The coalescent process in models with selection, recombination and geographic subdivision

NORMAN KAPLAN[1],* RICHARD R. HUDSON[2] AND MASARU IIZUKA[3]

[1] *National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, N.C. 27709 USA*

[2] *Department of Ecology and Evolutionary Biology, University of California at Irvine, Irvine, CA 92717 USA*

[3] *General Education Course, Chikushi Jogakuen Junior College, Ishizaka 2-12-1, Dazaifu-shi, Fukuoka-ken 818-01, Japan*

(*Received 14 February 1990 and in revised form 23 April 1990*)

Summary

A population genetic model with a single locus at which balancing selection acts and many linked loci at which neutral mutations can occur is analysed using the coalescent approach. The model incorporates geographic subdivision with migration, as well as mutation, recombination, and genetic drift of neutral variation. It is found that geographic subdivision can affect genetic variation even with high rates of migration, providing that selection is strong enough to maintain different allele frequencies at the selected locus. Published sequence data from the alcohol dehydrogenase locus of *Drosophila melanogaster* are found to fit the proposed model slightly better than a similar model without subdivision.

1. Introduction

As DNA sequence data become available for samples of genes from populations, it becomes more important to understand what patterns of nucleotide variation are expected under a variety of population genetics models. As a step in that direction, Kaplan, Darden & Hudson (1988) and Hudson & Kaplan (1988) studied the distribution of the number of polymorphic nucleotide sites in a sample of genes at selectively neutral sites linked to a selected site at which a polymorphism is maintained in a single panmictic population. Since many natural populations are geographically subdivided, it is desirable to generalize the analysis of Kaplan *et al.* (1988) and Hudson & Kaplan (1988) to a model with geographic structure. In this paper we show how to analyse such a model, focusing on the coalescent process of samples of genes from a subdivided population.

The behaviour of models with geographic subdivision is of immediate interest for the analysis of sequence data from the alcohol dehydrogenase (*Adh*) region of *Drosophila melanogaster* (Kreitman, 1983). Recent studies of variation at this locus have suggested the presence of a balanced polymorphism (Hudson, Kreitman & Aguadé, 1987; Kreitman & Aguadé, 1986; Oakeshott *et al.* 1982), and it is well known that there are latitudinal clines in the frequencies of electrophoretic variants at the *Adh* locus of *D. melanogaster* (Oakeshott *et al.* 1982; Simmons *et al.* 1989; Vigue & Johnson, 1973). Hudson & Kaplan (1988) compared the observed pattern of nucleotide variation in the *Adh* region to that predicted by a model in which the protein polymorphism at the *Adh* locus is maintained by balancing selection in a single panmictic population. The observed patterns were largely consistent with the balancing selection model without geographic subdivision. The purpose of this paper is to investigate the effects of geographic subdivision on the predictions of a balancing selection model. We will only consider the simplest case of a population with two subpopulations.

The theoretical approach is based on the coalescent process which is related to the genealogical history of a sample of genes (see review of Tavaré, 1984). The motivation for this approach stems from the work of Watterson (1975) who showed the connection between the coalescent process and the distribution of the number of selectively neutral polymorphic sites in a random sample of genes, assuming an infinite-sites model (Kimura, 1969). Recently, Kaplan *et al.* (1988) considered the coalescent process for a sample of selectively neutral genes at a locus which is tightly linked to a selected locus. Hudson & Kaplan (1988) extended this work and studied the coalescent process for a random sample of genes at a selectively neutral locus that is not tightly linked to a locus at which balancing selection maintains two alleles in the

* Corresponding author.

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

population. In this paper the analysis of Hudson and Kaplan is generalized to allow for geographic subdivision, as well as recombination, mutation and balancing selection. We focus on the case where the migration rate between the subpopulations is high, but selection is sufficiently strong and different in the two subpopulations to maintain substantially different allele frequencies in the two subpopulations. The results of this analysis are used to assess the effects of migration on the predicted variation in the *Adh* region of *Drosophila melanogaster*.

2. Theory

We begin by reviewing the connection between the coalescent process and S, the number of polymorphic sites in a random sample of size n, ($n \geqslant 2$) at a small region of the genome containing L nucleotide sites. Suppose we have a randomly mating diploid population of size N. If μ, the expected number of selectively neutral mutations per nucleotide site per chromosome per generation is sufficiently small, then with high probability, at most one selectively neutral mutation will have occurred at each nucleotide site since the most recent common ancestor. In this case the distribution of S can be approximated by

$$P(S=k) = \int_0^\infty e^{-\mu t}\frac{(\mu t)^k}{k!}dF(t) \quad (k \geqslant 0), \tag{1}$$

where

$$F(t) = P\left(\sum_{i-1}^{L} t_i \leqslant t\right) \quad (t \geqslant 0), \tag{2}$$

and t_i is the sum of the lengths (measured in generations) of all the branches of the ancestral tree describing the genealogical history of the ith nucleotide site, $1 \leqslant i \leqslant L$ (Watterson, 1975). If the L nucleotides are completely linked, then there is only one ancestral tree and so $\sum_{i-1}^{L} t_i = LT$, where T is the sum of the lengths of the branches of the ancestral tree of a single nucleotide site. Unless otherwise stated, we will assume that the region is completely linked. In this case it is appropriate to think of the region as a single locus. If μ is sufficiently small, then the probability of more than one mutation at a single site in the genealogy of the sample is negligible, and the representation in (1) is a reasonable approximation even if $L = 1$. The case $L = 1$ is important in the next section.

It follows from (1) that quantities describing the distribution of S can be calculated from quantities describing the distribution of T. For example

$$E(S) = L\mu E(T), \quad \text{Var }(S) = L\mu E(T) + (L\mu)^2 \text{ var }(T) \tag{3a}$$

and

$$P(S=0) = E(e^{-L\mu T}). \tag{3b}$$

The distribution of T is determined from the coalescent process. If the region is selectively neutral and isolated, i.e. not linked to a selected locus, then Watterson (1975) showed that T can be written as

$$T = \sum_{j-2}^{n} jY(j), \tag{4}$$

where the $\{Y(j)\}$ are independent random variables. Furthermore, if time is measured in units of $2N$ generations, then for large N, the distribution of $Y(j)$ can be approximated by an exponential distribution with parameter $j(j-1)/2$, $2 \leqslant j \leqslant n$.

Hudson & Kaplan (1988) used the coalescent process to study the distribution of T, assuming that the selectively neutral region is linked to a locus at which natural selection operates. There is no simple representation for the distribution of T in this case. However, if the allelic frequencies at the selected locus are tightly regulated, i.e. mutation and selection act in such a way that the frequencies of the alleles are essentially constant for long periods of time, then the Markov property of the coalescent process can be exploited to calculate quantities describing the distribution of T such as those in (3). Our goal is to generalize the analysis of Hudson and Kaplan for the tightly regulated case so as to include geographic subdivision.

Suppose a randomly mating diploid population of size N has two subpopulations, subpopulation 1 of size fN diploids and subpopulation 2 of size $(1-f)N$, $0 \leqslant f \leqslant 1$. At the selected locus A, it is assumed that there are two alleles, A_1 and A_2 and that their frequencies are tightly regulated. The frequency of the A_1 allele in the jth subpopulation is denoted by x_j, $j = 1, 2$.

Each generation the daughter population is obtained by random sampling with replacement after mutation, recombination, selection and migration have occurred. Since selection, migration, recombination and mutation are all effects of order $1/N$, the order of these processes in the life cycle makes no significant difference in our approximations (Hudson & Kaplan, 1988). The fitnesses of the three genotypes A_1A_1, A_1A_2 and A_2A_2 are w_{11}, w_{12} and w_{22}, respectively. The rates of mutation per gamete per generation at the selected locus are $u(A_1 \rightarrow A_2)$ and $v(A_2 \rightarrow A_1)$. Migration is such that in each generation, a proportion m_{12} of subpopulation 1 is made up of migrants from subpopulation 2, and a proportion m_{21} of subpopulation 2 is made up of migrants from subpopulation 1. Finally, the average number of crossovers per generation between the selectively neutral region and the selected locus is r. It is assumed that

$$w_{11} = w_{12} = w_{22} = 1 + O\left(\frac{1}{N}\right),$$

$$u = \frac{\beta}{2N} + O\left(\frac{1}{N^2}\right), \; v = \frac{\nu}{2N} + O\left(\frac{1}{N^2}\right),$$

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

$$m_{12} = \frac{\lambda_{12}}{2N} + O\left(\frac{1}{N^2}\right),\ m_{21} = \frac{\lambda_{21}}{2N} + O\left(\frac{1}{N^2}\right)$$

and

$$r = \frac{R}{2N} + O\left(\frac{1}{N^2}\right),$$

where $\beta > 0$, $\nu > 0$, $\lambda_{12} > 0$, $\lambda_{21} > 0$ and $R > 0$.

The generation from which the sample is taken is referred to as the 0th ancestral generation and the population t generations back in time as the tth ancestral generation. Each ancestor of each sampled gene (referred to as an ancestral gene) is in subpopulation 1 or 2 and is linked to either an A_1 or A_2 allele. We therefore define $Q(t) = (i_1, j_1, i_2, j_2)$ if in the tth ancestral generation, $t > 0$, i_k of the ancestral genes of the sample are in subpopulation k and are linked to the A_1 allele and j_k of the ancestral genes are in subpopulation k and linked to the A_2 allele, $k = 1, 2$. The value of $Q(0)$ depends on how the genes were sampled, e.g., if the sample is of size 2 and both genes are linked to an A_1 allele, but one is from subpopulation 1 and the other from subpopulation 2, then $Q(0) = (1, 0, 1, 0)$. The total number of ancestral genes in the tth ancestral generation is denoted by $|Q(t)| = \sum_{k=1}^{2}(i_k + j_k)$.

The number of ancestral genes does not increase as one goes back in time and so the Q process eventually reaches a state where $|Q(t)| = 1$, i.e. there is a single ancestor of the sample. The ancestral generation when

Table 1. *The conditional distribution of $Q(t)$ given $Q(t-1)$ (up to order $1/N$)*

Transition	Probability $(q_k(i_1, j_1, i_2, j_2)/2N)$
	1. Transitions resulting from coalescence
$(i_1 - 1)$	$\dfrac{\binom{i_1}{2}}{fx_1}\dfrac{1}{2N}$
$(j_1 - 1)$	$\dfrac{\binom{j_1}{2}}{f(1-x_1)}\dfrac{1}{2N}$
$(i_2 - 1)$	$\dfrac{\binom{i_2}{2}}{(1-f)x_2}\dfrac{1}{2N}$
$(j_2 - 1)$	$\dfrac{\binom{j_2}{2}}{(1-f)(1-x_2)}\dfrac{1}{2N}$
	2. Transitions resulting from mutation and recombination
$(i_1 + 1, j_1 - 1)$	$\dfrac{j_1(\beta + R(1-x_1))x_1}{(1-x_1)}\dfrac{1}{2N}$
$(i_1 - 1, j_1 + 1)$	$\dfrac{i_1(\nu + Rx_1)(1-x_1)}{x_1}\dfrac{1}{2N}$
$(i_2 + 1, j_2 - 1)$	$\dfrac{j_2(\beta + R(1-x_2))x_2}{(1-x_2)}\dfrac{1}{2N}$
$(i_2 - 1, j_2 + 1)$	$\dfrac{i_2(\nu + Rx_2)(1-x_2)}{x_2}\dfrac{1}{2N}$
	3. Transitions resulting from migration
$(i_1 + 1, i_2 - 1)$	$\dfrac{i_2\lambda_{21}x_1}{x_2}\dfrac{1}{2N}$
$(i_1 - 1, i_2 + 1)$	$\dfrac{i_1\lambda_{12}x_2}{x_1}\dfrac{1}{2N}$
$(j_1 + 1, j_2 - 1)$	$\dfrac{j_2\lambda_{21}(1-x_1)}{(1-x_2)}\dfrac{1}{2N}$
$(j_1 - 1, j_2 + 1)$	$\dfrac{j_1\lambda_{12}(1-x_2)}{(1-x_1)}\dfrac{1}{2N}$

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

this first occurs, T_0, is that generation which has the most recent common ancestor of the sample.

The Q process is a jump process and so we define T_1, $T_2, \ldots$ to be the number of generations between successive jumps and $Z_1, Z_2, \ldots$ the successive states to which the process moves. The forces of mutation, recombination and migration change the state of the Q process without changing the number of ancestral genes. A common ancestor event, on the other hand, decreases the number of ancestral genes by 1. The conditional distribution of $Q(t)$ given $Q(t-1)$ can be calculated up to order $1/N$ using the arguments in Kaplan *et al.* (1988) and Hudson & Kaplan (1988). There are twelve transitions whose probabilities are of order $1/N$. Each of the twelve transitions corresponds to state changes where i_1, j_1, i_2, and j_2 either increase by 1, decrease by 1 or remain the same. To simplify the notation only those components that change will be indicated. For each (i_1, j_1, i_2, j_2) the 12 transitions are labelled from 1 to 12 in some specified order and the conditional probability of the kth transition is denoted by

$$q_k(i_1, j_1, i_2, j_2)/2N \quad (1 \leqslant k \leqslant 12).$$

The $\{q_k(i_1, j_1, i_2, j_2)/2N\}$ are given in Table 1.

If time is measured in units of $2N$ generations, then the Q process can be approximated by a continuous time finite state Markov process whose parameters can be obtained from Table 1. Indeed, let

$$q_{i_1 j_1 i_2 j_2} = \sum_{k=1}^{12} q_k(i_1, j_1, i_2, j_2)$$

and

$$p_{i_1 j_1 i_2 j_2}(k) = \frac{q_k(i_1, j_1, i_2, j_2)}{q_{i_1 j_1 i_2 j_2}} \quad (1 \leqslant k \leqslant 12).$$

It follows by standard arguments that the holding time in state (i_1, j_1, i_2, j_2) has an exponential distribution with parameter $q_{i_1 j_1 i_2 j_2}$ and when a jump does occur, the probability that it is the kth transition is

$$p_{i_1 j_1 i_2 j_2}(k) \quad (1 \leqslant k \leqslant 12).$$

Let T denote the sum of the lengths (measured in units of $2N$ generations) of the branches of the ancestral tree for the region in question. It is not hard to show that

$$T = \int_0^{T_0} |Q(u)| du, \tag{6}$$

where T_0 is the time (measured in units of $2N$ generations) of the most recent common ancestor of the sample. Using the representation of T in (6) together with the Markov property of the Q process, it is not difficult to derive equations to calculate the quantities in (3). As an example we will show how to calculate $P(S=0)$ for a sample of size 2, a quantity which we will need in the next section.

For any (i_1, j_1, i_2, j_2) such that $i_1 + j_1 + i_2 + j_2 = 2$, we define

$$H(i_1, j_1, i_2, j_2) = E(e^{-\frac{1}{2}\theta T} | Q(0) = (i_1, j_1, i_2, j_2)),$$

where $\theta = 4N\mu$. It follows from (6) that

$$T = 2T_1 + \int_{T_1}^{T_0} |Q(u)| du.$$

Using the Markov structure of the Q process, we obtain

$$H(i_1, j_1, i_2, j_2) = \left(\frac{q_{i_1 j_1 i_2 j_2}}{q_{i_1 j_1 i_2 j_2} + \theta}\right) \times E(H(Z_1) | Q(0) = (i_1, j_1, i_2, j_2)), \tag{7}$$

where the distribution of Z_1 given $Q(0) = (i_1, j_1, i_2, j_2)$ is given by the $\{p_{i_1 j_1 i_2 j_2}(k), 1 \leqslant k \leqslant 12\}$. Equation (7) defines a system of ten equations that can be solved numerically for any choice of the parameters.

We next consider the case of high migration. More specifically suppose that

$$\lambda_{12} = M\delta_{12} \quad \text{and} \quad \lambda_{21} = M\delta_{21},$$

where $\delta_{12} > 0$, $\delta_{21} > 0$, and M is large. To simplify the discussion we assume that the sample is of size 2. For $t > 0$, (measured in units of $2N$ generations) let $W(t) = (i_1 + i_2, j_1 + j_2)$ whenever $Q(t) = (i_1, j_1, i_2, j_2)$. This process keeps track of how many ancestral genes are linked to the A_1 and A_2 alleles and ignores which subpopulation the genes are in. The W process moves between the states $(2,0)$, $(1,1)$ and $(0,2)$ until a common ancestor event occurs and then the process moves to either of the absorbing states $(1,0)$ or $(0,1)$. Whenever a mutation event, recombination event or a common ancestor event occurs, the W process changes state, but it does not change state when a migration event occurs.

We now show how to calculate the infinitesimal probabilities of the W process. Let $t > 0$. To demonstrate the argument, we assume for definiteness that $W(t) = (2,0)$. In the next small interval of time the process can remain at $(2,0)$, jump to $(1,1)$ if a mutation or recombination event occurs or jump to $(1,0)$ because of a common ancestor event. Suppose for example, that a mutation or recombination event occurs in $(t, t+\Delta)$, $\Delta > 0$. It follows from the dynamics of the Q process that

$$\begin{aligned} P(W(t+\Delta) = (1,1) \mid W(t) = (2,0)) \\ = P(W(t+\Delta) = (1,1) \mid Q(t) = (2,0,0,0))\, P_1(t) \\ + P(W(t+\Delta) = (1,1) \mid Q(t) = (1,0,1,0))\, P_2(t) \\ + P(W(t+\Delta) = (1,1) \mid Q(t) = (0,0,2,0))\, P_3(t), \end{aligned} \tag{8}$$

where

$$P_1(t) = P(Q(t) = (2,0,0,0) \mid W(t) = (2,0)),$$

$$P_2(t) = P(Q(t) = (1,0,1,0) \mid W(t) = (2,0))$$

and

$$P_3(t) = P(Q(t) = (0,0,2,0) \mid W(t) = (2,0)).$$

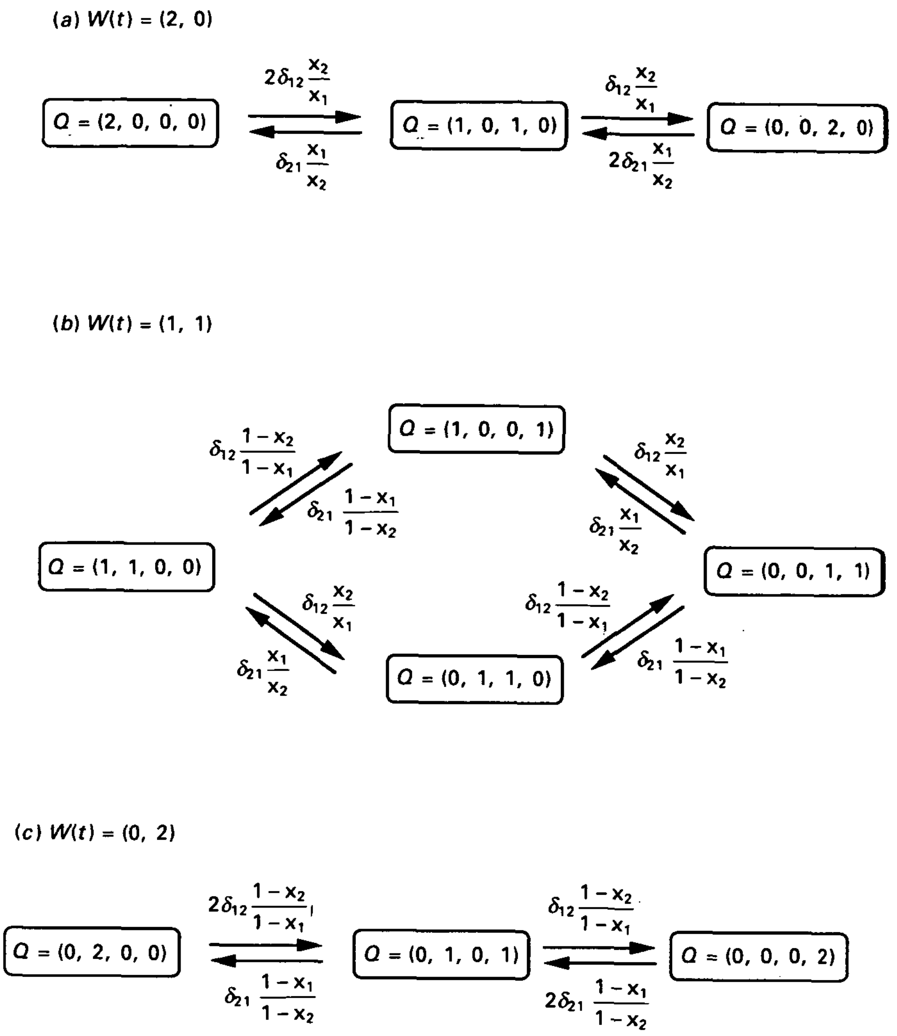
(a) W(t) = (2, 0)
Q = (2, 0, 0, 0)
$2\delta_{12}\frac{x_2}{x_1}$
$\delta_{21}\frac{x_1}{x_2}$
Q = (1, 0, 1, 0)
$\delta_{12}\frac{x_2}{x_1}$
$2\delta_{21}\frac{x_1}{x_2}$
Q = (0, 0, 2, 0)
(b) W(t) = (1, 1)
Q = (1, 0, 0, 1)
$\delta_{12}\frac{1-x_2}{1-x_1}$
$\delta_{21}\frac{1-x_1}{1-x_2}$
$\delta_{12}\frac{x_2}{x_1}$
$\delta_{21}\frac{x_1}{x_2}$
Q = (1, 1, 0, 0)
Q = (0, 0, 1, 1)
$\delta_{12}\frac{x_2}{x_1}$
$\delta_{21}\frac{x_1}{x_2}$
$\delta_{12}\frac{1-x_2}{1-x_1}$
$\delta_{21}\frac{1-x_1}{1-x_2}$
Q = (0, 1, 1, 0)
(c) W(t) = (0, 2)
Q = (0, 2, 0, 0)
$2\delta_{12}\frac{1-x_2}{1-x_1}$
$\delta_{21}\frac{1-x_1}{1-x_2}$
Q = (0, 1, 0, 1)
$\delta_{12}\frac{1-x_2}{1-x_1}$
$2\delta_{21}\frac{1-x_1}{1-x_2}$
Q = (0, 0, 0, 2)

Fig. 1. The transition rates of the migration process for different states of $W(t)$.

Using the transition probabilities of the Q process in Table 1, we obtain

$$
\begin{aligned}
&P(W(t+\Delta) = (1,1) \mid W(t) = (2,0))\\
&= \Delta\left[\frac{2(\nu+Rx_1)(1-x_1)}{x_1}P_1(t)\right.\\
&+\frac{2(\nu+Rx_2)(1-x_2)}{x_2}P_3(t)\\
&\left.+\left(\frac{(\nu+Rx_1)(1-x_1)}{x_1}+\frac{(\nu+Rx_2)(1-x_2)}{x_2}\right)P_2(t)\right]\\
&+O(\Delta^2). \qquad (9)
\end{aligned}
$$

To complete the calculation of $P(W(t+\Delta) = (1,1) \mid W(t) = (2,0))$ we must evaluate the $\{P_i(t)\}$, and to do this we use the assumption of high migration. Let T' denote the time of the change in state of the W process that is closest to t, as one approaches t from the present (i.e. $T' < t$). If M is sufficiently large, then with high probability a large number of migration events will occur in the interval (T', t). Thus the $\{P_i(t)\}$ can be approximated by the stationary distribution of the three state Markov process describing the dynamics of the migration process of the lineages of two genes, each linked to an A_1 allele. The transition rates of this Markov process are given in Fig. 1*a*. The stationary probabilities can be obtained using standard arguments (Karlin, 1969) and are

$$P_1(t) = \Omega_1^2, \quad P_2(t) = 2\Omega_1(1-\Omega_1)$$

and

$$P_3(t) = (1-\Omega_1)^2,$$

where

$$\Omega_1 = \frac{\delta_{21}x_1^2}{\delta_{21}x_1^2+\delta_{12}x_2^2} = \frac{x_1^2}{x_1^2+(\delta_{12}/\delta_{21})x_2^2}.$$

The quantity Ω_1 can be interpreted as the proportion of the time that the lineage of an A_1 allele spends in subpopulation 1. If we substitute these values of $P_1(t)$, $P_2(t)$ and $P_3(t)$ in (9), then we obtain

$$
\begin{aligned}
&P(W(t+\Delta) = (1,1) \mid W(t) = (2,0))\\
&= q_{[2,0]}[1,1]\Delta+O(\Delta^2),
\end{aligned}
$$

where

$$
\begin{aligned}
&q_{[2,0]}[1,1]\\
&= \frac{2(\nu+Rx_1)(1-x_1)}{x_1}\Omega_1^2\\
&+\frac{2(\nu+Rx_2)(1-x_2)}{x_2}(1-\Omega_1)^2
\end{aligned}
$$

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

$$+\left(\frac{(\nu+Rx_1)(1-x_1)}{x_1}+\frac{(\nu+Rx_2)(1-x_2)}{x_2}\right)2\Omega_1(1-\Omega_1). \qquad (10)$$

The other five infinitesimal probabilities are calculated in the same way as (10) and these quantities are given in Table 2. It should be noted that the stochastic process describing the dynamics of the migration process depends on the state of $W(t)$. In Fig. 1 (*b*, *c*) the transition rates of the Markov processes describing the dynamics of the migration processes are given for the cases $W(t) = (1,1)$ and $W(t) = (0,2)$ respectively. The stationary probabilities for the Markov process in Fig. 1 *c* are the same as those for the process described in Fig. 1 *a* except that Ω_1 is replaced by Ω_2 where

$$\Omega_2 = \frac{\delta_{21}(1-x_1)^2}{\delta_{21}(1-x_1)^2+\delta_{12}(1-x_2)^2} = \frac{(1-x_1)^2}{(1-x_1)^2+(\delta_{12}/\delta_{21})(1-x_2)^2}.$$

The quantity Ω_2 can be interpreted as the proportion of the time that the lineage of an A_2 allele spends in subpopulation 1. The stationary probabilities for the process described in Fig. 1 *b* are

$\Omega_1\Omega_2$, $\Omega_1(1-\Omega_2)$, $\Omega_2(1-\Omega_1)$ and $(1-\Omega_1)(1-\Omega_2)$.

The advantage of the high migration case is that only three equations are needed to calculate each quantity in (3). For example, let

$$H[i,j] = E(e^{-\frac{1}{2}\theta T}\mid W(0) = (i,j)) \quad (i+j = 2).$$

Then,

$$H[2,0] = \frac{q_{[2,0]}[1,0]}{q_{[2,0]}+\theta}+\frac{q_{[2,0]}[1,1]}{q_{[2,0]}+\theta}H[1,1],$$

$$H[0,2] = \frac{q_{[0,2]}[0,1]}{q_{[0,2]}+\theta}+\frac{q_{[0,2]}[1,1]}{q_{[0,2]}+\theta}H[1,1],$$

and

$$H[1,1] = \frac{\dfrac{q_{[1,1]}[2,0]\,q_{[2,0]}[1,0]}{q_{[2,0]}+\theta}+\dfrac{q_{[1,1]}[0,2]\,q_{[0,2]}[0,1]}{q_{[0,2]}+\theta}}{q_{[1,1]}+\theta-\dfrac{q_{[1,1]}[2,0]\,q_{[2,0]}[1,1]}{q_{[2,0]}+\theta}-\dfrac{q_{[1,1]}[2,0]\,q_{[2,0]}[1,1]}{q_{[0,2]}+\theta}}, \qquad (11)$$

where

$$q_{[2,0]} = q_{[2,0]}[1,1]+q_{[2,0]}[1,0],$$
$$q_{[0,2]} = q_{[0,2]}[1,1]+q_{[0,2]}[1,0]$$

and

$$q_{[1,1]} = q_{[1,1]}[2,0]+q_{[1,1]}[0,2].$$

For a population with two subpopulations, the coalescent process for a sample of size 2 at a selectively neutral locus, not linked to any selected locus, behaves, in the high migration case, as if the sample was taken from a panmictic population with an effective population size equal to

$$2N\left(\frac{\Omega^2}{f}+\frac{(1-\Omega)^2}{1-f}\right)^{-1},$$

where $2Nf$ and $2N(1-f)$ are the sizes of the subpopulations and Ω and $1-\Omega$ are the stationary

Table 2. *The infinitesimal probabilities of the W process*

$$\begin{aligned}
q_{[2,0]}[1,1] &= \frac{2(\nu+Rx_1)(1-x_1)}{x_1}\Omega_1^2+\frac{2(\nu+Rx_2)(1-x_2)}{x_2}(1-\Omega_1)^2\\
&\quad+\left(\frac{(\nu+Rx_1)(1-x_1)}{x_1}+\frac{(\nu+Rx_2)(1-x_2)}{x_2}\right)2\Omega_1(1-\Omega_1)\\
q_{[2,0]}[1,0] &= \frac{\Omega_1^2}{fx_1}+\frac{(1-\Omega_1)^2}{(1-f)\,x_2}\\
q_{[0,2]}[1,1] &= \frac{2(\beta+R(1-x_1)\,x_1}{(1-x_1)}\Omega_2^2+\frac{2(\beta+R(1-x_2))\,x_2}{(1-x_2)}(1-\Omega_2)^2\\
&\quad+\left(\frac{(\beta+R(1-x_1))\,x_1}{(1-x_1)}+\frac{(\beta+R(1-x_2))\,x_2}{(1-x_2)}\right)2\Omega_2(1-\Omega_2)\\
q_{[0,2]}[0,1] &= \frac{\Omega_2^2}{f(1-x_1)}+\frac{(1-\Omega_2)^2}{(1-f)(1-x_2)}\\
q_{[1,1]}[2,0] &= \frac{(\beta+R(1-x_1))\,x_1}{(1-x_1)}\Omega_2+\frac{(\beta+R(1-x_2))\,x_2}{(1-x_2)}(1-\Omega_2)\\
q_{[1,1]}[0,2] &= \frac{(\nu+Rx_1)(1-x_1)}{x_1}\Omega_1+\frac{(\nu+Rx_2)(1-x_2)}{x_2}(1-\Omega_1)
\end{aligned}$$

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

probabilities of the associated migration process. In this case $\Omega = m_{21}/(m_{21}+m_{12})$. If the selectively neutral locus is linked to a selected locus, then the probabilities in Table 2 show that the high migration limit does not, in general, behave as if the sample was taken from a panmictic population with a different effective population size. The case where $x_1 = x_2$ is the one case where samples do behave like samples from a single panmictic population.

3. An Application

Hudson & Kaplan (1988) recently examined the spatial distribution of nucleotide variation in the *Adh* region of *D. melanogaster*. The data analysed was that of Kreitman (1983) which consists of the sequences of eleven cloned alleles from flies collected from widely separated localities around the world, including localities in Japan, the United States and France. Six of the sequences coded for a slow electromorph of *Adh*, and five coded for a fast electromorph. These will be referred to as the Slow and Fast sequences, respectively. Hudson and Kaplan compared the observed distribution of variation in the *Adh* region to predictions based on a model with balancing selection acting on a single nucleotide polymorphism in a single panmictic population. In their model, the Slow allele is maintained at a frequency of $x_0 = 0{\cdot}7$. The assumption of panmixis with a single allele frequency is unrealistic given the well documented latitudinal cline of frequencies of *Adh* electromorphs (Oakeshott *et al.* 1982). The observation of similar latitudinal clines in the frequency of the *Adh* electromorphs on three continents is one of the strongest lines of evidence for the importance of selection in the maintenance of the *Adh* polymorphism. With the theory presented in the previous section, we can examine a more realistic model than that considered by Hudson and Kaplan, namely a model in which there are two subpopulations, each with a different frequency of the Slow electromorph.

The predictions of the high-migration model analysed in the previous section will be compared to the observed variation. The use of the high migration model is motivated by the observation of Simmons *et al.* (1989) (see also Slatkin, 1987) that, other than the electromorph frequency differences between populations, there is little apparent differentiation between populations, at least in the USA. This pattern of variation suggests that migration rates are high, but that selection is sufficiently strong to maintain the allele frequency cline despite the migration.

To compare the predictions of the single panmictic population model and the geographic subdivision model to the observed pattern of variation, the 'sliding window' method employed by Hudson and Kaplan will be used. In this method, one calculates for each nucleotide site, a measure π (defined below) of the observed variation in a small window centred on that nucleotide site. If one numbers the nucleotide sites in the sequence consecutively, then $\pi(k)$, the value of π for the kth nucleotide site, can be plotted as a function of k. As Hudson and Kaplan demonstrate, this graphical presentation is very effective in displaying regions of excessively high or low variation. Since the region sequenced by Kreitman contained both coding and noncoding sequences, Hudson and Kaplan varied the size of the window around each nucleotide site so as to keep the number of possible silent changes in the window constant. In Fig. 2, the window size is adjusted so that there are always 150 possible silent nucleotide changes in the window.

For Kreitman's *Adh* data, Hudson and Kaplan considered three definitions of $\pi(k)$: $\pi_{FS}(k)$, $\pi_{FF}(k)$ and $\pi_{SS}(k)$, which are, respectively, the average number of pairwise differences in a window centred on nucleotide k between Fast and Slow sequences, the average number between only Fast sequences and the average number between only Slow sequences. The plot of the observed values of $\pi_{FS}(k)$ in Fig. 2 shows a very large peak in a small region containing the Fast/Slow polymorphism at codon 192.

Assuming a balanced polymorphism at the Fast/Slow locus and a uniform rate of recombination throughout the region (i.e. the average number of crossovers between nucleotide site i and j is $|i-j|R_0/2N$), Hudson and Kaplan calculated the expectation of $\pi_{FS}(k)$, $\pi_{SS}(k)$ and $\pi_{FF}(k)$ assuming a single panmictic population with the frequency of the slow allele maintained at a frequency of $x_0 = 0{\cdot}7$, and assuming $\beta = \nu = 0{\cdot}001$, $\theta = 0{\cdot}006$ and $R_0 = 0{\cdot}012$. (Hudson & Kaplan denoted β and ν by β_1 and β_2, respectively.) The value of x_0 was chosen rather arbitrarily as an approximate average frequency of the slow allele in many populations. The other parameter values were rough estimates obtained from prior analyses. The predicted values of $\pi_{FS}(k)$ under this model are shown in Fig. 2 by the lower hatched curve, labelled 'One population ($R = 0{\cdot}012$)'. With these parameter values the predicted level of variation is considerably smaller than the observed variation in the vicinity of the Fast/Slow polymorphism. A much better fit was obtained with a much lower level of recombination, namely when $R_0 = 0{\cdot}002$, as shown by the upper hatched curve in Fig. 2, labelled 'One population ($R = 0{\cdot}002$)'.

Using the results of the previous section, we can investigate the consequences of geographic subdivision on the expected level of variation in the neighbourhood of a balanced polymorphism. Indeed, all that has to be done is replace $H(1,1)$, $H(2,0)$ and $H(0,2)$ by $H[1,1]$, $H[2,0]$ and $H[0,2]$ respectively, in the equations given by Hudson and Kaplan for $E(\pi_{FS}(k))$, $E(\pi_{FF}(k))$ and $E(\pi_{SS}(k))$. Thus

$$E(\pi_{FS}(k)) = \sum_{m \in W_k} (1 - H_m[1,1]),$$

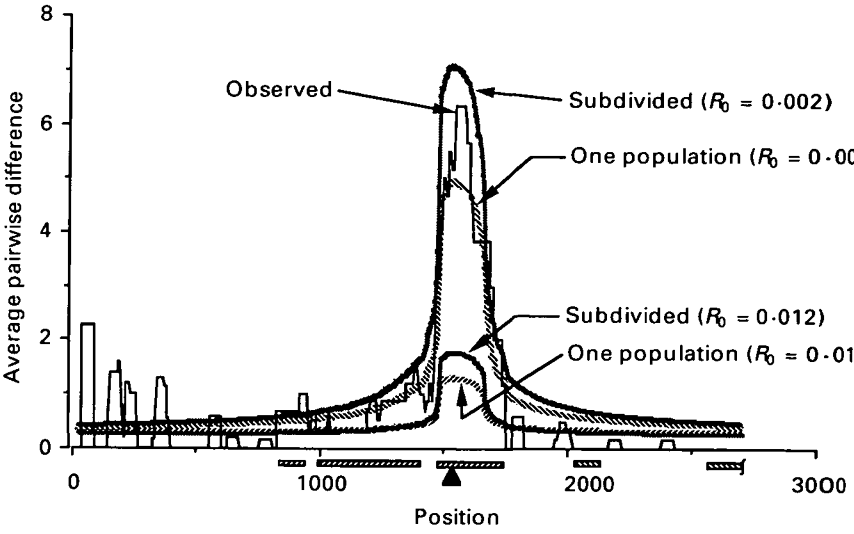

Fig. 2. The observed and predicted nucleotide variation in the *Adh* region of *D. melanogaster*. The observed variation, shown by the thin solid line, is the average pairwise number of differences between a Slow and Fast allele in a small sliding window (see text). The horizontal axis is the nucleotide position of the centre of the window. The predicted pairwise numbers of differences for the subdivided population models are shown by the thick grey lines labelled 'Subdivided'. The predictions of the single panmictic population model are shown by the thick hatched lines, labelled 'One population'. The hatch bars below the position axis show the coding exons of the *Adh* locus and of another apparent coding region 3′ to *Adh*. The triangle marks the position of the Fast/Slow protein polymorphism of *Adh*. The parameter values of the predicted curves of the single panmictic population model are $\beta = 0{\cdot}001$, $\theta_0 = 0{\cdot}006$, $x_0 = 0{\cdot}7$, and R_0 equal to either 0·002 or 0·012. For the subdivided population model, the parameters are the same except $x_1 = 0{\cdot}95$, $x_2 = 0{\cdot}5$, $f = 0{\cdot}5$, and $\delta_{12}/\delta_{21} = 1{\cdot}0$.

where W_k is the set of sites in the window centred at k and $H_m[1,1]$ is the value of $H[1,1]$ for nucleotide m. Similarly,

$$E(\pi_{FF}(k)) = \sum_{m \in W_k} (1 - H_m[2,0]),$$

and

$$E(\pi_{SS}(k)) = \sum_{m \in W_k} (1 - H_m[0,2]).$$

To evaluate $H_m[1,1]$, $H_m[2,0]$ and $H_m[0,2]$ we need, in addition to the values of β, ν, θ and R_0 given by Hudson and Kaplan, estimates of x_1, x_2, f and δ_{12}/δ_{21}. If x_1 denotes the frequency of Slow in subpopulation 1 (a southern population, say) and x_2, the frequency of Slow in subpopulation 2 (a northern population), then reasonable estimates are $x_1 = 0{\cdot}95$ and $x_2 = 0{\cdot}5$ (Oakeshott *et al.* 1982; Simmons *et al.* 1989; Vigue & Johnson 1973). In Fig. 2, $E(\pi_{FS}(k))$ is plotted for $x_1 = 0{\cdot}95$, $x_2 = 0{\cdot}5$, $f = 0{\cdot}5$, $\delta_{12}/\delta_{21} = 1{\cdot}0$ and $R_0 = 0{\cdot}012$ (lower thick grey curve) and $R_0 = 0{\cdot}002$ (upper thick grey curve). The subdivision model provides a slightly better fit with the *a priori* estimate of $R_0 = 0{\cdot}012$, but still does not predict a peak as high as the observed peak. At $R_0 = 0{\cdot}002$, the subdivision model predicts a considerably higher peak than the one panmictic population model. A curve like that predicted under the panmictic model with $R_0 = 0{\cdot}002$, can be obtained under the subdivision model with $R_0 = 0{\cdot}003$, a value of R_0 somewhat closer to the *a priori* estimate of this parameter.

The above values of f and δ_{12}/δ_{21} were chosen rather arbitrarily due to the lack of information concerning these parameters. One might surmise that the southern population of *D. melanogaster* has a larger effective population size and that migration rates to the north might be higher than in the other direction. Indeed, $f = \delta_{12}/\delta_{21} = 0{\cdot}6$, gives a slightly higher predicted peak than when these parameters are 0·5, the values used above. With $f = \delta_{12}/\delta_{21} = 0{\cdot}6$, the best fit to the observed variation is obtained with $R_0 = 0{\cdot}0035$, rather than 0·003, slightly closer to the prior estimate of R_0. However, no further improvement of the fit was obtained with larger values of f or smaller values of δ_{12}/δ_{21}. For the parameter values examined, the effect of subdivision on $E(\pi_{SS}(k))$ and $E(\pi_{FF}(k))$ is relatively minor and is not shown.

We conclude that the subdivision model gives only a slightly better fit to the observed variation in the *Adh* region than the single panmictic population model. More precisely, under the subdivision model, the value of the recombination parameter which gives the best fit is slightly closer to our prior estimate of this parameter.

The effects of migration can be much larger when the frequencies of the selected alleles are much more different in the two subpopulations. For example, if the frequencies of the Slow allele were 0·95 and 0·1 in the two subpopulations (instead of 0·95 and 0·5), then

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

the effect of subdivision on recombination would be much greater, with the best fit to the data obtained with R_0 approximately equal to 0·006, approximately three times the best fitting value found by Hudson and Kaplan. Subdivision reduces the effective recombination rate. This is because with subdivided populations, a higher proportion of the total population is homozygous at the selected locus due to the Wahlund effect (Hartl & Clark, 1989). Since only recombination in heterozygous individuals is effective in reducing the divergence between selected alleles, effective recombination is reduced in the subdivided populations.

4. Conclusion

The preceding analysis indicates how the coalescent process can be studied in models with both selection and migration, as well as with recombination and mutation. Each of these forces introduces parameters into the model, and assigning them values can be problematic in any application. Some simplification is possible if the migration rates are high, and so we have limited our attention to this case.

The analysis shows that the predicted patterns of variation in the *Adh* region of *Drosophila melanogaster* are not greatly changed by incorporating subdivision into the model of Hudson and Kaplan. Under the single panmictic population model of Hudson and Kaplan, the best fit of the model to the data was obtained with $R_0 = 0{\cdot}002$, a factor of six smaller than the *a priori* estimate of this parameter. Under the subdivision model, an equally good fit to the data is obtained with a recombination rate, $R_0 = 0{\cdot}003$, which is somewhat closer to the *a priori* estimate of this recombination parameter. The small difference between the two values of R_0 suggests that geographic subdivision does not account for the difference between the *a priori* estimate of R_0 and the value predicted by Hudson and Kaplan using a single panmictic population model.

Subdivision has the effect of reducing the effective recombination rate when the frequencies of the selectively maintained alleles are very different in the different subpopulations. This is because, in this context, recombination has an effect only when it occurs between gametes with different alleles at the selected locus. When the population is subdivided with different frequencies of the selected allele, a higher proportion of individuals are homozygous at the selected locus (Wahlund effect), and thus effective recombination occurs less frequently than it would in a single population with average allele frequencies at the selected locus.

References

Hartl, D. L. & Clark, A. G. (1989). *Principles of Population Genetics*, 2nd edn. Sunderland, Mass.: Sinauer Associates.

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

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

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

Karlin, S. (1969). *A First Course in Stochastic Processes*, New York: Academic Press.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Kreitman, M. & Aguadé, M. (1986). Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. *Genetics* **114**, 93–110.

Oakeshott, J. G., Gibson, J. B., Anderson, P. R., Knibb, W. R., Anderson, D. G. & Chambers, G. K. (1982). Alcoholdehydrogenaseandglycerol-3-phosphatedehydrogenase clines in *Drosophila melanogaster* on three continents. *Evolution* **36**, 86–96.

Simmons, G. M., Kreitman, M. E., Quattlebaum, W. F. & Miyashita, N. (1989). Molecular analysis of the alleles of alcohol dehydrogenase along a cline in *Drosophila melanogaster*. I. Maine, North Carolina, and Florida. *Evolution* **43**, 393–409.

Slatkin, M. (1987). The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42–49.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.

Vigue, C. L. & Johnson, F. M. (1973). Isozyme variability in species of the genus *Drosophila*. VI. Frequency – Property – Environment relationships of allelic alcohol dehydrogenases in *D. melanogaster*. *Biochemical Genetics* **9**, 213–227.

Watterson, G. A. (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **10**, 256–276.

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press

6–11 October 1991 Washington, DC

The 8th International Congress of Human Genetics, sponsored by the American Society of Human Genetics, will be held at the Washington, DC Convention Center. The program will include 9 plenary lectures, 16 symposia and more than 50 workshops. Approximately 3,000 submitted abstracts will be presented in varying formats including poster, slide, workshop and poster symposium sessions. All aspects of both clinical and basic human genetics research, diagnosis and treatment will be covered. The deadline for receipt of abstracts is 1 April 1991. To obtain detailed information and official forms for abstract submission and registration contact M. Ryan, Meetings Manager, ICHG, 9650 Rockville Pike, Bethesda, MD 20814 USA [Telephone (301) 571-1825; Fax (301) 530-7079].

https://doi.org/10.1017/S0016672300029074 Published online by Cambridge University Press