Replicability of Introgression Under Linked, Polygenic Selection

Himani Sachdeva and Nicholas H. Barton¹

Institute of Science and Technology Austria, A-3400 Klosterneuburg, Austria

ORCID ID: 0000-0002-8548-5240 (N.H.B.)

ABSTRACT We study how a block of genome with a large number of weakly selected loci introgresses under directional selection into a genetically homogeneous population. We derive exact expressions for the expected rate of growth of any fragment of the introduced block during the initial phase of introgression, and show that the growth rate of a single-locus variant is largely insensitive to its own additive effect, but depends instead on the combined effect of all loci within a characteristic linkage scale. The expected growth rate of a fragment is highly correlated with its long-term introgression probability in populations of moderate size, and can hence identify variants that are likely to introgress across replicate populations. We clarify how the introgression probability of an individual variant is determined by the interplay between hitchhiking with relatively large fragments during the early phase of introgression and selection on fine-scale variation within these, which at longer times results in differential introgression probabilities for beneficial and deleterious loci within successful fragments. By simulating individuals, we also investigate how introgression probabilities at individual loci depend on the variance of fitness effects, the net fitness of the introduced block, and the size of the recipient population, and how this shapes the net advance under selection. Our work suggests that even highly replicable substitutions may be associated with a range of selective effects, which makes it challenging to fine map the causal loci that underlie polygenic adaptation.

KEYWORDS introgression; Hill–Robertson interference; polygenic adaptation

THE extent to which phenotypic and genetic changes are replicated during adaptation across closely related populations has generated much interest (Conte *et al.* 2012; Storz 2016), and is part of the broader question of the predictability of evolutionary change (Lässig *et al.* 2017). Parallel evolution can be investigated at multiple scales, and may refer to the involvement of the same nucleotides, the same genes, or even the same pathways during adaptive responses in different populations (Manceau *et al.* 2010). An important challenge is to interpret highly replicable genetic loci; do such loci necessarily make large contributions to the selected trait? Conversely, for different architectures of the selected trait, how often are the same genetic loci implicated across replicate populations?

Several factors can influence the replicability of allele frequency changes or adaptive substitutions at individual loci (Stern 2013). More shared variation between replicate populations makes it more likely that the same variants respond to selection

across different replicates, but the response is then associated with a weaker reduction in neutral diversity, [i.e., “soft sweeps,” see Hermisson and Pennings (2005)]. By contrast, adaptive substitutions that arise *de novo* are rarely shared across replicates, but can be more easily identified as being under selection because of the associated hard sweep patterns.

Factors influencing genetic parallelism at the level of individual variants remain poorly understood for selection on highly polygenic traits, despite several studies with natural and laboratory replicates (Burke *et al.* 2010; Chan *et al.* 2012; Yeaman *et al.* 2016). Polygenic traits are often characterized by high genotypic redundancy, with multiple genotypes corresponding to approximately the same phenotype. For traits under stabilizing selection, a shift in the selection optimum can thus result in a highly heterogeneous response at the genotypic level, with selection amplifying initial random fluctuations in allele frequencies, to produce quite different outcomes across replicates, even if these initially share the same variation.

In general, adaptation in complex traits involves a large number of partially linked and weakly selected variants, and is thus characterized by pervasive hitchhiking, where effectively neutral or deleterious variants are swept to high frequencies

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.118.301429>

Manuscript received July 28, 2018; accepted for publication September 28, 2018; published Early Online October 1, 2018.

¹Corresponding author: Institute of Science and Technology Austria (IST Austria), Am Campus 1, A-3400 Klosterneuburg, Austria. E-mail: nick.barton@ist.ac.at

along with clusters of positively selected variants (Barton 1995). While the effects of hitchhiking on neutral diversity have been studied extensively in the theoretical literature, most of these studies assume that selected variants are either all deleterious or all beneficial in a given population (e.g., Barton and Bengtsson 1986). Moreover, most empirical studies also separately consider the effect of positive or background selection on patterns of diversity, divergence, or introgression [though see Elyashiv *et al.* (2016) for an investigation of the joint effects of the two]. However, a typical genome is likely to harbor many linked variants with a range of (positive and negative) selective effects. The efficacy of selection in discriminating between multiple, tightly linked beneficial and deleterious loci is an essential determinant of the total phenotypic response to selection in a finite population (Robertson 1970, 1977). It is also key to assessing whether an allele frequency change (even one that occurs across replicates) is itself adaptive and not a consequence of hitchhiking.

The extent of hitchhiking (and the associated patterns of neutral diversity) are strongly influenced by the density of selected polymorphisms on the genome. For example, when the rate of deleterious mutations is higher than the typical selective effect per mutation, polymorphic variants are sufficiently common that neutral diversity no longer depends on the selective effects of individual variants (Good *et al.* 2014). This results in a qualitatively different shape of the neutral site frequency spectrum, which cannot be predicted by standard models of background selection (e.g., Nordborg *et al.* 1996), but is better described by an *infinitesimal* framework, which is parametrized by the variance in population fitness (or the variance per unit map length), rather than the variance of fitness effects at individual loci (Neher *et al.* 2013).

In this paper, we focus on a similar scenario of a complex trait determined by a large number of weakly selected variants uniformly spread across the genome [see also Sachdeva and Barton (2018)]. The main goal is to explore how the interplay between linkage, polygenic selection, and genetic drift shapes the long-term introgression probability of different fragments of such a genome when it is introduced into a genetically homogeneous recipient population. For simplicity, we consider the introgression of a medium-sized block of map length y_0 rather than the whole genome. This allows us to ignore certain complications, such as multiple crossovers within the selected region, in analytical calculations. It is also representative of a scenario where a genome that is repeatedly back-crossed into a recipient population breaks up into several medium-sized segments that evolve more or less independently, while they are rare in the recipient population. Most importantly, as shown below, the introgression probability of any variant depends primarily on the effect of variants within a characteristic linkage scale, and not on loosely linked regions, as long as the introduced genome has *no net* selective effect in the recipient population. This again suggests that studying the introgression of medium-sized blocks can provide useful insight into the more general case.

The introduced block is assumed to carry a large number L of loci with a range of selective effects, uniformly spaced over

map length y_0 . These loci contribute to an additive trait that is under directional selection in the recipient population. The trait value associated with any portion of the block is then just the sum of effects of all the loci it contains. The effect sizes of loci on the introduced block are drawn from a distribution with mean μ and variance σ^2 ; the effect sizes of variants fixed in the recipient population can be set to zero without loss of generality. For simplicity, we assume that μ and σ^2 do not vary across the introduced block. However, most of our analytical results hold even when this condition is not met, for instance, if there is a statistically significant clustering of large-effect variants in the donor genome.

If the donor population is in linkage equilibrium, such that allelic states of different loci along the introduced block are statistically uncorrelated, then different segments of this block (each with l_1 loci) have random (and typically unequal) contributions, with mean μl_1 and variance $\sigma^2 l_1$. Note that the variance of these contributions, or more generally the variance per unit map length, given by $V_0 = \sigma^2 L / y_0$, is the same irrespective of whether the block contains a few, large-effect loci (small L and large σ^2) or many, small-effect loci (large L and small σ^2). Thus, varying L while keeping V_0 and μL constant tunes the extent to which individual loci interfere with each other within fragments with a given map length and total selective effect. If the number of loci is very large and the distribution of allelic effects is correspondingly narrow—i.e., in the limit $\mu \rightarrow 0$, $\sigma^2 \rightarrow 0$, and $L \rightarrow \infty$, with μL and $\sigma^2 L$ held fixed—this model approaches the *infinitesimal model with linkage* [see Robertson (1977)]. This model is similar to the well-known infinitesimal model (Bulmer 1980; Barton *et al.* 2017), but also accounts for linkage by considering loci on a linear map, rather than unlinked loci.

In this study, we explore the effects that shape the introgression dynamics of different segments of a *particular* introduced block, in contrast to our previous work [see Sachdeva and Barton (2018)], which analyzed statistical averages associated with an *ensemble* of such blocks (all characterized by the same net trait value z_0 and the same genic variance per unit map length V_0). We focus here on the following two questions. First, what is the expected *initial* growth rate of a fragment or locus embedded within the introduced block, under directional selection? Is the growth rate of an individual variant influenced more by its own effect or the effects of linked loci within a characteristic map distance? Second, to what extent do expected initial growth rates predict the ultimate fixation probability of different fragments in a finite population?

The distinction between initial and long-term introgression is related to the qualitatively different role played by recombination during these two phases. Recombination primarily breaks the introduced genome into smaller fragments during the initial phase, but as introgressing fragments become more abundant in the population, recombination can also bring together various successful fragments, countering Hill–Robertson interference. Thus, over longer time scales, recombination can uncover fine-scale variation within successful fragments (provided these have not fixed), allowing selection to discriminate between tightly linked variants. In a very

large population, this would allow selection to fix all positively selected variants and eliminate all deleterious variants within introgressing fragments that have survived the initial phase. In a smaller population, only part of this fine-scale discrimination can be achieved, which constrains the net response to selection. An important focus of our work is to explore how the ultimate fate of variants in a finite population is determined by the interplay between the early dynamics characterized by selection on relatively large genomic fragments and long-term dynamics governed by selection on fine-scale variation within these. This clarifies the conditions under which individual variants establish with high probability (and hence are likely to be replicated across populations that receive the same genome). Identifying the distribution of effects associated with highly replicable loci, even within this simple model, can shed light on the limitations of fine mapping the causal loci that underlie polygenic adaptation.

The expected growth rates of genomic segments during the initial phase of introgression are relatively easy to calculate, since introgressed fragments are rare, and hence unlikely to encounter each other during mating. Thus, any genome carries at most one fragment of the introduced block, which leads to explicit analytical expressions for the initial growth rate of fragments and for the probability of survival of at least some part of the introduced block. We then study long-term introgression by simulating individuals, and illustrate the correlation between initial growth rates and long-term fixation probability by means of a representative example. We then describe how introgression probabilities at individual loci depend on the variance of fitness effects, the total trait value associated with the introduced block, the size of the recipient population, and the extent of initial stochasticity (which is tuned in this model by introducing single vs. multiple identical copies of the block). The extent to which selection and recombination can pick out and amplify individual favorable variants is a basic determinant of the net response to selection, when selected variants are tightly linked. We address the question of selection limits, and their dependence on population size and linkage, at the end of the paper.

Methods

Consider a situation where N_0 identical copies of a block of genome are introduced into a very large population of

diploids at $t = 0$. Any diploid individual is assumed to carry at most one copy of the introduced block. The block has L loci, which are assumed to be evenly spaced on the genetic map, with recombination rate c between adjacent loci. The extension to unequal rates of recombination between loci is straightforward. The additive effect of the k th locus is denoted by s_k . The fitness of the block is multiplicative across loci and is given by $\exp[\sum_{k=1}^L s_k]$

Initial introgression into a large population

To derive analytical results for the initial dynamics of block fragments, we assume that the introduced block, with map length $y_0 = cL$, is short enough that multiple crossovers can be neglected. When the introduced block and its descendant fragments are sufficiently rare in the population (as expected during the initial phase), then the probability of recombination between fragments is negligible and any individual inherits at most one introgressed fragment. Therefore, we need only consider single fragments spanning loci $i \cdots j$, with selective effect $S_{i,j} = \sum_{k=i}^j s_k$, which break up by recombination with rate $c(j-i)$. For weak selection and recombination ($s_k, c \ll 1$), the expected numbers $N_{i,j}(t)$ of different fragments change approximately continuously through time, according to a set of linear equations:

$$\begin{aligned} \frac{\partial N_{i,j}}{\partial t} &= [S_{i,j} - c(j-i)]N_{i,j} + c \left[\sum_{m=1}^{i-1} N_{m,j} + \sum_{m=j+1}^L N_{i,m} \right], \\ N_{i,j} &= N_0 \delta_{i,1} \delta_{j,L} \text{ at } t = 0 \end{aligned} \quad (1)$$

The first term is the rate at which a fragment spreads intact (without being split), and is positive when the fragment is amplified by selection (at a rate proportional to its fitness $S_{i,j}$) faster than it is split by recombination (rate proportional to map length $c(j-i)$). The second and third terms are due to generation of the fragment by recombination from larger blocks (in which it is embedded).

Equation 1 can be solved explicitly by taking the Laplace Transform, solving for the expected number of single recombinants $N_{1,i}$ and $N_{i,L}$, and then solving for expected numbers of smaller blocks. The solution is a sum of exponentially decaying (or increasing) components:

$$\begin{aligned} N_{i,j}(t) &= N_0 \sum_{k=j}^L \sum_{m=1}^i A_{j+1,k} B_{k+1,L} \hat{A}_{1,m-1} \hat{B}_{m,i-1} \exp \left[t \left(c + \sum_{v=m}^k (s_v - c) \right) \right] \\ A_{m,k} &= \frac{c}{g_{m,k}} \left(1 + \frac{c}{g_{m+1,k}} \right) \cdots \left(1 + \frac{c}{g_{k,k}} \right) \\ B_{m,k} &= \frac{-c}{g_{m,k}} \left(1 - \frac{c}{g_{m,k-1}} \right) \cdots \left(1 - \frac{c}{g_{m,m+1}} \right) \left(1 - \frac{c}{g_{m,m}} \right) \end{aligned} \quad (2)$$

where $g_{m,k} = \sum_{v=m}^k (s_v - c)$; $\hat{A}$ is obtained by replacing c by $-c$ in A and $\hat{B}$ by replacing $-c$ by $+c$ in B . Also, $A_{m,k}$, $B_{m,k}$, $\hat{A}_{m,k}$, and $\hat{B}_{m,k}$ are all equal to 1 for $m > k$.

The sum is over all blocks $[m, k]$ (with $m \leq i \leq j \leq k$) that contain the fragment $[i, j]$, including $[i, j]$ itself. For large t , the exponential term with the largest growth rate will dominate the dynamics of the block. Thus, at long times, any block grows at the same rate as the fastest growing parent block from which it can be generated. Note that if the intrinsic growth rate $S_{i,j} - c(j - i)$ of the block is larger than the growth rate of any parent block containing it, then the long-term growth rate of the block is just $S_{i,j} - c(j - i)$. An important consequence of Equation 2 is that deleterious fragments may also spread through the population, if generated at a constant rate by a beneficial, exponentially growing parent block. In fact in this case, the deleterious subblock will grow at the same rate as the beneficial parent block, as is evident from Equation 2. Thus, each such exponentially increasing block generates various descendant subblocks of varying fitness, giving rise to a family of blocks, akin to a quasi-species (Eigen *et al.* 1988).

Figure 1A depicts the long-term growth rates of all possible 55 fragments of a block with 10 loci, while Figure 1B shows the expected dynamics of these fragments. Note the spread of different families of subblocks at different rates, where each such family (represented by lines of one color in Figure 1B) consists of subblocks that are all contained within a particular, fast-growing parent block. After an initial transient phase, all subblocks within a family increase at the rate of growth of this parent block. In the following, we approximate the growth rate of any fragment by the rate associated with the fastest-growing term in Equation 2, *i.e.*, by the intrinsic growth rate of the fastest-growing parent block that contains this fragment. We refer to this as the expected growth rate of the fragment, since it describes the dynamics of the expected number of copies of the fragment in the population, averaging over all possible stochastic histories, including those in which the fragment is lost.

The key assumption underlying Equation 1 is that descendants of the introduced fragment form a negligible fraction of the population, and thus never encounter each other during mating. Then, the spread of introgressed genetic material through the population can be formulated as a *branching process* (BP) that treats the (forward in time) lineages of different descendants as being independent of each other, but having a common dependence on $\{s_1, s_2, \dots, s_L\}$ (Sachdeva and Barton 2018). The BP framework is very powerful and can, in principle, yield the full distribution of introgressing fragments as a function of time. Here, we write down the equation for the probability $P_{1,L}$ that at least *some* part of a block $\{1, L\}$ spanning loci $i = 1, \dots, L$ survives at long times, given that it is initially introduced as a single copy. To work with dimensionless quantities, we scale both selection and survival probability relative to the map length: $\tilde{P}_{i,j} = P_{i,j}/Lr$, and $\tilde{S}_{i,j} = S_{i,j}/Lr$.

The long-term scaled survival probability of a block $\{1, L\}$ satisfies:

$$0 = \frac{1}{L} \left(\sum_{k=1}^{L-1} \tilde{P}_{1,k} + \sum_{k=2}^L \tilde{P}_{k,L} \right) + (\tilde{S}_{1,L} - 1) \tilde{P}_{1,L} - \frac{1}{2} \tilde{P}_{1,L}^2 \quad (3)$$

This equation is of the form $0 = f + a\tilde{P}_{1,L} - \tilde{P}_{1,L}^2/2$, where the first term f is due to the survival of smaller subblocks and the coefficient a of the second term is the intrinsic growth rate of the full block. This equation is analogous to equation 3 in Barton (1995), but has an additional driving term f . Solving the quadratic, we have:

$$\tilde{P}_{1,L} = g \left[\frac{1}{L} \left(\sum_{k=1}^{L-1} \tilde{P}_{1,k} + \sum_{k=2}^L \tilde{P}_{k,L} \right), (\tilde{S}_{1,L} - 1) \right] \quad (4)$$

where $g[f, a] = a + \sqrt{a^2 + 2f}$; $g[0, a] = \max[0, 2a]$. This is easily evaluated, because $\tilde{P}$ for larger blocks depends only on that for smaller blocks. Note that blocks with a negative net rate of increase ($a < 0$) can survive (in part), if they contain smaller blocks with positive growth rates, such that $f > 0$. If the block $\{1, L\}$ is introduced in N_0 copies, then the survival probability of each such block is roughly independent of other blocks, as long as N_0 is much smaller than the size of the recipient population. Then, the overall survival probability is just $1 - (1 - P_{1,L})^{N_0}$.

Long-term introgression

Equation 1 is a valid description of the dynamics of genomic fragments only while any individual carries at most one such fragment. As shown in Sachdeva and Barton (2018), this is true over a time scale that scales weakly ($\sim \log(N)$) with the size N of the recipient population. Beyond this initial time scale, mating between individuals carrying introgressed material becomes more frequent and genomes carrying multiple fragments of the introduced block emerge. This phase is thus characterized by competition between genomes bearing different mosaics of the fragments that have established and proliferated in the initial phase, and is no longer described by Equation 1.

To study long-term introgression, we simulate populations with N diploid individuals. The simulations are initialized by assuming that $2p_0N$ haploid genomes in the population carry identical introduced blocks, while all other genomes carry identical “native” blocks. For simplicity, we assume that each of the $2p_0N$ introduced blocks is present in a different individual. Each block has L loci, with rate of recombination c between adjacent loci. The allelic effect of each native locus is equal to zero, while effects of the introduced loci are drawn from a distribution with variance σ^2 and mean μ . The trait value S associated with any individual is then just the sum of effects of all introduced variants in its diploid genome; its fitness is e^S .

In each generation, $2N$ parents are drawn by randomly sampling individuals in proportion to their fitness. Each parent produces a gamete with recombination between parental

Figure 1 (A) Expected (asymptotic) growth rates of various fragments $[i, j]$ of a block with $L = 10$ loci, with fitness contributions drawn from a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = 0.1/L$, such that the contributions of all loci sum to 0; the rate of recombination between adjacent loci is $c = 0.01$. The two axes of the plot represent the two endpoints of the fragment; each location (i, j) in the upper triangle of the plot represents a particular fragment with endpoints i and $j \geq i$. The color associated with each location encodes the expected growth rate of the fragment (see adjoining color scale). Note that many of the 55 subblocks have the same expected asymptotic growth rate; thus, there are 15 distinct asymptotic growth rates (colors) in (A), of which five are positive. The black curve in the upper panel shows how the trait value varies along the introduced block. The x-axis represents the positions i of the loci on the block and the y-axis represents the cumulative contribution of all loci from $[0, i]$; thus, any segment with a positive slope represents a subblock with a net positive contribution to trait value and vice versa. (B) Expected numbers of each of the 55 fragments contained within the 10-locus block in (A), as a function of time t . Black curves represent fragments with asymptotically negative growth rates; curves in other colors represent fragments with asymptotically positive growth rates, which increase in number at long times. All fragments with the same asymptotic (positive) growth rate are represented by the same color. All curves are calculated by iterating Equation 1. (C) Asymptotic growth rates of various fragments of a block with $L = 2048$ loci and map length $y_0 = 0.4$. The net trait value associated with the block is 0 and variance per unit map length $\sqrt{V_0} = 0.05$ (which implies $\sigma^2 = 0.001/L$). The two axes of the plot represent the map positions y_1 and y_2 of the two endpoints of any fragment $\{y_1, y_2\}$ in the block. Only those fragments that have a positive expected growth rate in the long run are shown. The inset inside the main plot zooms into a particular region of the block (highlighted in green in the upper panel) and depicts fragments that have a positive growth rate within this region.

haplotypes: the number of crossover points is drawn from a Poisson distribution with mean y_0 ; the locations of the crossover points are chosen by uniformly sampling one of the $L - 1$ junctions between loci without replacement. The gametes are then paired to form N individuals of the next generation.

We compute introgression probabilities by performing 200–400 replicate simulations, each initialized by introducing

$2p_0N$ identical copies of the same block at $t = 0$. Thus, replicate populations differ only in the stochastic history of reproduction and recombination. The introgression probability of a particular fragment at time t is calculated as the fraction of replicates in which the fragment is present *either by itself or as part of a larger block*. Since any fragment must either fix or be completely eliminated at long times, the introgression

probability of a fragment in the large t limit, is just its fixation probability. A fragment with fixation probability P would be found in two replicate populations with probability P^2 ; thus, any fragment with high introgression probability is also a replicable fragment. Note that the introgression probability is not the same as the survival probability $P_{1,L}$, which was calculated in Equation 4. The latter refers to the probability that the block $\{1, L\}$, introduced as a single copy at $t = 0$, is not fully lost from a very large population, but has at least one surviving descendant subblock.

Data availability

FORTTRAN 95 codes used to generate the simulated data can be found at: https://git.ist.ac.at/himani.sachdeva/source_codes_replicability_introgression_patterns/snippets.

Results

To gain insight into how the ultimate fixation probability of different parts of the introduced genome is shaped by the early dynamics (as encapsulated by the expected initial growth rates) vs. long-term dynamics (characterized by selection on multiple, linked successful fragments), we first analyze one representative example in detail. This illustrates key features of the introgression process, which hold more generally.

Introgression probabilities of different fragments of the genome: an example

Consider an introduced block with $L = 2048$ loci, uniformly spaced over map length $y_0 = 0.4$. The contributions of different loci are drawn from a normal distribution with mean 0 and variance $\sigma^2 = 0.001/L$, using an iterative scheme that ensures that the contributions sum to $S_{1,L} = 0$ [see also Sachdeva and Barton (2018)]. Figure 1C depicts all fragments of this block that have a positive expected growth rate. Note that these fragments are quite small; longer fragments of a nearly neutral introduced block are split by recombination faster than they are amplified by selection, and hence have negative expected growth rates.

In this example, we simulate a diploid population of size $N = 4000$, with initial frequency $p_0 = 0.005$ of the introduced haplotype, such that there are $2Np_0 = 40$ introduced blocks in the population at $t = 0$, and, hence, 40 approximately independent realizations of the initial introgression process within any one population. The dynamics of early introgression in such a population would thus be close to the prediction of Equation 2. By contrast, with a single introduced block ($p_0 = 1/2N$), all or some of the introduced genome, including beneficial fragments, may be lost from the population while present in low numbers, which causes the dynamics of individual populations to differ markedly from the expected dynamics. The parameter $2Np_0$ thus governs the extent to which the initial dynamics of any individual population is deterministic.

Figure 2, A and B show the introgression probability of different fragments of the introduced genome vs. their expected growth rates (as calculated above) for two different time instants: $t = 600$ (Figure 2A) and $t = 24,000$ (Figure 2B). Each point represents a particular fragment; the color of the point encodes the length of the fragment or, alternatively, the number of equally spaced selected loci that it contains (see accompanying color scale). The introgression probability of any fragment is the fraction of replicates in which the fragment is present, considering only those replicates in which at least some part of the introduced block survives at long times. We only show introgressing fragments with positive expected growth rates, as fragments with negative growth rate fail to introgress in the long run within a large population.

The introgression probabilities of different fragments at $t = 600$ are highly correlated with their expected growth rates (correlation coefficient $r = 0.82$), while the correlation is somewhat weaker ($r = 0.46$) but still nonzero at $t = 24,000$. The weaker correlation at longer times is due to substantial differences in the long-term introgression probabilities of different fragments that have the *same* expected growth rate (note the much higher scatter along the y-axis in Figure 2B relative to Figure 2A). Note that we use the correlation coefficient purely as a descriptive statistic: thus, an r value of 0.8 implies that 64% of the total variance of introgression probabilities can be explained by a linear regression with the expected growth rates. However, this does not provide a complete description, since the introgression probabilities are not normally distributed about the linear prediction.

Figure 2B suggests that the introgression of shorter fragments is more replicable than that of longer fragments within any family of fragments with the same expected growth rate; note the higher introgression probabilities associated with black as compared to blue points within each vertical column of points. To investigate this in more detail, we zoom into a small window of the genome that contains a large number of replicable fragments. Figure 2, C and D show snapshots of this genomic window for 50 replicate populations at $t = 600$ and $t = 24,000$, respectively. Each horizontal band corresponds to one replicate population and the colors along each such band encode the frequencies (in that population) of introgressed variants at different map positions. The red triangles depict selective effects of individual variants.

At long times (Figure 2D corresponding to $t = 24,000$), multiple, disjointed fragments of the introduced genome are fixed in any population, even within this limited genomic region. We focus on the genomic window between the two vertical lines: this segment has the second highest expected growth rate among all the segments of the introduced genome. Note that many of the replicate populations lack a single, deleterious introgressed variant in the middle of this segment. These populations most probably witnessed one (or a few) recombination events that brought together two fragments of this segment onto a single genome *without the*

Figure 2 Introgression probabilities of different *fragments* of the genome shown in Figure 1C (with $S_{1,L} = 0$, $y_0 = 0.4$, $\sqrt{V_0} = 0.05$, and $L = 2048$) in a finite population of size $N = 4000$. (A and B) The probability that a subblock is present (either by itself or as part of a larger block) in a replicate at any frequency, at $t = 600$ (A) and at $t = 24,000$ (B), vs. the expected growth rate of the subblock. Each point represents a subblock; the color represents the length of the subblock (or alternatively the number of equally-spaced loci it contains). Note that at $t = 24,000$, subblocks with the same expected growth rate can have very different introgression probabilities, with longer subblocks within a family being less likely to introgress (as a whole) than shorter subblocks. Uncertainty in introgression probability estimates is $\sim 3\%$ for a probability of 0.8 and $\sim 25\%$ for a probability of 0.05. Subblocks with negative growth rates have long-term introgression probabilities close to zero and are not shown. Expected growth rates are calculated as described in the text. (C and D) Snapshots of 50 replicates that carry at least some introgressed material at $t = 24,000$ (D) and the same replicates at $t = 600$ (C). The x-axis denotes the map position along the genome (for a limited genomic stretch), each horizontal line depicts a different replicate population (numbered 1–50 along the y-axis), and the colored segments within each horizontal line represent fragments of the introduced block present in the population at time t ; the color represents the frequency of the introgressed variant at that genomic position in that population. All replicate populations receive the same introduced genome at $t = 0$; the selective effects of loci on this genome are depicted by the red triangles. All triangles lying below $s = 0$ (the dashed black line) are deleterious. The window between the two vertical lines represents the subblock with the second-highest expected growth rate (among all subblocks contained in the introduced genome). All plots are obtained from individual-based simulations of replicate populations that receive $2Np_0 = 40$ identical copies of the same introduced block at $t = 0$.

deleterious variant in the middle. This new combination of fragments would then have out-competed the original segment, as it is associated with a larger selective advantage than the original segment, while having the same (low) probability of being split by recombination. In a subsequent section, we analyze this process in more detail for a toy example with three loci.

To avoid comparing fragments of different sizes, many of which are overlapping or even fully contained within other fragments, we compare introgression probabilities of different *single-locus variants* embedded within the introduced genome. Figure 3, A and B show introgression probabilities at single loci (which constitute a subset of the fragments shown in Figure 2, A and B) vs. their expected growth rates, for

$t = 600$ and $t = 24,000$. As before, the expected growth rate of a locus is the growth rate of the fastest-growing subblock that contains it. The introgression probability at individual loci shows high correlation with the expected growth rate, at both $t = 600$ (with correlation coefficient $r = 0.74$) and at $t = 24,000$ (with $r = 0.82$). However, at longer times, there is a substantial scatter among loci with the same expected growth rate, with deleterious loci (purple dots) having lower introgression probabilities than beneficial loci (orange dots) within any particular cluster of such loci.

To explore how long-term introgression probabilities at individual loci depend on their own selective effects, we consider three separate windows of the introduced genome (shown in green, red, and blue in the inset of Figure 3C).

Figure 3 Introgression probabilities of different *single-locus variants* embedded in the genome shown in Figure 1C (with $S_{1,L} = 0$, $y_0 = 0.4$, $\sqrt{V_0} = 0.05$, and $L = 2048$), when it is introduced into a finite population of size $N = 4000$. (A and B) The probability that a particular single-locus variant (that was originally part of the introduced block) is present either by itself or as part of a larger block in a replicate (at any frequency), at $t = 600$ (A) and at $t = 24,000$ (B), vs. its *expected growth rate*. Each point represents a locus; the color represents the selective effect of the introduced variant at this locus. Note that at $t = 24,000$, loci with the same growth rate can have very different introgression probabilities, which appear to be negatively correlated with the selective effect of the loci. (C and D) The probability that a particular introduced, single-locus variant is present (either by itself or as part of a larger block) at $t = 24,000$, vs. the *selective effect* associated with the locus for: (C) single-locus variants present in three windows or regions within the introduced block. (D) all single-locus variants contained in the introduced block. The three genomic windows are represented by colored segments in the inset of (C). The inset shows the variation of trait value along a part of the introduced block, with the green, blue, and red segments depicting subblocks that are associated with expected growth rates 0.01078, 0.00987, and 0.00715, respectively. The green, red, and blue points in the main plot represent introgression probabilities of single-locus variants contained in these three segments. Note that the introgression probability is very weakly correlated with selective effect across all loci (D), but decreases as selective effect becomes more negative *within* clusters of loci that have the same expected growth rate (C). All plots are obtained from individual-based simulations of replicate populations that receive $2Np_0 = 40$ identical copies of the same introduced block at $t = 0$. Uncertainty in introgression probability estimates is $\sim 3\%$ for a probability of 0.8 and $\sim 25\%$ for a probability of 0.05.

The main plot in Figure 3C shows the long-term introgression probabilities of single-locus variants contained within each of these three segments vs. their selective effects, with green, red, and blue points depicting loci contained in the corresponding segments shown in the inset. Figure 3C suggests that within loci with the same expected growth rate, long-term introgression probabilities are sensitive to the selective effect of the locus. In particular, the introgression probabilities of deleterious loci decline with their selective disadvantage (note especially the red and blue points).

In the early phases of introgression, tightly linked loci contained within the same exponentially increasing subblock have the same fate (as evident in the limited scatter of introgression probabilities among loci with the same expected

growth rate in Figure 3A). At early times, selection does not “see” individual loci, since recombination has not had sufficient time to separate tightly linked variants. However, at longer times, recombination generates smaller and smaller fragments of successful subblocks, and also reconstructs various combinations of these fragments, some of which would lack deleterious variants present in the original parent block. Such combinations would supplant the original successful block, thus also eliminating some of the deleterious loci contained within this block. Note that this kind of fine-grained separation of tightly linked beneficial and deleterious variants within a block over long time scales can only occur if the block has not already fixed in the population. For instance, the subblock corresponding to the green segment in Figure 3C fixes very rapidly, which does not allow enough time for

embedded deleterious variants to recombine away and get eliminated.

Does the substantial decline in introgression probability with the selective effect of individual loci within limited genomic windows lead to a correspondingly strong dependence on individual effects across larger map distances? Figure 3D shows introgression probabilities (at $t = 24,000$) of single-locus variants across the entire block vs. their selective effects. Even in the long run, single-locus introgression probabilities are only weakly correlated with selective effect across the whole block (correlation coefficient $r = 0.37$ in Figure 3D), in contrast to the much stronger correlation with expected growth rates ($r = 0.82$ in Figure 3B). This is consistent with our explanation that the ultimate fixation probability is shaped by linked selection (or selection on clusters of loci) in the early phases of introgression, which then constrains the extent to which selection can distinguish between individual variants within these clusters in the later phases of introgression. Note further that the most replicable variants are associated with a range of selective effects (see also Figure 4) and can even be deleterious, which makes it misleading, at least within this model, to infer that high replicability implies adaptive significance.

Dependence of introgression probabilities on the distribution of fitness effects

This raises the question: are long-term introgression probabilities of single-locus variants more strongly correlated with their own selective effects when these effects are larger? In other words, as the distribution of fitness effects (DFE) becomes wider (larger σ^2), is there an approach to an effective single-locus regime where selection on the focal locus, rather than linked selection, determines introgression outcomes? In examining this question, we must distinguish between the variance of fitness effects σ^2 , and the genic variance per unit map length $V_0 = \sigma^2/c$, which depends on both σ^2 and the density of selected loci on the genome. The genic variance V_0 influences the extent of linked selection; a fragment of length y (emerging from an introduced genome with $S_{1,L} \sim 0$) has a typical (positive or negative) contribution that scales as $\sqrt{V_0 y}$. Thus, successful fragments have higher expected growth rates (on average) and correspondingly shorter fixation times when V_0 is larger. On the other hand, the per locus variance σ^2 determines the extent to which selection can discriminate between loci of different effects at fine recombination scales at long times. Thus, large σ^2 and large V_0 are expected to influence the correlation between introgression probabilities and selective effect differently. While larger σ^2 implies higher efficacy of selection in separating tightly linked beneficial and deleterious loci, higher V_0 results in higher growth rates of successful blocks, which causes tightly linked loci to fix together before they can be taken apart by recombination and seen individually by selection.

To clarify the roles of σ^2 and V_0 in determining long-term introgression probabilities, we compare two blocks with the same density of selected variants (*i.e.*, with the same number

Figure 4 The distribution of selective effects of single-locus variants that introgress with probability $> 50\%$. The three distributions correspond to three different introduced blocks with map length $y_0 = 0.4$ and $S_{1,L} = 0.0$. The first block (distribution in black) has $L = 2048$ loci with effects drawn from a distribution with variance $\sigma_{DFE}^2 = 0.001/L$, resulting in genic variance per unit map length $V_0 = \sigma_{DFE}^2 L / y_0 = 0.0025$. The second block (blue) has $L = 2048$ loci, with the allelic effect of each locus exactly twice the effect of the corresponding locus in block 1; thus, block 2 has $\sigma_{DFE}^2 = 0.004/L$ and $V_0 = 0.01$. The third block (red) has $L = 512$ loci. The allelic effect of the i th locus in this block is $\sum_{k=1}^4 \gamma_{Ai-k}$, where γ_i is the effect of the i th locus on block 1; thus, block 3 has $\sigma_{DFE}^2 = 0.004/L$ but $V_0 = 0.0025$. Note that loci with negative selective effects are least likely to establish for the third block (red) that has fewer and larger effect loci. All plots are obtained from simulations of replicate populations with $N = 4000$ individuals; $2Np_0 = 40$ copies of the same block are introduced into each replicate population at $t = 0$.

L of selected loci and the same recombination rate c per locus), but with the variance of fitness effects of individual loci on one block being four times the corresponding variance for the other block. It then follows that the genic variance per unit map length V_0 for the first block is also four times the corresponding variance for the second block. We consider all variants with long-term introgression probabilities > 0.5 within each block and plot the distribution of their selective effects (Figure 4). Strikingly, the distribution has a very similar shape in the two cases (black vs. blue curves), but is shifted toward slightly *weaker* selective effects (relative to σ_{DFE}) when the block has a wider DFE and a correspondingly high genic variance V_0 (blue), than when the block has a more narrow DFE and a lower value of V_0 (black). The mean selective effect of such strongly replicable loci is $0.662\sigma_{DFE}$ in the first case (blue curve) and $0.754\sigma_{DFE}$ in the second case (black). Moreover, the fraction of strongly replicable loci that are *deleterious* is slightly higher (0.156 vs. 0.117) for the introduced block with larger σ_{DFE}^2 and V_0 . This suggests that for a given (high) density of selected variants on the genome, a wider DFE results in more extensive hitchhiking of deleterious loci during introgression and, thus, reduced power to pinpoint adaptive loci from patterns of replicability.

We also compare two blocks characterized by the same V_0 , but with different values of σ_{DFE}^2 and L . Since V_0 governs initial introgression, both blocks are expected to show very

similar short-term patterns of introgression. However, a population that receives the block with wider DFE (and lower density of selected variants) should undergo more efficient elimination of deleterious loci in the long run. This is indeed observed in simulations; the distribution of selective effects of strongly replicable variants is shifted toward higher effects (mean selective effect $1.0285\sigma_{DFE}$) for the block with wider DFE (red curve), as compared to the block that has a narrower DFE but the same V_0 (black curve). Further, the fraction of strongly replicable loci that are deleterious is negligible for the red curve, as compared to 0.117 for the black curve. This is consistent with the general expectation that direct selection on loci should affect their introgression outcomes more strongly than linked selection when the contribution of a genomic region is determined by fewer, larger effect loci.

Note that, in the present example, the recipient population is sufficiently large or recombination between neighboring selected variants sufficiently frequent that individual variants can be isolated. More generally, selection and recombination are expected to weed out regions of some typical length l' from within successful blocks, where l' would depend on population size N and the growth rate of the successful block. The selective effects associated with replicable loci would then depend on the DFE of segments of length l' , rather than the DFE of individual variants. Understanding how l' depends on population size N and V_0 is nontrivial, and we do not attempt to address this question here.

Dependence of introgression probabilities on population size

The size of the recipient population is important in determining the probability of stochastic loss and hence the extent to which the spread of an advantageous allele is predictable. The fixation probability of an unlinked allele with selective effect s , introduced into a population of size N at a frequency p_0 , was derived by Kimura (1957). This fixation probability approaches a limit that is independent of N for $Ns \gg 1$: when a single copy of the allele is introduced ($p_0 = 1/2N$), the asymptotic fixation probability is $2s$ for $s \ll 1$. Thus, most beneficial alleles are lost in a single population. On the other hand, when a beneficial allele is introduced at a fixed frequency p_0 , its asymptotic fixation probability is 1 for large N .

To what extent are these single-locus predictions relevant to a scenario where the introduced genome consists of multiple, linked, beneficial and deleterious loci? We investigate the dependence of long-run introgression probabilities on N by comparing introgression of the same genome into populations of different sizes for two scenarios: one where a single copy of the block is introduced into the population ($p_0 = 1/2N$, see Figure 5A) and the second where multiple copies of the same block are introduced at a fixed initial frequency (p_0 independent of N , Figure 5B). Figure 5, A and B show the introgression probabilities (at $t = 24,000$) of single-locus variants vs. their expected growth rates in the two cases, for various population sizes. Solid lines show the

predictions of a modified version of Kimura's formula, where the introgression probability P_i of the introduced variant at locus i is assumed to be:

$$P_i = \frac{(1 - e^{-4p_0 s_{eff,i} N})}{(1 - e^{-4s_{eff,i} N})} \quad (5)$$

where $s_{eff,i}$ is the expected growth rate at locus i , rather than the selective effect of the locus.

Interestingly, Equation 5 can explain 40–70% of the variance of introgression probabilities (see also caption of Figure 5), suggesting that clusters of loci identified as having the same expected initial growth rate are, in some sense, natural linkage units into which the genome can be decomposed. To a first approximation, each such unit can be considered as unlinked to others. However, note that there is considerable scatter of the introgression probabilities (about the prediction of Equation 5), especially in large populations. As we argue above, the differential introgression of tightly linked loci within a genomic fragment depends on its time to fixation relative to the time scale at which fine-scale recombination within the fragment becomes effective, which in turn depends on population size N .

We now examine this dependence on N in detail for an introduced block with $L = 3$ loci. Assume that the selective effects at the three loci satisfy $s_1 > 0$, $s_2 < 0$, and $s_3 > 0$, such that the expected growth rate of the full block is higher than the growth rates of any of its single- or double-locus fragments. However, if recombination were to bring together introgressing variants at the first and third locus without the deleterious ($s_2 < 0$) variant in the middle, then this double recombinant would be fitter than the introduced parent block and can supplant the latter.

An approximate expression for the fixation probability of this “101” recombinant (carrying introgressed variants at the first and third loci) can be obtained using the semideterministic approach of Hartfield and Otto (2011) (details in Appendix). The 101 recombinant is mostly generated by recombination between the full parent block, and a descendant block carrying an introgressed variant at just the first or the third locus. The frequency $P_{101}^{emerge}(t)$ of such recombination events (that lead to the emergence of the 101 recombinant in generation t) can be computed by assuming deterministic dynamics for genotypic frequencies. We then use a BP framework to calculate the probability $P_{fix}(t)$ that a 101 recombinant, emerging at time t , fixes in the population in the long run. Note that the time at which the recombinant emerges determines its fitness relative to the recipient population, since the mean fitness of the recipient population itself changes as the introgressed genome spreads through it.

The probability that none of the $NP_{101}^{emerge}(t)$ recombinants (generated in generation t) fix is then: $[1 - P_{fix}(t)]^{NP_{101}^{emerge}(t)}$, which is approximately $\sim \exp[-NP_{101}^{emerge}(t)P_{fix}(t)]$. Further, if double recombinants emerge and establish rarely, then the net probability that no double recombinant fixes in a

Figure 5 Introgression probabilities of single-locus variants vs. their expected growth rates, for recipient populations of different sizes when (A) a single copy of the introduced block enters the population at $t = 0$. (B) Multiple ($2p_0N$ with $p_0 = 0.005$) copies of the block are introduced at $t = 0$. Each point represents the introgression probability of a single-locus variant, as obtained from individual-based simulations for $N = 100$ (red), $N = 1000$ (green), and $N = 4000$ (blue). Note that the blue points in (B) are the same as those shown in Figure 3B. Lines represent predictions of Equation 5. These predictions explain 52.4, 63.4, and 67.8% of the variance of introgression probabilities for $N = 100$, 1000, and 4000, respectively, in (A), and 52.4, 65.6, and 42.6% of the corresponding variance in (B). The same block (with $S_{1,L} = 0$, $y_0 = 0.4$, $\sqrt{V_0} = 0.05$, and $L = 2048$) is introduced in each simulation. Note the greater scatter in introgression probabilities (among loci with the same expected growth rate) for larger N .

population of size N is just $\sim \exp[-N \int_{t=0}^{\infty} dt P_{101}^{emerge}(t) P_{fix}(t)]$. This implies that the double recombinant fixes with probability $1 - \exp[-N \int_{t=0}^{\infty} dt P_{101}^{emerge}(t) P_{fix}(t)]$, which increases with the size of the recipient population and approaches 1 for large N , for a fixed p_0 . The fixation probability of the 101 recombinant in individual-based simulations matches the semideterministic prediction quite well (see Appendix).

Thus, large populations typically have a larger number of recombinants that combine various fit fragments of a successful subblock (before it fixes). This makes it increasingly probable that at least one of these recombinants establishes and supplants the parent subblock in the population, thus eliminating deleterious variants embedded in the parent subblock.

Role of net selective effect of introduced block

So far, we have considered the case where the introduced block is neutral with respect to the recipient population ($S_{1,L} = 0$). However, when the introduced block is deleterious as a whole ($S_{1,L} < 0$), any of its constituent fragments (including those with a high, positive expected growth rate in the long run) experience an initial selective disadvantage. Many of the $2p_0N$ copies of the introduced block are lost from the population before the beneficial fragment is separated from the deleterious background and amplified by selection, which suggests that the actual introgression probability in this case would be lower than that predicted by Equation 5.

Figure 6 compares introgression probabilities obtained from simulations with the predictions of Equation 5 for a neutral (Figure 6A), beneficial (Figure 6B), and deleterious (Figure 6C) block. As expected, Equation 5 broadly captures the variation of single-locus introgression probabilities along the neutral block, but systematically overestimates (or underestimates) these when the block is deleterious (or beneficial).

Note that introgression probabilities of single-locus variants are highly correlated with the expected growth rate, even for blocks with $S_{1,L} \neq 0$ (where the correlation coefficient is $r = 0.8$ for the beneficial block shown in Figure 6B and somewhat lower at $r = 0.58$ for the deleterious block in Figure 6C). However, Equation 5 no longer accurately describes the relationship between the two quantities.

More generally, since the rate of increase of a subblock itself changes over time, before approaching a constant (asymptotic) value (see Figure 1B), a more accurate predictive approach needs to explicitly consider the establishment of any fragment under a time-varying effective selection coefficient (e.g., see Uecker and Hermisson 2011) instead of a constant s_{eff} (as used in Equation 5). An interesting question is whether the effect of total fitness $S_{1,L}$ of the introduced block can be partially captured by a constant factor (corresponding to a genome-wide barrier to gene flow) that attenuates (or inflates) the introgression probabilities of all fragments of the block by the same amount. Explicit expressions for barrier strength have been derived by Bengtsson (1985) and Barton and Bengtsson (1986) for the simpler case of a neutral locus embedded in a genome with all deleterious loci.

Dependence of long-term selection response on population size and linkage

Directional selection on standing genetic variation results in a net phenotypic advance, which depends on population size N . When selection acts on a very large number of polymorphic *unlinked* loci (as in the standard infinitesimal model), additive genetic variance declines at a rate $\sim 1/2N$ per generation due to drift. This constrains the net advance, which scales *linearly* with N (Robertson 1960). The situation is more complex for selection on linked loci (Robertson 1970, 1977). Tightly linked variants would fix together rapidly in a small population, but be broken apart and seen by selection in a

Figure 6 Introgression probability of the single-locus variant at locus i on the introduced block vs. i , as obtained from individual-based simulations (red) and as predicted by Equation 5 (green), for an introduced block that is (A) neutral with respect to the recipient population ($S_{1,L} = 0$), (B) beneficial ($S_{1,L} = 0.05$), or (C) deleterious ($S_{1,L} = -0.05$). Each of the three blocks has map length $y_0 = 0.4$ and $L = 2048$ loci, with selective effects drawn from a distribution with mean $\mu = S_{1,L}/L$ and variance $\sigma^2 = 0.001/L$ (subject to the constraint that the effects of all loci add up to $S_{1,L}$). In each set of simulations, $2Np_0 = 40$ identical copies of the block are introduced into each replicate population of size $N = 4000$. For ease of visualization, only half of the genome is shown. Note that Equation 5 broadly captures variation of introgression probability along the introduced block for the $S_{1,L} = 0$ block (explaining 42.6% of the total variance in probabilities), but significantly underestimates or overestimates probabilities when $S_{1,L} \neq 0$.

larger population. Thus, in this case, increasing population size tunes drift as well as the extent of hitchhiking.

We investigate the scaling behavior of net advance for different degrees of linkage by simulating a particular set of $L = 2048$ loci, with net trait value $S_{1,L} = 0$, spread across blocks of different map lengths y_0 . Multiple (*i.e.*, $2p_0N$) copies of the block are introduced into a population of size N at $t = 0$. We then measure the net advance under selection (defined as the average trait value in the limit $t \rightarrow \infty$) in a population and average over replicate populations. To assess the effect of linkage on net advance, we do simulations where the same $L = 2048$ loci are uniformly spaced across blocks with different y_0 ; the allelic effects of the loci are scaled with $\sqrt{y_0}$ such that the genic variance per unit map length ($V_0 = \sigma^2 L / y_0$) and the trait value relative to the total genic variance ($S_{1,L} / \sqrt{V_0 y_0}$) is the same for blocks of different map lengths. Figure 7 shows the average net advance vs. population size N for blocks of different lengths. The net advance is normalized by the probability of survival of at least some part of the introduced block.

Figure 7 shows that the net advance scales sublinearly with N for the shortest blocks (tightly linked loci), but approaches a linear dependence with increasing block length. Based on the above arguments, efficient fine-scale separation of beneficial and deleterious loci should also occur within short blocks in a sufficiently large population, resulting in asymptotically linear scaling of net advance with N . However, this is difficult to confirm in simulations. More importantly, this limit might not be realized in typical scenarios involving populations of a few thousand (as in Figure 7), when introduced variants are tightly linked. Thus, sublinear scaling of net advance with the size of the recipient population may be commonly observed.

Discussion

Selection on linkage blocks

The nature of genetic change (substitutions vs. minor shifts in allele frequency) at individual loci during polygenic

adaptation and the resultant genomic signatures of such adaptation have attracted much interest recently (Pritchard *et al.* 2010). A key challenge is to disentangle the effect of selection on the allele frequency dynamics of the selected locus from the effect of linked selection due to nearby loci. Our analysis shows that these two kinds of selection act over different time scales and together shape the fixation probability of different variants on a single genome introduced into a genetically homogeneous population. We identify sets of loci that share the same short-term fate, by decomposing the genome into contiguous blocks that grow faster (under directional selection) than any parent block containing them. Each such block acts as a linkage unit: the initial introgression of loci within a block is governed by the rate of spread of the block (rather than the selective effects of individual loci) and is largely independent of other such blocks.

While the expected growth rates of linkage blocks provide a useful first estimate for the variation of introgression probability along the genome, a more refined estimate needs to account for several other factors, such as the fine-scale variation within successful fragments, the effects of linkage between multiple successful fragments, and the effect of the background genome, which could attenuate or inflate the effective number of blocks introduced into the population.

The idea that a recombining genome under selection can be viewed as a collection of unlinked, effectively asexual segments or linkage blocks is quite powerful, and allows for the calculation of neutral diversity using theoretical results for asexual populations (Neher *et al.* 2013; Good *et al.* 2014; Weissman and Hallatschek 2014). In our formulation, the intrinsic growth rate of a fragment of map length y and selective effect S_y is proportional to $S_y - y$, which is of the order of $\sim \sqrt{V_0 y} - y$, if the full introduced block is neutral ($S_{1,L} = 0$). This is the maximum for a characteristic map length $y_* \propto V_0$ or, alternatively, for a characteristic number of loci given by $L_* \sim \sigma^2 / c^2$, where σ^2 is the variance of selective effects of individual variants and c the rate of recombination between adjacent selected variants. Thus, on average, blocks of map length y_* should grow faster than any parent block

Figure 7 The net advance relative to $\sqrt{V_0 y_0}$, vs. size N of the recipient population, for introduced blocks of different lengths y_0 . A given block consists of L uniformly spaced loci; the allelic effects are drawn from a Gaussian distribution with mean 0 and variance $\sigma^2 = 0.4/L$ (for the block of length $y_0 = 0.05$), such that the effects across all loci sum to 0. For the other (longer) blocks, we scale the allelic effects with $\sqrt{y_0}$, so that the genic variance per unit map length is the same for blocks of different map lengths. All points are obtained from individual-based simulations. The net advance scales sublinearly with N for small N , but approaches a linear dependence for larger N ; the approach to $\sim N$ behavior is faster for longer blocks.

containing them. This heuristic derivation for the length of linkage blocks is consistent with the expectations of Neher *et al.* (2013) (see Equation 5 in that study), modulo a logarithmic correction arising from the dependence of fitness variance on population size. However, unlike earlier work, our analysis does not just yield an *average* linkage scale, but explicitly decomposes a specific genome into linkage blocks of different sizes by accounting for local variations of selective effect along the genome.

Selection on individual loci and net advance under selection

While the expected growth rate of linkage blocks shows a high correlation with the long-term fixation probability of individual variants within these (Figure 3B) and can capture the coarse-grained variation of introgression probabilities along the introduced block reasonably well (see Figure 6A), there can be considerable fine-scale variation of introgression probabilities of individual variants within each such linkage unit. In fact, this fine-scale variation often reflects differences in selective effect of variants *within* a successful block (Figure 3C). As we illustrate by means of a toy example, the efficacy of selection at fine scales depends on the rate at which a fit block fixes in the population, relative to the rate at which recombination can generate and bring together fragments of the block into even fitter combinations that lack some of the deleterious variants present in the original block. In larger populations, the latter rate is larger, resulting in finer-scale sifting of deleterious variants with increasing population size at long time scales.

The extent to which selection can eliminate individual deleterious variants embedded within a favorable background

determines the net response under selection. Robertson (1970) derived approximate expressions for the net advance due to selection acting on standing variation contributed by linked loci of equal effects (spread over map length l) and found that the net advance in a population of size N is significantly reduced relative to the advance under free recombination, while l is less than the selection intensity. Importantly, for tight linkage, the net advance scales sublinearly with N for small N , but approaches a linear dependence for larger N . Our results are broadly consistent; the net advance increases with increasing map length. Further, it scales approximately linearly with N for the longest block (weakest linkage), but shows a much weaker (sublinear) dependence on N for tighter linkage between loci (Figure 7).

Selective effects associated with highly replicable loci

Our analysis shows that highly replicable substitutions may be associated with a wide range of selective effects, and can even be deleterious if the density of selected variants on the genome is high (see Figure 3D or Figure 4). The fraction of deleterious substitutions during adaptation is lower when the trait value is determined by fewer, widely-spaced loci of stronger effect. In this case, a typical linkage block contains fewer ($\propto \sigma^2/c^2$) loci; thus, the selective effect of a locus makes a larger proportional contribution to, and hence shows a stronger correlation with, its expected growth rate. Then, deleterious loci are less likely to be present in an initially successful fragment and to hitchhike during the initial phase of introgression. Moreover, even those deleterious variants that attain high frequencies have a greater chance of being eliminated in the long run (once fine-scale recombination starts isolating individual loci), due to stronger selection against individual variants. However, note that a wider DFE actually results in a higher fraction of deleterious substitutions (due to stronger hitchhiking), unless the density of selected variants is correspondingly low (Figure 4).

Interestingly, differences in replicability among *tightly linked* variants are informative about the effect sizes of these variants; missing or polymorphic variants (that decline in frequency) within genomic regions with high overall introgression probability are typically associated with deleterious effects. By contrast, differences in introgression probabilities of loosely linked loci (that belong to different linkage blocks) reflect how the strength of linked selection varies along the genome and cannot be explained by differences in selective effects of individual loci (contrast Figure 3, C and D). In general, our study underlines the need for a cautious approach toward interpreting patterns of introgression or adaptation in replicates, and suggests that it is more difficult to ascribe adaptive significance to individual substitutions than to clusters of such substitutions. While genomic regions with multiple replicated substitutions (which correspond to successful blocks in our model) are typically important for adaptation, individual substitutions within these may be nearly neutral (even if they occur across replicates, see Figure 3D).

Interpreting genomic islands of divergence

Genomic regions where the native haplotype persists (*i.e.*, for which the introduced haplotypes fail to introgress) would appear as islands of divergence between the recipient and source population (Strasburg *et al.* 2012). In our model, such regions are associated with introduced fragments that have a *negative* expected growth rate in the recipient population. Note that such fragments need not have deleterious effects; negative growth rates can also be associated with moderately beneficial fragments that are nevertheless broken up faster by recombination than they are amplified by selection. Such fragments may be completely lost from the population, especially when introduced at a low frequency into the recipient population. Thus, in this model at least, the hybridization history (steady gene flow *vs.* sporadic migration events) can be an important factor determining whether extended regions of differentiation are necessarily correlated with barriers to gene flow.

Replicability during adaptation to a phenotypic optimum

Our model considers an additive trait under directional selection, and thus ignores epistasis and genetic redundancy, which can strongly influence replicability (Yeaman 2015). A common scenario is one where the additive trait is under stabilizing selection toward a different optimum in the recipient population. If introduced haplotypes are far from the fitness optimum, then they effectively experience directional selection, causing adaptive fragments within these to be amplified, much as in the present model. In the long run, closer to the optimum, the redundancy of different genomic fragments that make similar contributions to the phenotype may play an important role. Understanding how the initial spread of adaptive fragments constrains long-term introgression patterns along the genome in this case is an interesting direction for future work.

Genetically heterogeneous donor and recipient populations

Our model assumes that the recipient population is genetically homogeneous and that the initial hybridization event involves introduction of multiple copies of the *same* genomic block into the recipient population. Further, all replicate populations receive *identical* genomes from the donor population. Thus, these assumptions effectively ignore variation within the source and recipient population, and just focus on fixed differences between the two. Most theoretical work on introgression and barriers to gene flow makes similar assumptions (Barton and Bengtsson 1986; Uecker *et al.* 2015). However, genetic heterogeneity within the donor and recipient populations is likely to qualitatively impact patterns of introgression and the extent to which these are convergent across replicate populations, and is thus a natural extension to consider.

The present analysis provides some intuition in a scenario where a few nonidentical haplotypes are introduced into a genetically homogeneous recipient population: genomic fragments that have high expected growth rates and are present in multiple copies among the introduced haplotypes

are expected to introgress with high probability. However, genetic variation within the donor and recipient populations can have other important consequences. Small recipient populations may exhibit significant inbreeding depression due to segregation of deleterious recessive alleles. Alleles from a diverged population then have a heterotic advantage, even if they are deleterious within the recipient population. In fact, heterosis has been suggested as a possible reason for the persistence of Neanderthal-derived variants in modern human populations (Harris and Nielsen 2016).

Efficacy of selection

We have considered the contribution that a single introduced genome makes to a homogeneous population. A natural extension is to consider the efficacy of selection within a heterogeneous population. Initially, there is a set of genomes, each carrying alleles at very many loci that have a distribution of effects, both positive and negative. We can focus on one of these genomes and ask what contribution it makes to the final selection response; the expected response of the whole population is the sum of the expected contributions of each initial genome. The key differences from the analysis presented here are that a genome finds itself competing against other genomes that are themselves improving under selection, and that it finds itself associated with other genomes that have random effects on the trait. A first step might be to simply treat the rest of the population as having a mean and variance determined by classical quantitative genetics, and ask whether an analysis focused on a single genome in this variable background can give a good approximation.

We have contrasted two regimes: one where the fate of an allele depends on its own selective effect and another where selection acts on fragments of genome whose effects on fitness depend on very many loci. This contrast is closely related to the more practical question of when we can detect individual causal alleles, by genetic mapping or an association study, or instead can only attribute genetic variance to regions of the genome. An important question for the future is to find ways to determine when it is appropriate to deal with discrete loci rather than take a statistical approach; or better, how to combine these two representations of genetic variation to understand real genetic architectures.

Acknowledgments

We thank Christelle Fraisse for useful comments on the manuscript.

Literature Cited

- Barton, N. H., 1995 Linkage and the limits to natural selection. *Genetics* 140: 821–841. <http://www.genetics.org/content/140/2/821>
- Barton, N. H., and B. O. Bengtsson, 1986 The barrier to genetic exchange between hybridizing populations. *Heredity* 57: 357–376. <https://doi.org/10.1038/hdy.1986.135>

- Barton, N. H., A. M. Etheridge, and A. Véber, 2017 The infinitesimal model: definition, derivation, and implications. *Theor. Popul. Biol.* 118: 50–73. <https://doi.org/10.1016/j.tpb.2017.06.001>
- Bengtsson, B. O., 1985 The flow of genes through a genetic barrier, pp. 31–42 in *Evolution Essays in Honour of John Maynard Smith*, edited by J. J. Greenwood, P. H. Harvey, and M. Slatkin. Cambridge University Press, Cambridge, UK.
- Bulmer, M. G., 1980 *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.
- Burke, M. K., J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose *et al.*, 2010 Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* 467: 587–590. <https://doi.org/10.1038/nature09352>
- Chan, Y. F., F. C. Jones, E. McConnell, J. Bryk, L. Bünger *et al.*, 2012 Parallel selection mapping using artificially selected mice reveals body weight control loci. *Curr. Biol.* 22: 794–800. <https://doi.org/10.1016/j.cub.2012.03.011>
- Conte, G. L., M. E. Arnegard, C. L. Peichel, and D. Schluter, 2012 The probability of genetic parallelism and convergence in natural populations. *Proc. Biol. Sci.* 279: 5039–5047. <https://doi.org/10.1098/rspb.2012.2146>
- Eigen, M., J. McCaskill, and P. Schuster, 1988 Molecular quasi-species. *J. Phys. Chem.* 92: 6881–6891. <https://doi.org/10.1021/j100335a010>
- Elyashiv, E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker *et al.*, 2016 A genomic map of the effects of linked selection in *Drosophila*. *PLoS Genet.* 12: e1006130. <https://doi.org/10.1371/journal.pgen.1006130>
- Good, B. H., A. M. Walczak, R. A. Neher, and M. M. Desai, 2014 Genetic diversity in the interference selection limit. *PLoS Genet.* 10: e1004222. <https://doi.org/10.1371/journal.pgen.1004222>
- Harris, K., and R. Nielsen, 2016 The genetic cost of Neanderthal introgression. *Genetics* 203: 881–891. <https://doi.org/10.1534/genetics.116.186890>
- Hartfield, M., and S. P. Otto, 2011 Recombination and hitchhiking of deleterious alleles. *Evolution* 65: 2421–2434. <https://doi.org/10.1111/j.1558-5646.2011.01311.x>
- Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335–2352. <https://doi.org/10.1534/genetics.104.036947>
- Kimura, M., 1957 Some problems of stochastic processes in genetics. *Ann. Math. Stat.* 28: 882–901. <https://doi.org/10.1214/aoms/1177706791>
- Lässig, M., V. Mustonen, and A. M. Walczak, 2017 Predicting evolution. *Nat. Ecol. Evol.* 1: 77. <https://doi.org/10.1038/s41559-017-0077>
- Manceau, M., V. S. Domingues, C. R. Linnen, E. B. Rosenblum, and H. E. Hoekstra, 2010 Convergence in pigmentation at multiple levels: mutations, genes and function. *Philos. Trans. R. Soc. Lond. B* 365: 2439–2450. <https://doi.org/10.1098/rstb.2010.0104>
- Neher, R. A., T. A. Kessinger, and B. I. Shraiman, 2013 Coalescence and genetic diversity in sexual populations under selection. *Proc. Natl. Acad. Sci. USA* 110: 15836–15841. <https://doi.org/10.1073/pnas.1309697110>
- Nordborg, M., B. Charlesworth, and D. Charlesworth, 1996 The effect of recombination on background selection. *Genet. Res.* 67: 159–174. <https://doi.org/10.1017/S0016672300033619>
- Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20: R208–R215. <https://doi.org/10.1016/j.cub.2009.11.055>
- Robertson, A., 1960 A theory of limits in artificial selection. *Proc. R. Soc. Lond. B Biol. Sci.* 153: 234–249. <https://doi.org/10.1098/rspb.1960.0099>
- Robertson, A., 1970 A theory of limits in artificial selection with many linked loci, pp. 246–288 in *Mathematical Topics in Population Genetics*, edited by K. Kojima. Springer, Berlin. https://doi.org/10.1007/978-3-642-46244-3_8
- Robertson, A., 1977 Artificial selection with a large number of linked loci, pp. 307–322 in *Proceedings of the 1st International Conference on Quantitative Genetics*, edited by E. Pollak, O. Kempthorne, and T. B. Bailey Jr. Iowa State University Press, Ames, IA.
- Sachdeva, H., and N. H. Barton, 2018 Introgression of a block of genome under infinitesimal selection. *Genetics* 209: 1279–1303. <https://doi.org/10.1534/genetics.118.301018>
- Stern, D. L., 2013 The genetic causes of convergent evolution. *Nat. Rev. Genet.* 14: 751–764. <https://doi.org/10.1038/nrg3483>
- Storz, J. F., 2016 Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* 17: 239–250. <https://doi.org/10.1038/nrg.2016.11>
- Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis *et al.*, 2012 What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367: 364–373. <https://doi.org/10.1098/rstb.2011.0199>
- Uecker, H., and J. Hermisson, 2011 On the fixation process of a beneficial mutation in a variable environment. *Genetics* 188: 915–930. <https://doi.org/10.1534/genetics.110.124297>
- Uecker, H., D. Setter, and J. Hermisson, 2015 Adaptive gene introgression after secondary contact. *J. Math. Biol.* 70: 1523–1580. <https://doi.org/10.1007/s00285-014-0802-y>
- Weissman, D., and O. Hallatschek, 2014 The rate of adaptation in large sexual populations with linear chromosomes. *Genetics* 196: 1167–1183. <https://doi.org/10.1534/genetics.113.160705>
- Yeaman, S., 2015 Local adaptation by small-effect alleles. *Am. Nat.* 186: S74–S89. <https://doi.org/10.1086/682405>
- Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, L. K. Gray *et al.*, 2016 Convergent local adaptation to climate in distantly related conifers. *Science* 353: 1431–1433. <https://doi.org/10.1126/science.aaf7812>

Communicating editor: R. Nielsen

Appendix: Introgression of a Block with Three Loci: Semideterministic Approximations

Consider a genotype “111” introduced at frequency p_0 in a population of N individuals, all having genotype “000.” For simplicity, we consider haploid individuals, but the analysis can be generalized to the diploid case. In the following, “1” will always denote the introduced allele and “0” the native allele. The native alleles at each of the three loci are associated with zero selective effect. The introduced alleles at the three loci have selective effects s_1 , s_2 , and s_3 ; the net selective effect of the introduced genotype is $s = s_1 + s_2 + s_3$. The introduced genotype is of the type $+-+$, i.e., with $s_1 > 0$, $s_2 < 0$, and $s_3 > 0$. We now derive an approximate expression for the fixation probability of the double recombinant 101 (i.e., $+0+$), which is the fittest possible recombinant, using the semideterministic approach employed by Hartfield and Otto (2011).

We consider frequencies of genotypes generated by a single recombination event. These single recombinants are: 100, 110, 001, and 011. The genotypes 110 and 011 will typically be quite rare, as they have lower fitness than the genotypes 001 and 100, as well as the introduced genotype 111. As a first approximation, we assume that their frequency is zero. Then, we need only track three genotypes: 001, 100, and 111 (along with the native genotype 000), before the genotype 101 emerges. Under deterministic dynamics and assuming $s, c \ll 1$, such that all second-order terms in s and c can be neglected, the frequencies of these genotypes evolve as:

$$\begin{aligned}\partial_t p_{111}(t) &= (s - 2c)p_{111} - p_{111}[sp_{111} - s_1p_{100} - s_3p_{001} - c(p_{001} + p_{100} + 2p_{111})] \\ \partial_t p_{100}(t) &= s_1p_{100} + cp_{111} - c(p_{001} + p_{111})(2p_{100} + p_{111}) - p_{100}[sp_{111} - s_1p_{100} - s_3p_{001}] \\ \partial_t p_{001}(t) &= s_3p_{001} + cp_{111} - c(p_{100} + p_{111})(2p_{001} + p_{111}) - p_{001}[sp_{111} - s_3p_{001} - s_1p_{100}]\end{aligned}\quad (6)$$

These can be solved numerically to obtain various genotypic frequencies as a function of time.

The fraction of mating events that generate the recombinant 101 in generation t is $cp_{111}(t)[p_{100}(t) + p_{001}(t)] + 2cp_{100}(t)p_{001}(t)$. Since the frequencies $p_{001}(t)$ and $p_{100}(t)$ are typically quite small, we can approximate the probability of emergence $P_{101}^{emerge}(t)$ of the recombinant 101 by $P_{101}^{emerge}(t) \sim cp_{111}(t)[p_{100}(t) + p_{001}(t)]$. Note that $P_{101}^{emerge}(t)$ is the probability that a mating event in generation t results in a 101 recombinant; it is not the frequency of 101 genotypes in generation t .

We can now calculate the fixation probability $P_{fix}(t)$ of a 101 recombinant generated in generation t by approximating its spread as a time-inhomogeneous BP, where the relative selective advantage of this recombinant changes over time due to the changing mean population fitness (Hartfield and Otto 2011). Then, $P_{fix}(t)$ follows [see Barton (1995)]:

$$\begin{aligned}\partial_t P_{fix}(t) &= -[s(t) - 2c]P_{fix}(t) + P_{fix}^2(t)/2 \\ P_{fix}(t = \infty) &= 2(s_1 + s_3 - s - 2c) \\ s(t) &= s_1 + s_3 - sp_{111}(t) - s_1p_{100}(t) - s_3p_{001}(t)\end{aligned}\quad (7)$$

Equation 7 can be solved numerically to obtain $P_{fix}(t)$. Then, the probability that none of the $NP_{101}^{emerge}(t)$ recombinants (that emerge in generation t) fix is just: $[1 - P_{fix}(t)]^{NP_{101}^{emerge}(t)}$, which can be approximated by $\sim \exp[-NP_{101}^{emerge}(t)P_{fix}(t)]$. Further, the net probability that no double recombinant fixes in a population of size N is $\sim \exp[-N \int_{t=0}^{\infty} dt P_{101}^{emerge}(t)P_{fix}(t)]$. The key assumption is that instances of double recombinants emerging and establishing are rare, so that all such instances can be treated as independent.

Figure A1 Probability that the “101” recombinant fixes in a population with N individuals vs. N . The results of individual-based simulations (points) show good agreement with the prediction of the semideterministic analysis described above (dashed line). The introduced block has three loci with $s_1 = 0.0012$, $s_2 = -0.0006$, and $s_3 = 0.0014$; the rate of recombination between adjacent loci is $c = 0.0002$; the initial frequency of the introduced genotype is $p_0 = 0.006$. The fixation probability is normalized with respect to the probability that at least some part of the introduced block survives. The error bars show uncertainty estimates for the fixation probabilities obtained from individual-based simulations.

Predictions of the semideterministic analysis are in very good agreement with results of individual-based simulations of populations of different sizes (Figure A1). Note that the analysis above assumes deterministic dynamics of the introduced block and other genotypes produced by single crossover events. A more careful analysis can be done, accounting for initial stochasticity, when the introduced block and its descendants are rare and may be lost by chance.