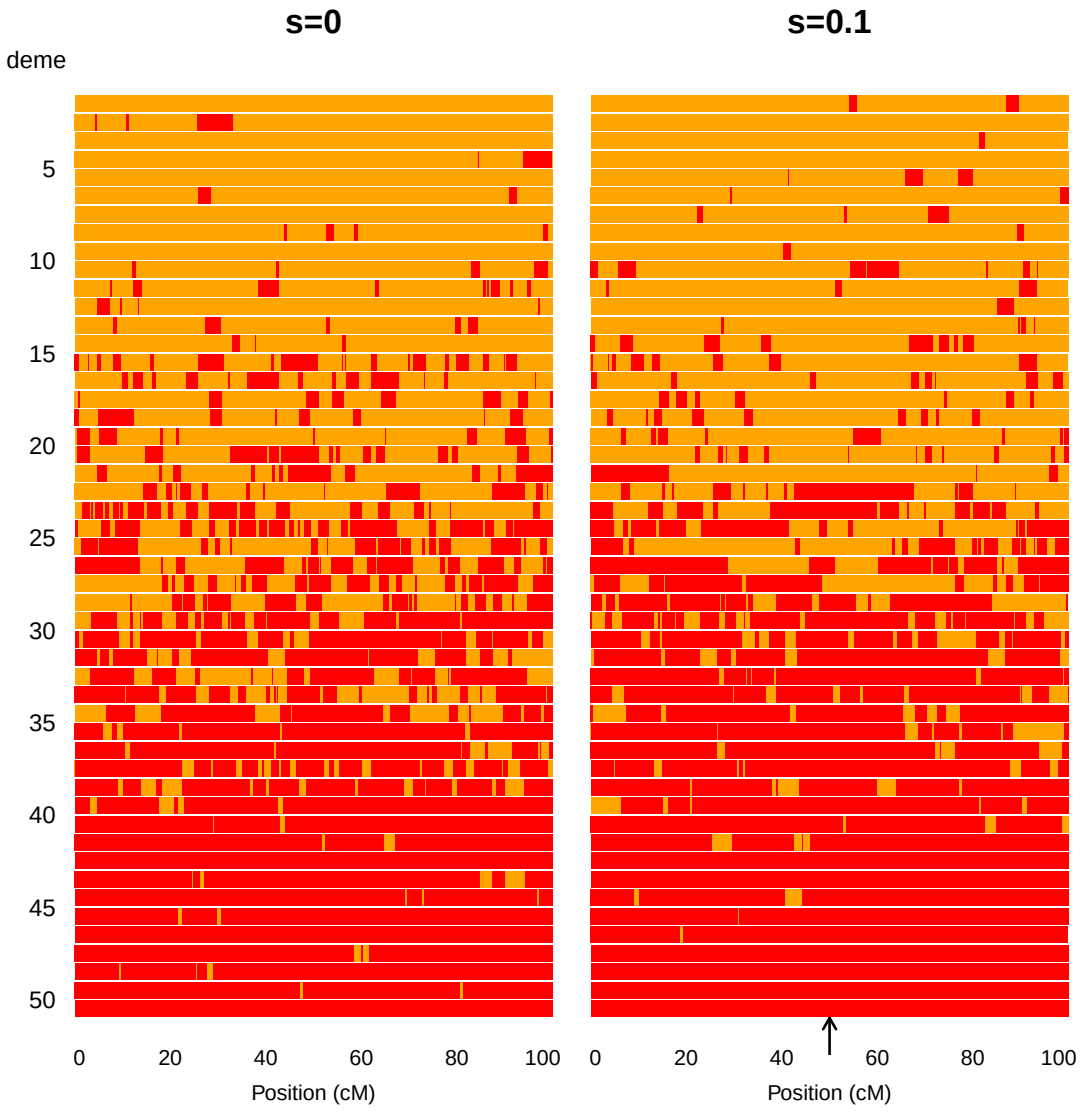

Figure S5: Randomly sampled chromosomes across a hybrid zone of age $T = 100$. Here we compare chromosomes of length 1M from a neutral zone to one that has a single under-dominant locus ($s = 0.1$) in the middle of the chromosome (indicated by black arrow). Red blocks along the chromosome denote ancestry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

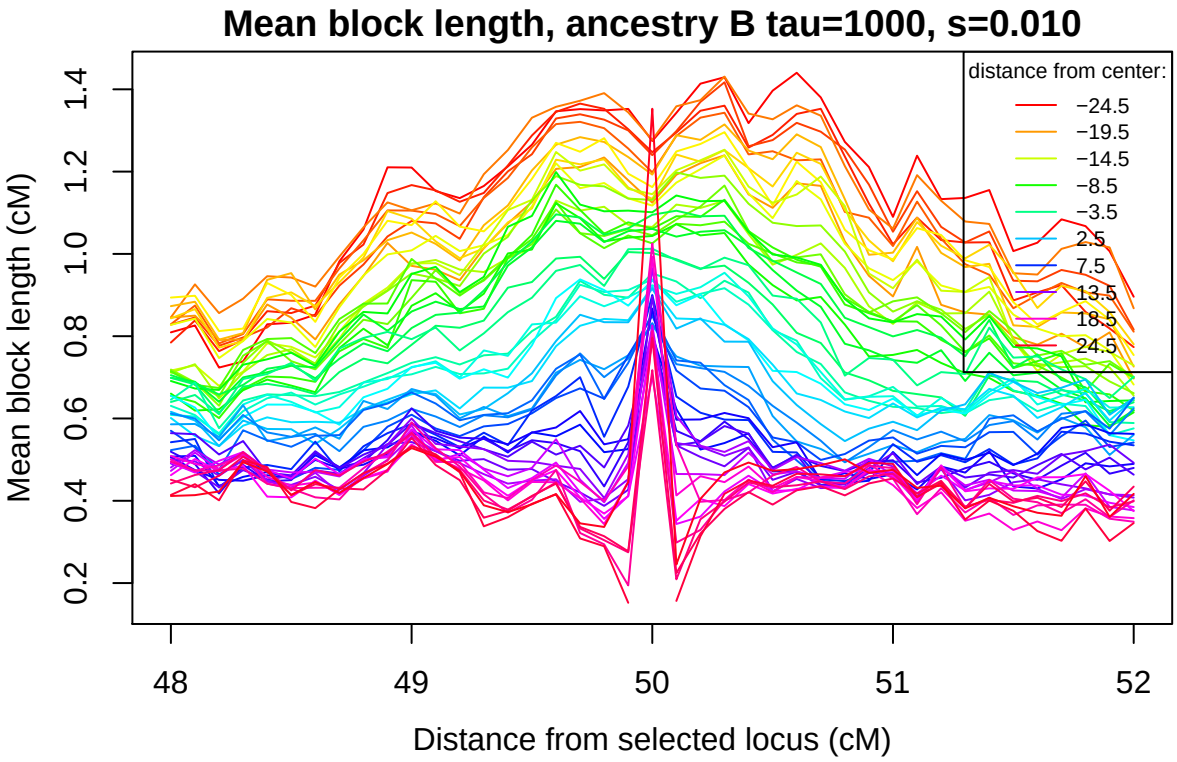

Figure S6: **Mean haplotype lengths of B haplotypes**, $l_B(x, m)$, across the genome (horizontal axis) and at differntial spatial locations (colored lines), from a simulation with 50 demes having 500 individuals each, $s = 0.01$, $\sigma = 1$, and after $T = 1000$ generations.

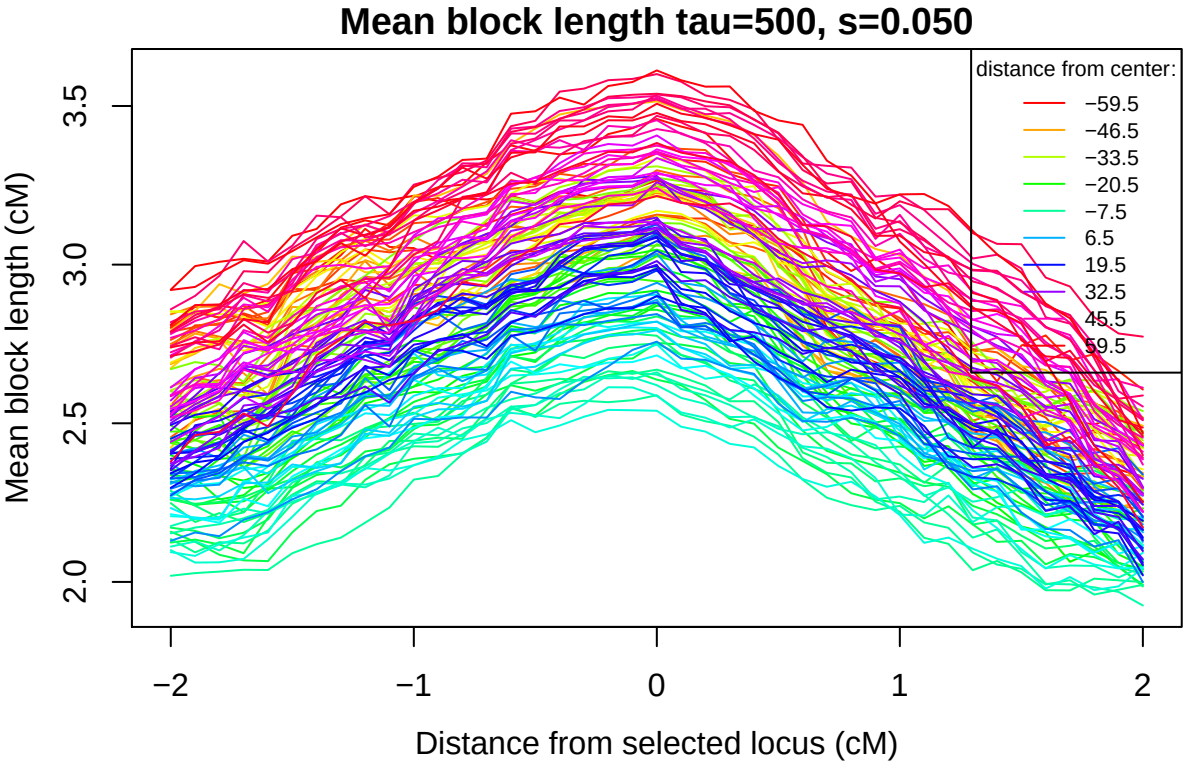

Figure S7: **Mean enclosing block length** $l(m, x)$, across the genome (horizontal axis) and at different geographic positions (different colored lines). Results are from a simulation with 120 demes of 200 diploids each, selection $s = .05$, dispersal $\sigma = 3$, and after $T = 500$ generations.

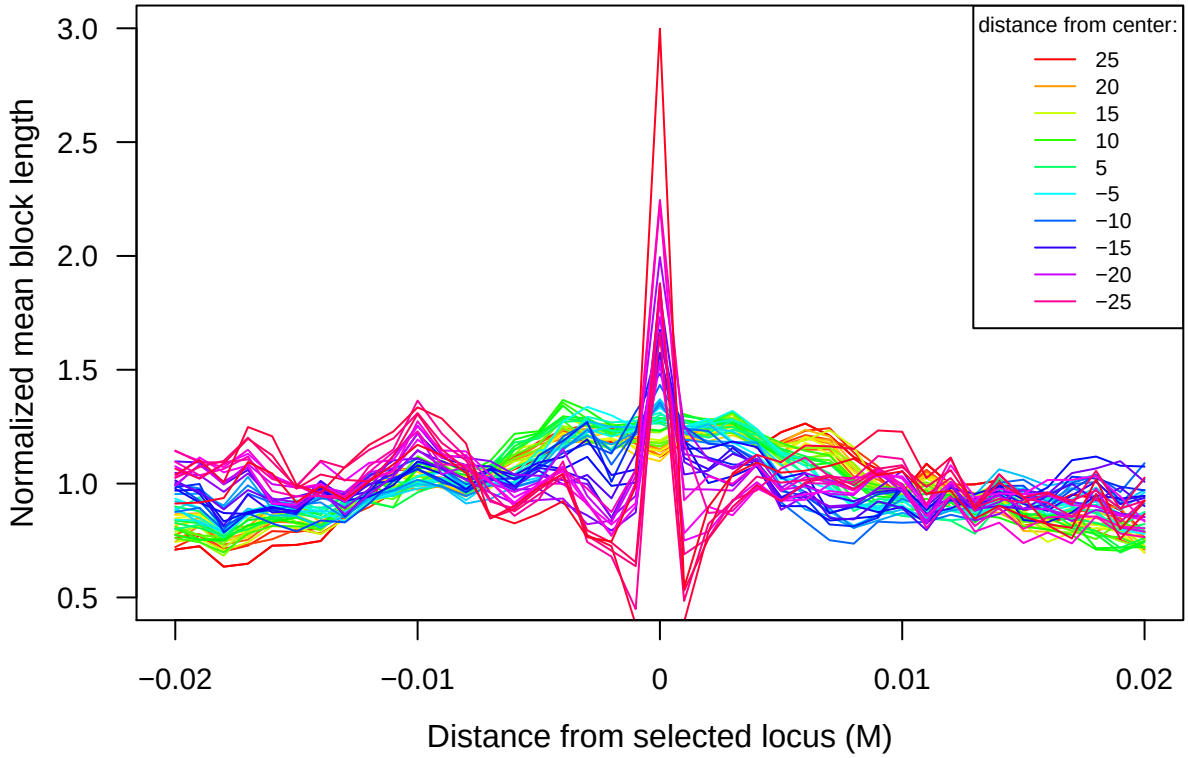

Figure S8: **Normalized mean enclosing block length**, $l_B(m, x)$, after $T = 1000$ generations, against position relative to the selected locus (horizontal axis) located in the center of a 1M chromosome. Each line shows the mean block length at that spatial and genomic position divided by the mean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

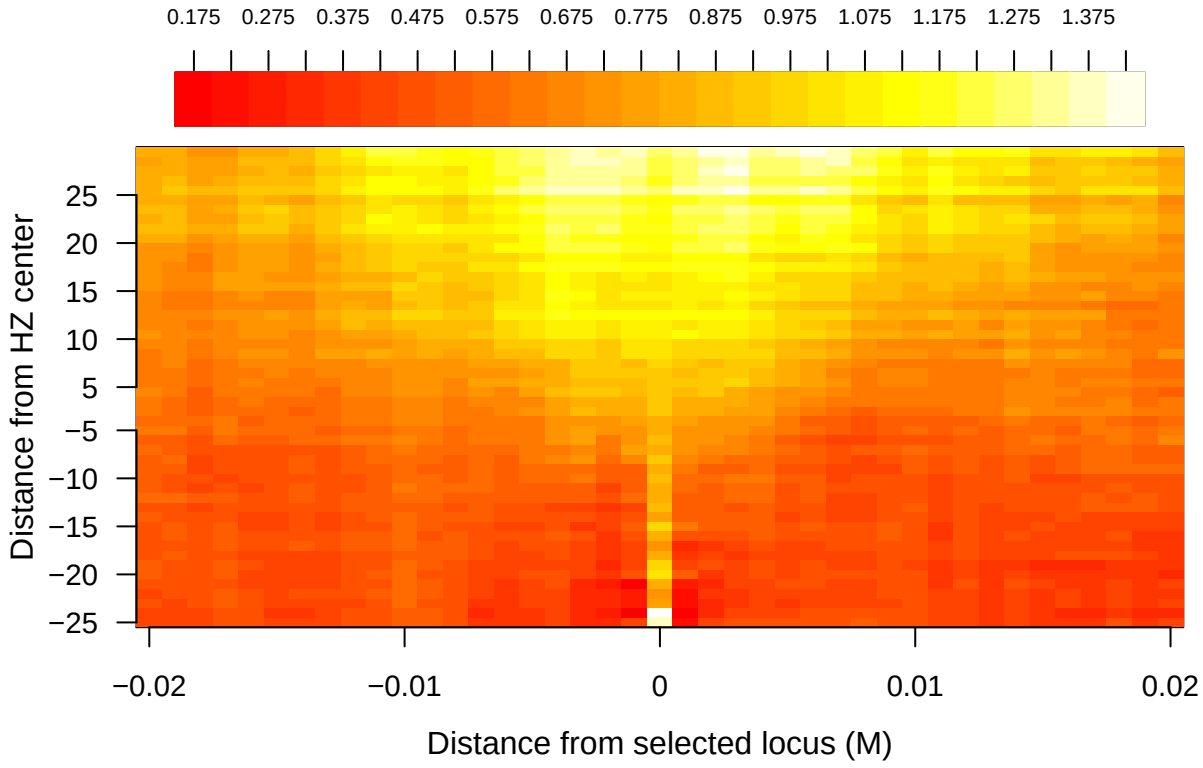

Figure S9: Heatmap of mean block length $l_B(m)$ along a simulated chromosome.

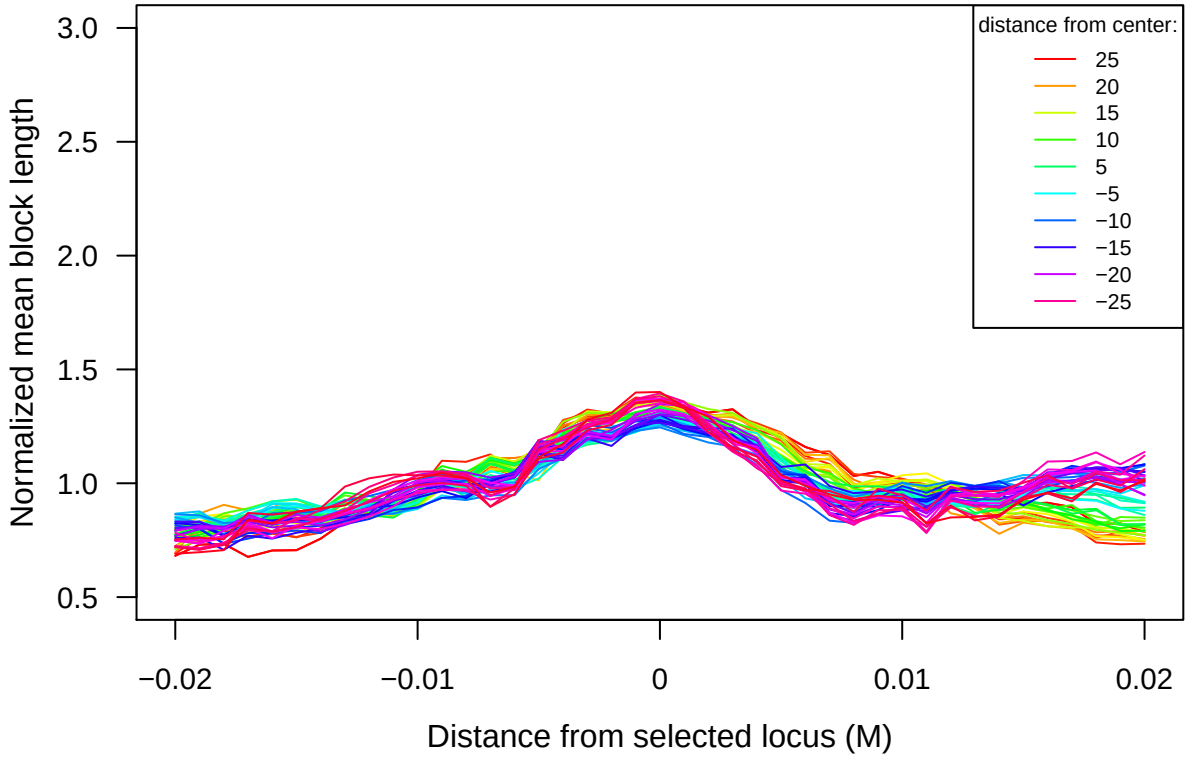

Figure S10: Heatmap of mean block length $l(m)$ surrounding a given position along the genome with a single underdominant site ($s = 0.01, T = 1000$).

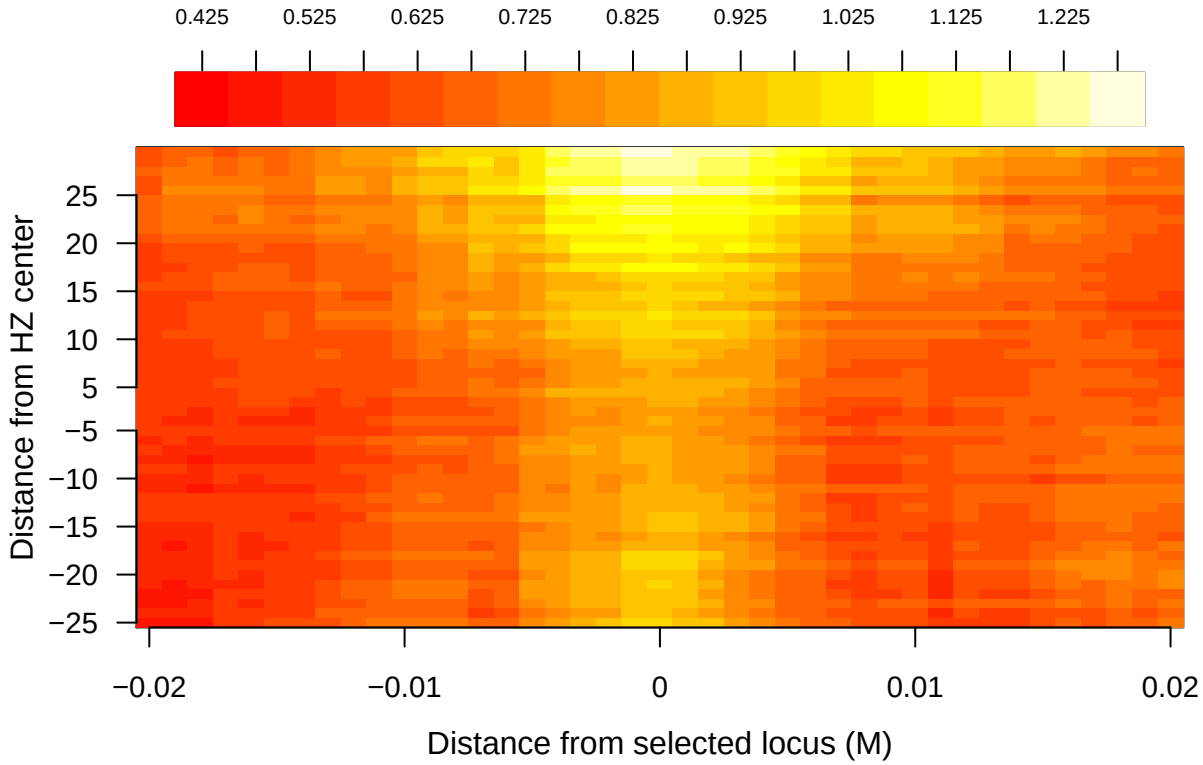

Figure S11: Heatmap of mean block length $l(m)$ along a simulated chromosome with a single underdominant site ($s = 0.01, T = 1000$)

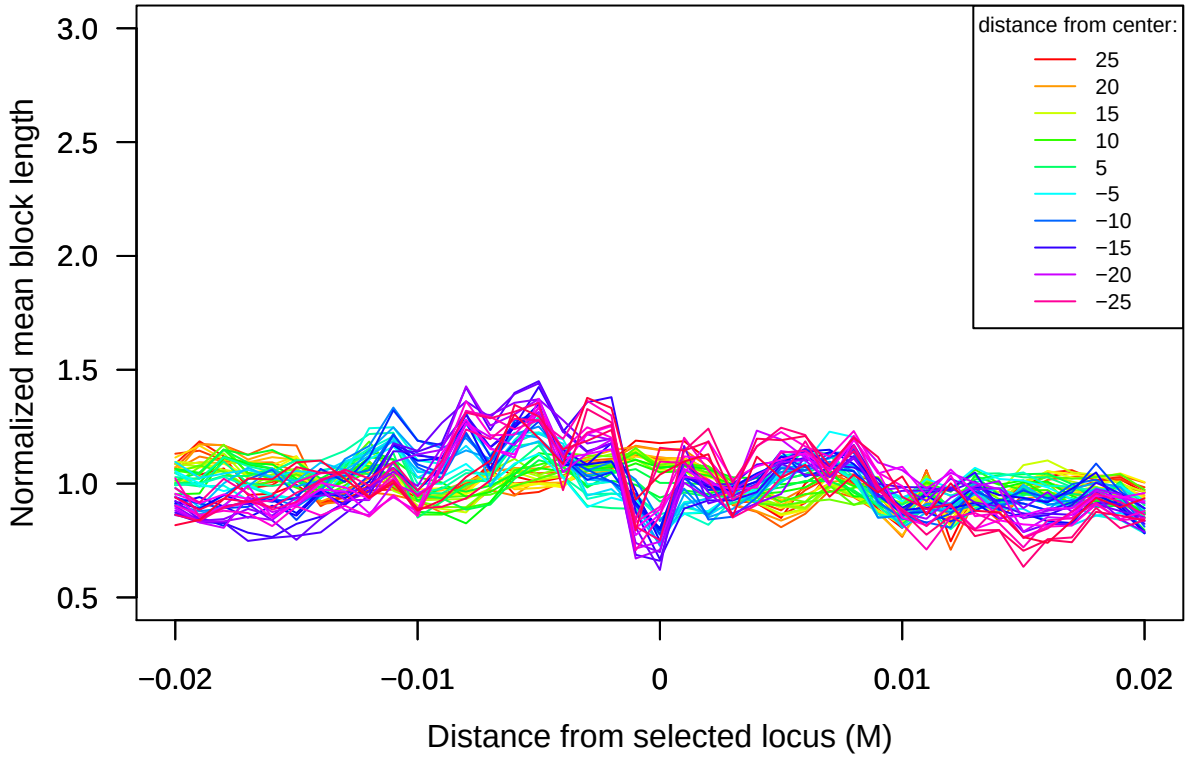

Figure S12: Mean block length of $l_A(m\pm)$ across chromosome with single under dominant site, conditioning on ancestry B at the selected locus. $(s = 0.01, T = 1000)$

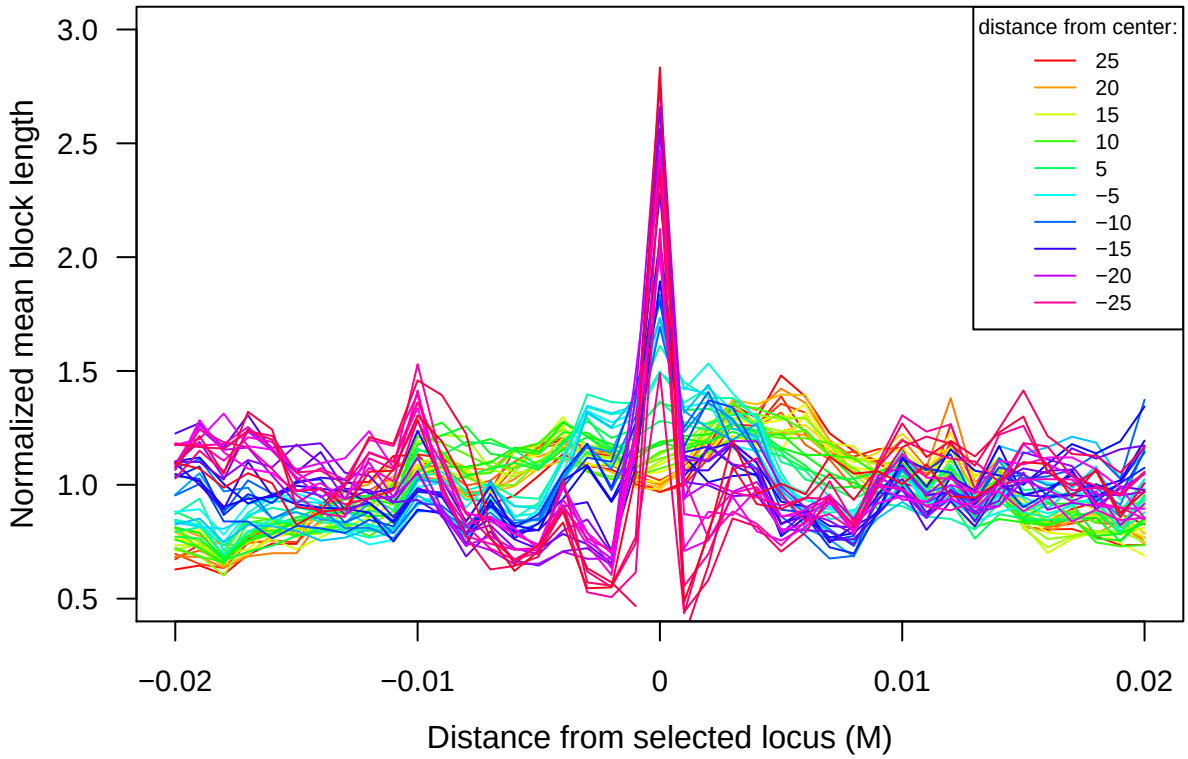

Figure S13: Ratio $\frac{2\sum l_B(m_i)}{\sum l_A(m_i-)+I_A(m_i+)}$ of mean block length and adjacent block lengths across a simulated chromosome with a single underdominant site and conditioning on ancestry B at the selected site $(s = 0.01, T = 1000)$. Each line represents a deme and is normalized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

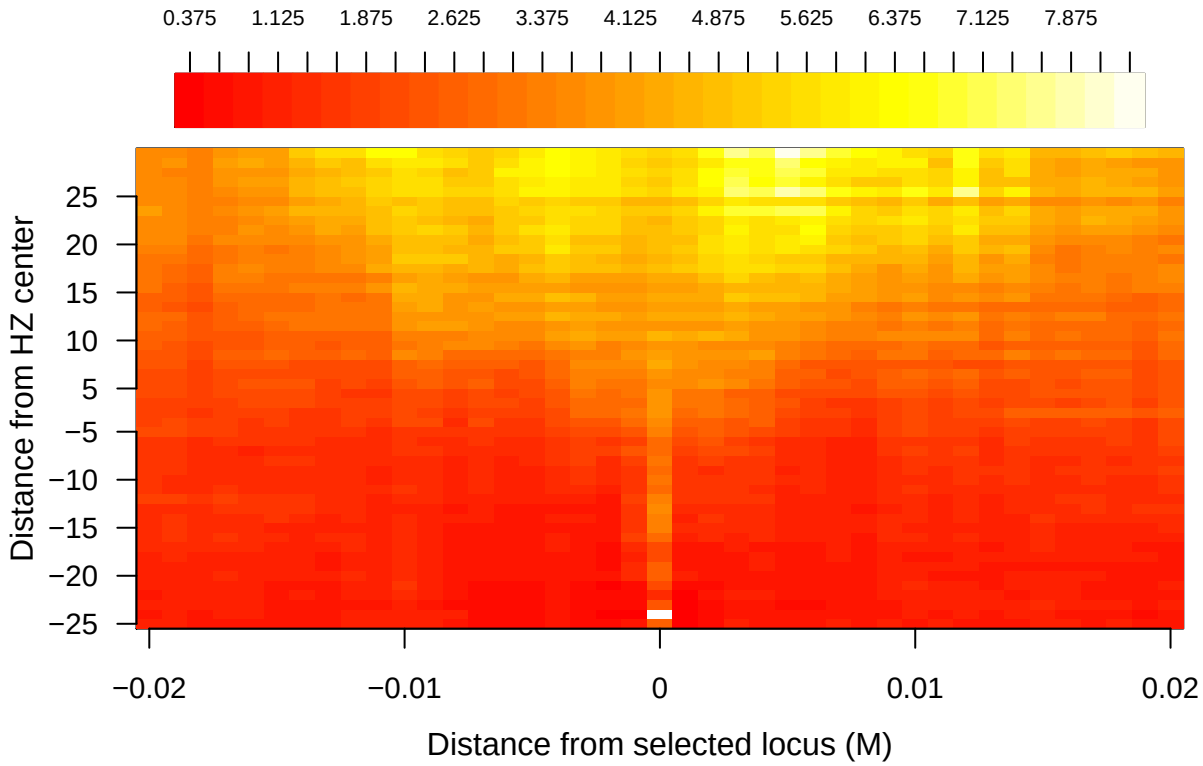

Figure S14: Heatmap of $\frac{2\sum l_B(m_i)}{\sum l_A(m_i-)+l_A(m_i+)}$ across a simulated chromosome with a single underdominant site and conditioning on ancestry B at the selected site ($s = 0.01, T = 1000$).

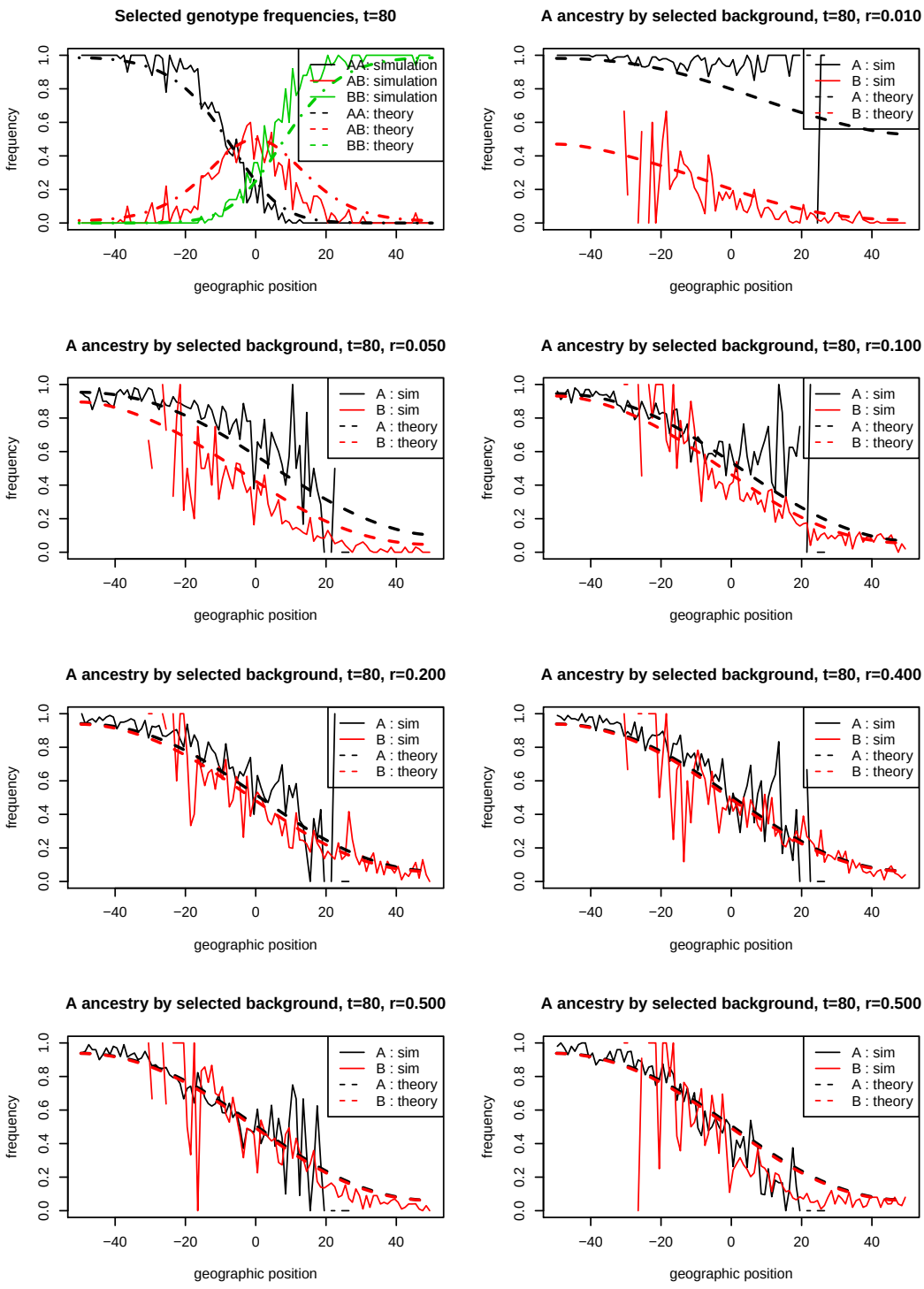

Figure S15: **Conditional frequencies of ancestry A at $\tau = 80$**, comparing simulation and theory, with $\sigma = 3$ deme spacings, $s = 0.05$, and 50 individuals per deme. The top left figure shows observed and expected genotype frequencies for the two homozygotes and the heterozygote at the selected locus; expected genotype counts were obtained assuming random mating, and by solving equation (1) 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.

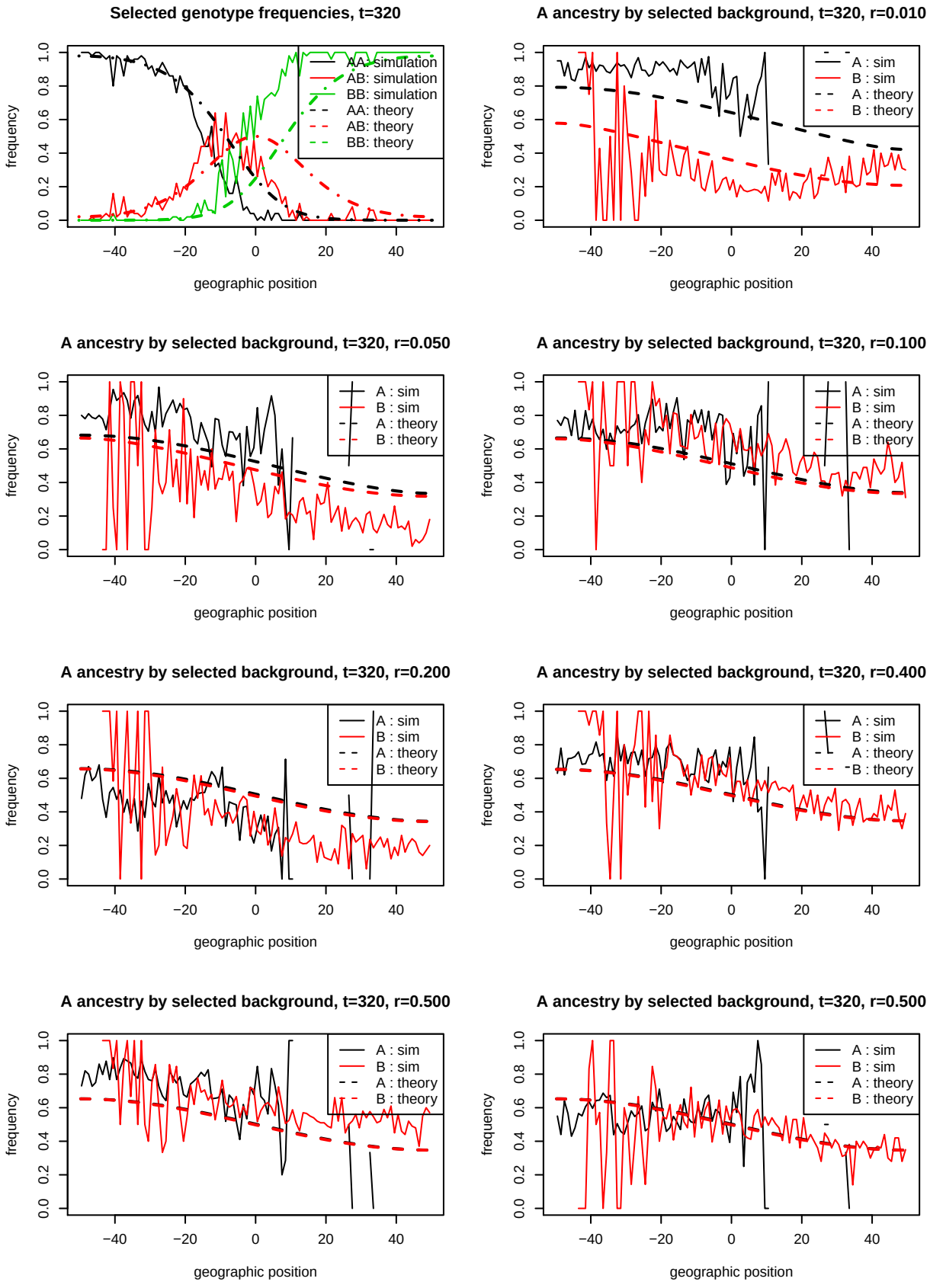

Figure S16: **Conditional frequencies of ancestry A at $\tau = 320$**, as in figure S15. Deviations are larger than at $\tau = 80$, due to genetic drift.

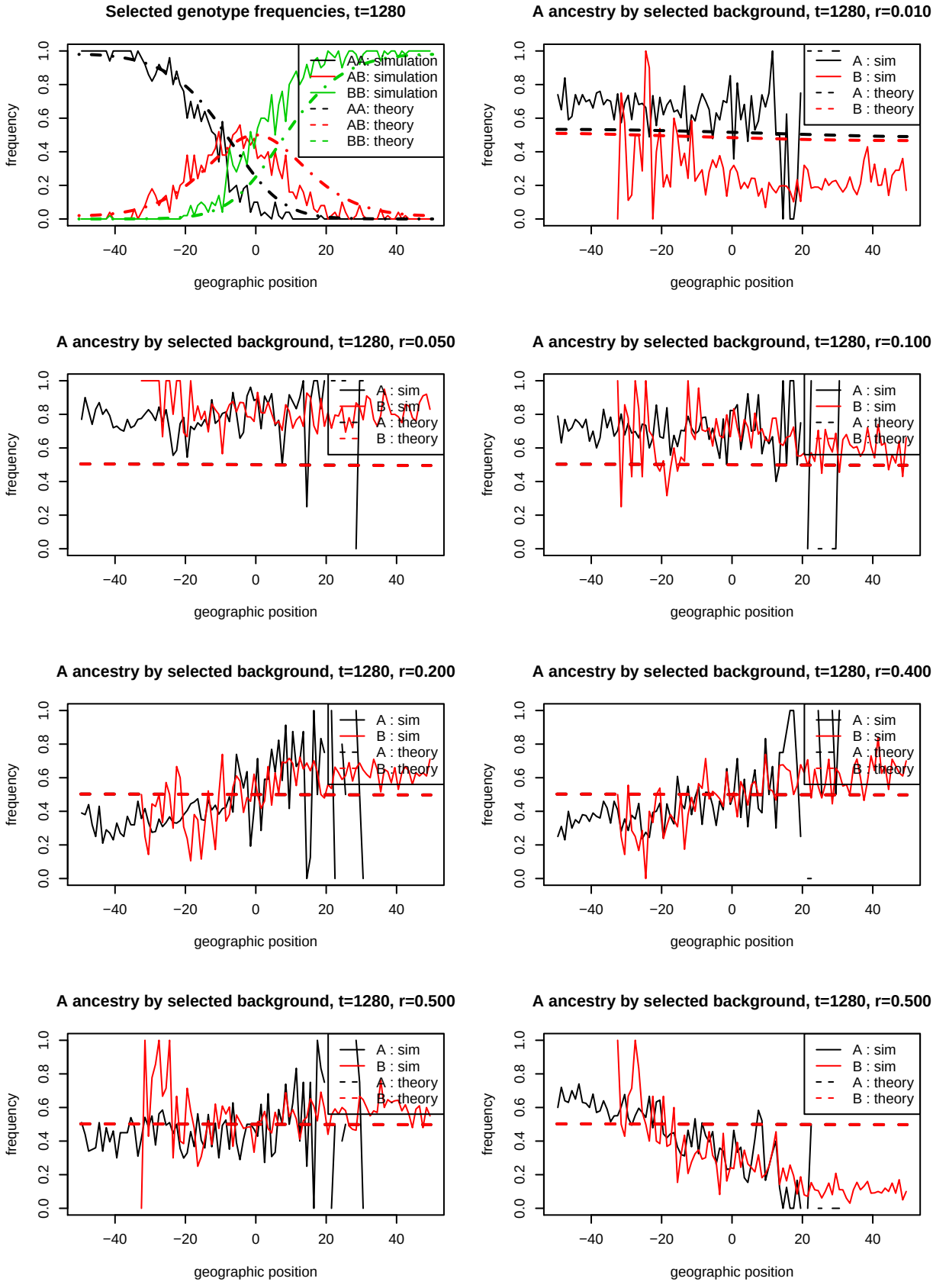

Figure S17: **Conditional frequencies of ancestry A at $\tau = 1280$**, as in figure S15. Deviations are larger still than at $\tau = 320$, due to genetic drift.