January 1999

Adapt globally, act locally: The effect of selective sweeps on bacterial sequence diversity

J. Majewski
McGill University

Frederick M. Cohan
Wesleyan University, fcohan@wesleyan.edu

Adapt Globally, Act Locally: The Effect of Selective Sweeps on Bacterial Sequence Diversity

Jacek Majewski[1] and Frederick M. Cohan

Department of Biology, Wesleyan University, Middletown, Connecticut 06459-0170

Manuscript received September 20, 1996
Accepted for publication April 23, 1999

ABSTRACT

Previous studies have shown that genetic exchange in bacteria is too rare to prevent neutral sequence divergence between ecological populations. That is, despite genetic exchange, each population should diverge into its own DNA sequence-similarity cluster. In those studies, each selective sweep was limited to acting within a single ecological population. Here we postulate the existence of *globally* adaptive mutations, which may confer a selective advantage to all ecological populations constituting a metapopulation. Such adaptations cause global selective sweeps, which purge the divergence both within and between populations. We found that the effect of recurrent global selective sweeps on neutral sequence divergence is highly dependent on the mechanism of genetic exchange. Global selective sweeps can prevent populations from reaching high levels of neutral sequence divergence, but they cannot cause two populations to become identical in neutral sequence characters. The model supports the earlier conclusion that each ecological population of bacteria should form its own distinct DNA sequence-similarity cluster.

IT is becoming increasingly clear that a full accounting of ecological diversity in the bacterial world requires a molecular approach. Molecular techniques have demonstrated that only a small fraction of bacterial species are culturable (Amman *et al.* 1995; Huber *et al.* 1995; Ohkuma and Kudo 1996), so our best hope of identifying the full scope of bacterial biodiversity is to characterize the sequence diversity of genes that can be amplified directly from natural habitats (Knight *et al.* 1992; Pace 1997). Such surveys typically yield clusters of organisms with similar sequences, and each sequence-similarity cluster is typically interpreted as a distinct ecological population (Britschgi and Giovannoni 1991; Murray and Stackebrandt 1995; Boivin-Jahns *et al.* 1996).

This interpretation is justified because in studies of more familiar and culturable taxa, bacterial systematists have found an empirical correspondence between ecologically distinct populations and sequence-similarity clusters. That is, groups of bacteria known to be ecologically different generally fall into separate sequence-similarity clusters (Vandamme *et al.* 1996; Palys *et al.* 1997); conversely, ecologically uncharacterized strains that fall into separate sequence clusters have subsequently been found to have different ecological properties (Balmelli and Piffaretti 1996; Normand *et al.* 1996). Sequence surveys appear to be an efficient method for discovering the ecological diversity of culturable as well as nonculturable bacteria (Vandamme *et al.* 1996; Palys *et al.* 1997).

While ecologically distinct groups of bacteria are frequently distinguishable as separate sequence-similarity clusters, it is important to find a strong theoretical basis for this observation. If there are times when multiple ecological populations of bacteria fall together into the same sequence cluster, molecular approaches may severely underestimate bacterial biodiversity (Cohan 1994a,b, 1995, 1996, 1999).

Recent theory has shown why ecological populations should correspond to sequence clusters (Cohan 1994a,b; Palys *et al.* 1997). In this theory, ecological populations are defined so that (1) each adaptive mutation confers a benefit only in the genetic background of its original population, and (2) mutant cells bearing an adaptive mutation can outcompete only members of their own population. Natural selection favoring adaptive mutants within a particular population purges that population of genetic diversity *at all loci*, owing to the low rate of recombination in bacteria. [Each such purging event is called a "selective sweep" (Guttman and Dykhuizen 1994b); we refer to recurrent selective sweeps as "periodic selection" (Atwood *et al.* 1951; Koch 1974; Levin 1981).] Because an adaptive mutant does not outcompete cells from other populations, periodic selection purges only the diversity within populations and not the divergence between populations. Each round of periodic selection thereby enhances the distinctness of ecological populations at all loci and fosters the divergence

Corresponding author: Frederick M. Cohan, Department of Biology, Wesleyan University, Middletown, CT 06459-0170.
E-mail: fcohan@wesleyan.edu

[1] *Present address:* Laboratory of Statistical Genetics, Box 192, Rockefeller University, 1230 York Ave., New York, NY 10021.

of different ecological populations into separate sequence-similarity clusters.

The tendency for bacterial populations to form separate sequence clusters is opposed by recombination between populations (Cohan 1994a). Depending on the rates of interpopulation recombination and the intensity of periodic selection, the model has shown three possible classes of outcomes of neutral sequence divergence between populations: (1) under extremely low rates of recombination, populations will diverge without bound, so that every nucleotide site that can be substituted harmlessly will eventually become substituted; (2) under higher rates of recombination, populations will reach an equilibrium level of divergence, so that populations fall into distinct sequence clusters, but divergence between them never becomes saturated; and (3) under yet higher rates of recombination, ecologically distinct populations will not be distinguishable by neutral sequence data, as the levels of divergence within and between populations will be nearly equal. Given the low rates of recombination estimated thus far for bacteria (Selander and Musser 1990; Maynard-Smith *et al.* 1993; Whittam and Ake 1993; Guttman and Dykhuizen 1994a; Roberts and Cohan 1995), the model predicts that each population should be distinct as a separate sequence-similarity cluster, as described in cases 1 and 2 above (Cohan 1994a, 1995; Palys *et al.* 1997).

Nevertheless, it is not clear that the existing model adequately predicts the degree of sequence divergence between ecological populations. Here we present an alternative and more general model for periodic selection, in which some mutations may be adaptive outside of the context of their original populations. In this model, the domain of competitive superiority of an adaptive *mutant* (*i.e.*, the cell) is still its own ecological population, but the adaptive *mutation* (*i.e.*, the allele) can be recombined into other populations, where it can confer higher fitness and cause a local selective sweep within each recipient population (Figure 1). This process may homogenize the populations for any segment that is cotransferred between populations along with the adaptive mutation. We have hypothesized that globally adaptive mutations could homogenize populations for neutral sequence diversity at all gene loci, provided that the size of fragments recombined is large enough and that universally adaptive mutations recur throughout the genome.

In this article, we present a coalescence model to explore the conditions under which universally adaptive mutations can homogenize neutral sequence diversity across ecological populations. We tested whether different ecological populations might fail to diverge into separate sequence-similarity clusters under the rates of recombination observed in bacteria. We also tested whether universally adaptive mutations may prevent populations with low recombintion rates from diverging without bound.

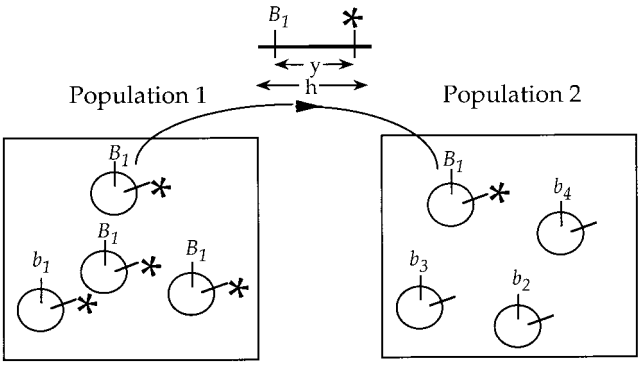

Figure 1.—Transmission of a globally adaptive mutation from population 1, where the mutation originally occurred, to population 2. The asterisk represents an adaptive mutation; the *B* locus is a gene of interest, whose allelic diversity we are monitoring. The B_1 allele at this locus was originally associated with the adaptive mutation in population 1. The diagram shows a possible scenario for coalescence of gene segments from different populations. A selective sweep has just been completed in population 1: most of the local diversity at locus *B* has been purged and the population contains mostly the allele originally associated with the adaptive mutation (allele B_1). The remaining diversity in population 1 is represented by b_1. The B_1 allele is then cotransferred along with the adaptive allele into population 2. The fraction of the genome that is cotransferred is *h*, and *y* is the distance between the target of selection and the gene of interest. When the selective sweep is complete in population 2, most of the diversity at the *B* locus will also be purged by the B_1 allele (provided that recombination rates are low enough). Hence, the selective sweep leads to coalescence of neutral gene segments within the same population and also in different populations.

THE MODEL

Ecological populations and adaptive mutations: A metapopulation consists of *n* closely related ecological populations, each containing *N* cells (Table 1). Each population is adapted to a different ecological niche. Recombination occurs rarely within and between these populations, and the metapopulation is closed to recombination with other such metapopulations.

Following Cohan (1994a), we define an ecological population as the domain of competitive superiority of an adaptive mutant. Thus, an adaptive mutant would outcompete to extinction all other strains from the same population (because they are adapted to the same niche) but would not drive to extinction strains from other populations. While an adaptive *mutant* (*i.e.*, the cell) has a competitive advantage only within its own ecological population, an adaptive *mutation* (*i.e.*, the mutant allele) may be either locally or globally adaptive. A locally adaptive mutation confers a benefit only in the genetic background of its original population, whereas a globally adaptive mutation can confer a benefit in any genetic background within the metapopulation. A global selective sweep occurs when a globally adaptive mutation recombines from its original population into other populations: any cell receiving the adaptive muta-

TABLE 1

Definitions of parameters

n	Number of populations in the metapopulation
N	Number of cells in each population
c	Rate (per gene segment per genome per generation) at which individuals of a given population integrate DNA from any other individual in the metapopulation
c_s	Rate (per gene segment per genome per generation) at which individuals of a given population integrate DNA from other individuals of the same population
c_d	Rate (per gene segment per genome per generation) at which individuals of a given population integrate DNA from any other population
c_δ	Rate (per gene segment per genome per generation) at which individuals of a given population integrate DNA from one particular population (other than their own)
σ_g, σ_l	Per population rate of global and local selective sweeps, respectively
μ_g, μ_l	Per capita rate of globally and locally adaptive mutations, respectively
z	Selective advantage of an adaptive mutant
h	Average size of a recombining DNA fragment, measured as a fraction of the genome
y	Distance between the adaptive mutation and the gene segment of interest, measured as fraction of the genome, for a particular selective sweep
p_l	Probability that the local selective sweep results in coalescence of two segments from the same population
p_{gs}	Probability that a global selective sweep results in coalescence of two segments from the same population
p_{gd}	Probability that a global selective sweep results in coalescence of two segments from different populations
P	Average of p over many selective sweeps, representing all possible values of y: P_l, P_{gs}, P_{gd} are the average values of p_l, p_{gs}, p_{gd}, respectively
q	Probability of corecombination of the adaptive allele and the segment of interest
μ_0	Neutral mutation rate per third-base site
Δ_s, Δ_d	Number of nucleotide substitutions per third-base site, for two gene segments chosen randomly from the same population and different populations, respectively
π_s, π_d	Nucleotide divergence over all sites within and between populations, respectively

tion from the original population is then able to outcompete other members *of its own population* (Figure 1). Whereas a globally adaptive mutation confers fitness globally to all cells in the metapopulation, natural selection acts only locally to favor the adaptive genotype *within each population*.

We assume that selective sweeps are rare events and that the duration of the sweep is short relative to the time between sweeps.

Rate of fixation of adaptive mutations: Following Cohan (1994a), adaptive mutation is modeled as a one-step process that occurs randomly over time at a rate μ_g (for global adaptations) or μ_l (for local adaptations) per capita per generation. Each adaptive mutation confers a selective advantage z. Taking into account that only the fraction $2z$ of adaptive mutations is expected to become fixed (assuming that the population size $N \gg 1/z$; Wright 1931), locally adaptive mutations are fixed within a population by directional selection at a rate $\sigma_l = 2z\mu_l N$. It is assumed that once a globally adaptive mutation becomes fixed in its original population (with probability $2z$), recurrent recombination and subsequent selection will cause the mutation to eventually become fixed in all populations. Therefore, globally adaptive mutations are fixed at a rate $\sigma_g = 2z\mu_g nN$.

Recombination within and between populations: Recombination in bacteria is unidirectional and the segment recombined is usually a small fraction of the genome (Smith 1988). We therefore model recombination as a gene conversion process in which a segment of the recipient DNA is replaced with the homolog of the donor. The model is concerned with recombination at two loci: a "gene of interest," whose sequence divergence we wish to predict, and a selected gene, whose adaptive mutation is favored by selection. Each gene is assumed to be short enough so that it is not split by recombination. A single recombination event may involve one or both of the genes, depending on the size of the recombining fragment (h) and the distance between the loci (y).

Recombination follows a modified island model, where c_s is the rate (per gene segment per genome per generation) at which individuals integrate (as recipients) DNA at a gene segment of interest from other individuals of the same ecological population; c_d is the rate at which individuals integrate DNA from any other ecological population; c is the total rate of recombination at which an individual integrates DNA from any other individual in the metapopulation; thus $c = c_s + c_d$. The value c_δ is the rate at which individuals integrate DNA from a *particular* ecological population (other than their own). In a metapopulation consisting of n ecological populations, $c_\delta = c_d/(n - 1)$.

Probability that a selective sweep leads to coalescence: Our model determines the expected time (going backward from the present) to coalescence into a common ancestor for two homologous gene segments occurring today in two different individuals. These individuals may be cells from the same or different ecological populations of the metapopulation.

We define p as the probability that a selective sweep leads to coalescence at a gene segment of interest. This is the probability that two cells chosen from the metapopulation immediately following a selective sweep are identical by descent for the gene segment of interest. Whether a selective sweep results in coalescence at a

gene of interest depends on the relative magnitudes of the selective advantage of the adaptive mutation, the rate at which recombination separates the gene of interest from the selected gene, and the population size. If the rate of recombination is high and the selective advantage low, the event is unlikely to lead to coalescence.

We consider several instances of the variable p, corresponding to the probabilities of coalescence within and between populations, for globally and locally adaptive mutations: p_l is defined as the probability that a *local* selective sweep within a population leads to coalescence of segments from that population; p_{gs} is the probability that a *global* selective sweep leads to coalescence of segments from the *same* population; and p_{gd} is the probability that a *global* selective sweep leads to coalescence of segments from *different* populations of the metapopulation.

In appendix a, we derive a method (adapted from Kaplan *et al.* 1989) for calculating p_l, p_{gs}, and p_{gd}, for the special case of two ecological populations, *i.e.*, $n = 2$. The variables as p_l, p_{gs}, and p_{gd} are functions of c_s, c_d, N, and q, where q is the probability that a recombination event results in corecombination of the adaptive allele with the segment of interest. This probability is a function of the length h of the DNA taken up by a recipient cell during a recombination event and of the distance y between the adaptive mutation and the segment of interest (both h and y are measured as fractions of the genome):

$$q = \begin{cases} (h - y)/h & \text{if } h \le 1/2;\ y \le h, \\ 0 & \text{if } h \le 1/2;\ y > h, \\ (h - y)/h & \text{if } h > 1/2;\ y \le 1 - h, \\ (2h - 1)/h & \text{if } h > 1/2;\ y > 1 - h. \end{cases} \quad (1)$$

Because the coalescence of homologous segments from different populations requires that the transfer of the adaptive mutation from population 1 to population 2 includes the segment of interest (Figure 1), p_{gd} will be highly dependent on the probability of cotransfer.

We assume that the size (h) of the recombining DNA fragment is constant, while adaptive mutations occur randomly throughout the genome. Because we are interested in modeling the consequences of many selective sweeps, we need to calculate the mean probability (P) that a selective sweep leads to coalescence, averaged over all possible distances (y) between the neutral marker and the adaptive mutations (*i.e.*, between 0 and ½ because the bacterial chromosome is circular):

$$P = 2\int_0^{1/2} p(y)dy. \quad (2)$$

Both the probabilities p and the above integral were evaluated numerically.

The coalescence model: Our coalescence model calculates the expected time that two homologous gene segments (occurring in different organisms) have diverged since their last common ancestor. These gene segments are postulated to be short enough so that they are not split by recombination. The following are the expected times to coalescence for two strains from the same and different ecological populations, $E(t_s)$ and $E(t_d)$ (derived in appendix b):

$$E(t_s) = \frac{1 + \sigma_l(1 - P_l)E(t_s) + \sigma_g(1 - P_{gs})E(t_s) + 2c_dE(t_d)}{1/N + \sigma_l + \sigma_g + 2c_d} \quad (3)$$

$$E(t_d) = \frac{1 + \sigma_g(1 - P_{gd})E(t_d) + 2c_\delta E(t_s)}{\sigma_g + 2c_\delta}. \quad (4)$$

These equations were solved using Mathematica (Wolfram 1991). The solutions are too long to present here. Numerical representations of the probability density functions, $P_{ts}(t)$ and $P_{td}(t)$, of the respective times to coalescence were also calculated using a method outlined in appendix b.

Calculation of the expected nucleotide divergence: The expected nucleotide sequence divergence is predicted using the probability density functions for t_s and t_d following Cohan (1994a). Nucleotide substitutions are postulated to consist only of synonymous mutations, and every third base substitution is taken to be synonymous, with no synonymous substitutions allowed at the first or second bases of codons. The number of neutral substitutions per third base site (Δ) is then obtained by multiplying the time to coalescence by twice the per third base site rate of mutation (μ_0),

$$\Delta = 2\mu_0 t, \quad (5)$$

where $t = t_s$ or t_d are the times until coalescence of segments in the same population or different populations, respectively.

The nucleotide sequence divergence over all sites, π, may be calculated by correcting for multiple substitutions per site (Jukes and Cantor 1969) and correcting for substitutions occurring only at third base sites:

$$\pi = \frac{1}{3} \cdot \frac{3}{4}(1 - e^{-4\Delta/3}). \quad (6)$$

We used the probability density functions $P_{ts}(t)$ and $P_{td}(t)$ to calculate the expected divergence

$$E(\pi) = \frac{1}{3} \cdot \frac{3}{4}\int_0^\infty (1 - e^{-4\Delta/3}) \cdot P_t(t)dt. \quad (7)$$

The integration was carried out numerically. The expected nucleotide divergence between segments from the same population, $E(\pi_s)$, and different populations, $E(\pi_d)$, was calculated using $P_{ts}(t)$ and $P_{td}(t)$, respectively. The neutral nucleotide divergence after infinite time reaches a level of ¼, which we refer to as "unbounded divergence."

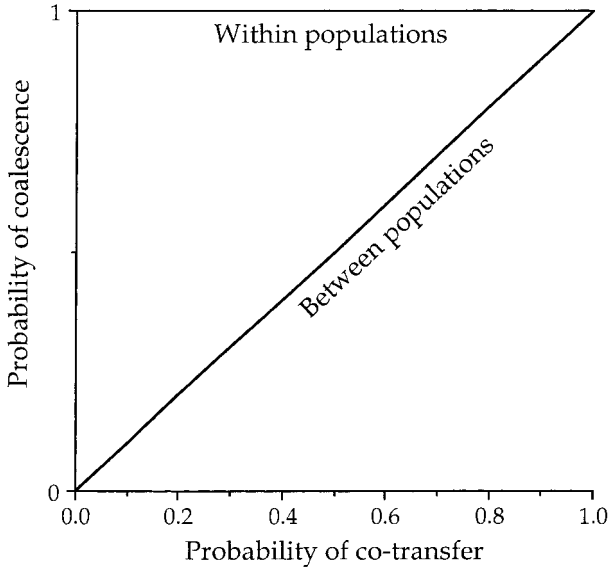

Figure 2.—The probability of coalescence at a locus of interest during a global selective sweep, within (p_{gs}) and between (p_{gd}) populations, as a function of the probability of cotransfer (q) of the gene of interest with the adaptive mutation. The total rate of recombination was set to $c = 10^{-8}$ (based on estimates in bacteria, see discussion). The selective advantage, $z = 10^{-2}$; the population size, $N = 5 \times 10^{14}$; and the number of populations in the metapopulation, $n = 2$. Recombinations rates within and between populations were set as equal ($c_s = c_\delta$). The values p_{gs} and p_{gd} were calculated using Equation A4.

RESULTS

The following parameter values were used in all numerical calculations: the neutral mutation rate per third base site, $\mu_0 = 3 \times 10^{-10}$; the selective advantage, $z = 10^{-2}$; the population size, $N = 5 \times 10^{14}$; and the number of populations in the metapopulation, $n = 2$. Recombination rates within and between populations were set as equal ($c_s = c_\delta$; *i.e.*, no sexual isolation between populations) to maximize the homogenizing effect of recombination.

The diversity-purging effect of an adaptive mutation: The probability that a particular global selective sweep causes coalescence, within or between populations, is shown in Figure 2. The probability of coalescence within a population, p_{gs}, is always near 1 because recombination is so rare in bacteria (see also Cohan 1994b). The probability of coalescence of segments from different populations, p_{gd}, is approximately equal to q (Figure 2). This is because a globally adaptive mutation causes coalescence between populations at a gene of interest only when it causes coalescence *within* each population (occurring with probability p_{gs}) and the gene of interest is cotransferred between populations along with the adaptive mutation (occurring with probability q). Thus, the probabilities of coalescence within and between populations are most similar for genes most closely linked to the adaptive mutation (*i.e.*, $q = 1$).

The ratio of globally to locally adaptive mutation rates: We next explore the effect of recurrent adaptive mutations on population structure. We focus on the significance of the ratio of globally to locally adaptive mutations. We maintain the total frequency of adaptive mutations constant, while allowing the ratio of global:local adaptations to vary. We consider three relative frequencies of global:local events, 1:0, 1:1, and 0:1 (Figure 3).

Figure 3 shows that globally adaptive mutations reduce neutral sequence divergence between populations compared to the case with only local adaptations. This effect is most pronounced at low recombination rates. When only local selective sweeps are possible, the model shows that a recombination rate of 10^{-10} leads to unbounded neutral divergence between populations (*i.e.*, $\pi_d \approx 1/4$). Increasing the global:local ratio decreases the divergence between populations by up to 50-fold. The divergence within populations also decreases, but to a much lower extent. Thus, increasing the proportion of globally adaptive mutations makes the populations less distinct in neutral characters.

Consider next whether globally adaptive mutations can prevent different ecological populations from diverging into separate sequence clusters. We define populations as falling into separate sequence-similarity clusters when $E(\pi_d) > 2E(\pi_s)$ (Palys *et al.* 1997). Using this criterion, Figure 4 shows that the critical recombination rates necessary for populations to diverge into separate clusters are nearly the same whether or not globally adaptive mutations occur.

Analysis of a simplified model with no locally adaptive mutations: We concentrated on the effect of globally adaptive mutations by considering the special case of a two-component metapopulation in which all the adaptive mutations are global (Figure 5). For this special case we treated the coalescence equations analytically to gain further insight into the behavior of the sequence divergence functions presented in Figure 3. Noting that bacterial populations are always large enough so that the probability of coalescence by drift is negligible relative to coalescence by periodic selection (*i.e.*, $1/N \ll \sigma P$), Equations 3 and 4 reduce to

$$E(t_s) = \frac{4c_\delta + \sigma_g P_{gd}}{(2c_\delta + \sigma_g P_{gs})(2c_\delta + \sigma_g P_{gd}) - 4c_\delta^2} \qquad (8)$$

$$E(t_d) = \frac{4c_\delta + \sigma_g P_{gs}}{(2c_\delta + \sigma_g P_{gs})(2c_\delta + \sigma_g P_{gd}) - 4c_\delta^2}. \qquad (9)$$

We may consider $\sigma_g P_{gs}$ and $\sigma_g P_{gd}$ as pseudoparameters, representing the diversity-purging effect of periodic selection (*i.e.*, the rate of selective sweeps times the probability of coalescence within each sweep). The times to coalescence are then determined by only three factors: c_δ, the rate of recombination between populations; $\sigma_g P_{gs}$, the within-population diversity-purging effect of global

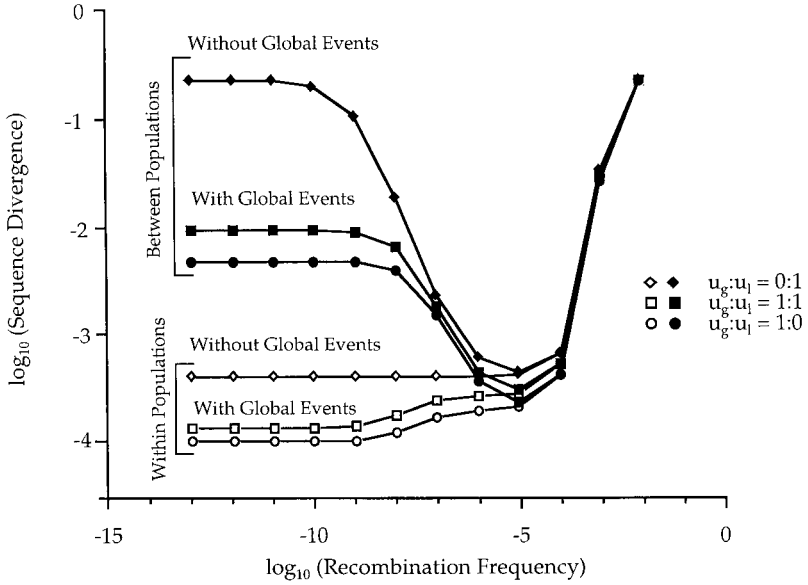

Figure 3.—Expected sequence divergence within ($E[\pi_s]$) and between ($E[\pi_d]$) populations as a function of the recombination rate ($c = c_s + c_\delta$), for different ratios of global:local selective sweeps. The total rate of selective sweeps is maintained constant at $\mu_l + \mu_g = 10^{-19}$. The size of recombination fragments is 2% of the genome ($h = 0.02$). The neutral mutation rate per third base site, $\mu_0 = 3 \times 10^{-10}$; the selective advantage, $z = 10^{-2}$; the population size, $N = 5 \times 10^{14}$; and the number of populations in the metapopulation, $n = 2$. Recombination rates within and between populations were set as equal ($c_s = c_d$). The graph is based on the solution of Equation 7.

periodic selection; and $\sigma_g P_{gd}$, the between-population diversity-purging effect of global periodic selection.

Consider the relative magnitudes of $\sigma_g P_{gs}$ and $\sigma_g P_{gd}$. We used Equation A4 of appendix a to calculate the values of P_{gs} and P_{gd} across the range of frequency of recombination (c) and selective advantage (z) considered in this article, and we found that $P_{gs} \geq P_{gd}/h$. We assume that the size of the recombination fragment h is usually <10% for the genome (see discussion). Therefore $\sigma_g P_{gs} \gg \sigma_g P_{gd}$. This leaves four regions of magnitude for c_δ: $c_\delta \gg \sigma_g P_{gs}$, $c_\delta \sim \sigma_g P_{gs}$, $c_\delta \sim \sigma_g P_{gd}$, and $c_\delta \ll \sigma_g P_{gd}$. These regions correspond to regions I through IV, respectively, of Figure 5.

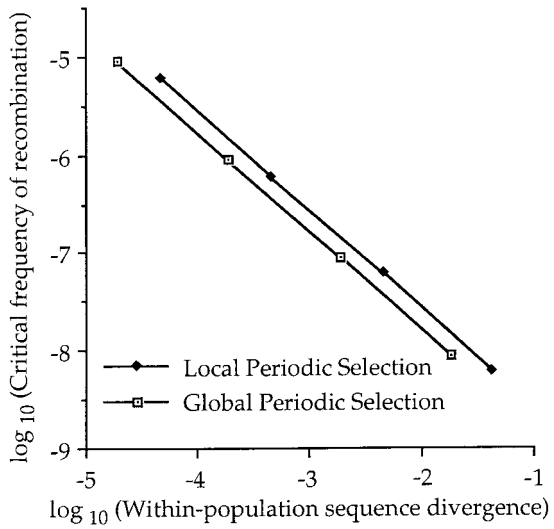

Figure 4.—The maximum (critical) recombination frequency c that allows ecological populations to fall into distinct sequence-similarity clusters as a function of the expected sequence divergence within populations $E(\pi_s)$. The critical recombination frequency is defined here as that which yields $E(\pi_d) = 2E(\pi_s)$. Any recombination rate higher than this would yield less distinct populations. The curves for local and global selective sweeps represent the situations where all adaptive mutations are locally or globally adaptive, respectively, with μ_l or μ_g equal to 10^{-19}. The size of the recombination fragment is set at $h = 10\%$; other parameters are as in Figure 3. Note that, even for this large recombination fragment size, the critical recombination rate yielding distinct sequence clusters is virtually the same whether or not globally adaptive mutations occur. The graph is based on the solution of Equation 7.

Region I of Figure 5, $c_\delta \gg \sigma_g P_{gs} \gg \sigma_g P_{gd}$, corresponds to very high recombination rates, yielding the following approximation of Equations 8 and 9:

$$E(t_d) \approx E(t_s) \approx \frac{2}{\sigma_g P_{gs}}. \quad (10)$$

Region I of the graph corresponds to the case where high rates of recombination within populations (c_s) diminish the diversity-purging effect of periodic selection (P_{gs}), while high values of interpopulation recombination (c_δ) further homogenize the populations, making them indistinguishable (such that expected divergence levels within and between populations are equal).

The conditions in region II, $c_\delta \sim \sigma_g P_{gs} \gg \sigma_g P_{gd}$, yield

$$E(t_s) \approx \frac{2}{\sigma_g P_{gs}} \quad (11)$$

$$E(t_d) \approx \frac{2}{\sigma_g P_{gs}} + \frac{1}{2c_\delta}. \quad (12)$$

In region II of Figure 5, recombination is no longer sufficient to prevent populations from diverging. The divergence between populations is greater than that within populations and is determined by the equilibrium between recombination (which acts to homogenize the populations) and local diversity-purging events (which tend to keep the populations distinct).

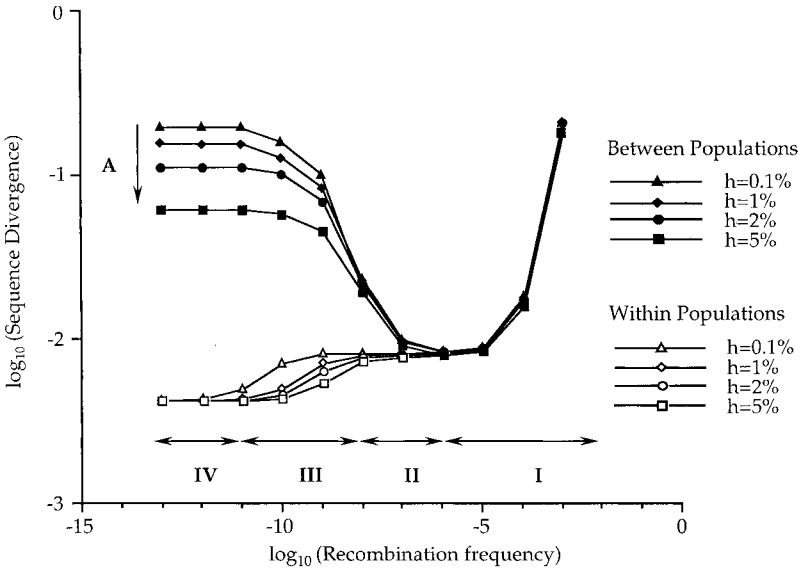

Figure 5.—Divergence within populations ($E[\pi_s]$ and between populations ($E[\pi_d]$) as a function of the recombination frequency (c), for different sizes of the recombining fragment (h). All selective sweeps are global ($\mu_l = 0$, $\mu_g = 2 \times 10^{-21}$); other parameter values are as in Figure 3. A corresponds to the decrease of between-population divergence with increasing recombination-fragment size h. The graph can be divided into four regions, representative of the relative effects of different key events on the divergence. The graph is based on the solution of Equation 7.

The conditions of region III, $\sigma_g P_{gs} \gg \sigma_g P_{gd} \sim c_\delta$, yield

$$E(t_s) \approx \frac{4c_\delta + \sigma_g P_{gd}}{\sigma_g P_{gs}(2c_\delta + \sigma_g P_{gd})} \quad (13)$$

$$E(t_d) \approx \frac{1}{2c_\delta + \sigma_g P_{gd}}. \quad (14)$$

Region III reflects the increasing significance of global periodic selection. Divergence between populations is determined by the combined homogenizing effects of recombination (c_δ) and global periodic selection ($\sigma_g P_{gd}$).

Region IV corresponds to the case of extremely rare recombination, $\sigma_g P_{gs} \gg \sigma_g P_{gd} \gg c_\delta$, yielding

$$E(t_s) \approx \frac{1}{\sigma_g P_{gs}} \quad (15)$$

$$E(t_d) \approx \frac{1}{\sigma_g P_{gd}}. \quad (16)$$

This is the limiting case, where recombination between populations becomes so infrequent that its effects are entirely overwhelmed by periodic selection. In this limit, the divergence within populations is determined solely by the intensity of *local* purging of diversity, while the divergence between populations is only limited by the intensity of *global* purging of diversity.

Under the conditions of rare recombination (region IV), the ratio of the times to coalescence (*i.e.*, $E[t_d]$:$E[t_s]$) approaches $1/h$. This follows from two consequences of rare recombination. First, because the locus of interest and the adaptive mutation are rarely separated by recombination, a selective sweep almost certainly leads to coalescence of gene segments from the same population (*i.e.*, $P_{gs} \approx 1$). Second, the transmission of the adaptive mutation from population 1 to population 2 is likely to be the result of a single transfer event. Hence, the probability of coalescence of two gene segments from different populations (P_{gd}) approaches the probability that the transfer event was a cotransfer of the adaptive mutation and the segment of interest (averaged over all distances between the two loci). That is, $P_{gd} \approx h$, and $E[t_d]/E[t_s] \approx 1/h$.

Effect of recombination fragment size on population divergence: We consider next the effect of the recombination fragment size (h) on the distinctness of ecological populations. In general, larger recombination fragments increase the probability (q) that a gene of interest will cotransfer across populations with an adaptive mutation (Equation 1), thus fostering coalescence of segments between populations (Figure 2). Hence, larger sizes of recombination fragments tend to make ecological populations appear less distinct in neutral characters (A, Figure 5). The effect of h on the distinctness of populations is most important at low between-population recombination rates (Figure 5).

The effect of h on population distinctness (quantified as $E[\pi_d]/E[\pi_s]$) is shown explicitly in Figure 6. Under very low rates of between-population recombination, the distinctness ratio approaches $1/h$ for large fragment sizes (*i.e.*, $h > 10\%$; Figure 6). Thus, global periodic selection alone (*i.e.*, with little recombination between populations) cannot reduce the distinctness ratio of populations to 1 (so that $E[\pi_d] \approx E[\pi_s]$ unless the recombination fragment size reaches 100% of the genome.

We explored in more detail the effect of h on population distinctness for the case when within-population divergence levels are 1% (*i.e.*, $E[\pi_s] = 0.01$), because this is the divergence level frequently observed within bacterial sequence-similarity clusters (Palys *et al.* 1997).

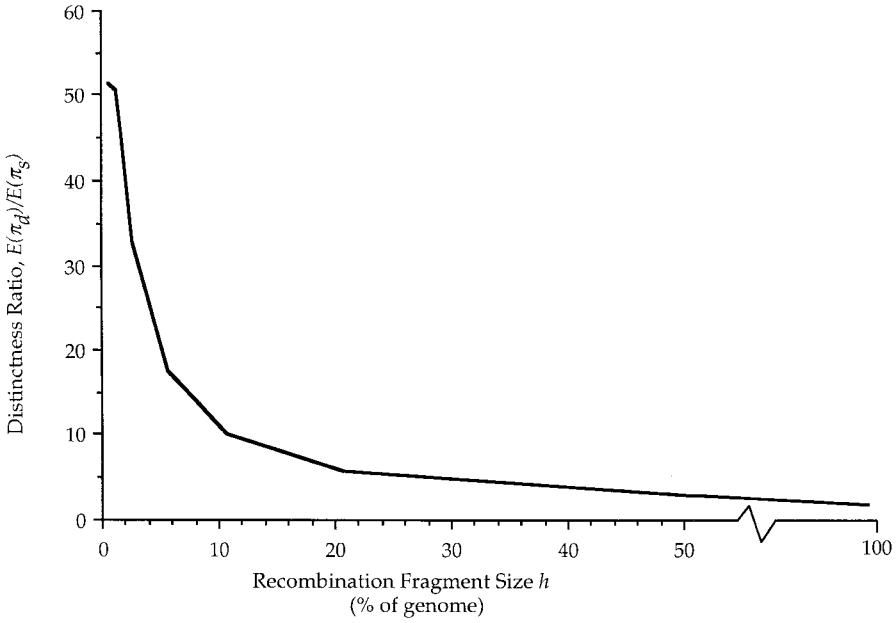

Figure 6.—The distinctness of populations as a function of the size of recombination fragment *h*. The distinctness of populations is quantified as the ratio of between:within population divergence ($E[\pi_d]/E[\pi_s]$). Note that for large fragment sizes ($h > 10\%$), the distinctness ratio approaches $1/h$. All selective sweeps are global ($\mu_l = 0$, $\mu_g = 2 \times 10^{-21}$); other parameters are as in Figure 3. The rate of between-population recombination is set to $c_\delta = c_s = 10^{-12}$, a value too low to produce a homogenizing effect on the metapopulation, while sufficiently high to allow globally adaptive mutations to transfer across populations. Divergence levels within and between populations are based on the solution of Equation 7; P_{gs} and P_{gd} are calculated from N, c_s, q, and z, using Equations 2 and A4.

With this level of divergence, global periodic selection is quite ineffective in reducing divergence between populations when recombination fragments are small (Figure 7). For example, global periodic selection with a recombination fragment of 1% of the genome cannot reduce the between-population divergence by >4%; however, with larger recombination fragments (*e.g.*, $h =$ 5%), global periodic selection may significantly reduce the between-population divergence from unbounded neutral divergence ($\pi_d = 0.25$) to a much more limited level of divergence ($\pi_d = 0.09$; Figure 7).

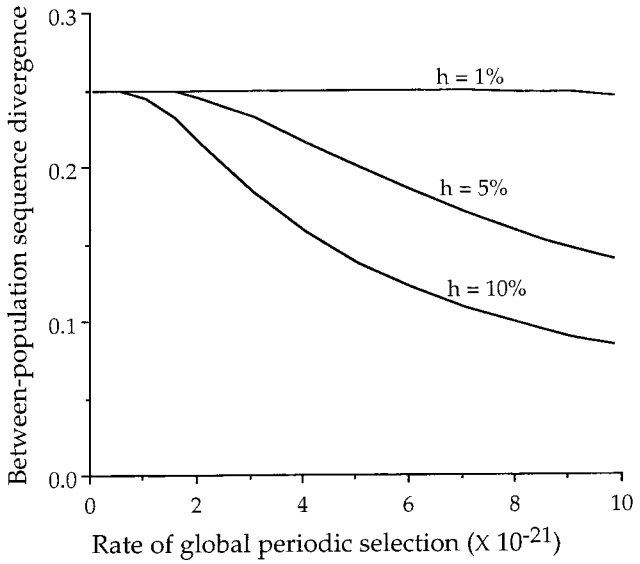

Figure 7.—Expected sequence divergence between populations ($E[\pi_d]$), as a function of the rate of globally adaptive mutations (μ_g), when the within-population divergence level is set at $E(\pi_s) = 0.01$. Note that for small sizes of the recombining fragment (*i.e.*, $h = 1\%$) the divergence between populations is virtually unaffected by the rate of globally adaptive mutations. For each value of μ_g, the rate μ_l was chosen to yield $E(\pi_s) = 0.01$ (using the solution of Equation 7). Other parameter values are as in Figure 6.

DISCUSSION

This study presents a coalescence model for investigating the effect of globally adaptive mutations on neutral sequence divergence in bacteria. We used this model to test whether interpopulation transfer of globally adaptive mutations might prevent neutral sequence divergence between ecologically distinct populations of bacteria.

Assumptions of the model: If globally adaptive mutations are to reduce divergence between ecological populations at every locus in the genome, we must assume that every gene locus has the opportunity to hitchhike from population to population along with globally adaptive mutations (Figure 1). We therefore assume that globally adaptive mutations that confer benefits in more than one population exist, that they are numerous, and that they appear throughout the genome. The latter two assumptions are required because only a limited fraction of the genome can be cotransferred (and subsequently homogenized) across populations with any given adaptive mutation: the segments transferred in bacterial recombination are generally small (Smith 1988), and the transfer of large segments across populations is probably disfavored by natural selection (Cohan 1994b; Zawadzki and Cohan 1995).

Consider next the central premise of the model, that globally adaptive mutations exist and are numerous. The likelihood of globally adaptive mutations must de-

pend on the degree of ecological divergence between populations. In the early stages of population divergence, a mutation that is adaptive in one population is likely to be adaptive in others. As the populations become progressively more finely tuned to their respective niches, accumulating many niche-specific adaptations, we should see fewer adaptive mutations that can benefit more than one population. We therefore expect globally adaptive mutations to prevent neutral sequence divergence genome-wide only between the most closely related populations.

Does a typical adaptive mutation confer a benefit in more than one population? Recently, Guttman and Dykhuizen (1994b) provided evidence that one adaptive mutation precipitated selective sweeps in all the ecological populations included within *Escherichia coli*. A selective sweep apparently purged sequence diversity within a small chromosomal region from all the various sequence clusters of *E. coli*, while these clusters retained their distinctness for all other chromosomal regions studied. This is exactly the pattern expected soon after a global selective sweep. As shown in Figure 2, for genes that are closely linked to the adaptive mutation ($q \approx 1$), there is nearly total purging of diversity both within and between populations; for genes that are not linked to the adaptive mutation ($q \approx 0$), there is purging of diversity within populations but none between populations. Provided that each of the *E. coli* sequence clusters is actually a separate ecological population (Cohan 1994a,b, 1999; Palys *et al.* 1997), the selective sweep demonstrated by Guttman and Dykhuizen (1994b) appears to have been driven by a globally adaptive mutation.

Globally adaptive mutations as a homogenizing force in neutral sequence evolution: Analysis of our model has shown that, in general, globally adaptive mutations tend to make populations less distinct. Especially under extremely low recombination rates, globally adaptive mutations severely depress neutral sequence divergence between populations while having only a minor effect on within-population diversity (Figure 3). Populations that would diverge without bound in the absence of global periodic selection may be prevented from diverging without bound in the presence of global periodic selection.

Nevertheless, global periodic selection does not homogenize neutral sequence divergence to the extent that populations become indistinguishable. Consider, for example, ecological populations whose average within-population sequence divergence is ~1%, a value typical for sequence-similarity clusters in bacteria (Palys *et al.* 1997). In the absence of global periodic selection, such populations diverge into separate sequence-similarity clusters whenever the between-population recombination rate is $<10^{-7.6}$; in the presence of global periodic selection, even for a large recombination fragment ($h = 10\%$), the critical recombination rate decreases only slightly to $10^{-7.8}$ (Figure 4). Recombination rates between most bacterial populations are unlikely to exceed either critical value (Whittam and Ake 1993; Roberts and Cohan 1995; Palys *et al.* 1997; Cohan 1999). We therefore conclude that in spite of the homogenizing effect of global periodic selection, ecological populations should diverge into separate sequence-similarity clusters.

Analysis of the model has shown that the effect of global adaptations on between-population divergence is highly dependent on the size of the fragment recombined (Figure 7). If the recombination fragment is small (<1% of the genome), global periodic selection is virtually ineffective in reducing between-population divergence; however, if the recombining fragment is large (*e.g.*, 5% of the genome), global periodic selection may significantly reduce the between-population divergence (Figure 7).

The effect of global periodic selection on sequence divergence may therefore depend on the mode of genetic transfer between populations, because the various modes of transfer differ greatly in the length of DNA recombined. In naturally competent taxa, such as Streptoccus and Bacillus, transformation may be the predominant mode of DNA exchange. The average fragment of DNA incorporated in both Streptococcus and Bacillus transformation is <1% of the genome (Humbert *et al.* 1995; Zawadzki and Cohan 1995). To the extent that transformation is the primary mode of transferring adaptive mutations across populations in these taxa, global periodic selection should have virtually no effect on sequence divergence (Figure 7).

Other modes of recombination, such as transduction and conjugation, can transfer much larger segments of DNA. A generalized transducing phage can, in principle, transfer segments as large as the phage's own genome, which could be ~10% of the bacterium's genome (Fraenkel-Conrat 1985; Arber 1994). Conjugating plasmids can transfer even larger segments: in the case of the *Hfr* plasmid of *E. coli*, most of the genome can be transferred (but there may be additional fitness constraints on the size of large transferred fragments; see above). Therefore, when transduction and conjugation are the principal means of transfer of adaptive mutations, global periodic selection can have an important role in reducing divergence between populations.

In summary, global periodic selection can limit the sequence divergence between ecological populations. The effect of global periodic selection is most pronounced for groups of populations with low between-population recombination, such that global periodic selection is the only constraint on divergence between populations. Global periodic selection is unlikely to prevent the divergence of ecological populations into separate sequence clusters. A quantitative prediction of the homogenizing effect of global periodic selection would require more information about the rate of mutations that confer adaptations in multiple populations, infor-

mation about how evenly globally adaptive mutations are distributed throughout the genome, and information about the size of fragments that can be transferred between populations and then successfully accommodated by the receiving population.

We thank Michael Feldgarden for suggesting that we explore a model of global periodic selection and Richard Hudson for suggesting important improvements to the model. This work was supported by Environmental Protection Agency grants R82-1388-010 and R82-5348-010 and by research funds from Wesleyan University.

LITERATURE CITED

Amann, R. I., W. Ludwig and K. H. Schleifer, 1995 Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:** 143–169.

Arber, W., 1994 Bacteriophage transduction, pp. 107–113 in *Encyclopedia of Virology*, edited by R. C. Webster and A. Granoff. Academic Press, London.

Atwood, K. C., L. K. Schneider and F. J. Ryan, 1951 Periodic selection in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **37:** 146–155.

Balmelli, T., and J. Piffaretti, 1996 Analysis of the genetic polymorphism of *Borrelia burgdorferi* sensu lato by multilocus enzyme electrophoresis. Int. J. Syst. Bacteriol. **46:** 167–172.

Boivin-Jahns, V., R. Ruimy, A. Bianchi, S. Daumas and R. Christen, 1996 Bacterial diversity in a deep-subsurface clay. Appl. Environ. Microbiol. **62:** 3405–3412.

Britschgi, T. B., and S. J. Giovannoni, 1991 Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. **57:** 1707–1713.

Cohan, F. M., 1994a The effects of rare but promiscuous genetic exchange on evolutionary divergences in prokaryotes. Am. Nat. **143:** 965–986.

Cohan, F. M., 1994b Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol. Evol. **9:** 175–180.

Cohan, F. M., 1995 Does recombination constrain neutral divergence among bacterial taxa? Evolution **49:** 164–175.

Cohan, F. M., 1996 The role of genetic exchange in bacterial evolution. ASM News **62:** 631–636.

Cohan, F. M., 1999 Genetic structure of prokaryotic populations, pp. 475–489 in *Evolutionary Genetics from Molecules to Morphology*, edited by R. Singh and K. Krimbas. Cambridge University Press, New York (in press).

Fraenkel-Conrat, H., 1985 *The Viruses: Catalogue, Characterization, and Classification*. Plenum, New York.

Guttman, D. S., and D. E. Dykhuizen, 1994a Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266:** 1380–1383.

Guttman, D. S., and D. E. Dykhuizen, 1994b Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics **138:** 993–1003.

Huber, R., S. Burggraf, T. Mayer, S. M. Barns, P. Rossnagel *et al.*, 1995 Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. Nature **376:** 57–58.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Humbert, O., M. Prudhomme, R. Hakenbeck, C. G. Dowson and J. Claverys, 1995 Homeologous recombination and mismatch repair during transformation in *Streptococcus pneumoniae*: saturation of the Hex mismatch repair system. Proc. Natl. Acad. Sci. USA **92:** 9052–9056.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kaplan, N. L., T. Darden and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

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

Knight, I. T., W. E. Holben, J. M. Tiedje and R. R. Colwell, 1992 Nucleic acid hybridization techniques for detection, identification, and enumeration of microorganisms in the environment, pp. 65–91 in *Microbial Ecology: Principles, Methods, and Applications*, edited by M. A. Levin, R. J. Seidler and M. Rogul. McGraw-Hill, New York.

Koch, A. L., 1974 The pertinence of the periodic selection phenomenon to prokaryotic evolution. Genetics **77:** 127–142.

Levin, B. R., 1981 Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. Genetics **99:** 1–23.

Maynard-Smith, J., N. H. Smith, M. O'Rourke and B. G. Spratt, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:** 4348–4388.

Murray, R. G. E., and E. Stackebrandt, 1995 Taxonomic note: implementation of the status *Canditatus* for incompletely described procaryotes. Int. J. Syst. Bacteriol. **45:** 186–187.

Normand, P., S. Orso, B. Cournoyer, P. Jeannin, C. Chapelon *et al.*, 1996 Molecular phylogeny of the genus *Frankia* and related genera and emendation of the family Frankiaceae. Int. J. Syst. Bacteriol. **46:** 1–9.

Ohkuma, M., and T. Kudo, 1996 Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. Appl. Environ. Microbiol. **62:** 461–468.

Pace, N. R., 1997 A molecular view of microbial diversity and the biosphere. Science **276:** 734–740.

Palys, T., L. K. Nakamura and F. M. Cohan, 1997 Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. Int. J. Syst. Bacteriol. **47:** 1145–1156.

Roberts, M. S., and F. M. Cohan, 1995 Recombination and migration rates in natural populations of *Bacillus subtilis* and *Bacillus mojavensis*. Evolution **49:** 1081–1094.

Selander, R. K., and J. M. Musser, 1990 Population genetics of bacterial pathogenesis, pp. 11–36 in *Molecular Basis of Bacterial Pathogenesis*, edited by B. H. Iglewski and V. L. Clark. Academic Press, San Diego.

Smith, G. R., 1988 Homologous recombination in procaryotes. Microbiol. Rev. **52:** 1–28.

Vandamme, P., B. Pot, M. Gillis, P. De Vos, K. Kersters *et al.*, 1996 Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Rev. **60:** 407–438.

Whittam, T. S., and S. E. Ake, 1993 Genetic polymorphisms and recombination in natural populations of *Escherichia coli*, pp. 223–245 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, edited by N. Takahata and A. G. Clark. Japan Scientific Society Press, Tokyo.

Wolfram, S., 1991 *Mathematica®: A System for Doing Mathematics by Computer*, Ed. 2. Addison-Wesley, Reading, MA.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Zawadzki, P., and F. M. Cohan, 1995 The size and continuity of DNA segments integrated in *Bacillus* transformation. Genetics **141:** 1231–1243.

Communicating editor: R. R. Hudson

APPENDIX A: Probability That a Periodic Selection Event Leads to Coalescence

We consider the special case of a metapopulation consisting of two ecological populations. The adaptive mutation driving the periodic selection event begins in population 1 and is subsequently passed into population 2 by recombination. We use a two-locus, four-allele model. A is the locus under selection, while B is the segment of interest whose neutral sequence divergence we are investigating. Alleles in population 1 are designated by subscript 1; those in population 2 are designated by subscript 2. Within population 1, the advantageous allele is designated as A_1, and all other alleles at the selected locus are designated a_1. A_2 is the advantageous allele in population 2; a_2 designates all the other alleles at the selected locus in this population. An allele

at locus B can be attached to any of the four A alleles, *i.e.*, A_1, a_1, A_2, a_2. The frequencies of the alleles A_1 and A_2 in their respective populations are x_1 and x_2.

Let $g_X(Y,t)$ be the conditional probability that if a randomly selected gene B from generation t of the metapopulation is attached to the allelic type Y (at locus A), its ancestor in generation $(t-1)$ was attached to allelic type X. [$g_X(Y,t)$ is equivalent to the quantity $f_X(Y,t)/f(Y,t)$ of Hudson and Kaplan (1988).] Following Hudson and Kaplan (1988), and keeping only the highest order terms in N (where N is the population size), 8 of the 16 g probabilities are

$$g_{A_1}(A_1,t) = 1 + O\left(\frac{1}{N}\right)$$

$$g_{A_1}(a_1,t) = \frac{x_1(t-1)R_{11}}{N} + O\left(\frac{1}{N^2}\right)$$

$$g_{A_1}(A_2,t) = \frac{x_1(t-1)R_{12}}{2N} + \frac{x_1(t-1)R_2}{Nx_2(t-1)} + O\left(\frac{1}{N^2}\right)$$

$$g_{A_1}(a_2,t) = \frac{x_1(t-1)R_{12}}{2N} + O\left(\frac{1}{N^2}\right)$$

$$g_{a_1}(a_1,t) = 1 + O\left(\frac{1}{N}\right)$$

$$g_{a_1}(A_1,t) = \frac{(1-x_1(t-1))R_{11}}{N} + O\left(\frac{1}{N^2}\right)$$

$$g_{a_1}(A_2,t) = \frac{(1-x_1(t-1))R_{12}}{2N} + O\left(\frac{1}{N^2}\right)$$

$$g_{a_1}(a_2,t) = \frac{(1-x_1(t-1))R_{12}}{2N} + \frac{(1-x_1(t-1))R_2}{N(1-x_2(t-1))} + O\left(\frac{1}{N^2}\right),$$

where $R_{11} = R_{22} = 2Nc_s(1-q)$, the per-population rate at which the two loci are separated by recombination within a population; $R_{12} = R_{21} = 2Nc_\delta(1-q)$, the per-population rate at which the two loci are separated by recombination between populations; and $R_1 = R_2 = Nc_\delta q$, the per-population rate at which a DNA segment containing both loci is transferred between populations. The remaining 8 g values may by obtained by substituting 2 for 1 and 1 for 2 in all the indices of the above equations, *e.g.*,

$$g_{A_2}(a_2,t) = \frac{x_2(t-1)R_{22}}{N} + O\left(\frac{1}{N^2}\right).$$

We now define the Q process. Suppose that m B genes are selected at random at the end of the selective sweep (time $t = 0$). Let $Q(0) = (i, j, k, l)$, where i, j, k, l represent the number of B genes attached to A_1, a_1, A_2, a_2, respectively. Going back in time, $Q(t)$ describes the number of ancestral B genes attached to each A allele at time t (*i.e.*, t generations before time 0). The total number of ancestral B genes in generation t is denoted by $|Q(t)|$. Note that $|Q(t)|$ never increases, because the number of ancestral alleles can only stay constant or decrease (if two or more of the sampled alleles had a common ancestor in the previous generation). We are interested in the cases where $Q(t)$ changes states, *i.e.*, $Q(t-1) \neq Q(t)$. There are two possible cases.

Case 1. $|Q(t-1)| = |Q(t)|$: The only possible state changes allowed by this condition result from recombination between parental genes. Given $Q(t) = (i, j, k, l)$, there are 12 possible states of $Q(t-1)$:

$\{(i+1, j-1, k, l); (i+1, j, k-1, l); (i+1, j, k, l-1);$
$(i, j+1, k-1, l); (i, j+1, k, l-1); (i-1, j+1, k, l);$
$(i, j, k+1, l-1); (i-1, j, k+1, l); (i, j-1, k+1, l);$
$(i-1, j, k, l+1); (i, j-1, k, l+1); (i, j, k-1, l+1)\}.$

Note that all other jumps would require more than a single recombination event and their probabilities are therefore of the order $1/N^2$ and are negligible. We are interested in the probabilities that the process jumps from (i, j, k, l), to any of the above states, *e.g.*,

$$P[Q(t-1) = (i+1, j-1, k, l) \mid Q(t) = (i, j, k, l)].$$

The probability that a selected gene B from generation t is attached to a_1 while its ancestor was attached to A_1 is given by $g_{A_1}(a_1,t)$. Because we are sampling j a_1 alleles,

$$P[Q(t-1) = (i+1, j-1, k, l) \mid Q(t) = (i, j, k, l)] = j \cdot g_{A_1}(a_1,t).$$

Equations of the same form can be obtained for the remaining 11 jumps.

Case 2. $|Q(t-1)| \neq |Q(t)|$: We have already noted that the number of ancestral alleles can only decrease going backward in time. This case implies that some of the genes sampled at time t must have a common ancestor at time $(t-1)$. Kaplan *et al.* (1988) have shown that the probability that two genes of a particular allelic type at time t have a common ancestor in generation $(t-1)$ is, to the first order in N, given by the probability of coalescence by drift, *i.e.*,

$$\frac{1}{Nx_a(t-1)} + O\left(\frac{1}{N^2}\right),$$

where $x_a(t-1)$ is the frequency of the allelic type a in the parental generation. Thus, the probability that two B alleles attached to A_1 at t have a common ancestor at $(t-1)$ is

$$\frac{1}{Nx_1(t-1)} + O\left(\frac{1}{N^2}\right).$$

Because i B genes attached to A_1 are being sampled,

$$P[Q(t-1) = (i-1, j, k, l) \mid Q(t) = (i, j, k, l)]$$

$$= \binom{i}{2}\left[\frac{1}{Nx_1(t-1)} + O\left(\frac{1}{N^2}\right)\right],$$

where if $i < 2$, $\binom{i}{2}$ is interpreted as 0. Similarly,

$$P[Q(t-1) = (i, j-1, k, l) \mid Q(t) = (i, j, k, l)] = \binom{j}{2}\left[\frac{1}{N(1-x_1(t-1))} + O\left(\frac{1}{N^2}\right)\right]$$

$$P[Q(t-1) = (i, j, k-1, l) \mid Q(t) = (i, j, k, l)] = \binom{k}{2}\left[\frac{1}{Nx_2(t-1)} + O\left(\frac{1}{N^2}\right)\right]$$

$$P[Q(t-1) = (i, j, k, l-1) \mid Q(t) = (i, j, k, l)] = \binom{l}{2}\left[\frac{1}{N(1-x_2(t-1))} + O\left(\frac{1}{N^2}\right)\right],$$

and because the chance of more than one coalescence event per generation is of the order of $1/N^2$, jumps of >1 state are ignored. We have now defined the probabilities of every possible state change of $Q(t)$. The probability that the Q process does not change state can thus be written as

$$P(Q(t-1) = Q(t) \mid Q(t) = (i, j, k, l)) = 1 - h_{ijkl}(x_1(t-1), x_2(t-1)) + O\left(\frac{1}{N^2}\right),$$

where h_{ijkl} is the total probability of the Q process changing states:

$$\begin{aligned} h_{ijkl}(x_1(t-1), x_2(t-1)) = {} & \frac{\binom{i}{2}}{Nx_1(t-1)} + \frac{\binom{j}{2}}{N(1-x_1(t-1))} \\ & + \frac{\binom{k}{2}}{Nx_2(t-1)} + \frac{\binom{l}{2}}{N(1-x_2(t-1))} \\ & + i\,(g_{a_1}(A_1,t) + g_{A_2}(A_1,t) + g_{a_2}(A_1,t)) \\ & + j\,(g_{A_1}(a_1,t) + g_{A_2}(a_1,t) + g_{a_2}(a_1,t)) \\ & + k\,(g_{A_1}(A_2,t) + g_{a_1}(A_2,t) + g_{a_2}(A_2,t)) \\ & + l\,(g_{A_1}(a_2,t) + g_{a_1}(a_2,t) + g_{A_2}(a_2,t)). \end{aligned}$$

Calculation of p_{gd} and p_{gs}: At the end of the selective sweep we sample two *B* alleles from the metapopulation. We want to know the probability that the two alleles had a common ancestor during the selective sweep. We consider two cases:

Case 1: the two genes are sampled from different ecological populations. The probability of coalescence during the sweep is p_{gd}.

Case 2: the two genes are sampled from the same ecological population. The probability of coalescence during the sweep is p_{gs}.

We begin by considering case 1, calculation of p_{gd}. We follow the Q process back in time, going from $t = \tau_f$ (end of the selective sweep) to $t = \tau_b$ (beginning of selective sweep; Figure 8). We can write $(1 - p_{gd})$ as the probability of escaping coalescence, *i.e.*, leaving the selective sweep ($t = \tau_f$) with one *B* gene in population 1 (attached to A_1) and one *B* gene in population 2 (attached to A_2) and entering it ($t = \tau_b$) with two ancestral *B* genes:

$$1 - p_{gd} = P[Q(\tau_b) = (i, j, k, l) \mid Q(\tau_f) = (1,0,1,0)] \quad (i + j + k + l) = 2. \tag{A1}$$

We also need to define $P_{ijkl}(t)$, the probability of finding the Q process in the state (i, j, k, l) at time (t), given that at the end of the selective sweep $Q(\tau_f) = (1,0,1,0)$,

$$P_{ijkl}(t) = P[Q(t) = (i, j, k, l) \mid Q(\tau_f) = (1,0,1,0)].$$

Hence, we can rewrite Equation A1 as

$$1 - p_{gd} = \sum_{i+j+k+l=2} P_{ijkl}(\tau_b). \tag{A2}$$

To calculate all the relevant $P_{ijkl}(\tau_b)$ values, we use the differential equations governing their behavior:

$$\begin{aligned} \frac{dP_{1100}}{dt} = {} & -h_{1100}P_{1100} \\ & + 2P_{2000}\cdot g_{a_1}(A_1,t) + 2P_{0200}\cdot g_{A_1}(a_1,t) \\ & + P_{1010}\cdot g_{a_1}(A_2,t) + P_{1001}\cdot g_{a_1}(a_2,t) \\ & + P_{0110}\cdot g_{A_1}(A_2,t) + P_{0101}\cdot g_{A_1}(a_2,t) \end{aligned}$$

$$\begin{aligned} \frac{dP_{1010}}{dt} = {} & -h_{1010}P_{1010} \\ & + 2P_{2000}\cdot g_{A_2}(A_1,t) + 2P_{0020}\cdot g_{A_1}(A_2,t) \\ & + P_{1100}\cdot g_{A_2}(a_1,t) + P_{1001}\cdot g_{A_2}(a_2,t) \\ & + P_{0110}\cdot g_{A_1}(a_1,t) + P_{0011}\cdot g_{A_1}(a_2,t) \end{aligned}$$

$$\begin{aligned} \frac{dP_{1001}}{dt} = {} & -h_{1001}P_{1001} \\ & + 2P_{2000}\cdot g_{a_2}(A_1,t) + 2P_{0002}\cdot g_{A_1}(a_2,t) \\ & + P_{1100}\cdot g_{a_2}(a_1,t) + P_{1010}\cdot g_{a_2}(A_2,t) \\ & + P_{0101}\cdot g_{A_1}(a_1,t) + P_{0011}\cdot g_{A_1}(A_2,t) \end{aligned}$$

$$\begin{aligned} \frac{dP_{0110}}{dt} = {} & -h_{0110}P_{0110} \\ & + 2P_{2000}\cdot g_{a_2}(A_1,t) + 2P_{0002}\cdot g_{A_1}(a_2,t) \\ & + P_{1100}\cdot g_{a_2}(a_1,t) + P_{1010}\cdot g_{a_2}(A_2,t) \\ & + P_{0101}\cdot g_{A_1}(a_1,t) + P_{0011}\cdot g_{A_1}(a_2,t) \end{aligned}$$

$$\frac{dP_{0101}}{dt} = -h_{0101}P_{0101}$$

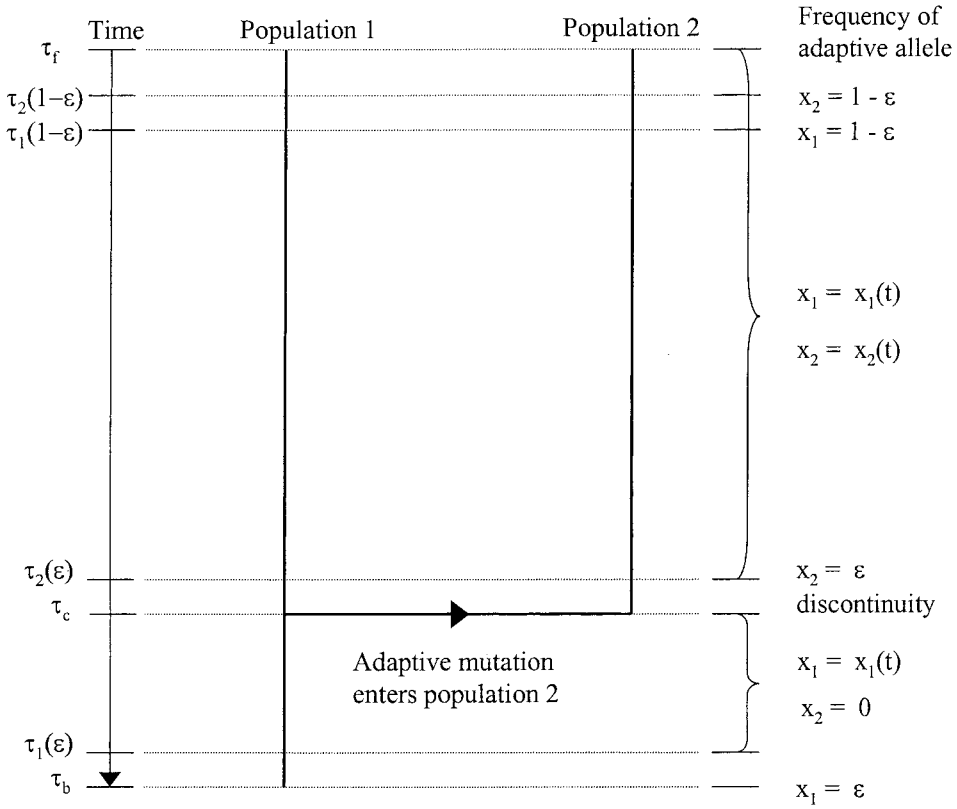

Figure 8.—The course of a global selective sweep. The adaptive mutation occurs in population 1 at time $\tau_{b'}$ is transferred to population 2 at $\tau_{c'}$ and completes the selective sweep of the metapopulation at τ_f. Shown are the intervals where the differential Equations A3 may be treated deterministically, $(\tau_1(\varepsilon), \tau_c)$ and $(\tau_2(\varepsilon), \tau_2(1-\varepsilon))$ and the discontinuity $(\tau_c, \tau_2(\varepsilon))$, where additional boundary conditions are required $(\varepsilon = 5/Nz)$. The values x_1 and x_2 are the frequencies of the adaptive allele in populations 1 and 2, respectively.

$$
\begin{aligned}
&+ 2P_{2000} \cdot g_{a_2}(A_1,t) + 2P_{0002} \cdot g_{A_1}(a_2,t)\\
&+ P_{1100} \cdot g_{a_2}(a_1,t) + P_{1010} \cdot g_{a_2}(A_2,t)\\
&+ P_{0101} \cdot g_{A_1}(a_1,t) + P_{0011} \cdot g_{A_1}(a_2,t)\\
\frac{dP_{0011}}{dt} &= -h_{0011}P_{0011}\\
&+ 2P_{2000} \cdot g_{a_2}(A_1,t) + 2P_{0002} \cdot g_{A_1}(a_2,t)\\
&+ P_{1100} \cdot g_{a_2}(a_1,t) + P_{1010} \cdot g_{a_2}(A_2,t)\\
&+ P_{0101} \cdot g_{A_1}(a_1,t) + P_{0011} \cdot g_{A_1}(a_2,t)\\
\frac{dP_{2000}}{dt} &= -h_{2000}P_{2000} + P_{1100} \cdot g_{A_1}(a_1,t)\\
&+ P_{1010} \cdot g_{A_1}(A_2,t) + P_{1001} \cdot g_{A_1}(a_2,t)\\
\frac{dP_{0200}}{dt} &= -h_{0200}P_{0200} + P_{1100} \cdot g_{A_1}(a_1,t)\\
&+ P_{1010} \cdot g_{A_1}(A_2,t) + P_{1001} \cdot g_{A_1}(a_2,t)\\
\frac{dP_{0020}}{dt} &= -h_{0020}P_{0020} + P_{0110} \cdot g_{A_2}(a_1,t)\\
&+ P_{1010} \cdot g_{A_2}(A_1,t) + P_{0011} \cdot g_{A_2}(a_2,t)\\
\frac{dP_{0002}}{dt} &= -h_{0002}P_{0002} + P_{1001} \cdot g_{a_2}(A_1,t)\\
&+ P_{0101} \cdot g_{a_2}(a_1,t) + P_{0011} \cdot g_{a_2}(A_2,t). \qquad \text{(A3)}
\end{aligned}
$$

We also need the equations describing changes in the frequencies of the adaptive alleles A_1 and A_2, $x_1(t)$ and $x_2(t)$:

$$\frac{dx_1}{dt} = -zx_1(1 - x_1)$$

$$\frac{dx_2}{dt} = -zx_2(1 - x_2).$$

To be able to treat the model deterministically, we follow Kaplan *et al.* (1989) and limit t to

$$\tau_2(1 - \varepsilon) \le t \le \tau_1(\varepsilon),$$

where $\tau(\varepsilon)$ corresponds to the time in the early phase of the selective sweep, where $x_1 = \varepsilon$, and $\tau_2(1 - \varepsilon)$ corresponds to a time near the end of the selective sweep, where $x_2 = 1 - \varepsilon$. We take $\varepsilon = 5/Nz$.

It remains for us to establish the boundary conditions

$$P_{1010}(\tau_2(1 - \varepsilon)) = 1,$$

while all other $P_{ijkl}(\tau_2(1 - \varepsilon))$ are zero. Also,

$$x_2(\tau_2(1 - \varepsilon)) = 1 - \varepsilon.$$

Because the adaptive mutation enters population 2 later than it enters population 1, we must determine the value of x_1 at the time when $x_2 = 1 - \varepsilon$. For that purpose we need to first run the frequency equations forward in time (starting at $x_1 = \varepsilon$, $x_2 = 0$). We then allow for the transfer of a single adaptive mutation to

population 2. This takes place at time $E(\tau_c)$, which is the expectation of the transfer of the first adaptive allele that will become fixed in population 2. (This is in fact equal to the time at which $1/2z$ alleles have crossed over.) We run the equations forward in time until $\tau_2(1-\varepsilon)$ to establish the end boundary conditions for x_1; then we can run both x_1 and x_2 backward, along with the equations for P_{ijkl}, following Equation A3 (see Figure 8).

The transfer of the first adaptive allele from population 1 to population 2 is a stochastic event and hence introduces a discontinuity in the solutions to differential equations for P_{ijkl}. At the time immediately preceding the initial transfer of the adaptive allele into population 2 (time τ_c+), no A_2 alleles existed. Therefore, at time τ_c+ the probabilities of a B allele being attached to an A_2 allele [$P_{ijkl}(\tau_c+)$ for $k \neq 0$] must be zero. However, these P terms might have nonzero values at $\tau_2(\varepsilon)$. Following Kaplan *et al.* (1989), we may neglect the contribution of $P_{0020}(\tau_2[\varepsilon])$ to p_{gd}. However, a nonzero value of $P_{ijkl}(\tau_2[\varepsilon])$ implies that immediately after the transfer event, one of the sampled B genes is attached to an A_2 allele. In this case, there exist two possible states of the Q process immediately preceding the transfer: $Q(\tau_c+) = (i+1, j, 0, l)$ if the transfer was a corecombination of A and B (with a probability q); or $Q(\tau_c+) = (i, j, 0, l+1)$ if only the A allele was transferred (with a probability $q-1$). Hence, the additional boundary conditions at τ_c needed to account for the transfer are

$$\left.\begin{aligned} P_{ij0l}(\tau_c) &= P_{ij0l}(\tau_2(\varepsilon)) + q \cdot P_{i-1j1l}(\tau_2(\varepsilon)) \\ &\quad + (1-q) \cdot P_{ij1l-1}(\tau_2(\varepsilon)) \\ P_{ij1l}(\tau \geq \tau_c) &= P_{0020}(\tau \geq \tau_c) = 0 \end{aligned}\right\} \text{ for } i+j+l=1$$

$$x_2(\tau \geq \tau_c) = 0.$$

Using the above boundary conditions, we can obtain the values $P_{ijkl}(\tau_1[\varepsilon])$ at the beginning of the selective sweep. Then, using the approximation $P_{ijkl}(\tau_1[\varepsilon]) = P_{ijkl}(\tau_b)$, Equation A2 becomes

$$\begin{aligned} 1 - p_{gd} &= P_{1001}(\tau_1(\varepsilon)) + P_{1100}(\tau_1(\varepsilon)) \\ &\quad + P_{0200}(\tau_1(\varepsilon)) + P_{0002}(\tau_1(\varepsilon)) + P_{0101}(\tau_1(\varepsilon)). \end{aligned} \tag{A4}$$

The above analysis also applies to case 2, the calculation of p_{gs}. We only need to alter the initial conditions, *i.e.*, allowing $Q(\tau_f) = (2,0,0,0)$ (if we are sampling two alleles from the original population where the adaptive mutation first occurred) or $Q(\tau_f) = (0,0,2,0)$ (if we are sampling two alleles from the population to which the adaptive mutation has been transferred). Because the probabilities of choosing either population are equal, we calculate p_{gs} as the average of the two values.

APPENDIX B: The Expected Time to Coalescence

Coalescence within populations

Following Cohan (1994a), we first consider the time to coalescence t_s for two segments currently residing in cells of the same ecological population. The statistical properties of t_s are investigated by dividing t_s into two constituent quantities: the time k_s necessary to go back into the past of the two lineages to reach a "key event" (defined below) and the additional time necessary to go beyond the key event to reach a coalescence of the two lineages into a common ancestor (if the key event did not result in coalescence). A key event is any event that changes the expected time to coalescence. In the case of segments from the same ecological population, a key event may be any of the following: coalescence by drift acting on N cells of the population, a local selective sweep, a global selective sweep, or a genetic exchange event in which one of the segments is transferred into its present population from another ecological population.

The variable t_s is thus defined below, where the first term k_s represents the time to reach the most recent key event, and the other three terms represent the additional time necessary to go beyond the key event to reach a coalescence (Table 2):

$$\begin{aligned} t_s &= k_s + \chi_{\sigma l}(\chi_{pl} t_s' + \chi_{\gamma l} t_d') \\ &\quad + \chi_{\sigma g}(\chi_{pgs} t_s'' + \chi_{\gamma g} t_d'') + \chi_{cd} t_d'''. \end{aligned}$$

Local selective sweep as the key event: The random

TABLE 2

Indicator variables (for appendix b)

$\chi_{\sigma l}$	Indicator for whether the key event is a local selective sweep
$\chi_{\sigma g}$	Indicator for whether the key event is a global selective sweep
χ_{cd}	Indicator for whether the key event is a recombination event with any other population
χ_{pl}	Indicator for whether segments from the same population at the end of a local selective sweep (1) escape coalescence in the selective sweep and (2) begin the selective sweep in the same population
$\chi_{\gamma l}$	Indicator for whether segments from the same population at the end of a local selective sweep begin the selective sweep in different populations
χ_{pgs}	Indicator for whether segments from the same population at the end of a global sweep (1) escape coalescence and (2) begin the sweep in the same population
$\chi_{\gamma g}$	Indicator for whether segments from the same population at the end of a global selective sweep begin the selective sweep in different populations
χ_{pgd}	Indicator for whether segments from different populations at the end of a global sweep (1) escape coalescence and (2) begin the sweep in different populations
$\chi_{\gamma d}$	Indicator for whether segments from different populations at the end of a global selective sweep begin the selective sweep in the same population
$\chi_{c\delta}$	Indicator for whether the key event is a between-population recombination event, such that two segments in different populations after the event are in the same population as before

variable $\chi_{\sigma l}$ indicates whether the key event was a local selective sweep ($\chi_{\sigma l} = 1$ if the key event was a local selective sweep; $\chi_{\sigma l} = 0$ else). Two lineages that are in the same population at the end of the selective sweep may begin the sweep in one of three states: a single lineage (*i.e.*, the lineages coalesce); two lineages in the same population; or they may begin as two lineages in different populations. The random variable χ_{pl} indicates whether the lineages escape coalescence and begin in the same population (1 if yes, 0 else, as above), and t_s' is the additional time to coalescence in this case. The random variable $\chi_{\gamma l}$ indicates whether the lineages begin the selective sweep in different populations (1 if yes, 0 else), and t_d' is the additional time to coalescence in this case. Accordingly, the sum $(\chi_{pl}t_s' + \chi_{\gamma l}t_d')$ represents the additional time to coalescence when the key event is a local sweep.

Global periodic selection as the key event: The random variable $\chi_{\sigma g}$ indicates whether the key event was a global selective sweep (1 if yes, 0 else). The random variable χ_{pgs} indicates whether the lineages escape coalescence and begin the sweep in the same population (1 if yes, 0 else), and the random variable t_s'' represents the additional time to coalescence in this case. Similarly, $\chi_{\gamma g}$ indicates whether the lineages begin the selective sweep in different populations (1 if yes, 0 else), and t_d'' represents the additional time to coalescence in this case. The sum $(\chi_{pgs}t_s'' + \chi_{\gamma g}t_d'')$ represents the additional time to coalescence if the event is a global selective sweep.

Recombination between populations as the key event: The random variable $\chi_{c\,d}$ indicates whether the key event is a recombination between populations, in which one of the lineages enters the current population (1 if yes, 0 else). The random variable t_d''' indicates the additional time to coalescence beyond this key event.

Coalescence between populations

Next consider the time to coalescence, t_d, for segments that are now in two cells belonging to different ecological populations. As above, we divide t_d into the time k_d to go back to the most recent key event (a global selective sweep or a genetic exchange event in which one of the lineages enters its current population from the population of the other lineage) and the additional time required to go beyond the key event to reach coalescence:

$$t_d = k_d + \chi_{\sigma g}(\chi_{pgd}t_d'''' + \chi_{\gamma d}t_s''') + \chi_{c\delta}t_s''''.$$

Global selective sweep as the key event: The random variable χ_{pgd} indicates whether the two lineages presently in different populations escape coalescence and begin the sweep in different populations, and t_d'''' is a random variable for the additional time to coalescence in this case. The random variable $\chi_{\gamma d}$ indicates whether the two lineages escape coalescence and begin the sweep in the same population, and t_s''' represents the additional time to coalescence in this case. The sum $(\chi_{pgd}t_d'''' + \chi_{\gamma d}t_s''')$ represents the additional time to coalescence when the key event is a global selective sweep.

Between-population recombination as the key event: The random variable $\chi_{c\,\delta}$ indicates whether the key event was a recombination event between populations, such that before the event the lineages were in the same population and afterward were in different populations (1 if yes, 0 else). The random variable t_s'''' represents the additional time to coalescence in this case.

Expected values of t_s and t_d: The expected values of all indicator variables were calculated following Cohan (1994a), except when the key event is a selective sweep. The expectation of χ_{pl} is the probability that two segments from the same population escape coalescence and begin the sweep in the same population $(1 - P_l)$ and likewise for the expected values of other indicator variables: $E(\chi_{pgs}) = 1 - P_{gs}$; $E(\chi_{pgd}) = 1 - P_{gd}$; $E(\chi_{\gamma l}) = P_{lc}$. We also define $E(\chi_{\gamma g}) = P_{gsc}$ and $E(\chi_{\gamma d}) = P_{gdc}$. All the relevant expected values are calculated in appendix a.

The expected values of t_s and t_d are as follows:

$$\begin{aligned} E(t_s) = {} & [1 + \sigma_l((1 - P_l - P_{lc})E(t_s) + P_{lc}E(t_d)) \\ & + \sigma_g((1 - P_{gs} - P_{gsc})E(t_s) + P_{gsc}E(t_d)) \\ & + 2c_dE(t_d)]/(1/N + \sigma_l + \sigma_g + 2c_d) \\ E(t_d) = {} & [1 + \sigma_g((1 - P_{gd} - P_{gdc})E(t_d) \\ & + P_{gdc}E(t_s)) + 2c_\delta E(t_s)]/(\sigma_g + 2c_\delta). \end{aligned}$$

It can be shown that if the selective sweeps are rapid and infrequent (*i.e.*, duration of a sweep is much greater than the interval between sweeps) the above equations reduce to

$$\begin{aligned} E(t_s) = {} & [1 + \sigma_l(1 - P_l)E(t_s) + \sigma_g(1 - P_{gs})E(t_s) \\ & + 2c_dE(t_d)]/(1/N + \sigma_l + \sigma_g + 2c_d) \\ E(t_d) = {} & [1 + \sigma_g(1 - P_{gd})E(t_d) \\ & + 2c_\delta E(t_s)]/(\sigma_g + 2c_\delta). \end{aligned}$$

Probability density functions of t_s and t_d: We define $P_{ts}(t)$ as the probability that the time t_s to coalescence for two segments currently in the same population is equal to t. Similarly, $P_{td}(t)$ is the probability that the time t_d to coalescence for two segments currently in different populations is equal to t. The values $P_{ts}(t)$ and $P_{td}(t)$ can then be expressed as the probability that the most recent key event occurred at exactly time t and led to coalescence or, if the key event occurred at any other time $\tau < t$ and did not lead to coalescence, the additional time necessary to reach coalescence was $(t - \tau)$. We can consider the key events to be exponentially distributed (Hudson and Kaplan 1988). Hence,

$$\begin{aligned} P_{ts}(t) = {} & (1/N + \sigma_l(P_l) + \sigma_g(P_{gs})) \\ & \cdot e^{-(1/N+\sigma_l+\sigma_g+2c_d)t} \\ & + (\sigma_l(1 - P_l) + \sigma_g(1 - P_{gs})) \end{aligned}$$

$$\cdot \int_0^t e^{-(1/N+\sigma_l+\sigma_g+2c_d)\tau} P_{ts}(t-\tau)d\tau$$

$$+ 2c_d \int_0^t e^{-(1/N+\sigma_l+\sigma_g+2c_d)\tau} P_{td}(t-\tau)d\tau$$

and

$$P_{ts}(t) = \sigma_g P_{gd} \cdot e^{-(\sigma_g+2c_\delta)t}$$

$$+ \sigma_g(1-P_{gd}) \int_0^t e^{-(\sigma_g+2c_\delta)\tau} P_{td}(t-\tau)d\tau$$

$$+ 2c_\delta \int_0^t e^{-(\sigma_g+2c_\delta)\tau} P_{ts}(t-\tau)d\tau.$$

The above recursions were solved numerically to yield numerical representations of the two probability distribution functions.