

## 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

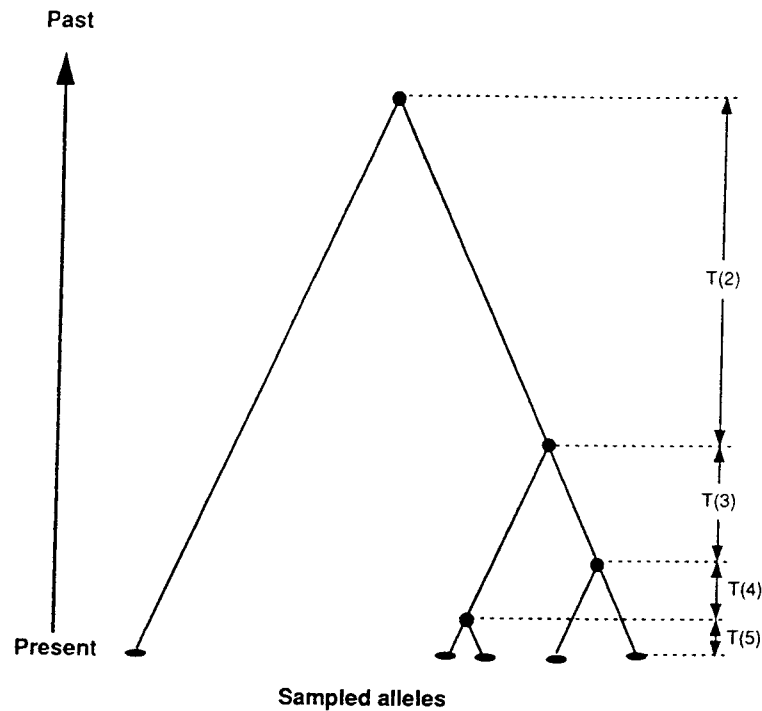


Fig. 1. An example of a genealogy of a sample of five alleles, showing the time intervals between coalescent events. In this figure, the intervals,  $T(i)$ , are shown with lengths proportional to their expected values as given by eqn (5).

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, 1982*a,b*) and the review by Tavaré (1984). Much of the very elegant and useful work of Griffiths (1980), Watterson (1984) and Padmadasastra (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; Donnelly 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 1982*a*, 1982*b*). Also, the important work on the relationship between gene trees and species trees will not be discussed (Hudson 1983*b*; 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 1982*a,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 effect 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 transversions 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,  $\mu$ , 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  $\mu t$ . 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  $\mu t$ , regardless of these other factors.

This property of the constant-rate neutral mutation process will be exploited in the following way. Let  $T_{\text{tot}}$  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  $T_{\text{tot}}$ , is Poisson distributed with mean  $\mu T_{\text{tot}}$ . Once the distribution of  $T_{\text{tot}}$  is determined under a particular model, the distribution of  $S$  can easily be obtained. For example, if the first two moments of  $T_{\text{tot}}$  are determined, then the first two moments of  $S$  can be calculated using properties of compound distributions as:

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

and

$$\text{Var}(S) = \mu E(T_{\text{tot}}) + \mu^2 \text{Var}(T_{\text{tot}}) \quad (2)$$

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  $T_{\text{tot}}$  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  $T_{\text{tot}}$ . Two different models that lead to the same mean value of  $T_{\text{tot}}$  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, 1982*a,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  $\times$  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:

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

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:

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

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:

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

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  $\mu$  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  $T_{\text{tot}}$ , 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

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

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):

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

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 sites 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:

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

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  $\theta$  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  $2\mu$ . 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:

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

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):

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

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

$$\begin{aligned} Q_n(j) &= \left( \frac{n\mu}{n\mu + \frac{\binom{n}{2}}{2N}} \right)^j \frac{\frac{\binom{n}{2}}{2N}}{n\mu + \frac{\binom{n}{2}}{2N}} \\ &= \left( \frac{\theta}{\theta + n - 1} \right)^j \frac{n - 1}{\theta + n - 1} \end{aligned} \quad (11)$$

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