Genet. Res., Camb. (1992), **60**, *pp.* 221–234

Do Eve's Alleles Live On?

G. A. WATTERSON[1,*] AND P. DONNELLY[2]
[1] *Department of Mathematics, Monash University, Victoria 3168, Australia*
[2] *School of Mathematical Sciences, Queen Mary and Westfield College, London, E1 4NS, U.K.*

(*Received 2 March 1992 and in revised form 2 June 1992*)

Summary

Consider a random sample of genes at a locus, drawn from a population evolving according to the infinitely many, neutral, alleles model. The sample will have a most recent common ancestor gene, which we shall call 'Eve'. The probability distribution, for the number of genes of oldest allelic type in a sample, is known and has a neat form. Rather less is known about the distribution for the number of genes in the sample which are of the same allelic type as Eve possessed. If the latter number is positive, then these genes are automatically of the oldest type in the sample. But Eve may have no non-mutant descendants in the sample; then, the oldest allele will be a mutant arising in a line of descent after Eve. The paper studies the number of non-mutant descendants from Eve, its distribution and moments. It seems that there may be few neat results. In large samples, the proportion of genes of Eve's type has an approximate β-like density, together with a discrete probability atom at zero, if the mutation rate parameter is low. Extinction of the allele of even the *population's* common ancestor is possible, but not certain, and bounds are obtained for its probability. Some comments are made about the applications and implications of the results for human mitochondrial DNA.

1. Introduction

We use the term 'Eve' to denote the most recent common ancestor (MRCA) of a random sample of genes drawn from one locus in the present generation. How many of these descendant genes will have inherited Eve's allele, without mutation? We study this question, using the coalescent process as a model for the stochastic development of the sample's genealogical tree. For a discussion of the coalescent, see e.g. Kingman (1982), Watterson (1984), Tavaré (1984), and Griffiths (1989).

Consider a sample of n genes, one from each of n homologous chromosomes, taken at a particular instant. Looking backwards in time, usually one sees that each gene has descended from its own parent gene, without mutation. However, occasionally, two genes will have descended from a common parent gene (a coalescing of their ancestral lines) or, again occasionally, a gene will have been subject to mutation, changing its previous allelic type to a completely new allelic type. We call either of these occurrences an 'event' in the genealogy.

When there are n genes under discussion, the model assumes that the most recent event in the past was either a coalescent event, with probability $(n-1)/(n+\theta-1)$, or a mutation event, with probability $\theta/(n+\theta-1)$. Here, θ is a mutation rate parameter; if the population size is $2N$ genes and u is the probability of any one gene mutating in one generation, then θ is twice the mutation rate per gene per unit time, where time is measured in units of $2N$ generations. Thus $\theta = 4Nu$. (For haploid mitochondrial 'genes' discussed in Section 4, however, we define $\theta = 2Mu$, where M is the size of the female population, and measure time in units of M generations.)

The sample of genes may be traced back, through various coalescent events, to a single ancestor, 'Eve'. Were it not for mutation, the sample genes would all be of the same allelic type as Eve's type. Of course, the possibility of mutation means that many, if not all, of the sample genes may be of new allelic types, arising in the lines of descent since Eve. All mutations are assumed to result in new alleles.

* Corresponding author.

Kelly (1979), Exercise 7.2.6, showed that the *oldest* allele present in a sample would have i representative genes in the sample with probability

$$p_n(i) = \frac{\theta}{n}\binom{n}{i}\Big/\binom{\theta+n-1}{i}$$
$$= \frac{\theta}{n+\theta-i}\prod_{k=n-i+2}^{n}\frac{k-1}{k+\theta-1}, \qquad (1.1)$$

for $i = 1, 2, \ldots, n$, and, of course, $p_n(0) = 0$. See also Donnelly (1986).

The oldest allele in the sample may be descended from a mutant ancestor *after* Eve, or could possibly be descended from Eve's gene without mutation. We denote the number of genes in the sample which are of Eve's type by Y_n, and the number of genes of the oldest type in the sample by X_n. Thus

$$p_n(i) = \Pr(X_n = i),$$

and let

$$q_n(i) = \Pr(Y_n = i).$$

Of course, if it happens that the oldest allele is present in *all* the genes in a sample, then it must be that Eve had that allele. Hence

$$q_n(n) = p_n(n) = \prod_{k=2}^{n}\frac{k-1}{k+\theta-1} = \frac{\Gamma(n)\Gamma(1+\theta)}{\Gamma(n+\theta)}. \qquad (1.2)$$

More generally, because

$$Y_n = X_n \quad \text{if} \quad Y_n \geqslant 1,$$

then for $i \geqslant 1$,

$$q_n = p_n(i)\Pr(\text{oldest allele} = \text{Eve's allele} \mid X_n = i), \qquad (1.3)$$

but the second probability on the right is apparently not easy to evaluate.

It is the aim of the present paper to investigate the distribution $q_n(.)$. Because of a theorem of Shimizu (1987), see also Watterson (1989), this distribution applies also to the number of non-mutant copies of an MRCA in a multi-gene family of n genes on *one* chromosome, subject to mutation and gene conversion.

An explicit, but complicated, formula for $q_n(i)$ is given by (2.14) and (2.15) below, and various other aspects of the distribution are studied in Section 2. In Section 3, limiting results are found when $n \to \infty$. In Section 4, we make some remarks about the application of our results to human mitochondrial DNA.

Further work could be done on the other distributions, introduced by Griffiths (1989), for the numbers of non-mutant copies of ancestors at the nodes of the genealogical tree subsequent to the MRCA.

Later, we will use the notations

$$\theta_{(n)} = \theta(\theta+1)(\theta+2)\ldots(\theta+n-1),$$

and

$$\theta_{[n]} = \theta(\theta-1)(\theta-2)\ldots(\theta-n+1).$$

2. The distribution

(i) *A recurrence relation*

Griffiths (1989), eqn (3.7), obtained the recurrence relation

$$[n(n-1)+i\theta]\,q_n(i) = n(i-1)\,q_{n-1}(i-1)$$
$$+n(n-i-1)\,q_{n-1}(i)+(i+1)\,\theta q_n(i+1) \qquad (2.1)$$

as a special case of a more general result. The equation holds for $i = 0, 1, \ldots, n$ [provided we interpret $q_n(i)$ as zero for i outside this range] and for $n = 2, 3, \ldots$.

The equation (2.1) can be derived directly as follows. In order for the sample of size n to have i genes of Eve's type, the immediately preceding event was either a mutation or a coalescent. If it were a mutation [probability $\theta/(n+\theta-1)$] then either there were $i+1$ genes of Eve's type [probability $q_n(i+1)$] and one of them mutated [probability $(i+1)/n$], or there were i genes of Eve's type [probability $q_n(i)$], and one of the *other* genes mutated [probability $(n-i)/n$]. If the most recent event had been a coalescence [probability $(n-1)/(n+\theta-1)$], then there were $n-1$ ancestors at that stage. If i of these were of Eve's type [probability $q_{n-1}(i)$] then one of the *other* genes must leave two descendants [probability $(n-i-1)/(n-1)$], while if $i-1$ of these ancestors were of Eve's type [probability $q_{n-1}(i-1)$] then one of *these* must leave two descendants [probability $(i-1)/(n-1)$]. Putting the various possibilities together, we get

$$q_n(i) = \frac{\theta}{n+\theta-1}\left[q_n(i+1)\frac{i+1}{n}+q_n(i)\frac{n-i}{n}\right]$$
$$+\frac{n-1}{n+\theta-1}\left[q_{n-1}(i)\frac{n-i-1}{n-1}+q_{n-1}(i-1)\frac{i-1}{n-1}\right].$$

This can be re-arranged to yield (2.1).

The probabilities (2.1) can, of course, be computed in succession, for $n = 2, 3, \ldots$ and, for each n, for $i = n-1, n-2, \ldots, 1, 0$. The boundary condition (1.2) applies (or can be deduced as the $i = n$ case), as does the initial condition $q_1(i) = \delta_{i,1}$, the Kronecker delta.

(ii) *Some moments*

Griffiths (1986), see also Beder (1988), showed that the mean number of genes of Eve's type, μ_n say, is given by

$$\mathrm{E}(Y_n) = \mu_n = \sum_{i=0}^{n} i q_n(i) = n\prod_{i=2}^{n}\frac{i(i-1)}{i(i-1)+\theta}. \qquad (2.2)$$

The mean number of genes of the oldest allele is simpler:

$$\mathrm{E}(X_n) = \sum_{i=1}^{n} i p_n(i) = \frac{n+\theta}{1+\theta}, \qquad (2.3)$$

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

Table 1. *Distributions of numbers of genes of MRCA allele, $q_{10}(.)$, and of oldest allele, $p_{10}(.)$, in sample of 10 genes*

Tabulated: $q_{10}(i)$ above $p_{10}(i)$.

	θ					
i	0·01	0·1	0·5	1	2	5
0	0·0000	0·0056	0·0897	0·2307	0·4717	0·8182
	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000
1	0·0011	0·0105	0·0414	0·0628	0·0754	0·0507
	0·0011	0·0110	0·0526	0·1000	0·1818	0·3571
2	0·0012	0·0117	0·0446	0·0649	0·0720	0·0397
	0·0012	0·0122	0·0557	0·1000	0·1636	0·2473
3	0·0014	0·0132	0·0484	0·0672	0·0683	0·0302
	0·0014	0·0138	0·0594	0·1000	0·1455	0·1648
4	0·0017	0·0152	0·0532	0·0698	0·0641	0·0223
	0·0017	0·0158	0·0640	0·1000	0·1273	0·1049
5	0·0020	0·0180	0·0593	0·0727	0·0593	0·0158
	0·0020	0·0186	0·0698	0·1000	0·1091	0·0629
6	0·0025	0·0220	0·0675	0·0761	0·0538	0·0106
	0·0025	0·0227	0·0776	0·1000	0·0909	0·0350
7	0·0033	0·0286	0·0791	0·0800	0·0473	0·0066
	0·0033	0·0292	0·0887	0·1000	0·0727	0·0175
8	0·0049	0·0410	0·0978	0·0848	0·0397	0·0037
	0·0049	0·0418	0·1064	0·1000	0·0545	0·0075
9	0·0097	0·0752	0·1351	0·0909	0·0303	0·0017
	0·0097	0·0759	0·1419	0·1000	0·0364	0·0025
10	0·9722	0·7591	0·2838	0·1000	0·0182	0·0005
	0·9722	0·7591	0·2838	0·1000	0·0182	0·0005
Mean	9·9105	9·1522	6·5809	4·5507	2·4031	0·5482
	9·9109	9·1818	7·0000	5·5000	4·0000	2·5000
Var.	0·4439	3·9266	11·5740	12·6064	8·7791	2·0575
	0·4394	3·5773	8·4000	8·2500	6·0000	2·6786

see Kelly (1979), Exercise 7.2.3. It can also be shown that the variance of the number of genes of the oldest allele is

$$\mathrm{Var}\,(X_n) = \frac{\theta(n-1)(n+\theta)}{(2+\theta)(1+\theta)^2}. \tag{2.4}$$

(iii) *Examples*

In Table 1 we show some examples of the distributions $p_n(.)$ and $q_n(.)$, together with their means and variances, in the case when $n = 10$. In particular, we note that in conformity with (1.3),

$$q_n(i) < p_n(i), \quad i = 1, 2, \ldots, n-1,$$

although the two agree to 4 decimal places when $n = 10$ and $\theta = 0{\cdot}01$, for instance. Note also that $p_n(0) = 0$ whereas $q_n(0) > 0$, and that $p_n(n) = q_n(n)$.

The $p_n(.)$ distribution is

(i) J-shaped when $\theta < 1$,

(ii) uniform on $1, 2, \ldots, n$ when $\theta = 1$,

(iii) L-shaped when $\theta > 1$.

The $q_n(.)$ distribution can be similar, but can also be bimodal (modes at 0 and n) for a range of intermediate θ values. The uniform solution $q_n(i) \equiv 1/(n+1)$ for $i = 0, \pm 1, \pm 2, \ldots$ solves (2.1) when $\theta = 2$, but it does not satisfy the boundary conditions of our problem in general. The one case when $q_n(.)$ is correctly uniform over $0, 1, \ldots, n$ is when $n = 2, \theta = 2$.

We note from Table 1 that the means decrease, and the variances at first increase and then decrease, as θ increases. These results are to be expected.

While it is possible to obtain, from (2.1), explicit expressions for $q_n(i)$ for values of i close to n, the only simple ones seem to be the expression (1.2) for $q_n(n)$, and

$$q_n(n-1) = \frac{\theta n}{n+\theta} q_n(n) = \theta \prod_{k=2}^{n+1} \frac{k-1}{k+\theta-1}. \tag{2.5}$$

For instance, the next most simple is

$$q_n(n-2) = A_n\, q_n(n) \tag{2.6}$$

where, for $n = 2$,

$$A_2 = \theta^2/(2+\theta),$$

and for $n > 2$,

$$A_n = \theta(1+\theta) \sum_{j=3}^{n} \frac{j^2}{(j+\theta)[j(j-1)+(j-2)\theta]} B_n(j)$$

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

with

$$B_n(j) = \prod_{k=j+1}^{n} \frac{k(k-3)(k+\theta-1)}{(k-1)[k(k-1)+(k-2)\theta]}.$$

The product here is interpreted as 1 when $j = n$.

It is clear that we need some further information to gain insight into the distribution $q_n(.)$ for Y_n. We now describe some equations which pertain to this distribution.

(iv) *An integral equation*

It is possible to obtain, from (2.1), an integral equation for the probability generating function (p.g.f)

$$Q_n(s) = \mathrm{E}(s^{Y_n}) = \sum_{i=0}^{n} q_n(i)\, s^i,$$

namely

$$Q_n(s) = \frac{ns}{\theta} Q_{n-1}(s) + \int_s^1 \left[1 + \frac{1-v}{n-1} - \frac{nv}{\theta}\right] \times Q_{n-1}(v) f_{n,s}(v)\, dv, \tag{2.7}$$

where

$$f_{n,s}(v) = a\frac{(1-v)^{a-1}}{(1-s)^a}, \quad s < v < 1, \tag{2.8}$$

is a probability density function, and where

$$a \equiv a_n = \frac{n(n-1)}{\theta}.$$

(v) *Another recurrence relation*

The moments of the $f_{n,s}(v)$ distribution in (2.8) can be shown to be

$$\int_s^1 v^k f_{n,s}(v)\, dv = \sum_{l=0}^{k} V_{k,l}^{(n)} s^l, \quad k = 0, 1, 2, \ldots,$$

say, where

$$V_{k,l}^{(n)} = a\frac{k!}{l!(a+l)_{(k-l+1)}}, \quad k \geqslant l.$$

If we use these moments, and expand both sides of (2.7) in powers of s and equate coefficients, we find the recurrence

$$\begin{aligned} q_n(i) &= \frac{n}{\theta} q_{n-1}(i-1) + \frac{1}{n-1} \\ &\quad \times \sum_{k=i-1}^{n-1} q_{n-1}(k)[nV_{k,i}^{(n)} - (a+1)\, V_{k+1,i}^{(n)}] \\ &= \frac{n(i-1)}{n(n-1)+i\theta} q_{n-1}(i-1) + \frac{n(n-1)}{i!\theta^2} \\ &\quad \times \sum_{k=i}^{n-1} q_{n-1}(k) \frac{k![n(n-k-1)+(k+1)\theta]}{(a+i)_{(k-i+2)}}. \end{aligned} \tag{2.9}$$

Another derivation will be mentioned later.

The equation (2.9) involves more terms than does (2.1); on the other hand it is an equation for $q_n(i)$ using $q_{n-1}(.)$ terms only. While the latter may be of some theoretical advantage, it is clear that (2.1) is simpler for computation.

(vi) *Factorial moments*

Another deduction which can be made from (2.7) is an equation for all the factorial moments of Y_n. Using the notation

$$\mu_{n,k} = E[Y_n(Y_n-1)(Y_n-2)\ldots(Y_n-k+1)] = E[(Y_n)_{[k]}]$$

for the kth factorial moment, we have

$$Q_n(s) = 1 + \sum_{k=1}^{n} (1-s)^k(-1)^k \mu_{n,k}/k!$$

which, when substituted into (2.7) and after equating powers of $1-s$, yields, for $k \geqslant 1$,

$$\mu_{n,k} = \frac{n}{n(n-1)+k\theta}[(n+k-1)\mu_{n-1,k} + k(k-1)\mu_{n-1,k-1}]. \tag{2.10}$$

[The integral of $(1-v)^k$ with respect to the density (2.8) is $a(1-s)^k/(a+k)$, and is needed to obtain (2.10).]

It may be noticed that (2.10) is rather simpler than either (2.1) or (2.9). Of course, because $Y_n \leqslant n$, $\mu_{n,k} = 0$ for all $k > n$. This can be taken as a boundary condition, together with $\mu_{n,0} = 1$. Solving (2.10) in succession, both from $k = 0$ upwards and from $k = n$ downwards, yields explicit solutions for the factorial moments. But they get progressively more complicated, the further is the order away from 0 or n.

The first order moment is, of course, given by (2.2):

$$\mu_{n,1} \equiv \mu_n = n\prod_{i=2}^{n} \frac{i(i-1)}{i(i-1)+\theta}.$$

The second order factorial moment is

$$\mu_{n,2} = 2n(n+1)\sum_{k=1}^{n-1} \frac{1}{(k+1)(k+2)}\left[\prod_{j=2}^{k} \frac{j(j-1)}{j(j-1)+\theta}\right] \times \left[\prod_{j=k+1}^{n} \frac{j(j-1)}{j(j-1)+2\theta}\right],$$

in which the empty product when $k = 1$ is interpreted as 1. At the other end, we have

$$\mu_{n,n} = \frac{n!(n-1)!}{(1+\theta)_{(n-1)}},$$

and

$$\mu_{n,n-1} = \frac{n!(n-1)!(n+2\theta)}{(1+\theta)_{(n)}}.$$

We shall show in the next section how the general form for the factorial moments may be obtained.

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

(vii) *Probabilities from moments*

In principle, it is possible to obtain expressions for the $q_n(i)$s by the use of factorial moments, via the formula

$$q_n(i) = \frac{1}{i!}\sum_{j=0}^{n-i}\frac{(-1)^j}{j!}\mu_{n,i+j}. \tag{2.11}$$

For instance,

$$q_n(n) = \frac{1}{n!}\mu_{n,n},$$

and

$$q_n(n-1) = \frac{1}{(n-1)!}[\mu_{n,n-1}-\mu_{n,n}],$$

which agree with (1.2) and (2.5).

The factorial moments, $\mu_{n,k}$ can be very large; it is numerically preferable to use scaled moments

$$r_n(k) = \mu_{n,k}/n_{[k]} \tag{2.12}$$

which must lie in $[0,1]$. Indeed, since

$$r_n(k) = E\left(\frac{(Y_n)_{[k]}}{n_{[k]}}\right),$$

we see that $r_n(k)$ is the probability that, if we choose k genes without replacement from the sample of n, then all k genes would be of Eve's allelic type. Substituting from (2.12) into (2.10) yields the recurrence

$$[n(n-1)+k\theta]\,r_n(k) = k(k-1)\,r_{n-1}(k-1) + [n(n-1)-k(k-1)]\,r_{n-1}(k), \tag{2.13}$$

with boundary condition

$$r_n(0) = 1, \quad n = 1,2,\ldots.$$

The distribution for Y_n can then be found, in principle from

$$q_n(i) = \sum_{j=0}^{n-i}(-1)^j\frac{n_{[i+j]}}{i!j!}r_n(i+j). \tag{2.14}$$

While (2.13) is a good method for calculating the probabilities $r_n(.)$ in succession, it is possible to obtain explicit, but complicated, algebraic expressions.

PROPOSITION 1. *The scaled factorial moments are given by*

$$r_n(i) = \frac{2n_{(i)}\,i!(i-1)!}{n_{[i]}} \times \sum\sum\ldots\sum \times \prod_{j=1}^{i}\left\{\frac{1}{(k_{j-1}+j-1)(k_{j-1}+j)}\prod_{l=k_{j-1}+1}^{k_j}\frac{l(l-1)}{l(l-1)+j\theta}\right\}, \tag{2.15}$$

where the $i-1$-fold summation is over

$$1 = k_0 \leqslant k_1 < k_2 < \ldots < k_{i-1} < k_i = n.$$

Proof. In the coalescent process describing the way in which equivalence classes of genes coalesce in backwards time because of sharing common ancestors, the jump chain is independent of the time variables between the jumps; Kingman (1982). Let K_j denote the number of ancestors of the whole sample of n genes when the number of ancestors of a sub-sample of i genes has just reduced to $j, 0 < j \leqslant i$. During the drop in the sub-sample from j to $j-1$, the number of ancestors of the whole sample drops from K_j to K_j-1, then to $K_j-2,\ldots,$ to $K_{j-1}+1$ and then to K_{j-1}. The last transition must involve a coalescing of two sub-sample ancestors, with probability

$$j(j-1)/[(K_{j-1}+1)\,K_{j-1}]$$

because of the *random* pairing of ancestors. The earlier transitions, e.g. from m to $m-1$ ancestors, must involve the coalescing of two ancestors which are not both sub-sample ancestors, with probability

$$[m(m-1)-j(j-1)]/[m(m-1)] = (m-j)(m+j-1)/[m(m-1)].$$

Thus

$$\Pr(K_{j-1} = k_{j-1}\,|\,K_j = k_j) = \frac{j(j-1)}{(k_{j-1}+1)\,k_{j-1}}\prod_{m=k_{j-1}+2}^{k_j}\frac{(m-j)(m+j-1)}{m(m-1)},$$

which is essentially (2.7) in Saunders *et al.* (1984) with their i, j, k and l replaced by k_j, j, $j-1$ and k_{j-1} respectively. Hence,

$$\Pr(K_1 = k_1, K_2 = k_2,\ldots,K_{i-1} = k_{i-1}) = \prod_{j=2}^{i}\left[\frac{j(j-1)}{(k_{j-1}+1)\,k_{j-1}}\prod_{m=k_{j-1}+2}^{k_j}\frac{(m-j)(m+j-1)}{m(m-1)}\right] = \frac{n_{(i)}\,i!(i-1)!}{n_{[i]}}\prod_{j=2}^{i}\frac{1}{(k_{j-1}+j-1)(k_{j-1}+j)}.$$

Here, we assume that

$$1 \leqslant k_1 < k_2,\ldots,k_{i-1} < k_i = n.$$

In order that all i genes, in our sub-sample from n, are of Eve's type, no mutations can occur in their lines of descent. Now the probability that no mutations occur while they had j ancestors, over a time interval of length T_j say, is the Poisson probability

$$\exp(-\tfrac{1}{2}j\theta T_j).$$

Supposing that $K_j = k_j$ and $K_{j-1} = k_{j-1}$, then the time, T_j, over which our i genes had j ancestors is the sum of independent, exponentially distributed random variables with means

$$2/[k_j(k_j-1)], 2/[(k_j-1)(k_j-2)],\ldots,2/[(k_{j-1}+1)\,k_{j-1}].$$

Averaging the Poisson probability over the distribution of T_j yields the probability of no mutation during this period as

$$\prod_{l=k_{j-1}+1}^{k_j}\frac{l(l-1)}{l(l-1)+j\theta}.$$

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

Putting these facts together, and introducing the conventions $k_0 = 1$ and that empty products are 1, we have proven (2.15). □

The explicit expression (2.15) is obviously not very attractive! In principle, it yields expressions for the factorial moments, using (2.12), and for the distribution $q_n(.)$, via (2.14).

(viii) *Two stochastic processes*

Our Y_n need not be considered as a once-only random variable, but also as one term in a stochastic process $(Y_n, n = 1, 2, \ldots\}$. Thus Y_n is the number of genes of Eve's allele in the genealogy, just prior to each change in descendant number from n to $n+1$. The initial state is $Y_1 = 1$, as Eve's gene has Eve's allele! The transition probabilities can be read off (2.9), or derived by considering the evolution of the coalescent in forward time:

$$\Pr(Y_n = l \mid Y_{n-1} = i) = \frac{i}{n-1} V^{(n)}_{i+1,l} + \left(1 - \frac{i}{n-1}\right) V^{(n)}_{i,l}.$$

These are, unfortunately, functions of n rather than being time-homogeneous.

The corresponding conditional factorial moments are

$$\mathrm{E}((Y_n)_{[k]} \mid Y_{n-1} = i) = n[(n+k-1)\, i_{[k]} + k(k-1)\, i_{[k-1]}/[n(n-1)+k\theta],$$

which are consistent with (2.10). In particular,

$$\mathrm{E}(Y_n \mid Y_{n-1} = i) = in^2/[n(n-1)+\theta]. \tag{2.16}$$

A bivariate process in *continuous* time, which allows for the exponential waiting times between coalescent events, is $\{N(t), Y(t)\}$, in which $N(t)$ plays the role of n, the number of descendants at time t. The transition rates, from state (n, i), are to

$$\left.\begin{array}{ll} (n+1, i+1): & \dfrac{n(n-1)}{2}\dfrac{i}{n}, \\ (n+1, i): & \dfrac{n(n-1)}{2}\left(1 - \dfrac{i}{n}\right), \\ (n, i-1): & \theta i/2. \end{array}\right\} \tag{2.17}$$

This process has initial state $N(0) = 2$, $Y(0) = 2$, because we take $t = 0$ to be just *after* Eve's gene split into two offspring genes, which will both be of Eve's allelic type except for a negligible probability of a mutation. Y_n in the previous process corresponds to $Y(t)$ here, *just before* $N(t)$ jumps from n to $n+1$.

We will make some use of these processes in the next section, in connection with asymptotic behaviour.

3. Some asymptotics

The rather unsatisfactory state of the 'solution' so far found for our problem suggests that we should attempt to find asymptotic results when n is large. Limiting results may be interpreted as applying to an infinite *population*. We will assume from now on that θ remains fixed as n increases. We start by summarizing some limiting results for X_n and $p_n(.)$.

(i) *The oldest allele*

It is easy to check that for $p_n(.)$ given by (1.1),

$$p_n(i) \sim \begin{cases} \dfrac{\theta}{n}\left(1 - \dfrac{i}{n}\right)^{\theta-1} & \text{if} \quad n - i \to \infty \\ \dfrac{\theta\Gamma(n-i+\theta)}{n^\theta \Gamma(n-i+1)} & \text{if} \quad n - i = O(1) \quad \text{as } n \to \infty. \end{cases} \tag{3.1}$$

(3.2)

The mean (2.3) is given, fairly accurately, by an integral approximation based on (3.1):

$$\mathrm{E}(X_n) = \sum_{i=1}^{n} i p_n(i) \approx \theta n \int_0^1 v(1-v)^{\theta-1}\, dv = \frac{n}{1+\theta}.$$

The corresponding variance is asymptotically

$$\mathrm{Var}(X_n) \sim \frac{\theta n^2}{(2+\theta)(1+\theta)^2},$$

which may be calculated either from (2.4) or from the continuous $\beta(1, \theta)$ density

$$\theta(1-v)^{\theta-1}, \quad 0 < v \leqslant 1.$$

This density is the density of the oldest of the population's age-ordered allele frequencies,

$$\lim_{n\to\infty} \frac{X_n}{n} = V, \quad \text{say,}$$

(now said to have the GEM distribution); see Donnelly (1986) and Hoppe (1987).

Thus the oldest allele's behaviour is well understood.

(ii) *Eve's allele*

PROPOSITION 2. *There exists a random variable, W say, such that*

$$\frac{Y_n}{n} \to W \text{ a.s. as } n \to \infty.$$

Proof. From (2.16) we see that

$$\mathrm{E}\left(\frac{Y_n}{n} \,\middle|\, Y_{n-1}\right) = \frac{n(n-1)}{n(n-1)+\theta}\frac{Y_{n-1}}{n-1} \leqslant \frac{Y_{n-1}}{n-1},$$

so that $Y_n/n, n = 1, 2, 3, \ldots$ is a supermartingale, bounded on $[0, 1]$, and that

$$\frac{Y_n}{\mu_n} = \frac{Y_n}{n}\prod_{j=2}^{n}\frac{j(j-1)+\theta}{j(j-1)}$$

is a (bounded) martingale. The expression (2.2) can be derived this way, noting the initial value $Y_n/n = 1$ when $n = 1$. Doob's Martingale Convergence Theorem can now be applied to obtain the result. □

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

We will now discuss various aspects of the asymptotic behaviour of the distribution of Y_n, and hence of W's distribution. We will show, *inter alia*, that W has an absolutely continuous distribution on $(0, 1]$, together with an atom of discrete probability at 0.

(iii) *Moments*

It is known, Beder (1988), that the limit of the mean of Y_n/n, that is, of μ_n/n from (2.2), is

$$\mathrm{E}(W) = \lim_{n\to\infty} \mu_n/n \sim \begin{cases} \theta\pi/\cos[\frac{1}{2}\pi(1-4\theta)^{\frac{1}{2}}] & \text{if} \quad \theta \leqslant \frac{1}{4}, \\ \theta\pi/\cosh[\frac{1}{2}\pi(4\theta-1)^{\frac{1}{2}}] & \text{if} \quad \theta > \frac{1}{4}. \end{cases} \tag{3.3}$$

Higher order moments are given by

$$\mathrm{E}(W^i) = \lim_{n\to\infty} r_n(i),$$

$$= 2i!(i-1)!\sum\sum\ldots\sum\prod_{j=1}^{i} \times\left[\frac{1}{(k_{j-1}+j-1)(k_{j-1}+j)}\prod_{l=k_{j-1}+1}^{k_j}\frac{l(l-1)}{l(l-1)+j\theta}\right],$$

from (2.15), where the $i-1$ fold summation is over $1 = k_0 \leqslant k_1 < k_2 < \ldots < k_{i-1} < k_i = \infty$.

(iv) *Some particular probabilities*

For the distribution of main interest, $q_n(.)$, we see from (1.2) and (3.2) that

$$q_n(n) \sim \Gamma(1+\theta)n^{-\theta} \quad \text{as} \quad n\to\infty. \tag{3.4}$$

Then, (2.5) shows that

$$q_n(n-1) \sim \theta\Gamma(1+\theta)n^{-\theta} \quad \text{as} \quad n\to\infty.$$

While it may be possible to find an asymptotic expression for $q_n(n-2)$ using (2.6), the continuation of this line of investigation seems not to be promising.

(v) *Extinction probability*

When $\{Y_n\}$ is viewed as a stochastic process, the state in which *no* genes are of Eve's type, $Y_n = 0$, is absorbing. Therefore, W has a distribution with an atom of probability at 0. Putting $i = 0$ in (2.1) yields

$$q_n(0) = q_{n-1}(0) + \frac{\theta}{n(n-1)}q_n(1)$$

so that

$$q_n(0) = \theta\sum_{m=2}^{n}\frac{q_m(1)}{m(m-1)}.$$

This shows that $q_n(0)$ is monotonically increasing in n. The transition from $n-1$ ancestors to n descendants can only *increase* the probability that there are none of Eve's allele remaining. (Paradoxically, the larger sample might be thought to have *less* chance of having none of Eve's allele; but then its 'Eve', i.e. its most recent common ancestor, may be further back in time.) There must be convergence to some positive limit, say $q_\infty(0) = \Pr(Y_n = 0 \text{ for some } n)$;

$$q_\infty(0) = \theta\sum_{m=2}^{\infty}\frac{q_m(1)}{m(m-1)}.$$

Crude bounds on $q_\infty(0)$ are

$$\tfrac{1}{2}\theta q_2(1) = \frac{\theta^2}{(2+\theta)(1+\theta)} < q_\infty(0) < \theta\sum_{m=2}^{\infty}\frac{1}{m(m-1)} = \theta. \tag{3.5}$$

The bounds establish that the probability, that the population has none of Eve's allele, approaches 0 or 1 as $\theta\to 0$ or ∞, respectively.

A more accurate upper bound can be obtained from the bivariate process (2.17).

PROPOSITION 3. *The extinction probability of Eve's allele has the upper bound*

$$q_\infty(0) \leqslant \frac{\theta e^\theta - \theta}{\theta e^\theta + 1} \sim \theta^2 \quad \text{as} \quad \theta\to 0. \tag{3.6}$$

Proof. $q_\infty(0)$ is the probability that $Y(t)$ ever hits 0, starting from $Y(0) = 2$. This is smaller than the corresponding probability for the following modification of the $\{N(t), Y(t)\}$ process. Let $\{Z(t)\}$ be a *univeriate* birth and death process, with $Z(0) = 2$ and when $Z(t) = i$, with birth and death rates being in the ratio

$$i : \frac{i\theta}{i-1} \quad \text{if} \quad i > 1,$$

and

$$i : i\theta \quad \text{if} \quad i = 1.$$

We note that this process has, if anything, a higher relative death rate than $\{Y(t)\}$ had in (2.17), for which the ratio was

$$i : \frac{i\theta}{n-1}, \quad \text{where } i \leqslant n, \text{ and } n \geqslant 2.$$

For the $\{Z(t)\}$ process, define

$$M(i) = \begin{cases} \left[\theta e^\theta - \theta\sum_{l=0}^{i-2}\frac{\theta^l}{l!}\right]/[\theta e^\theta + 1], & \text{if } i \geqslant 2, \\ \theta e^\theta/[\theta e^\theta + 1], & \text{if } i = 1, \\ 1, & \text{if } i = 0. \end{cases}$$

Then it may be checked that $\{M[Z(t)]\}$ is a martingale. But because $Z(t)$ approaches either 0 or ∞, and hence $M[Z(t)]$ approaches either 1 or 0 respectively, then

$$\Pr(Z(t)\to 0) = \lim_{t\to\infty}\mathrm{E}(M[Z(t)] \mid Z(0) = 2),$$

$$= M(2),$$

$$= \frac{\theta e^\theta - \theta}{\theta e^\theta + 1}.$$

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

Table 2. *Probability of loss of MRCA's allele, $q_n(0)$*

	θ					
n	0·01	0·1	0·5	1	2	5
1	0	0	0	0	0	0
2	$4{\cdot}93\times10^{-5}$	$4{\cdot}33\times10^{-3}$	0·0667	0·1667	0·3333	0·5952
3	$5{\cdot}75\times10^{-5}$	$5{\cdot}08\times10^{-3}$	0·0799	0·2024	0·4083	0·7170
4	$6{\cdot}03\times10^{-5}$	$5{\cdot}34\times10^{-3}$	0·0845	0·2156	0·4372	0·7628
5	$6{\cdot}15\times10^{-5}$	$5{\cdot}46\times10^{-3}$	0·0867	0·2219	0·4514	0·7852
10	$6{\cdot}31\times10^{-5}$	$5{\cdot}61\times10^{-3}$	0·0897	0·2307	0·4717	0·8182
20	$6{\cdot}35\times10^{-5}$	$5{\cdot}64\times10^{-3}$	0·0904	0·2329	0·4772	0·8279
30	$6{\cdot}36\times10^{-5}$	$5{\cdot}65\times10^{-3}$	0·0905	0·2333	0·4783	0·8299
40	$6{\cdot}36\times10^{-5}$	$5{\cdot}65\times10^{-3}$	0·0906	0·2335	0·4787	0·8306
50	$6{\cdot}36\times10^{-5}$	$5{\cdot}65\times10^{-3}$	0·0906	0·2336	0·4788	0·8309
100	$6{\cdot}36\times10^{-5}$	$5{\cdot}66\times10^{-3}$	0·0906	0·2337	0·4791	0·8314
$\frac{\theta e^{\theta}-\theta}{\theta e^{\theta}+1}$	$9{\cdot}95\times10^{-5}$	$9{\cdot}47\times10^{-3}$	0·1778	0·4621	0·8099	0·9919

Hence the required bound (3.6) is proved. □

PROPOSITION 4. *Extinction of Eve's allele is neither certain, nor impossible, for any positive value of θ. Hence there is no sub-critical, super-critical phenomenon here.*

Proof. The upper bound (3.6) is always less than 1 and the lower bound in (3.5) is always positive, for all $\theta > 0$. □

In Table 2 we show $q_n(0)$ values, and the upper bound (3.6), for various values of n and θ. It seems that the limit is attained quite quickly, at least for these θ values. The upper bound is not a very accurate approximation for intermediate values of θ. It could be improved by conditioning on the first step(s) of $Y(t)$ and starting the $Z(t)$ approximation at a later stage. Another approximation to the limit will be considered later, in connection with Table 5.

(vi) *The continuous part*

PROPOSITION 5. *There exists some function, $f(w) \geqslant 0$ say, such that for any Borel set, B, in $(0, 1]$,*

$$\Pr(W \in B) = \int_B f(w)\,\theta(1-w)^{\theta-1}\,dw, \tag{3.7}$$

where $0 \leqslant f(w) \leqslant 1$. W also has an atom of probability at 0.

Proof. The limiting proportions of the oldest allele, V, and of Eve's allele, W, are equal so long as Eve's allele is still present. Thus

$$\{W \in B\} \subseteq \{V \in B\},$$

so that

$$0 \leqslant \Pr(W \in B) \leqslant \Pr(V \in B), \tag{3.8}$$

in particular when the right side is zero. Thus over $(0, 1]$, W is absolutely continuous with respect to V. The Radon Nikodym theorem yields the existence of f, and (3.8) yields the bounds for $f(w), 0 < w \leqslant 1$. □

The equation (3.7) is analogous to (1.3); but there may be no neat formula for the continuous part of W's distribution. We now develop an approximate formula when θ is small.

(vii) *An empirical approximation for small θ*

Numerical studies of successive values of $r_n(i)$, for i near n, suggested that the following approximate ratio holds:

$$r_n(i)/r_n(i+1) \approx (i+1+2\theta)/(i+1+\theta).$$

This leads to the approximation, $\hat{r}_n(i)$, say,

$$\hat{r}_n(i) = \Gamma(1+\theta)\frac{\Gamma(n)\,\Gamma(n+1+2\theta)\,\Gamma(i+1+\theta)}{\Gamma(n+\theta)\,\Gamma(i+1+2\theta)\,\Gamma(n+1+\theta)}$$

$$= \frac{(1)_{(n-1)}\,(i+1+2\theta)_{(n-i)}}{(1+\theta)_{(n-1)}\,(i+1+\theta)_{(n-i)}}. \tag{3.9}$$

There are several virtues in (3.9).

(i) When (3.9) is substituted into (2.14), it gives a (rather surprisingly) tractable approximation to $q_n(i)$. We discuss this a little later.

(ii) $\hat{r}_n(n) = r_n(n)$ and $\hat{r}_n(n-1) = r_n(n-1)$ exactly, for each n.

(iii) Using somewhat inconsistent applications of Stirling's formula, if we approximate

$\Gamma(n)/\Gamma(n+\theta)$ by $n^{-\theta}$,

but

$\Gamma(n+1+2\theta)/\Gamma(n+1+\theta)$

by

$n^{\theta}(1+\frac{1}{2}\theta(1+3\theta)/n)$

and

$\Gamma(i+1+\theta)/\Gamma(i+1+2\theta)$

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

Table 3. *Accuracy of approximating* $r_n(i)$ *by* $\hat{r}_n(i)$

Tabulated: $r_n(i)$ above $\hat{r}_n(i)$, for various values of $w = i/n$.

w										
0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1
					$n = 10, \quad \theta = 0.01$					
1·0000	0·9911	0·9861	0·9829	0·9804	0·9785	0·9768	0·9754	0·9742	0·9732	0·9722
1·0008	0·9910	0·9861	0·9829	0·9804	0·9785	0·9768	0·9754	0·9742	0·9732	0·9722
					$n = 100, \quad \theta = 0·01$					
1·0000	0·9713	0·9648	0·9610	0·9583	0·9562	0·9545	0·9530	0·9517	0·9506	0·9496
0·9999	0·9713	0·9648	0·9610	0·9583	0·9562	0·9544	0·9530	0·9517	0·9506	0·9496
					$n = 10, \quad \theta = 0·1$					
1·0000	0·9152	0·8726	0·8450	0·8247	0·8088	0·7957	0·7847	0·7751	0·7667	0·7591
0·9965	0·9135	0·8720	0·8447	0·8246	0·8087	0·7957	0·7846	0·7751	0·7667	0·7591
					$n = 100, \quad \theta = 0·1$					
1·0000	0·7518	0·7036	0·6764	0.6575	0·6432	0.6317	0.6222	0·6140	0·6068	0·6005
0·9868	0·7517	0·7036	0·6764	0·6575	0·6432	0·6317	0·6222	0·6140	0·6068	0·6005
					$n = 10, \quad \theta = 0·5$					
1·0000	0·6581	0·5367	0·4655	0·4171	0·3815	0·3538	0·3315	0·3130	0·2973	0·2838
0·8436	0·6327	0·5273	0·4614	0·4152	0·3806	0·3534	0·3313	0·3129	0·2973	0·2838
					$n = 100, \quad \theta = 0·5$					
1·0000	0·2673	0·1939	0·1598	0·1390	0·1247	0·1141	0·1058	0·0991	0·0935	0·0887
0·7913	0·2662	0·1937	0·1597	0·1390	0·1247	0·1141	0·1058	0·0991	0·0935	0·0887
					$n = 10, \quad \theta = 1$					
1·0000	0·4551	0·3196	0·2482	0·2037	0·1731	0·1507	0·1336	0·1201	0·1091	0·1000
0·6000	0·4000	0·3000	0·2400	0·2000	0·1714	0·1500	0·1333	0·1200	0·1091	0·1000
					$n = 100, \quad \theta = 1$					
1·0000	0·0866	0·0466	0·0319	0·0243	0·0196	0·0165	0·0142	0·0124	0·0111	0·0100
0·5100	0·0850	0·0464	0·0319	0·0243	0·0196	0·0165	0·0142	0·0124	0·0111	0·0100
					$n = 10, \quad \theta = 2$					
1·0000	0·2403	0·1350	0·0883	0·0630	0·0476	0·0374	0·0303	0·0251	0·0212	0·0182
0·2758	0·1655	0·1103	0·0788	0·0591	0·0460	0·0368	0·0301	0·0251	0·0212	0·0182
					$n = 100, \quad \theta = 2$					
1·0000	0·0127	0·0040	0·0019	0·0011	0·0007	0·0005	0·0004	0·0003	0·0002	0·0002
0·1768	0·0117	0·0038	0·0019	0·0011	0·0007	0·0005	0·0004	0·0003	0·0002	0·0002
					$n = 10, \quad \theta = 5$					
1·0000	0·0548	0·0201	0·0094	0·0051	0·0030	0·0019	0·0013	0·0009	0·0007	0·0005
0·0307	0·0168	0·0098	0·0060	0·0039	0·0026	0·0018	0·0013	0·0009	0·0007	0·0005
					$n = 100, \quad \theta = 5$					
1·0000	0·0001	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000
0·0053	0·0001	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000	0·0000

by

$$i^{-\theta}(1-\tfrac{1}{2}\theta(1+3\theta)/i),$$

then we have, for large i and n,

$$\hat{r}_n(i) \approx \Gamma(1+\theta)\, i^{-\theta}\left[1-\tfrac{1}{2}\theta(1+3\theta)\left(\frac{1}{i}-\frac{1}{n}\right)\right].$$

This approximation is reasonable for i large, but less so for small i.

(iv) When (3.9) is substituted into both sides of (2.13), the ratio of the right to the left sides is

$$1+\frac{\theta^2(1+\theta)(n-i)(n-i-1)}{[n(n-1)+i\theta](n-1)(n+2\theta)(i+\theta)},$$

which is exactly 1 (as would be needed for (3.9) to be a perfect solution) when $\theta = 0$ or $i = n$ or $i = n-1$, and which is

$$1+\theta^2(1+\theta)\,O(n^{-5}) \quad \text{if } n-i = O(1) \quad \text{as } n\to\infty,$$

and

$$1+\theta^2(1+\theta)\,O(n^{-2}(i+\theta)^{-1}) \quad \text{if } n-i = O(n) \quad \text{as } n\to\infty.$$

Thus (3.9) is likely to be a very good approximation to $r_n(i)$ when θ is very small, for all i and n, and also for all θ, when i is close to n.

In Table 3, we illustrate the above remarks by comparing $r_n(i)$ and $\hat{r}_n(i)$ for various n, i, and θ values. The approximation is remarkably accurate for many entries in the table, and improves as n increases,

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

provided w and θ are held fixed, where $w = i/n > 0$. However, when i is fixed, particularly when it is close to 0, the approximation gets progressively worse as θ or n increase.

Analytical results for the worst cases of the approximation may be helpful outside the range of Table 3; (3.9) yields

$$\hat{r}_n(0) = \frac{(1)_{(n-1)}(1+2\theta)_{(n)}}{(1+\theta)_{(n-1)}(1+\theta)_{(n)}} \sim \frac{[\Gamma(1+\theta)]^2}{\Gamma(1+2\theta)},$$

whereas the exact value is

$$r_n(0) = 1,$$

and, further, (3.9) says

$$\hat{r}_n(1) = \prod_{i=2}^{n} \frac{(i+2\theta)(i-1)}{(i+\theta)(i-1+\theta)} \sim \frac{(1+\theta)[\Gamma(1+\theta)]^2}{(1+2\theta)\,\Gamma(1+2\theta)},$$

whereas (2.12) and (2.2) show that

$$r_n(1) = \prod_{i=2}^{n} \frac{i(i-1)}{i(i-1)+\theta},$$

with asymptotic behaviour as for μ_n/n in (3.3).

Let us denote by $\hat{q}_n(i)$ the approximation for $q_n(i)$, obtained by substituting $\hat{r}_n(i)$ from (3.9) into (2.14). Then we obtain

$$\hat{q}_n(i) = \theta \frac{\Gamma(n+1)\,\Gamma(n)\,\Gamma(n-i+\theta)\,\Gamma(i+1+\theta)}{\Gamma(i+1)\,\Gamma(n-i+1)\,\Gamma(n+\theta)\,\Gamma(n+1+\theta)} \tag{3.10}$$

$$= \frac{\theta}{n-i+\theta} \prod_{j=n-i+2}^{n} \frac{j-1}{j-1+\theta} \prod_{j=i+2}^{n+1} \frac{j-1}{j-1+\theta} \quad (\text{if } i>0)$$

$$= p_n(i) \prod_{j=i+2}^{n+1} \frac{j-1}{j-1+\theta} \quad (\text{if } i>0). \tag{3.11}$$

In Table 4 we show the comparison of $nq_n(i)$ with $n\hat{q}_n(i)$ for a range of θ, n and i values, with $w = i/n > 0$. Some comments about the results (and others not shown in the Table) are

(i) $\hat{q}_n(i) = q_n(i)$ exactly for $i = n-1$ and n.

(ii) If $\hat{q}_n(i)$, as given by (3.10), is substituted into both sides of (2.1), the ratio of the right to left sides is

$$1 + \frac{\theta(1+\theta)(n-i)(n-i-1)(n-1+2\theta)}{[n(n-1)+i\theta](n-1)(n-i-1+\theta)(i+\theta)}.$$

This should be 1, for $\hat{q}_n(i)$ to be an exact solution of (2.1). We see that this is so for $i = n$ or $n-1$, or if $\theta = 0$. Otherwise, we have that the ratio is

$$1 + \theta(1+\theta)\,O(n^{-3}) \quad \text{if } n-i = O(1) \quad \text{as } n \to \infty,$$

and

$$1 + \theta(1+\theta)\,O(n^{-1}(i+\theta)^{-1}) \quad \text{if } n-i = O(n) \quad \text{as } n \to \infty.$$

Unfortunately, these depart from 1 by terms which are larger than those established earlier for the accuracy of $\hat{r}_n(i)$, indeed by factors of n^2 and n respectively. It is therefore not surprising that $\hat{q}_n(i)$ does not, in general, agree as well with $q_n(i)$ as did $\hat{r}_n(i)$ with $r_n(i)$. We might judge the agreement as reasonable for $\theta \leqslant 0{\cdot}1$ and all i, but for larger θ, only for i close to n.

(iii) $\hat{q}_n(0)$ is certainly not a satisfactory approximation to $q_n(0)$, in general. We have

$$\hat{q}_n(0) = \theta \frac{\Gamma(n)\,\Gamma(1+\theta)}{\Gamma(n+1+\theta)} \sim \theta\Gamma(1+\theta)\,n^{-1-\theta},$$

which decreases as n increases, whereas $q_n(0)$ itself increases with n. This is most unfortunate, because one of the most interesting aspects of the whole problem is the asymptotic behaviour of $q_n(0)$.

(iv) It should be noted that (3.11) holds only for $i > 0$, because $p_n(0) = 0$. (3.11) gives an approximate solution to the problem posed by (1.3), namely the evaluation of the probability that the oldest allele in the sample *is* Eve's allele, given X_n.

(v) We see from Table 4 that, for fixed $w = i/n$, $0{\cdot}1 \leqslant w \leqslant 0{\cdot}9$, $nq_n(wn)$ has already come close to its limit when $n = 10$. There is little change when n increases to 100, except for the higher w values when θ is high. It might be noted, however, that when $\theta < 1$, $nq_n(n)$ diverges as n increases, in accordance with (3.4).

It is suggested by (3.10) that $n\hat{q}_n(i)$ might itself be approximated, as $n \to \infty$ with $w = i/n$, by

$$n\tilde{q}_n(i) = \theta w^{\theta}(1-w)^{\theta-1}, \quad 0 < w \leqslant 1, \tag{3.12}$$

an approximating, non-normalized, β density for $W = \lim Y_n/n$. This is dominated by the $\beta(1,\theta)$ density, as it should be, and suggests the approximation

$$f(w) \approx w^{\theta}, \quad 0 < w \leqslant 1,$$

in (3.7). The difference, between the integral of the density (3.12) and 1, would give us an approximation for the atom at 0:

$$\tilde{q}_n(0) = 1 - \int_0^1 \theta w^{\theta}(1-w)^{\theta-1}\,dw$$

$$= 1 - [\Gamma(1+\theta)]^2/\Gamma(1+2\theta), \tag{3.13}$$

$$\approx 1{\cdot}643\theta^2 \quad \text{when } \theta \ll 1. \tag{3.14}$$

The last approximation comes from 6.1.36 in Abramowitz & Stegun (1972), but it exceeds the approximate bound θ^2 in (3.6).

We might also hope that (3.12) would provide an approximation to the mean of Y_n, μ_n, by

$$\tilde{\mu}_n = n \int_0^1 \theta w^{1+\theta}(1-w)^{\theta-1}\,dw,$$

$$= \theta n \frac{\Gamma(2+\theta)\,\Gamma(\theta)}{\Gamma(2+2\theta)}, \tag{3.15}$$

and to the variance, σ_n^2 say, by

$$\tilde{\sigma}_n^2 = n^2 \int_0^1 \theta w^{2+\theta}(1-w)^{\theta-1}\,dw - \tilde{\mu}_n^2$$

$$= n^2\Gamma(1+\theta)\left[\frac{\Gamma(3+\theta)}{\Gamma(3+2\theta)} - \Gamma(1+\theta)\left(\frac{\Gamma(2+\theta)}{\Gamma(2+2\theta)}\right)^2\right]. \tag{3.16}$$

Clearly (3.15) does not equal (2.2), although both are $n(1-\theta)$ to first order terms in θ.

In Table 5 we show comparisons of the actual

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

Table 4. *Accuracy of approximating* $nq_n(i)$ *by* $n\hat{q}_n(i)$

Tabulated: $nq_n(i)$ above $n\hat{q}_n(i)$, for various values of $w = i/n$.

w									
0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1
				$n = 10, \quad \theta = 0{\cdot}01$					
0·0110	0·0124	0·0142	0·0165	0·0198	0·0247	0·0328	0·0490	0·0971	9·7218
0·0109	0·0123	0·0141	0·0164	0·0197	0·0246	0·0328	0·0490	0·0971	9·7218
				$n = 100, \quad \theta = 0{\cdot}01$					
0·0110	0·0124	0·0142	0·0165	0·0198	0·0247	0·0329	0·0491	0·0976	94·9620
0·0109	0·0123	0·0141	0·0164	0·0197	0·0247	0·0328	0·0491	0·0976	94·9620
				$n = 10, \quad \theta = 0{\cdot}1$					
0·1046	0·1166	0·1319	0·1520	0·1796	0·2201	0·2856	0·4105	0·7516	7·5914
0·0909	0·1060	0·1234	0·1452	0·1742	0·2160	0·2827	0·4089	0·7516	7·5914
				$n = 100, \quad \theta = 0{\cdot}1$					
0·1049	0·1171	0·1325	0·1529	0·1809	0·2221	0·2891	0·4185	0·7845	60·0533
0·0878	0·1043	0·1223	0·1446	0·1741	0·2167	0·2849	0·4156	0·7828	60·0533
				$n = 10, \quad \theta = 0{\cdot}5$					
0·4139	0·4459	0·4845	0·5322	0·5932	0·6747	0·7915	0·9784	1·3513	2·8377
0·2134	0·2824	0·3514	0·4258	0·5109	0·6150	0·7530	0·9601	1·3513	2·8377
				$n = 100, \quad \theta = 0{\cdot}5$					
0·4203	0·4534	0·4935	0·5433	0·6075	0·6944	0·8214	1·0326	1·4997	8·8734
0·1722	0·2536	0·3300	0·4102	0·5012	0·6127	0·7628	0·9959	1·4838	8·8734
				$n = 10, \quad \theta = 1$					
0·6281	0·6491	0·6723	0·6982	0·7274	0·7610	0·8003	0·8480	0·9091	1·0000
0·1818	0·2727	0·3636	0·4545	0·5455	0·6364	0·7273	0·8182	0·9091	1·0000
				$n = 100, \quad \theta = 1$					
0·6444	0·6658	0·6893	0·7154	0·7448	0·7782	0·8170	0·8632	0·9206	1·0000
0·1089	0·2079	0·3069	0·4059	0·5050	0·6040	0·7030	0·8020	0·9010	1·0000
				$n = 10, \quad \theta = 2$					
0·7538	0·7203	0·6828	0·6407	0·5928	0·5377	0·4735	0·3969	0·3030	0·1818
0·0826	0·1488	0·2204	0·2893	0·3471	0·3857	0·3967	0·3719	0·3030	0·1818
				$n = 100, \quad \theta = 2$					
0·7792	0·7363	0·6882	0·6336	0·5712	0·4989	0·4137	0·3115	0·1849	0·0198
0·0231	0·0719	0·1354	0·2019	0·2600	0·2981	0·3046	0·2681	0·1770	0·0198
				$n = 10, \quad \theta = 5$					
0·5068	0·3966	0·3025	0·2233	0·1582	0·1061	0·0659	0·0365	0·0167	0·0050
0·0071	0·0173	0·0307	0·0440	0·0528	0·0538	0·0461	0·0321	0·0167	0·0050
				$n = 100, \quad \theta = 5$					
0·4930	0·3527	0·2405	0·1539	0·0902	0·0465	0·0196	0·0057	0·0007	0·0000
0·0001	0·0012	0·0042	0·0087	0·0124	0·0126	0·0090	0·0039	0·0007	0·0000

Table 5. *Comparison of true and approximate* $q_n(0)$, μ_n, *and* σ_n^2 *values for samples of size* $n = 100$

θ	$q_{100}(0)$	$\tilde{q}_{100}(0)$	μ_{100}	$\tilde{\mu}_{100}$	σ^2_{100}	$\tilde{\sigma}^2_{100}$
0·01	0·0001	0·0002	99·02	99·00	48·73	49·64
0·1	0·0057	0·0142	90·70	90·36	423·88	460·45
0·5	0·0906	0·2146	62·92	58·90	1164·92	1438·96
1·0	0·2337	0·5000	41·60	33·33	1169·51	1388·89
2·0	0·4791	0·8333	20·09	10·00	697·77	566·67
5·0	0·8314	0·9960	3·51	0·22	106·16	12·58

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

atomic probability $q_n(0)$, the mean μ_n and the variance σ_n^2, with the approximations in (3.13), (3.15) and (3.16) respectively, using $n = 100$ as sample size. It is clear that $\tilde{q}_n(0)$ seriously over-estimates the actual probability of loss of Eve's allele, and it exceeds the bound $(\theta e^\theta - \theta)/(\theta e^\theta + 1)$ in Table 2. The mean and variance approximations are fairly reasonable (on a proportional basis) for $\theta \leqslant 1$, say, but are poor when $\theta = 5$. It would be preferable if more accurate approximations could be found.

4. Application to human mitochondrial DNA

It is natural to ask about the extent to which the results of the previous sections apply to aspects of human evolution, particularly in the case of (selectively neutral) haploid mitochondrial DNA. Interest in the most recent common ancestor, Eve, of human mitochondrial DNA was stimulated by Cann *et al.* (1987). We may ask, what proportion of the extant human population carries her genes? (Here, we think of a gene as being any sequence of bases in mitochondrial DNA; the fact that our model does not allow recombination within a gene now becomes a virtue!)

There are at least three aspects of the preceding analysis which are likely to be unrealistic for modelling human evolution: the infinite alleles assumption, and the (implicit) assumptions that the population size is constant over the period under consideration and that the population is panmictic rather than spatially structured. Nevertheless, we would argue that some broad conclusions may still be drawn.

The central feature of the infinite alleles assumption is that it excludes the possibility of back-mutation. If this is unrealistic in a particular case (for example, if attention is focused on a particular nucleotide site) then the analysis applies to the number, or proportion, of descendants who are *identical by descent* at that locus with the MRCA. In the context, this is of more interest than being just *identical in state*. But to study the latter concept, progress could be made by modifying the bivariate process $[N(t), Y(t)]$ of Section 2, so that its transitions, in (2.17), would also include a jump from state (n, i) to $(n, i+1)$ because of mutations back to the type of the common ancestor. Of course, the number of descendants with their common ancestor's allele is always bounded below by the number of individuals who are identical by descent with that ancestor. We do not pursue back mutation here.

Suppose now that the population size is not constant over the period back to the MRCA, but for the moment continue to assume random mating. Denote the present time by 0 and the population size t generations ago by M_t. (For human mitochondrial DNA the appropriate population consists only of human females. Note also that similar results apply to models with overlapping generations.) In the case of constant population size $M_t = M$, the genealogy of the sample converges to the coalescent provided we rescale the way we measure time: in this case the step from generation $t-1$ to t in the past contributes M^{-1} units of time in the new time scale. If the population size is variable (provided, as seems reasonable for human evolution since Eve, that the variability is not dependent on the gene frequencies at the locus in question) the same result is true (Kingman, 1982) if we adopt a new, non-linear, time scale in which the step from generation t to $t-1$ in the past contributes M_t^{-1} units of time in the new time scale. Thus if an event of interest occurred at time T in the coalescent, this would correspond to generation t' given by

$$t' = \min\left\{t: \sum_{k=1}^{t} M_k^{-1} > T\right\}. \tag{4.1}$$

Perhaps the most natural way to think about mutation in neutral models is first to choose a realization of the genealogical tree of the sample, according to the coalescent, and then to superimpose mutation onto this genealogical tree. In the infinite alleles case with constant population size, with $M_t = M$ as the number of haploids, we define $\theta = 2Mu$, where u is the mutation rate per gene per generation. Then we put mutations on each branch of the tree at the points of a Poisson process of rate $\theta/2$, with the processes on distinct branches being independent. But in the variable population size case, the Poisson processes of mutation are not time homogeneous on the coalescent branches. Denote the time to a common ancestor in the coalescent timescale by τ_c. Running forward from this point, through the genealogy, after a time $\tau\,(<\tau_c)$ (which corresponds to time $T = \tau_c - \tau$ in the evolution of the coalescent) the Poisson processes on each branch of the genealogy are running at rate $\theta_T/2$, where $\theta_T = 2M_{t'}u$ for t' given by (4.1), again independently for each branch.

An exact analysis, analogous to that given earlier, is thus substantially complicated by the effect of the new time scaling. Further, it depends crucially on knowledge of the size (or possibly effective size) of the population of human females throughout its evolution since Eve. And, were one to embark on such a task, it is not clear how, or to what extent, one should incorporate available knowledge about the time since our mitochondrial MRCA. Instead, we make a number of qualitative points.

Whatever the exact details, it is clear that for human populations the sequence $M_1, M_2, \ldots$ is decreasing, with a substantial change (of perhaps between three and five orders of magnitude) over the period back to the alleged mitochondrial Eve. This means that in converting from real time to the appropriate timescale for the coalescent, initially (small T) we greatly speed up time, but this speeding up factor decreases markedly as we go further into the past. Under the assumption that u, the mutation rate

https://doi.org/10.1017/S0016672300030974 Published online by Cambridge University Press

per gene per generation, is constant, this means that as we go forward through the genealogy from the MRCA, the rate of the mutation processes increases by the same substantial factor. One very crude approach then would be to bound aspects of the distribution of the proportion of the population who are identical by descent with the MRCA between two extreme cases, one for the smallest value of θ (presumably θ_{τ_c}) and one for the largest (θ_0).

Observe, however, that in the constant population size case, the fate of the MRCA's allele is determined by mutations very early on in the genealogy (i.e. close to the MRCA). This is evident from the dynamics of the bivariate process $[N(t), Y(t)]$ and for example from Tables 2 and 4 where aspects of the distribution do not change substantially as the genealogy grows from, say, size 4 or 10 to 100, even for large θ. This suggests that in the variable population size case one should approximate the distribution by using the value of θ appropriate early on in the genealogy. It is problematical as to the effective population size then, especially if allowance is made for overlapping generations and spatial subdivision. Takahata (1986) estimated that the common ancestral *species* for humans, rat, mouse and bovine had an effective population size of about 10^7 at about 75 million years ago. On the other hand, Cann *et al.* (1987) estimated that the human mitochondrial Eve may have lived only about 200000 years ago. They, and Wainscoat (1987), speculated that a population bottleneck may have occurred around the speciation time or around Eve's time. Cavalli-Sforza (1991) surveyed results from various studies of genetic and language evolution, which support the hypothesis of Eve existing in Africa some 100000–200000 years ago, although the issue remains contentious. For argument's sake, we consider the range 10^2 to 10^6 for the effective female population size around Eve's time.

The appropriate value of u will depend on the region of the genome and the number of bases in the gene in question. For a single nucleotide site one might take $u = 10^{-9}$ while for a gene of 1000 bases somewhere of the order of 10^{-6} may be more natural. A referee has suggested considerably higher u values, for primate mtDNA, of 6×10^{-7} per site per generation, i.e. 6×10^{-4} for a 1000-base gene. Depending on the exact choices, one might thus consider approximating by the constant population size results for θ in the range $2 \times 10^2 \times 10^{-9} = 2 \times 10^{-7}$ to $2 \times 10^6 \times 6 \times 10^{-7} = 1{\cdot}2$ for a site, and between $2 \times 10^2 \times 10^{-6} = 2 \times 10^{-4}$ to $2 \times 10^6 \times 6 \times 10^{-4} = 1200$ for a 1000-site gene. Our general point is that if small (or large) θ values were correct then, their use in our theory should give some idea of the current distribution of the proportion of alleles now, identical by descent with Eve's alleles, even at loci for which larger θ values would be appropriate for current evolution.

If we focus on small values of θ, then, by (3.6), θ^2 over-estimates the probability that Eve's allele (at a particular locus) would be missing from a population. For instance, for $\theta < 10^{-1}$ say, there is very little chance that Eve's allele *would* be lost from the population, and (3.12) is an approximation for the density of the proportion of genes identical by descent with Eve's gene. On the other hand, if large values of θ were appropriate, say $\theta > 50$, the lower bound in (3.5) shows that there would be a very high probability of Eve's allele being lost.

Of course, the evolution of real human populations since the MRCA is likely to have been affected by spatial population structure. Genealogy in the presence of population subdivision has recently been studied (e.g. Takahata, 1988; Takahata & Slatkin, 1990; Notohara, 1990; Hey, 1991) and although more complicated than the panmictic case, one could superimpose mutation and try to repeat the earlier analysis, before again taking account of variable population (and subpopulation) sizes. This is well beyond our current scope. More qualitatively, it should again be the case that the fate of the MRCA's allele depends on events early in the genealogy. The time for which the sample (or population) has exactly two ancestors (and hence the 'length' of the early genealogy) is more variable and also (in some sense) longer if the population is geographically structured. Thus such structure increases the chance that this early part of the genealogy will be quite long, in which case Eve's allele is more likely to be lost. While it is not entirely clear, it seems that population structure (with or without variability in population sizes) may well increase the probability that Eve's alleles are lost.

G.A.W. thanks the Science and Engineering Research Council, grant GR/F/94019, for providing a Visiting Fellowship at Queen Mary and Westfield College, University of London, and the College for its kind hospitality. P.D. was supported in part by SERC Advanced Fellowship B/AF/1255 and grants GR/F 32561 and GR/F 98727. We thank Bob Griffiths for suggesting this problem, and for some helpful comments, and a referee for suggesting possible values for primate mutation rates.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.

Beder, B. (1988). Allelic frequencies given the sample's common ancestral type. *Theoretical Population Biology* **33**, 126–137.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31–36.

Cavalli-Sforza, L. L. (1991). Genes, peoples and languages. *Scientific American* (November), 72–78.

Donnelly, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theoretical Population Biology* **30**, 271–288.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, vol. 1, 3rd edn. New York: J. Wiley.

Griffiths, R. C. (1986). Family trees and DNA sequences. In *Proceedings of the Pacific Statistical Congress* (ed. I. S. Francis, B. F. J. Manly and F. C. Lam), pp. 225–227. Elsevier Science Publishers.

Griffiths, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-sites model. *Journal of Mathematical Biology* **27**, 667–680.

Hey, J. (1991). A multi-dimensional coalescent process applied to multiallelic selection models and migration models. *Theoretical Population Biology* **39**, 30–48.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology* **25**, 123–159.

Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. New York: J. Wiley

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* **29**, 59–75.

Saunders, I. W., Tavaré, S. & Watterson, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471–491.

Shimizu, A. (1987). Stationary distribution of a diffusion process taking values in probability distributions on the partitions. In *Stochastic Models in Biology* (ed. M. Kimura, G. Kallianpur and T. Hida). Berlin/New York: Springer-Verlag.

Takahata, N. (1986). An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. *Genetical Research, Cambridge* **48**, 187–190.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research, Cambridge* **52**, 213–222.

Takahata, N. & Slatkin, M. (1990). Genealogy of neutral genes in two spatially isolated populations. *Theoretical Population Biology* **38**, 331–350.

Tavaré, S. (1984). Line-of-descent and genealogical processes and their applications in population genetics. *Theoretical Population Biology* **26**, 119–164.

Wainscoat, J. (1987). Out of the garden of Eden. *Nature* **325**, 13.

Watterson, G. A. (1984). Lines of descent and the coalescent. *Theoretical Population Biology* **26**, 77–92.

Watterson, G. A. (1989). Allele frequencies in multigene families. I. Diffusion equation approach. *Theoretical Population Biology* **35**, 142–160.