A Coalescent Model of Ancestry for a Rare Allele

Jinko Graham*†,1 and Elizabeth A. Thompson*‡,§

*National Institute of Statistical Sciences, Research Triangle Park, North Carolina 27709, †Program in Statistical Genetics, North Carolina State University, Raleigh, North Carolina 27695, ‡Department of Statistics and
§Department of Biostatistics, University of Washington, Seattle, Washington 98195

Manuscript received June 29, 1999
Accepted for publication May 12, 2000

ABSTRACT

In disequilibrium mapping from data on a rare allele, interest may focus on the ancestry of a random sample of current descendants of a mutation. The mutation is assumed to have been introduced into the population as a single copy a known time ago and to have reached a given copy number within the population. Theory has been developed to describe the ancestral distribution under arbitrary patterns of population expansion. Further results permit convenient realization of the ancestry for a random sample of copies of a rare allele within populations of constant size or within populations growing or shrinking at constant exponential rate. In this article, we present an efficient approximate method for realizing coalescence times under more general patterns of population growth. We also apply diagnostics, checking the age of the mutation. In the course of the derivation, some additional insight is gained into the dynamics of the descendants of the mutation.

GENETIC linkage studies based on pedigree data have limited resolution due to relatively small numbers of segregations. One possible strategy for narrowing the candidate region is disequilibrium mapping, first applied successfully to localization of genes for cystic fibrosis (Cox et al. 1989) and Huntington’s disease (Snell et al. 1989; Theilmann et al. 1989). The premise behind disequilibrium mapping is the decay over time of linkage disequilibrium between a disease mutation and marker alleles on the background haplotype. The extent of decay depends on the recombination frequency between the disease and marker loci, and the number of segregations in the ancestry of descendants of the mutation (Arnason et al. 1977; Thompson 1978). The more segregations that relate these descendants, the more fragmentation of the ancestral haplotype on which the disease mutation was introduced, and the finer the scale of mapping (Thompson 1997).

A key concept in disequilibrium mapping is thus the ancestry of a sample of descendants of a disease mutation (e.g., Kaplan et al. 1995). Thompson (1975) used a linear birth-and-death process to model the dynamics of descendants and derived the joint distribution of coalescence times, conditional on the time of the most recent common ancestor. Slatkin (1996) derived means and variances of the total ancestral tree length, pairwise differences, and time of most recent common ancestor of a sample, conditional on the time of the mutation.

Building on this work, Slatkin and Rannala (1997) and Rannala and Slatkin (1998) developed theory for the distribution of coalescence times (see also Griffiths and Tavaré 1998). These results are particularly useful in the special case of a population of constant size or of size growing or shrinking at constant exponential rate, since coalescence times may then be conveniently realized as the order statistics of an independent identically distributed sample. In this article, we describe an efficient approximate method of realizing coalescence times under more general patterns of population growth, implemented in a previous report on disequilibrium mapping (Graham and Thompson 1998). The method assumes a rare disease allele, present as a single copy a known time ago. Diagnostics based on subtree coalescent methods (Griffiths and Tavaré 1998) may be applied to check the assumption of the time of a single copy. In the course of the derivation, some additional insight is gained into the dynamics of the descendants of the mutation. Throughout, we refer to these descendants as the disease subpopulation.

The starting point for our development is a Moran model of gene reproduction in a population of constant size (Moran 1962). At each successive Moran event, one copy of a gene from the population is randomly selected to give birth, and one is randomly selected to die. Thus, sampling processes for birth and for death events are the same. To relate the Moran model to coalescence times, we must choose the number of Moran events that represent a generation. Consider all copies that give birth at a fixed time. A natural candidate for the scaling of a generation is the expected age of such a parent, which is N Moran events in a population of size N copies. Equating N Moran events to a genera-
tion leads to a one-generation inbreeding ratio that is approximately the same as that for a Wright-Fisher population of half the size (Crow and Kimura 1970). The faster rate of inbreeding in the Moran population derives from a larger variance in progeny number than the binomial variance in a Wright-Fisher population of the same size. In fact, extrabinomial variance in progeny is observed in most natural populations, owing to a few copies with relatively large numbers of offspring (Crow and Kimura 1970). However, the scaling of generations is a matter of choice and depends on the application. For example, Graham and Thompson (1998) scaled generations as \( N/2 \) Moran events to eliminate differences due to rates of inbreeding between the proposed disequilibrium-mapping method and other methods based on a Wright-Fisher model of reproduction.

The scaling of a generation determines the rate of events in the analogous continuous-time reproductive model. Thus, if a generation is scaled as \( N/2 \) Moran units in the discrete-time model, the continuous-time analog will have Moran events occurring randomly at rate \( N \) per generation. Consider a population that reproduces according to the continuous-time model. Then a random sample of \( K \) copies from the population coalesces with rate \( K(K - 1)/N^2 \) per Moran unit (Felsenstein 1971) or equivalently with rate \( K(K - 1)/N \) per generation. Coalescent times \( T_i, k = 2 \ldots K \), in generations, during which a sample of size \( K \) from the current population has \( k \) ancestors, are then independent and exponentially distributed with rate \( k(k - 1)/N \) (Kingman 1982). Figure 1 illustrates the notation for \( K = 6 \) sampled alleles.

For a population with known demographic history, size fluctuations may be modeled by allowing additional copies to enter the population or extant copies to leave at the appropriate rate. Migrating copies are assumed to have the same age structure as the population as a whole. The coalescent in a population of changing size may be obtained by generating coalescence times under a constant-sized population and by rescaling to account for size fluctuations, as described by Griffiths and Tavaré (1994). Alternatively, in exponentially growing or shrinking populations, coalescence times may be obtained directly, as described in Slatkin and Hudson (1991). An interesting feature of the coalescent in growing populations is the negative dependence among coalescence times. In growing populations, each coalescence time depends on the sum of the more recent coalescence times, since this sum determines the size of the population at that epoch. The negative correlation among coalescence times increases with the rate of growth, since higher growth rates impose more restrictive constraints on past population size. As the size of the population decreases, so too does the variability in coalescence times.

The remainder of the article is organized as follows. In the next section, we describe a coalescent for a random sample of descendants of a disease mutation, conditional on past disease-allele counts. Just as coalescent rates for a population random sample depend on past copy numbers in the population, coalescent rates for a random sample from the disease subpopulation depend on past disease-allele counts. Following this, we investigate the properties of these counts under a Moran + birth model of population reproduction, assuming a single copy of the disease allele a known time \( T \) ago. We show that prospective simulation of disease-allele counts from a single copy at \( T \) is ineffective for realizing past counts conditioned on the current count because of the high probability of extinction and the variability in current copy number given survival. However, the counts follow very closely a birth-and-death process, provided the disease allele is rare. The birth-and-death process leads to analytic expressions for the conditional moments of past counts. These moments allow approximation, to arbitrary precision, of the conditional distribution. We next discuss efficient realization of the counts from this approximating distribution, since conditioned counts determine coalescence rates. Finally, we examine the assumption of a single copy at \( T \), given the current count, and population demographic information. This is accomplished by studying the age distribution of a selectively neutral mutation, using subtree coalescent methods proposed by Griffiths and Tavaré (1998).

Throughout, we illustrate ideas using the Finnish population, currently of size \( 5 \times 10^6 \) people or \( 10^7 \) haplotypes (Hästbacka et al. 1992). Mutations for several rare recessive diseases are enriched in this population, suggesting the presence of disease-predisposing mutations for each a single or a small number of founding individuals. Ancestors of the Finns are thought to have immigrated to the southwest of the country some 80 generations before present (gpb) or 2000 years ago. We have assumed 1000 founding individuals, or 2000 haplotypes, in our simulations. Subsequent to founding,
Coalescents for a Rare Allele

COALESCENCE RATES

In this section, we adopt the perspective of Thompson and Neel (1997) and consider the coalescent for a random sample of descendants of a disease mutation, conditional on disease-allele counts. These present descendants define the disease subpopulation, which is embedded within the total population. The historical pattern of growth in the total population is assumed known, and the disease mutation is selectively neutral.

Let \( N(t) \) and \( D(t) \) denote the copy numbers in the population and the disease subpopulation at \( t \) gbp, respectively. Suppose \( k(0) = K \) copies of the disease allele are sampled at present, and let \( k(t) \) be the number of ancestors of the sample at \( t \) gbp. In the total population, the instantaneous rate of Moran events at \( t \) gbp is \( N(t) \) per generation. Lineages split via Moran births, and coalescences occur when a splitting lineage and its offspring are among the lineages of the sample from the disease subpopulation. Given \( k(t), D(t), \) and \( N(t) \), Moran coalescences occur with instantaneous rate

\[
\frac{N(t) - 1}{N(t)} \times \frac{D(t)}{N(t)} \times \left( \frac{k(t)}{2} \right) \times N(t) = \frac{k(t)(k(t) - 1)}{N(t)} \times \frac{N(t) - 1}{D(t) - 1}
\]

provided \( N(t) \) is of reasonable size. The first factor on the left-hand side is the chance that a different copy is chosen to die than to give birth, and the second is the chance that a copy giving birth belongs to the disease subpopulation. The third factor is the chance that the reproducing copy and its offspring are among the pairs of lineages from the sample. The product of these three terms is the probability that a Moran event at \( t \) leads to a coalescence in the sample from the disease subpopulation. The fourth term is the Moran event rate at \( t \) gbp.

This overall approach to calculating coalescence rates by successive conditioning was taken by Felsenstein (1971) for a random sample of copies from a population of constant size. The factor \( (N(t) - 1)/D(t) - 1 \) in Equation 1 increases the rate of coalescence relative to a random sample from the total population and may be viewed as an ascertainment correction. The ascertainment correction accounts for sampling of the disease rather than the total population. Also, since \( N(t) \) is typically large, \( (N(t) - 1)/N(t) = 1 \). Thus, except when \( D(t) \) is very small, the rate is close to the rate \( k(t)(k(t) - 1)/D(t) \) that would be obtained if the disease subpopulation reproduced under a continuous-time Moran model, independently of the total population. Early in the history of the disease subpopulation, when \( D(t) \) is small, coalescences occur at a faster rate.

These coalescence rates allow realization of the ancestry of a random sample from the disease subpopulation, conditional on past disease-allele counts \( D(t) \). To obtain realizations of \( D(t) \), more careful modeling of reproduction in the total population is required. In the next section, we investigate the distribution of \( D(t) \) under a Moran model with additional births.

DISEASE-ALLELE COUNTS

Patterns of growth in the overall population are not necessarily reflected in the disease subpopulation. Rapid initial growth of disease-allele counts is expected, conditional on survival of the disease mutation (Thompson and Neel 1997). To capture conveniently such growth patterns in disease-allele counts, we switch from a deterministic to a stochastic model of growth in the total population. We continue to assume a Moran model of reproduction but model size fluctuation through additional births or deaths, introduced via a pure birth or a pure death process.

Growth of human populations since the advent of agriculture could be attributable to increased birth rates, decreased death rates, or a combination. Increased birth rates may more realistically reflect increased family sizes after the transition from a nomadic to an agricultural lifestyle, however (Curtis and Barnes 1989). Introducing extra births into the Moran model decreases the age of a parent and shortens the generation time. By contrast, the magnitude of an age-independent death rate does not affect the stable age distribution of a population because death events affect all age classes equally. Copies are randomly selected from the population to die, without regard to their age. We choose to model growth by increasing the birth rate, since the effect of increased births on generation time will be slight, provided the rate of growth of the population is small. We parameterize additional births by \( \lambda(t) \), the instantaneous rate of population increase at \( t \) gbp.

Populations of shrinking size may be similarly accommodated by introducing additional death events to the Moran model, although these are not discussed here. The rate of additional births (or deaths) does not have to be constant over time. Modeling growth by reducing the death rate rather than increasing the birth rate could be easily adopted instead, if desired. For example, under selective neutrality, the birth and death process derived in Slatkin and Rannala (1997), with birth rate \( 1/2 \) and death rate \( 1/2 - \lambda(t) \), accommodates population expansion by reducing by \( \lambda(t) \) the intrinsic death rate of \( 1/2 \) (scaled to a Wright-Fisher reproductive model).

We consider evolution forward in time from the point \( T \) gbp at which a single copy of the disease allele is assumed. The jump times for disease-allele counts \( D(t) \)
are generated successively from $t = T$ to $t = 0$. For example, labeling jump times $t_k$, $i = 1, 2, \ldots$, in the order they occur forward in time, and starting at $T$ with $D(T) = 1$, $t_1$ is realized. Then, given $t_1$ and $D(t_1) > 0$, $t_2$ is realized, etc. At each $t_i$ the disease-allele count either increases or decreases by 1. This process conditions on $D(T) = 1$, but not on the present count $D(0)$ or survival of the disease allele to present. As before, time is expressed in gbp. Moran and additional birth events impacting $D(t)$ are treated separately; combining them yields the overall process.

We consider first the instantaneous rate of Moran events impacting $D(t)$. Given a Moran event at time $t_i$ and $D(t) = N(t)$, the chance that the disease-allele count changes is

$$2 \frac{D(t)}{N(t)} \left( 1 - \frac{D(t)}{N(t)} \right),$$

with the count being equally probable to increase or decrease by 1. The instantaneous rate of such Moran events is thus

$$r_b(t) = 2D(t) \times \left( 1 - \frac{D(t)}{N(t)} \right),$$

per generation. The instantaneous rate of birth events impacting the size of the disease subpopulation is

$$r_b(t) = \lambda(t)D(t).$$

The rates specify the distribution of times at which the disease-allele count changes. These counts may thus be realized prospectively, conditional on a single copy of the disease allele at $T$; further details are given in Appendix A.

This approach is computationally intensive because of the potentially large number of events between $T$ and the present. More importantly, the scheme is inefficient, because it does not permit conditioning on $D(0)$, the current disease-allele count, except by rejection sampling of realizations. The next section describes a birth-and-death process that approximates reproduction in the disease subpopulation. The birth-and-death approximation allows past disease-allele counts to be generated retrospectively, conditional on $D(0)$. The retrospective scheme saves computational time when realizing $D(t)$ conditioned on $D(0)$ and on $D(T) = 1$.

A BIRTH-AND-DEATH APPROXIMATION

Provided the disease allele is rare in the population, the counts $D(t)$ may be approximated by a birth-and-death process with instantaneous death rate $\mu^* = 1$ and birth rate $\lambda^*(t) = \lambda(t) + 1$. This can be seen by comparing the Moran event rate $r_b(t)$ in (3) to the total event rate $D(t)$ in a birth-and-death model with birth and death rates of 1 per generation. The only difference is the correction factor $1 - D(t)/N(t)$, which is due to Moran dependence among copies of the disease allele imposed by the constraints of total population size. This dependence reduces the rate of events compared to a birth-and-death model, in which alleles reproduce independently. However, for a rare disease allele, $D(t) \ll N(t)$ throughout history, and copies of the disease allele reproduce essentially independently.

The birth-and-death process also approximates the evolution of the disease allele in the Moran + birth model. With a rare disease allele, Moran events almost never involve copies of the disease allele both dying and reproducing. Jump times for disease-allele counts are thus essentially the same as the times of birth or death in the disease subpopulation.

APPENDIX B reviews the distribution of the number of descendants of a mutation under a birth-and-death approximation. Using the descendant distribution, we find that, under constant exponential growth of the Finnish population, a single copy of a disease mutation introduced at founding $T = 80$ generations ago has probability $P(D(0) = 0) = 0.904$ of extinction by the present. This is in excellent agreement with simulations under the Moran + birth model: of 10,000 simulation replicates, 9056 (90.6%) went extinct. Figure 2 shows the conditional distribution of the current disease-allele count given survival to the present, along with the approximating geometric distribution from the birth-and-death process. The simulations show that, conditional on present survival, there is a probability of 0.82 of achieving a current count $D(0)$ of at least $b = 10^4$ in Finns; unconditionally, the probability is only 0.077. Given survival to present, the chance that there are between 8000 and 12,000 current copies of the mutation is about 0.067. Without conditioning on survival of the mutation, the chance is only 0.006.

A key characterization of the birth-and-death approximation is the marginal probability generating function (pgf). In Appendix B, we use the pgf to derive the mo-
ment generating function and hence moments of past disease-allele counts at a time $t$, conditional on the count $D(t)$ at a more recent time $t_b$, and on $D(T) = 1$. The derivation follows Thompson et al. (1992), but is for a birth-and-death process in continuous time rather than for a branching process in discrete time.

Figure 3 shows expected past disease-allele counts in Finns, in the log₁₀ scale, given a present count of $D(0) = b$, for $b = 10,000$, 5,000, and 500, respectively. Expectations are calculated assuming a single copy of the disease allele at founding $T = 80$ gbp and constant exponential growth of the total population. For all three curves, initial expected growth of disease-allele counts is faster than exponential, since surviving disease mutations are those that have, on average and by chance, high initial growth. For a disease with low present copy number such as $b = 500$, slower-than-exponential growth is expected to follow the initial burst.

Conditional standard deviations are shown in Figure 4. Variability is relatively low at first due to small counts but increases over time with the number of copies before decreasing to zero at present, as required by the conditioning. In general, variability is small compared to the mean (e.g., the maximum SD for $b = 10,000$ is $\sim 220$ when expected copy number is $\sim 3,600$) and increases with the current count $b$ because there are more copies of the disease allele.

Conditional skewness coefficients may also be calculated. Figure 5 illustrates one-generation skewness coefficients for $b = 10,000$ in Finns. One-generation skewness is defined as

$$E[(D(t) - \mu(t))^3 | D(t-1), D(T) = 1],$$

with $\mu(t) = E[D(t) | D(t-1), D(T) = 1]$. Coefficients at $t$ gbp are conditioned on counts at $t - 1$ gbp and on $D(T) = 1$. We have set these to be $D(t - 1) = E[D(t-1) | D(0) = 10,000, D(T) = 1]$. The skewness of a gamma distribution with matching first two moments is also shown. The conditional distribution is positively skewed early in the history, when the count is low, since the mutation survives to the present. At more recent times, an increased count leads to more symmetric distributions and less skewness.

**REALIZATION OF COUNTS**

This section describes how past counts $D(t)$ may be realized, conditional on the present count $D(0)$ and on a single allele a known time $T$ ago ($D(T) = 1$). We realize counts backward in time at one-generation intervals. Increments of one generation provide sufficient resolution for the coalescent rates of Equation 1. Thus, for example, in the Finns, we look at times of $t = 1, \ldots, 79, 80$ gbp, realizing backward, successively, from
$D(t)|D(t - 1), D(T) = 1$. Given $D(0)$ and $D(T) = 1$, $D(1)$ may be realized. Similarly, given $D(1)$ and $D(T) = 1$, $D(2)$ may be realized. Continuing this process back in time from 0 to $T$ yields a sample path for $\{D(t)|D(0), D(T) = 1\}$ at one-generation resolution. Conditional moments of past copy numbers are obtained from the joint pgf in (B5), Appendix B. These moments allow construction of the conditional distribution at $t$ gbp, given the count at $t - 1$ gbp. The distribution is constructed to arbitrary precision by matching on an appropriate number of moments.

Figure 6 shows five sample paths for an allele with a current count of $D(0) = 10,000$ copies in the Finnish population. To illustrate ideas, we have assumed constant exponential growth of the population, but more flexible growth patterns may also be accommodated. For example, Graham and Thompson (1998) realized past counts in the Japanese population, which was assumed to have three historical eras with distinct growth rates. Disease-allele counts are generated from a normal-gamma mixture having the same first three moments as the conditional distribution. The normal and gamma components of the mixture are completely specified by their means and variances, both of which have been selected to match the mean and variance of the conditional distribution. The mixing proportion has been selected so that the third moment of the mixture distribution matches the third moment of the conditional distribution. Counts are plotted in the $\log_{10}$ scale and cannot go below 1 because the disease mutation survives to the present. One of the sample paths hits the single-copy boundary several times. The most recent of these times provides a tighter upper bound than $T$ on the time to the most recent common ancestor of the sample. More such hits will tend to occur if the current count is lower than expected given survival to the present.

Hence, to some extent, stochastic modeling of $D(t)$ adjusts for current counts that are more consistent with a younger mutation than $T$. We discuss diagnostics to check adequacy of the assumed $T$ in the next section. Realized disease-allele counts $D(t)$ determine the conditional coalescence rates in equation (1). Coalescence rates, in turn, permit realization of the ancestry, conditional on current disease-allele count and on a single copy of the mutation at $T$.

**AGE OF MUTATIONS**

In this section, subtree coalescent methods (Griffiths and Tavaré 1998) are used to examine the age distribution of a selectively neutral mutation, with current copy number $b = 10,000$. Inferences of the age of the mutation provide insight into the existence of a single copy at the assumed time $T = 80$ gbp.

In a population currently of size $n = N(0)$ copies in which there are $b = D(0)$ copies of the disease mutation, let $J$ denote the number of ancestors of the total population when the disease mutation arose. Consider the coalescent for the total population, and let $T$ be the time during which there were $j$ of its ancestors (see Figure 1). Then the distribution of $J$ given $b$ and $n > b$ is

$$P(J = j) = \begin{cases} \frac{j \rho_n(b) ET_j}{\sum_{k=0}^{n-b} k \rho_n(b) ET_k}, & 2 \leq j \leq n - b + 1, \\ 0, & \text{otherwise}, \end{cases}$$

(4)

where

$$\rho_n(b) = \frac{(n - b - 1)}{j - 2} \frac{n - 1}{j - 1}$$

denotes the probability of $b$ current descendants of a random copy chosen from among $j$ ancestors at some time in the past (Griffiths and Tavaré 1998).

To obtain information about the age of the mutation from these equations, we introduce $n(t)$, the number of ancestors of the $n$ current copies in the total population at time $t$ gbp. The maximum number of ancestors at time $t$ is the total population size $N(t)$ at that time. When $J$, the number of ancestors at the time the mutation arose, is $>n(t)$, the disease mutation must be younger than $t$. However, when the mutation is younger than $t$, $J$ must be $\geq n(t)$. Hence, the probability that the mutation is younger than $t$ is at least $P(J > n(t))$, and, at most, $P(J \geq n(t))$. Solving these equations for $t$ yields bounds for the $\alpha$th quantiles $t_\alpha$ of the age distribution. These bounds are typically quite narrow, provided $n(t_\alpha)$ is not too small. For example, solving $P(J \geq n(t)) = 0.5$ and $P(J > n(t)) = 0.5$ gives a lower and upper bound,
Coalescents for a Rare Allele

381

respectively, on the median age of the mutation. These probabilities can be expanded and evaluated empirically. One possible expansion is

\[ P(J > n(t)) = \sum_{i=1}^{N(t)} P(J > n(t)) = \hat{a} P(n(t) = \hat{a}) \]

Similarly,

\[ P(J \geq n(t)) = \sum_{i=1}^{N(t)} P(J = i) n(t) = \hat{a} P(n(t) = \hat{a}) \]

The events \( n(t) = \hat{a} \) and \( J > \hat{a} \) or \( J > i - 1 \) are independent because the number of ancestors at time \( t \) offers no information about the relative lengths of coalescence times and hence about the number of ancestors \( J \) when the mutation arose. The probabilities \( P(J > \hat{a}) \) can be evaluated via Equation 4, while

\[ P(n(t) = \hat{a}) = \begin{cases} P(S_{i+1} = \hat{a} - 1) P(S_i = \hat{a}, i \geq 2, S_{i+1} = \hat{a}, i = 1, \end{cases} \]

where \( n = N(0) \) is the total number of Finnish copies at present, and \( S_{i} = \Sigma_{t=1}^{T} T \). Although not pursued here, we note that, in addition to the quantiles of the age distribution, it is also possible to estimate the mean age of the statistics of an independent, identically distributed sample. In this continuous-time model, expected growth of surviving mutations is very slow; rates of growth. For more accuracy at small numbers \( j \) of lineages, the following relation

\[ ET_j = E(T|S_{i+1} \geq T) P(S_{i+1} \geq T) + E(T|S_{i+1} < T) P(S_{i+1} < T) \]

was used, where \( E(T|S_{i+1} \geq T) = \frac{2N(T)/(j(j - 1))}{\Sigma_{i=2}^{n} k} \) because the coalescent process has moved past founding. The size of the Finnish population was assumed to be constant at \( N(T) = 2000 \) copies prior to founding at \( T = 80 \) gbp. Under the Moran birth model, the probability that a disease allele with current count \( b = 10,000 \) is younger than \( T = 80 \) generations is \( \approx 0.96 \), and the median age of the mutation is \( \approx 61 \) generations. These calculations show that a single selectively neutral mutation at founding would in fact be expected to have current copy number \( \geq b = 10,000 \), suggesting that the mutation is younger than founding. However, as discussed in the section on realizing disease-allele counts, stochastic modeling of counts provides robustness against a mutation that is younger than the assumed \( T \), as long as the observed count at present is more consistent with the true (younger) age than the assumed age.

DISCUSSION

We have derived coalescence rates for a random sample of descendants of a selectively neutral, nonrecurrent disease mutation, conditional on past disease-allele counts. We discuss the distribution of these counts, assuming a single copy of the disease allele a known time ago. Disease-allele counts, further conditioned on the current count, may be realized by use of rejection sampling, but this is ineffective. We thus describe an efficient approximate approach to their realization, wherein disease-allele counts are simulated retrospectively in single-generation steps on the basis of their conditional moments. These moments can be computed for any assumed demographic history. Hence, coalescence times for a sample from the disease subpopulation can be simulated conveniently under more general demographic conditions than constant exponential growth. Slatkin and Rannala (1997; Rannala and Slatkin 1998) derive the joint density of coalescence times under constant exponential growth. While the derivation may be extended to varying growth rates, it is unclear whether the resulting density would still permit convenient realization of coalescence times as the order statistics of an independent, identically distributed sample. We conclude with an application of diagnostic checks of the assumption of a single copy a known time ago; diagnostics are based on subtree coalescent methods (Griffiths and Tavaré 1998).

In a population reproducing according to the Moran + birth model, copy numbers of a rare disease allele are found to be well approximated by a birth-and-death process. The birth-and-death process decouples Moran birth and Moran death events in the disease subpopulation. Hence, Moran dependence among copies of the disease allele can be ignored. In this continuous-time model, expected growth of surviving mutations is very slow; initial, consistent with the observations of Thompson and Neel (1997) in a discrete-time model. Immediately after the initial burst, disease subpopulations with current copy numbers that are lower than expected, given survival, have slower rates of growth. The more variable the birth-and-death process, the more pronounced the initial burst of growth.

Our approach to realizing coalescence times involves two stages. In the first, we use a continuous-time Moran model to compute conditional coalescence rates given past disease-allele counts \( D(t) \). Then, we describe a birth-and-death process to approximate past counts \( D(t) \) at one-generation intervals, conditional on the current count \( D(0) \) and also on \( D(T) = 1 \), an assumed time \( T \) ago. Since \( D(T) = 1 \), \( T \) can be equal to or more recent...
than $T_1$, the time of origin of the mutation, but must be more ancient than the time of the most recent common ancestor of the disease subpopulation. We may condition the coalescent where $T_1$ is random on an assumed age $t_1$ by using past counts $D(t)$ conditional on $D(0)$ and on $D(t_1) = 1$. SLATKIN and RANNALA (1997) and RANNALA and SLATKIN (1998) use a reconstructed birth and death process (Nee et al. 1994) to derive the distribution of coalescence times directly, by conditioning on $T_1$ and treating it as a known parameter. GRIFFITHS and TAVARE (1998), on the other hand, view $T_1$ as a random quantity and study its distribution given $D(0)$. Randomness in $T_1$ could be incorporated into the present approach by first realizing $T_1$ using the results of GRIFFITHS and TAVARE (1998) and then realizing coalescence times conditional on $D(T_1) = 1$.

This research was supported in part by National Institutes of Health grants DK-42654 (J.G.) and GM-46255 (E.A.T.).

LITERATURE CITED


Communicating editor: S. TAVARE

APPENDIX A:

EVENT TIMES FOR DISEASE-ALLELE COUNTS

Waiting times in a time-dependent Poisson process with rate $r(t)$ may be transformed to waiting times in a regular Poisson process with constant rate $r(0)$. Let $t$ and $t^*$ denote real and transformed time, respectively. Then the appropriate rescaling (Cox and Miller 1977) is given by

$$t^* = \int_0^t \frac{r(s)}{r(0)} \, ds. \quad (A1)$$

As an application, consider the jump times for disease allele counts. Let $T$ denote the random time of the $i$th (Moran or pure birth) event impacting the count, and define $T_0 = T$. Although event times $T_i$ are indexed in the order they occur forward in time, from the introduction of the disease mutation at $T_0 = T$, time is measured in generations before present, as illustrated in Figure A1. Given $T_i = t_i$ and $D(t_i)$, the time $T_{i+1}$ for the next event is

$$t_i = \min(M_{i+1}, B_{i+1}),$$

where $M_{i+1}$ and $B_{i+1}$ are, respectively, the waiting times, in generations, to the next Moran and the next pure-birth event impacting the count. In the interim, for $t \in (T_{i+1}, t_i)$, the count is constant at $D(t_i)$. Given $t_i$ and $D(t_i)$, realization of $B_{i+1}$ using $r_0(t)$ is thus straightforward. Generating $M_{i+1}$ is more complicated, because $N(t)$ and hence $r_0(t)$ in Equation 3 change over this interval.

We generate $M_{i+1}$ by transforming an exponential variate $M^*$ with constant rate $r_0(t)$ given by Equation 3. From Equation (A1),
time \( t \) (gbp) \hspace{2cm} T_0 = T

\[ \begin{array}{c}
T_i \\
M_{i+1} \\
T_{i+1} = T_i - M_{i+1} \\
T_i - B_{i+1}
\end{array} \]

Present \( (t=0) \)

**Figure A1.**—Schematic of the time \( T_{i+1} \) to the next Moran or pure birth event impacting the disease-allele count, given the \( i \)th event occurred at \( T_i \). Events are indexed in the order they occur forward in time, starting from \( T_0 = T \), but time is measured in gbp. Waiting times, \( M_{i+1} \), to the next Moran event and \( B_{i+1} \) to the next birth event impacting the count, are measured in generations.

\[
M^n = \int_{t-M_{i+1}}^{t} \frac{r_2(s)}{N^*(t)} ds.
\]  

\( (A2) \)

Under constant exponential growth of the population at rate \( \lambda \), we may substitute \( N(s) = N(t) \exp(\lambda (t-s)) \) for \( s < t \) into the rates from (3), and rewrite Equation A1 as

\[
M^n = \frac{2D(t)}{r_2(t)} \int_{t-M_{i+1}}^{t} \left[ 1 - \frac{D(t)}{N(t) \exp(\lambda (t-s))} \right] ds = \frac{2D(t)}{r_2(t)} \left[ M_{i+1} - \frac{D(t)}{N(t) \lambda} \left( 1 - e^{-\lambda M_{i+1}} \right) \right].
\]

This implies

\[
[N(t) - D(t)] M^n + \frac{D(t)}{\lambda} = \frac{D(t)}{\lambda} \exp(-\lambda M_{i+1}) + M_{i+1} N(t).
\]  

\( (A3) \)

Solving for \( M_{i+1} \) yields the waiting time to the next Moran event impacting the disease-allele count.

**APPENDIX B:**

**ATTRIBUTES OF THE BIRTH-AND-DEATH PROCESS**

Consider the generalized birth-and-death process that approximates growth of a disease subpopulation described earlier. Let \( s \) be a time increment in generations. Then an allelic copy at \( t \) gbp has \( Y(t) \) descendants \( s \) generations later at \( t-s \) gbp with zero-modified geometric distribution (Kendall 1948), such that

\[
P(Y(s) = 0) = \xi(s) = 1 - \frac{\exp(-\rho(s))}{W(s)},
\]  

\( (B1) \)

and

\[
P(Y(s) = n) = [1 - \xi(s)][1 - \eta(s)]\eta(s)^{n-1}, \quad n \geq 1,
\]  

\( (B2) \)

where

\[
\rho(s) = \int_{0}^{s} [\mu^*(t) - \lambda^*(t)] dt, \quad \eta(s) = 1 - \frac{1}{W(s)},
\]

\[
W(s) = \exp(-\rho(s))\left[ 1 + \int_{0}^{s} \exp(\rho(t))\mu^*(t) dt \right].
\]

Since \( Y_0(s) = D(T-s) \), this also gives the unconditional distribution of the size of the disease subpopulation at \( T-s \) gbp.

Under constant exponential growth of the total population at rate \( \lambda \), \( \lambda^*(t) = \lambda^* = \lambda + 1 \), and the descendant distribution reduces to

\[
P(Y(s) = 0) = \mu^*B(s),
\]

\[
P(Y(s) = n) = [1 - \mu^*B(s)][1 - \mu^*B(s)][\lambda^*B(s)]^{n-1}
\]

(Feller 1968), where

\[
B(s) = \frac{1 - \exp(\lambda^*(\mu^* - r_s))}{\mu^* - \lambda^* \exp(\lambda^*(\mu^* - r_s)).}
\]

For example, assuming constant exponential growth of the Finnish population, a disease mutation introduced as a single copy at founding 80 generations ago has probability \( P(Y_0(80) = 0) = P(D(0) = 0) = 0.905 \) of extinction by the present.

A key characterization of the birth-and-death approximation is the marginal pgf. For the process starting at \( t_0 = T \) gbp and running \( s \) generations to \( T-s \) gbp, this is

\[
g_\phi(x, s) = \frac{\xi(s) + [1 - \xi(s) - \eta(s)]x}{1 - \eta(s)x}.
\]  

\( (B3) \)

(Kendall 1948). In general, the marginal pgf for the process starting at \( t \) gbp and running \( s \) generations to \( t-s \) gbp is

\[
g_\phi(x, s) = \frac{\xi(s) + [1 - \xi(s) - \eta(s)]x}{1 - \eta(s)x}.
\]  

\( (B4) \)

When the growth rate is constant, \( g_\phi(x, s) \) simplifies to

\[
\mu^*(1-x) - (\mu^* - \lambda^* x) e^{-(\lambda^* - \mu^*)s},
\]

\[
\lambda^*(1-x) - (\mu^* - \lambda^* x) e^{-(\lambda^* - \mu^*)s}.
\]

\[
= \frac{[xA_1(s) + A_1(s)]}{[xA_2(s) + A_2(s)]}.
\]

\( (Cox and Miller 1977) \), where \( A_1(s) = \exp(-\lambda^* s) \), \( A_2(s) = \lambda^* A_1(s) - \mu^* \), \( A_3(s) = \mu^* - \mu^* A_1(s) \), \( A_4(s) = \lambda^* A_3(s) - \lambda^* \), and \( A_5(s) = \lambda^* - \mu^* A_1(s) \).

Following THOMPSON et al. (1992), we now develop
expressions for the moments for past numbers of a rare disease allele at \( t_i \), conditional on the number \( D(t_j) \) at more recent \( t_j \), and on \( D(T) = 1 \).

The joint pgf for \( D(t_i) \) and \( D(t_j) \), \( t_i > t_j \), is

\[
g(z, w, t_i, t_j) = E(z^{D(t_i)} w^{D(t_j)})
= \sum_{r} \sum_{j} z^r w^j P(D(t_i) = r, D(t_j) = j).
\]

This may be written

\[
g(z, w, t_i, t_j) = g_0(g_1(w, \delta) z, t_i), \tag{B5}
\]
in terms of the time difference \( \delta = t_i - t_j \), and the marginal pgfs \( g_0 \) and \( g_1 \) in (B3) and (B4), respectively.

For a general random variable \( X \), the \( n \)th factorial moment is \( E \prod_{i=1}^{n}(X - i) \). We obtain the \( n \)th factorial moment for \( D(t_i) \) given \( D(t_j) = j \) by dividing the coefficient \( C_j \) of \( w^j \) in the series expansion of

\[
\frac{\delta^n}{\delta z^n} g(z, w, t_j, t_i) \bigg|_{z=1}
\]

by \( P(D(t_j) = j) \) or \( P(Y_{\hat{0}}(\delta) = j) \) in Equation B2. For example, under constant exponential growth, we obtain

\[
E[D(t_j)|D(t_i) = j, D(T) = 1] = c_1(t_i, t_j) \times j + c_2(t_i, t_j),
\]
where

\[
c_1(t_i, t_j) = [A_1(t_i - t_j)(l \times \lambda(t_i - t_j)) + A_1(\delta)A_1(T - t_j)]
- A_1(\delta)E[\delta(l \times \lambda(T - t_j))]/[A_1(\delta)(l \times \lambda)^{-1}];
\]

\[
c_2(t_i, t_j) = [A_1(\delta)A_1(\delta)A_1(T - t_j) - A_1(\delta)A_1(T - t_j)]/[A_1(\delta)(l \times \lambda)^{-1}].
\]

The coefficient \( c_1(\cdot, t_j) \) increases monotonically from 0 at \( t_i = T \) to 1 at \( t_i = t_j \); the coefficient \( c_2(\cdot, t_j) \) is quadratic with a maximum at \( t_i = t_j/2 \) and with boundary values \( c_2(T, t_j) = 1 \) and \( c_2(t_i, t_j) = 0 \). Although not reported, similar expressions may be derived for higher moments or for moments under changing patterns of growth. When growth rates change over time, these moments indicate that for \( D(t_j) \) moderately large (e.g., >500), and for \( t_i \) and \( t_j \), close (e.g., \( \delta = 1 \) generation apart), local rates apply, and event rates earlier in history have little influence.