Genet. Res., Camb. (1998), **72**, *pp.* 123–133. *With* 10 *figures.* Printed in the United Kingdom © 1998 Cambridge University Press 123

The effect of hitch-hiking on neutral genealogies

N. H. BARTON*
Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland

(*Received 18 December 1997 and in revised form 12 May 1998*)

Summary

When a favourable mutation sweeps to fixation, those genes initially linked to it increase in frequency; on average, this reduces diversity in the surrounding region of the genome. In the first analysis of this 'hitch-hiking' effect, Maynard-Smith and Haigh (1974) followed the increase of the neutral allele that chanced to be associated with the new mutation in the first generation, and assumed that the subsequent increase was deterministic. Later analyses, based on either coalescence arguments, or on diffusion equations for the mean and variance of allele frequency, have also made one or both of these assumptions. In the early generations, stochastic fluctuations in the frequency of the selected allele, and coalescence of neutral lineages, can be accounted for correctly by following relationships between genes conditional on the number of copies of the favourable allele. This analysis shows that the hitch-hiking effect is increased because an allele that is destined to fix tends to increase more rapidly than exponentially. However, the identity generated by the selective sweep has the same form as in previous work, $h[r/s]\ (2\ Ns)^{-2r/s}$, where $h[r/s]$ tends to 1 with tight linkage. This analysis is extended to samples of many genes; then, genes may trace back to several families of lineages, each related through a common ancestor early in the selective sweep. Simulations show that the number and sizes of these families can (in principle) be used to make separate estimates of r/s and Ns.

1. Introduction

The coalescent process provides a powerful method for approximating the effects of random drift on a sample of neutral genes (Hudson, 1990; Donnelly & Tavare, 1995). As one traces a set of lineages back in time, pairs of lineages will coalesce into one ancestral line at a rate which is the inverse of the effective number of genes in the population ($1/2N_e$). This simple approximation to the distribution of genealogies is accurate provided that the lineages of interest make up a small fraction of the population. The coalescent provides a natural tool for understanding the statistics of samples of DNA sequences and, moreover, allows very efficient computer simulations.

Real populations are structured, both spatially and genetically. Even if a gene does not itself affect fitness, its fate depends on where it is, and on which genes it is associated with. Classical population genetics shows that spatial subdivision has two opposing effects on neutral variation. Subdivision into partially isolated demes of constant size preserves variation, and so increases the effective size of the whole (Wright, 1939; Nagylaki, 1982). However, if populations fluctuate in size, then genes which happen to be in successful locations contribute disproportionately, leading to loss of variation and a smaller effective size (Whitlock & Barton, 1997). In genealogical terms, subdivision increases mean coalescence times, whilst fluctuations reduce them.

Provided demes are sufficiently large, the ancestry of samples drawn from a subdivided population can be approximated by the 'structured coalescent' (Kaplan *et al.*, 1991; Notohara, 1990; Hudson, 1990; Hey, 1991; Herbots, 1995; Nordborg, 1997). Tracing backwards in time, each neutral lineage is associated with a location which can change as a result of migration. At any given time, a lineage may coalesce with other lineages that are currently in the same

* e-mail: n.barton@ed.ac.uk.

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

place, with a probability inversely proportional to the number of genes reproducing there. Exactly the same process can describe the evolution of a gene through a population which is structured by the existence of gene combinations with different fitnesses (Kaplan *et al*., 1988; Hudson & Kaplan, 1988). A neutral gene may find itself associated with various combinations of selected genes, each with a different fitness. Tracing back, each neutral lineage may be associated with a genetic background that can change with time as a result of recombination and mutation of the selected genes. At any given time, a lineage may coalesce with other lineages that are currently in the same background. This approach has been used to derive analytical and simulation results for the effect of various kinds of selection: balancing selection, where a constant allele frequency is maintained (Kaplan *et al*., 1988; Hudson & Kaplan, 1988); 'selective sweeps', where a single favourable mutation sweeps through the population (Kaplan *et al*., 1989); and 'background selection', where selection continually eliminates deleterious mutations (Charlesworth *et al*., 1995; Hudson & Kaplan, 1995).

Application of this approach to genetic structure raises several mathematical difficulties that are not immediately apparent from the analogy with spatial structure. First, transfer of the neutral marker between different selected backgrounds occurs by recombination, rather than migration. Though both processes can be described by a linear matrix equation, the geometry of recombination is considerably less tractable than for migration on a one- or two-dimensional lattice: transfer can be to any of very many gene combinations. Second, and more seriously, almost all existing results are for one selected locus. However, very large numbers of genes are under selection, and substantial effects are only likely to be observed through the cumulative effects of many loci. Then, each genetic background may be present in small numbers, if at all. This leads to two related violations of the assumptions underlying the classical coalescent: the numbers of each particular genotype actually found in a population may be small, and may vary randomly. In the classical models, an individual's ancestor can be sampled from a known deterministic distribution; with genetic structure, this approximation is only reasonable if the total population size is large compared with the number of genotypes. When the approximation fails, we must incorporate stochastic fluctuations of the background genotypic array, and must allow for interactions between the lineages that descend through a rare genotypic class.

In this paper, I explore these two difficulties in the simple case of a 'selective sweep' at a single locus. In their original analysis of this problem, Maynard Smith & Haigh (1974) supposed that the only randomness was in which neutral allele was first associated with the new mutant; subsequently, the selected allele increases exponentially at a rate s, and its association with the neutral marker decays geometrically by recombination, at a rate r. However, subsequent random drift, and random fluctuations in the numbers of the selected allele, can have a substantial effect. In particular, the expected frequency of an allele *given that it will fix* is accelerated above the unconditional expectation of $e^{st}/2N$ by a factor $1/2s$. This problem was discussed by Maynard Smith & Haigh (1974) and Kaplan *et al*. (1989); however, while Otto & Barton (1997) take it into account in calculating the effect of a selective sweep on the expected frequencies of alleles at linked loci, its effects on higher-order measures such as heterozygosity have not been analysed. A further problem, which becomes important when more than two lineages are considered, is that these lineages may interfere with each other during the early stages, when they may be associated with small numbers of the favourable mutant. While these factors do not have a very large influence on the effect of a selective sweep on pairwise measures such as the distribution of coalescence times or heterozygosity, a proper analysis of the process may help in understanding the much harder case, where there are many selected loci.

2. Analysis of pairwise relationships

(i) *The four phases of a 'selective sweep'*

For simplicity, only the case of selection strong relative to drift will be considered. (If Ns were moderate or small, the more elaborate algorithm of Neuhauser & Krone (1997) would be required to describe the genealogy of the selected locus.) Assuming $1 \gg s \gg 1/N$, a 'selective sweep' can be divided into four phases. After a favourable mutation, P, arises, its numbers, k, fluctuate in a branching process with expected growth rate s. Only the fraction $\sim 2s$ of processes which lead to ultimate fixation need be considered. After some randomly distributed time, these will reach a large enough number ($ks > 2$, say) to increase deterministically; in this, the second phase, numbers are large enough for drift to be negligible, but the allele frequency is still low ($k \ll 2N$). In the third phase, which lasts $\sim 1/s$ generations, the new allele sweeps to high frequency. Finally, the original allele, Q, is eliminated, leading to fixation. Now, trace lineages at a neutral locus backwards through these four phases. Any sample of lineages must begin in association with P. However, in the third phase, as Q becomes common, lineages cross back and forth between backgrounds P and Q by recombination. Tracing back to the second phase, some fraction of lineages will now be associated with Q. Since P is now rare, these lineages are unlikely to return to association

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

with P, and will coalesce only in the distant past, $\sim 2N$ generations back. During the second phase, more lineages escape by recombination into background Q, never to return. Lineages associated with allele P now also have an appreciable chance of coalescing during the brief duration of the sweep. Finally, in the first phase, all lineages which remain in association with P must either coalesce or escape by recombination. This first phase is the most complicated, since one must consider genealogies conditional on the random sequence of numbers of the P allele, and then average over this sequence.

The following analysis considers first the chance that a pair of lineages will coalesce while associated with the new allele P, some time after the substitution, rather than tracing back to a common ancestor in the more distant past. I set out a series of approximations of increasing complexity, leading up to the exact solution, and compare their accuracy. Next, the distribution of pairwise coalescence times is derived, in essentially the same way. Finally, the distribution of genealogies which relate samples of several genes is considered. Viewed in the long term, over timescales of $\sim 2N$ generations, a brief selective sweep appears to induce near-simultaneous coalescences; the problem is to find their statistical distribution, and to find how closely this is approximated by the classical coalescent, which treats lineages as evolving independently of each other.

(ii) *Coalescence in the first generation only*

The simplest approach is to find the probability f_P that a neutral marker currently associated with the favourable allele P traces back to the single ancestral allele which was associated with the original mutant, P. Since recombination transfers markers between genetic backgrounds, we must also follow the probability f_Q that a marker currently associated with Q traces back to association with the original mutant. The probability that two alleles presently associated with P are identical by descent as a result of the selective sweep can then be approximated as $f_{PP} = f_P^2$. This is equivalent to allowing for coalescence only in the first generation, and is essentially the approach originally taken by Maynard Smith & Haigh (1974).

The recursions for f_P, f_Q are

$$\begin{aligned} f'_P &= (1-rq)f_P + rqf_Q, \\ f'_Q &= (1-rp)f_Q + rpf_P. \end{aligned} \tag{1}$$

The difference between backgrounds, $\delta = f_P - f_Q$, decreases by a factor $(1-r)$ due to recombination, whilst the average $\hat{f} = pf_P + qf_Q$ increases by $\Delta\hat{f} = \delta\Delta p$ due to the increase Δp at the selected locus. Initially, $f_P = 1, f_Q = 0$, and so $\delta_0 = 1, \hat{f}_0 = p_0 = 1/2N$. Assume that $s, r \ll 1$, so that time can be taken to be continuous. Then, integrating over the timecourse of the substitution:

$$\begin{aligned} \delta &= \delta_0 \, \mathrm{e}^{-rt}, \\ f_P &= \hat{f} = \frac{1}{2N} + \int_{1/2N}^{1} \mathrm{e}^{-rt} \mathrm{d}p. \end{aligned} \tag{2}$$

If the favourable allele increases deterministically, with constant genic selection s, as $(p/q) = \mathrm{e}^{st}/2N$, and letting $f_{PP} = f_P^2$:

$$f_{PP} = (2N)^{-2\rho}\,\Gamma[1+\rho]^2\,\Gamma[1-\rho]^2 \qquad (\rho = r/s < 1, 2N \gg 2Ns \gg 1). \tag{3}$$

where Γ is the Gamma function.

We next take account of random fluctuations during the establishment of the new mutant, given that it is destined to be fixed. Since fluctuations in p occur at a time when Δp is very small, they do not contribute significantly to the increase in identity, $\Delta\hat{f} = \delta\Delta p$. Fluctuations have only an indirect effect, by altering the time between the occurrence of the mutant and the time at which the substitution occurs; during this time, the association between marker and favourable allele is dissipating by recombination. Let the allele frequency during the deterministic phase be $(p/q) = \mathrm{e}^{s(t+\tau)}/2N$. The quantity $z = 2s\mathrm{e}^{s\tau}$ has an exponential distribution with mean 1 (see appendix C of Otto & Barton, 1997); thus, the expected frequency of an allele destined for fixation is accelerated by a factor $(1/2s)$ relative to that expected in the absence of stochastic effects. Averaging f_{PP} over the distribution of τ gives:

$$f_{PP} = (4Ns)^{-2\rho}\,\Gamma[1+\rho]^2\,\Gamma[1-\rho]^2\,\Gamma[1+2\rho] \qquad (\rho = r/s < 1, 2N \gg 2Ns \gg 1). \tag{4}$$

Note that this is greater than the expectation of f_P, as calculated in appendix C of Otto & Barton (1997), since it includes the variance in f_P. Comparing (3) with (4), we see that allowing for the random stochastic acceleration increases identity by a factor $(2s)^{-2\rho}$, which may be substantial when linkage is loose.

(iii) *Coalescence in all generations*

Next, we allow for the possibility that lineages may coalesce during any of the early generations, when background P is present in small numbers. Let f_{PP} be the chance that two genes, presently associated with P, are identical by descent; similarly for f_{PQ}, f_{QQ}. For small r and large N, the recursions simplify to:

$$\begin{aligned} \Delta f_{PP} &= \frac{(1-f_{PP})}{2Np} + 2rq(f_{PQ} - f_{PP}), \\ \Delta f_{PQ} &= r(qf_{QQ} - f_{PQ} + pf_{PP}), \\ \Delta f_{QQ} &= \frac{(1-f_{QQ})}{2Nq} + 2rp(f_{PQ} - f_{QQ}). \end{aligned} \tag{5}$$

When allele P is present in large numbers, coalescence is negligible, and so terms in $1/2N$ can be dropped.

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

Then, lineages descend independently, and identities can be calculated from (2), in terms of f_P, f_Q. Suppose we start at some low frequency ϵ $(1/Ns \ll \epsilon \ll 1)$, when $f_{PQ,\epsilon}, f_{QQ,\epsilon} = 0$. Integrating (5) up to fixation, the final identity is

$$f_{PP,1} = f_{PP,\epsilon}(\epsilon^{\rho}\Gamma[1+\rho]\,\Gamma[1-\rho])^2. \tag{6}$$

In the early stages, any lineages which escape from background P are unlikely to trace back into it. Setting $f_{PQ} = 0, q = 1$, we need only consider f_{PP} during this period. Integrating the equation for Δf_{PP} in (5), on the deterministic assumption $p = \mathrm{e}^{st}/2N$, and splicing the solution onto (6) at some arbitrary ϵ, gives

$$f_{PP,1} = (2Ns)^{-2\rho}\,\Gamma[1+\rho]^2\,\Gamma[1-\rho]^2\,\Gamma[1+2\rho]. \tag{7}$$

This expression is based on the same assumptions as Stephan *et al.*'s (1992) analysis. Their diffusion approximation (9) for the variance of neutral allele frequencies in the two backgrounds is equivalent to (5), and their approximate solution (19) corresponds to (7), though without the factor $\Gamma[1+\rho]^2\,\Gamma[1-\rho]^2$.

Comparison of (7) with (3) shows that allowing for coalescence throughout the time when the allele is rare has introduced a factor $s^{-2\rho}$. However, (7) does not allow for the stochastic fluctuations in numbers $(2Np)$ in the early stages of increase, or for the expected acceleration of an allele that is destined to fix. We deal with this problem in the next section.

(iv) *The exact solution*

With strong selection, stochastic fluctuations influence the numbers of the favourable allele only when it is at very low frequency. Then, we need only consider pairs of lineages that are both within the rare background (f_{PP}), and must average the equation for f_{PP} in (5) over random sequences of p. Since these sequences form a Markov chain, f_{PP} must depend solely on the present numbers of copies of P, $k = 2Np$. Let this conditional identity be f_k. This changes according to

$$f_k^* = (1-r)^2 \sum_{j=1}^{\infty} \Gamma_{jk}\left(\frac{1-f_j}{j} + f_j\right). \tag{8}$$

Here, Γ_{jk} is the probability that if there are k copies of P in the present generation, there were j in the previous generation; the probability of identity in the previous generation is proportional to $1/j$.

The backwards transition matrix, P_{jk}, can be calculated from the forwards matrix using the relation $\Gamma_{jk} = (\psi_j/\psi_k)\,P_{jk}$, where ψ_j is the leading left eigenvector of P_{jk}. For the Wright–Fisher model, with expected offspring number $\lambda = (1+s)$, ψ_j is close to its diffusion approximation, $2/j$. Using this approximation:

$$\gamma_{jk} = \frac{\lambda \mathrm{e}^{-j\lambda}(j\lambda)^{k-1}}{(k-1)!}. \tag{9}$$

The sum of Γ_{jk} over j is 1; the approximation (9) sums to $[(1+s)^k\,\mathrm{Li}_{1-k}(\mathrm{e}^{-(1+s)})]/(k-1)!$, which is very close to 1 except for small k. For $s = 0{\cdot}1, k = 1, 2, 3, 4\ldots$, the sum is 0·549, 0·905, 0·994, 1·001 …, and thereafter deviates by $< 0{\cdot}0001$.

The moments of the backwards distribution can be found from (9):

$$\left\langle \frac{1}{j} \right\rangle = \sum_k \frac{1}{j}\Gamma_{jk} = \frac{\lambda}{k-1},$$

$$\langle j \rangle = \sum_k j\Gamma_{jk} = \frac{k}{\lambda}, \tag{10}$$

$$\mathrm{var}\,(j) = \sum_k (j - \langle j \rangle)^2\,\Gamma_{jk} = \frac{k}{\lambda^2}.$$

Again, this is accurate except for $k = 1, 2$. For $s = 0{\cdot}1$, the errors for $\langle 1/j \rangle$ are 59·9 %, 6·9 %, −0·67 % for $k = 2, 3, 4$; for $\langle j \rangle$, 19·8 %, −6·4 %, 0·7 % for $k = 1, 2, 3$; for $\mathrm{var}\,(j)$, −8·1 %, 11·1 %, 0·3 % for $k = 1, 2, 3$.

This recursion can be solved numerically by setting $f_k^* = f_k$. The equations can be closed by supposing that f_j converges to its asymptotic form $\sim j^{-2\rho}$ above some large j, and using (9) to approximate Γ_{jk}. It can be approximated by expanding f_j as a Taylor series around k, and substituting for the moments from (10). This leads to the diffusion:

$$\begin{aligned}
\Delta f_k = 0 = &-2rf_k + \left\langle \frac{1}{j} \right\rangle (1-f_k) \\
&+ f_k' \left\langle \left(\frac{j-1}{j}\right)(j-k) \right\rangle \\
&+ \frac{f_k''}{2} \left\langle \left(\frac{j-1}{j}\right)(j-k)^2 \right\rangle \\
= &-2rf_k + \frac{(1-f_k)}{k-1} - f_k'\left(\frac{1+ks(k-2-s)}{(k-1)(1+s)}\right) + \frac{kf_k''}{2} \\
&\times \left(\frac{(ks)^2 - (ks+1)(1+s)^2 - s^2 + (k-1)(1+s)}{(1+s)^2(k-1)}\right).
\end{aligned} \tag{11}$$

In the limit of small s, large k, with $y = ks \sim 1$, this simplifies to the diffusion limit:

$$0 = -2\rho f + \frac{(1-f)}{y} - yf' + \frac{y}{2}f'' + O(s). \tag{12}$$

In the limit of large y, the solution to (12) converges to $h[\rho]\,y^{-2\rho}$; it can be solved numerically by imposing this constraint for large y, and setting $f[0] = 1$. (It is possible, but unhelpful, to express the $h[\rho]$ explicitly as a double integral of a MeijerG function.) This solution is valid for $y \ll 2N$; by splicing it onto (6), the net effect of the selective sweep can be found as

$$f_{PP} = h[\rho]\,(2Ns)^{-2\rho}\,\Gamma[1+\rho]^2\,\Gamma[1-\rho]^2. \tag{13}$$

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

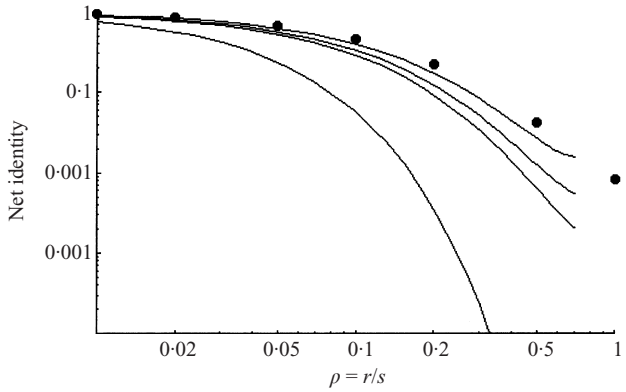

Fig. 1. Net identity generated by a selective sweep, plotted against r/s. The curves are (from bottom to top): coalescence in the first generation, exponential increase (3); coalescence in the first generation, stochastic increase (4); coalescence in all generations, deterministic increase (7); and coalescence in all generations, stochastic increase. $Ns = 100$. Note that these solutions are valid for $\rho < 1$ in the limit $Ns \to \infty$, but break down near $\rho = 1$ for finite Ns. Filled circles show simulated results, for $s = 0{\cdot}1$, $N = 10^3$, 100000 replicates. Standard errors are indistinguishable on this scale.

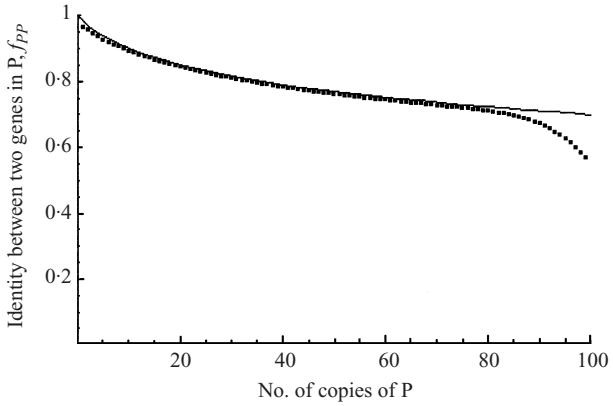

Fig. 2. The identity between two genes associated with P, f_{PP}, is plotted against the number of copies of P, for the discrete solution to (8) (squares), and the diffusion approximation (12, 13) (line). Note that f_2 is slightly in error, because the backwards matrix was approximated by (9); the discrete solution also fails near $k = 100$, where the calculation was truncated. Otherwise, agreement is close.

This has the same form as (4, 7), but is larger than either (Fig. 1). The diffusion approximation (12, 13), which takes into account both the stochastic increase in the selected allele, and coalescence in all generations, is in excellent agreement with the discrete equation (8) (Fig. 2).

(v) *The distribution of coalescence times*

Thus far, we have calculated the probability that two genes both descend from a common ancestor which was associated with the favourable mutant, P. In large populations, this gives the net identity generated by the selective sweep, since ancestors that were associated with allele Q are likely to be much more ancient ($\sim 2N$ generations back). We now find the distribution of coalescence times, by considering the probability that two genes are identical in stale, despite mutation to novel alleles at a rate μ. The identity in state, considered as a function of $z = (1-\mu)^2$, gives the generating function for the distribution of coalescence times.

The recursion for the identity in state is now found simply by allowing for a decrease in identity by a factor $(1-\mu)$ per generation, down each of the two lineages. In the early stochastic stages of increase, the exact recursion (8) is multiplied by a factor $(1-\mu)^2$. The diffusion approximation to this recursion is then found simply by replacing ρ by $(\rho+\mu/s)$ in (13), and has the solution $h[\rho+\mu/s](2Ns)^{-2(\rho+\mu/s)}$, where $h[\ ]$ must be found numerically from (8), or its approximation, (12). After coalescence has become negligible, the change in identity, f_{PP}, is proportional to f_P^2, where f_P, f_Q are given by a modification of (1):

$$\begin{aligned} f'_P &= (1-rq-\mu)f_P + rqf_Q, \\ f'_Q &= (1-rp-\mu)f_Q + rpf_P. \end{aligned} \tag{14}$$

This has solution $f_{P,1} = \mathrm{e}^{-\mu T_\epsilon}\Gamma[1+\rho+\mu/s]\,\Gamma[1-\rho-\mu/s]\,\epsilon^\rho f_{P,\epsilon}$, where T_ϵ is the time from the present, back to when $p = \epsilon$. Combining this solution with that for (12) raises a delicate question. Tracing back from the present, the favourable allele will have reached frequency ϵ at some definite time, T_ϵ. If its increase were deterministic throughout, it would have originated at some earlier time T, with $\epsilon = \mathrm{e}^{s(T-T_\epsilon)}/2N$. However, the time when it actually originated is randomly distributed and, on average, somewhat more recent; moreover, the stochastic acceleration will be correlated with the identity generated by the selective sweep. We will avoid this complication by working in terms of the *effective* time of origin, T, which can be calculated from the observed deterministic increase of the favourable allele, and which is on average somewhat more distant than the actual origin of the successful mutation. On this interpretation, ϵ cancels as the two solutions are spliced together, and we have

$$f_{PP} = h\left[\rho+\frac{\mu}{s}\right](2Ns)^{-2\rho}s^{-\frac{2\mu}{s}}\Gamma\left[1+\rho+\frac{\mu}{s}\right]^2\Gamma\left[1-\rho-\frac{\mu}{s}\right]^2 \mathrm{e}^{-2\mu T}, \tag{15}$$

where $h[\ ]$ is given by (8), or its diffusion approximation (12). This is the generating function for the distribution of coalescence times, measured relative to the effective start of the selective sweep. Note that, because coalescence occurs early in the selective sweep, the distribution of coalescence times is independent of population size, N (except through the strength of selection relative to drift, Ns).

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

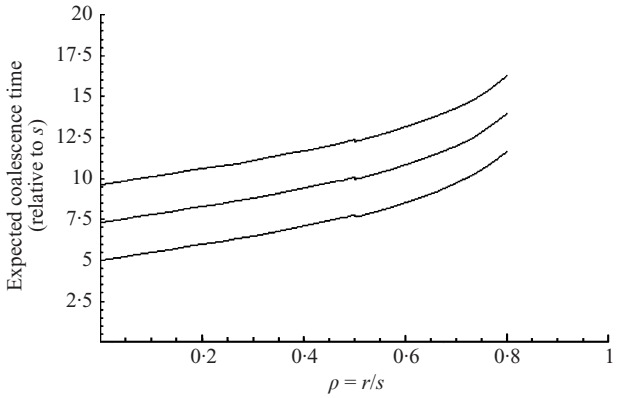

Fig. 3. The expected time of coalescence, counted from the effective start of the selective sweep, plotted against recombination rate. Both are scaled relative to s; from (16). From bottom to top, the three curves are for $s = 0{\cdot}01$, $0{\cdot}001$, $0{\cdot}0001$.

The expected coalescence time, conditional on the coalescence having occurred as a result of the bottleneck, is

$$E[t] = -\frac{\partial \log[f_{PP}]}{2\partial\mu}$$
$$= T - \frac{1}{s}\left(\log\left[\frac{1}{s}\right] + \psi[1+\rho] - \psi[1-\rho] + \frac{1}{2}\frac{\partial \log[h]}{\partial\rho}\right), \quad (16)$$

where $\psi[z] = \partial \log[\Gamma[z]]/\partial z$. For tight linkage, the average coalescence time, counting from the effective start of the selective sweep, is close to $\log[1/s]/s$; it increases approximately linearly with recombination rate (Fig. 3). The variance of the coalescence time can be found by taking the second differential of (16), and the full distribution by taking its inverse Laplace transform.

3. The structures of genealogies

With strong selection ($Ns \gg 1$), the coalescence events caused by a selective sweep occur almost simultaneously relative to the timescale of random drift ($t \sim N$); thus, the distribution of coalescence times derived above could not in practice be observed directly. However, even disregarding this information, a sample of genes contains more information than just the average pairwise identity. We now consider whether this information could be used to distinguish a selective sweep from a brief reduction in population size, and to disentangle the several parameters which describe the selective sweep. Kaplan et al. (1989) considered the distribution of the number of lineages which coalesce during a selective sweep. However, this does not provide a complete description of the process. Tracing a sample of lineages back, some subset will share a common ancestor associated with allele P, which will then escape into background Q. Other subsets may coalesce to share a *different* ancestor. Thus, a brief bottleneck generates some number of *families* of lineages, each sharing a different common ancestor at approximately the same time, and related to each other in the more distant past. The structure of the genealogy can therefore be represented by a list of family sizes, $\mathbf{n} = \{n_1, n_2, \ldots\}$.

An obvious approximation is to suppose that lineages coalesce randomly during the bottleneck, with some probability. On this view, all the statistical effects of a brief selective sweep depend on a single parameter; it gives a good approximation to Tajima's D (Tajima, 1989; J. M. Braverman, personal communication). This approach simplifies inferences from sequence data, but on the other hand would make it impossible to infer more than one parameter; for example, it would not be possible to disentangle N, s and ρ. In the following section, therefore, we find the distribution of the numbers and sizes of families of lineages generated by a selective sweep. We investigate first whether this distribution can be encapsulated in a single parameter, and second whether the usual coalescent approach (e.g. Kaplan & Hudson, 1989) gives an adequate approximation to the exact branching process.

It would be simple to extend (8) to give the chance that a set of i genes are all identical by descent, given that they are associated with a selected allele present in k copies. However, this would not give enough information to reconstruct the full distribution of family sizes, $\mathbf{n}$. Moreover, since coalescence occurs at times when the number of sampled lineages may approach the number of favoured alleles, it is more straightforward to follow the relationships among *all* those neutral alleles that are associated with the favourable mutation, P.

These relationships can be constructed as follows. First, establish the numbers of the selected allele, P. This can be done either by starting at some large number of copies of P, and working back using the backwards matrix Γ_{jk}, or by starting with a single copy, and working forwards using the forwards transition matrix conditioned on ultimate fixation, P^*_{jk}. Second, propagate the relationships forwards in time, given this random sequence. In the first generation there is a single copy of P, associated with a single family of neutral alleles, of size 1; this is represented as $\mathbf{n} = \{1\}$. In the next generation, there are k copies of P. Each of these descends from a lineage that was associated with Q in the previous generation with probability r and, with probability $(1-r)$, descends from the family that was associated with P in the previous generation. Therefore, there is a single family whose size n_1 is given by binomial sampling from k genes with probability $(1-r)$, and a

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

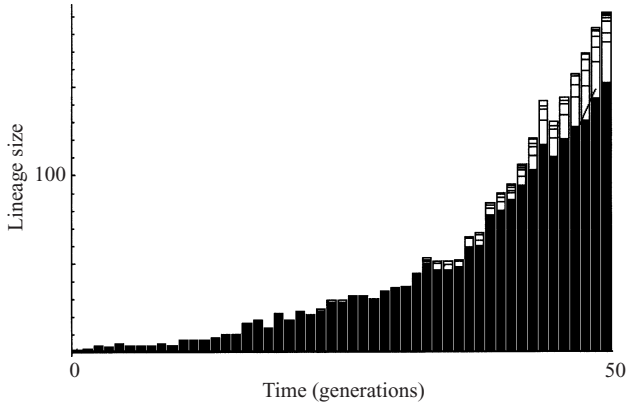

Fig. 4. The size of a set of lineages, plotted over 50 generations since the origin of an allele P with advantage $s = 0{\cdot}1$. The recombination rate is $r = 0{\cdot}01$. Each segment of the bar chart represents the numbers of each distinct family of lineages, including those lineages that go extinct. The continuous curve gives the expected deterministic rate of increase, which in this case is somewhat slower than the actual rate. The black bars show the size of the largest lineage, descending from the gene associated with the original favourable mutation. The white segments show lineages which later become associated with the favourable mutation by recombination, and thus are also amplified as it increases.

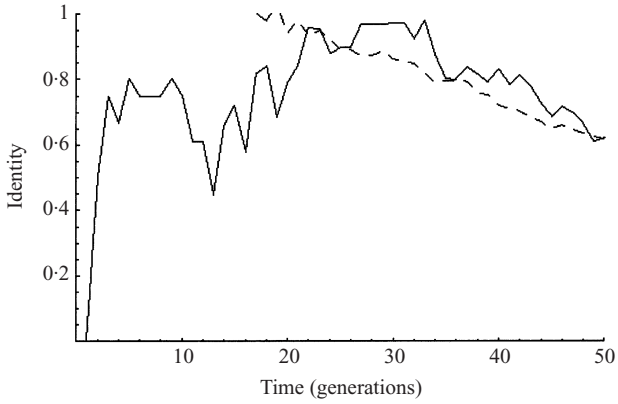

Fig. 5. The probability that two distinct genes associated with the favourable allele P are identical by descent, plotted against time. The continuous line shows the actual identity in the example of Fig. 4, whilst the dashed line shows the prediction $h[\rho](ks)^{-2\rho}$ for large numbers of copies (from 12).

set of $(k-n_1)$ unrelated families of size 1. In general, suppose that in one generation there are j copies of P, divided among a set of families of size $\mathbf{n}\,(j = \Sigma n_i)$. In the next generation, there are k copies of P, divided into families of size $\mathbf{n}'$. the distribution, of family sizes, $\mathbf{n}'$, is given by multinomial sampling; the previous families, $\mathbf{n}$, have probability $\mathbf{n}(1-r)/j$ of being sampled, whilst the remainder is made up by unrelated lineages that have just become associated with P, and form families of size 1. The third and final step is to derive the ultimate distribution of family sizes from the distribution of family sizes at a time when P is present in many copies ($k^* \gg 1/s$) but is nevertheless at low frequency ($k^* \ll 2N$). The probability that a gene sampled after the sweep has been completed traces back to one of the k^* copies of P is $\theta = (k^*/2N)^{\rho}\,\Gamma[1+\rho]\,\Gamma[1-\rho]$ (from 3). Therefore, the final distribution of family sizes is given by multinomial sampling from the distribution at k^*, with probabilities $n\theta/k^*$.

Figure 4 shows the sizes of a set of lineages sampled in this way, plotted against time. In this example, there is tight linkage ($\rho = r/s = 0{\cdot}1$), and so the allele which was associated with the initial favourable mutation leaves the largest number of descendants: 152 out of 192 after 50 generations. However, there are other families with significant representation: one with 23 members, one with 7, one with 5, and one with 2. The 3 remaining neutral alleles are unrelated. The actual change in identity among distinct pairs chosen from among those alleles associated with P is plotted against time in Fig. 5; it converges to decay at a rate proportional to $k^{-2\rho}$ (dashed line) as k becomes large.

In the above example, results were presented only for the set of genes present after the initial establishment of the favourable allele. The ultimate effect of the bottleneck involves a sampling from this set, with probability θ. The statistical distribution of ultimate family sizes was investigated by taking 100 replicates, in each of which 100 genes were sampled after the selective sweep had been completed. As above, $s = 0{\cdot}1$, $r = 0{\cdot}01$; population size was taken as $N = 10^6$. The calculation was terminated at $k = 1000$ copies, at which time a gene sampled after the selective sweep has a chance $\theta = 0{\cdot}509$ of tracing back to this initial set. The mean pairwise identity in the sample, averaged over replicates, was 0·090, compared with a prediction from (13) of 0·099. However, identity varied widely between replicate sweeps, depending on chance associations between neutral and selected allele; the standard deviation was 0·045 (upper bars in Fig. 6).

Two factors affect the reduction in diversity caused by a selective sweep. The relatedness among the cohort of neutral alleles associated with the favourable mutation depends solely on the relative rates of recombination and selection, $\rho = r/s$. The consequence of this relatedness for the whole population depends on the fraction of lineages that trace back to this initial cohort, θ, which decreases with population size, N. This is because in a large population it takes many generations for the new mutation to reach high frequency, during which time associations are broken down by recombination. From (13), pairwise identity depends mainly on the factor $(2Ns)^{-2\rho} = \exp(-2\rho \ln(2Ns))$. Thus, identity can be produced either by moderately tight linkage in a moderately large population, or very tight linkage in a very large

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

Fig. 6. The distribution of pairwise identity among samples of 100 genes, taken after an allele P with advantage $s = 0{\cdot}1$ has fixed. Results are from 100 independent replicates in each case. The upper bars are for $r = 0{\cdot}01$, $N = 10^6$, whilst the lower bars are for $r = 0{\cdot}00453$, $N = 10^{12}$. Both combinations are predicted to give mean identity 0·099 (from 13).

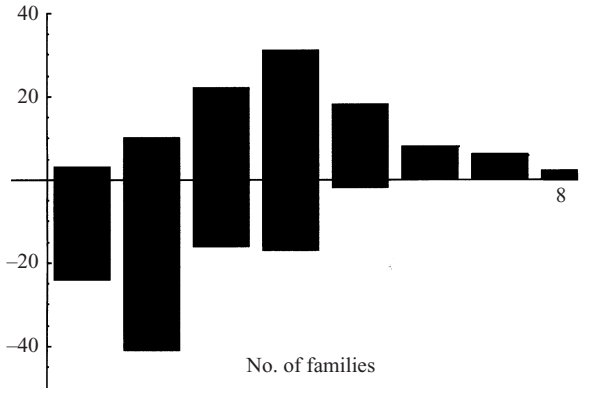

Fig. 7. The distribution of numbers of families of related alleles (discounting singlets) among a sample of 100 genes following a selective sweep ($s = 0{\cdot}1$). The upper bars are for $r = 0{\cdot}01$, $N = 10^6$, whilst the lower bars are for $r = 0{\cdot}00453$, $N = 10^{12}$, as in Fig. 6.

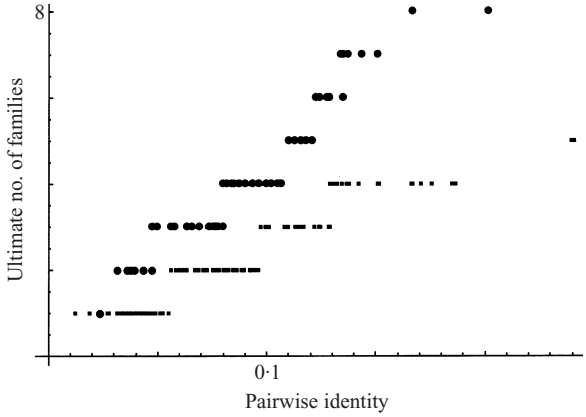

Fig. 8. Scatter plot of the ultimate number of families versus the pairwise identity. Parameters are as in Figs. 6 and 7; filled circles: $r = 0{\cdot}01$, $s = 0{\cdot}1$, $N = 10^6$; squares: $r = 0{\cdot}00453$, $s = 0{\cdot}1$, $N = 10^{12}$.

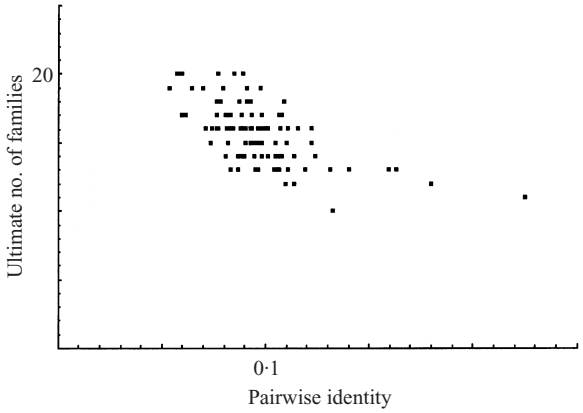

Fig. 9. Scatter plot of the ultimate number of families versus the pairwise identity, under simple random drift. One hundred replicates were simulated, after $t = 0{\cdot}208N$; this gives an expected pairwise identity 0·099, as in Figs. 6–8.

population. In the first case, many families may emerge from the initial bottleneck, and be represented in the final sample. In contrast, in the second case only a single family is likely to emerge, so that all the identity generated by the selective sweep will be due to a single coalescence. These two extremes are compared in Figs. 6 and 7. The upper bars show results for $N = 10^6$, $Ns = 10^5$, $\rho = 0{\cdot}1$, whilst the lower bars are for $N = 10^{12}$, $Ns = 10^{11}$, $\rho = 0{\cdot}0453$. These parameter values were chosen to give the same average pairwise identity, 0·099, and in fact give similar distributions of pairwise identity across replicates (Fig. 6). In contrast, the number of families differs substantially between the two cases (Fig. 7). With a moderate population size and moderate linkage (upper bars), the mean number of families is 4·05 (SD 1·50), whilst with a large population and tighter linkage (lower bars), the mean number of families is 2·32 (SD 1·08).

Figure 8 shows the joint distribution of identity and number of families for the two examples. Since both quantities can be estimated, given sufficient sequence diversity, the lack of overlap in Fig. 8 indicates that, in principle at least, these parameter values could be distinguished. However, the feasibility of estimating Ns as well as ρ is limited in practice by whether the effects of individual selective sweeps can be isolated; whether the number of sets of lineages that coalesce during each sweep can be estimated from sequence data; and by the weak (logarithmic) dependence of the structure of the genealogy on Ns.

The question now arises as to whether the effects of a brief selective sweep are equivalent to a brief reduction in population size. Provided that population size, N, does not become so small as to approach the number of sampled lineages, lineages can be taken as coalescing independently, at a rate inversely proportional to N. On this approximation, time can be

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

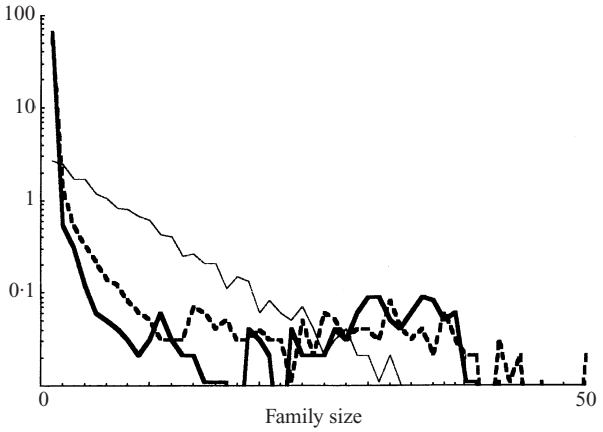

Fig. 10. The distribution of family size, averaged over 100 replicates. Comparison is between simple random drift (thin line); a selective sweep with moderate linkage in a moderate sized population ($r = 0{\cdot}01$, $s = 0{\cdot}1$, $N = 10^6$; thick dashed line); and a selective sweep with tighter linkage in a very large population ($r = 0{\cdot}00453$, $s = 0{\cdot}1$, $N = 10^{12}$; thick continuous line). Parameters are as in Figs. 6–9, and are chosen to give expected identity 0·099. Note that the vertical scale is logarithmic; the lower limit of 0·01 corresponds to a single occurrence of a family of that size amongst the 100 replicates.

rescaled, such that a population bottleneck is equivalent to a prolonged period at larger numbers (Felsenstein, 1992). Figure 9 shows the relationship between pairwise identity and the number of families of lineages, for a bottleneck in which drift generates expected identity 0·099, as in Figs. 6–8. The pattern is quite different; identity is due to a larger number of families (mean 15·7, SD 2·22), and the number of families decreases with pairwise identity, rather than increasing. Comparison of Figs. 8 and 9 suggests that a population bottleneck could readily be distinguished from a selective sweep if one knew the pattern of coalescence generated amongst a sample of 100 genes.

The number of families of related genes is just one statistic that describes the distribution of family sizes. The full distribution, averaged over 100 replicates, is shown in Fig. 10, for each of the three examples discussed above. For a given expected pairwise identity, simple random drift generates many clusters of relatives, each typically consisting of a few genes (thin line in Fig. 10); for this example, only 2·6% of genes are unrelated to any others ($n_i = 1$; left intercept on Fig. 10). In contrast, the identity generated by a selective sweep is typically due to a few families, each involving many genes; most genes remain unaffected by the sweep, since they do not trace back to an association with the favourable mutation. Thus, 63·8% of genes are unrelated to any others in the example of moderate linkage ($r = 0{\cdot}01$, $N = 10^6$; thick continuous line in Fig. 10), and 57·8% with tighter linkage ($r = 0{\cdot}004531$, $N = 10^{12}$; thick dashed line in Fig. 10). The distribution of family sizes allows different kinds of sweep to be distinguished: with tighter linkage, there is a broader spread of family sizes, whereas with looser linkage, family size clusters around sets of approximately 30 genes.

4. Discussion

In order to calculate properly the chance that a pair of genes will become identical by descent as a result of their association with a favourable mutation, the initial fluctuations in the frequency of that mutation must be accounted for. In particular, an allele that is destined to be fixed will increase substantially faster than expected from its selective advantage, thereby increasing its effect on linked loci. Such stochastic effects can be treated by following identity conditional on the numbers of the favourable genetic background, rather than taking the usual approach of following coalescence within a deterministically evolving population. For a selective sweep at a single locus, the increase in identity is due to events in just the first few generations, and so is not large (Fig. 1). However, stochastic fluctuations in the genetic background may become crucial when many loci are involved, so that each background genotype is rare. To see this, consider balancing selection. This maintains alternative genetic backgrounds in the population, and hence inflates diversity at linked neutral loci (Hudson *et al*., 1987). Calculations which assume that selection maintains two alleles at constant frequency at each of n equally spaced loci predict an implausibly large effect. However, this assumption cannot be correct for large n, because each of the 2^n backgrounds then becomes extremely rare; this is so even if selection is strong enough that allele frequencies hardly fluctuate. (For example, in a population of $< 10^6$, any particular 20 locus genotype is unlikely to be present at all.) A proper treatment of genealogies embedded in multilocus genetic backgrounds, whether by mathematical analysis or by simulation, must therefore take account of the effect of random drift on those backgrounds. This greatly complicates the problem, because the whole population must be considered, rather than just the sample of neutral alleles.

A selective sweep caused by the spread of a favourable mutation through a large population has similar effects on neutral diversity to a founder effect. In particular, the increase in numbers of a new mutation which is destined to fix has the same distribution as the numbers of a haploid subpopulation, founded by a single colonist and limited by logistic density-dependence. If the whole metapopulation is very large, neutral genealogies within an expanding subpopulation have the same distribution as do those within the favourable genetic background, at least during the early stages of the selective sweep. Then, any lineages which leave the subpopulation

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

through immigration are unlikely to return, just as any lineages which leave the favourable background by recombination are unlikely to trace back into it. The models are not quite identical, since during a selective sweep, lineages may cross back and forth between backgrounds P and Q at a different rate from an island model with extinction and recolonization. Nevertheless, the same methods apply and, in particular, an exact solution requires that the distribution of coalescence times (or identity by descent) must be calculated conditional on deme sizes or genotype numbers (see Whitlock & Barton, 1997).

The analytical results of this paper are restricted to the relation between pairs of genes. A sample of many genes contains more information than just pairwise relationships: it is this information which might allow different evolutionary processes to be distinguished. A selective sweep causes a burst of coalescence events, in which sets of lineages trace back to a single ancestral allele that was associated with the favourable mutation during its initial increase. If the mutation increased very rapidly ($r \ll s$), then one family is likely to emerge; it will include many genes that trace back to an association with the one gene that was associated with the initial mutation. However, if linkage is looser, several different families may emerge, tracing back to distinct ancestors that were lucky enough to become associated with the favourable mutation as it increased from low numbers. Thus, the relationship between the number of families of related genes, and the pairwise identity, can in principle be used to distinguish the parameters (r/s) and Ns which determine the nature of a selective sweep (Figs. 8, 9). It is not obvious that the number of families would be the best estimator for this purpose: the full distribution of family sizes differs between a founder event and selective sweeps, and between different kinds of founder event (Fig. 10), and so various statistics based on this distribution might be considered.

The discussion so far has been of what might be inferred, given a dated genealogy for a single neutral locus. One can also ask what might be inferred, in principle, from genealogies at multiple loci. Individual selective sweeps might be identified through bursts of coalescence that occur at the same time at different loci. For each sweep, the pairwise identity at each locus gives an estimate of $(2Ns)^{-2r/s}$. However, examining genealogies at linked loci could locate the sweep and separate all the parameters. For example, a plot of log(pairwise identity) against map distance would have slope $1/s$. This would locate the selected locus, and determine the strength of selection; together with genealogical structure, it would also allow a population bottleneck to be rejected. Examination of the distribution of family sizes at the various marker loci would determine Ns, and would be a test of the model of a simple selective sweep against alternatives such as the spread of a gene through a spatially structured population. Further information could come from linkage disequilibrium, or the concordance between linked genealogies.

In practice, of course, one cannot directly observe genealogies; inferences must be made from the DNA sequence, which has evolved as neutral mutations arise at random on the genealogy. Given a sufficiently long sequence, with no recombination, the genealogy can be reconstructed. However, it is not clear whether this is possible, since in outcrossing species recombination may not be sufficiently rare, relative to mutation. Moreover, if one concentrates on regions of the genome with little crossing over, hitch-hiking may reduce intra-population variation (Charlesworth, 1996), making it difficult to estimate the genealogy, and restricting information to the recent past.

Most work on distinguishing the various causes of reduced sequence diversity has concentrated on Tajima's (1989) D statistic, which is based on the discrepancy between estimates of $4N\mu$ from pairwise comparisons, and from the number of segregating alleles. A brief reduction in population size, or a selective sweep, is expected to cause an excess of rare alleles, and hence a negative value of D; this is because alleles which arise as the population recovers from a loss of diversity are likely to be recent, and hence at low frequency. This is not observed in *Drosophila* data, suggesting that reduced diversity in regions of reduced recombination may be due to some other process, such as the steady hitch-hiking effect of linked deleterious mutations ('background selection') (Braverman *et al*., 1995; Charlesworth *et al*., 1995). Tajima's D depends on the number of segregating alleles, which depends in turn on the total length of the tree. This is determined by the number of lineages which were extant before the selective sweep, or in other words on the number of coalescences that occurred. The number of lineages is distinct from the number of families of related genes discussed above, since it includes those singlet lineages that escape coalescence. Since these are the most common class (Fig. 10), the number of lineages is a less sensitive statistic for distinguishing different causes of reduced variability than the number of families. However, there is still the difficult question of whether the number of families of related genes (or some similar description of genealogical structure) can in practice be estimated from sequence data.

Fu (1997) compares several statistical tests, including Tajima' D, for their ability to detect various different processes that may cause allele frequency distributions to deviate from neutrality; however, he does not investigate whether they can be used to distinguish *amongst* these alternatives. The results presented here show that neutral genealogies contain considerable information about the nature of selection

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press

at linked loci. The question now is to find whether enough of this information is preserved in the DNA sequence to tell us how selection has acted on the surrounding genome.

S. P. Otto kindly supplied the simulation results shown in Fig. 1. I would like to thank B. Charlesworth, A. Etheridge and S. P. Otto for their comments on the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council, and by the Darwin Trust of Edinburgh.

References

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**, 783–796.

Charlesworth, B. (1996). Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genetical Research* **68**, 131–150.

Charlesworth, D., Charlesworth, B. & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**, 1619–1632.

Donnelly, P. & Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences; inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research* **59**, 139–147.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Herbots, H. M. (1995). Stochastic models in population genetics: genealogy and genetic differentiation in structured populations. PhD thesis, University of London.

Hey, J. (1991). A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys on Evolutionary Biology* **7**, 1–44.

Hudson, R. R. & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–839.

Hudson, R. R. & Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics* **141**, 1605–1617.

Hudson, R. R., Kreitman, M. & Aguade, M. (1987). A test for neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Kaplan, N., Hudson, R. R. & Iizuka, M. (1991). The coalescent process in models with selection, recombination and geographic subdivision. *Genetical Research* **57**, 83–91.

Kaplan, N. L., Darden, T. & Hudson, R. B. (1988). The coalescent process in models with selection. *Genetics* **120**, 819–829.

Kaplan, N. L., Hudson, R. R. & Langley, C. H. (1989). The hitch-hiking effect revisited. *Genetics* **123**, 887–899.

Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

Nagylaki, T. (1982). Geographical invariance in population genetics. *Journal of Theoretical Biology* **99**, 159–172.

Neuhauser, C. & Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics* **145**, 519–534.

Nordborg, M. (1997). Structured coalescent processes on different timescales. *Genetics* **146**, 1501–1514.

Nordborg, M., Charlesworth, B. & Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research* **67**, 159–174.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* **29**, 59–75.

Otto, S. P. & Barton, N. H. (1997). The evolution of recombination: removing the limits to natural selection. *Genetics* **147**, 879–906.

Stephan, W., Wiehe, T. H. & Lenz, M. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *Theoretical Population Biology* **41**, 237–254.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.

Whitlock, M. C. & Barton, N. H. (1997). The effective size of a subdivided population. *Genetics* **146**, 427–441.

Wright, S. (1939). *Statistical Genetics in Relation to Evolution*. Actualités scientifiques et industrielles 802. Exposés de biometrie et de la statistique biologique XIII. Paris: Hermann et Cie.

https://doi.org/10.1017/S0016672398003462 Published online by Cambridge University Press