Genetic hitch-hiking in a subdivided population

MONTGOMERY SLATKIN* AND THOMAS WIEHE
Department of Integrative Biology, University of California, Berkeley, CA 94720-3140, USA
(Received 13 August 1997 and in revised form 9 December 1997 and 13 January 1998)

Summary
The problem of genetic hitch-hiking in a geographically subdivided population is analysed under the assumption that migration rates among populations are relatively small compared with the selection coefficient for a newly arising advantageous allele. The approximate method used in the paper is valid when the number of emigrants per generation (Nm) is less than one. The approximate analysis shows that hitch-hiking can result in substantial differences among populations in the frequencies of neutral alleles closely linked to the advantageous allele. Thus, in cases for which genetic hitch-hiking is thought to be responsible for low levels of genetic variability in regions of the genome with restricted crossing over, it might be possible to find confirmatory evidence for that hypothesis by finding unusual patterns of geographic differentiation in the same regions of the genome.

1. Introduction
Genetic hitch-hiking is a term introduced by Maynard Smith & Haigh (1974) to describe the effect of a gene substitution at a selected locus on a closely linked neutral locus. Maynard Smith & Haigh found that heterozygosity would be reduced significantly at the neutral locus provided that the selection coefficient, s, is larger than the recombination rate, c, between the selected and neutral loci. These results have been confirmed and extended by Kaplan et al. (1988), Stephan et al. (1992) and Wiehe & Stephan (1993). Reduced heterozygosity in regions of low recombination in Drosophila, found by Begun & Aquadro (1991) and others, has been attributed to the effects of hitch-hiking, although Charlesworth et al. (1993) have proposed an alternative explanation.

The existing theory of hitch-hiking suggests that it is a process that leads to uniformity and homogeneity at neutral loci. That is true for a single population but, as suggested to us by C. H. Langley (personal communication), it is not necessarily true when considering a geographically subdivided population. We will show that under some conditions, hitch-hiking can lead to substantial population differentiation, as measured by Wright’s Fst, even when there is significant gene flow, in spite of the fact that hitch-hiking will reduce heterozygosity within each population. A complete model of this problem is intractable so we will begin with a simple model that illustrates our main point. We will then present some numerical analysis that provides a more detailed picture of the process.

2. Deterministic analysis of a single population
In this section, we follow the deterministic analysis of Maynard Smith & Haigh (1974). We assume that there is a haploid population and consider two diallelic loci with recombination rate c between them. At the selected locus, allele B has a selective advantage s over the alternative allele b. Initially, the population is fixed for b but a single B allele is introduced. We assume that selection is sufficiently strong that B increases deterministically in frequency:

$$p(t) = \frac{p_0(1+s)^t}{1-p_0+p_0(1+s)^t},$$

where $p_0$ is the initial frequency of B and t is the number of generations. The choice of $p_0$ is problematic. We assume that there is one copy initially, so it would seem reasonable to assume $p_0 = 1/N$, where N is the haploid population size. When there are only one or a few copies of B, however, stochastic events dominate.
ultimately goes to fixation is initially on a chromosome in frequency \( q \) (\( 1 - q \)). For our purposes, the dependence on \( p_0 \) is weak and is absorbed into \( q^* \) and \( q^{**} \) defined below, so the choice will not affect our conclusions.

At the neutral locus, we assume that one allele, \( A \), is initially in frequency \( q \) and that the copy of \( B \) that ultimately goes to fixation is initially on a chromosome carrying \( A \). In that case, Maynard Smith & Haigh (1974, eqn 8) show that \( q \) will increase to a value \( q^* \).

\[
q^* = 1 - c(1 - q)(1 - p_o) \sum_{n=0}^{\infty} \frac{(1 - c)^n}{1 - p_o + p_o(1 + s)^{n+1}}.
\]  

(2)

If \( c \) is sufficiently small, (2) can be approximated by

\[
q^* \approx 1 + \frac{c(1 - q)}{s} \ln (p_o)
\]  

(3)

(Maynard Smith & Haigh, 1974, eqn 14). As mentioned above, the value of \( p_0 \) affects \( q^* \), but (3) shows that for small \( c \) the dependence is only logarithmic.

If instead, \( B \) is initially on an \( A \) chromosome, then the frequency of \( A \) will be reduced from \( q \) to \( q^{**} \) given by

\[
q^{**} = cq(1 - p_o) \sum_{n=0}^{\infty} \frac{(1 - c)^n}{1 - p_o + p_o(1 + s)^{n+1}},
\]  

(4)

whose value also depends only weakly on \( p_o \).

### 3. Island model

We now consider a collection of \( d \) subpopulations in an island model of migration. In each generation, a fraction \( m \) of each subpopulation is replaced by immigrants drawn randomly from the \( d - 1 \) other subpopulations. We assume that \( m \) is sufficiently small that only one subpopulation is going to fixation at any time. In that case, we can treat each subpopulation as fixed either for \( B \) or for \( b \), and ignore the details of the fixation process in each subpopulation.

Assume first that the frequency of \( A \) in all subpopulations is initially \( q \) and that \( B \) is introduced by mutation on an \( A \)-bearing chromosome in one of the subpopulations. Substitution of \( B \) proceeds, after which the frequency of \( A \) is increased to \( q^* \) in that subpopulation. In that subpopulation, there are two kinds of chromosomes: \( AB \), in frequency \( q^* \), and \( aB \), in frequency \( 1 - q^* \). Assuming that emigrant chromosomes are chosen randomly, those are the probabilities that emigrants will carry each of those chromosomes.

If an \( AB \) chromosome arrives at a subpopulation that has not undergone substitution of \( B \), then the process will be the same as in the first subpopulation: \( B \) will be substituted and afterwards the frequency of \( AB \) chromosomes will be \( q^* \). If instead an \( aB \) chromosome arrives, the results will be different: \( B \) will be substituted and the frequency of \( AB \) chromosomes will be \( q^{**} \).

We can proceed to construct a simple model of the process by noting that there are three kinds of subpopulations: those in which \( b \) is fixed (type 1), those in which fixation of \( B \) resulted from the arrival of an \( AB \) chromosome (type 2), and those in which fixation of \( B \) resulted from the arrival of an \( aB \) chromosome (type 3).

We let \( x_i \) be the number of type 2 subpopulations and \( y_i \) be the number of type 3 subpopulations after the \( i \)th substitution. The number of type 1 subpopulations is \( d - i \) and initially \( x_1 = 1 \) and \( y_1 = 0 \). Although there appear to be two random variables, \( x_i \) and \( y_i \), only one is independent because \( y_i = i - x_i \). We can develop a simple model to predict the distribution of \( x_i \) by ignoring events between substitutions of \( B \) in different subpopulations. Given \( x_i \) and \( y_i \), the probability that a type 1 population becomes type 2 is \( (x_i q^* + y_i q^{**})/(x_i + y_i) \), and hence this is the probability that \( x_{i+1} = x_i + 1 \).

Treating \( x_i \) as a random variable with possible values \( 1, 2, \ldots, l \), we can describe its distribution by a vector \( g_i \), whose elements are the probabilities that \( x_i = 1, 2, \ldots, l \). We can then compute the elements of \( g_{i+1} \) by multiplying \( g_i \) by an \( i + 1 \) by \( i \) matrix, \( T_i \), whose non-zero elements reflect the two possible transitions for each value of \( x_i \):

\[
T_{i,j+1}^{(i)} = \frac{f q^* + (i - j) q^{**}}{i}, \quad T_{i,j}^{(i)} = \frac{j(1 - q^*) + (i - j) (1 - q^{**})}{i}.
\]  

(5)

In (5), \( T_{i,j+1}^{(i)} \) is the probability that the next fixation of \( B \) begins with an \( AB \) chromosome and \( T_{i,j}^{(i)} \) is the probability that it begins with an \( aB \) chromosome. By successive matrix multiplication, we obtain \( g_i \):

\[
g_d = T_i^{(d-1)} T_i^{(d-2)} \cdots T_i^{(1)} g_1,
\]  

(6)

where \( g_1 \) is a vector whose only element is 1, reflecting the initial condition we assumed, \( x_1 = 1 \).

Although \( g_i \) cannot be expressed in a simpler form, it is easy to evaluate (6) numerically. Note that the result depends only on \( q^* \), \( q^{**} \) and \( d \). Furthermore, because the elements of \( T_i \) are linear in \( j \), we can obtain a recursion equation for the expectation of \( x_i \):

\[
E(x_{i+1}) = q^{**} + 1 + \frac{q^* - q^{**}}{i} E(x_i),
\]  

(7)

where \( E(.) \) denotes expectation. The solution to (7) with the initial condition \( E(x_1) = 1 \) is

\[
E(x_i) = q^{**} \sum_{j=2}^{d} \pi_j + \pi_1,
\]  

(8)
In this case, rate, for the low migration model described in the text. In this case, \( q = 0.2, p_h = 0.001, d = 20 \) and \( s = 0.1 \).

where \( \pi_d = 1 \) and \( \pi_j = \Pi_{d,j}^k(1 + (q^* - q^{**})/k) \) for \( j < d \). Similar equations and solutions can be found for the variance and higher moments, but it is just as easy to use (6) to find the entire distribution.

In this simple model, hitch-hiking in space results in local differentiation of subpopulations whenever \( \chi_d \neq d \). We can quantify the extent of differentiation using Wright's \( F_{ST} \). We computed \( F_{ST} \) by taking the ratio of the expected value of the variance in the frequency of \( A \) across subpopulations to \( E(\chi_d)[1 - E(\chi_d)] \).

Fig. 1 shows some typical results: \( F_{ST} \) is quite large for very small values of \( c \) and decreases as \( c \) increases to \( s \). In this graph, \( F_{ST} \) has the value it would have immediately after \( B \) was substituted throughout the population. Subsequent gene flow would reduce \( F_{ST} \) to its equilibrium value under gene flow and genetic drift, \( 1/(1 + 4Nm) \). The approach to that equilibrium would be on a time scale of \( 1/m \) generations.

Our approximate analysis is valid when only one subpopulation is going to fixation of \( B \) at any time. We can determine the upper limit on \( m \) by recalling that the approximate time of fixation of a strongly selected allele is roughly \( 1/s \) generations. The number of emigrants from subpopulations fixed for \( B \) is of the order of magnitude of \( Nm \) and the probability that each emigrant copy of \( B \) goes to fixation in a subpopulation fixed for \( b \) is approximately \( 2s \). Thus, the time between successive fixations of \( B \) is approximately \( 1/(2Nm) \), which must be greater than \( 1/s \) for our approximate analysis to be valid. Thus, we require that \( 2Nm \) be less than 1. When \( 2Nm \) is much greater than 1, then intuition suggests that fixation of \( B \) in most subpopulations would result from emigrants from the subpopulation in which \( B \) was initially fixed. Little differentiation among subpopulations at the linked marker would result and the effect of genetic hitch-hiking would be roughly the same as in a single panmictic population.

If \( 2Nm \ll 1 \), then substantial differentiation among neutral loci would be expected even in the absence of hitch-hiking, provided that subpopulations had been relatively isolated for a long time. The effect modelled in this paper would augment differentiation and neutral loci linked to loci at which substitutions occurred and could also lead to differentiation more rapidly than would occur under the effects of drift alone.

4. Stepping stone model

If we imagine a one-dimensional array of subpopulations with gene flow only between adjacent subpopulations and assume that \( B \) is introduced by mutation at one end, then the analysis in the previous section allows us to make a relatively simple prediction about what will happen. In the first population, hitch-hiking will increase the frequency of \( A \) to \( q^* \), which is also the probability that a successful \( AB \) chromosome will arrive at the second subpopulation. The process will continue until the successful chromosome arriving at the next subpopulation is \( aB \). The probability that the first \( i + 1 \) subpopulations will have \( A \) in frequency \( q^* \) is a geometric distribution with parameter \( q^* : \Pr(i + 1) = (1 - q^*)q^*(i \geq 0) \).

Similarly, once a successful \( aB \) chromosome arrives at one subpopulation, the probability that the next \( i \) subpopulations also have frequency \( q^{**} \) is a geometric with parameter \( q^{**} \). Thus, hitch-hiking in space will result in patches of relatively high and low frequency of \( A \) with patch size determined by the initial frequency of \( A \), \( q \) and the ratio \( c/s \). Only if \( q^* \) is nearly 1 will hitch-hiking preserve genetic uniformity.

A two-dimensional stepping stone model would be analysed in the same way. The patterns generated are much more complicated but the conclusion is the same. The spread of \( B \) among subpopulations will result in patches of high and low frequency of \( A \) even if the frequency was the same initially.

5. Detailed analysis of two populations

The preceding sections all relied on the simplification that the substitution of \( B \) took place instantaneously. It is appropriate to look in more detail at the case with intermediate levels of migration. We will derive some analytic approximations and then compare the approximate results with exact numerical analysis.

We consider two populations, labelled \( S_1 \) and \( S_2 \), and assume that the frequency of allele \( A \) in both is initially \( q \). At time \( t = 0 \) the favourable mutant \( B \) is introduced in frequency \( p_h \) to population \( S_1 \). If selection is strong \( (2Ns \gg 1) \) the fixation process is
Continuous black lines, \( F_{ST} \) as obtained by numerical integration of the exact system of equations for \( u_i \) and \( v_i \) \((i = 1, 2)\); continuous grey line, \( F_{ST} \) according to (20); dotted grey line, \( F_{ST}^{\max} \) according to (22). Parameters: \( q = 1/2, p_u = 10^{-3}, d = 2, s = 10^{-2} \) and \( c = 10^{-2} \).

nearly deterministic. The time for \( B \) to be fixed (i.e. going from frequency \( p_o \) to \( 1 - p_o \)) in one population takes approximately

\[
t_1 = \frac{-2}{s} \log(p_o) \tag{9}
\]
generations. Eventually, a \( B \)-carrying individual will migrate to \( S_2 \) and take either allele \( A \) or \( a \) allele a long. In the absence of recombination no differentiation is to be expected because the same haplotype (either \( AB \) or \( aB \)) will eventually be fixed in both populations. With \( c > 0 \), after fixation of \( B \) in both, population \( S_2 \) may be of one type while the other population is of the other type. We will characterize the effect of hitchhiking by deriving some approximate results for \( F_{ST} \).

Following Maynard Smith & Haigh (1974) we write the variables in terms of conditional frequencies for the \( A \) allele and of the frequency of \( B \). We define \( u_i \) and \( u_z \) to be the frequencies of \( A \) on \( B \)-bearing chromosomes in populations \( S_1 \) and \( S_2 \) and \( v_i \) and \( v_z \) be the frequencies of \( A \) on \( b \)-bearing chromosomes. To derive our analytic results, we first assume a continuous time approximation for (1):

\[
p_1(t) = \frac{p_0}{p_o + (1 - p_o)e^{-\lambda t}}, \tag{10}
\]
where \( p_1(t) \) is the frequency of \( B \) in population \( S_1 \).

In \( S_2 \), we assume that \( B \) will also increase deterministically but with a time delay \( \tau \):

\[
p_2(t) = \frac{p_0}{p_o + (1 - p_o)e^{-\lambda(t - \tau)}}, \tag{11}
\]
To derive \( \tau \), let time be rescaled in units of \( t_i : \tau' = t/t_i \), where \( t_i \) is defined by (9). The scaled migration rate is \( m_i = t_i m \). At time \( t' \), the number of migrants that carry \( B \) is approximately Poisson-distributed with a parameter \( \lambda(t') \). Note that \( \tau' = \tau/t_i \) is the first time when the expected number of \( B \) carrying migrants exceeds 1. Rescaling (11),

\[
p_1(t) = \frac{p_0}{p_o + p_0 \exp(-\lambda t)\tau'}.
\]

Assuming that the number of gametes in \( S_1 \) is \( 1/p_o \), the number of \( B \)-carrying migrants is \( m_i p_1(t')/p_o \). Therefore, determining the minimum value of \( \tau' \) for which \( m_i p_1(t')/p_o \geq 1 \), we obtain

\[
\tau' \geq \frac{\log(m_i - p_o)}{2\log(p_o)}.
\]

which, in terms of the original parameters, becomes

\[
\tau = \max 0, -\frac{\log(m_i - p_o)}{s} \approx \max 0, -\frac{\log(m_i)}{s}.
\]

When roughly \( m < s \), the effect of migration on population \( S_1 \) is negligible during the selective phase. As before, let \( q \) be the allele frequency of \( A \) in \( S_1 \) before \( B \) is introduced. After finishing the selective phase the frequency of \( A \) in \( S_1 \) is \( q' \), which is approximately

\[
q' = q + (1 - q)p_o^{\lambda \tau}
\]
if \( B \) was initially linked to \( A \) and

\[
q'' = q(1 - p_o^{\lambda \tau})
\]
if \( B \) was linked to a (see Appendix).

To find an approximation for \( F_{ST} \) as a function of time, we assume \( u_i(0) = 1 \). Numerical analysis of the exact equation for \( u_i \) and \( p_i \) \((i = 1, 2)\) suggests that for \( t > t_i/2 \) \( u_i \) is roughly constant and approximately equal to \( q' \). To calculate \( u_i(t_1 + \tau) \) we first need to determine the probabilities \( q^* \) and \( 1 - q^* \) with which an \( AB \) or an \( aB \) haplotype migrates to \( S_2 \). If this happens early \((\tau < t_i/2)\) then \( q^* \) will differ from \( q' \). As shown in the Appendix, one finds for this case

\[
q^* = q + (1 - q)(\exp(-\tau s))^{\lambda \tau} = q + (1 - q)\exp(-\tau c).
\]

If, on the other hand, \( \tau > t_i/2 \) then \( q^* = q' \). Similarly, if \( u_i(0) = 0 \), then

\[
q'' = q(1 - \exp(-\tau s))^{\lambda \tau} = q(1 - \exp(-\tau c))
\]
if \( \tau < t_i/2 \), and \( q'' = q'' \) if \( \tau \geq t_i/2 \). Recall that \( \tau \) depends on \( m \), and therefore \( q^* \) and \( q'' \) depend on \( m \) also. After \( B \) is substituted in both populations, the frequency of \( A \) in \( S_1 \) is \( u_i \) \((i = 1, 2)\). Therefore, when \( u_i(0) = 1 \), \( u_i(t_1 + \tau) = q' \) and \( u_i(t_1 + \tau) = q^* q' + (1 - q^*) q'' \). With
There are values of $t$ in $[0, t_1 + \tau]$, such that $p_1 - p_2 \approx 1$. In particular, if $m = 0$, then $p_1 - p_2 = 1$ for all $t > t_1$. On the other hand, with high migration rates the selective substitution in $S_2$ is initiated only shortly after $t = 0$. Differentiation between the populations is possible only if a recombination event takes place in the short time interval $[0, \tau]$. Such an event is more unlikely the smaller $\tau$ is and both populations tend to be of the same type (the same allele is hitch-hiking). In this case, $p_1 - p_2 \approx 0$ for all $t$ in $[0, t_1 + \tau]$. Finally, intermediate migration rates result in an intermediate value of $\max_{t \in [0, t_1 + \tau]} (p_1 - p_2)$. Combining the two, we define

$$F_{ST}^{\max}(m) = kF_{ST}(m = 0) + (1 - k)F_{ST}(m)$$

with

$$k = \max_{t \in [0, t_1 + \tau]} (p_1(t) - p_2(t)).$$

For the special case $m = 0$ one has from (19)

$$F_{ST} = \frac{(1 - q)p_0^{2c/s} - (1 - q)p_0^s}{(2 - p_0^{2c/s})(2q + (1 - q)p_0^s)}.$$  

The time at which the difference $p_1(t) - p_2(t)$ is maximized is $t = (t_1 + \tau)/2$. Therefore,

$$k = \frac{\tanh^{ST}}{4}.$$  

Fig. 2 shows plots of $F_{ST}$, $F_{ST}^{\max}$ and their analytic approximations for some parameter choices.

6. Discussion and conclusions

We have shown that genetic hitch-hiking in a subdivided population can lead to a substantial temporarily increased differentiation at neutral loci. Increased differentiation would be expected when $2Nm \ll 1$ and $c \ll s$. Ultimately, low levels of gene flow could reduce the extent of differentiation (Fig. 2), but that could occur on a sufficiently long time scale that it would not be important. We have made many restrictive assumptions in order to obtain relatively simple results, and so we cannot claim that our analysis applies to all situations. The assumption that $2Nm \ll 1$ is particularly limiting. If the migration rate is higher, then it is not possible to assume that fixation within each subpopulation occurs much more rapidly than dispersal of the mutant among subpopulations. Stochastic effects of the kind discussed by Otto & Barton (1997) are also potentially important. Our analysis is intended to show that hitch-hiking in space can under some circumstances lead to an increase in population differentiation. To provide an accurate quantitative picture of the extent of differentiation under general conditions a much more detailed analysis and extensive simulations would be required.
The qualitative patterns in our results suggest that, if subpopulations are completely isolated, greater differentiation would be found in regions of a genome with low rates of recombination than in regions of higher rates. Low levels of gene flow change this pattern: differentiation is expected to be largest in regions of intermediate levels of recombination (Fig. 3). Very low or high recombination rates both lead to lower levels of $F_{ST}$. Therefore, completely isolated populations should produce a different relationship between $F_{ST}$ and recombination rate compared with populations in which there is some gene flow.

Two studies have found increased differentiation in a subdivided population in regions of low recombination. Stephan & Mitchell (1992) found greater differentiation between populations of *Drosophila ananassae* in India and Burma near the centromere on the X chromosome, where there is reduced recombination, than in two regions of the genome with higher rates of recombination. Begun & Aquadro (1993) found similar patterns when comparing populations of *D. melanogaster* in Zimbabwe with other populations. In both cases, these authors invoked independent hitch-hiking events in each geographic area to account for the greater extent of differentiation in regions of low recombination (Stephan, 1994). Our results here show that substitutions of the same advantageous allele spread by gene flow to different subpopulations could produce the observed patterns. At present, we cannot distinguish between these possibilities because we do not know the actual levels of gene flow in these two cases. If there is any ongoing gene flow, then eventual dispersal of an advantageous mutation arising in one population to another seems likely.

Appendix. Derivation of $q^*$ and $q^∗$

The conditional allele frequency $u(t)$ can be written in integral form (Stephan *et al.*, 1992) as

$$u(t) = u(0) - c(u(0) - q) \int_0^t \frac{(1-p_0)e^{-(t-s)p_0}z}{p_0 + (1-p_0)e^{-s}z} dz.$$  \hspace{1cm} (A 1)

For $t < t_1/2$, one may replace the denominator under the integral by $e^{-s}$. Upon integrating one obtains

$$u(t) \approx u(0) - c(u(0) - q) \frac{(1-e^{-s})}{e}.$$  \hspace{1cm} (A 2)

Replacing $t$ by $\tau$ according to (5) and $u_i(0)$ by 1, one arrives at

$$u_i(\tau) = q + (1-q)\pi_1(\tau).$$

If instead $t$ is replaced by $t_1/2$ the expression becomes

$$u_i(t_1/2) = q + (1-q)\pi_0^{(0)}.$$  \hspace{1cm} (A 3)

Because of the observation that $u_i$ does not change if $t > t_1/2$, the latter holds for $u_i(t_1)$.

This research was supported in part by grants from the National Institutes of Health (R01-GM40282 and R01-GM28428) to M.S. and by a research grant from the German Academic Exchange Service (DAAD) to T.W. We thank N. H. Barton and W. Stephan for helpful discussions of this topic and comments on an earlier version of this paper.

References


