DOI: 10.1534/genetics.109.109447

Signatures of Recent Directional Selection Under Different Models of Population Expansion During Colonization of New Selective Environments

Yuseob Kim*,[1] and Davorka Gulisija†

**Center for Evolutionary Functional Genomics, The Biodesign Institute, and School of Life Sciences, Arizona State University, Tempe, Arizona 85287-5301 and †Center of Rapid Evolution (CORE), Department of Zoology, University of Wisconsin, Madison, Wisconsin 53705*

Manuscript received September 4, 2009
Accepted for publication December 2, 2009

ABSTRACT

A major problem in population genetics is understanding how the genomic pattern of polymorphism is shaped by natural selection and the demographic history of populations. Complex population dynamics confounds patterns of variation and poses serious challenges for identifying genomic imprints of selection. We examine patterns of polymorphism using computer simulations and provide analytical predictions for hitchhiking effects under two models of adaptive niche expansion. The population split (PS) model assumes the separation of a founding population followed by directional selection in the new environment. Here, the new population undergoes a bottleneck and later expands in size. This model has been used in previous studies to account for demographic effects when testing for signatures of selection under colonization or domestication. The genotype-dependent colonization and introgression (GDCI) model is proposed in this study and assumes that a small number of migrants carrying adaptive genotype found a new population, which then grows logistically. The GDCI model also allows for constant migration between the parental and the new population. Both models predict reduction in variation and excess of high frequency of derived alleles relative to neutral expectations, with and without hitchhiking. Under comparable conditions, the GDCI model results in greater reduction in expected heterozygosity and more skew of the site frequency spectrum than the PS model. We also find that soft selective sweeps (fixation of multiple copies of a beneficial mutation) occurs less often in the GDCI model than in the PS model. This result demonstrates the importance of correctly modeling the ecological process in inferring adaptive evolution using DNA sequence polymorphism.

THE pattern of genetic variation within a population is determined by its evolutionary history. The density of polymorphic sites along the chromosomes, the distribution of allele frequencies at those sites, and the statistical association of polymorphism at different sites are influenced by events of natural selection and population (demographic) dynamics (Rosenberg and Nordborg 2002; Nielsen 2005). Population genetic theory allows us to predict the pattern of genetic variation under specific models of selection and demography and, inversely, to infer the evolutionary history from a sample of DNA sequences within a population. A recent event of directional selection is often detected when a sudden removal of polymorphism is observed at a genomic location, due to the hitchhiking effect of a rapidly spreading beneficial mutation that wipes out preexisting variation (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992; Barton 2000). Numerous surveys of DNA sequence polymorphism revealed local reductions of variation clearly due to hitchhiking or selective sweeps (Wootton *et al.* 1999; Nair *et al.* 2003; Schlenke and Begun 2004; Sabeti *et al.* 2006; Macpherson *et al.* 2007; Thornton *et al.* 2007; Williamson *et al.* 2007; Akey 2009). From such findings, it has become evident that directional selection plays a major role in shaping the genomic pattern of sequence variation in natural populations (Gillespie 2000; Begun *et al.* 2007; Hahn 2008). A recent selective sweep also provides basic information regarding directional selection, such as the strength and fixation time of beneficial mutations (Wang *et al.* 1999; Kim and Stephan 2002; Przeworski 2003). However, such inference is not robust to deviation from the standard model of hitchhiking—the fixation of a new codominant beneficial mutation in a constant-sized random-mating population. Fixations of beneficial mutations in real populations are not likely to occur under simple demography or simple models of directional selection (Innan and Kim 2004; Jensen *et al.* 2005; Teshima and Przeworski 2006; Chevin and Hospital 2008).

The sensitivity of the pattern of selective sweeps to biological details poses serious problems for studying

[1]*Corresponding author:* Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University, P.O. Box 875301, Tempe, AZ 85287-5301. E-mail: yuseob.kim@asu.edu

adaptive evolution using genetic data. However, at the same time, it opens the possibility of capturing information that allows the inference of biological context in which adaptive evolution occurs, beyond merely confirming that a certain allele in a genomic region spread quickly in the recent past. Among numerous biological complications, recent studies have focused on the effects of complex demography on the pattern of selective sweeps. Methods have been proposed to extract the signal of genetic hitchhiking from the background pattern of polymorphism shaped by demography (Jensen *et al.* 2005; Nielsen *et al.* 2005) or to estimate the joint parameters of demography and directional selection (Wright *et al.* 2005; Li and Stephan 2006). Approaches of the latter studies would generate information regarding the biological context of adaptive evolution. Such studies, however, require either novel theory of genetic hitchhiking or efficient methods of computer simulation that could predict and generate detailed patterns of polymorphism under models of directional selection in the biological setting of interest.

Many well-known and important examples of adaptive evolution occur during or after the establishment of a new population in a new environment. It is believed that the migration of humans out of Africa was followed by repeated episodes of directional selection. For example, strong selective sweeps at pigmentation genes in some non-African human populations demonstrate the history of those populations' adaptation after migration into new environments (Lamason *et al.* 2005; Myles *et al.* 2007). Likewise, evolution of agronomic traits in domesticated plants and animals involves the establishment of small (cultivated) populations derived from wild ancestors followed by strong directional (artificial) selection (Doebley *et al.* 2006). Other examples include the invasion of a nonnative species into new habitats following human-caused disturbance (Lee 2002; Lee and Gelembiuk 2008) and host switching of pathogens (Parrish *et al.* 2008). In all of these examples, the genetic footprint of directional selection here should overlap with that created by the demographic process of founding and expanding a new population. This ubiquitous mode of evolution, encompassing all the examples above, might be called "adaptive niche expansion."

In this study, we investigate the pattern of genetic variation under two models of adaptive niche expansion. The first model assumes a simple split of ancestral populations into parental and derived populations (Figure 1A). The population split is followed by directional selection on adaptive alleles in the derived population. This model, referred to here as population split (PS), has been used in previous studies that account for demographic effects when testing for signatures of selection under colonization or domestication (Wright *et al.* 2005; Li and Stephan 2006; Thornton and Jensen 2007). The PS model is simple enough to allow the application of standard approximations in

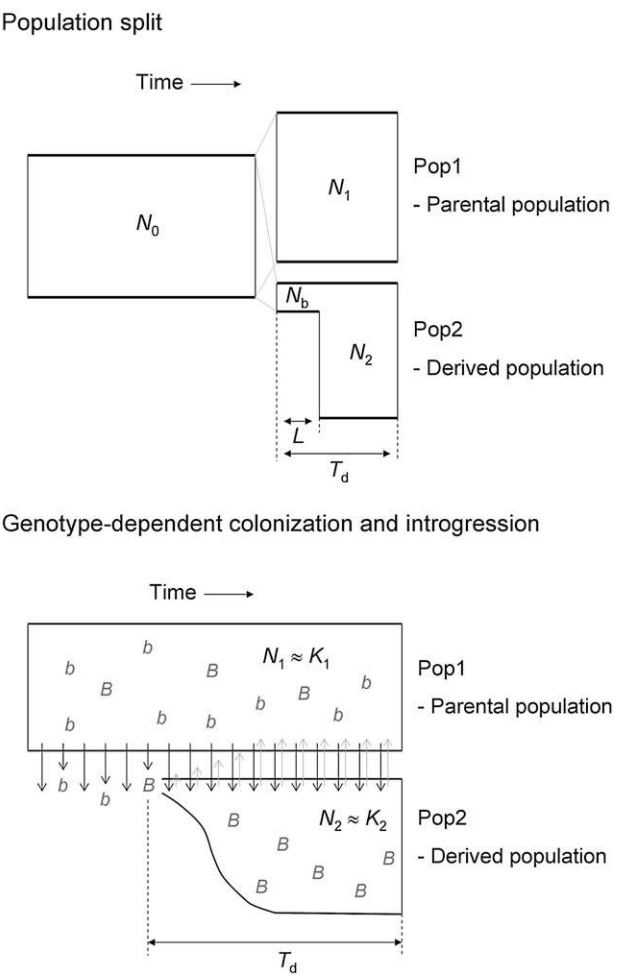

Figure 1.—Schemes for the PS (**A**) and GDCI (**B**) models of adaptive niche expansion.

population genetics: we may use the Wright–Fisher model of reproduction and the coalescent (diffusion) approximation. Using coalescent simulation, the patterns of genetic variation in this model have been extensively studied (Innan and Kim 2004; Wright *et al.* 2005; Thornton and Jensen 2007; Innan and Kim 2008).

While the PS model assumes an instantaneous establishment of a new derived population, the natural processes of colonization may be more gradual and complicated. We therefore consider a different model of adaptive niche expansion in which a small number of migrants carrying certain alleles successfully initiate a logistic growth in a new habitat (Figure 1B; see below for further description). These two models may look similar, the first model being the approximation for the second. However, the ecological processes explicitly modeled are clearly different. Our major interest is whether such subtle ecological/demographic differences leave distinct signatures on the patterns of genetic variation. In this study, we present the two models of adaptive niche expansions and expound on their analytic predictions on the pattern of genetic variation. Distinct impacts on the patterns of genetic variation

between the two models would emphasize the importance of accounting for details in the demographic scenario when testing for signatures of selection.

POPULATION SPLIT MODEL

This model, shown in Figure 1A, was first proposed and investigated in the context of plant domestication (Eyre-Walker *et al.* 1998; Wright *et al.* 2005; Innan and Kim 2008). However, it might be the simplest model that can generally be applied to various scenarios of population expansion into new environment (*e.g.*, Li and Stephan 2006). Here we do not present all theoretical results of the PS model, as basic results were obtained in other studies and the main purpose of this study is to compare this model to a new model of adaptive niche expansion proposed below. We, however, provide full derivation for the sampling probability of sequence polymorphism under the PS model in appendix a.

For simplicity, we consider populations of haploid individuals that undergo recombination upon random union. The ancestral population maintains a constant effective population size (number of haploids) of N_0. At time T_d, the ancestral population splits into two daughter populations. They remain diverged for T_d generations without exchanging migrants. The effective size of the first population, pop1, remains constant at N_1 between time 0 (present) and T_d (time is counted backward). The second population (pop2) is a small founding population of size N_b ($\ll N_0$). This population bottleneck of size N_b lasts for L_b generations. Then, at time $T_d - L_b$, the size of pop2 increases to N_2.

It is assumed that pop2 occupies an environment different from that of ancestral population (pop1). Therefore, directional selection on beneficial alleles, advantageous in the new environment with selection coefficient s, begins in pop2 immediately following the divergence at T_d, as modeled in Innan and Kim (2008). We examine the pattern of genetic variation at a neutral site, which recombines with the site of beneficial mutation with probability r per generation, conditional on the fixation of the beneficial allele. The beneficial allele, denoted B, may originate from the standing genetic variation in the ancestral population or through a single mutation at time T_d in the derived population. In the former scenario, it is assumed that the relative frequency of B is f_0 both in the ancestral population immediately before the population split and in pop2 immediately after the split. A key assumption of the PS model is that the $N_b f_0$ copies of the B allele simultaneously turn beneficial, enjoying the same selective advantage, at the founding of the derived population.

A soft selective sweep (Innan and Kim 2004; Hermisson and Pennings 2005; Pennings and Hermisson 2006) can occur with $N_b f_0 \gg 1$; if two or more copies of B, which may be linked to distinct sequences, increase to high frequency and contribute to the fixation, the reduction of genetic variation is expected to be less severe than it is when a single copy sweeps through the population (hard selective sweep). Note that a hard selective sweep can occur even if $N_b f_0 \gg 1$, because only one copy may survive the stochastic loss in the early phase of increase (Orr and Betancourt 2001). In this model, we quantify the expected prevalence of a soft sweep by H_B, *i.e.,* the probability that two randomly chosen copies of B at present are not identical by descent when the lineages of the two allele are traced back to time T_d. Namely, we define that all $N_b f_0$ copies of the beneficial allele are distinct. (Note that the actual outcome of a soft sweep will depend on whether the two "distinct" copies of B at time T_d have single or multiple mutational origins. However, we do not make an assumption about it. Therefore, our definition of a soft selective sweep remains inclusive for both cases.) The expected frequency of B in pop2 at time t is given by $f = f_0/(f_0 + (1 - f_0)e^{-st})$ and it takes T_f generations to get fixed where $T_f < L_b$ with strong selection. Then, H_B is simply the probability that two randomly chosen lineages of B, starting at present, do not coalesce until time T_d in the past. Therefore,

$$H_B \approx \exp\left[-\frac{T_d - L_b}{N_2} - \frac{L_b - T_f}{N_b} - \int_0^{T_f} \frac{1}{N_b f}\,dt\right] \approx \exp\left[-\frac{L_b}{N_b} - \frac{1 - f_0}{N_b f_0 s}\right]. \tag{1}$$

This equation shows that the probability that the two lineages remain separate until T_d is reduced by the population-size bottleneck ($\exp(-L_b/N_b)$) and by decreasing frequency of B ($\exp(-(1-f_0)/N_b f_0 s)$). A soft selective sweep may occur when both factors are moderate. Therefore, a soft selective sweep can prevail only if f_0 is greater than $1/(N_b s)$.

In the case of hard selective sweep, the expected level of genetic variation can be obtained for a neutral locus linked to the target of selection. The derivation uses the diffusion approximation developed in Kim (2006). From Equations A10, A12, and A13 in appendix a, the expected heterozygosity at a neutral locus, which recombines with the selected locus at rate r per generation, in pop2 is given by

$$\begin{aligned} H(r)_{\text{hard}} &= \theta_0 \int_0^1 \frac{\Phi_{2hh}(2, 1, z, T_d)}{z}\,dz + \Phi_2(1)_{n_2=2} \\ &= \theta_0 e^{-T_d-(L_b)/N_2-L_b/N_b}(1 - y^2) + \boldsymbol{\theta}_b e^{-(T_d-L_b)/N_2}\left(1 - e^{-L_b/N_b}\right) \\ &\quad + \boldsymbol{\theta}_2\left(1 - e^{-(T_d-L_b)/N_2}\right), \end{aligned} \tag{2}$$

where $\theta_0 = 2N_0\mu$, $\theta_b = 2N_b\mu$, $\theta_2 = 2N_2\mu$, and μ is mutation rate at a neutral locus. y is the expected final frequency, after hitchhiking, of descendant copies that trace back to a single copy of the neutral allele on the chromosome where the beneficial mutation occurred (Birky and Walsh 1988; Gillespie 2000). Previous analyses suggest that $y \approx (2N_b s)^{-r/s}$, if $N_b s$ is sufficiently large (>100) (Stephan *et al.* 1992; Kim and Stephan

2002; Kim and Nielsen 2004). The three terms in the last line of the equation above correspond to the contributions of mutations that originate in the ancestral population, in pop2 during the bottleneck, and in pop2 after the bottleneck, respectively. If T_d is short relative to N_2 or N_0, allowing us to ignore genetic drift during the post-bottleneck period and the contribution of new mutations after the population split, the above equation is simplified to

$$H(r)_{\text{hard}} \approx \theta_0 e^{-L_b/N_b}\left(1 - (2N_b s)^{-2r/s}\right). \quad (3)$$

Therefore, the level of genetic variation found in the current pop2 is that of the ancestral population (θ_0) reduced by both population bottleneck ($\exp(-L_b/N_b)$) and the hitchhiking effect ($1 - y^2$).

GENOTYPE-DEPENDENT COLONIZATION AND INTROGRESSION MODEL

The PS model assumes that a new population of N_b individuals suddenly moves into a new habitat and become subject to directional selection. In the case in which the beneficial allele in pop2 originates from standing genetic variation in the ancestral population, $N_b f_0$ copies of this beneficial allele simultaneously become subject to directional selection with equal selective advantages. In reality, the process of establishing a new population might be more gradual than posited in the PS model. Our second model of adaptive niche expansion attempts to capture the gradual ecological process in the establishment of derived population. We consider a scenario in which a parental population (pop1) continuously sends migrants to a new habitat. It is assumed that most migrants die or fail to reproduce at a rate sufficient enough to establish a new population. However, if one or more migrants carry an allele that confers a reproductive success in the new environment, a new population (pop2) might be created by the descendants of the migrants that inherited the adaptive allele. It is assumed that migration from the main population continues after the establishment of the derived allele. Subsequently, the introgression of neutral alleles from the parental to the derived population will occur, homogenizing the pattern of variation in two populations except at loci closely linked to the adaptive locus. We refer to this scenario of adaptive niche expansion as the genotype-dependent colonization and introgression (GDCI) Model (Figure 1B). Depending on the cumulative effects of migration, the GDCI model may generate a signature of selection similar to that of the PS model.

The Wright–Fisher model, or other models of reproduction that require specifying the population size at a given time, is not convenient to be applied here because the growth of the derived population must be modeled separately. Furthermore, if the derived population is a mixture of migrants carrying different genotypes with different fitness, the growth rate of the population depends on the exact genetic composition of the population (each nonadaptive allele produces less than one descendant on average, and thus gets eliminated eventually. However, they do not disappear immediately). We thus use the following simple model of reproduction to allow the feedback between demography and selection. The evolutionary dynamics of different alleles is specified by the absolute, rather than relative, fitness that is a function of ecological parameters. Consider a population of N haploid individuals that reproduce in discrete generations and, during reproduction, may randomly pair and perform recombination. Let W_X be the absolute fitness (the expected number of its descendants into the next generation) of an individual carrying genotype X in the given environment. The number of offspring in the next generation from each individual is Poisson distributed with parameter W_X. Then, the total number of individuals may increase or decrease stochastically between generations. In the GDCI model above, all individuals in the parental population (pop1) are assumed to have the same fitness. We model the absolute fitness in pop1, for all genotypes, as

$$W_1 = 1 + \rho\left(1 - \frac{N_1}{K_1}\right), \quad (4)$$

where ρ is the intrinsic growth rate of the population, N_1 is the current size of pop1, and K_1 is the carrying capacity of pop1. At equilibrium, N_1 will fluctuate around K_1. With a large value of N_1, the reproduction in this population should approach that of the Wright–Fisher model, since the binomial distribution of offspring number in the latter model converges to a Poisson distribution.

In pop2, the absolute fitness depends on whether a haploid carries the adaptive (B) or nonadaptive (b) allele for the environment, if one locus is responsible for the adaptation. Then, we may specify the absolute fitness as

$$W_B = 1 + \rho\left(1 - \frac{N_2}{K_2}\right) \quad (5a)$$

and

$$W_b = (1 - s_b)\left(1 + \rho\left(1 - \frac{N_2}{K_2}\right)\right) = (1 - s_b)W_B, \quad (5b)$$

where N_2 and K_2 are population size and carrying capacity of pop2, respectively. It might be more realistic to assign separate growth rates (ρ) and carrying capacities for allele b rather than using Equation 5b. However, multiplying a single factor $1 - s_b$ in Equation 5b effectively reduces both ecological parameters simultaneously. As indicated above, s_b is given such that $W_b < 1$ for all values of N_2.

We assume that migration occurs in both directions and the number of migrants is proportional to the size

of the source population; at each generation, the expected number of migrants from pop1 to pop2 is $M_1 = N_1 m$ and that from pop2 to pop1 is $M_2 = N_2 m$. As in the case of the PS model, the adaptive allele in pop2, *B*, is assumed to segregate neutrally in pop1 with frequency f_0. Therefore, before pop2 is established, on average $M_1 f_0$ haploids with the adaptive allele arrive in the new environment each generation. Once individuals with the adaptive allele establish the initial population that survives stochastic loss, N_2 grows logistically until it reaches K_2.

We are interested in the pattern of genetic variation at neutral loci in pop2 observed shortly after the growth of pop2 is completed. The amount of variation depends on the linkage to the adaptive locus. If a neutral locus is closely linked to the adaptive locus, its expected heterozygosity in pop2 should be low because most neutral lineages originate from one or a few that migrated into pop2 along with the adaptive allele, *B*, on the same chromosome. This mechanism is fundamentally identical to the hitchhiking effect of a beneficial mutation, as first described in Maynard Smith and Haigh (1974), but in a different mode of directional selection. We thus aim to derive an approximation to $H(r)$, the expected heterozygosity, as a function of recombination rate r.

As our model of reproduction is similar to the Wright–Fisher model with respect to the offspring distribution, we may apply the methods of coalescent approximation that were used in other studies under the Wright–Fisher model. At time t, the derived population is composed of $n_B(t)$ haploids carrying *B* and $n_b(t)$ haploids carrying *b* at the selected locus ($n_B(t) + n_b(t) = N_2(t)$). Counting time backward from the present ($t = 0$), let T_d be the last generation when $n_B(t) > 0$. With limited migration, T_d corresponds to the point when the first successful *B* haploid (carrying the "founding" copy of *B*) starts growing in pop2 (Figure 1B). Then, we randomly pick two alleles at a neutral locus at $t = 0$ in the derived population (pop2) and follow their lineages backward in time. With limited migration and strong selection against *b*, $n_b(0)$ is far smaller than $n_B(0)$. Therefore, we consider only the case in which both neutral lineages are linked to the *B* allele at $t = 0$. The two lineages in pop2 may coalesce before T_d. This event occurs with probability $1/n_B(t)$ at time t, if both lineages remain in pop2 and are linked to *B*. However, if one of the lineages migrate to pop1 (if time is counted backward), the coalescent event cannot occur. There are two routes by which a lineage at the neutral locus can migrate from pop2 to pop1 before T_d. The first route is through recombination onto a chromosome carrying the *b* allele, which shortly moves to pop1 because haploids in pop2 carrying *b* must be recent migrants from pop1 due to selection against *b*. The second route is through direct migration to pop1 along with the linked *B* allele before T_d, which implies a soft selective sweep (this migrating copy of the *B* allele is different from the "founding" copy of *B* that entered pop2 at T_d).

To obtain the approximate probability of first-route migration, we consider the scenario in which, forward in time, most haploid migrants from pop1 to pop2 carry allele *b* at the selected locus ($f_0 \ll 1$). Neutral alleles carried by these migrant chromosomes may stay in pop2 only if they recombine with the *B* allele. Otherwise, they will be eliminated with rate $1 - W_b$. Let $M^*(t_1, t_2)$ be the expected number of neutral lineages that entered pop2 at time t_1 and still remain linked with *b* in pop2 at time t_2 ($T_d > t_1 \geq t_2 \geq 0$). Considering selection against *b*,

$$M^*(t_1, t_2) = M_1(1-r)^{t_1 - t_2} \prod_{t=t_2}^{t_1} W_b(t).$$

Ignoring short-term change in $W_b(t)$, we may use $W_b(t) \approx W_b(t_2)$ for $t_2 \leq t \leq t_1$. Therefore, $M^*(t_1, t_2) \approx M_1((1-r)W_b(t_2))^{t_1 - t_2}$. Then, backward in time, a neutral lineage that is in pop2 and linked to *B* can migrate to pop1 if it recombines with *b*. This happens with probability $rn_b^*(t)/(n_B(t) + n_b(t))$, where $n_b^*(t)$ is the number of neutral lineages that are linked with *b* and shortly migrate back to pop1. $n_b^*(t)$ is different from $n_b(t)$ because some of lineages that are currently linked with *b* may recombine back with *B* before migrating to pop1. We find that

$$n_b^*(t) = \sum_{j=0}^{\infty} M^*(t+j, t) \approx M_1 \sum_{j=0}^{\infty} ((1-r)W_b(t))^j = \frac{M_1}{1-(1-r)W_b(t)}. \tag{6}$$

Considering that migration is not frequent, the probability that either one of the two lineages migrates in this route is approximately $2rn_b^*(t)/(n_B(t) + n_b(t))$.

The probability of second-route migration is simply the proportion of *B* alleles that just migrated from pop1 (forward in time) among all copies of *B* in pop2. Therefore, the probability for a given lineage is $M_1 f_0 / n_B(t)$, because the expected number of *B* allele migrating (forward in time) into pop2 each generation is $M_1 f_0$.

The probability that two lineages coalesce in pop2 is then approximately

$$P_{\text{coal}} = \sum_{t=1}^{T_d} \prod_{i=0}^{t-1} \left(1 - \frac{1}{n_B(i)} - \frac{2rn_b^*(i)}{n_B(i) + n_b(i)} - \frac{2M_1 f_0}{n_B(i)}\right) \frac{1}{n_B(t)}. \tag{7}$$

The expected value of $n_B(t)$ is given by the logistic growth of *B* haploids. Namely,

$$n_B(t) \approx \frac{K_2}{1 + (K_2 - 1)e^{-(T_d - t)}}. \tag{8}$$

$n_b(t)$ is given by $M_1 \sum_{i=0}^{\infty} \prod_{j=1}^{i} W_b(t-j)$. Again, ignoring the short-term change in W_b, $n_b(t) \approx M_1/(1 - W_b(t))$. Using these approximations for $n_b(t)$ and $n_B(t)$, Equation 7 can be calculated. Further simplification of Equation 7 is possible if we note that most coalescent

events would occur when t is reasonably close to T_d. Then, we may substitute $W_b(t)$ in the equations above with $W_b(T_d) = (1-s_b)(1+\rho)$ and also assume that $n_b(t) \ll n_B(t)$ for all t. Therefore, using $\delta = 1-(1-r)(1-s_b)(1+\rho)$,

$$\begin{aligned} P_{coal} &\approx \sum_{t=1}^{T_d}\prod_{i=0}^{t-1}\left(1-\frac{1}{n_B(i)}-\frac{2rM_1/\delta}{n_B(i)}-\frac{2M_1f_0}{n_B(i)}\right)\frac{1}{n_B(t)} \\ &= \frac{1}{1+2rM_1\,\delta+2M_1f_0}\sum_{t=1}^{T_d}\prod_{i=0}^{t-1}\left(1-\frac{1+2rM_1/\delta+2M_1f_0}{n_B(i)}\right)\frac{1+2rM_1/\delta+2M_1f_0}{n_B(t)} \\ &\approx 1\Big/\left[1+2M_1\left(\frac{r}{1-(1-r)(1-s_b)(1+\rho)}+f_0\right)\right] \end{aligned} \tag{9}$$

The summation in the second line becomes one because any one of the three events must happen before T_d. Once one lineage migrates to pop1, with probably $1-P_{coal}$, the remaining lineage also migrates to pop1 before or at T_d. We may ignore the possibility that these two lineages relocated to pop1 are identical by descent (originating from one haploid chromosome at T_d). For the case of a soft sweep (migration through the second route), this requires $K_1f_0 \gg 1$. The expected heterozygosity for these two neutral lineages, given their independent migrations to pop1, is thus identical to that of two lineages randomly chosen at $t=0$ from pop1. This means that the expected heterozygosity in pop2 is identical to that in pop1 unless the coalescent event occurs. Assuming that T_d is very short relative to K_1, which is the coalescent time scale for pop1, we may ignore mutations during the period between $t=0$ and T_d. Then, the expected heterozygosity in pop2 is approximately

$$H(r) = \theta_1(1-P_{coal}), \tag{10}$$

where $\theta_1 = 2K_1\mu$ is the expected heterozygosity in pop1 and P_{coal} is given by either Equation 7 or 9. Figure 2 shows that these analytic approximations are reasonably close to the result of individual-based forward-in-time simulations, which is described in APPENDIX B. Both analytic solution and simulations assume that the allelic difference between two neutral lineages linked to different copies of B at T_d is equal to that between two randomly chosen neutral lineages in pop1. This is not realistic unless recombination rate between two loci is very large or the recurrent mutations between b and B are very frequent. Therefore, the above equation overestimates the level of sequence variation in real data in the case of soft selective sweeps. However, it is currently not feasible to obtain the expected heterozygosity between two neutral alleles that are linked to an identical (by state) neutral allele at another locus (as in the case of B in pop1) that has drifted to frequency f_0. In addition, we note that, if the allele B is deleterious, rather than neutral, in the ancestral population, $H(r)$ given above would further overestimate the actual level of variation, since a deleterious allele has a recent origin.

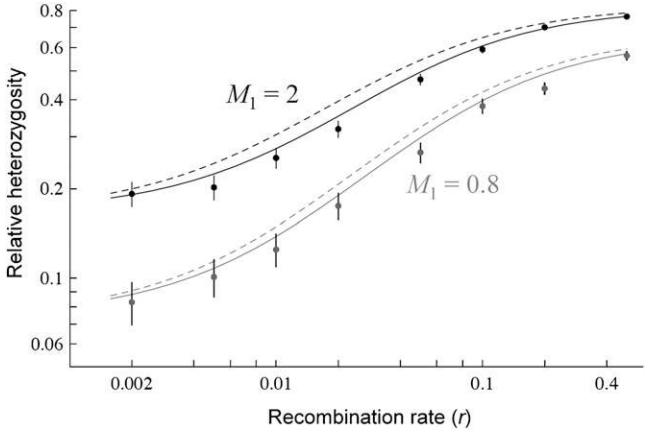

FIGURE 2.—Expected heterozygosities in the derived population relative to the ancestral population in the GDCI model are given for $M_1 = 0.8$ (gray) and 2 (black) as functions of the recombination distance from the adaptive locus. Analytic approximations (Equation 10) using P_{coal} given by Equations 7 (solid curve) and 9 (dashed curve) are shown along with simulation results for eight different recombination rates (mean ± 2 SE). Simulation results are based on 500 replicates of individual-based simulations for each parameter set. Other parameters: $K_1 = 40{,}000$, $K_2 = 50{,}000$, $T_d = 300$, $\rho = 0.05$, $s_b = 0.2$, and $f_0 = 0.05$.

We can isolate the probability of soft selective sweeps, equivalent to Equation 1 for the PS model, by choosing $r=0$ in Equation 7. This leads to the solution identical to the probability of a soft selective sweep due to recurrent migration obtained first by PENNINGS and HERMISSON (2006). Namely,

$$H_B = \frac{2M_1f_0}{1+2M_1f_0}. \tag{11}$$

Note that this probability does not depend on the strength of selection on the B allele in pop2. We can also obtain the effect of hard selective sweeps by letting $f_0 \to 0$. For example, using Equation 9,

$$H(r)_{hard} \approx \theta_1\frac{2M_1r}{2M_1r+1-(1-r)(1-s_b)(1+\rho)}. \tag{12}$$

This equation is equivalent to Equation 3 of the PS model. Equation 12 is compared to the individual-based simulation results of hard selective sweeps in Figure 3. This approximation is more accurate for larger s_b and smaller ρ. Using P_{coal} by Equation 7 yields much better agreement to simulation results than by Equation 9, which presumably reflects an error introduced by the assumption that $n_b(t) \ll n_B(t)$ for all t. Looking backward in time, the recombination events by which a neutral lineage linked to the adaptive allele escapes coalescence ("first-route migration") occur when $n_b^*(t)$ is not too much smaller than $N_2(t)$. This happens in a much shorter window of time compared to the standard model of selective sweeps (KAPLAN *et al.* 1989), as the probability of equivalent recombination event in the latter model is proportional to $1-x$, where x is the

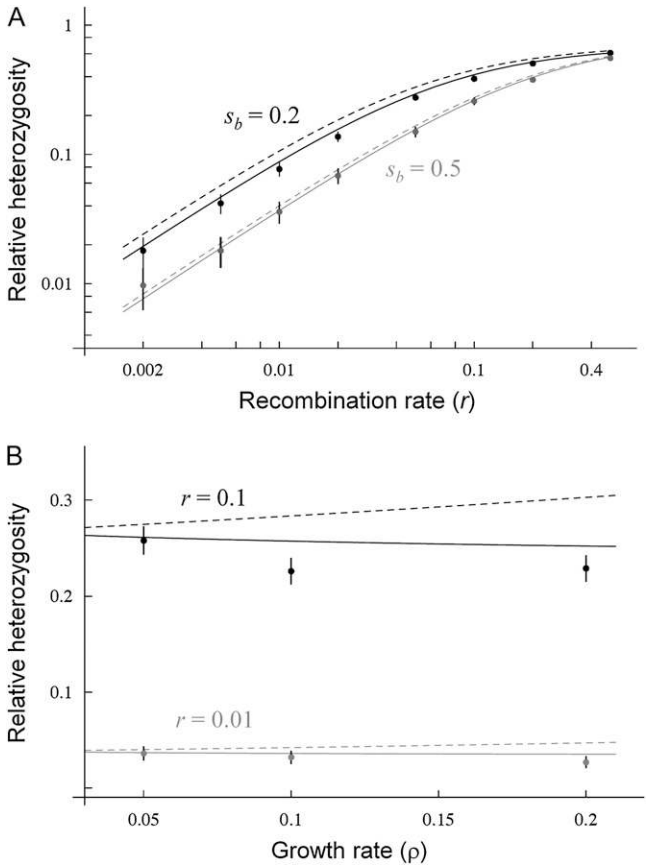

FIGURE 3.—Expected heterozygosities in the derived population relative to the ancestral population after hard selective sweeps in the GDCI model as a function of (A) recombination rate ($s_b = 0.2$ or 0.5) and (B) the haploid growth rate (ρ; $r = 0.01$ or 0.1). Analytic approximations for $H(r)_{\text{hard}}$ using Equation 7 ($f_0 = 0$) and Equation 10 (solid curve) and using Equation 12 (dashed curve), are shown along with simulation results (mean ± 2 SE). Other parameters: $K_1 = 40{,}000$, $K_2 = 50{,}000$, $T_d = 300$, $M_1 = 1$, $\rho = 0.05$ (A), and $s_b = 0.2$ (B).

relative frequency of the beneficial mutation: while $1 - x$ changes gradually according to the logistic function, $n_b^*(t)/N_2(t)$ changes drastically when t is close to T_d. It is thus important to correctly describe $n_b^*(t)/N_2(t)$ for $t \approx T_d$ in Equation 7. Therefore, assuming $N_2(t) = n_B(t)$ when $n_b(t)$ is not much smaller than $n_B(t)$ may cause a significant error. This explains that the approximation using Equation 9 gets worse for smaller s_b (Figure 3A). The approximations and simulation results also indicate that the hitchhiking effect depends very little on the growth rate (ρ) of B haploids in pop2 (Figure 3B).

COMPARISON OF THE TWO MODELS

Both the PS and the GDCI model predict a local reduction of genetic variation due to the hitchhiking effect of an allele favored under the new selective environment in the derived population. However, for comparable effective population bottlenecks and strengths of selection, the two models predict quite different degrees of reduction in expected heterozygosity. Figure 4 shows that, in comparison between equation 1 and 11, the probability of a soft selective sweep is much higher in the PS model than in the GDCI model when we use comparable conditions (L_b and N_b [PS] and M_1 and s_b [GDCI] were adjusted so that both models yield $H(0.5) \approx 0.6$ and s [PS] = s_b [GDCI]). Therefore, the weakened signature of genetic hitchhiking due to soft selective sweeps, considered as a potential difficulty in detecting selection in plant domestication and other models of adaptive niche expansion (INNAN and KIM 2004; HERMISSON and PENNINGS 2005; PRZEWORSKI *et al.* 2005), might be a smaller problem in the GDCI model than in the PS model. With weak selection ($N_b s \ll 1/f_0$), the probability of a soft sweep in the PS model may become as low as that in the GDCI model. [Note that Equation 1 is a function of the strength of selection but Equation 11 is not (HERMISSON and PENNINGS 2005, PENNINGS and HERMISSON 2006).] However, such weak selection may produce a very narrow region of the selective sweep. With a large value of f_0 (> 0.01), there is a limit to which the strength of selection can be reduced and, at the same time, a distinct local reduction of variation (which requires $N_b s \gg 1$) can be produced.

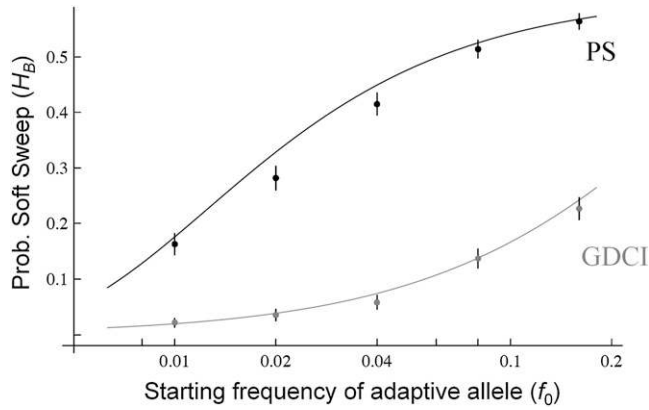

FIGURE 4.—Comparison of the probability of a soft selective sweep (H_B) between the PS and the GDCI model. Analytic approximation for the PS (Equation 1) and the GDCI (Equation 11) are shown by solid curves. Results of individual-based simulation, measured by Equation B1, for eight different recombination rates are also shown (mean ± 2 SE). Parameters: $N_0 = N_1 = 40{,}000$, $N_2 = 50{,}000$, $T_d = 400$, $N_b = 400$, $L_b = 200$, $s = 0.2$ for the PS model and $K_1 = 40{,}000$, $K_2 = 50{,}000$, $T_d = 400$, $M_1 = 1$, $\rho = 0.1$, and $s_b = 0.2$ for the GDCI model.

When we consider hard selective sweeps only, the GDCI model predicts more severe and wider reduction of variation around the adaptive locus (Figure 5). To compare the extent of local selective sweeps, let us define r_c to be a recombination rate that satisfies

$$\frac{H(r_c)_{\text{hard}}}{H(0.5)} = c,$$

where $0 < c \ll 1$. Assuming $2N_b s \gg 1$, which is a condition necessary for producing a distinct local reduction of variation, and using Equation 3, yields

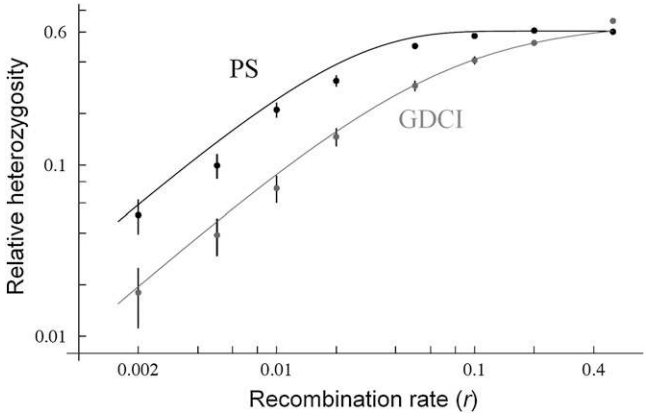

FIGURE 5.—Comparison of the expected heterozygosity (relative to pop1) between the PS and the GDCI model. Analytic approximation for the PS (Equation 3) and the GDCI (Equations 7 and 10) are shown by solid curves. Results of individual-based simulation, measured by Equation B1, for eight different recombination rates are also shown (mean ± 2 SE). Parameters: $N_0 = N_1 = 40{,}000$, $N_2 = 50{,}000$, $T_d = 400$, $N_b = 400$, $L_b = 200$, $s = 0.2$ for the PS model and $K_1 = 40{,}000$, $K_2 = 50{,}000$, $T_d = 400$, $M_1 = 1$, $\rho = 0.05$, and $s_b = 0.2$ for the GDCI model.

$$r_{c,\mathrm{PS}} = -\frac{s\ln(1-c)}{\ln(2N_b s)} \approx \frac{cs}{\ln(2N_b s)} \tag{13}$$

for the PS model. For the GDCI model, using Equation 11,

$$r_{c,\mathrm{GDCI}} \approx \frac{c(1-W_b(T_d))}{2(1-c)M_1 + 2 - (1+c)W_b(T_d)} \approx \frac{c(1-W_b(T_d))}{2M_1 + 2 - W_b(T_d)} = \frac{c(s_b - \rho + s_b\rho)}{1 + 2M_1 + s_b - \rho + s_b\rho}. \tag{14}$$

Then, assuming that the numerators of these formulas (*i.e.,* strength of selection) are comparable, the PS model produces a smaller hitchhiking effect (narrower span of sweep) than the GDCI if $\ln(2N_b s) > 1 + 2M_1 + s_b - \rho + s_b\rho$. This condition may be met for a wide range of reasonable scenarios, for example, if $2N_b s > 100$ and $M_1 \sim 1$ (see DISCUSSION).

Both the probability of soft sweeps and the extent of reduced variation by hard selective sweeps suggest that the GDCI model can produce a greater reduction in polymorphism than the PS model with comparable parameter strength of selection. However, since both models predict V-shaped patterns of local reduction in polymorphism around the locus under selection, it may not be possible to determine which model is more compatible with a given observation of reduced heterozygosity, unless the fitness effect of the nonadaptive allele in the pop2 and effective migration rates can be correctly measured experimentally. We therefore explored whether the site frequency spectrum (SFS; GRIFFITHS 2003) might allow us to distinguish between the two models. The frequency spectrum in the GDCI model can be obtained using a frequency-based forward-in-time simulation (APPENDIX B). Figure 6 shows that, for similar reduction in the heterozygosity of pop2 relative to pop1 after a hard selective sweep, the SFS in the two models are quite different. We find that, relative to the PS model, the hitchhiking effect creates a greater excess of high-frequency derived alleles. Even with moderate reduction of expected heterozygosity ($H(r)/\theta_1 = 0.1 \sim 0.3$), the GDCI model produces an almost U-shaped SFS. It should be noted that a U-shaped distribution can also be produced under the PS or standard hitchhiking model (*e.g.,* KIM 2006, Figure 1), but with a much smaller ratio r/s and thus accompanying a greater reduction in the expected heterozygosity than that shown in Figure 6. A coalescent-based explanation for this difference in the SFS between the PS and the GDCI model is offered below.

DISCUSSION

Strong directional selection is expected to occur during population expansion into new environments. This study investigated genetic hitchhiking under two different models of founding derived populations. The PS model has been used to approximate the evolutionary history of plant domestication (EYRE-WALKER *et al.* 1998; WRIGHT *et al.* 2005; INNAN and KIM 2008) and the

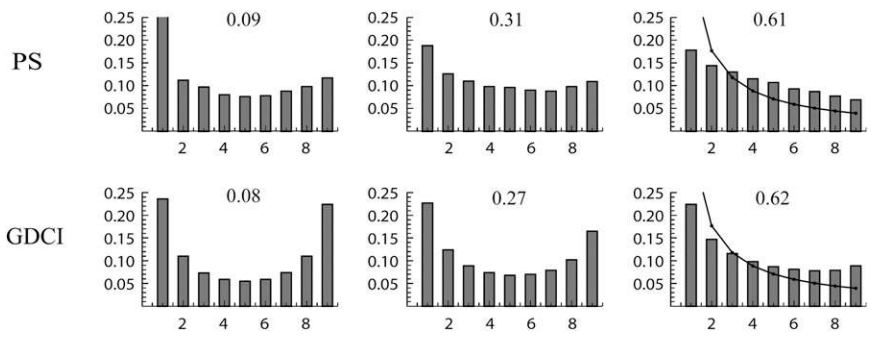

FIGURE 6.—Comparison of the site frequency spectrum in derived populations in the PS and GDCI models. Results of coalescent simulation (INNAN and KIM 2008) for the PS model with three recombination rates ($r = 0.001$, 0.005, and 0.5) are shown on the top row. Other parameters: $N_0 = N_1 = 40{,}000$, $N_2 = 50{,}000$, $T_d = 400$, $N_b = 400$, $L_b = 200$, and $s = 0.1$. Results of frequency-based simulation for the GDCI model with three recombination rates ($r = 0.01$, 0.05, and 0.5) are shown on the bottom row. Other parameters: $K_1 = 40{,}000$, $K_2 = 50{,}000$, $M_1 = 1$, $T_d = 400$, $\rho = 0.1$, $s_b = 0.2$. The expected heterozygosities relative to the ancestral are shown above each histogram.

adaptive expansion of Drosophila populations (Li and Stephan 2006). As the population size is piecewise constant in this model, the application of existing mathematical and computational tools, such as the Wright–Fisher model of reproduction and coalescent simulation, is straightforward. However, a sudden foundation of a derived population and its immediate isolation from the parental population might not be realistic. The model of directional selection starting in this derived population is particularly problematic: positive selection starts simultaneously on all $N_b f_0$ copies of a beneficial allele, where N_b is the initial size of the derived population and f_0 is the frequency of this allele in the ancestral population, immediately after the population split. This scenario may unrealistically maximize the likelihood of soft selective sweeps as it assumes that many copies of beneficial alleles start increasing in frequency simultaneously. In such a case, the signature of hitchhiking, measured by the reduction of variability, is predicted to be significantly weakened since multiple heterogeneous haplotypes increase to high frequency (Innan and Kim 2004; Hermisson and Pennings 2005). However, the simultaneous start of directional selection on multiple copies of a beneficial mutation might not happen in real populations. During plant domestication, for example, selection for a favorable trait (thus for a beneficial allele) may start when an ancestral farmer obtains one or a few individual plants displaying this trait. Then, even if there are many other copies of the same allele in the founding (cultivated) population, only those chosen by this particular farmer would get a head start in reproducing faster and disproportionately contribute to the final fixation of this allele in the population, which would significantly reduce the degree of soft selective sweeps.

On the other hand, the process of founding a derived population is gradual in the GDCI model. Since the change in size of the derived population depends on its genetic composition, the Wright–Fisher model is not adequate. We therefore constructed a model in which each allele leaves descendants according to its absolute fitness, which is determined by ecological parameters. In this model, establishment of the derived population starts with the growth of one or a few migrants carrying the adaptive allele whose absolute fitness in the new environment is greater than one. Even though the constant rate of migration allows late-coming migrants to leave descendants in the new population, their contribution to the final population size is small relative to the early founding migrants. Therefore, genetic variation is less likely to show the pattern of soft selective sweeps at linked neutral loci than in the PS model.

The establishment of a derived population due to continuous migration in the GDCI model might be more realistic than the PS model in many cases of habitat expansion (*e.g.*, freshwater invasion of marine copepods) or certain cases of domestication (*e.g.*, domestication of dogs from gray wolfs). However, it is likely that a real biological process of adaptive niche expansion is more complicated and the PS and GDCI models simply offer two different approximations to the same evolutionary event. It should be noted that the two models studied here are mainly concerned with the adaptive evolution that is critical for the *initial* foundation of the derived population, most likely due to positive selection on the preexisting mutations in the parental population. Therefore, the hitchhiking effect in the GDCI model is necessarily restricted to only one or a few loci in the genome that played the most important role in the initial growth of the derived population. On the other hand, the PS model (with a hard selective sweep) can be used to analyze the fixation of new adaptive mutations that arose after the initial establishment of the new population, even though this population was founded through continuous migration. In this sense, the PS model might be more general. However, if a new mutation that occurs after the foundation of a small population acts to greatly increase the size of this population (*i.e.,* a mutation conferring large absolute fitness), the signature of hitchhiking around this allele might be closer to the GDCI than to the PS model.

Although these two models may approximate the same process of adaptive niche expansion, the pattern of genetic variation at both linked and unlinked neutral loci are strikingly different. With the condition that $\ln(2N_b s) > 1 + 2M_1 + s_b - \rho + s_b\rho$, the expected heterozygosity is much lower in the GDCI model than in the PS model. From $H(0.5) \approx \theta_1 M_1/(M_1 +)1 - 0.5W_b(T_d) \approx \theta_1 M_1/(M_1 +)0.5$ and the fact that most domesticated populations harbor about 30–80% of ancestral variation, a reasonable range of M_1 (the expected number of migrants from pop1 to pop2) for the GDCI model, at least in the case of domestication, might be from 0.2 to 2. This might also be true for most other cases of adaptive niche expansion. Then, considering $2N_b s$ cannot be lower than ~50 to produce a distinct local reduction of genetic variation in the PS model (Kim and Stephan 2002), the above inequality will generally hold. Therefore, for a comparable strength of selection, the reduction of polymorphism caused by the hitchhiking effect will be much more severe in the GDCI model. This severity of hitchhiking effect is mainly explained by the fact that opportunity for the decay of beneficial-neutral allele association exists only briefly in the GDCI model: a given lineage of a adaptive allele that enters pop2 can recombine onto a different neutral allele with probability $rn_b/(n_B + n_b)$, which decreases much more rapidly than the equivalent probability [*i.e.,* $r(1 - x)$, where x is the relative frequency of beneficial mutation] in the PS model, when time is counted forward.

Furthermore, the two models predict different allele frequency distribution in genomic regions that are either close or distant from the locus of selection (Figure 6). In particular, the excess of high-frequency

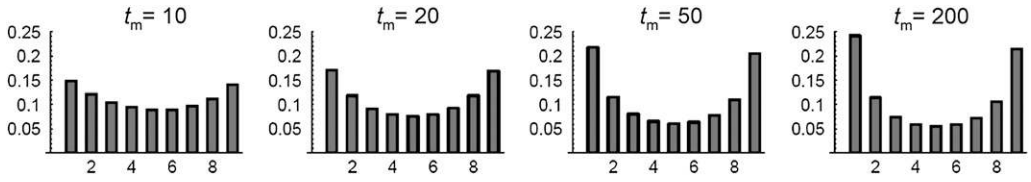

FIGURE 7.—Site frequency spectrum in the GDCI model with limited migration after the establishment of the derived population. Four different lengths of migration period (t_m = 10, 20, 50, and 100) were simulated. Other parameter values are identical to those that produced (T_d = 400, r = 0.01) of Figure 6.

derived allele is much greater in the GDCI model than in the PS model. Interestingly, this excess is also predicted in genomic regions that are unlinked to the locus under selection ($r = 0.5$; Figure 6). This may increase false-positive detection of selective sweeps in a SFS-based analysis assuming the PS model when the actual evolutionary process is closer to the GDCI model.

Here we consider possible explanations for the origin of this unique pattern of frequency spectrum in the GDCI model. First, the excess of high-frequency derived alleles might be due to recurrent migration that continues after the growth of pop2 is completed: neutral variants that hitchhike along the adaptive allele reach high frequencies in pop2 but probably never become fixed in the population due to recurrent migration of ancestral variant from pop1. On the other hand, once neutral variants reach fixation in the PS model by hitchhiking they cannot become polymorphic again. If this explanation is correct, the excess of high-frequency derived allele in the GDCI model should diminish as we limit the migration between two populations after the initial establishment of pop2. Figure 7 shows the SFS obtained by frequency-based forward simulation in which recurrent migration between populations lasts only for 10, 20, 50, and 200 generations after the first founding copy of B enters pop2 ($T_d = 400$). Contrary to the expectation, U-shaped SFS persists for all lengths of period in which migration is allowed. Therefore, the excess of high-frequency derived allele in GDCI model is not explained by continuous migration after the founding of pop2.

A more plausible explanation for the proportion of high-frequency derived alleles might be offered considering the major difference in the expected shapes of neutral genealogies subject to hitchhiking in the PS *vs.* the GDCI model. With small rates of recombination, the genealogy at the linked neutral locus is expected to be either one of three types illustrated in Figure 8 if it could leave nonzero polymorphism in the sample of DNA sequences (it is assumed that the contribution of mutations occurring during or after the process of a selective sweep can be ignored). In the standard model of genetic hitchhiking or the PS model, recombination events during a selective sweep are likely to produce genealogies similar to the type I or type II trees shown in Figure 8: looking backward in time, each lineage may recombine onto a chromosome carrying the nonbeneficial allele (b) and escape the coalescence to other lineages linked to the beneficial allele. Then, the separate lineages that exit the selective phase undergo the neutral coalescent process, leading to long inner branches in the genealogy (FAY and WU 2000; KIM and NIELSEN 2004). Because the rate of recombination event is proportional to $1 - x$ and that of a coalescent event is proportional to $1/x$, where x is the frequency of the beneficial allele (KAPLAN *et al.* 1989), while x is decreasing backward in time, recombination events occur earlier than coalescent events on average. Therefore, each lineage that "migrates" to pop1 by recombination, not having experienced coalescence, is ancestral to only one chromosome in the current sample, thus producing type I or II tree. On the other hand, in the GDCI model, both recombination and coalescent events occur at rates inversely proportional to the number of B haploids [$(2rM_1/\delta)/n_B(i)$ and $1/n_B(i)$, respectively, in the derivation of Equation 9]. As these two events occur concurrently, when a lineage escapes the hitchhiking effect by a rare event of recombination onto a b chromosome, this lineage may be the common ancestor of a variable number of neutral lineages. This process can thus create the type III genealogy, in which the two lineages that exit the selective phase are ancestral to similar numbers of chromosomes in the sample. Both type I and III trees can produce a distribution of derived-allele frequencies that is symmetrical around 0.5, because there are only two long inner branches where a mutation can occur and therefore the expected frequency of the mutant allele in the sample is 0.5. However, the expected heterozygosity is much lower with type I than with type III tree because the former results in derived alleles only in extreme frequencies (singleton polymorphism) in the sample. Then, it will be a type II tree (or any other with multiple independent lineages escaping coalescence by recombination), rather than type I, in the PS model that would produce the level of expected heterozygosity similar to that produced by a type III genealogy in the GDCI model. Type II genealogy does not produce a U-shaped distribution of derived-allele frequency: since new mutations are descended onto

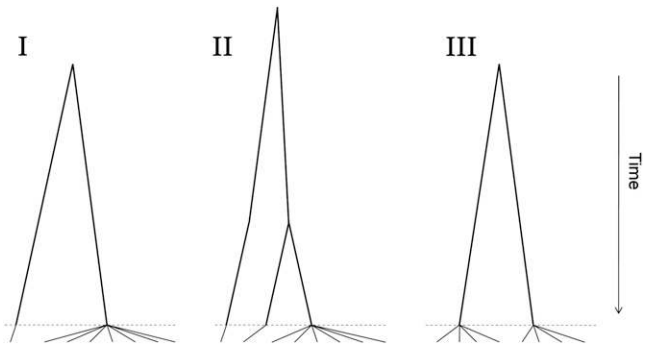

Figure 8.—Types of gene genealogies (coalescent trees) that are expected at a neutral locus that is partially linked to the locus under selection (hard selective sweep) in the derived population (pop2) of the PS or the GDCI model. Dashed horizontal lines represent the time in the past when the first successful *B* haploid appeared in pop2 (thus T_d in the PS and GDCI model). The period below this dashed line is defined to be the "selective phase." In the type I tree, a single lineage avoids coalescing to any other lineage by recombining onto *b*-carrying chromosome during the selective phase. Therefore, two lineages remain at time T_d, which are subject to the coalescent process under the standard neutral model. Another lineage independently escapes coalescence during the selective phase in the type II tree, leaving three lineages at T_d. In the type III tree, neutral lineages start coalescing before one remaining lineage recombines onto a *b*-carrying chromosome. Therefore, two lineages remain at time T_d and they are ancestral to similar numbers of genes in the present-day sample.

only one of three lineages that are connected by inner branches (*i.e.,* three lineages that exit the selective phase) more often than they are descended onto two of the three lineages, the expected frequency of the derived allele in the sample is less than 0.5. We argue that this explains why there is a greater excess of high-frequency derived alleles in the GDCI model than in the PS model for a comparable reduction in the expected heterozygosity.

As a skew of the site frequency spectrum (deviation from the neutral equilibrium) and the pattern of linkage disequilibrium produced after a selective sweep are intimately related to each other due to a common underlying genealogy (Kim and Nielsen 2004), it is also expected that a unique pattern of linkage disequilibrium will be generated under the GDCI model. In summary, our analyses predict significant differences in many aspects of genetic variation between the PS and GDCI model. This result further highlights the importance of correctly modeling the demographic/ecological background in the analysis of selective sweeps. For example, assuming the PS model, one may greatly overestimate the strength of selection based on the chromosomal span of reduced variation. It should also be noted that the estimation of demographic history from the genome-wide SFS (neutral variation at loci unlinked to the locus of selection) will be erroneous if the correct model is not explored (Figure 6).

We thank Dr. Joachim Hermisson for his comments that significantly improved analysis in the manuscript. This study was supported by National Science Foundation grant DEB-0449581 and National Institutes of Health grant R01GM084320 to Y.K., a Hartley Corporation Fellowship from the Graduate Women in Science to D.G., and by National Science Foundation grant DEB-0745828 and University of Wisconsin Vilas Award to Carol Eunmi Lee.

LITERATURE CITED

Akey, J. M., 2009 Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. **19:** 711–722.

Barton, N. H., 2000 Genetic hitchhiking. Philos. Trans. R. Soc. B. Biol. Sci. **355:** 1553.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. **5:** e310.

Birky, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. **85:** 6414–6418.

Chevin, L-M., and F. Hospital, 2008 Selective sweep at a quatitative trait locus in the presence of background genetic variation. Genetics **180:** 1645–1660.

Doebley, J. F., B. S. Gaut and B. D. Smith, 2006 The molecular genetics of crop domestication. Cell **127:** 1309–1321.

Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc. Natl. Acad. Sci. **95:** 4441–4446.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

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

Griffiths, R. C., 2003 The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor. Popul. Biol. **64:** 241–251.

Hahn, M. W., 2008 Toward a selection theory of molecular evolution. Evolution **62:** 255–265.

Hermisson, J., and P. S. Pennings, 2005 Soft sweeps molecular population genetics of adaptation from standing genetic variation. Genetics **169:** 2335–2352.

Innan, H., and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. **101:** 10667.

Innan, H., and Y. Kim, 2008 Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics **179:** 1713–1720.

Jensen, J. D., Y. Kim, V. Bauer DuMont, C. F. Aquadro and C. D. Bustamante, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics **170:** 1401–1410.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kim, Y., 2006 Allele frequency distribution under recurrent selective sweeps. Genetics **172:** 1967–1978.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics **167:** 1513–1524.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

Kim, Y., and W. Stephan, 2003 Selective sweeps in the presence of interference among partially linked loci. Genetics **164:** 389–398.

Kim, Y., and T. Wiehe, 2009 Simulation of DNA sequence evolution under models of recent directional selection. Briefings Bioinformatics **10:** 84–96.

Lamason, R. L., M. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton *et al.*, 2005 SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science **310:** 1782–1786.

Lee, C. E., 2002 Evolutionary genetics of invasive species. Trends Ecol. Evol. **17:** 386–391.

Lee, C. E., and G. W. Gelembiuk, 2008 Evolutionary origins of invasive populations. Evolutionary Appl. **1:** 427–448.

Li, H., and W. Stephan, 2006 Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. **2:** e166.

Macpherson, J. M., G. Sella, J. C. Davis and D. A. Petrov, 2007 Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. Genetics **177:** 2083.

Marth, G. T., E. Czabarka, J. Murvai and S. T. Sherry, 2004 The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics **166:** 351–372.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Myles, S., M. Somel, K. Tang, J. Kelso and M. Stoneking, 2007 Identifying genes underlying skin pigmentation differences among human populations. Hum. Genet. **120:** 613–621.

Nair, S., J. T. Williams, A. Brockman, L. Paiphun, M. Mayxay *et al.*, 2003 A selective sweep driven by pyrimethamine treatment in Southeast Asian malaria parasites. Mol. Biol. Evol. **20:** 1526–1536.

Nielsen, R., 2005 Molecular signatures of natural selection. Annu. Rev. Genet. **39:** 197–218.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res. **15:** 1566.

Orr, H. A., and A. J. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875–884.

Parrish, C. R., E. C. Holmes, D. M. Morens, E. C. Park, D. S. Burke *et al.*, 2008 Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol. Mol. Biol. Rev. **72:** 457–470.

Pennings, P. S., and J. Hermisson, 2006 Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. Mol. Biol. Evol. **23:** 1076–1084.

Przeworski, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667–1676.

Przeworski, M., G. Coop and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution **59:** 2312–2323.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380–390.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. Science **312:** 1614–1620.

Schlenke, T. A., and D. J. Begun, 2004 Strong selective sweep associated with a transposon insertion in Drosophila simulans. Proc. Natl. Acad. Sci. **101:** 1626–1631.

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

Teshima, K. M., and M. Przeworski, 2006 Directional positive selection on an allele of arbitrary dominance. Genetics **172:** 713–718.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive discovery rate in multilocus genome scans. Genetics **175:** 737–750.

Thornton, K. R., J. D. Jensen, C. Becquet and P. Andolfatto, 2007 Progress and prospects in mapping recent selection in the genome. Heredity **98:** 340–348.

Wang, R. L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. Nature **398:** 236–239.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante *et al.*, 2007 Localizing recent adaptive evolution in the human genome. PLoS Genet. **3:** e90.

Wootton, J. C., X. Feng, M. T. Ferdig, R. A. Cooper, J. Mu *et al.*, 1999 Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Plant Mol. Biol. **50:** 333–359.

Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310–1314.

Communicating editor: R. Nielsen

APPENDIX A: DERIVATION OF ANALYTIC SOLUTIONS FOR THE PS MODEL

Sampling probability in the derived population: Under the infinite site model of molecular evolution, the probability of observing k neutral variants at a nucleotide site when n sequences are sampled ($0 < k < n$) is given by

$$P_k = \int_0^\infty N_t \mu \phi\left(n, k, \frac{1}{N_t}, t\right) dt, \tag{A1}$$

where N_t is the number of haploid individuals at time t (counting generations backward from the present), μ is the neutral mutation rate per generation, and $\phi(n, k, z, t)$ is the probability of a neutral mutant that starts at frequency z at time t and is found at frequency k/n in a sample of n sequences at present. Namely, $\phi(n, k, z, t)$ is the expected contribution of neutral mutants at time t to the current polymorphism of size k. It decreases with increasing genetic drift between time t and present and thus depends on the profile of the effective population size during this period. With constant population size N, Kim (2006) found that

$$\phi(n, k, z, t) = \binom{n}{k} \sum_{i=0}^{n} \sum_{j=0}^{i} c_{ij}^{(k,n-k)} z^j e^{-\binom{i}{2}(t/N)}, \tag{A2}$$

using a diffusion approximation. Coefficients $c_{ij}^{(a,b)}$ are obtained from recursions (Kim 2006). This solution can be easily generalized to populations experiencing step-wise size changes.

In the PS model given in Figure 1A, the probability that a neutral variant segregating at time t (originating from either the ancestral population or pop1) is found at frequency k/n_1 in the current sample of n_1 sequences in pop1 is

$$\phi_1(n_1,k,z,t)=\binom{n_1}{k}\sum_{i=0}^{n_1}\sum_{j=0}^{i}c_{ij}^{(k,n_1-k)}z^j e^{-\binom{i}{2}\tau_1}, \tag{A3}$$

where

$$\tau_1=\frac{t}{N_1}I_{(0,T_d)}+\left(\frac{t-T_d}{N_0}+\frac{T_d}{N_1}\right)I_{(T_d,\infty)}. \tag{A4}$$

($I_{(a,b)}=1$ if $a<t\le b$ and 0 otherwise.) Namely, we normalize time by the effective population size. The sampling probability of neutral variants in pop1 is obtained by integrating the contributions by all mutations in the past.

$$\begin{aligned}P_k^{(1)}&=N_0\mu\int_{T_d}^{\infty}\phi_1\left(n_1,k,\frac{1}{N_0},t\right)dt+N_1\mu\int_0^{T_d}\phi_1\left(n_1,k,\frac{1}{N_1},t\right)dt\\&=\binom{n_1}{k}\sum_{i=0}^{n_1}\sum_{j=0}^{i}c_{ij}^{(k,n_1-k)}\left[\frac{\theta_0}{2}\left(\frac{1}{N_0}\right)^j\int_{T_d}^{\infty}e^{-\binom{i}{2}\tau_1}dt+\frac{\theta_1}{2}\left(\frac{1}{N_1}\right)^j\int_0^{T_d}e^{-\binom{i}{2}\tau_1}dt\right].\end{aligned} \tag{A5}$$

where $1\le k\le n_1-1$ and $\theta_0=2N_0\mu$ and $\theta_1=2N_1\mu$ are scaled mutation rates for ancestral population and pop1, respectively. Similarly, the sampling probability in pop2 (without selection at time T_d) is

$$\begin{aligned}P_k^{(2)}=&\binom{n_2}{k}\sum_{i=0}^{n_2}\sum_{j=0}^{i}c_{ij}^{(k,n_2-k)}\\&\times\left[\frac{\theta_0}{2}\left(\frac{1}{N_0}\right)^j\int_{T_d}^{\infty}e^{-\binom{i}{2}\tau_2}dt+\frac{\theta_b}{2}\left(\frac{1}{N_b}\right)^j\int_{T_d-L_b}^{T_d}e^{-\binom{i}{2}\tau_2}dt\right.\\&\left.+\frac{\theta_2}{2}\left(\frac{1}{N_2}\right)^j\int_0^{T_d-L_b}e^{-\binom{i}{2}\tau_2}dt\right],\end{aligned} \tag{A6}$$

where

$$\tau_2=\frac{t}{N_2}I_{(0,T_d-L_b)}+\left(\frac{t-T_d+L_b}{N_b}+\frac{T_d-L_b}{N_2}\right)I_{(T_d-L_b,T_d)}+\left(\frac{t-T_d}{N_0}+\frac{L_b}{N_b}+\frac{T_d-L_b}{N_2}\right)I_{(T_d,\infty)} \tag{A7}$$

($\theta_b=2N_b\mu$ and $\theta_2=2N_2\mu$). The numerical solution to Equation A5 or A6 is very close to that of Equation 1 in Marth *et al.* (2004), which was derived using the coalescent theory. As in the case of their solution, ours can be extended to single populations of more complicated demography as long as the population sizes change in steps. Furthermore, the current formula is more flexible than the solution of Marth *et al.* (2004) in that it can be easily modified to include population divergence and selective sweeps (see below). We obtain this flexibility mainly because the derivation is based on the allele frequency dynamics forward in time, which allows more intuitive arrangement of terms.

Joint frequency spectrum from two divergent populations: Let P_{ij} be the probability that, at a given site, i copies of a derived allele are found in a sample of n_1 sequences in pop1 and j derived alleles in a sample of n_2 sequences in pop2. Here, in addition to the case of simultaneous polymorphism ($0<i<n_1$ and $0<j<n_2$), we consider segregation in one population only ($0<i<n_1$ and $j=0$ or n_2, or $i=0$ or n_1 and $0<j<n_2$) and fixed difference ($i=0$ and $j=n_2$, or $i=n_1$ and $j=0$). First, we examine the expected contribution of neutral mutations that arose in the ancestral population to the joint sampling probability. If such a mutation is either lost or fixed in the population before T_d, it cannot generate any polymorphism or difference between two populations. Only an allele segregating in the ancestral population at time T_d can thus contribute. The probability of finding a derived allele at a frequency interval $[p, p+dp]$ in the ancestral population at T_d is given by $(\theta_0/p)\,dp$ because the ancestral population is assumed to be in neutral equilibrium. Then the contribution to P_{ij} is given by

$$\Lambda(i,j)=\theta_0\int_0^1\frac{\phi_1(n_1,i,z,T_d)\phi_2(n_2,i,z,T_d)}{z}\,dz. \tag{A8}$$

$\Lambda(i,j)$ is well defined for all $0\le i\le n_1$, $0\le j\le n_2$.

Derived alleles that originate after T_d can also be sampled in pop1 or pop2. Define

$$\phi_1(k) = N_1\mu \int_0^{T_d} \phi_1\left(n_1, k, \frac{1}{N_1}, t\right) dt = \frac{\theta_1}{2}\binom{n_1}{k} \sum_{i=0}^{n_1}\sum_{j=0}^{i} c_{ij}^{(k,n_1-k)} \left(\frac{1}{N_1}\right)^j \int_0^{T_d} e^{-\binom{i}{2}\tau_1} dt \tag{A9}$$

for $0 < k \leq n_1$. This is the probability of sampling k neutral mutations that occurred in pop1 between T_d and present. We consider sufficiently small T_d and small θ_1 so that $\sum_{i=1}^{n_1} \phi_1(i) < 1$ (therefore, $\Phi_1(n_1)$ is well defined). Similarly, for pop2

$$\phi_2(k) = \binom{n_2}{k} \sum_{i=0}^{n_2}\sum_{j=0}^{i} c_{ij}^{(k,n_2-k)} \left[\frac{\theta_b}{2}\left(\frac{1}{N_b}\right)^j \int_{T_d-L_b}^{T_d} e^{-\binom{i}{2}\tau_2} dt + \frac{\theta_2}{2}\left(\frac{1}{N_2}\right)^j \int_0^{T_d-L_b} e^{-\binom{i}{2}\tau_2} dt\right]. \tag{A10}$$

Finally, since we assume the infinite sites model,

$$\begin{array}{ll} P_{i0} \approx \Lambda(i,0) + \phi_1(i) & (0 < i \leq n_1), \\ P_{0j} \approx \Lambda(0,j) + \phi_2(j) & (0 < j \leq n_2), \\ P_{ij} \approx \Lambda(i,j) & (i,j > 0). \end{array} \tag{A11}$$

This completes the joint SFS at two populations shown in the PS model without selection. It should be noted that this approach can be extended to models in which more than two populations split from the common ancestor, because of the simplicity of Equation A8.

Adding selective sweeps: Next, we add directional selection to the model. We consider a hard selective sweep caused by a beneficial mutation arising at T_d in pop2. The probability of joint polymorphism (sampling i mutants in pop1 and j mutants pop2) at a linked neutral locus is given by

$$\Lambda_{hh}(i,j) \approx \theta_0 \int_0^1 \frac{\phi_1(n_1, i, z, T_d)\phi_{2hh}(n_2, i, z, T_d)}{z} dz, \tag{A12}$$

where

$$\phi_{2hh}(n,j,z,t) = \binom{n}{j} \sum_{k=0}^{n}\sum_{i=0}^{k} c_{ki}^{(j,n-j)} \{z(y + (1-y)z)^i + (1-z)(1-y)^i z^i\} e^{-\binom{k}{2}\tau_2} \tag{A13}$$

and

$$y = (2N_b s)^{-r/s}.$$

Here, r is the recombination fraction between the neutral and the selected loci and s is the selection coefficient of the beneficial mutation. The joint SFS with a selective sweep is therefore given by replacing $\Lambda(i, j)$ in Equation A11 by $\Lambda_{hh}(i, j)$.

APPENDIX B: SIMULATION METHODS

While the PS model allows a straightforward design of coalescent simulations (Innan and Kim 2008), the coalescent simulations for the GDCI model may not be feasible unless simplifying assumptions on the growth of pop2 are made. We therefore use forward-in-time whole-population simulations to examine the pattern of variation in the GDCI model. To reduce the simulation time, which is a notorious problem for forward-in-time simulations (Kim and Wiehe 2009), we use the method of individual-based simulation for selective sweeps in Kim and Stephan (2003) as well as the frequency-based simulation.

Individual-based simulation: This method of simulating the GDCI model assumes that, when the successful migration of B haploids to pop2 occurs (T_d generations back from the present in Figure 1B), the pattern of variation in pop1 follows that of the standard neutral equilibrium. Then, to obtain the expected heterozygosity at present, we need to observe only the increased identity by descent (due to events of coalescence) in both pop1 and pop2 in the period of T_d generations. The simulation thus starts with $N_1 = K_1$ haploid individuals, represented as chromosomes carrying alleles at one neutral and one selected locus, in pop1. Then, M_1 haploids are randomly chosen from pop1 and move to pop2, where M_1 is Poisson distributed with mean $N_1 m$. Then, in both pop1 and pop2, each haploid produces the Poisson number of offspring according to the absolute fitness given by Equations 3 and 4. While the allele at the selected locus is inherited to every offspring, the allele at the neutral locus is inherited to the offspring with probability $1 - r$. With probability r, the offspring receives the neutral allele of a randomly chosen individual at the parental generation. This process of reproduction is repeated for T_d generations. (If $N_2 > 0$, M_2 haploids are randomly chosen from pop2 and move to pop1, where M_2 is Poisson distributed with mean $N_2 m$.)

The expected heterozygosity at the neutral locus is determined by measuring the increase of the identity by descent during siumulation as described in Kim and Stephan (2003) and Kim and Wiehe (2009). We obtain ϕ_j, the probability that two randomly selected gene lineages in population j do not coalesce between time T_d and 0 (when time runs backward). This quantity determines the expected heterozygosity at present: if new mutations at neutral loci can be ignored between time T_d and 0, two randomly selected alleles from population j are different only if the two lineages do not coalesce between time T_d and 0 and if the two ancestral alleles at time 0 are different (with probability $2K_1\mu$, where μ is the mutation rate, assuming that the expected heterozygosity is approximated by that in the Wright–Fisher population with size K_1). Therefore, the expected heterozygosity in population j is given by $2K_1\mu\phi_j$.

Frequency-based simulation: Next, to obtain the frequency spectrum in the GDCI model, we use the method of frequency-based forward-in-time simulation (Kim and Wiehe 2009). However, the two alleles at the neutral locus, A and a, here are defined by identity-by-state. Each generation, the numbers of individuals, n_1, n_2, n_3, and n_4 corresponding to AB, Ab, aB, and ab haploids, in both populations are updated to n_1', n_2', n_3', n_4' according to the deterministic equations for recombination and migration. Then, the number in the next generation, $n''_i\,(i = 1\ to\ 4)$, is determined by Poisson distribution with mean $n_i'W_i$ for haplotype i, where W_i is obtained from the absolute fitness in Equations 4 and 5. We simulate hard selective sweeps in this model. Therefore, we start the simulation when there is one founding copy of B in pop2: at time 0, with the frequency of A being p_0, $\{n_1, n_2, n_3, n_4\} = \{1, M_1p_0, 0, M_1(1 - p_0)\}$ with probability p_0 (the founding B copy is linked to A) or $\{0, M_1p_0, 1, M_1(1 - p_0)\}$ with probability $1 - p_0$ (the founding B copy is linked to a) in pop2. The initial frequencies for pop1 are $\{n_1, n_2, n_3, n_4\} = \{0, K_1p_0, 0, K_1(1 - p_0)\}$; we simply ignore the frequency of B in pop2 to make sure the hard sweep happens in pop2. We draw p_0 from the standard distribution of derived-allele frequency under neutral equilibrium (probability density of $p_0 \sim 1/p_0$). We run the simulation for T_d generation, conditional on B being established in pop2, and then observe the final frequency of A $(= p)$. By repeating this procedure, the site frequency spectrum at the neutral locus is obtained: the probability of observing j copies of A in a sample of k sequences is $\int_0^1 \binom{k}{j} p^j(1 - p)^{k-j} f(p)\,dp$, where $f(p)$ is the empirical distribution of p obtained in the simulation.