Theoretical Population Biology **59**, 287–296 (2001)
doi:10.1006/tpbi.2001.1523, available online at http://www.idealibrary.com on IDEAL®

Rare Alleles and Selection

Carsten Wiuf

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England

Received August 23, 2000

A subpopulation $\mathscr{D}$ of rare alleles is considered. The subpopulation is part of a large population that evolves according to a Moran model with selection and growth. Conditional on the current frequency, q, of the rare allele, an approximation to the distribution of the genealogy of $\mathscr{D}$ is derived. In particular, the density of the age, T_1, of the rare allele is approximated. It is shown that time naturally is measured in units of $qN(0)$ generations, where $N(0)$ is the present day population size, and that the distribution of the genealogy of $\mathscr{D}$ depends on the compound parameters $\rho = rqN(0)$ and $\sigma = sqN(0)$ only. Here, s is the fitness per generation of heterozygote carriers of the rare allele and r is the growth rate per generation of the population. Amongst more, it is shown that for constant population size ($\rho = 0$) the distribution of $\mathscr{D}$ depends on σ only through the absolute value $|\sigma|$, not the direction of selection.

Key Words: **birth–death process; exponential growth; genealogy; rare allele; selection.**

INTRODUCTION

In the present paper, the genealogy of a subpopulation $\mathscr{D}$ of a rare allele in a large population is studied. A two-allele Moran model that allows for selection and exponential growth of the population is developed and, based on this model, an approximation to the distribution of the genealogy of $\mathscr{D}$ is derived. It is assumed that the rare allelic class is the result of a single mutational event in the entire population's history.

Recently, several authors have drawn attention to this problem: Slatkin and Rannala (1997), Thompson and Neel (1997), Rannala (1997). In these approaches the age, T_1, of the rare variant is treated as a (nonstochastic) parameter and, accordingly, the distribution of the genealogy of a sample of rare variants is given in terms of this parameter T_1. Wiuf and Donnelly (1999) argued that the correct interpretation of T_1 is to consider it a stochastic variable and not a parameter. The mutation having given rise to the rare variant is more likely to have occurred in genealogical trees with a long branch between the most recent common ancestor (MRCA) of $\mathscr{D}$, and the ancestry of the rest of the population. As a consequence, conditioning on the mutation having occurred has the effect of stochastically increasing the length of this branch. This effect turned out to be important and has not been captured in the previously published approaches.

The age, considered as a stochastic variable, of a *neutral* variant found in frequency q, $0 < q < 1$, in the population has been studied among others by Kimura and Ohta (1973), Maruyama (1974a), Griffiths and Tavaré (1998), Wiuf and Donnelly (1999), and Stephens (2000). Based on an exact coalescent analysis presented in Wiuf and Donnelly (1999), Wiuf (2000) developed approximations, conditional on the frequency q, to the distribution of the genealogy of a neutral rare variant (say, $q < 5$–10%) and to the distribution of the time the rare variant arose.

This analysis is here extended to cover the case where the rare allelic class is evolving under selective pressure. Time is measured backwards starting at the present time, $t = 0$, and the population is assumed to be growing exponentially at a constant rate, r, per generation. T_1 generations back in time a mutation gave rise to a new type of allele. At present time the variant allele is found in low frequency q. Further back in time than T_1 the

0040-5809/01 $35.00

population consisted of one allelic type only, called the *normal* allele, and the new type, called the *variant* allele, has (heterozygote) fitness s compared to the normal homozygote. As q becomes small, homozygote carriers of the rare allele can effectively be ignored; even in cases where the homozygote is deleterious or has a fitness much different from the heterozygote. To ensure that no further mutations occur as the population evolves up to present time, the mutation rate, u, per generation between the two alleles is set to ≈ 0. The situation is depicted in Fig. 1.

In this paper, approximate distributions of T_0, T_1, and the genealogy of $\mathscr{D}$, conditional on q, are derived. This is done in two steps: First, the limit distributions are found for $N(0) \to \infty$ and $N(0)\, u \to 0$, where $N(0)$ is the current population size. The number, $k = qN(0)$, of variant alleles is held fixed. In the second step, this limit is considered for large k. It is shown that time naturally is measured in units of $qN(0)$ generations and that the approximate distributions depend on the compound parameters $\sigma = sqN(0)$ and $\rho = rqN(0)$ only. This is similar to the standard coalescent process where the scaling is in units of $N(0)$ generations. Maruyama (1974b) found, using a diffusion approach, the expected age of a rare variant under genic selection in a constant population, and it is shown that his result agrees with what is found in this paper. Also the results are consistent with Wiuf (2000). As a second application of the approximations we find the extinction probability of the rare allele given its current frequency.

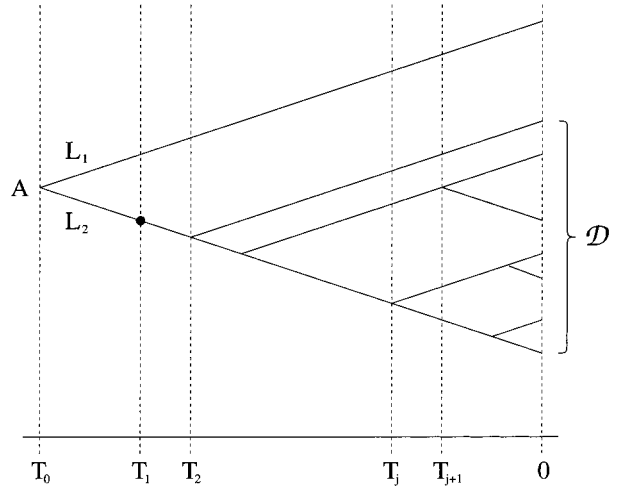

FIG. 1. The genealogy of the rare variant. At time T_0 in the past a normal allele, A, give, birth to a lineage (L_2) in which the mutation, having given rise to the variant allele, arises. The ancestor A is the first MRCA of $\mathscr{D}$ and the class of present-day normal alleles. T_1 is the time the mutation arises and T_j, $j \geqslant 2$, denotes the time while there are at least j ancestors of $\mathscr{D}$. In the example, T_j and T_{j+1} are shown for $j = 3$ and the size of $\mathscr{D}$ is $k = 7$. All other lineages, than those shown, born in L_2 die out before present time.

An approach using branching processes (in the present paper a birth–death process approximation to the Moran model) is not new; both Maruyama (1974b) and Slatkin and Rannala (1997) developed approximations based on birth–death processes, whereas Thompson (1976) and Rannala (1997) adopt models based on related processes. The current approach differs from the approaches by Slatkin and Rannala (1997), Thompson (1976), and Rannala (1997) in that it allows for T_1 to be stochastic, and in that effects of the demography of the entire population are taken into account. It differs from Maruyama (1974b) in that explicit results are found for the density of the genealogy of $\mathscr{D}$ and T_1; Maruyama (1974b) derived the expectation of T_1 in a constant population only.

THEORY

Our starting point is a two-allele diploid Moran model with selection and exponential growth of the entire population. This model is introduced and discussed in Appendix 1. It comprises four parameters: the fitness of heterozygotes, s, the fitness of homozygotes, s', the growth rate, r, and the mutation rate, u. Denote the normal allele by A_2 and the variant allele by A_1. All time variables in the Moran model have superscript *, e.g., T_0^*, to distinguish them from variables measured in real non-overlapping generations. Let T_0^* and T_1^* be defined similar to T_0 and T_1 in the previous paragraph, and let T_j^*, $2 \leqslant j \leqslant k$, denote the time while there are at least j ancestors of $\mathscr{D}$, with $|\mathscr{D}| = k$ (Fig. 1). Further, let $M(\mathscr{D})$ denote the event that the mutation giving rise to the A_1 allele has only happened once in the history of the entire population, that no back mutations have occurred, and that the present number of the A_1 allele is k. Note that all the T_j^*'s are uniquely defined on the event $M(\mathscr{D})$.

We seek the probability distribution of the genealogy of $\mathscr{D}$, conditional on $M(\mathscr{D})$, i.e., the probability

$$P(T_j^* = \tau_j, 0 \leqslant j \leqslant k \mid M(\mathscr{D})) = \frac{P(T_j^* = \tau_j, 0 \leqslant j \leqslant k, M(\mathscr{D}))}{P(M(\mathscr{D}))}, \qquad (1)$$

for $0 \leqslant \tau_k \leqslant \tau_{k-1} \leqslant \cdots \leqslant \tau_1 \leqslant \tau_0$. For any specific choice of the present-day population size, $N(0)$, the number of A_1 alleles, k, and the parameters s, s', r, and u, Eq. (1) can be computed using the Markov property of the Moran model. The denominator is found summing the

numerator over all possible τ_j, $0 \leqslant j \leqslant k$. The numerator is more involved, but can in principle be calculated splitting it in a sum over all possible population histories that coincide with $T_j^* = \tau_j$, $0 \leqslant j \leqslant k$, and $M(\mathscr{D})$. If the population size is constant over time ($r = 0$), the distribution of allele frequencies at time τ in the past is the stationary distribution of the process. If the population is growing in size ($r > 0$), we simply start the process from the stationary distribution of a population of size 1 and mutation rate u.

However, in practice, such a summation is not feasible and we must resort to approximations. Let,

$$t(v) = \sum_{\iota=0}^{\lfloor N(0)\, v \rfloor} \frac{2}{N(\iota)}, \tag{2}$$

where $N(\tau)$ denotes the population size at generation τ and $\lfloor x \rfloor$ denotes the integer part x. Equation (2) defines the way time is transformed from overlapping generations in the Moran model to nonoverlapping real generations (see Appendix 1). Further, define U_j^*, $0 \leqslant j \leqslant k$, by $T_j^* = N(0)\, U_j^*$. The variable U_j^* measures time in the Moran model in units of $N(0)$ overlapping generations, and the transformed variables $t(U_j^*)$ measures time in real generations.

THEOREM 1. *Assume k is fixed. The process $(t(U_0^*), t(U_1^*), ..., t(U_k^*))$ converges in distribution for $N(0) \to \infty$ and $N(0)\, u \to 0$ to a continuous time Markov chain $(T_0, T_1, ..., T_k)$, such that*

$$0 < P(T_0 = T_1 \mid M(\mathscr{D})) < 1. \tag{3}$$

Proof. See Appendix 2.

The parent of the mutant offspring might survive till present time, therefore, $P(T_0 = T_1 \mid M(\mathscr{D})) > 0$. However, as k increases the probability of $T_0 = T_1$ vanishes. Analytical expressions for (3) and the transition probabilities are given in Appendix 1; they are in general intractable and do not seem to provide further insight into the process. Further, the distribution of $(T_0, T_1, ..., T_k)$ depends on r and s, but not the fitness, s', of homozygotes.

If X is a stochastic variable, let $X \sim f(x)$ denote that X has density $f(x)$. Define $q(x; \beta)$, $\beta \geqslant 0$, by

$$q(x; \beta) = \frac{2\beta}{e^{\beta x} - 1}, \tag{4}$$

if $\beta > 0$, and

$$q(x; 0) = \frac{2}{x}. \tag{5}$$

We have $q(x; 0) = \lim_{\beta \to 0} q(x; \beta)$.

THEOREM 2. *The process $(T_0/k, T_1/k, ..., T_k/k)$ converges in distribution to a continuous time Markov chain V_j, $j \geqslant 0$, for $k \to \infty$, such that $rk \to \rho$ and $sk \to \sigma$. In fact,*

$$V_1 \sim C_1 q(v; |\alpha|)^2 \times \exp\{ -q(v; |\alpha|) + (|\alpha| - \rho)\, v \}, \quad v > 0, \tag{6}$$

$$V_{j+1} \mid V_j = u \sim \tfrac{1}{2} q(v; |\alpha|)^2 \times \exp\{ -[q(v; |\alpha|) - q(u; |\alpha|)] + |\alpha|\, v \}, \quad u > v, \tag{7}$$

and

$$V_0 \mid V_1 = u \sim \frac{e^{\rho v} q(v; \rho)^3}{q(u; \rho)^2}, \qquad v > u. \tag{8}$$

The variables V_j, $j \geqslant 1$, can be represented in the form

$$V_j = \frac{1}{|\alpha|} \log\left(1 + \frac{2\,|\alpha|}{X_1 + \cdots + X_j}\right), \tag{9}$$

such that X_j, $j \geqslant 1$, forms a series of independent variables,

$$X_1 \sim C_2 \frac{x^{\rho/|\alpha|} e^{-x}}{(2\,|\alpha| + x)^{\rho/|\alpha|}}, \tag{10}$$

and for $j \geqslant 2$,

$$X_j \sim Exp(1). \tag{11}$$

Here, $\alpha = \sigma + \rho$, C_1 and C_2 are normalizing constants depending on $|\alpha|$ and ρ only, and $Exp(\beta)$ denotes an exponential variable with rate β.

Proof. See Appendix 2.

As a consequence, the natural scaling of time is in units of $k = qN(0)$ generations. Further, the genealogy of $\mathscr{D}$ can be simulated starting with T_1 and building the

genealogy up going towards the present time. From Theorem 1 and 2 the conditional distribution of the genealogy of $\mathscr{D}$ given $T_1(V_1)$ could be derived. In general, this distribution is not expected to agree with the similar results in Thompson (1976), Nee *et al.* (1994), and Slatkin and Rannala (1997) because the mutation process having given rise to the variant allele is not modelled in any of these papers.

Below, comments to special cases of Theorem 2 are listed (1–5). In 6, the genealogy of a subsample $\mathscr{D}_0$ drawn randomly from $\mathscr{D}$ is discussed. Data about all carriers of a variant allele ($\mathscr{D}$) will rarely be available, one can only hope to obtain a finite sample, $\mathscr{D}_0$, of these. Finally in 7, the extinction probability of $\mathscr{D}$ is found.

1. *No selection, and no expansion.* The results in this case are in agreement with results in Wiuf (2000). From Theorem 2 the general form of the density of V_j, $j \geqslant 1$, is found,

$$V_j \sim \frac{2^j}{(j-1)!\, v^{j+1}}\, e^{-2/v}, \qquad v > 0, \tag{12}$$

and, further, $X_j \sim Exp(1)$ for all $j \geqslant 1$. The series relates to the number of ancestors, A_j, of the entire population the first time there are j ancestors of $\mathscr{D}$; $A_{j-1} \approx (X_1 + \cdots + X_j)/q$ (Wiuf, 2000).

2. *No selection, but expansion.* Also the results found in this case are in agreement with the results in Wiuf (2000). There is no simple expression for the densities of V_j, $j \geqslant 1$. In special cases they can be computed, e.g.,

$$V_2 \sim C_3 \left(\frac{1}{e^{\rho v} - 1} - \log \frac{e^{\rho v}}{e^{\rho v} - 1} \right) \times \frac{e^{\rho v}}{(e^{\rho v} - 1)^2} \exp\left\{ -\frac{2\rho}{e^{\rho v} - 1} \right\}, \qquad v > 0 \tag{13}$$

(the density of V_1 is given in Theorem 2). Here C_3 is a normalizing constant depending on ρ only. The variable X_1 can be simulated using an acceptance–rejection scheme (Bratley *et al.*, 1983); for example let the proposal Y be gamma distributed with parameters 2 and 1, $Y \sim \Gamma(2, 1)$, and let the acceptance probability be $1/(1 + Y/2\rho)$. If ρ is small the proposal might be taken to be $Y \sim Exp(1)$ and the acceptance probability $Y/(2\rho + Y)$. As ρ approaches infinity X_1 becomes gamma distributed, $X_1 \sim \Gamma(2, 1)$. Also in this case, Wiuf (2000) found that $A_{j-1} \approx (X_1 + \cdots + X_j)/q$.

3. *Selection, and no expansion.* The density of V_j, $j \geqslant 1$ is given by

$$V_j \sim \frac{2^j\, |\sigma|^{j+1}\, e^{|\sigma| v}}{(j-1)!\, (e^{|\sigma| v} - 1)^{j+1}} \times \exp\left\{ -\frac{2\,|\sigma|}{e^{|\sigma| v} - 1} \right\}, \qquad v > 0, \tag{14}$$

and, further, $X_j \sim Exp(1)$ for all $j \geqslant 1$.

It is of particular interest that the distributions in (14) depend on σ through $|\sigma|$ only; that is, one cannot tell from the structure of the genealogy alone whether the variant allele is under positive or negative selection. The expectation of V_1 is given by

$$E(V_1) = \frac{1}{|\sigma|} \{ \gamma + \log(2\,|\sigma|) \} - \frac{e^{2\,|\sigma|}}{|\sigma|}\, Ei(-2\,|\sigma|), \tag{15}$$

where $\gamma \approx 0.58$ is Euler's constant and $-Ei(-z) = \int_z^\infty e^{-t}/t\, dt$ the exponential integral. Equation (15) agrees with Maruyama (1974b).

4. *Selection, and expansion.* The variable X_1 can be simulated using an acceptance–rejection scheme; e.g., the proposal Y might be taken to be gamma distributed with parameters $\rho/|\alpha| + 1$ and 1, $Y \sim \Gamma(\rho/|\alpha| + 1, 1)$, and the acceptance probability $1/(1 + Y/2\,|\alpha|)^{\rho/|\alpha|}$. In contrast to case 3, we find that σ and $-\sigma$ (with ρ fixed) give rise to different distributions of the genealogy of $\mathscr{D}$. The distribution with growth ρ and selection σ is identical to that with growth ρ and selection $-2\rho - \sigma$.

5. *Expansion, and $\rho = -\sigma$.* This case is not realistic biologically, but is here considered for the sake of completeness. The variable X_1 follows a generalized inverse Gaussian distribution, $X_1 \sim GIG(1, 4\rho, 2)$ (in the notation of Seshadri, 1993, p. 27)

$$X_1 \sim C_4 e^{-2\rho/x - x}, \tag{16}$$

where C_4 is a normalizing constant. It can be simulated using an acceptance–rejection scheme with acceptance probability $\exp(-2\rho/Y)$ and proposal $Y \sim Exp(1)$. If ρ is large, algorithms to simulate a *GIG* variable can be used (Bratley *et al.*, 1983).

6. *Samples from $\mathscr{D}$.* Genealogies of a sample $\mathscr{D}_0$ of size n taken from the $\mathscr{D}$ can be simulated using the sample scheme proposed in Wiuf (2000). First, the numbers, $J_k, J_{k-1}, ..., J_2$, of ancestors of $\mathscr{D}$ at the times of coalescence events in the genealogy of $\mathscr{D}_0$ are simulated according to results in Saunders *et al.* (1984), see also Wiuf (2000). The distribution of $(J_k, J_{k-1}, ..., J_2)$ does not depend on the demography or the selection coefficient and applies to general binary trees (see Griffiths and Tavaré, 1998), in particular to the genealogy of $\mathscr{D}$.

Second, the times, $T_{J_k}, T_{J_{k-1}}, ..., T_{J_2}$, until there are $k-1$, $k-2, ..., 1$ ancestors of $\mathscr{D}_0$, respectively, are simulated according to the formulas given in this paper.

The approach to sampling taken by Nee *et al.* (1994) is somewhat different, though it also relies on a birth–death process approximation to the genealogy of the entire population (a metapopulation of species). Their results are not a priori expected to agree with those presented here, where a mutation process causes a variant class to exist.

7. *Extinction probabilities.* Let E denote the time of extinction in units of k generations. The probability that $\mathscr{D}$ is extinct at time v in the future is

$$P(E \leqslant v) = \exp\left\{-\frac{2\alpha e^{\alpha v}}{e^{\alpha v}-1}\right\}, \qquad (17)$$

if $\alpha \neq 0$, and

$$P(E \leqslant v) = \exp(-2/v), \qquad (18)$$

if $\alpha = 0$ (Appendix 2). In particular, $P(E=\infty)=0$, if $\alpha \leqslant 0$, and $P(E=\infty) = 1-\exp(-2\alpha)$, otherwise. Thus, E is possibly infinite. If $E=\infty$, $\mathscr{D}$ might either go to fixation or become extinct. In the latter case, it persists in the population for a long time, longer than is measurable on a time scale in units of k generations. For example, if the rare allele is selectively neutral, the chance, p_{ext}, of eventual extinction of the rare allele is $p_{ext} = 1-q = 1-k/N(0)$ (Ewens, 1979, this also holds under expansion). It goes to fixation with probability $p_{fix} = 1-p_{ext}$. Thus, for example, a neutral rare allele in an expanding population might persist (with probability $P(E=\infty) = 1-\exp(-2\rho)$) in the population for a significant amount of time before extinction.

Conditional on $E<\infty$, the density of E agrees with that of V_1. For $\rho=0$, this result is known for a number of related processes (e.g., Ewens, 1979), whereas for $\rho>0$, this appears to be new.

DISCUSSION

Expressions for the densities of the coalescence times in the genealogy of $\mathscr{D}$ and the age of the rare variant have been derived. Wiuf (2000) showed that the approximations in the neutral case are good provided $q<5$–$10\,\%$. He compared his results with the exact coalescent analysis in Wiuf and Donnelly (1999). In the general case, where the heterozygotes have a small selective (dis) advantage, a similar degree of accuracy is expected.

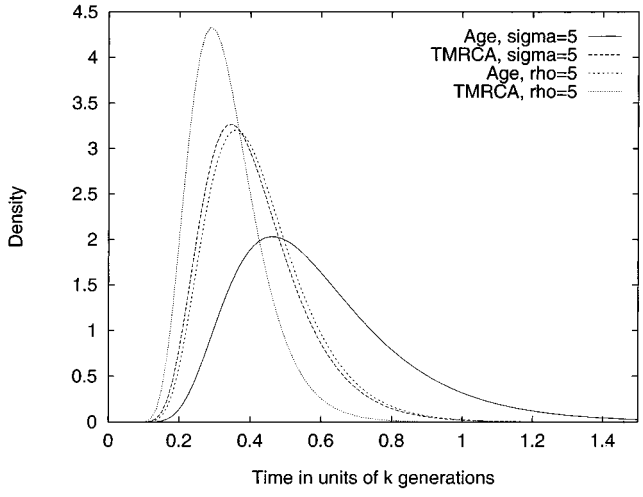

FIG. 2. Comparing selection and expansion. Shown in the figure are the densities (in units of k generations) of the age, V_1, of the rare allele and the time, V_2, of the MRCA of $\mathscr{D}$, respectively, in two different scenarios. In the first, the population size is constant, but the rare allele has a selective advantage, $\sigma=5$, and in the second, the population is growing exponentially with $\rho=5$, but the rare allele is selectively neutral. Both V_1 and V_2 are shorter under growth than under selection.

M. Stephens (personal communication) has kindly provided simulation results that support this; based on these simulations it is suggested that the results in this paper are accurate for $q<5\,\%$ and moderate values of σ. He used a MCMC method to simulate genealogies of large samples in which the rare allele is found in frequency q.

Effects of positive selection and effects of an expanding population are often equated; e.g., Slatkin and Rannala (1997). The present study shows that compared to a neutral population of constant size both selection and expansion have the effect of shortening the branches in the genealogy of $\mathscr{D}$ (comparing times in units of $k=qN(0)$ generations). But the shortening of branches happens differently in the two cases. This is illustrated in Fig. 2; if $\rho=|\sigma|$, the MRCA of $\mathscr{D}$ is younger under expansion than under selection. Apparently, one reason for this is that under expansion the entire population decreases in size (going backwards in time) and coalescence events are thus forced to happen at a higher rate than in a population of constant size. Denote by $P_{\sigma,\rho}$ the probability measure corresponding to a model with selection σ and expansion ρ. The observation in Fig. 2 can mathematically be put in the following way (Theorem 2). Let $j \geqslant 1$ and $v>0$. Then, if $c>0$,

$$P_{0,c}(V_j > v) < P_{c,0}(V_j > v) < P_{0,0}(V_j > v). \qquad (19)$$

As a consequence, V_j under $P_{c,0}$ is stochastically greater than V_j under $P_{0,c}$.

APPENDIX 1

1. *Time.* First a digress on time. In a Moran model of constant population size, N, $N/2$ (overlapping) generations correspond roughly to one (nonoverlapping) generation in a Wright–Fisher (or similar) model. This is justified comparing the mean and variance of offspring number per individual per $N/2$ generations in a Moran model with the same quantities per generation in a Wright–Fisher model. Asymptotically for large N these are all one. We call generations in a Moran model for *Moran* generations and nonoverlapping generations for *real* generations. That is, one Moran generation is about $2/N$ real generations. If the population size varies with time we find that locally the τth Moran generation corresponds to $2/N(\tau)$ real generations, where $N(\tau)$ denotes the total population size at time τ in the Moran model. In effect, the relation between time in the Moran model and in a model with nonoverlapping generations is not linear, but given by

$$t=\sum_{\iota=0}^{\tau}\frac{2}{N(\iota)}, \tag{20}$$

where t denotes time in the nonoverlapping generation model and is counted backwards from the present.

2. *A Moran model.* Consider the following diploid two-allele Moran model with selection and expanding population size. On average $2r$ new genes (genes, alleles, and individuals will be used synonymously) are added to the population per Moran generation. This is accomplished in the following way,

$$N(\tau)=N(0)-\lfloor 2r\tau\rfloor, \tag{21}$$

where $\lfloor x\rfloor$ denotes the integer part of x. If $N(\tau)=1$, we let $N(\tau')=1$ for all $\tau'>\tau$. If $N(0)$ is large, Eq. (21) corresponds to an exponential increase in population size at rate r per real generation. In fact, solving

$$\begin{aligned}t&=\sum_{\iota=0}^{\tau}\frac{2}{N(\iota)}\approx\int_0^{2\tau/N(0)}\frac{1}{1-rx}\,dx\\&=\frac{1}{r}\log\left(1-\frac{2r\tau}{N(0)}\right)\end{aligned} \tag{22}$$

with respect to τ gives that $\tau=N(0)(1-e^{rt})/2r$ generations counted backwards from time 0 correspond to t real generations. The population size at time τ is $N(\tau)=N(0)-\lfloor 2r\tau\rfloor$ which is $\approx N(0)\,e^{-rt}$, as claimed. The error term in (22) is of order $o(1/N(0))$, where $o(x)$ denotes a term such that $o(x)/x$ vanishes for $x\to 0$.

Let the two alleles be named A_1 and A_2 and let the selection coefficients of the genotypes be

$$\begin{array}{ccc}A_1A_2 & A_1A_2 & A_2A_2\\ \hline 1+2s' & 1+2s & 1\end{array}. \tag{23}$$

A new Moran generation is formed from the previous generation by choosing 1 or 2 new offspring and one gene to die; one if $\lfloor 2r\tau\rfloor=\lfloor 2r(\tau-1)\rfloor$ and two otherwise. This is done according to (23) and such that each gene has the same chance of dying. Mutation between A_1 and A_2 occurs at rate u, with u being of order less than $1/N(0)$, that is $u=o(1/N(0))$. This makes mutations very rare and it becomes unlikely that more than one mutation has happened in the populations history.

Put

$$\begin{aligned}Q_1=&\frac{j^2}{N(\tau)^2}(1+2s')(1-u)+\frac{j(N(\tau)-j)}{N(\tau)^2}(1+2s)\\&+\frac{(N(\tau)-j)^2}{N(\tau)^2}u,\end{aligned} \tag{24}$$

and

$$\begin{aligned}Q_2=&\frac{(N(\tau)-j)^2}{N(\tau)^2}(1-u)+\frac{j(N(\tau)-j)}{N(\tau)^2}(1+2s)\\&+\frac{j^2}{N(\tau)^2}(1+2s')\,u.\end{aligned} \tag{25}$$

Denote the sum of Q_1 and Q_2 by Q,

$$Q=1+4\frac{j}{N(\tau)}s+2\frac{j^2}{N(\tau)^2}(s'-2s). \tag{26}$$

A new individual is of type A_1 with probability Q_1/Q and of type A_2 with probability Q_2/Q. If two new individuals are required, they are drawn independently of each other. The individual that dies chosen randomly amongst all individuals.

3. *Convergence to a birth–death process.* Assume the current number of A_1 alleles in generation τ is j. The

probability that there are $j+1$ A_1's in the next Moran generation is

$$P_\tau(j \to j+1) = \frac{j}{N(\tau)}(1+2s)\, c_r(\tau) + o(j/N(\tau)), \quad (27)$$

and that of $j+2$ A_1's is

$$P_\tau(j \to j+2) = o(j^2/N(\tau)^2). \quad (28)$$

In (27), $c_r(\tau)=1$ if $\lfloor 2r\tau \rfloor = \lfloor 2r(\tau-1) \rfloor$, and $c_r(\tau)=2$ otherwise. The probability that there are $j-1$ A_1's in the next Moran generation is

$$P_\tau(j \to j-1) = \frac{j}{N(\tau)} + o(j/N(\tau)). \quad (29)$$

Let β_j be the time until a birth occurs, and δ_j the time until a death occurs, given there are j A_1's at time τ. We find

$$\begin{aligned} &P_\tau(\beta_j > \tau') \\ &\quad = \prod_{\iota=0}^{\tau'-1} \left(1 - \frac{j}{N(\tau-\iota)}(1+2s)\, c_r(\tau-\iota) \right. \\ &\qquad \left. + o(j/N(\tau-\iota)) \right) \\ &\quad \approx \exp\left\{ -\frac{j}{2} \sum_{\iota=0}^{\tau'-1} \frac{2}{N(\tau-\iota)}(1+2s)\, c_r(\tau-\iota) \right\}, \quad (30) \end{aligned}$$

for $N(\tau)$ large. Again the error term is of order $o(j/N(\tau))$. However, $N(\tau-\iota)$ is constant over $\lfloor 1/2r \rfloor$ Moran generations, yielding

$$\begin{aligned} &P_\tau(\beta_j > \tau') \\ &\quad \approx \exp\left\{ -\frac{j}{2} \sum_{\iota=0}^{\tau'-1} \frac{2}{N(\tau-\iota)}(1+2s)(1+2r) \right\}, \quad (31) \end{aligned}$$

or

$$P_t(B_j > t') \approx \exp\left\{ -\frac{j}{2}(1+2s)(1+2r)\, t' \right\}, \quad (32)$$

for τ' large, such that $\tau'/N(\tau)$ is constant, see (22). That is, B_j is exponential with parameter $j(1+2s)(1+2r)/2$. In Eq. (32), t (t') is τ (τ') expressed in real generations and B_j is β_j measured in real generations. Similarly, we find

$$P_\tau(\delta_j > \tau') = P_t(D_j > t') \approx \exp\left\{ -\frac{j}{2} t' \right\}, \quad (33)$$

where D_j is δ_j measured in real generations.

We conclude that the evolution of A_1 lineages in the Moran model converges in distribution for $N(0) \to \infty$ to a continuous-time linear birth–death process with birth rate $\lambda = (1+2s)(1+2r)/2$ and death rate $\mu = 1/2$, provided the number of A_1 lineages is small compared to $N(0)$ and the mutation rate fulfills $u = o(1/N(0))$. Note that the expected offspring number of an allele in one real generation is $\exp(\lambda - \mu) = \exp(s + r + 2rs) \approx 1 + s + r$, if r and s both are small (Kendall, 1948). This allows s to be interpreted as the selective (dis) advantage of heterozygotes. Further, as seen from (32) and (33), the fitness, s', of homozygotes is asymptotically insignificant.

The same result is true (with $s=0$) for the evolution of a small number of A_2 lineages provided the number of variant alleles, A_1, is small compared to $N(0)$. The argument for this is similar to the one given above and will, therefore, not be presented again.

In one Moran generation, one or two individuals give birth. Because the relation between Moran generations and real time is not linear, the birth rate per real time unit depends on t. Consider the number, $n_B(t, t')$, of births in the real time interval $[t, t')$,

$$n_B(t, t') = \sum_{\iota=\tau}^{\tau'-1} c_r(\iota), \quad (34)$$

where τ (τ') is t (t') in Moran time. Using (22), the birth rate, $b(t)$, per real time unit at time t is, for large $N(0)$,

$$b(t) = \lim_{N(0)\to\infty} \frac{1}{N(0)} n_B(t, t + 1/N(0)) \propto \exp\{-rt\}. \quad (35)$$

APPENDIX 2

Figure 3 illustrates the genealogy of $\mathscr{D}$ going forwards in time. Consider an individual A in some generation τ_0. If A is the first common ancestor of $\mathscr{D}$ and a normal allele, then the following is true: At generation $\tau_0 = T_0^*$ individual A give birth to a new lineage, L_2, in which the mutation arise. Note that there is at most two individuals that give birth in each generation (see Appendix 1). The individual A has descendants in lineage L_1 (normal

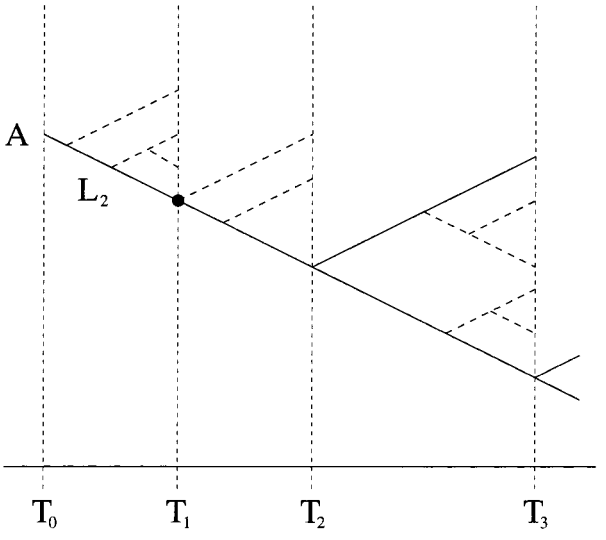

FIG. 3. Descendants of A. Solid lines represent lineages that survive until present time and dotted lines represent lineages that die out before present time. At time T_1^*, there are $k_1 = 4$ descendants of A in lineage L_2 (lineage L_1 is not shown). None of the three dotted lineages survive because otherwise A could not be the first MRCA of $\mathscr{D}$ and a present-day normal allele. They die out at some time after T_1^*. The parent of the variant survives at least until time T_2^* along with one other lineage. In total there are $k_2 = 3$ descendants of the mutant at time T_2^*. The two mutant lineages at time T_2^* has $k_3 = 6$ descendants at time T_3^*, of which only 2 survive till present time.

alleles) in the present-day population. From generation T_0^* to generation T_2^* no new lines that survive until the present-day are born in lineage L_2. Otherwise $\mathscr{D}$ could not have a MRCA at time T_2^* (but at some generation $\tau > T_2^*$) or A could not be the first common ancestor of $\mathscr{D}$ and the class of normal alleles (the first common ancestor would be at some generation $\tau < T_0^*$). At time T_2^* the lineage L_2 splits into two lineages, both with descendants in the present-day population, and so forth (Fig. 3).

The probability $P(T_j^* = \tau_j,\ 0 \leqslant j \leqslant k,\ M(D))$ is split in a sum of probabilities, one for each population history that coincide with $T_j^* = \tau_j,\ 0 \leqslant j \leqslant k$, and $M(\mathscr{D})$. If $T_0^* = \tau_0$ and τ_0 is a generation in which two individuals have offspring, the summation is (amongst more) over these two individuals. If only one individual has offspring, the summation is over this individual. We proceed in the following way. Define the events, E_j, $0 \leqslant j \leqslant k-1$, and the events F_j, $1 \leqslant j \leqslant k$, by (Fig. 3)

- $E_0(k_0)$ is the event that A has offspring (lineage L_2) at time τ_0 and the parent lineage L_1 has $k_0 > 1$ descendants in the present day population;
- $E_1(k_1)$ is the event that L_2 has k_1 descendants at time $\tau_1 + 1$, one of them gives birth to a mutant, the parent and the other $k_1 - 1$ lineages die out before present time;
- $E_{j+1}(k_{j+1})$ is the event that the $j \geqslant 1$ mutants at time τ_j evolve into $k_{j+1} \geqslant j$ lineages at time $\tau_{j+1} + 1$, such that the number, $k_{(j+1)i}$, of descendants of mutant i, $i = 1, 2, ..., j$, is at least one and $k_{(j+1)i} - 1$ of them die out before present time. Additionally one of the remaining j mutants give birth to a new lineage at time $\tau_{j+1} + 1$;
- $F_j(k)$ is the event that the $j \geqslant 1$ mutants at time τ_j evolve into k lineages at present time, such that all j mutants have at least one descendant each.

If $\tau_0 = \tau_1 + 1$, the parent of the mutant offspring survives and E_0 and E_1 are replaced by

- $E_{01}(k_0)$ is the event that A gives birth to a mutant (lineage L_2) at time τ_0 and the parent lineage L_1 has $k_0 \geqslant 1$ descendants in the present day population.

Double branching events in one generation (which occur with probability $o(1/N(\tau)^2)$) are ignored. The events E_j and F_j give all possible population histories that agree with $T_j^* = \tau_j$ and $M(\mathscr{D})$. Note that two events, E_i and E_j, $i < j$, or E_i and F_j, $i \leqslant j$, either deal with different lineages or with lineages in different time epochs. Now,

$$M_j = \bigcap_{i=0}^{j} \{T_i^* = \tau_i\} \cap M(\mathscr{D}) = \left(\bigcap_{i=0}^{j} \bigcap_{k_i} E_i(k_i) \right) \cap F_j(k) \cap M(\mathscr{D}), \qquad (36)$$

and the probability of M_j can be obtained as a sum of probabilities over i and k_i, each addend is of the form $P(E_0(k_0) \cap \cdots \cap E_j(k_j) \cap F_j(k) \cap M(\mathscr{D}))$ for some k_i, $0 \leqslant i \leqslant j$.

1. *Proof of Theorem* 1. In Appendix 1 it is shown that the evolution of a finite number of lineages can be approximated by linear birth–death (b–d) processes with time scaled in real generations. If this approximation is applied to each addend in (36) separately, we obtain an approximation to the probability of M_j. Formally, this is justifiable because the approximation applies to the evolution of the ancestor A itself (Appendix 1). One important feature of b–d processes is that lineages evolve independently of each other, such that each addend in the probability of M_j splits up in a product of terms and such that summation over k_i can be performed before multiplication of terms. Further, the mutation rate u cancels in $P(M_j \mid M(\mathscr{D})) = P(M_j)/P(M(\mathscr{D}))$ and will henceforth be ignored, e.g., in $P(M_j)$ and $P(M(\mathscr{D}))$.

Up to real generation T_1 lineage L_2 evolves with birth rate λ_1 and death rate μ_1 determined by the demographic structure of the whole population. After time T_1, the birth rate, λ_2, is also determined by the selection coefficient, whereas the death rate, μ_2, is left unchanged, $\mu_2 = \mu_1$. In lineage L_1, the process evolves with rates determined by the demographic structure solely; that is, with rates λ_1 and μ_1. In Appendix 1, it is shown that the birth and death rates are given by

$$\lambda_1 = \frac{1}{2} + r, \quad \lambda_2 = \frac{(1+2r)(1+2s)}{2}, \quad \text{and}$$

$$\mu_1 = \mu_2 = \frac{1}{2}, \tag{37}$$

where s and r are the selection coefficient of the variant allele and the rate of increase in population size per real generation, respectively. In the absence of selection, Slatkin and Rannala (1997) derived λ_1 and μ_1 using heuristic arguments.

Now, consider a linear b–d process with constant birth and death rates, λ and μ. The general theory for b–d processes was developed by Kendall (1948) and a notation similar to his is adopted. If $\mu \neq \lambda$; let

$$p_t = \frac{\mu\{e^{(\lambda-\mu)t} - 1\}}{\lambda e^{(\lambda-\mu)t} - \mu}, \tag{38}$$

and

$$\eta_t = \frac{\lambda}{\mu} p_t. \tag{39}$$

If $\mu = \lambda$, let

$$p_t = \eta_t = \frac{\lambda t}{1 + \lambda t}. \tag{40}$$

The probability, $P_k(t)$, that a single lineage has k, $k \geqslant 1$, descending lineages t generations later is given by

$$P_k(t) = (1 - p_t)(1 - \eta_t)\, \eta_t^{k-1}, \tag{41}$$

and the probability, $P_0(t)$, that the lineage is extinct t generations later, i.e., it has no descending lineages, is

$$P_0(t) = p_t. \tag{42}$$

Let p_{it} (η_{it}) be p_t (η_t), if $\lambda = \lambda_i$ and $\mu = \mu_i$, $i = 1, 2$. Applying (35), (41), and (42), the probabilities $p_{01}(t_0)$, $p_{j+1}(t_j, t_{j+1})$, and $q_j(t_j)$, respectively of E_{01}, E_{j+1}, and F_j, respectively, are given by

$$p_{01}(t_0) = p_0(t_0) = b(t_0)(1 - p_{1t_0})\, dt_0, \tag{43}$$

$$\begin{aligned} p_1(t_0, t_1) &= \sum_{k_1=1}^{\infty} g_1(t_0 - t_1)\, \eta_{1(t_0-t_1)}^{k_1-1}\, p_{1t_1}^{k_1} k_1\, dt \\ &= g_1(t_0 - t_1)\, h_1(t_0, t_1)^2\, p_{1t_1}\, dt_1, \end{aligned} \tag{44}$$

$$\begin{aligned} p_{j+1}(t_j, t_{j+1}) &= \sum_{k_{(j+1)\,i} \geqslant 1} g_2(t_j - t_{j+1})^j\, \eta_{2(t_j - t_{j+1})}^{k_{j+1}-j} \\ &\quad \times p_{2t_{j+1}}^{k_{j+1}-j} \left[\prod_{j=1}^{j} k_{(j+1)\,i} \right] dt_{j+1} \\ &= g_2(t_j - t_{j+1})^j\, h_2(t_j, t_{j+1})^{2j}\, dt_{j+1}, \end{aligned} \tag{45}$$

$$\begin{aligned} q_j(t_j) &= \sum_{m_i \geqslant 1} g_2(t_j)^j\, \eta_{2t_j}^{k-j} \\ &= C(k)\, g_2(t_j)^j\, \eta_{2t_j}^{k-j}. \end{aligned} \tag{46}$$

The sum in (46) is over all tuples $(m_1, ..., m_{j+1})$ such that $\sum_i m_i = k$ and $m_i \geqslant 1$ for all i and $C(k) = \sum_{m_i \geqslant 1} 1$. Further, the functions g_i and h_i, $i = 1, 2$, are defined by $g_i(t) = (1 - \eta_{it})(1 - p_{it})$ and $h_i(t, u) = 1/(1 - \eta_{i(t-u)} p_{iu})$, and $b(t)$ in (43) is given by $b(t) \propto \exp\{-rt\}$, Eq. (35).

The Markov property of $(T_0, T_1, ..., T_k)$ now readily follows from writing $P(M_j)$ as the product $q_j(t_j)\, p_0(t_o) \prod_{i=1}^{j} p_i(t_i, t_{i+1})$ (or $q_j(t_j)\, p_{01}(t_0) \prod_{i=2}^{j} p_i(t_i, t_{i+1})$). Also the inequality $0 < P(T_0 = T_1 \mid M(\mathscr{D})) < 1$ is a consequence of (43)–(46).

2. *Proof of Theorem* 2. Upon transformation of variables, $v_j = t_j/k$, it follows that

$$\begin{aligned} &P(T_0 = T_1 \mid M(\mathscr{D})) \\ &\quad = \frac{k \int p_{01}(kv_0)\, q_1(kv_0)}{k \int p_{01}(kv_0)\, q_1(kv_0) + k^2 \int p_0(kv_0)\, p_1(kv_0, kv_1)\, q_1(kv_1)} \to 0, \end{aligned} \tag{47}$$

for $k \to \infty$. Equations (6)–(8) are easily obtained using (47), (38)–(46), and standard limit considerations. Equations (9)–(11) follow from showing that if X_j, $j \geqslant 1$, is a series of independent variables that fulfill (10)–(11), then V_j, $j \geqslant 1$, defined by (9), fulfill (6) and (7).

ACKNOWLEDGMENTS

M. Stephens is thanked for reading and commenting on an early version of the manuscript. He is also thanked for his help in providing simulations of genealogies under circumstances similar to those described in this paper. The Mathematical Genetics Group at Oxford

University is thanked for commenting on an oral presentation of the paper. The author was supported by grant BBSRC 43/MMI09788 and by the Carlsberg Foundation, Denmark.

REFERENCES

Bratley, P., Fox, B. L., and Schrage, L. E. 1983. "A Guide to Simulation," Springer-Verlag, New York.

Ewens, W. J. 1979. "Mathematical Population Genetics," Springer-Verlag, New York.

Griffiths, R. C., and Tavaré, S. 1998. The age of a mutant in a general coalescent tree, *Stoch. Models* **14**, 273–295.

Kendall, D. G. 1948. On the generalized birth-and-death process, *Ann. Math. Stat*. **19**, 1–15.

Kimura, M., and Ohta, T. 1973. The age of a neutral mutant persisting in a finite population, *Genetics* **75**, 199–212.

Maruyama, T. 1974a. The age of an allele in a finite population, *Genet. Res* **23**, 137–143.

Maruyama, T. 1974b. The age of a rare mutant gene in a large population, *Am. J. Hum. Genet*. **26**, 669–673.

Nee, S., May, R. M., and Harvey, P. H. 1994. The reconstructed evolutionary process, *Philos. Trans. R. Soc. London, Ser B* **344**, 305–311.

Rannala, B. 1997. On the genealogy of a rare allele, *Theor. Pop. Biol*. **52**, 216–223.

Saunders, I. W., Tavaré, S., and Watterson, G. A. 1984. On the genealogy of nested subsamples from a haploid population, *Adv. Appl. Prob*. **16**, 471–491.

Seshadri, V. 1993. "The Inverse Gaussian Distribution," Clarendon Press, Oxford.

Slatkin, M., and Rannala, B. 1997. Estimating the age of alleles by use of intraallelic variability, *Am. J. Hum. Genet*. **60**, 447–458.

Stephens, M. 2000. Times on trees, and the age of an allele, *Theor. Pop. Biol*. **57**, 109–119.

Thompson, E. A. 1976. Estimation of age and rate of increase of rare variants, *Am. J. Hum. Genet*. **28**, 442–452.

Thompson, E. A., and Neel, J. V. 1997. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history, *Am. J. Hum. Genet*. **60**, 197–204.

Wiuf, C. 2000. On the genealogy of a sample of neutral rare alleles, *Theor. Pop. Biol*. **58**, 61–75.

Wiuf, C., and Donnelly, P. 1999. Conditional genealogies and the age of a neutral mutant, *Theor. Pop. Biol*. **56**, 183–201.