Gene genealogies and the coalescent process

RICHARD R. HUDSON

1. INTRODUCTION

When a collection of homologous DNA sequences are compared, the pattern of similarities between the different sequences typically contains information about the evolutionary history of those sequences. Under a wide variety of circumstances, sequence data provide information about which sequences are most closely related to each other, and about how far back in time the most recent common ancestors of different sequences occurred. If the sequences were obtained from distinct species, then the information is frequently extracted and displayed in the form of an inferred phylogenetic tree, which may represent the evolutionary relationships of the species from which the sequences were sampled. If, instead of being from different species, the sequences are from different individuals of the same population, the information is genealogical, and in this case gene trees can sometimes be inferred. A gene tree shows which sampled sequences are most closely related to each other and perhaps the times when the most recent common ancestors of different sequences occurred. A hypothetical gene tree, or genealogy, of five sampled sequences is shown in Fig. 1. In the absence of recombination, each sequence has a single ancestor in the previous generation. (It is important to distinguish a gene tree of sampled sequences from the pedigree of a sample of diploid individuals, in which the number of ancestors grows as one proceeds back in time, because each diploid individual has two parents.) The possibility of obtaining detailed information about the genealogy of sampled genes dramatically changes the situation for molecular population geneticists.

Before the DNA era, molecular polymorphism data were primarily in the form of frequencies of electromorphs, alleles distinguished by their mobility on electrophoretic gels. With protein electrophoresis, two homologous copies of a gene could be classified as being the same or different. If they were different, one could not measure how different; if the two copies were the same, one could not with confidence distinguish whether
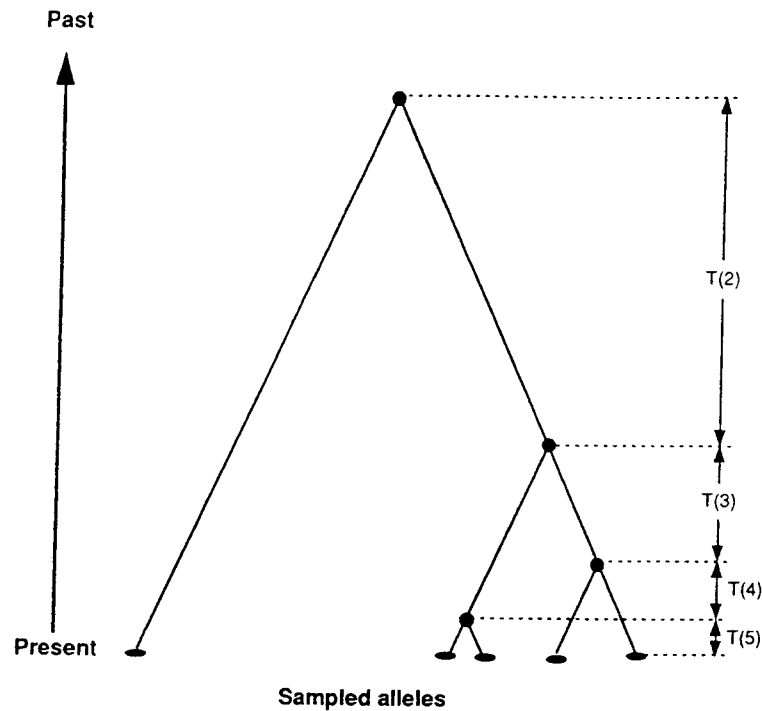
Past

Fig. 1. An example of a genealogy of a sample of five alleles, showing the time intervals between coalescent events. In this figure, the intervals, $T(i)$, are shown with lengths proportional to their expected values as given by eqn (5).

they were really the same or simply convergent in certain physical properties leading to similar electrophoretic mobility. Thus detailed information about the genealogies of genes could not be extracted from data on electromorph frequencies. With modern DNA techniques, sequences of homologous regions of many individuals are obtainable and detailed information about the genealogy of sampled genes will be obtained. Examples of genealogies inferred from sampled alleles are given in Stephens and Nei (1985), Aquadro et al. (1986), Bermingham and Avise (1986), Avise et al. (1987) and Cann et al. (1987).

The obvious challenge for molecular population geneticists is: How can we utilize this information to increase our understanding of the forces acting on molecular variation in natural populations? From the theory side, we can begin by examining the properties of genealogies that arise under a variety of population genetic models. It is important to ask: Are genealogies expected to be very different under different competing models? Can we devise statistical tests that take advantage of the different genealogies expected? To proceed with this task, one needs to examine

the statistical properties of genealogies of sampled genes under different models.

In the following, I will describe a variety of circumstances in which properties of genealogies can be derived analytically or by computer simulation. This will not constitute a comprehensive review of gene genealogy theory, but rather a very personal view that concentrates on the infinite-site model. Some properties of genealogies will be described under selectively neutral models, with and without recombination, and with and without geographic structure. The effects of some forms of selection will also be described. I will indicate some applications of this genealogical approach for carrying out statistical tests or estimating parameters or simply allowing an 'eye-ball' test of the fit of observations to data. I will also indicate how simulations based on the coalescent process can be constructed and used to investigate a variety of models.

This will not be a rigorous mathematical treatment. Those interested in a more precise analysis should consult the seminal work of Kingman (1980, 1982a,b) and the review by Tavaré (1984). Much of the very elegant and useful work of Griffiths (1980), Watterson (1984) and Padmadisastra (1987, 1988) on coalescents and lines of descent that focus on the infinite allele model will not be covered. This includes a large body of work on the ages of alleles (Donnelly 1986; Donelly and Tavaré 1986; Tavaré *et al.* (1989) that is reviewed by Ewens (1989). The infinite-allele models and the infinite-site models are very closely related, as will be described later, and results from one can often be used immediately to answer questions about the other. However, the questions asked and the parameter values considered are often quite distinct for the two models. In this chapter, I will concentrate on results that directly concern infinite-site models, which I feel are most useful in the interpretation of nucleotide variation in populations.

I will focus on properties of relatively small samples of alleles. The work on properties of genealogies of entire populations, including fixation times, will not be considered (Donnelly and Tavaré 1987; Watterson 1982a, 1982b). Also, the important work on the relationship between gene trees and species trees will not be discussed (Hudson 1983b; Neigel and Avise 1986; Pamilo and Nei 1988; Takahata 1989).

Statistical properties of genealogies depend very strongly on the kind of sampling that occurs to produce one generation from the last. In this chapter, only the Wright-Fisher (W-F) model will be considered. The sampling that produces one generation from the last under this model is described briefly in the next section. A range of alternative neutral models have been found that have essentially the same genealogical properties as the W-F model, with only a change of time-scale (Kingman 1982a,b; Watterson 1975; see also the reviews by Tavaré, 1984, and Ewens, 1989).

2. SEPARATING THE GENEALOGICAL PROCESS FROM THE NEUTRAL MUTATION PROCESS

As will be discussed in great detail in the following pages, the statistical properties of genealogies depend on such factors as population size, geographic structure and the presence of selectively maintained alleles. That properties of genealogies should depend on these demographic properties is obvious, because actual genealogies depend on who had offspring and who did not, who migrated and to where, and whose offspring bore selectively important mutations. It should also be clear that strictly neutral mutations – mutations that have not and will not affect fitness – should have no affect on the genealogies of random samples. This is because, by definition, neutral mutations do not affect the number of offspring or tendency to migrate of individuals bearing those mutations. That being the case, we can study the properties of genealogies without regard to a specific mutation model for *neutral* variants. So, for example, the statistical properties of genealogies do not depend on whether neutral mutations are more frequently transitions than tranversions or whether an infinite-site, finite-site or infinite-allele model is most appropriate. Of course, the statistical properties of our inferences about the genealogical process are likely to depend strongly on the mutation process. For example, if the neutral mutation rate is very low, all the sequences in a sample may be identical and we could get no information about the genealogy of the sample.

With the neutral mutation process that we will consider, each offspring differs from its parent at the locus under consideration by a Poisson distributed number of mutations. The mean number of mutations, μ, will be assumed constant, independent of genotype, population size and time. The mutations are assumed to occur independently in different individuals and different generations. This mutation model will be referred to as the constant-rate neutral mutation process. This is the standard neutral mutation model (Kimura 1983; Watterson 1975). Under these assumptions, mutations accumulate *along lineages* in an inexorable fashion independent of, for example, population size or selection events at linked loci. Given t, the number of generations since the most recent common ancestor of two sampled homologous sequences, S, the number of mutations that have occurred in the descent to the two descendent sequences, is Poisson distributed with mean $2\mu t$. When t is a random quantity, the mean and the variance – in fact all the moments of S – are determined by the moments of t assuming the constant-rate neutral mutation process.

To emphasize this point, consider a population that at time 0 is completely homozygous at a locus at which only neutral mutations occur. After t generations of evolution, one examines the sequence at the locus in a single randomly selected individual. Under the mutation scheme we

have described in the previous paragraph, the number of mutations that will have occurred to distinguish our randomly sampled individual from the individuals in the population at time 0, is just the number of mutations that have occurred along a particular lineage of length t. This number of mutations is Poisson distributed with mean μt. It does not matter what the population size has been, whether selection has been occurring at linked loci, or whether there is population subdivision. This is the basis for the results of Birky and Walsh (1988) concerning the rate of accumulation of neutral mutations when selection is occurring at linked loci. In the example above, the number of mutations that have fixed in the entire population between time 0 and time t will depend on these demographic aspects of the population. Similarly, the amount of polymorphism in the population at time t will depend on population size and other demographic factors, but the number of mutations that will have occurred along individual lineages in the past t generations, that distinguish a sampled sequence from their ancestors t generations back, is Poisson distributed with mean μt, regardless of these other factors.

This property of the constant-rate neutral mutation process will be exploited in the following way. Let T_{tot} denote the sum of the lengths of the branches of the genealogy of a sample. As discussed in the previous paragraph, S, the number of mutations on the genealogy, given T_{tot}, is Poisson distributed with mean μT_{tot}. Once the distribution of T_{tot} is determined under a particular model, the distribution of S can easily be obtained. For example, if the first two moments of T_{tot} are determined, then the first two moments of S can be calculated using properties of compound distributions as:

$$E(S) = \mu E(T_{tot}) \tag{1}$$

and

$$\mathrm{Var}(S) = \mu E(T_{tot}) + \mu^2\, \mathrm{Var}(T_{tot}) \tag{2}$$

Reiterating, under the models that we will consider, the properties of genealogies do not depend on the neutral mutation process, and therefore can be studied without precise specification of the neutral mutation process. For example, we can study the statistical properties of T_{tot} without specifying the rate or pattern of neutral mutation. Furthermore, statistical properties of neutral variation in samples are completely determined by the statistical properties of the genealogies and the neutral mutation process. In other words, if two different models make the same assumptions about the neutral mutation process and if the two different models lead to the same distribution of genealogies, then the pattern of neutral variation will be the same for the two models. For example, if the neutral

mutation process is as we have described above, the mean value of S is completely determined by the mean value of T_{tot}. Two different models that lead to the same mean value of T_{tot} will have the same mean value of S.

Throughout this chapter, we will consider an ideal W-F model, with either N haploids or N diploids. Briefly, this is a discrete generation model in which, for the haploid version, the N haploids of an offspring generation are obtained by sampling (and replicating possibly with mutation) N times with replacement from the parent generation. In the selectively neutral version, all parents are equally likely as parents of each of the N haploid offspring. A detailed description of this model is contained in Ewens (1979). We will assume that N is large and constant, in which case individuals have approximately Poisson distributed numbers of offspring. Most of the results concerning this model will be approximate, ignoring terms of order $(1/N^2)$ relative to $(1/N)$. This corresponds to the usual assumptions made for using diffusion approximations and will be referred to as the diffusion approximation. In contrast to the W-F model, exact results can often be obtained for the Moran model (see, for example, Watterson 1975). The Moran model will not be considered here.

3. THE SIMPLEST CASE: NO SELECTION AND NO RECOMBINATION

Although genealogical processes are implicit in much of the work on identity coefficients that has been carried on for many years, it was the knowledge of the nature of the genetic material and the possibility of obtaining sequence data (or restriction map data) that stimulated some of the earliest work that considers the genealogical process directly. Watterson's (1975) remarkable paper describes the basic properties of genealogies under neutral models and marks the beginning of modern coalescent theory. The following description of the no-recombination genealogy under the W-F neutral model draws heavily from the work of Watterson (1975), Kingman (1980, 1982a,b) Griffiths (1980) and Tajima (1983).

To begin, we consider an ideal haploid species without recombination, without geographic subdivision and without selection – a typical garden-variety haploid species. We wish to examine properties of the genealogy of a random sample of n individuals from this population. Let us label the population from which the sample was drawn, generation 0. The ancestral population t generations back in time will be referred to as generation t.

The basic property of a sample drawn from such a population, upon which much of the following is based, concerns the probability, $P(n)$, that all the n sampled individuals have separate distinct ancestors in the

preceding generation. Consider first a sample of two individuals. The probability that the second individual sampled has the same parent as the first is $1/N$, as under the W-F neutral model each individual of the previous generation is equally likely to be the parent of any individual of the current generation. Thus $P(2)$ is $1-1/N$. If three individuals are sampled, the probability that all three have distinct ancestors in the previous generation, is the probability that the first two have distinct parents × the probability that the parent of the third individual drawn is distinct from the first two parents. As there are $N-2$ individuals that are distinct from the parents of the first two sampled individuals, the probability that the third individual has a distinct parent from the first two, given that the first two have distinct parents, is $(N-2)/N = 1-2/N$. In general, the probability that n sampled individuals have n distinct parents in the previous generation is:

$$P(n) = \prod_{i=1}^{n-1} (1-i/N) \approx 1 - \frac{\binom{n}{2}}{N} \tag{3}$$

We can ask the same question about these n distinct ancestors: What is the probability that they have n distinct ancestors one generation earlier? Clearly, this is also $P(n)$. This means that the probability that the n sampled individuals have n distinct ancestors in each of the preceding t generations, and that in the $t + 1$ generation back in time, two or more of the sampled individuals have common ancestors is:

$$P(n)^t [1-P(n)] \approx \frac{\binom{n}{2}}{N} e^{-\frac{\binom{n}{2}}{N} t} \tag{4}$$

In words, the time back until the first occurrence of a common ancestor is geometrically distributed and will be approximated by an exponential distribution with mean $N/\binom{n}{2}$. For large N and small n, as we will assume throughout, the probability that more than two individuals of our sample have common ancestors in a single generation is very small and will be ignored. Thus with high probability, the recent history of our sample consists of t generations in which n distinct lineages exist, and then at generation $t + 1$, a single pair of lineages 'coalesce' at the most recent common ancestor of two of the sampled individuals. Each of the $\binom{n}{2}$ possible pairs of lineages are equally likely to form the coalescing pair. To continue tracing the history of our sample back in time, we note that

in the generations preceding the first coalescence, there are $n - 1$ ancestors or lineages to follow. The probability – each generation – that all of these ancestors have distinct ancestors in the preceding generation is $P(n-1)$. So the time to the next coalescence is approximately exponentially distributed with mean $N/\binom{n-1}{2}$. At this coalescence, each of the $\binom{n-1}{2}$ possible pairs of lineages are equally likely to coalesce at this node.

Note that one of these $(n-1)$ lineages has two descendants in our original sample, the other lineages having a single descendant in the sample. We can continue in this way until all the lineages have coalesced into a single lineage, the common ancestor of the entire sample of n individuals.

A genealogy of five sampled alleles is shown in Fig 1. The stochastic process that generates a genealogy, referred to as the coalescent process, can be summarized very briefly. The time, $T(j)$, during which there are j distinct lineages is approximately exponentially distributed, and if time is measured in units of N generations, the mean of $T(j)$ is:

$$E[T(j)] = 1/\binom{j}{2} \tag{5}$$

The two lineages that coalesce at a node in the genealogy, say in generation $t + 1$, are two lineages randomly chosen from the lineages present in generation t. Notice that we have not had to concern ourselves with lineages other than those that are ancestral to our sample. Also note that the intervals between coalescences, the $T(j)$'s, are statistically independent of each other. Also, it is important to note that the older parts of the genealogy (the upper parts of the genealogy in Fig. 1), are identical in statistical properties to the genealogies of smaller samples. For example, the part of the genealogy above the most recent coalescent event in the history of a sample of size n, is distributed exactly as the genealogy of a sample of size $n - 1$. Generating such genealogies on a computer is trivial (an example of a program is given in the Appendix).

These properties of genealogies apply to mitochondrial genomes as well as to garden-variety haploid organisms. If mitochondrial inheritance is strictly maternal and polymorphism within individual females is negligible, then N is the number of females.

For a large population of N diploids, under the W-F model with random mating, no recombination and no selection, the results are also the same, except that N is replaced by $2N$. The genealogy in this case should be thought of as the genealogy for a specific locus within which no recombination occurs. The locus might consist of a single nucleotide site or, if the recombination rate is sufficiently low, of many contiguous nucleotide

sites that can be considered completely linked. For the model being considered, sufficiently low means that $Nr \ll 1$, where r is the recombination rate per generation between the ends of the region being considered. If time is measured in units of N generations for haploid models, and in units of $2N$ generations for diploid models, the results are exactly the same for haploids and diploids, i.e. the mean of $T(j)$ is given by eqn (5).

Unlinked loci in large populations are essentially independent and will have their own independent genealogies. Linked loci, which have correlated genealogies, will be considered later.

4. ADDING NEUTRAL MUTATIONS TO THE GENEALOGY

Given the properties of the genealogies just described, we can predict properties of samples under various mutation schemes. As discussed in the previous section, we will assume a constant-rate neutral mutation process, in which each offspring gamete differs from its parent by an average of μ mutations. In addition, we will assume an infinite-site model (Kimura 1969). Under this model, the locus is composed of many sites, so that no more than one mutation occurs at any site in the genealogy of our sample. The oft-employed infinite-allele model (Kimura and Crow 1964) is similar, assuming that each mutation produces a new allele, not present anywhere else in the genealogy of the sample. For our purposes, the infinite-site model and the infinite-allele model are essentially the same but under the infinite-allele model one ignores how many mutations distinguish alleles and notes only whether alleles are the same or different.

The first properties to be considered concern the distribution of the number of mutations that occur on the branches of the genealogy of a sample. Under the infinite-site model, this number of mutations is identical to the number of nucleotide sites that would be polymorphic in the sample. The number of polymorphic sites in the sample, denoted S, is often referred to as the number of segregating sites in the sample. First, we consider the expected value of S.

From eqn (1) we can calculate the expectation of S from the expectation of T_{tot}, the total length of the genealogy. It follows easily from the definition of $T(j)$, that the sum of the lengths of the branches of the genealogy is $\sum_{i=2}^{n} iT(i)$. Therefore, from eqn (5), now measuring time in units of $2N$ generations, it follows that

$$E(S) = \frac{\theta}{2} \sum_{i=2}^{n} iE(T(i)) = \theta \sum_{i=1}^{n-1} 1/i \qquad (6)$$

where $\theta = 4N\mu$ (Watterson 1975). The variance of the total time is also

easily obtained, and using eqns (2) and (6), one obtains (Watterson 1975):

$$\text{Var}(S) = \theta \sum_{i=1}^{n-1} 1/i + \theta^2 \sum_{i=1}^{n-1} 1/i^2 \tag{7}$$

In fact, any moment of S can be expressed in terms of the moments of the T_i. Watterson also showed that the number of segregating sizes is approximately normally distributed in samples of sufficient size.

We can obtain the entire distribution of S, but first we consider the probability that $S = 0$, for a sample of size 2. This is equivalent to the expected homozygosity, $E(F)$, or the probability that two sampled alleles are identical. This probability will be derived in two ways. For two sampled alleles to be identical under the infinite-site model (or the infinite-allele model), it must be the case that no mutations have occurred on the lineages that descend to them from their most recent common ancestor (denoted MRCA). Given t, the number of generations back to their MRCA, the probability that no mutations have occurred in the descent to the sampled alleles is $e^{-2\mu t}$. This follows from our Poisson assumption about mutation. Therefore, if we take the expectation of $e^{-2\mu t}$, over the distribution of t, which is exponential with mean $2N$ in the diploid model, we find:

$$E(F) = E(e^{-2\mu t}) = \int_0^\infty \frac{e^{-t/2N}}{2N} e^{-2\mu t}\, dt = \frac{1}{1 + \theta} \tag{8}$$

This is a classic result (Kimura and Crow 1964) that can, of course, be derived from recursions, but here one gets a sense of its connection to the genealogy.

Equation (8) also illustrates a general connection between the infinite-allele model and the coalescent process. For any model of the population process, which determines the genealogical process, if the mutation process is the infinite-allele constant-rate neutral mutation process that we have been assuming, then the probability that two randomly sampled alleles are identical is $C(\theta) = E(e^{-\theta t})$, where this expectation is with respect to the distribution of t, the time back to the most recent common ancestor of two random alleles measured in units of $2N$ generations. The identity coefficient with $-\theta$ as argument, $C(-\theta)$, is also the moment-generating function of t. The moments of t, and consequently moments of S, are easily obtained from $C(\theta)$ by standard methods. For example, $E(t)$ is $-C'(0)$ and $E(S)$ is $-\theta C'(0)$, where $C'(0)$ represents the derivative of $C(\theta)$ with respect to θ evaluated at $\theta = 0$. This is quite general. For example, in models of gene conversion in multigene families, identity coefficients have been obtained for pairs of alleles sampled in various ways (Nagylaki and Petes 1982). The moments of the number of sites

that would distinguish these alleles under an infinite-site model, can be calculated as just described by taking derivatives of the identity coefficients.

An alternative derivation of eqn (8) involves tracing the history of the two sample alleles back in time, until either the MRCA of the alleles is found or a mutation on one of the lineages is found. In each generation, the probability, P_{CA}, that the MRCA occurs is $1/2N$. Also, in each generation, the probability, P_{mut}, that one or the other of the two lineages experiences a mutation is 2μ. The two alleles can be identical if, and only if, the first event encountered is a common ancestor event. Given that one or the other event has occurred, and ignoring the possibility that both occur in the same generation, the probability that the first event encountered is the common ancestor event is:

$$E(F) \approx \frac{P_{CA}}{P_{CA} + P_{mut}} = \frac{1/2N}{1/2N + 2\mu} = \frac{1}{1 + \theta} \tag{9}$$

In a similar fashion, one can derive the entire distribution of the number of mutations that have occurred since the MRCA of the sample of size 2. The probability, $P_2(j)$, of j mutations occurring on the lineages since the MRCA, is the probability that the first j events, as we trace backwards in time, are mutations and the $(j + 1)^{st}$ event is a common ancestor event. Thus, we have (Watterson, 1975):

$$P_2(j) = \left(\frac{\theta}{1+\theta}\right)^j \frac{1}{1+\theta} \tag{10}$$

Using a similar argument, we can obtain the probability, $Q_n(j)$, that j mutations occur in the time in which there are n ancestral lineages. To get j mutations during this time, the first j events, during the time there are n lineages, must be mutations, and the $(j + 1)^{st}$ event must be a common ancestor event. Hence, this probability is

$$Q_n(j) = \left(\frac{n\mu}{n\mu + \dfrac{\binom{n}{2}}{2N}}\right)^j \frac{\dfrac{\binom{n}{2}}{2N}}{n\mu + \dfrac{\binom{n}{2}}{2N}}$$

$$= \left(\frac{\theta}{\theta+n-1}\right)^j \frac{n-1}{\theta+n-1} \tag{11}$$

The number of segregating sites in a sample of size n is the sum of the

number that occur while there are n lineages, and the number during the rest of the genealogy distributed just like the number in a sample of size $n-1$. It follows that $P_n(j)$, the probability of j segregating sites in a sample of size n, can be written as:

$$P_n(j) = \sum_{i=0}^{j} P_{n-1}(j-i)Q_n(i)$$ (12)

The distribution of the number of segregating sites can quickly be calculated using this recursion. Tavaré (1984) obtained an explicit expression for $P_n(j)$. The distribution of S is shown in Fig. 2 for $\theta = 5$ and $n = 20$.

The use of eqn (12) is illustrated by the following example. Recent surveys of polymorphism in the yellow-achaete-scute region of *Drosophila melanogaster* revealed 9 polymorphic sites in 2112 nucleotide sites in 64 chromosomes examined (Aguadé *et al.* 1989). Estimates of θ per base pair from other regions of the *D. melanogaster* genome have averaged about 0.005. Aguadé *et al.* wanted to determine if the observation of 9 polymorphic sites was consistent with the hypothesis that θ per base pair in the yellow-achaete-scute region is 0.005. Using eqn (12), we can calculate that the probability of 9 or fewer polymorphisms, in a sample of 64
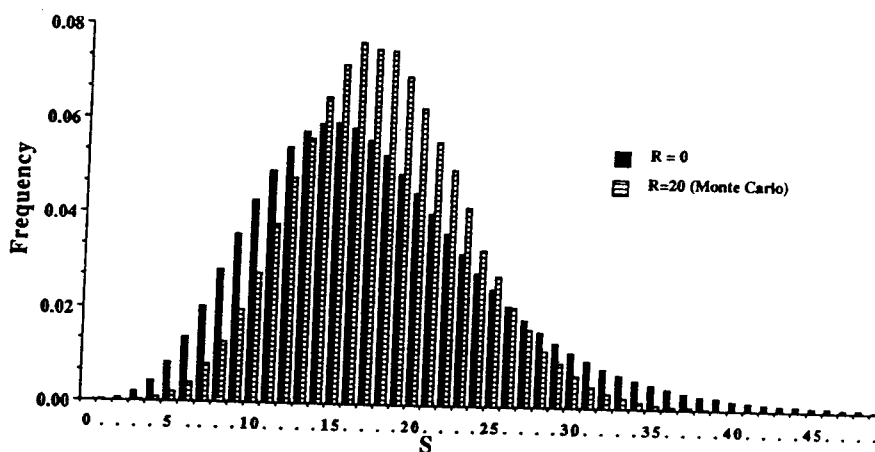

Fig. 2. The distribution of S, the number of segregating sites, in a sample of 20 alleles with θ ($=4N\mu$) $= 20$. The no-recombination distribution ($R = 4N\mu = 0$) was calculated with eqn (12). For $R = 20$, the distribution is an estimate obtained by generating 100 000 replicates by a Monte Carlo method described in the text. The expected value of S for both distributions is 17.7, which can be calculated using eqn (6).

with θ = 2112(0.005) = 10.6, is approximately 2×10^{-6}. Assuming the equilibrium neutral model is correct, one must reject a value of 0.005 as the per base pair mutation parameter for this region. If one assumed that some recombination occurs in this region, the probability of 9 or fewer polymorphic sites is even smaller.

5. RECOMBINATION

Let us consider first two loci. It is assumed that no recombination occurs within each locus but, between the two loci, the probability of recombination is r per generation per offspring produced. If $r = 0$, the two loci will always have the same genealogy. If r is large, in a large random mating population, the genealogies of the two loci will be essentially independent (see eqn 13). The difficult case is with intermediate levels of recombination, when the genealogies at the two loci are correlated. Clearly, the marginal distribution of genealogies for each locus under a neutral model, is the single locus no-recombination distribution described above. The only effect of linkage is to produce a correlation between the genealogies for the two loci.

Let us begin by describing how one might simulate on a computer the genealogy of a sample of two gametes, denoted $a_1(0)b_1(0)$ and $a_2(0)b_2(0)$. We proceed, as before, backward in time. We trace the two lineages back until either a coalescent occurs (probability $1/2N$ per generation) or a recombination event occurs (probability $2r$ per generation). The time back until one of these events is exponentially distributed with mean $2N/(1+R)$, where R is $4Nr$. The probability that the first event is a coalescent event is $1/(1+R)$. In this case, both loci have their MRCA at this time and the genealogies are complete. The other possibility is that the first event is a recombination event. The first event is a recombination event with probability $R/(1+R)$. In this case, one of the two lineages splits in two as illustrated by the genealogy in Fig. 3. In this example, the first event, as one traces backward in time, is a recombination event that occurs in generation t_1. In this example, the ancestral gamete, $a_2(t_1-1)b_2(t_1-1)$, is the recombinant descendant of two individuals in generation t_1, which are denoted $a_2(t_1)$- and -$b_2(t_1)$. At this point, there are three lineages to follow back in time from the three ancestral gametes in generation t_1. One ancestral gamete, denoted $a_1(t_1)b_1(t_1)$, is an ancestor at both loci to one of the sampled gametes. One of the ancestral gametes, denoted $a_2(t_1)$-, is an ancestor of the a_2 allele in the sample, but the b allele of this ancestral gamete, indicated by a hyphen, has no descendant in the sample. The history of this allele represented by the hyphen is of no direct interest. The third ancestral gamete, -$b_2(t_1)$, is the ancester at the b locus of the b_2 allele in the sample. We continue back in time until the next event,

Past
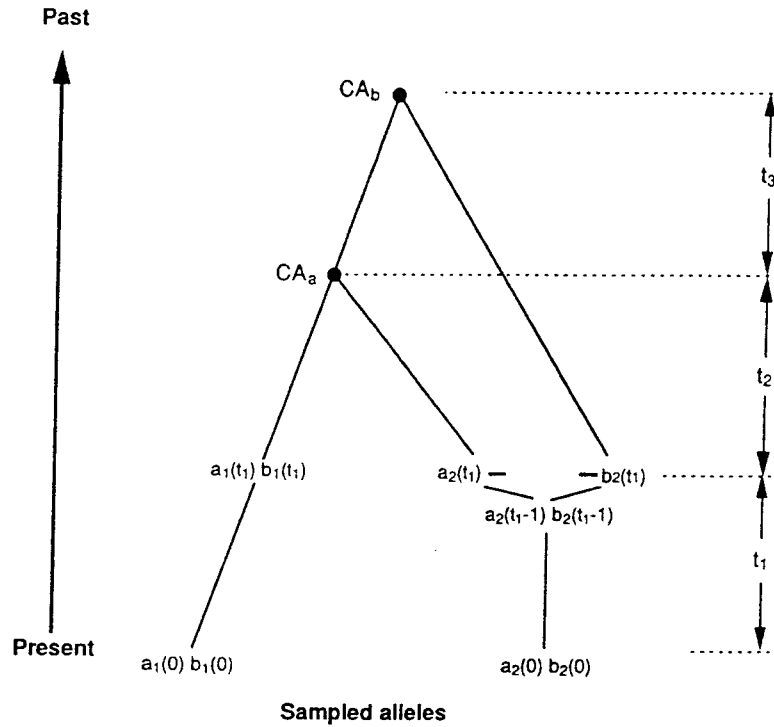

Present

Sampled alleles

Fig. 3. An example two-locus genealogy for a sample of size 2. In this case, the first event, which occurs in generation t_1, is a recombination event such that the ancestor gamete $a_2(t_1-1)b_2(t_1-1)$ is the recombinant descendant of the two gametes $a_2(t_1)$- and -$b_2(t_1)$. The second event is a common ancestor event, labeled CA_a, at which time, the lineages of $a_1(0)b_1(0)$ and $a_2(t_1)$- coalesce. It is at this point in time, t_1+t_2 generations ago, that the most recent common ancestor of the sampled 'a' locus alleles occurred. The next event is a common ancestor event, labeled CA_b. At this time, $t_1+t_2+t_3$ generations ago, the most recent common ancestor of the sampled 'b' alleles occurred.

Fig. 4. An example genealogy for an infinite-site recombination model. The two samples gametes, labeled 1 and 2, are represented by the hatched and dotted bars. Recombination events can occur anywhere along the bars. There are five events in this genealogy, designated RE_1, RE_2, CA_1, CA_2 and CA_3, in order from most recent to most ancient. The most recent event, RE_1, is a recombination event that brought two segments together to form the ancestor of gamete 2. Following lineages backward in time, as usual, the result of RE_1 is the splitting of the lineage of gamete 2 into two parts, one being the lineage of the left end of the gamete, and the other being the lineage of the right part of the gamete. The next event back in the genealogy, labeled RE_2, is also a recombination event
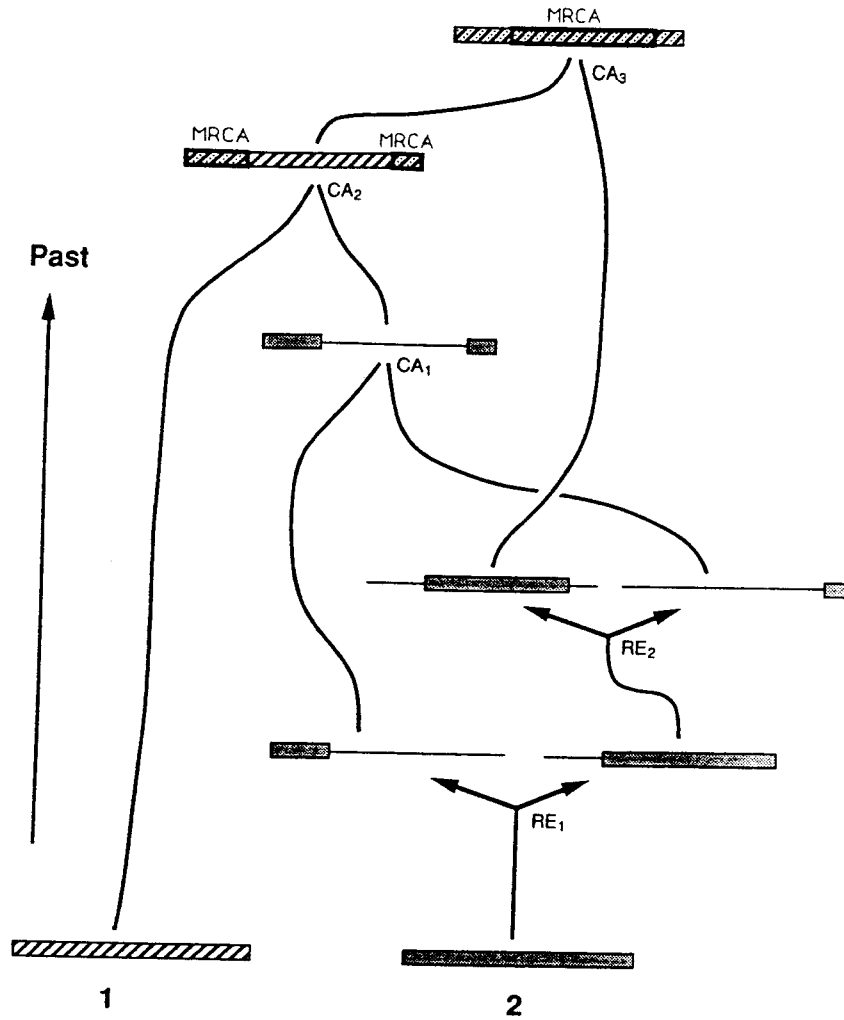
with a crossover in the right-hand segment of an ancestor of gamete 2. At this point in time, there are three distinct ancestors of gamete 2, each being an ancestor of a different part of gamete 2. In contrast, gamete 1 still has a single ancestor. The next event, CA_1, is a common ancestor event involving two ancestors of gamete 2. At this point, one of the two ancestors of gamete 2 is an ancestor for two non-contiguous portions of gamete 2. The next event, CA_2, is a common ancestor event where finally the most recent common ancestor of parts of gametes 1 and 2 occur. The segments with most recent common ancestor at this point are the left end, marked MRCA, and the right end also marked MRCA. The last event, is a common ancestor event where the most recent common ancestor of the sample gametes for the middle segment occurred.

either a coalescent event between any of the three lineages (probability $(\frac{3}{2})/2N$ per generation) or a recombination event (probability r per generation). Note that, during this part of the genealogy, only recombinations involving the lineage of $a_1(t_1)b_1(t_1)$ are relevant. Recombinations in the lineage of $a_2(t_1)$- do not result in any change in the state of the process and are irrelevant to the genealogy of the sampled alleles. Eventually, the two alleles at the **a** locus will coalesce and the two alleles at the **b** locus will coalesce, and the two-locus genealogy will be complete.

By consideration of this two-locus process, it is possible to derive various properties of the joint distributions of the times, t_a and t_b, back to the most recent common ancestors of the **a** and **b** alleles, respectively.

Griffiths (1981a) derived properties of the joint distribution of the number of segregating sites at each locus in samples of size 2, when each locus is assumed to be an infinite-site locus. From Griffiths' result, the correlation of t_a and t_b, the times to the MRCA at locus a and b can be found (Hudson 1983a; Kaplan and Hudson 1985):

$$\text{Cor}(t_a, t_b) = \frac{R+18}{R^2+13R+18} \tag{13}$$

Consideration of this two-locus coalescent shows that the probability that $t_a = t_b$ is exactly the same as the correlation of t_a and t_b (Hudson, unpublished).

Simulations based on the two-locus coalescent were used by Hedrick and Thomson (1986) to study two-locus sampling properties of the neutral model. Kaplan and Hudson (1985) considered the coalescent process for several linked loci to calculate the homozygosity at a global locus made up of several sub-loci between which recombination could occur.

Hudson (1983a) and Kaplan and Hudson (1985) also considered an infinite-site version of the above coalescent process, in which recombination could take place anywhere on a continuous interval that represents a contiguous stretch of nucleotide sites. Figure 4 shows a representation of the genealogy of a sample of two gametes under this model. The process is very similar to the preceding two-locus case, except that recombination takes place at random positions along the continuous interval that represents the sequence. In this case, small contiguous segments are likely to have similar genealogies, but the segments farther apart would be likely to have quite different genealogies. The details of how to carry out such a simulation are described in Hudson (1983a).

In the genealogy in Fig. 4, the MRCA of the segment of DNA in the middle occurs farther back in time than the MRCA of the end segments. In this sense, the size of the genealogy is larger for the middle segment than for the end segments, and assuming that the neutral mutation rate is the same all along the segment, we would expect the number of neutral

mutations per unit length to be greater in the middle segment. In Fig. 5, the outcome of a single realization of this genealogical process is shown for a large contiguous chunk of DNA for a sample size 10. This figure indicates how much the size of the genealogy, as measured by T_{tot}, can vary from one segment to the next. The size of the segment of DNA considered in Fig. 5 is such that $4Nr$ equals 100, where r is the recombination rate per generation between the ends of the region. Although estimates are very rough, this has been estimated to correspond to approximately 5000 base pairs in *D. melanogaster*. (This number can be obtained from estimates of per base pair recombination rate 0.5×10^{-8} and effective population size 10^6: (Hudson and Kaplan 1988; Hudson 1987.)

As before, the total number of segregating sites in a sample, S, conditional on the genealogies of all the segments, is Poisson distributed with mean $\theta T/2$, where in this case T is an average of the sizes of the genealogies of each of the segments weighted by their lengths and θ is $4N$ times the mutation rate for the entire sequence. As the recombination rate increases, the weighted average, T, is made up of greater numbers of
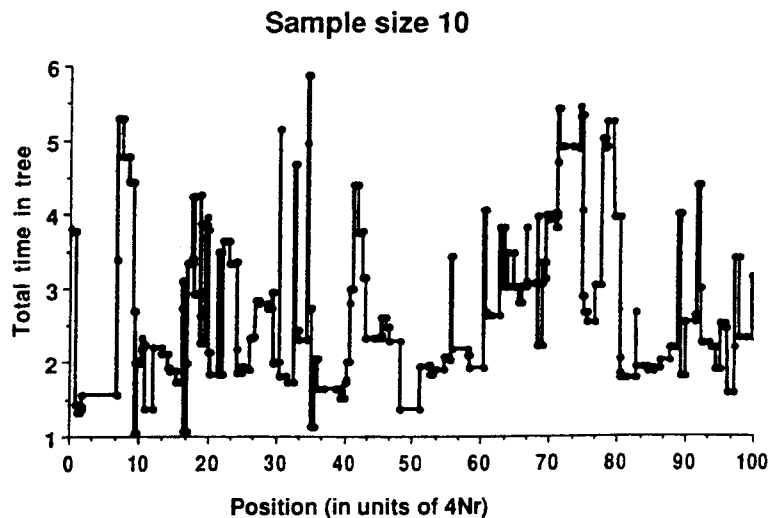

Sample size 10

Fig. 5. The total time in the genealogy of the sample, T_{tot}, measured in units of $4N$ generations, plotted as a function of position, for a single realization of the coalescent process for a neutral infinite-site recombination model. The total length of the region of DNA considered is such that the $4Nr = 100$, where r is the recombination rate between the ends of the region. The horizontal axis is the nucleotide position, as measured by the product of the $4N$ and the recombination rate between the site and the left end of the region considered. Evidently, T_{tot} varies considerably from site to site, over a region this size.

relatively smaller segments that have less correlated genealogies. The result is that the variance of T tends to zero, and S becomes Poisson as the recombination parameter (R) tends to infinity (see Ewens 1979, p. 276). Kaplan and Hudson (1985) showed that the variance of S is

$$\mathrm{Var}(S) \approx \theta \left(\sum_{i=1}^{n-1} 1/i \right) + \theta^2 \, \mathrm{Var}(T) \tag{14}$$

and that

$$\mathrm{Var}(T) \approx \frac{2\left(\sum_{i=1}^{n-1} 1/i^2 \right)}{R^2} \left(-R + \frac{23R+101}{2\sqrt{97}} \log \left(\frac{2R+13-\sqrt{97}}{2R+13+\sqrt{97}} \frac{13+\sqrt{97}}{13-\sqrt{97}} \right) \right.$$
$$\left. + \frac{R-5}{2} \log \left(\frac{R^2+13R+18}{18} \right) \right) \tag{15}$$

For sample size 2, the approximation for Var(T) was based on the usual 'diffusion approximations', but for larger sample sizes there is no theoretical justification for the approximation, except that Monte Carlo simulations indicated that it works quite well in the cases examined, namely with small to moderate values of R (Kaplan and Hudson 1985). The number of recombination events in the genealogy of a sample has been examined by Hudson and Kaplan (1985), and an estimator of R based on inferred numbers of events was investigated. A recombination event was inferred to have occurred between two polymorphic sites when all four possible gametic types (haplotypes) involving the two sites were present in the sample.

The distribution of S in a sample of size 20 for $\theta = 5$ and with $R = 0$ and $R = 20$ are shown in Fig. 2. The mean of S does not depend on R, but this figure shows clearly how recombination can reduce the variance in S. The distribution shown for $R = 20$ is based on 100 000 samples generated by the algorithm described above. The variance of S in the Monte Carlo samples was 28.04, whereas the variance calculated with eqns (14) and (15) is 28.28.

6. ESTIMATING θ OR N

One can use S to estimate θ or, if the neutral mutation rate (μ) is known, the population size N. The two commonly used methods are moment estimators. Because the expected number of differences between two alleles is θ, an obvious estimator of θ is $\bar{\theta}$, the average pairwise number of differences between alleles in a sample (see Nei 1987, eqn 10.6). This

is an unbiased estimator of θ. Tajima (1983) showed that under the W-F model with no recombination, the variance of this estimator is (see also Nei 1987, eqn 10.9):

$$\text{Var}(\bar{\theta}) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \tag{16}$$

Watterson (1975) suggested an estimator based on eqn (6), namely:

$$\hat{\theta} = \frac{S}{\sum\limits_{i=1}^{n-1} 1/i} \tag{17}$$

This estimator is clearly unbiased. Under the no-recombination model, the variance of this estimator can easily be calculated using eqn (7), because:

$$\text{Var}(\hat{\theta}) = \frac{\text{Var}(S)}{\left(\sum\limits_{i=1}^{n-1} 1/i\right)^2} \tag{18}$$

The variance of θ is always less than the variance of $\hat{\theta}$. With recombination, both of these estimators have substantially reduced variance. The variance of $\hat{\theta}$ in the presence of recombination can be estimated using eqns (14), (15) and (18).

In some circumstances, the reduced variance of S in the presence of recombination may be justification for considering nuclear genes instead of mitochondrial genes for certain problems. For example, recent studies (Avise et al. 1988) of mitochondrial genes were used to estimate effective population sizes, using prior estimates of μ. Although practical considerations concerning the relative ease of isolation of mtDNA compared to nuclear DNA may mitigate against the use of nuclear DNA, more precise estimates might be obtained with nuclear data.

For the no-recombination model, maximum likelihood estimates of θ based on S can be obtained, and it has been shown that the maximum likelihood estimates always exceed $\hat{\theta}$ (Tavaré 1984). I have examined a small number of cases and always found that the mean square error of the maximum likelihood estimate exceeds the mean square error of $\hat{\theta}$.

7. MIGRATION AND GEOGRAPHIC STRUCTURE

A number of authors have utilized the genealogical approach to consider properties of samples when there is geographic structure (Griffiths 1981b;

Slatkin 1987, 1989; Strobeck 1987; Tajima 1989, Takahata 1988). To illustrate the concepts, let us consider a two-population symmetric island model. Each subpopulation consists of N diploids. Each generation, a small fraction m of each subpopulation is made up of migrants from the other subpopulation. In other words, each individual's parent was resident in the same population with probability $1-m$, and in the other subpopulation with probability m. As with the panmictic model, the probability that two alleles from the same subpopulation have a common ancestor in the previous generation is $1/2N$. Two alleles from different subpopulations have negligible probability of having a common ancestor in the previous generation. Putting these properties together, we can describe the genealogical process for a sample of alleles, n_1 from subpopulation 1 and n_2 from subpopulation 2. We denote the state of the ancestral lineages by an ordered pair, (i,j), indicating that i ancestors reside in subpopulation 1 and j reside in subpopulation 2. As usual, we trace the lineages back in time, in this case until either a common ancestor occurs or one of the lineages changes residence. This time is exponentially distributed with mean

$$\frac{1}{\left(\binom{n_1}{2} + \binom{n_2}{2} + (n_1+n_2)\frac{M}{2}\right)}$$

measuring time in units of $2N$ generations and where $M = 4Nm$. Given that one of the two events occurs, the probability that it is a common ancestor event among the n_i lineages in subpopulation i is:

$$\frac{\binom{n_i}{2}}{\left(\binom{n_1}{2} + \binom{n_2}{2} + (n_1+n_2)\frac{M}{2}\right)}$$

If the common ancestor event occurs in subpopulation 1, the state of the ancestral lineages changes to (n_1-1, n_2). The probability that the event is a change of residence of a lineage in subpopulation i is:

$$\frac{n_i\frac{M}{2}}{\left(\binom{n_2}{2} + \binom{n_2}{2} + (n_1+n_2)\frac{M}{2}\right)}$$

If a lineage changes from subpopulation 1 to subpopulation 2, working

backward in time, then the state of the ancestral lineages changes to (n_1-1, n_2+1). And the process continues.

As described, the process is amenable to implementation as a Monte Carlo simulation. Strobeck (1987), Tajima (1989) and Slatkin and Maddison (1989) have carried out Monte Carlo simulations based on this approach.

To illustrate how analytical results can be obtained by this approach, we calculate the probability of identity of two alleles sampled from the same subpopulation, $P_s(\theta)$, and the probability of identity of two alleles from different subpopulations, $P_d(\theta)$. As noted earlier, we can calculate the moments of S once these identity coefficients are obtained. We assume a symmetric island model, as above, except with n subpopulations. We trace backward in time in the genealogy of two alleles from the same subpopulation, until either a coalescent, mutation or a migration event occurs. If the first event is a coalescent event, probability $1/(1+\theta+M)$, the two alleles are identical. If the first event is a mutation, probability $\theta/(1+\theta+M)$, the two alleles are not the same. If the first event is a migration, then the probability of identity of the two alleles is $P_d(\theta)$. This leads to the following equation for $P_s(\theta)$:

$$P_s(\theta) = \frac{1}{1+\theta+M} \cdot 1 + \frac{\theta}{1+\theta+M} \cdot 0 + \frac{M}{1+\theta+M}P_d(\theta) \qquad (19)$$

For two alleles from two distinct populations, only mutations and migration events that bring the two lineages into the same subpopulation need to be considered. If the first event is a mutation event, probability $\theta/(\theta+M/n)$, the two alleles are different. If the first event is a migration event that takes one of the lineages into the subpopulation of the other, probability $(M/n)/(\theta+M/n)$, the probability of identity is $P_s(\theta)$. This leads to the following equation for $P_d(\theta)$:

$$P_d(\theta) = \frac{\dfrac{M}{n}}{\theta+\dfrac{M}{n}} P_s(\theta) \qquad (20)$$

Solving eqns (19) and (20),

$$P_s(\theta) = \frac{(n-1)\theta+M}{(n-1)\theta^2+\theta(n-1+Mn)+M} \qquad (21)$$

and

$$P_d(\theta) = \frac{M}{(n-1)\theta^2+\theta(n-1+Mn)+M} \qquad (22)$$

These results are not new, having been obtained by several others without consideration of the coalescent process (see Crow and Aoki, 1984, and references therein). To obtain the expectation of the times to the common ancestor, t_s and t_d, for two alleles from the same subpopulation and different subpopulations, respectively, we can use the method described earlier in Section 4. Treating the identity coefficients $P_s(\theta)$ and $P_d(\theta)$ as moment-generating functions, the expectations of t_s and t_d are:

$$E(t_s) = P'_s(0) = n \tag{23}$$

and

$$E(t_d) = -P'_d(0) = n + \frac{n-1}{M} \tag{24}$$

The expected number of differences between two alleles from the same subpopulation is $\theta E(t_s) = n\theta$, and for two alleles from different subpopulations $\theta E(t_d) = n\theta + (n-1)\theta/M$ (Li 1976; Slatkin 1987; Strobeck 1987). Therefore, the expected time to the common ancestor of two alleles sampled from one subpopulation, as well as the expected number of differences, is independent of migration rate. If M is small, the expected time to the common ancestor of two alleles from different populations is relatively large, as is their divergence. This is consistent with our intuition that if the migration rate is low, the two subpopulations will be substantially differentiated. This is illustrated by the genealogies in Fig. 6. Tajima (1989) has used the coalescent approach to study the expected number of segregating sites in samples larger than 2.

Although the mean number of differences between two alleles from the same subpopulation does not depend on the migration rate, other aspects of the distribution do depend on the migration rate. In Fig. 7, the distribution of the average pairwise difference between 10 alleles sampled from the same subpopulation is shown. In this case, there were a total of three subpopulations and $M = 4Nm = 0.2$, and $\theta = 5.0$. Also shown is the distribution of the same statistic when $M = \infty$, i.e. a panmictic population with $\theta = 15.0$, and for a panmictic population with $\theta = 5.0$. The distributions with $M = 0.2$ and $M = \infty$ have the same mean, but otherwise the distributions are quite different. The $M = 0.2$ case has its mode and much of its mass around 5, with a very long tail. Except for the long tail, the distribution looks much like the distribution for a panmictic population with $\theta = 5.0$. This is because with the small migration rate, most of the time coalescent events occur within the subpopulation without any migration, and therefore the sample is like a sample from a single population with parameter $\theta = 5.0$. In contrast, the $M = \infty$ case has its mode around 15.
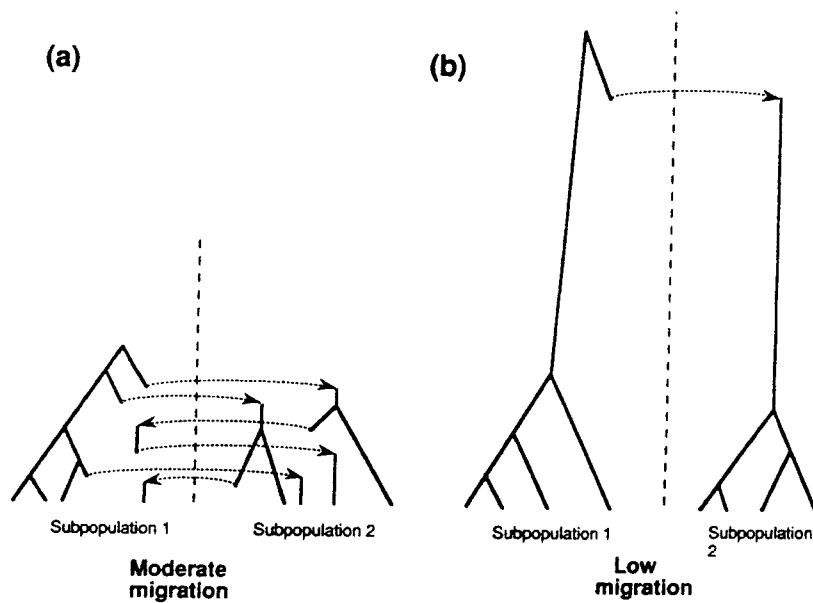
Fig. 6. (a) An example of a genealogy for a sample of size 8, 4 from each of 2 subpopulations, when the migration rate is moderately high. Each migration event is indicated by a dotted line with an arrow that indicates the actual direction of movement of an individual migrant. In this case, there would be relatively little differentiation of the two subpopulations. (b) An example genealogy with low migration rate. In this genealogy there is a single migration event. Alleles from within a subpopulation will be much more similar than alleles from different subpopulations.

These genealogies can also be interpreted as genealogies of gametes bearing different selected alleles (see Section 8). Subpopulation 1 would represent the pool of S-bearing gametes, and subpopulation 2 would represent the pool of F-bearing gametes. In this case, the dotted lines with arrows indicate mutations making an F allele into an S allele, and vice versa. If the mutation rate between the selected alleles is high, sequences bearing different alleles will be no more diverged than alleles bearing the same allele. If the mutation rate between F and S is low, S- and F-bearing gametes will be relatively diverged from each other. The genealogies could also represent the genealogy of a site linked to the selected locus. In this case, the dotted lines with arrows would represent mutations between the selected alleles and/or recombination events between the site and the selected locus.
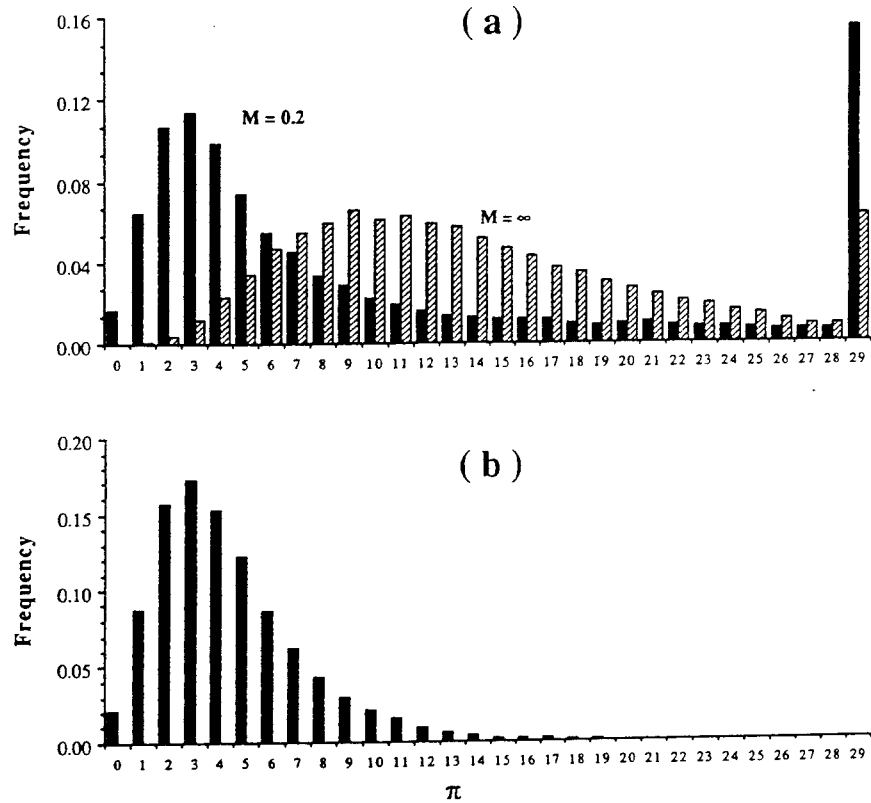
Fig. 7. (a) The distribution of π, the average pairwise number of differences between alleles in a sample of size 10 from a single subpopulation. The population is made up of three subpopulations, each of diploid size N, with $\theta = 4N\mu = 5$, and with $M = 4Nm = 0.2$ (solid bars) and $M = \infty$ (hatched bars). The mean for both distributions is approximately 15. (b) The distribution of the same quantity, for a single panmictic population, with $\theta = 4N\mu = 5$. Note the similarity with the low migration case in (a).

8. BALANCING SELECTION

Kaplan *et al.* (1988) have shown how the coalescent process can be analyzed under models with certain forms of selection. They focus primarily on the case where some form of balancing selection maintains a two-allele polymorphism at a particular nucleotide site, the 'selected site'. It is assumed that recurrent mutation between the two 'selected' alleles, designated F and S, occurs at rate v per replication. The analysis addresses the question: For sites completely linked to the selected site, how is the genealogy different from a genealogy of a neutral site isolated from any

selection? When selection is weak and the frequency of the alleles at the selected site can drift considerably, numerical results can be obtained with some pain (Darden *et al.* 1989). Results are fairly simple when selection is strong and unchanging, so that the frequencies of the selected alleles, S and F, remain constant.

In the case of strong and constant selection, the coalescent process of sampled alleles is analogous to the coalescent process for the subdivided population model considered earlier, except that migration is no longer symmetric. If the frequencies of S and F are p and q, respectively, then one can consider the population to be subdivided into two subpopulations of size $2Np$ and $2Nq$. Mutation plays the role of migration. Each generation, an average of $2Nqv$ F alleles mutate (migrate) to the S allele (subpopulation) and $2Npv$ alleles mutate in the other direction. This means that a fraction, $2Nqv/2Np$, of the S alleles in each generation, approximately, are descendants of F alleles of the previous generation. In other words, an S allele of one generation has as parent an F allele with probability qv/p. If one is considering n_1 S alleles, the probability, P_{SF}, that one of them has as parent an F allele is, approximately:

$$P_{SF} = n_1 \frac{qv}{p}$$

Similarly, the probability, P_{FS}, that one of n_2 F alleles has an S allele as parent in the previous generation is:

$$P_{FS} = n_2 \frac{pv}{q}$$

The quantities pv/q and qv/p are the analogues of migration in the subdivided population model. In this case, 'migration' is not symmetric and the sizes of the two 'subpopulations' are not equal.

The probability of coalescent events are functions of the size of each subpopulation of alleles. For example, the probability that two S gametes have a common ancestor in the previous generation is approximately $1/2Np$, and the corresponding probability for two F gametes if $1/2Nq$. More generally, the probability, $P_{CA,S}$, that for n_1 S alleles some pair will have a common ancestor in the previous generation is:

$$P_{CA,S} = \frac{\binom{n_1}{2}}{2Np}$$

Similarly, for n_2 F alleles, the probability, $P_{CA,S}$, that some pair of the alleles will have a common ancestor in the previous generation is: