Gene Genealogies Within Mutant Allelic Classes

Montgomery Slatkin

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Manuscript received September 27, 1995
Accepted for publication February 12, 1996

ABSTRACT

A coalescent theory of the gene genealogy within an allelic class that arises by a unique mutational event is developed and analyzed. To interpret this theory it was necessary to expand on existing theory for populations of varying size. Two features of the gene genealogy—the average pairwise distance and the total tree length—within the mutant class and within the nonmutant class are found. An index, I, is proposed that describes the extent to which a genealogy is similar to one from a population of constant size (for which $I = 0$) or to a star genealogy (for which $I = 1$). The value of I is positive in growing populations and is generally positive for the gene genealogy for the mutant class. The value of I is negative for a population decreasing in size and for the nonmutant class, if the mutant arose recently. The results are discussed in the context of the infinite sites model of mutation, which is appropriate for nucleotide sequence data, and the generalized stepwise mutation model, which is appropriate for microsatellite loci. The same genealogical methods are used to find the probability of at least one recombination event between a nucleotide that defines an allelic class and a marker at a nearby linked site.

THE widespread use of methods from molecular biology permits the examination of genetic variation at different hierarchical levels. In this paper, I will be concerned with the variation within an allelic class that has arisen by a unique mutation at some time in the past. The goal is to develop the theory of intraallelic variation that can be used to learn about the history of the allelic class and possibly about the selective forces it has experienced. This paper will concentrate on some features of the gene genealogy within an allelic class and the similarity of that gene genealogy to one from a population that has changed in size.

For the purposes here, an allelic class is one in which every copy carries a sequence of DNA that distinguishes it from all other copies. That sequence could be a single nucleotide in a coding region that results in a detectably different phenotype, as in the case of the F/S polymorphism in Adh in *Drosophila melanogaster* (KREITMAN 1983). The F allele has lysine at codon position 192 while the S allele has threonine. KREITMAN (1983) found substantial variation at other nucleotide positions within each allelic class. Alternatively, an allelic class could be distinguished by more complex changes in sequence. GARZA *et al.* (1995) found that alleles at a microsatellite locus in chimpanzees differed by more than just the number of dinucleotide repeats.

The coalescent theory of intraallelic variation was initiated by SAUNDERS *et al.* (1984) and HUDSON and KAPLAN (1986). Those papers computed the expected number of segregating sites or the expected tree length of an allelic class conditioned on the configuration of allelic classes. In later work, HUDSON and KAPLAN (1988) and KAPLAN *et al.* (1988) developed a theory of genealogical structure of selected alleles, with the goal in part to extract more information from the KREITMAN (1983) data set. In that theory, the nucleotide position that defined the allelic class was subject to reversible mutation. The model became equivalent to a model of two subpopulations with gene flow between them caused by mutation at the critical nucleotide position. HUGHES and NEI (1988) used the extent of intraallelic variation in HLA class I loci within and between species to argue that overdominant selection affected those loci. CLARK (1993) and WEKEMANS and SLATKIN (1994) carried out simulation studies of gene genealogies within allelic classes of gametic self-incompatibility alleles. ETHIER and SHIGA (1994) develop some formal theory in intraallelic gene genealogies under the assumption that alleles arise via the infinite alleles model of mutation.

In this paper I will develop analytic methods for predicting some features of the gene genealogies of allelic classes, under the assumption that the allelic class arose by a unique mutation at a known time in the past. The goal of this analysis is similar to that of HUDSON and KAPLAN (1986) but differs in assuming that the time at which the allelic class arose by mutation is known. HUDSON and KAPLAN (1986) assumed an infinite alleles model of mutation and assumed that the allelic class of interest was randomly chosen. The time or origin of an allelic class by mutation is usually unknown, but a theory in which the time is assumed known is necessary to estimate the time of origin from information about genetic variation within the allelic class and to predict

Corresponding author: Montgomery Slatkin, Department of Integrative Biology, VLSB 3060, University of California, Berkeley, CA 94720-3140. E-mail: slatkin@garnet.berkeley.edu

the extent of linkage disequilibrium with closely linked markers.

In this paper, I will concentrate on the average pairwise coalescence time of two randomly chosen copies within an allelic class and on the total length of the gene genealogy of an allelic class. It is possible to find the expected values and variances of these quantities both within the mutant allelic class and within the nonmutants. I have chosen these quantities because under the assumptions of some mutation models, particularly the infinite sites model and the stepwise mutation model, one or both of them can be estimated from data without the necessity of inferring the gene genealogy itself. I will also consider the problem of recombination between a nucleotide that defines an allelic class and a linked marker.

One question will be the extent to which an intraallelic gene genealogy is similar to one generated by the coalescent process in a population of constant size or to a star genealogy. I introduce an index to quantify the degree of similarity. Before considering intraallelic genealogies, it will be useful to review and extend existing results for gene genealogies in populations of varying size.

FLUCTUATING POPULATION SIZE

Total tree length: We are concerned with the ancestry of a sample of i copies of a nonrecombining locus in a population whose size has changed in the past. Let $t = 0$ be the present and $N(t)$ be the population size at time t in the past. $N(t)$ must be sufficiently large that the diffusion approximation, defined below, is accurate for all t. The theory developed by KINGMAN (1982) and summarized by TAVARÉ (1984) shows that the ancestry can be described by a simple Markov process, for which the general solution can be obtained. Let $g_{ij}(t)$ be the probability that there are j ancestors at t, given that i copies are sampled at 0. Under the assumption that $N(t)$ is large for all t, the equations satisfied by these probabilities are

$$2N(t)\frac{dg_{ii}(t)}{dt} = -\binom{i}{2} g_{ii}(t)$$

$$2N(t)\frac{dg_{ij}(t)}{dt} = -\binom{j}{2} g_{ij}(t) + \binom{j+1}{2} g_{i,j+1}, \quad 2 \le j < i$$

$$2N(t)\frac{dg_{i1}(t)}{dt} = g_{i2}(t). \tag{1}$$

The diffusion approximation mentioned above requires that $N(t)$ be sufficiently large that no more than one coalescent event can occur in any generation. If that approximation cannot be made, than additional terms in (1) are needed to allow for the occurrence of two or more coalescent events.

The solution to (1) can be found by first changing the time scale to remove the $N(t)$ (GRIFFITHS and TAVARÉ 1994, Equation 3). Let

$$\tau(t) = \int_0^t \frac{dt'}{2N(t')} \tag{2}$$

so $\tau = 0$ corresponds to $t = 0$ and

$$2N(t)\frac{dg_{ij}(t)}{dt} = \frac{dg_{ij}(\tau)}{d\tau}. \tag{3}$$

With this change of variables, the analytic results developed for populations of constant size can be used to provide the general solution to (1):

$$g_{ij}(\tau) = \sum_{k=j}^{i} \rho_k(\tau)\frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)!\, i_{(k)}}, \quad 2 \le j \le i$$

$$g_{i1}(\tau) = 1 - \sum_{k=2}^{i} \rho_k(\tau)\frac{(2k-1)(-1)^{k} i_{[k]}}{i_{(k)}} \tag{4}$$

where $\rho_k(t) = e^{-k(k-1)t/2}$, $a_{(j)} = a(a+1)\cdots(a+j-1)$, and $a_{[j]} = a(a-1)\cdots(a-\mathrm{j}+1)$ (TAVARÉ 1984, Equations 6.1 and 6.2). To express g_{ij} as functions of t, we substitute for $\tau(t)$ using (2).

We will be concerned with the total tree length of a sample of size i, L_i, which is the sum of all the branch lengths. The value of L_2, which is twice the average pairwise coalescence time, is of particular interest. Clearly

$$L_i = \sum_{j=2}^{i} jT_j \tag{5}$$

where T_j are time intervals during which there are j ancestors (TAVARÉ 1984). When $N(t)$ is constant, the T_j are independent and have exponential distributions with means $4N/[j(j-1)]$, and hence $E(L_i) = 2N[1 + {}^1/_2 + {}^1/_3 + \cdots + 1/(\mathrm{i}-1)]$ (TAVARÉ 1984). For any other $N(t)$, the expectation of L_i is

$$E(L_i) = \sum_{j=2}^{i} j\int_0^\infty g_{ij}(\tau(t))\,dt. \tag{6}$$

Although (6) cannot be expressed in closed form, it can be evaluated numerically using a computer algebra program.

Average pairwise coalescence time: For many purposes, the average pairwise coalescence time, which is half of the average branch length separating two randomly chosen copies in a sample, will be of interest. If the model allows us to consider samples of arbitrary size, then the expected pairwise coalescence time in a sample is half the expected tree length in a sample of size 2, $E(L_2)/2$. In the model of intraallelic diversity considered below, the sample size is a random variable and hence cannot be assigned a value arbitrarily. For that reason, we need an alternative way to compute the average directly from the g's or their equivalent.

The question is, in a sample of i copies, what is the

average coalescence time of a randomly chosen pair. As above, we let T_j be the time interval during which there are j ancestral lineages present and let t_j be the time at which the coalescence from j to $j-1$ copies occurred ($t_j = T_i + T_{i-1} + \cdots + T_j$). Our randomly chosen pair of copies will be the first pair to coalesce (at t_i) with probability $\pi_i = 2/[i(i-1)] = 1/\binom{i}{2}$ and will not coalesce at that time with probability $1 - \pi_i$. If the coalescence does not occur at t_i, then before t_i, we have two randomly chosen copies from a sample of size $i-1$. By induction, the probability of coalescence at time t_j is $\pi_j(1-\pi_{j+1})\cdots(1-\pi_i)$ Therefore, the expected coalescence time between a pair of copies is the weighted average of the t_k, which reduces to

$$E(P_j) = \sum_{j=2}^{i} \prod_{k=j+1}^{i} (1-\pi_k) E(T_j) \qquad (7)$$

where the product is defined to be 1 for $j = i$. In this expression, the coefficient of $E(T_j)$ is the probability that the coalescence time of a randomly chosen pair depends on T_j. The expected value of P_i for a particular history of the population depends on the expected values of the T_j, the times during which there are j lineages remaining, which we can compute by integrating $g_{ij}(t)$ over all t. Note that for a population of constant size, $E[T_j] = 4N/[j(j-1)] = 2N\pi_j$ and hence $P_i = 2N$ for all $i \geq 2$, as required (TAJIMA 1983).

Stellate index: There is a wide variety of fluctuating population sizes that can be imagined, so it is impossible to generalize about the effects of variation in population size. If $N(t)$ fluctuates randomly and with no trend, then (2) tells us that the ancestry of the sample will be that of a sample from a population of constant size equal to the harmonic mean of $N(t)$. That will be true when we can replace the integral in (2) by $N_e t$, and it is simply a generalization of WRIGHT's (1938) classic result. The range of validity of this approximation could be explored by making different assumptions about patterns of fluctuations in $N(t)$, particularly the extent of temporal autocorrelation.

Here I will be concerned with a population that has grown from a small size to some larger size or the reverse. SLATKIN and HUDSON (1991) emphasized that sufficiently rapid population growth could create a gene genealogy that is starlike, meaning that the external branches are relatively long while the internal branches are relatively short. In contrast, the gene genealogy in a population of constant size has external branches that are relatively short and internal branches that become progressively longer towards the root. We showed that a starlike gene genealogy leads to patterns of nucleotide variation that are different from those in a population of constant size. DI RIENZO *et al.* (1994) reached a similar conclusion for variation in microsatellite allele frequencies. It is of interest then to know whether a gene genealogy is closer to being starlike or closer to being from a population of constant size. We can derive a simple index by noting that in a perfect star genealogy $L_i = iL_2/2$, because every terminal branch has the same length and there are no internal branches, while in a population of constant size, $E(L_i) = E(L_2)[1 + {}^1/_2 + {}^1/_3 + \cdots + 1/(i-1)]$:

$$I = \frac{\dfrac{L_i}{L_2\Sigma_i} - 1}{\dfrac{i}{2\Sigma_i} - 1} \qquad (8)$$

where $\Sigma_i = 1 + {}^1/_2 + {}^1/_3 + \cdots + 1/(i-1)$. Note that I is not defined if $i < 4$. For a randomly mating population, $I = 0$, and for a perfect star genealogy, $I = 1$, provided that the Ls take their expected values.

To illustrate the use of this index, consider the case examined by SLATKIN and HUDSON (1991), $N(t) = N_0 e^{-rt}$, which represents a population of size N_0 today that has grown exponentially at rate r. Figure 1a shows a graph of I as a function of i for different values of $\alpha = 2N_0 r$. We can see that with a reasonably large α I is substantially greater than 0. There is little dependence on i, the sample size. SLATKIN and HUDSON (1991) considered an idealized model for human populations, $N_0 = 10^9$, $r = 0.00488$ per generation (which assumes an exponential increase from 5000, 50,000 years ago with a 20-year generation time), $\alpha = 9.76 \times 10^6$, in which case $I = 0.822$ for $i = 20$. This result confirms the conclusion that very rapid population growth leads to a gene genealogy that is close to a star genealogy.

If the population has declined in size, I can be negative. We should note in this context that in order to obtain reasonable results, it is necessary to assume that $N(t)$ remains finite or increases sufficiently slowly that $\tau(t)$ defined by (2) goes to infinity as t increases. That condition is necessary to ensure that two randomly chosen copies have a most recent common ancestor at a finite time in the past with probability one (GRIFFITHS and TAVARÉ 1994). That condition is not satisfied for the model of exponential growth with $r < 0$, no matter how small r is, so we need to consider a different model of a declining population. A model that has been used in this context is one with a step change in population size at a time T in the past. In these calculations, the absolute population size does not matter because time is measured in units of $2N_0$ generations. Before time T in the past, the population size was rN_0. If $r > 1$, the population size decreased and if $r < 1$ it increased at T. Figure 1b shows the value of I as a function of r. Small values of r can lead to value of I comparable to those in the case of exponential growth. Values of r greater than one lead to negative values of I although for this model only small negative values are found. For given values of T and $r > 1$, I decreased with increasing i (results not shown).

HISTORY OF AN ALLELIC CLASS

Within the mutant class: The starting point is a sample of n copies of the locus at time 0. In that sample we find

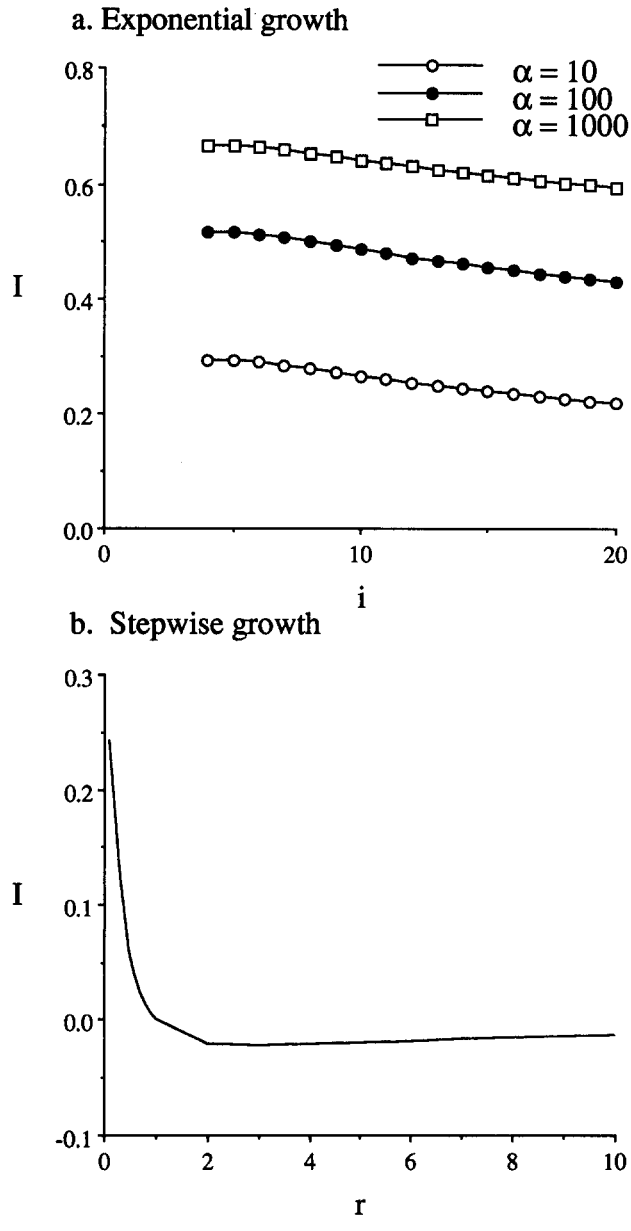

FIGURE 1.—Stellate index in growing and declining populations. (a) Values of I, defined by Equation 8, in a population that has undergone exponential growth, where i is the sample size, $\alpha = 2N_0r$, N_0 is the current population size and r is the rate of exponential increase. These results were obtained by evaluating Equation 6 using a simple program written in *Mathematica* (WOLFRAM 1991). (b) Values of I in a population that has undergone a step increase in proportion r at time $T = N/2$ in the past. Values of $r < 1$ indicate an increase in population size at T, and values of $r > 1$ indicate a decrease. These results were obtained by numerically evaluating Equation 6.

i copies of an allele that arose by a unique mutation at time t in the past. The goal is to find properties of the intraallelic gene genealogy. The theory in the preceding section tells us that the sample of size n will have j ancestral lineages at time t in the past with probability $g_{nj}(t)$. If the allelic class arose by mutation at time t, then one of those lineages, chosen at random, will have all of its descendants in the allelic class. We can derive the distribution of the number of descendants, given j, by using the result derived by KINGMAN (1982) that all configurations of numbers of descendants are equally likely. That allows us to reduce the problem to one of balls being randomly distributed into boxes, subject to the condition that none of the boxes be empty.

Consider the ancestry of the sample at a time t' between 0 and t, and let m be the number of ancestral lineages and k be the number of mutant lineages at t'. Both m and k are random variables, as is j. Our previous results tell us that $\Pr(m) = g_{nm}(t')$ and that $\Pr(j|m) = g_{mj}(t - t')$, where $\Pr(\cdot)$ indicates the probability, possibly conditioned in different ways. Given the values of j and m, the distribution of k is the distribution of the number of balls in a randomly chosen box when m balls are distributed randomly into j boxes, given than none of the boxes is empty. There are a total of $\binom{m-1}{j-1}$ ways in which the balls can be distributed so that no box is empty and of these there are $\binom{m-k-1}{j-2}$ ways that there are exactly k balls in the one box that represents the allelic class (FELLER 1957). Therefore

$$\Pr(k|m,j) = \frac{\binom{m-k-1}{j-2}}{\binom{m-1}{j-1}}. \tag{9}$$

The probability of i mutants at 0, given k and m at t' is obtained by noting that the problem is equivalent to finding the distribution of i, the numbers of balls in k boxes, when n balls are distributed randomly into m nonempty boxes. There are $\binom{n-i-1}{m-k-1}\binom{i-1}{k-1}$ ways in which i balls can be arrayed in k boxes (the mutants) and $n - i$ balls are arrayed in $m - k$ boxes (the nonmutants) and hence

$$\Pr(i|k,m) = \frac{\binom{n-i-1}{m-k-1}\binom{i-1}{k-1}}{\binom{n-1}{m-1}}. \tag{10}$$

Finally, the probability of i mutants at 0, given j, is by analogy with (9)

$$\Pr(i|j) = \frac{\binom{n-i-1}{j-2}}{\binom{n-1}{j-1}}. \tag{11}$$

Thus, the joint probability of i and k, given m and j is the product $\Pr(i|k, m)\ \Pr(k|m, j)$, and the unconditional joint probability is found by averaging over m and j

$$\Pr(i,k) = \sum_{m=1}^{n}\sum_{j=1}^{m} g_{nm}(t')g_{mj}(t - t')\Pr(k|m,j)\Pr(i|k,m). \tag{12}$$

The unconditional probability of i is

$$\Pr(i) = \sum_{j=1}^{n} g_{nj}(t)\Pr(i|j). \quad (13)$$

Note that in (12) and (13), we are allowing for the possibility that $i = n$ but not that $i = 0$, because the mutation that occurred at t is assumed to be represented in the sample.

The distribution of k at t', given i, is obtained by dividing (12) by (13) according to the usual rules of conditional probabilities. This quantity corresponds to the g's so we will write

$$f_{ik}(t') = \frac{\Pr(i,k)}{\Pr(i)} \quad (14)$$

to emphasize the similarity. Note that $f_{ik}(0) = \delta_{ik}$ and $f_{ik}(t) = \delta_{Ik}$, where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise.

The expected total tree length of the mutant class, given that mutant arose at t and that i copies are found in the sample is, by analogy with (6),

$$\mathrm{E}[L_i^{(m)}(t)] = \sum_{k=2}^{i} k \int_0^t f_{ik}(t')\,dt' \quad (15)$$

where the superscript (m) indicates the mutant class. Once again, this result cannot be expressed in closed form but it is relatively easy, if slow, to evaluate using a computer algebra program. Note that (15) can be used even if $N(t)$ is not constant. Variation in population size will be absorbed into the g's on which the f's depend. We can also compute the expected value of the pairwise coalescence time

$$E[P_i^{(m)}(t)] = \sum_{k=2}^{i} \prod_{l=k+1}^{i} (1 - \pi_l) \int_0^t f_{ik}(t')\,dt'. \quad (16)$$

To evaluate the right hand sides of (15) and (16), it is more efficient computationally to express f as a sum of the gs and then perform the integrations with respect to t' analytically.

The expected age of the root of the intraallelic gene genealogy, which we will denote by $G_i^{(m)}(t)$, is simply

$$E[G_i^{(m)}(t)] = \sum_{k=2}^{i} \int_0^t f_{ik}(t')\,dt'. \quad (17)$$

Although the mutant class is assumed to arise at t, there is likely to be a delay before the first branching event that determines the age of the root.

Because we know that the mutant arises as a single copy and then increases to i copies in the sample, the subpopulation of mutants can be thought of as an expanding population. Consequently we can anticipate that the gene genealogy of mutants might approach a star genealogy even if the overall population is of constant size. We can use the index defined by Equation 8

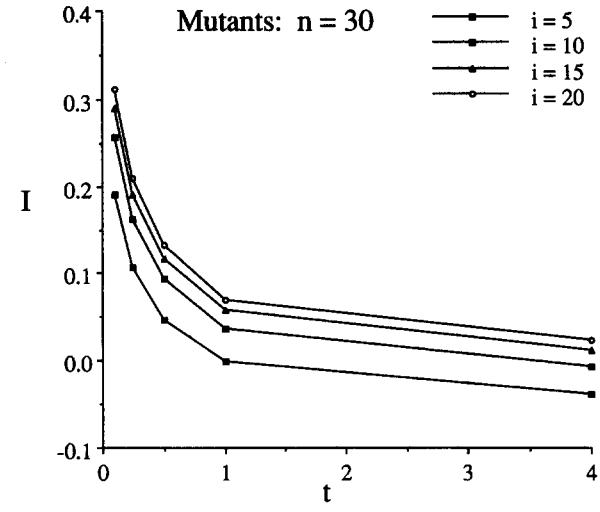

FIGURE 2.—Stellate index for intraallelic gene genealogies. Values of I defined by Equation 6, for the gene genealogy of the mutant class given that i copies of the mutant were found in a sample of $n = 30$ copies of the locus of interest and that the mutant class arose by a unique mutation at time t in the past. The population is assumed to be of constant size, N, and t is measured in units of $2N$ generations. The value of N does not affect the results provided that N is large enough that the diffusion approximation, defined in the text, can be assumed. These results were obtained from a C program that evaluates the equivalent of Equations 16 and 17 in the text.

to describe the degree to which that is so. Figure 2 shows the value of this index for different value of i and t in a population of constant size. This index is positive for small t but becomes slightly negative for small i and relatively large t. The reason for these negative values seems to be that, in those cases, most of the coalescent events occur quickly but the genealogy still has to have a long internal branch because there must be at least two lineages present at t. The resulting genealogy resembles two broccoli spears and, in extreme cases of that type, I can be negative. It is, however, very unlikely that the mutant would not be lost or fixed in $8N$ generations so negative values would be very unlikely to be found.

For a given value of i, I decreases as n increases, and the value of I depends primarily on i/n, the observed allele frequency (results not shown). The pairwise coalescence times and the total tree lengths both depend on n for a given value of i, as illustrated in Figure 3 for the pairwise values. Note that the expected pairwise coalescence times are always much less than 1, the value in scaled units for the sample from the whole population, even when the mutant arose before $t = 1$.

The nonmutant allelic class: The genealogical approach provides us with additional information about the history of the sample. We can view the gene genealogy as having two parts. One monophyletic clade contains i copies that are in the mutant class. The remaining $n - i$ are in another group that may or may not be monophyletic but must be at least paraphyletic. Using

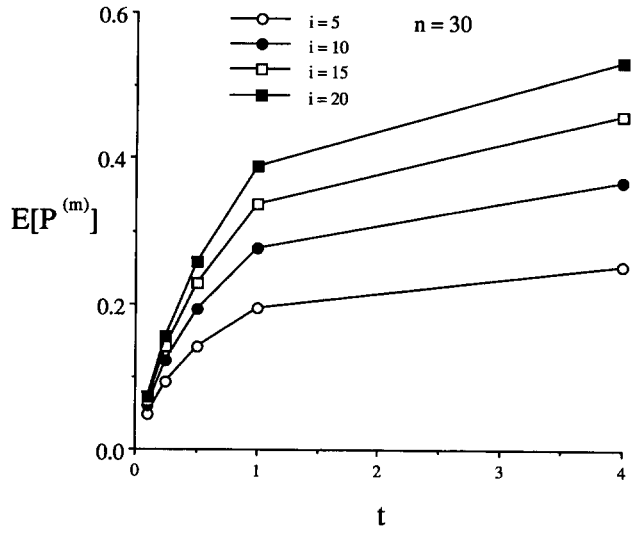

FIGURE 3.—Average pairwise coalescent times within an allelic class. Values of $P^{(m)}$, the expected pairwise coalescence time of two different copies of a mutant allele that arose at time t in the past, given that i copies of the mutant were found in a sample of size $n = 30$. The population was assumed to be of constant size N and both t and $P^{(m)}$ are measured in units of $2N$ generations. These results were obtained from a C program that evaluates the equivalent of Equation 16 in the text.

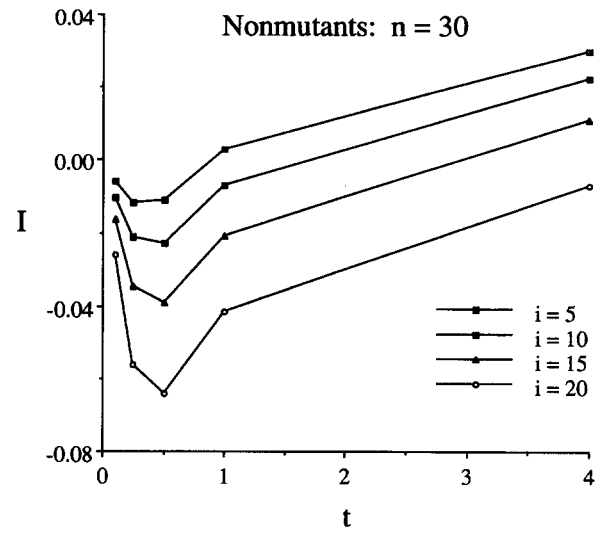

FIGURE 4.—Stellate index within the nonmutant allelic class. Values of I, defined by Equation 6, for the nonmutant class when the mutant arose at time t in the past. The sample size was $n = 30$ and i is the number of mutants found in the sample. These results were obtained by evaluating Equations 19 and 20 in the text.

the methods in the previous sections, we can find the total tree length and the average pairwise coalescence time in the nonmutant class. The only difference between this analysis and that for the mutant class is that the nonmutant class traces its ancestry to $j - 1$ lineages at t instead of 1, so, when $j > 2$, we have to add a component to account for the genealogy before t.

Let l be the number of nonmutants at t' $(k + l = m)$, and let h be the number of nonmutants at 0 $(h + i = n)$. Clearly, $\Pr(n - l|j, m) = \Pr(k|j, m)$ and $\Pr(n - h|l, m) = \Pr(i|k, m)$. The total tree length between 0 and t is found from an integral analogous to (15).

Immediately before t, there are j lineages but only $j - 1$ of them represent ancestors of the nonmutants in the sample. To find the contribution to the total tree length before t, we cannot use the unconditional probability of j at t, $g_{nj}(t)$, because we want the expectation of the total tree length in the nonmutants, given the value of i. Proceeding with the same type of Bayesian argument, Equation 11 gives us the probability of i, given j and n, and $g_{nj}(t)$ provides the unconditional probability of j, so the joint probability of i and j is the product of these two functions. We also know the unconditional probability of i, Equation 13. Therefore, the probability of j, given i, is

$$\Pr(j|i) = \frac{\Pr(i|j)g_{nj}(t)}{\sum_{j=1}^{n} \Pr(i|j)g_{nj}(t)} \tag{18}$$

and thus

$$E[L_i^{(nm)}(t)] = \sum_{l=2}^{m} l \int_0^t f_{i,m-l}(t')\,dt' + \sum_{j=3}^{n} \Pr(j|i)\Sigma_{j-1} \tag{19}$$

where the superscript (nm) indicates nonmutants. In (19) we can use the same function as in (15) because $l = m - k$. Similarly, the expectation of the average pairwise coalescence time.

$$E[P_i^{(nm)}(t)] = \sum_{l=2}^{m} \prod_{s=2}^{l} (1 - \pi_s) \int_0^t f_{i,m-l}(t')\,dt' + \sum_{j=3}^{n-1} \prod_{s=3}^{n-1} (1 - \pi_s)E(P_{j-1})\Pr(j|i) \tag{20}$$

where P_{j-1} is defined by (7). In the second term on the right hand side of (20), the product is the probability that a randomly chosen pair of copies has not coalescenced when j ancestral lineages remain. In general, $E(P_{j-1}) = E(L_2)/2$, independently of j and if $N(t)$ is constant $E(P_{j-1}) = 2N$. Equation 20 is written in the form it has to emphasize the similarity with (19).

The genealogy of the nonmutants is somewhat different from that for the mutants. The stellate index, I, is negative for small values of t, as shown in Figure 4. That occurs because the effective population size of the nonmutants is increasing in the past as the mutant allele frequency decreases. The ratio of the expected pairwise coalescence time within the mutant class, $E[P_i^{(m)}]$, to the value within the nonmutant class, $E[P_i^{(nm)}]$, is generally small and increases with the values of both i and t (Figure 5).

Variances in tree length and pairwise differences: This same approach can be used to compute the variances of the total tree length, pairwise differences, and age of the root. The calculations are somewhat messy but the idea is simple. We need to consider the ancestry of the allelic class at two intermediate times in the past, t' and t'', in order to find the joint probabilities that

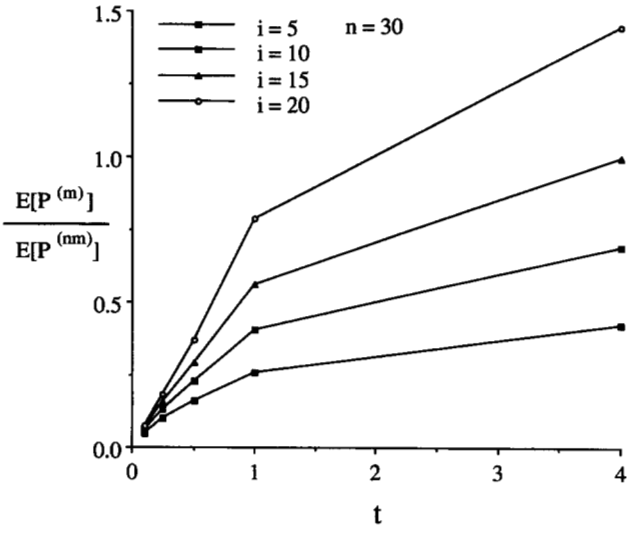

FIGURE 5.—Ratio of average pairwise coalescent times. The ratio of the average pairwise coalescence times within the mutant class and within the nonmutant class $P^{(m)}/P^{(nm)}$, given that the mutant arose at time t in the past and that i copies of the mutant were found in the sample of size $n = 30$. These results were obtained by evaluating Equations 16 and 20 in the text.

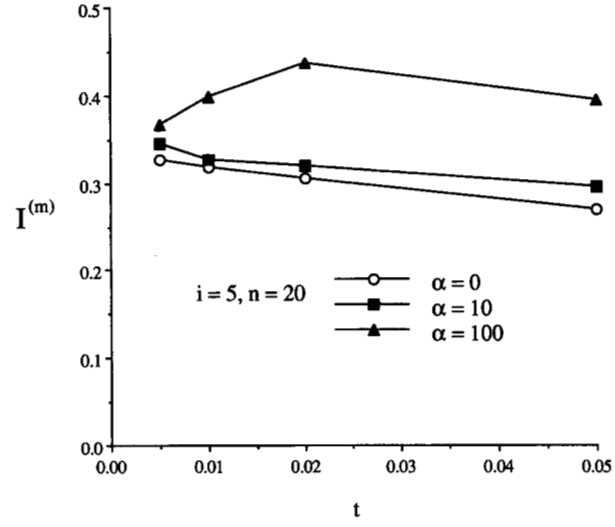

FIGURE 6.—Stellate index within an allelic class in an exponentially growing population. Values of I, defined by Equation 8, for a mutant found in $i = 5$ copies in a sample of size $n = 20$ for a population of constant size and a population that has undergone exponential growth with $\alpha = 2N_0r = 10$. and $\alpha = 100$, where N_0 is the current population size and r is the rate of exponential population growth in the past. These results were obtained using a C program to evaluate Equations 15 and 16.

there are k ancestral lineages at t' and l ancestral lineages at t'', $f_{kl}(t', t'')$. The second moment of the total tree length is then found by integrating with respect to t' and t'':

$$E[(L_i^{(m)}(t))^2] = \sum_{k=2}^{n} \sum_{l=2}^{k} \iint klf_{kl}(t',t'')\,dt'\,dt'' \quad (21)$$

with similar expressions for the second moments of pairwise differences and the age of the roots. The second moments for the nonmutant class are more difficult to find because the period before the mutant appeared also must be accounted for.

From (21), the variance is found by subtracting the square of the expectation. The quantitative results show that the gene genealogy of the mutation is somewhat less variable than a neutral genealogy. For example, with $i = 5$, $n = 20$, and $t = 0.1$, the coefficient of variation of the total tree length is 0.32, while for a neutral genealogy with five tips, the coefficient of variation is 0.57.

Exponential population growth: We have seen that in a population of constant size, the gene genealogy of a recently arisen mutant is starlike. We also know that a rapidly growing population tends to result in a starlike gene genealogy for the entire sample. We can use the methods developed in the previous sections to find how much more starlike the genealogy of a newly arisen mutant is in a rapidly growing than in a population of constant size. Figure 6 shows some results for the case of exponential population growth, as in Figure 1a. Time is measured in units of $2N_0$ generations where N_0 is the current population size. We can see that the exponential population growth increases the value of I from that in a population of constant size ($\alpha = 0$ in the figure).

RECOMBINATION

The preceding sections have assumed that there is no recombination. That is appropriate for some parts of the genome such as mtDNA in animals and the Y chromosome in mammals. If there is recombination and if the allelic class is defined by a single nucleotide substitution, these same methods can used to estimate the size of a nonrecombined region surrounding that nucleotide position.

We consider two sites with a recombination fraction c between them. At one site, there is a mutation that defines the new allelic class. The problem is to find the probability that, among the ancestral lineages in the allelic class, there has been no recombination event with the other site. If there has been no recombination, then the nucleotide at the linked site will be perfectly associated with mutant allele. When several polymorphic sites closely linked to the defining mutation, the haplotype associated with it may be unique. That is the basis for "disequilibrium mapping" (LANDER and BOTSTEIN 1989; HÄSTBACKA *et al.* 1992).

In a generation during which there are k lineages, the probability that none of them undergoes a recombination event is $(1 - c)^k$. Multiplying this quantity by itself T_k time, where T_k is the number of generations during which there are k ancestral lineages, and then adding over k, we find the total probability that there is no recombination to be $\phi = (1 - c)^{2NL} \approx e^{-RL}$, where

$L = \Sigma_{k=2}^{i} kT_k$ is total tree length of the mutant class measured in units of $2N$ generations and $R = 2Nc$. To find the expectation of ϕ in general, we need to find the generating function of L, which cannot be obtained using the above methods. For small values of R, $\phi \approx 1 - RE(L)$, where $E(L)$ is the expectation of the tree length of the mutant class, which is give by Equation 15. Numerical analysis shows that unless the mutant arose very recently (t small), $E(L)$ is of order 1 in magnitude and hence the probability of a recombination event between the mutant and a linked marker is of the same order of magnitude as the product $R = 2Nc$.

MUTATION MODELS

Given a gene genealogy, the mutation process determines the extent of genetic variation. I will consider two mutation models, the infinite sites model (WATTERSON 1975) and the generalized stepwise mutation model (OHTA and KIMURA 1973). The infinite sites model is appropriate for nucleotide sequences if it can be assumed that there is no recombination and that no site mutates more than once. Because the concern here is with within-species variation, the second assumption is a reasonable if imperfect starting point. The stepwise mutation model is appropriate for simple sequence repeat loci, also called microsatellites, which are currently being widely used as genetic markers both for gene mapping and for the analysis of population structure (FREIMER and SLATKIN 1996).

Infinite sites model: In this model, there is a single locus with a sufficiently large number of nucleotide sites that a mutation never affects the same site twice. The mutation rate per locus per generation is μ. TAVARÉ (1984) and HUDSON (1990) show that in a sample, the expected pairwise difference between two copies is 2μ multiplied by the average pairwise coalescence time in the sample and that the expected number of segregating sites in the sample is μ multiplied by the total tree length. As a consequence, the results from the previous sections can be used to predict the average pairwise difference and the numbers of segregating sites within the allelic class. Furthermore, those two quantities could be used to estimate the stellate index, I, defined Equation 8. Because the stellate index is generally positive for the mutant class and negative for the nonmutant class, there is some hope of distinguishing the two in cases where it is not clear from other considerations which descended from a recent mutation.

Stepwise mutation model: Because the application of the model is to microsatellites, it is easiest to describe the model in that context. A microsatellite locus contains several repeated units of two to five nucleotides. The size of an allele is the number of repeat units. In humans and other mammals, these loci are extremely abundant and polymorphic, and hence are useful for a variety of purposes. Microsatellites have been shown to fit the predictions of a stepwise mutation model in which allele size changes under mutation by one or a few repeat units (VALDES *et al.* 1993; SHRIVER *et al.* 1993; DI RIENZO *et al.* 1994), and some of the basic theory for that model has been developed (FREIMER and SLATKIN 1996). Here I will be concerned with a variation on the theme of the basic model and imagine that there are two kinds of mutations, mutations that change only repeat number and that occur relatively frequently and mutations that change the structure of the microsatellite locus and that occur relatively rarely. Sequences of microsatellite loci show that mutations occur that do more than change the number of repeats (GARZA *et al.* 1995). For example, there may be a change from a perfect allele, in which the variable portion contains only the repeated motif, for example $(CA)_n$, to an imperfect allele, for example $(CA)_nCCC(CA)_m$. The mutation rate may depend on whether an allele is perfect or imperfect (FREIMER and SLATKIN 1996).

Previous studies have shown that variation in repeat number at a microsatellite locus depends on the average pairwise coalescence time (FREIMER and SLATKIN 1996). For the stepwise mutation model, no simple relationship between total tree length and variation in repeat number has been found, so we will focus on only the pairwise differences. There are at least two possible applications of the theory of pairwise coalesce times developed in the preceding sections. If the mutation rate at a locus is known, as it is for some loci (WEBER and WONG 1993), and if one were willing to ignore mutations that change repeat number by more than one repeat unit, then the expected variance among the mutants and among the nonmutants are equal to the mutation rate multiplied by the appropriate expected coalescence times: $\mu E[P_i^{(m)}]$ and $\mu E[P_i^{(nm)}]$. On the other hand, if the mutation rate or mutation spectrum is not known, we can still predict the ratio of the variance in repeat number within the mutant class to the variance within the nonmutant class. Provided that the mutation rates are the same in the two classes, the mutation rate cancels from the ratio so the ratio of the variances is just the ratio of the pairwise coalescence times, as in Figure 5.

CONCLUSIONS

The goal of this paper was to show that a coalescent theory could be developed for allelic classes defined by unique mutations. The approach, which can be generalized, is to use a Bayesian argument to find the distribution of the numbers of ancestral lineages given the number observed in a sample and given that the class arose as a result of a mutation at a known time in the past. I have not attempted to address sampling or estimation problems here, although it seems that the general approach introduced by GRIFFITHS and TAVARÉ (1994) could be used for parameter estimation in this

context as well. There is as yet not much data to which the theory developed here can be applied, but the theory does show the potential gain in information about mutants from a detailed examination of variability within allelic classes.

I thank W. J. EWENS for helpful discussions, and R. R. HUDSON and S. TAVARÉ for several helpful comments on an earlier version of this paper. In particular, TAVARÉ pointed out the simple derivation of Equations 9–11. This research was supported in part by a grant from the National Institutes of Health (GM-40282).

LITERATURE CITED

CLARK A. G., 1993 Evolutionary inferences from molecular characterization of self-incompatibility alleles, pp. 79–108 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Sunderland, MA.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.* 1994 Mutational processes of simple sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

ETHIER, S. N., and T. SHIGA, 1994 Neutral allelic genealogy, pp. 87–97 in *Measure-Valued Processes, Stochastic Partial Differential Equations, and Interacting Systems*, edited by D. A. DAWSON. CRM Proceedings and Lecture Notes, Vol. 5, American Mathematical Society, Providence, RI.

FELLER, W., 1957 *An Introduction to Probability Theory and Its Applications*, Ed. 2, Wiley, New York.

FREIMER, N. B., and M. SLATKIN, 1996 Microsatellites: evolution and mutational processes, pp. 51–72 in *Variation in the Human Genome*, CIBA Foundation Symposium 197, Wiley & Co., Chichester.

GARZA, J. C., M. SLATKIN and N. G. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

GRIFFITHS, R. C. and S. TAVARÉ, 1994 Sampling theory for neutral alleles in a varying environment. Phil. Trans. Roy. Soc. Lond. B **344:** 403–410.

HÄSTBACKA, J., A. DE LA CHAPELLE A., I. KAITILA, P. SISTONEN, A. WEAVER *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genet. **2:** 204–211.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology **7:** 1–44.

HUDSON R. R., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

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

HUGHES A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167–70.

KAPLAN, N. L., T. DARDEN, and R. R. HUDSON, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.

KINGMAN, J. F. C., 1982 The coalescent. Stochast. Proc. Appl. **13:** 235–248.

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

LANDER E. S., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

SAUNDERS, I. W., S. TAVARÉ, and G. A. WATTERSON, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16:** 471–491.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY, and E. BOERWINKLE, 1993 VNTR allele frequency distributions under the stepwise mutation model. Genetics **134:** 983–993.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Popul. Biol. **26:** 119–164.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–276.

WEBER, J. L., and C. WONG, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

WEKEMANS, X., and M. SLATKIN, 1994 Gene and allelic genealogies at a gametophytic self-incompatability locus. Genetics **137:** 1157–1165.

WOLFRAM, S., 1991 *Mathematica: A System for Doing Mathematics by Computer.* Addison Wesley, Redwood City, CA.

WRIGHT, S., 1938 Size of population and breeding structure in relation to evolution. Science **87:** 430–431.

Communicating editor: G. B. GOLDING