Annu. Rev. Genet. 1995. 29:401–21

COALESCENTS AND GENEALOGICAL STRUCTURE UNDER NEUTRALITY

Peter Donnelly[1] *and Simon Tavaré*[2]

[1]Departments of Statistics, and Ecology and Evolution, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637

[2]Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, California 90089-1113

KEY WORDS: Mitochondrial Eve, variable population size, population structure, mutation rates

ABSTRACT

Genealogical or coalescent methods have proved very useful in interpreting and understanding a wide range of population genetic data. Our aim is to illustrate some of the central ideas behind this approach. The primary focus is genealogy in neutral genetic models, for which the effects of demography can be separated from those of mutation. We describe the coalescent for panmictic populations of fixed size, and its extensions to incorporate various assumptions about variation in population size and nonrandom mating caused by geographical population subdivision. The effects of such genealogical structure on patterns and correlations in genetic data are discussed. An urn model is useful for simulating samples at loci with complex mutation mechanisms. We give two applications of the genealogical approach. The first concerns methods for estimating the mutation rate from infinitely-many-sites data, and the second relates to inference about recent common ancestors and population history.

CONTENTS

0066-4197/95/1215-0401$05.00

INTRODUCTION

One of the important recent developments in population genetics modeling is the use of so-called coalescent or genealogical methods. In considering the structure of genetic data, these methods focus primarily on the genealogical tree of the sampled genes. They are attractive for several reasons: (*a*) Quantitative analysis of stochastic models is usually easier with genealogical methods than with traditional approaches; (*b*) the structure of genetic data reflects, in large part, the underlying genealogy, so that an understanding of genealogy enhances a qualitative understanding of the patterns of variation in genetic data; (*c*) use of the coalescent leads to extremely efficient simulation methods, and (*d*) it provides inference techniques for genetic data that allow, for the first time, full use of the information in that data.

Our aim here is to illustrate some of the central ideas underlying genealogical methods. It is neither possible, nor perhaps helpful, to be exhaustive, and we do not attempt a complete historical account even of the areas we do discuss. Earlier reviews of the approach may be found (see 7, 21, 22, 47, 48). A notable absentee from our coverage is genealogy in models with recombination (see 12, 21, 22). We have aimed for a middle course between giving free reign to intuition and giving precise mathematical statements of exact conditions under which the approximations apply.

Throughout, we focus on genealogy in neutral genetics models. Under the assumption of neutrality, the effects of population demography may be separated from those of mutation. The coalescent processes described below capture the relevant features of the demographic history of a sample, as far as its current genetic composition is concerned. In studying the evolution of different systems (for example, DNA sequences, multigene families, or mini-satellites) the underlying genealogy is probabilistically the same. The patterns of variation in such genetic data sets result from superimposing the effects of mutation on the sample genealogy. In contrast, for models incorporating natural selection, not only is relatively less known about genealogy [see for example (23, 24, 26)], but the application of genealogical methods is fundamentally more difficult, precisely because the effects of mutation and demography are inseparable.

The descriptions of genealogy discussed below relate fundamentally to haploid genetic units. In some contexts, perhaps most notably evolutionary modeling for mitochondrial DNA, the assumption of haploidy applies explicitly. In modeling diploid organisms, the analysis applies provided the focus is on regions of DNA within which recombination can be ignored. In this context, think of the DNA, or gene, itself as an "individual" that has a single "parent" (the gene from which it is copied) in the previous generation, and will have a number of "offspring" (genes that originate as copies of it) in the next generation. The fact that these haploid genes reside in pairs within organisms is not relevant in addressing certain evolutionary questions. Throughout, we frame the discussion in terms of haploid genes. As a consequence, for example, when we refer to the size of a population we mean the number of haploid genes rather than the number of diploid individuals.

THE COALESCENT

The ancestry of a random sample of n genes from a population that has evolved with constant size N over many generations is often modeled by a stochastic process known as a coalescent. This process was introduced by Kingman [(27–29), see also Hudson (20) and Tajima (44)], as an approximation, valid in the limit of large population size, to the ancestral structure of a wide variety of neutral demographic models.

We set the discussion in the context of models with discrete, nonoverlapping generations, but the result applies more generally (29). Suppose that in a particular generation the genes are labeled 1,2, . . . ,N, and let ν_1, ν_2, . . . ν_N be the respective numbers of descendants they have. The demographic structure of the population assumes panmixia, which is reflected in the assumption that the ν_i are exchangeable random variables. This assumption means that the joint distributions of these offspring numbers is unchanged under relabeling of the genes and, in particular, that the ν_i are identically distributed. We suppose also that these joint distributions do not change over time.

The simplest aspect of genealogy involves counting ancestors: Let $A_N(r)$ denote the number of distinct ancestors, r generations earlier, of a sample of n genes taken from the population in a particular generation. For a specific value of N and for a given distribution for the ν_i, the properties of this ancestral process are not in general easy to calculate. However, when the population size is large, and time is measured in units of N generations, there is a simple process that provides a good approximation to $A_N(\cdot)$. Together with an additional technical condition that is satisfied in cases of interest, Kingman (27, 29) showed that if

$$\lim_{N\to\infty} \mathrm{Var}(\nu_1) = \sigma^2, \ 0 < \sigma^2 < \infty, \qquad 1.$$

then the ancestral process $\{A_N(\lfloor Nt \rfloor), t \geq 0\}$, where $\lfloor x \rfloor$ denotes the integer part of x, converges (in distribution) to a process $\{A(\sigma^2 t), \mathrm{t} \geq 0\}$. It is worth emphasizing that the time scale of the approximation is determined by the parameter σ^2. For the Wright-Fisher model, in which the ν_i have a symmetric multinomial distribution, $\sigma^2 = 1$. The condition in Equation 1 ensures that on this new time scale, the genealogy does not degenerate.

The process $\{A(\sigma^2 t), t \geq 0\}$ is extremely simple: It is a continuous-time

Markov chain that starts from $A(0) = n$, and moves through the sequence of states $n, n - 1, \ldots, 3, 2$ as the number of ancestors decreases by exactly one each time it changes value. The process eventually reaches the value 1 at the time T_{MRCA} that the sample has been traced back to its most recent common ancestor (MRCA). T_{MRCA} is sometimes called the coalescence time of the sample. The amount of time T_j for which there are j distinct ancestors in the history of the sample has an exponential distribution with mean

$$\mathsf{E}(\mathrm{T}_j) = 2/(j(j-1)\sigma^2), j = 2,3, \ldots, n,$$

the T_j being independent for different j. Clearly, $T_{MRCA} = T_2 + T_3 + \cdots + T_n$, and

$$\mathsf{E}(T_{MRCA}) = \frac{2}{\sigma^2}(1 - \frac{1}{n}).$$

The role of σ^2 is intuitively reasonable: If the variance of the number of offspring is bigger, the time back to the MRCA should be shorter, and vice versa. The variance of T_{MRCA} is $1/\sigma^4$ for a sample of $n = 2$ genes. As n increases, this variance increases to about $1.16/\sigma^4$, with the latter value providing an accurate approximation to $\mathsf{Var}(T_{MRCA})$ for samples of size five or larger.

A full description of the genealogy of the sample requires more than just ancestral numbers. One natural way of picturing the genealogy is as a tree: The sampled genes are the "leaves", or tips, of the tree and the MRCA is the root. We think of, and draw, the trees "vertically," looking backwards in time from the sampled genes "up" to their common ancestor. See for example Figure 1.

The coalescent describes the probabilistic structure of the random tree, which arises as an approximation to the actual genealogical tree of a sample of genes from a large population, when time is measured in units of N generations. The tree has n leaves, one labeled by each of the sampled genes. At any time in the past (more recently than the common ancestor of the sample) there is one branch in the tree for each ancestor of the sample. As we move up the tree, the number of branches decreases each time the number of ancestors decreases. Thus, with time measured from the tips, there are j branches in the tree for time $T_{\mathrm{j}}, j = n, n - 1, \ldots, 2$. Each time the number of branches decreases, two existing branches are chosen uniformly at random and coalesced.

We shall see below that there are close relationships between the shape of the genealogical tree and patterns to be expected in genetic data from the sampled genes, so we turn now to several qualitative properties of coalescent trees. One of the most striking properties, which is evident in Figure 1, is the extent to which the tree is dominated by the last two branches. For a sample of size n, the mean time for which the tree has only two branches is $1/\sigma^2$. The

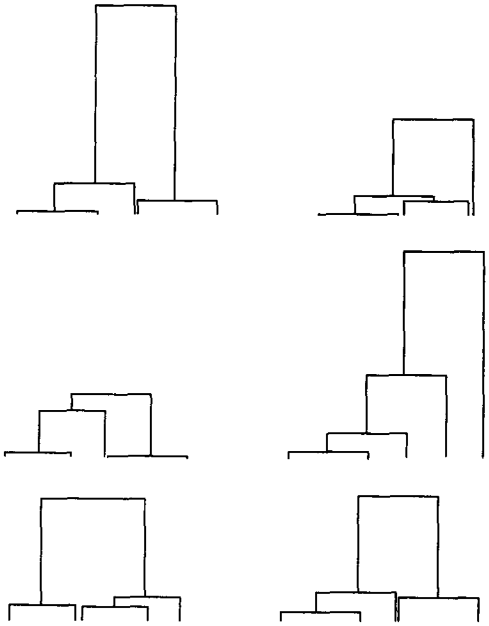

Figure 1 Six realizations, drawn on the same scale, of coalescent trees for a sample of $n = 5$. (In each tree the labels 1,2,3,4,5 should be assigned at random to the leaves.)

mean time for which the tree has more than two branches, namely $(1/\sigma^2)(1 - 2/n)$, is smaller: For much of the time since its common ancestor, the sample has only two ancestors. Further, for any sample size n, the variability in the time T_2 for which there are two branches in the tree accounts for most of the variability in the depth of the whole tree. In fact, the variability in the time for which there are 2, 3, 4, or 5 branches accounts for effectively all the variability in the depth of the tree.

Variable Population Size

In this section we discuss the behavior of the coalescent that arises as an approximation to the genealogical structure of populations with variable size. We consider first the case of deterministically varying population size.

Suppose that the generation at the time of sampling consisted of $M(0) = N$ genes and that the size of the population r generations earlier was $M(r)$. We assume the same sort of reproductive symmetry as earlier: In generation r the numbers $\nu_i^{(r)}$, $i = 1,2,\ldots,M(r)$, of descendants of the ith gene, are exchangeable random variables with sum $M(r-1)$ and variance $\mathrm{Var}(\nu_i^{(r)}) = \sigma^2(r)$. For the Wright-Fisher model, the $\nu_i^{(r)}$ have a symmetric multinomial distribution, and $\sigma^2(r) = M(r-1)(M(r)-1))/M(r)^2$.

The coalescent arises naturally when all the population sizes $M(r)$ are large. Once more we measure time in units of N generations, and we assume that there is an increasing continuous function $\Lambda(t)$ for which

$$\lim_{N\to\infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{\sigma^2(j)}{M(j-1)} = \Lambda(t), \qquad 2.$$

where $0 < \Lambda(t) < \infty$ for $t > 0$. For the Wright-Fisher case (39), the condition becomes

$$\lim_{N\to\infty} \sum_{j=1}^{\lfloor Nt \rfloor} \frac{1}{M(j)} = \Lambda(t).$$

In what follows we suppose that $\Lambda(t)$ has a density $\lambda(\cdot)$: $\Lambda(t) = \int_0^t \lambda(s)ds$. For theWright-Fisher model, $1/\lambda(t)$ may be thought of as the size of the population Nt generations ago, relative to its size N now.

Write $\{A^\nu(t), t \geq 0\}$ for the limiting process, which approximates ancestral numbers here. Its distribution can be defined (13, 14, 29) by a time change of the ancestral process $A(\cdot)$ that arose in the constant population-size case:

$$A^\nu(t) = A(\Lambda(t)),\ t \geq 0. \qquad 3.$$

Most interesting properties of the variable-size process can be calculated using the representation in Equation 3. For example, for $t > 0$,

$$P(T_j > t \mid \mathrm{T_n} + \ldots + T_{j+1} = s) = \exp(-\tbinom{j}{2})(\Lambda(t+s) - \Lambda(s))). \qquad 4.$$

This leads to a natural scheme for simulating the times $T_n, \ldots, T_2$ between successive coalescences in the genealogical tree. Let $U_n, \ldots, U_2$ be independent random variables uniformly distributed on (0,1), and let $F_{s,j}(t)$ denote the expression on the right of Equation 4. First simulate a value for T_n by setting $s = 0$ and $j = n$ and solving the equation

$$F_{s,j}(T_j) = U_j. \qquad 5.$$

To generate a value for T_{n-1}, update by setting $s = s + T_j$ and $j = n - 1$, and repeat the step in Equation 5. This procedure may be continued recursively until finally a value for T_2 is generated.

The approximation result says that with time measured in units of N generations, if there are k ancestors of the sample at some time s in the past, the rate at which the number of ancestors (equivalently, the number of branches in the genealogical tree) decreases is $k(k - 1)\lambda(s)/2$. Suppose, for example, that the population has increased in size, forward through time, with Wright-Fisher demography. In this case, $M(r) < N$ for all r, so that $\lambda(t) > 1$ for all t. That is, with time measured in units of the current population size, coalescences happen at a faster rate, relative to a population of constant size N, in a population whose size has grown to N. This is as we would expect intuitively: The smaller population sizes in the past make sharing of ancestors, and hence coalescences, more likely.

One particular growth scenario of some interest is that of a population whose size has grown exponentially forward in time. In this case, $\lambda(t) = e^{\beta t}$ for an appropriately scaled growth rate $\beta > 0$. Thus the pairwise coalescence rate in the genealogical tree grows exponentially. The effect of this on the general shape of the genealogical tree can be quite marked. Its relative effect is to stretch the tree near the leaves and to compress it substantially near the root. Intuitively, few coalescences happen while the population size is large. The rapid decrease backwards in time of the population size then induces all the remaining coalescences. For populations that have grown exponentially from a very small size, the resulting genealogical trees tend to resemble a star phylogeny (39), in stark contrast to their typical shape (Figure 1) in the case of populations of constant size. Note that this conclusion depends on the initial small size of the population. It is not true for populations that were initially of nontrivial size for some time before growing exponentially (32, 33).

Our discussion of genealogy with variation in population size applies in the case of deterministically changing population sizes. Such an assumption is unrealistic in most applications, in which the variation in the population size is the result of exogeneous and/or endogeneous stochastic effects. Nonetheless, in many applications, information may be available about the sizes of the population at various times in the past, so that the natural question concerns the structure of the genealogy conditional on observed values of the population sizes.

It turns out that for a large class of models (P Donnelly & TG Kurtz, unpublished) the structure of genealogy, conditional on the observed values of the population sizes, is the same as it is in the case of deterministic population sizes described above. This class includes the classical genetics models in which the distribution of offspring numbers is specified conditional on the values of exogenously varying population sizes, and models arising from branching processes in which, loosely speaking, numbers of descendants of different genes are independent.

On the other hand, the results described above do not apply in general. It is true for very general neutral models that unless there are discontinuities, i.e. sudden changes, in the processes governing the population size, the ancestral process can be represented as a time change of the process described in the previous subsection. However, the form of the time change, which is in general different from Equation 2, depends on properties of the random process governing the rate at which individuals are born in the population, about which little is known in many practical contexts. It thus appears that some caution is appropriate in applying the above results on the coalescent in populations of variable size.

For any of the models just described, the topology of the genealogical tree has the same structure in the variable population-size case as in the constant-size case: When the number of branches in the tree decreases by one, any two of the existing branches are equally likely to coalesce.

Genealogy in Geographically Structured Populations

The coalescent processes introduced above describe genealogical relationships in a randomly mating population. The assumption of panmixia may be unrealistic for many populations, notably those that exhibit geographical structure in which matings between "nearby" individuals are more likely than between "distant" individuals. We now examine genealogical structure in one class of models for geographically structured populations. [For further background, see for example (21, 35, 45)]. We suppose that the population consists of a discrete collection of subpopulations, or colonies, each of which is large and panmictic. The colonies are partially isolated from each other, with gene flow resulting from the migration between colonies. Such models may be inappropriate for some applications, in which case the genealogical structure is different from that described below.

Specification of the demography of the population requires a description of the reproductive mechanism within each colony and of the migration processes between colonies. Suppose the colonies are labeled, and write S for the set of labels. Denote by $N_i = c_iN$ the size of subpopulation i, where c_i is a positive integer constant. For definiteness we suppose that reproduction within each colony is governed by the neutral Wright-Fisher model, but the analysis extends, with the analogous dependence on the asymptotic variance of offspring numbers, to any of the exchangeable models for reproduction described above. In each generation, after reproduction, a proportion γ_{ij} of the genes born in colony i migrate to colony j ($\gamma_{ij} \geq 0$, $\sum_{j \neq i} \gamma_{ij} \leq 1$). We assume that the colony sizes are not changed by migration, so that for all $i \in \mathcal{S}$, $c_i \sum_{j \neq i} \gamma_{ij} = \sum_{j \neq i} c_j \gamma_{ji}$.

In this geographically structured setting, a description of ancestral relationships requires that we keep track not only of the number of ancestors of the sample of genes, but of the locations of each of the ancestors. In order to facilitate comparison with the panmictic case, we make the additional assumption here that the number of colonies is finite. We describe the approximation result somewhat informally. [For a formal statement, and details in the case of an infinite collection of colonies, see (19)]. Suppose the population has been evolving indefinitely in the manner just described, and consider a sample of n genes taken from a particular generation, which we call the present. When N is large, the genealogy of the sample is well approximated by a process (or random tree), which we call the structured coalescent. Write $C = \sum_{i \in \mathcal{S}} c_i$, so that CN is the total population size. With time measured into the past in units of CN generations, write $A_i(t)$, $i \in \mathcal{S}$, for the number of ancestors of the sample who were in the ith colony t time units ago. Two sorts of changes are possible in this structured ancestral process. It may be that two particular ancestors in one colony, say colony i, themselves share a common ancestor. In this case, the number of ancestors in colony i decreases by one. Alternatively, in a particular generation one of the ancestors in colony i may have migrated there from colony j. In this case, the number of ancestors in colony i decreases by one and the number of ancestors in colony j increases by one.

The structured ancestral process $\{A_i(t), t \geq 0, i \in \mathcal{S}\}$ is a continuous time Markov chain with transitions from a state $(A_1(t), A_2(t), \ldots)$ to a state of the form $(A_1(t), \ldots, A_{i-1}(t), A_i(t) - 1, A_{i+1}(t), \ldots)$ at rate $CA_i(t)(A_i(t) - 1)/(2c_i)$, and to a state of the form $(A_1(t), \ldots, A_{i-1}(t), A_i(t) - 1, A_{i+1}(t), \ldots, A_{j-1}(t), A_j(t) + 1, A_{j+1}(t), \ldots)$ at rate $A_i(t)M_{ij}/2$, where $M_{ij} = \lim_{N \to \infty} 2CNc_j\gamma_{ji}/c_i$. Its initial value is $A_i(0) = n_i$, where n_i is the number of genes sampled from colony i, $i \in \mathcal{S}$, and $\sum_{i \in S} n_i = n$.

To develop intuition on the effect of population structure on genealogy, focus attention on the case of C equal-sized colonies, for which $c_i = 1$ for all $i \in \mathcal{S}$. In the structured coalescent, lineages can coalesce only if they are in the same colony. If, as in the panmictic case, time is measured in units of the total population size, then each pair of lineages within a particular colony coalesces at rate C. In addition, each lineage in colony i "moves" to colony j at rate $M_{ij}/2$, where the values M_{ij}, some of which may be zero, reflect the migration rates.

Suppose, for definiteness, that the sampled individuals are all taken from a single colony. In this case, the initial coalescences happen C times more quickly than in the panmictic setting, thus compressing the branches near the leaves of the genealogical tree. If migration rates are small enough, it may happen that all sampled lineages coalesce before any lineage has moved to another colony. In this case, the entire tree is compressed. Alternatively, if some lineages do move from the sampled colony, then since the final coalescences (those near the root of the tree) cannot occur unless the lineages involved find themselves in the same colony, the top of the tree tends to be extended in comparison with the panmictic setting. The magnitude of this extension of the genealogical tree near the root depends on the migration rates and pattern and on the number of colonies, but it can be very marked. If the sampled individuals are not all taken from a single colony, then the effect just described, of waiting for lineages to find themselves in the same colony before the final coalescence can occur, is inevitable and leads to a stretching of the genealogical tree near its root.

To illustrate further the effect of geographical structure on genealogy, we consider briefly the coalescence time of pairs of genes, a problem for which exact results are available. It turns out that for models with considerable symmetry in migration rates and patterns, including, for example, the symmetric island and stepping-stone models, the mean coalescence time for a pair of genes taken from the same colony is the same as that for a panmictic population of the same total size (18, 43). This surprising result is somewhat misleading. Recall the intuition that while the two ancestral lineages are in the same colony, they coalesce at a rate C times higher than in the panmictic case. Thus in the structured population, if a coalescence occurs before either lineage leaves the colony from which the individuals were chosen, it is likely to occur more quickly than in the panmictic case. On the other hand, if one of the lineages does leave the original colony before coalescence, the coalescence time in the structured population will be much longer than in the panmictic case. Thus the coalescence time of two genes from the same colony in the structured case is more variable than in the panmictic setting, although the mean of each distribution is the same.

For illustration, consider the circular stepping-stone model that posits a collection of C equal-sized colonies arranged in a circle, with migration possible only between neighboring colonies. Write $M/2$ for the expected number of immigrant genes to (or migrant genes from) each colony per generation, and assume that a migrant chooses its destination uniformly from the two neighboring colonies. With time measured in units of CN generations, the mean and variance of the coalescence time for two genes from the same colony are 1 and $1 + (3M)^{-1}(C - C^{-1})$, respectively. In the panmictic case the mean and variance are both 1. One indication of the lengthening of the genealogical tree for a sample caused by the geographical structure in this model follows from the fact that for C even, the mean coalescence time for a pair of genes from "opposite sides" of the collection of colonies is $1 + C/(4M)$. If the number of colonies is large, and/or M is small, this can be much larger than the value of 2 for the expected coalescence time for the whole population in the panmictic case.

Explicit results for the means, variances, and moment generating functions of pairwise coalescence times are available (18) for a wide range of models, including several that are "asymmetric" in the sense of unequal colony sizes or asymmetric migration patterns. Apart from the inherent interest (see, for example, the discussion below of inference based on pairwise difference measures), these also allow a detailed analysis of Wright's measure, F_{ST}, of subpopulation differentiation (18, 38). It is noteworthy that many of these quantities depend in a sensitive way on exact details of the assumed population structure and migration patterns. As the latter are unlikely to be known in practice (if indeed this general class of model applies) considerable care is needed in interpreting inferences for structured populations, whether or not coalescent methods of analysis are used. For example, inferences and interpretations based on the assumption of a symmetric island model can be quite misleading if the true underlying population structure is different.

GENETICS AND THE COALESCENT

In different contexts it is natural to focus on different sorts of descriptions of genetic data. For example, one may wish to model genes as DNA sequences, as collections of discrete alleles (e.g. RFLPs), as unlabeled but distinguishable alleles (infinitely-many-alleles models), or in terms of the number of copies of a repeat sequence (e.g. VNTR data). Not only do different settings involve different descriptions of genetic type, but they also involve different specifications of the way in which mutation changes these types. These different descriptions do not, however, affect the underlying genealogy.

Extremely complicated mutation mechanisms may be studied via genealogy, but we motivate the method with a simple example. Consider a discrete generation model in which we assume that mutations occur independently to all genes, with the probability that a mutation occurs in a given gene in a given generation being u. In following along a lineage of length r generations, the number of mutations occurring along the lineage has a binomial distribution with parameters r and u. For the panmictic case, suppose that time is scaled in units of N generations, and u is of the order of $1/N$, in that

$$\lim_{N\to\infty} 2Nu = \theta. \qquad 6.$$

It follows from the Poisson approximation to the binomial that as $N \to \infty$ the number of mutations occurring along a lineage of length t in rescaled time has a Poisson distribution with mean $\theta t/2$. Indeed the process of mutations along the lineage is a Poisson process of rate $\theta/2$. This argument can be extended to show that conditional on the coalescent tree, mutations are laid down according to independent Poisson processes of rate $\theta/2$ in each branch. In particular, the total number of mutations occurring in the history of a sample of n genes since their MRCA has a compound Poisson distribution: given the total length $T = \sum_{j=2}^{n} jT_j$of the branches in the tree, it has a Poisson distribution with mean $\theta T/2$.

To simulate the distribution of types in a sample of genes under neutrality, (*a*) simulate the genealogical tree (allowing for population structure and variation in population size); (*b*) assign a type to the common ancestor of the sample (in most applications this type is random, being chosen from the equilibrium distribution of the mutation process); and (*c*) trace down the genealogical tree from the root, recording the effects of the mutations along the branches of the tree. The (random) types at the tips of the tree in this simulation represent a realization of a sample of n genes from the model in question. There are now many examples of this approach in the literature [reviewed in (21, 22)]. The same underlying methodology applies to simulation of nonequilibrium samples.

In the case of constant population sizes, there are very efficient methods for simulating stationary samples using genealogical urn models (5, 31, 48). Let $\mathcal{A}$ denote the set of possible types, and suppose that whenever a mutation occurs to a gene of type i it results in a gene of type j with probability p_{ij}. Note that $p_{ii} > 0$ allows the mutation rate to depend on the type of the gene. Suppose that $(\pi_m, m \in \mathcal{A})$ is the equilibrium distribution of the recurrent mutation process with transition matrix (p_{ij}).

1. Choose a type in $\mathcal{A}$ according to the distribution $(\pi_m, m \in \mathcal{A})$. The urn process begins with two balls in the urn, each of which is assigned this chosen type.
2. If there are currently k balls in the urn, choose one of these uniformly, independently of everything else, and denote its type by i.

   (*a*) With probability $(k - 1)\sigma^2/(\theta + (k - 1)\sigma^2)$, copy the chosen ball and return both to the urn;

   (*b*) With probability $\theta/(\theta + (k - 1)\sigma^2)$, mutate the chosen ball to type $j \in \mathcal{A}$ with probability p_{ij} and return the mutant ball to the urn.

3. If there are fewer than $n + 1$ balls in the urn, return to step 2. Otherwise, discard the last ball added to the urn and stop.

The types associated with the n balls in the urn correspond to a sample of size n from the stationary distribution of the underlying model.

Apart from its use in simulation, the underlying structure of these sampling distributions facilitates valuable qualitative insight. A fundamental observation, some of whose consequences will be seen in the next section, is that the types in a sample are dependent, precisely because of their shared genealogical history. Recall that in samples from constant-sized populations, typical genealogical trees are dominated by the time for which the sample has exactly two ancestors. During this time the types of these two ancestors tend to diverge because of mutation. For such trees, rather less divergence may happen from these two types to the tips, so that in such populations samples with two "clusters" of genetic types (with genes relatively similar within clusters and broadly less similar between clusters) should not be surprising. In particular, such clustering does not necessarily require that additional explanations be invoked.

The effect of continued rapid growth (forward in time) of a population can be to change the genealogical tree to be more star-shaped. This change increases the independence between genetic types in samples and tends to destroy the clustering effect just described. On the other hand, in structured populations the genealogical tree can be substantially stretched near the root and compressed near the tips, and this exacerbates the clustering effect.

One substantive application of the methods in this section arises in attempts to understand the mutation mechanisms at mini- and micro-satellite repeat, or VNTR, loci in human populations. [See (4, 36), and for related work (17, 40), and references therein.]

INFERENCE

Estimating the Mutation Rate

As an illustration of the use of coalescent methods in connection with genetic inference, we focus here on a particular problem that has received extensive attention in the literature: the estimation of the mutation rate θ on the basis of a sample of size n, taken at equilibrium, from the infinitely-many-sites model for DNA sequence data. This model assumes that each gene is a long sequence of completely linked sites. It assumes no back mutation, so that each mutation on the coalescent tree introduces a new segregating site. At any given site the sequences in the sample display one of two possible types: the ancestral type present in the MRCA of the sample, and the mutant type.

We consider the constant population–size case and assume that time has been scaled in units of $\sigma^{-2}N$ generations with the mutation parameter θ now defined as $\lim_{N\to\infty}\sigma^2 u/(2N)$. With this scaling, the parameter σ^2 makes no explicit appearance in what follows. However, if primary interest focuses on estimation of u itself, or in making time estimates in years rather than coalescent time units, then the value of σ^2 becomes crucial. The values of u, N, and σ^2 are confounded: Estimation of one of these requires independent information about the other two. Inferences about coalescence times require independent information on all three.

Define S_{ij} to be number of sites at which sequences i and j differ, for $i \neq j$. The sample homozygosity F_n may be written in the form

$$F_n = \frac{2}{n(n-1)}\sum_{i<j} I(S_{ij}=0), \qquad 7.$$

where $I(S_{ij} = 0) = 1$ if $S_{ij} = 0$, and 0 otherwise. The mean of F_n is $\mathsf{E}(F_n) = \mathsf{P}(S_{12} = 0) = 1/(1 + \theta)$, and this suggests the moment estimator $\theta_F = 1/F_n - 1$. Since the behavior of the allele frequencies in an infinitely-many-sites model is precisely the same as the behavior of the allele frequencies in the infinitely-many-alleles model, we recall Ewens' fundamental observation (6) that this estimator is using precisely that part of the data that is least informative for θ, and so it is likely to have poor statistical properties. Indeed, θ_F is not a consistent estimator because F_n has a nondegenerate limit distribution (41). (Recall that consistency means that as the sample size increases the estimator converges to θ.)

The maximum likelihood estimator θ_E for the infinitely-many-alleles model (6) is the solution of the equation $K_n = \sum_{k=0}^{n-1} \theta/(\theta + k)$, where K_n, the sufficient statistic for θ, is the number of alleles (distinct sequences) observed in the data. The estimator θ_E is consistent, and it has asymptotic variance

$$\mathrm{Var}(\theta_E) \approx \theta \Bigg/ \left(\sum_{k=1}^{n-1} \frac{k}{(\theta+k)^2}\right).$$

The number of segregating sites S_n in the sample equals the total number of mutations that occur on its genealogical tree. Watterson (51) proposed the unbiased estimator

$$\theta_W = S_n \Big/ \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)$$

of θ. The asymptotic variance of θ_W is

$$\operatorname{Var}(\theta_W) = \left(\theta \sum_{k=1}^{n-1} \frac{1}{k} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2} \right) \Big/ \left(\sum_{k=1}^{n-1} \frac{1}{k} \right)^2,$$

and so θ_W is also a consistent estimator of θ.

Tajima (44) suggested an unbiased estimator based on pairwise differences:

$$\theta_T = \frac{2}{n(n-1)} \sum_{i<j} S_{ij}, \qquad 8.$$

and he showed that the variance of θ_T is

$$\operatorname{Var}(\theta_T) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

Notice that $\operatorname{Var}(\theta_T)$ is asymptotic to $\theta(3 + 2\theta)/9 > 0$, so that θ_T is not consistent.

While consistency may not be the only, nor even the paramount, concern, the lack of consistency of some estimators highlights their inefficient use of the data. The explanation is partly genealogical. Both estimators θ_F and θ_T are based on pairwise comparisons within the sample. Exactly because all genes in the sample share the same overall genealogy, differences between pairs of genes are positively correlated, even when the four genes in the two pairs are distinct. This positive correlation of the terms in the sums in Equations 7 and 8 effectively limits the precision of these estimators.

The estimators θ_E and θ_W have variances that decay at a rate proportional to $1/\log n$, rather than the rate of $1/n$ familiar from the setting of estimation for independent and identically distributed random variables. This much slower decay rate is a consequence of the fact that the types of genes in a sample are not independent, because of their shared genealogy. As a consequence, estimates of mutation rates are going to be rather imprecise no matter how large the sample size n is. On the other hand, these two estimators are making rather more efficient use of information than those based on pairwise differences. These kinds of conclusions, depending as they do on genealogical structure, are not specific to the infinitely-many-sites model for mutation.

One common feature of these estimators is that they do not make full use of the information in the data. Some theoretical assessment of this problem is available (8, 10). In principle, it would be preferable to make inferences from the entire data set using a full likelihood-based approach. Even in the infinitely-many-sites model such an approach is difficult. No explicit expressions are available for the likelihood function. Further, effectively because of the enormous dimensionality of the space of possible genealogical trees, conventional simulation-based techniques do not appear helpful. However, using genealogical methods it is possible to derive recursive equations for the value of the likelihood function at particular parameter values (11, 16, 42). In turn, a Monte Carlo simulation method leads to numerical evaluation of the solution of the recursion, and thence to the evaluation, and, for example, the maximization of the likelihood function. [We do not give details here, but refer the reader to (13–16).] Methods of this kind, though extremely computationally intensive, now allow for optimal use of genetic data and appear extremely useful in practice. Further, they may be applied to different models and for different inference problems. For example, the method has been adapted for some variable population–size models (13, 14), to models with back mutation (15), and to models with geographical structure (34). See also (9, 31) for other approaches to such models. For another promising Monte Carlo inference method, see (30).

For comparative purposes, we apply the estimation methods described above to a mitochondrial data set from (50). The data are 360 base pairs of D-loop sequence from an Amerindian population. The sample comprised 55 individuals and had 14 alleles and 18 segregating sites. A detailed description of the sequences appears in (14). The MLE θ_{GT} based on the infinitely-many-sites model was found in (14) using the Monte Carlo approach. The estimate (with standard error in parentheses) was $\theta_{GT} = 4.76$ (1.48). For comparison, we obtained the values $\theta_E = 5.73$ (1.66), $\theta_W = 3.93$ (1.44), and $\theta_T = 3.24$ (1.88). The preceding standard errors for each method are increasing functions of the estimated values of θ. Thus if an estimate happens to be small, the same is true of the associated standard error. It may thus be more helpful to compare standard errors for a common value of θ. With $\theta = 4.76$, the estimated standard errors are 1.73 (E), 1.67 (W), and 2.29 (T), compared to 1.48 for (GT). The precision of the estimators, as measured by these standard errors, coincides with expectations from the theoretical discussion above.

Human Population Genetics

The previous section illustrated genealogical methods within the context of a formal estimation problem. Here we outline their use in two less structured inference problems from human population genetics: inference about the time since mitochondrial Eve and the use of genetic data to make inferences about population history. Our treatment here aims simply to indicate the flavor of certain approaches.

There has been considerable interest in inferences about the time T_{Eve} since "Eve," the MRCA of extant human mtDNA (2). Information about this time has an important bearing on competing theories for early human evolution. [See (49) for a critical review.] In fact, genealogical arguments have not figured greatly in this problem. Most authors regard T_{Eve} as a fixed parameter to be estimated from genetic data. Estimates are usually based on some measure of the "divergence" within the sample of sequences examined, calibrated by mutation rates estimated from comparisons with data from a separate species, assumed to have diverged from humans at a known time.

We have argued elsewhere (14, 33) that such methods are inappropriate in principle. The time T_{Eve} is not a parameter in the usual sense, rather it is the observed value of a random variable about whose distribution we have information before observing data. Inferential statements, whether classical or Bayesian, should relate to the conditional distribution of T_{Eve}, given the data. Its unconditional distribution is related to that of T_{MRCA} for a genealogical tree from a population model in which one hopes to capture the important features of human demography. This approach is implemented, via the Monte Carlo method referred to above, for a plausible but simple population model for a particular Amerindian population, in (14). It is, at least at the present time, impossible to implement in connection with demographic scenarios that aim to capture realistic features of general human demography. On the other hand, it appears difficult (33) even to construct simple neutral models of human populations, incorporating population growth and substructure, for which the unconditional distribution of T_{Eve} is consistent with current estimates of about 150,000 years. Our hope is that future developments in genealogy will allow full use in such problems of both the mitochondrial data and our background knowledge of human demography. In the interim, it should be remembered that most current estimates of T_{Eve} substantially underrepresent the uncertainty involved.

We saw earlier how aspects of population demography, such as population growth, can affect sample genealogies. This, in turn, changes the patterns to be expected in genetic data. It is thus natural to hope that one might learn about the underlying demographic history of a population from genetic data. A particular application concerns attempts to learn about the timing and nature of the growth of early human populations.

One summary of a DNA data set is obtained by counting the number of differences between each pair of sequences in the data and then plotting a histogram of all these pairwise difference counts. Attempts to reconstruct human population history [see (37) and references therein] have been based on such histograms, obtained from mitochondrial DNA data. The conclusion of this work is that the human population experienced an explosion during the late Pleistocene, between about 30,000 and 150,000 years ago. In related work (46, 49), genealogical arguments are used to study historical aspects of geographical population structure.

One problem with mitochondrial data is the uncertainty that follows from the fact that we have a single realization of the evolutionary process: Even if we knew the entire human mitochondrial genealogical tree, this would represent a sample of size one from a distribution (on trees) that would depend in a complicated way on aspects of the population's history and structure. Its use to estimate these underlying processes should then be cautious. In fact we have less information in the data, and we have already seen that statistics based on pairwise measures may not make particularly efficient use of available information.

This type of approach to understanding human (and other) population history appears to have considerable promise. As data from unlinked nuclear loci become available we will, in principle, have independent realizations of genealogy. It is not without difficulties however, and conclusions may depend sensitively on underlying assumptions. For example, in the mitochondrial context, it has been argued (32) that it is difficult to construct any plausible neutral models for human demography for which observed patterns of pairwise differences are likely. This and other arguments (1, 3) suggest possible non-neutrality of the human mitochondrial genome. Not only are models with selection more difficult to analyze, but conclusions from such models may differ substantively from those under neutrality: For example, the effects on pairwise differences of the sweep to fixation of a selectively favored mutant allele in a constant-size population may be similar to those of population explosion under neutrality. Further, current applications to human populations assume that the effects of population growth are as for the deterministic models described around Equations 2 and 3. As we noted there, this is true for some plausible models for human demography. However, it will be false, and the conclusions possibly quite misleading, for other plasuible models of human demography.

DISCUSSION

We have seen that genealogical methods apply readily to neutral models under a range of demographic assumptions. This allows a detailed study of such models. It also obviates the need to study more complicated demographic scenarios via (possibly misleading) effective population sizes. Further, a wide variety of behavior for genealogical trees, and hence patterns in genetic data, is possible under different assumptions about population demography. Many formal hypothesis tests of neutrality actually test a null hypothesis of panmixia, constant-size population, no recombination, and neutrality. Consequently, rejection of "neutrality" in such tests does not require selection to be operating. Similarly, the effects of selection could be confounded with other (demographic) departures from the null hypothesis in such a way as to lead to nonsignificant test statistics.

All coalescent theory relates to properties of a random sample of genes from the population. In practice, genetic data are typically obtained from convenience samples rather than proper random samples. There is an obvious danger that such data may contain individuals who share relatively too much ancestry on the relevant timescales. The extent to which application of coalescent (or traditional) methods to such convenience samples may be misleading remains an open, and potentially serious, question.

Our primary focus has been on samples taken from populations at equilibrium. However, there are many cases where other approximations are more appropriate. One example of great current interest involves linkage disequilibrium mapping, where equilibrium "diffusion time scale" approximations are known to be poor (25).

Coalescent methods provide powerful tools in a range of evolutionary problems. Our purpose here has been to outline some of the central ideas and results in the theory. In highlighting the applicability of the methodology in certain settings, we have also tried to indicate its limitations. Even in those cases in which the methodology is relevant, the interpretation of results from coalescent approaches can depend sensitively on particular assumptions. A general warning, along the lines of caveat emptor, may be appropriate: In practical applications, here as elsewhere, conclusions drawn from coalescent analyses can be misleading if the assumptions that underpin the analysis do not apply.

ACKNOWLEDGMENTS

PD was supported in part by a Block grant from the University of Chicago, and ST by NSF grant DMS90-05833.

Literature Cited

1. Ballard JWO, Kreitman M. 1995. The bell tolls for the neutral evolution of mitochondrial DNA. *Trends Ecol. Evol.* In press
2. Cann R, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36
3. Di Rienzo A, Wilson A. 1991. Branching in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 88:1597–601
4. Di Rienzo A, Petersen AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. 1994. Mutational processes of simple-

sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91: 3166–70

4a. Donnelly P, Tavaré S, eds. 1996.*Progress in Population Genetics and Human Evolution*. New York: Springer Verlag. In press

5. Ethier SN, Griffiths RC. 1991. The neutral two-locus model as a measure-valued diffusion. *Adv. Appl. Prob.* 22: 773–86

6. Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3:87–112

7. Ewens WJ. 1990. Population genetics theory—the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, ed. S Lessard, pp. 177–227. Amsterdam: Kluwer Dordrecht

8. Felsenstein J. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic analysis. *Genet. Res.* 59:139–47

9. Felsenstein J. 1992. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration approach. *Genet. Res.* 60: 209–20

10. Fu Y-X, Li W-H. 1993. Maximum likelihood estimation of population parameters. *Genetics* 134:1261–70

11. Griffiths RC. 1989. Genealogical-tree probabilities in the infinitely-many-sites model. *J. Math. Biol.* 27:667–80

12. Griffiths RC. 1991. The two-locus ancestral graph. In *Selected proceedings of the symposium on applied probability, Sheffield, 1989*, ed. IV Basawa, RL Taylor, 18:100–17. IMS Lecture Notes-Monogr. Ser.

13. Griffiths RC, Tavaré S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. London Ser. B* 344:403–10

14. Griffiths RC, Tavaré S. 1994. Ancestral inference in population genetics. *Stat. Sci.* 9:307–19

15. Griffiths RC, Tavaré S. 1994. Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* 46:131–59

16. Griffiths RC, Tavaré S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* 127:77–98

17. Harding RM, Boyce AJ, Martinson JJ, Flint J, Clegg JB. 1993. A computer simulation study of VNTR population genetics: constrained recombination rules out the infinite alleles model. *Genetics* 135:911–22

18. Herbots HM. 1994. *Stochastic models in population genetics: genealogy and genetic differentiation in structured populations*. PhD thesis. Univ. London, UK. 155 pp.

19. Herbots H. 1995. The structured coalescent. See Ref. 4a, In press

20. Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol.* 23:183–201

21. Hudson RR. 1991. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, ed. D Futuyma, J Antonovics, 7:1–44. Oxford: Oxford Univ. Press

22. Hudson RR. 1992. The how and why of generating gene genealogies. In *Mechanisms of Molecular Evolution*, ed. N Takahata, AG Clark, pp. 23–36. Sunderland: Sinauer

23. Hudson RR, Kaplan N. 1994. Gene trees with background selection. In *Alternatives to the Neutral Model*, ed. GB Golding, pp. 140–153. London/New York: Chapman Hall

24. Kaplan N, Darden T, Hudson RR. 1988. The coalescent process with selection. *Genetics* 120:819–29

25. Kaplan N, Hill WG, Weir BS. 1995. Likelihood methods for locating disease genes in nonequilibrium populations. *Am. J. Hum. Genet.* 56:18–32

26. Kaplan N, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. *Genetics* 123:887–99

27. Kingman JFC. 1982. On the genealogy of large populations. *J. Appl. Prob.* 19A: 27–43

28. Kingman JFC. 1982. The coalescent. *Stoch. Process. Appl.* 13:235–48

29. Kingman JFC. 1982. Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, ed. G Koch, F Spizzichino, pp. 97–112. Amsterdam: North-Holland

30. Kuhner MK, Yamato J, Felsenstein J. 1995. Applications of Metropolis-Hastings genealogy sampling. See Ref. 4a, In press

31. Lundstrom R, Tavaré S, Ward RH. 1992. Estimating mutation rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci. USA* 89:5961–65

32. Marjoram P, Donnelly P. 1994. Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* 136:673–83

33. Marjoram P, Donnelly P. 1995. Human demography and the time since mitochondrial Eve. See Ref. 4a. In press

34. Nath M, Griffiths RC. 1995. Estimation in an island model using simulation. *Theor. Pop. Biol.* In press

35. Notohara M. 1993. The genealogical process of neutral genes with mutation in geographically structured populations. *J. Math. Biol.* 31:123–32

36. Roe A. 1992. *Correlations and interactions in random walks and population genetics*. PhD thesis. Univ. London, UK. 329 pp.

37. Rogers A. 1995. Population structure and modern human origins. See Ref. 4a. In press

38. Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res.* 58: 167–75

39. Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555–62

40. Shriver MD, Jin L, Chakraborty R, Boerwinkle E. 1993. VNTR allele frequency distributions under the stepwise mutation model. *Genetics* 134:983–93

41. Stewart FM. 1976. Variability in the amount of heterozygosity maintained by neutral mutations. *Theor. Pop. Biol.* 9: 188–201

42. Strobeck C. 1983. Estimation of the neutral mutation rate in a finite population from DNA sequence data. *Theor. Pop. Biol.* 24:160–72

43. Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* 117:149–53

44. Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–60

45. Takahata N. 1991. Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* 129:585–95

46. Takahata N. 1993. Allelic genealogy and human evolution. *Mol. Biol. Evol.* 10:2–22

47. Tavaré S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Pop. Biol.* 46:119–64

48. Tavaré S. 1995. Calibrating the clock: using stochastic processes to measure the rate of evolution. In *Calculating the secrets of life*, ed. ES Lander, MS Waterman, pp. 114–52. Washington, DC: Natl. Acad. Press

49. Templeton AR. 1993. The "Eve" hypothesis: a genetic critique and reanalysis. *Am. Anthropol.* 91:51–72

50. Ward RH, Frazier BL, Dew K, Pääbo S. 1991. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA* 88: 8720–24

51. Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7:256–76