Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles

Robert C. Griffiths [a] Sabin Lessard [b]

[a] *Department of Statistics*
University of Oxford
1 South Parks Rd,
Oxford OX1 3TG, UK
email: griff@stats.ox.ac.uk

[b] *Département de mathématiques et de statistique*
Université de Montréal
C.P. 6128
Succursale Centre-ville
Montréal (Québec) H3C 3J7, Canada
email: lessards@dms.umontreal.ca

Abstract

Ewens' sampling formula, the probability distribution of a configuration of alleles in a sample of genes under the infinitely-many-alleles model of mutation, is proved by a direct combinatorial argument. The distribution is extended to a model where the population size may vary back in time. The distribution of age-ordered frequencies in the population is also derived in the model, extending the GEM distribution of age-ordered frequencies in a constant sized population model. The genealogy of a rare allele is studied using a combinatorial approach.

A connection is explored between the distribution of age-ordered frequencies and ladder indices and heights in an urn model, corresponding to the sample; and in a sequence of independent uniform random variables corresponding to the population.

Key words: Age distribution of alleles, Coalescent process, Ewens' sampling formula, GEM distribution, Infinitely-many-alleles model, Ladder indices and heights, Poisson Dirichlet process, Urn model.

Paper version 2.3

1 Introduction

Ewens' (1972) sampling formula (ESF) is the probability distribution of the number of different types of genes and their frequencies at a selectively neutral locus under the infinitely-many-alleles model of mutation. The coalescent process of Kingman (1982) describing the genealogy of a sample underlies the sampling distribution. Kingman (1978) relates the sampling distribution to partition structures. The population model under which the ESF holds can be described as a diffusion process which contains as its limit domain of attraction, when time is scaled appropriately, the Wright-Fisher model, the Moran model, and Cannings' (1974) exchangeable model generalizing the Wright-Fisher model. Earlier papers viewed the ESF as an approximate sampling formula in the Wright-Fisher model. The population gene frequencies are modelled as a diffusion process by Ethier and Kurtz (1981), and as a genealogical process by Griffiths (1980) and Donnelly and Tavaré (1987). Joyce and Tavaré (1987) relate the genealogical process to a linear pure birth and immigration process. Applied interest is that in the ESF the number of types is a sufficient statistic for the mutation rate. The stationary distribution of the population gene frequencies in the diffusion process model is the Poisson Dirichlet process. Donnelly and Kurtz (1986) study a particle process that relates the population frequencies, modelled as a measure valued diffusion process, to the coalescent process.

The distribution of non-mutant lines of descent from a given time in the past to the present time is studied by Griffiths (1980), Watterson (1984), Tavaré (1984) and Donnelly and Tavaré (1986), giving a generalization of the ESF to the distribution of allele types before and after the given time in the past. The distribution of age-ordered alleles in the ESF is derived in Donnelly and Tavaré (1986).

The age distribution of a mutation known to be of a given frequency in a population was first derived in a classic paper by Kimura and Ohta (1973). Recent papers studying the age distribution use a coalescent approach, a diffusion approach, or a combination of the two. (Slatkin and Rannala, 1997; Rannala and Slatkin, 1998; Griffiths and Tavaré, 1998; Wiuf and Donnelly, 1999; Stephens, 2000; Wiuf, 2000, 2001; Griffiths, 2003; Griffiths and Tavaré, 2003).

In this paper we present an elementary proof of the ESF based on combinatorial arguments in the framework of the coalescent process. The approach is used to extend the sampling formula and the Poisson Dirichlet distribution in the population to the case of a variable population size. The age-ordered distribution of gene frequencies in a sample and the population is also found in this case, extending known results.

The genealogy and age of a rare mutant type considered in Wiuf (2000, 2001) is studied by the combinatorial approach in this paper.

A connection is explored between the distribution of age-ordered frequencies and ladder indices and heights in an urn model, corresponding to the sample; and in a sequence of independent uniform random variables corresponding to the population.

2 Ewens' sampling formula: a combinatorial derivation

The ancestry of a random sample of n genes is described back in time by a coalescent tree, with vertices where lineages have a common ancestor (Kingman, 1982). Mutations occur along the edges of the coalescent tree according to a Poisson process of intensity $\theta/2$. Many discrete population models are in the domain of attraction of the coalescent when time is measured appropriately. In a Wright-Fisher model of constant size N with an infinitely-many-alleles model of mutation, novel mutant types are formed at a rate of u per gene per generation. The distribution of the ancestral tree of a sample of n genes converges to the coalescent when time is measured in units of N generations, where $\theta = 2Nu$ is the scaled mutation rate per gene per generation and $N \to \infty$. The number of non-mutant ancestors of a sample of n genes is a death process back in time, where ancestral lines are lost by either mutation or coalescence. Griffiths (1980) and Tavaré (1984) study this death process in the entire population and in a sample of n genes. Ewens (1990) calls the events in the death process defining events. Lines lost by mutation determine the family tree of the mutant in the sample with the number of members of the family as the number of leaves subtended by the mutation. Label the sample genes and list them in the order in which they are lost backward in time, following either a mutation or a coalescence. In the case of coalescence, one of the two genes involved is chosen at random to continue back in time as a parent gene and the other gene is lost. There are $n!$ different ordered loss lists. If there are k different types of genes represented in the sample with n_l genes of type l for $l = 1, \ldots k$, then there are $n!/[n_1! \cdots n_k!]$ possibilities for the positions of the genes of the different types in the list. However if the types that have the same number of genes in the sample are not labelled, then this number of possibilities is divided by $[b_1! \cdots b_n!]$, where b_j is the number of types represented j times in the sample for $j = 1, \ldots, n$ such that $\sum_j b_j = k$.

Now consider the probability of each particular sequence of events. When i genes remain the rate of mutation is $i\theta/2$ and the rate of coalescence is $i(i-1)/2$. The probability that a particular gene is the next one lost and that it is lost by mutation is $\theta/[i(\theta + i - 1)]$. Similarly the probability that a particular gene of a given type is the next one lost and that it is lost by

coalescence is $(j-1)/[i(\theta+i-1)]$, where j is the number of genes of the given type among the i remaining genes. Then it is clear that

$$n! \cdot \left(\frac{n!}{n_1! \cdots n_k!}\right) \cdot \frac{1}{b_1! \cdots b_n!} \cdot \frac{(n_1-1)! \cdot \theta \cdots (n_k-1)! \cdot \theta}{1 \cdot \theta \cdots n \cdot (\theta+n-1)}$$
$$= \frac{n!}{1^{b_1} \cdots n^{b_n}} \cdot \frac{1}{b_1! \cdots b_n!} \cdot \frac{\theta^k}{\theta \cdots (\theta+n-1)} \tag{1}$$

is the probability of having k types of genes with b_j types represented j times for $j = 1, \ldots, n$ in a sample of size $n = \sum_j j b_j$. This is the sampling formula conjectured by Ewens (1972) and proved by induction by Karlin and McGregor (1972). Hoppe (1984) derived the ESF from an urn model representation of sampling genes.

Notice that, if the n genes are labelled, then the probability that n_l given genes among these are of type l for $l = 1, \ldots, k$ is

$$\frac{n! \theta^k \prod_{l=1}^k (n_l-1)!}{\prod_{i=1}^n i(\theta+i-1)}.$$

Moreover, if the n sampled genes are labelled and the ancestry is traced back up to the point of m ancestral genes of given types, say types $1, \ldots, m$, then it suffices to choose these m ancestral genes and proceed as above for the others to find

$$\frac{(n-m)! \theta^{k-m} \prod_{l=1}^m n_l! \prod_{l=m+1}^k (n_l-1)!}{\prod_{i=m+1}^n i(\theta+i-1)} \tag{2}$$

for the probability of having n_l given genes of type l for $l = 1, \ldots, k$, types $1, \ldots, m$ being ancestral and types $m+1, \ldots, k$ being mutant. This is the formula originally given by Watterson (1984), which also extends Kingman's (1982) formula for the case $\theta = 0$.

3 Variable population size

If the population size is variable, the rate of mutation of i genes at time back t is $i\theta/2$ and the rate of coalescence is $i(i-1)/[2\lambda(t)]$, where $\lambda(t) = N(t)/N(0)$ and t is expressed in units of $N(0)$ generations. Starting with n genes at the current time 0 and measuring time backward, the genes are lost by mutation or coalescence at random times $T_n < T_{n-1} < \cdots < T_1$ according to a non-homogeneous death process of rate $i[\theta+(i-1)\lambda(t)^{-1}]/2$, where i is the number of remaining genes at time back t. At time $T_i = t$, the probability that a particular gene is lost by mutation is $[\theta\lambda(t)]/[i(\theta\lambda(t)+i-1)]$ and by coalescence

$(j-1)/[i(\theta\lambda(t)+i-1)]$, where j is the number of remaining ancestral lineages of the particular gene type in the sample.

Assume n distinct genes at time 0 and allocate each of them a type such that n_l are of type l for $l = 1, \ldots, k$. There are $n!/[n_1! \cdots n_k!]$ possibilities. The genes are lost in order and there are $n!$ cases in all to consider. Decompose these cases by looking at the position of the last gene of each type defined as the number of remaining genes the last time there remains one gene of the given type. The outcome will be an ordered sequence $1 = i_{l_1} < \ldots < i_{l_k} \leq n$, where i_{l_m} is the number of genes remaining just before the last gene of type l_m is lost, this type being the m-th oldest, for $m = 1, \ldots, k$, and $(l_1, \ldots, l_k)$ being a permutation of $(1, \ldots, k)$. This sequence configuration is possible if and only if the inequalities $i_{l_m} \leq \sum_{\nu=1}^{m-1} n_{l_\nu} + 1$ hold for $m = 1, \ldots, k$. The number of arrangements of the n genes satisfying these conditions is, starting from the last gene lost and ending with the first one,

$$\prod_{m=1}^{k} n_{l_m} \cdot \frac{(\sum_{\nu=1}^{m} n_{l_\nu} - i_{l_m})!}{\left(\sum_{\nu=1}^{m} n_{l_\nu} - i_{l_{m+1}} + 1\right)!}, \tag{3}$$

with the convention that $i_{l_{k+1}} = n$, since there are n_{l_m} possible genes for the last one of type l_m to be lost and $(\sum_{\nu=1}^{m} n_{l_\nu} - i_{l_m})$ possible genes for the ones lost between the last one of type l_m and the last one of type l_{m+1}, for $m = 1, \ldots, k$. The number of arrangements (3) can also be written as

$$\prod_{m=1}^{k} (n_{l_m}!) \cdot \begin{pmatrix} \sum_{\nu=1}^{m} n_{l_\nu} - i_{l_m} \\ n_{l_m} - 1 \end{pmatrix}.$$

The probability of each such sequence is

$$\mathbb{E}\left\{\frac{\theta^k \prod_{l=1}^{k}[(n_l - 1)!\lambda(T_{i_l})]}{n! \prod_{i=1}^{n}[\theta\lambda(T_i) + i - 1]}\right\}.$$

Finally, if the types with the same number of genes are not labelled, the probability of having k types of genes with b_j types represented j times for $j = 1, \ldots, n$ is

$$\frac{n! \cdot \theta^{k-1}}{\left(\prod_{l=1}^{k} n_l\right) \cdot \left(\prod_{j=1}^{n} b_j!\right)} \sum_{\mathbf{i},\mathbf{l}} a_{\mathbf{i},\mathbf{l}} \mathbb{E}\left\{\frac{\prod_{l=2}^{k} \lambda(T_{i_l})}{\prod_{i=2}^{n}[\theta\lambda(T_i) + i - 1]}\right\}, \tag{4}$$

where

$$a_{\mathbf{i},\mathbf{l}} = \frac{\prod_{m=1}^{k} \begin{pmatrix} \sum_{\nu=1}^{m} n_{l_\nu} - i_{l_m} \\ n_{l_m} - 1 \end{pmatrix}}{\begin{pmatrix} n \\ n_1, \ldots, n_k \end{pmatrix}},$$

with $\mathbf{i} = (i_1, \ldots, i_k)$ satisfying $1 = i_{l_1} < \cdots < i_{l_k} \leq n$ and $i_m \leq \sum_{l=1}^{m-1} n_l + 1$, for $m = 1, \ldots, k$, and $\mathbf{l} = (l_1, \ldots, l_k)$ being a permutation of $(1, \ldots, k)$. Note that

$$\sum_{\mathbf{i}} a_{\mathbf{i},\mathbf{l}} = \prod_{m=1}^{k} \left\{ \frac{n_{l_m}}{\sum_{\nu=m}^{k} n_{l_\nu}} \right\}, \tag{5}$$

which is the probability that type l_m is the m-th oldest, for $m = 1, \ldots, k$. The product on the right side of (5) is obtained by conditioning on the older types in sequential order from $m = 1$ to $m = k$. The sum on the left side of (5) is obtained by partitioning the event according to the positions of the last genes of the k different types then using the above number of possible arrangements of all the genes for each case divided by $\prod_{m=1}^{k}(n_{l_m}!)$, in order not to distinguish genes within types. Also $\sum_{\mathbf{i},\mathbf{l}} a_{\mathbf{i},\mathbf{l}} = 1$. Notice that

$$\frac{n! \cdot \theta^{k-1}}{\prod_{l=1}^{k} n_l} \sum_{\mathbf{i}} a_{\mathbf{i},\mathbf{l}} \mathbb{E} \left\{ \frac{\prod_{l=2}^{k} \lambda(T_{i_l})}{\prod_{i=2}^{n} [\theta\lambda(T_i) + i - 1]} \right\}, \tag{6}$$

is the probability of having n_{l_m} genes of type l_m, this type being the m-th oldest, for $m = 1, \ldots, k$. In the case of a constant population size, taking $l_k = k, \ldots, l_1 = 1$, without loss of generality, this probability reduces to

$$\frac{(n-1)!}{n_k \cdot (n_k + n_{k-1}) \cdots (n_k + \cdots + n_2)} \cdot \frac{\theta^k}{\theta \cdots (\theta + n - 1)}, \tag{7}$$

which is the formula for the age-ordered types given by Donnelly and Tavaré (1986).

4 Ladder indices and heights in an urn model

There is a representation of the conditional distribution of the partial sums $\sum_{\nu=1}^{m} n_\nu$, $m = 1, \ldots, k$, given $\mathbf{i}$ as the distribution of ladder heights given ladder indices $\mathbf{i}$ in an urn model.

The probability of a particular sequence $\mathbf{i}$ is

$$\begin{aligned} &\mathbb{E} \left\{ \frac{\prod_{c \notin \mathbf{i}} (c-1) \prod_{d \in \mathbf{i}} \theta\lambda(T_d)}{\prod_{i=1}^{n} [\theta\lambda(T_i) + i - 1]} \right\} \\ &= \frac{(n-1)!}{\prod_{l=2}^{k} (i_l - 1)} \mathbb{E} \left\{ \frac{\prod_{l=1}^{k} \theta\lambda(T_{i_l})}{\prod_{i=1}^{n} [\theta\lambda(T_i) + i - 1]} \right\}. \end{aligned} \tag{8}$$

The probability of an age-ordered configuration conditional on $\mathbf{i}$ from (6) and (8) is

$$h_{\mathbf{i}}(\mathbf{n}) = \frac{n\prod_{l=2}^{k}(i_l-1)}{\prod_{l=1}^{k} n_l}\cdot a_{\mathbf{i},\mathbf{l}}$$
$$= \frac{\prod_{l=2}^{k}(i_l-1)}{(n-1)!}\cdot\prod_{m=1}^{k}\frac{\left(\sum_{\nu=1}^{m} n_\nu - i_m\right)!}{\left(\sum_{\nu=1}^{m-1} n_\nu - i_m + 1\right)!}. \tag{9}$$

By convention the factorial term in the denominator is taken as 1 when $m = 1$. Note that $\sum h_{\mathbf{i}}(\mathbf{n}) = 1$, where summation is over $\sum_{l=1}^{k} n_l = n$ and $\sum_{l=1}^{m-1} n_l \geq i_m - 1$, $m = 2, \ldots, k$.

Label n balls in an urn $1, 2, \ldots, n$ and draw out balls at random sequentially. Let $\mathbf{i} = (i_1, \ldots, i_k)$ and $\mathbf{b} = (b_1, \ldots b_k)$ be ladder indices and ladder heights where successive maxima occur in the numbers on the balls which are drawn. The last index k is defined such that $b_k = n$. The probability of a particular initial pair $i_1 = 1, b_1$ is n^{-1}. Then for $m > 1$ the probability of obtaining i_m, b_m, conditional on $(i_1, \ldots, i_{m-1})$ and $(b_1, \ldots, b_{m-1})$ is

$$\frac{b_{m-1} - i_{m-1}}{n - i_{m-1}} \cdots \frac{b_{m-1} - i_m + 2}{n - i_m + 2} \cdot \frac{1}{n - i_m + 1}$$
$$= \frac{(b_{m-1} - i_{m-1})!}{(b_{m-1} - i_m + 1)!} \prod_{j=i_{m-1}}^{i_m - 1} \frac{1}{n-j}. \tag{10}$$

The joint probability of a configuration $\mathbf{i}, \mathbf{b}$ is therefore

$$P(\mathbf{i}, \mathbf{b}) = \frac{(n - i_k)!}{n!} \cdot \prod_{m=2}^{k} \frac{(b_{m-1} - i_{m-1})!}{(b_{m-1} - i_m + 1)!}$$
$$= \frac{1}{n!} \cdot \prod_{m=1}^{k} \frac{(b_m - i_m)!}{(b_{m-1} - i_m + 1)!}. \tag{11}$$

The distribution of the age-ordered frequencies $\mathbf{n} = (n_1, \ldots n_k)$ is identical to the ladder height distribution with $b_m = \sum_{\nu=1}^{m} n_\nu$. This follows because $h_{\mathbf{i}}(\mathbf{n})$ is proportional to $P(\mathbf{i}, \mathbf{b})$. Also comparing the two distributions the marginal ladder index distribution is

$$P(\mathbf{i}) = \frac{1}{n\prod_{m=2}^{k}(i_m - 1)}. \tag{12}$$

5 Population frequencies

5.1 GEM distribution

In a constant sized population model where $\lambda(t) = 1$, $t \geq 0$, the age-ordered frequencies $X_1, X_2, \ldots$ have a GEM distribution

$$Z_1, Z_2(1 - Z_1), Z_3(1 - Z_2)(1 - Z_1), \ldots, \tag{13}$$

where $\Big\{Z_i,\ i \geq 1\Big\}$ are mutually independent identically distributed (*i.i.d.*) random variables with density

$$\theta(1 - z)^{\theta - 1},\ 0 < z < 1,$$

(Donnelly and Tavaré, 1986, Ewens, 1990). This is a random partition representation (see Pitman (1996) and references therein). The unordered frequencies are distributed as a Poisson Dirichlet point process (Kingman, 1978).

The population analogue of (2) is derived in Griffiths (1980), Donnelly and Tavaré (1986).

Let $(n_1, \ldots, n_k)$ be a sample taken from the GEM distribution arranged in age order and $q(n_1, \ldots, n_k)$ the age-ordered distribution. It is known that the distribution is (7), however we give a short proof for completeness. Considering whether the oldest type in the sample is the oldest type in the population or not

$$\begin{aligned} q(n_1, \ldots, n_k) = &\binom{n}{n_1} \cdot \mathbb{E}\Big(Z_1^{n_1}(1 - Z_1)^{n - n_1}\Big) \cdot q(n_2, \ldots, n_k) \\ &+ \mathbb{E}\Big((1 - Z_1)^n\Big) \cdot q(n_1, \ldots, n_k) \\ = &\binom{n}{n_1} \cdot \frac{\theta\Gamma(n_1 + 1)\Gamma(n - n_1 + \theta)}{\Gamma(n + \theta + 1)} \cdot q(n_2, \ldots, n_k) \\ &+ \frac{\theta}{n + \theta} \cdot q(n_1, \ldots, n_k), \end{aligned} \tag{14}$$

where $q(n_2, \ldots, n_k)$ is intrepreted as 1 if $k = 1$. Simplifying (14)

$$q(n_1, \ldots, n_k) = \frac{\theta}{(n - n_1)} \cdot \frac{(n - 1)!}{(n - n_1 - 1)!} \cdot \frac{\Gamma(n - n_1 + \theta)}{\Gamma(n + \theta)} \cdot q(n_2, \ldots, n_k),$$

and (7) follows by recurrence. The distribution of the age-ordered relative frequencies $\mathbf{n}/n$ in a constant sized population model, (7), converges to the GEM distribution as $n \to \infty$ because of the fact that it is a sample distribution from the GEM distribution.

5.2 Variable population size

It is of interest to find the population distribution of age-ordered frequencies in a variable sized population model. This extends the Poisson Dirichet and GEM distributions. The limit distribution for the age-ordered relative frequencies conditional on $\mathbf{i}$ is intrepreted as the population distribution. This distribution can be described in terms of: (i) the event times $\{T_j\,, j \geq 1\}$; (ii) a stochastic sequence $\mathbf{i}$ generated by a mixture of Bernoulli trials $\{\chi_j,\, j \geq 1\}$ conditionally independent given $\{T_j,\, j \geq 1\}$ such that $P(\chi_j = 1|T_j) = \theta\lambda(T_j)/[\theta\lambda(T_j) + j - 1]$, $P(\chi_j = 0|T_j) = 1 - P(\chi_j = 1|T_j)$; and (iii) the age-ordered frequencies, conditional on $\mathbf{i}$.

The sequence $\{T_j, j \geq 1\}$ is a reverse Markov chain, with transition distributions

$$\begin{aligned} P(T_j > t \mid T_{j+1} = s) &= \exp\left\{-\frac{\theta}{2}(t-s)j - \binom{j}{2}\int_s^t \frac{du}{\lambda(u)}\right\} \\ &= \exp\left\{-\frac{j\theta}{2}\int_s^t \frac{du}{p_j(u)}\right\}, \end{aligned} \tag{15}$$

where $t > s$ and $p_j(u) = \theta\lambda(u)/[\theta\lambda(u)+j-1]$. In the limit there is an entrance boundary at infinity in the process. It follows from (15) and

$$P(\chi_j = 1|T_j = t) = p_j(t),$$

that

$$P(\chi_j = 1, T_j \in (t, t+dt) \mid T_j > t) = \frac{j\theta}{2}dt + o(dt).$$

The asymptotic form for $n^{k-1}h_{\mathbf{i}}(\mathbf{n})$ as $n \to \infty$, with $n_i/n \to x_i$ for $i = 1, \ldots, k$, is

$$\begin{aligned} n^{k-1}h_{\mathbf{i}}(\mathbf{n}) &= \prod_{m=2}^{k}(i_m - 1)\cdot n^{i_m-1}\frac{\Big(\sum_{\nu=1}^{m} n_\nu - i_m\Big)!}{\Big(\sum_{\nu=1}^{m} n_\nu - 1\Big)!} \\ &\qquad \cdot n^{-(i_m-2)}\frac{\Big(\sum_{\nu=1}^{m-1} n_\nu - 1\Big)!}{\Big(\sum_{\nu=1}^{m-1} n_\nu - i_m + 1\Big)!} \\ &\sim \prod_{m=2}^{k}(i_m - 1)\Big(\sum_{\nu=1}^{m} x_\nu\Big)^{-(i_m-1)}\Big(\sum_{\nu=1}^{m-1} x_\nu\Big)^{i_m-2} \\ &= \prod_{l=2}^{k}(i_l - 1)\cdot\prod_{l=1}^{k-1}\Big(\sum_{j=1}^{l} x_j\Big)^{i_{l+1}-i_l-1}. \end{aligned} \tag{16}$$

Note that $(n_1 - i_1)! = (n_1 - 1)!$ and $(\sum_{l=1}^{k} n_l - 1)! = (n-1)!$ in simplifying the first line of (16).

The distribution in (iii) is related to the distribution of ladder heights and indices in a sequence of *i.i.d.* uniform random variables on [0,1] as shown in the next subsection.

5.3 Ladder indices and heights in a sequence of uniform random variables

Let $\{U_l,\, l \geq 1\}$ be a sequence of *i.i.d.* uniform random variables on $[0,1]$ and $\{S_m,\, m \geq 1\}$ the successive maxima which occur in the sequence $\{U_l,\, l \geq 1\}$ at random indices $\mathbf{i}$ (with $i_1 = 1$) such that $S_m = U_{i_m}$, $m \geq 1$. Then the joint probability of the first k indices of successive maxima $(i_1, \ldots, i_k)$ and density of these maxima $\{S_m,\, 1 \leq m \leq k\}$ is, by direct argument,

$$\prod_{m=1}^{k-1} s_m^{i_{m+1}-i_m-1}. \tag{17}$$

The marginal probability of obtaining the indices is

$$\int \prod_{m=1}^{k-1} s_m^{i_{m+1}-i_m-1} \prod_{m=1}^{k} ds_m = \frac{1}{i_k} \prod_{m=2}^{k} \frac{1}{i_m - 1}, \tag{18}$$

where integration is over $0 < s_1 < s_2 < \cdots < s_k < 1$. The conditional distribution of $\{S_m,\, 1 \leq m \leq k\}$ given the indices is thus

$$i_k \prod_{m=1}^{k-1} (i_{m+1} - 1) s_m^{i_{m+1}-i_m-1}. \tag{19}$$

Rescaling

$$\eta_m = \frac{S_m}{S_k},\ 1 \leq m < k,$$

the joint density of $\{\eta_m,\, 1 \leq m < k\}$ and S_k is

$$\begin{aligned} & i_k \prod_{m=1}^{k-1} (i_{m+1} - 1) \eta_m^{i_{m+1}-i_m-1} \cdot s_k^{\sum_{\nu=1}^{k-1}(i_{\nu+1}-i_\nu-1)} \cdot s_k^{k-1} \\ & \quad = i_k \prod_{m=1}^{k-1} (i_{m+1} - 1) \eta_m^{i_{m+1}-i_m-1} \cdot s_k^{i_k-1}. \end{aligned} \tag{20}$$

The Jacobian of the transformation is s_k^{k-1}. Integrating with respect to $0 < s_k < 1$, the density of $\Big\{\eta_m,\ 1 \le m < k\Big\}$ is

$$\prod_{m=1}^{k-1} (i_{m+1} - 1)\eta_m^{i_{m+1}-i_m-1}, \tag{21}$$

which is identical to the density (16) of the partial sums $\{\sum_{\nu=1}^{m} X_\nu,\ 1 \le m < k\}$. In the limit as $n \to \infty$, $k \to \infty$ and $S_k \to 1$, so it follows that the distribution of $\{\sum_{\nu=1}^{m} X_\nu,\ m \ge 1\}$ given $\mathbf{i}$ is identical to the distribution of ladder heights in $\{U_l,\ l \ge 1\}$ given that they occur at ladder indices $\mathbf{i}$.

The distribution (16) is simplified by making a transformation to independent exponential random variables

$$\mathbf{X} = (X_1, \ldots, X_{k-1}) \to \mathbf{V} = (V_1, \ldots, V_{k-1}),$$

where

$$\sum_{l=1}^{m} x_l = \exp\Big\{ - \sum_{l=m}^{k-1} v_l \Big\}. \tag{22}$$

The Jacobian of the transformation is

$$\prod_{m=1}^{k-1} \exp\Big\{ - \sum_{l=m}^{k-1} v_l \Big\}, \tag{23}$$

and making the transformation in (16), the density of $\mathbf{V}$ over $\mathbb{R}_+^{k-1}$ is

$$\prod_{m=1}^{k-1} (i_{m+1} - 1) \exp\Big\{ - (i_{m+1} - 1)v_m \Big\}. \tag{24}$$

That is, $V_1, \ldots, V_{k-1}$ are *i.i.d.* exponential random variables with rates $i_2 - 1, \ldots, i_k - 1$. In the limit there is an infinite number of types, so the age-ordered population frequencies have a representation

$$\begin{aligned} X_1 &= e^{-\sum_{l=1}^{\infty} V_l}, \\ X_m &= \sum_{l=1}^{m} X_l - \sum_{l=1}^{m-1} X_l \\ &= e^{-\sum_{l=m}^{\infty} V_l} \Big(1 - e^{-V_{m-1}} \Big),\ m \ge 2. \end{aligned} \tag{25}$$

An equivalent representation to (25) is that, for $m \ge 1$,

$$-\log\left(\sum_{l=1}^{m} X_l \right) = \sum_{j=i_{m+1}}^{\infty} (j-1)^{-1} \chi_j W_j, \tag{26}$$

where $\{W_j, j > 1\}$ is a sequence of *i.i.d.* exponential random variables with rate parameters unity. A third representation of (25) as a random partition is

$$X_m = \xi_{m-1} \prod_{l=m}^{\infty} (1 - \xi_l),\ m \geq 1, \tag{27}$$

where $\{\xi_l, l \geq 0\}$ are mutually independent random variables, with $\xi_0 = 1$, and for $m \geq 2$, ξ_m having a density of

$$(i_{m+1} - 1)(1 - z)^{i_{m+1}-2},\ 0 < z < 1.$$

Equation (27) is obtained by setting, for $l > 1$, $\xi_l = 1 - e^{-V_l}$.

The mean values of the age-ordered frequencies, conditional on $\mathbf{i}$, from (25) are for $m \geq 1$

$$\mathbb{E}(X_m \mid \mathbf{i}) = \frac{1}{i_m} \prod_{l=m+1}^{\infty} \Big(1 - \frac{1}{i_l}\Big). \tag{28}$$

The unconditional mean frequencies can be partially found. We have

$$\begin{aligned}
\mathbb{E}(X_1) &= \mathbb{E}\Big[\mathbb{E}\Big(X_1 \mid \mathbf{i}\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{l=2}^{\infty} \Big(1 - \frac{1}{i_l}\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{j=2}^{\infty} \Big(1 - \frac{1}{j}\chi_j\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{j=2}^{\infty} \Big(1 - \frac{\theta\lambda(T_j)}{j(\theta\lambda(T_j) + j - 1)}\Big)\Big].
\end{aligned} \tag{29}$$

A similar calculation gives that for $m \geq 1$

$$\mathbb{E}(X_m) = \mathbb{E}\Big[\frac{1}{i_m} \prod_{j=i_m+1}^{\infty} \Big(1 - \frac{\theta\lambda(T_j)}{j(\theta\lambda(T_j) + j - 1)}\Big)\Big]. \tag{30}$$

An alternative expression to (30) is

$$\mathbb{E}(X_m) = \mathbb{E}\Big[\frac{1}{\theta\lambda(T_{i_m+1}) + i_m} \prod_{l=i_m+1}^{\infty} \frac{\theta\lambda(T_l) + l}{\theta\lambda(T_{l+1}) + l}\Big]. \tag{31}$$

Equation (31) is found by simplifying terms in the product of (30) and shifting the product index in the denominator by unity.

In the usual constant population size case when $\lambda(t) = 1$, $t > 0$, we have

$$\mathbb{E}(X_m) = \mathbb{E}\Big[\frac{1}{\theta + i_m}\Big]. \tag{32}$$

Remark. Convergence of the product in (31) needs justification. Let $\{\tau_l, l \geq 1\}$ be independent exponential random variables with rates $\{l(l+\theta-1)/2, l \geq 1\}$. In the constant sized population case (with notation T_l°), $T_l^\circ = \sum_{k=l}^\infty \tau_k$. As $l \to \infty$, $T_l^\circ, T_l \to 0$, and $T_l \sim T_l^\circ$ because $\lambda(0) = 1$. We assume here that $\lambda(t)$ is continuous at $t = 0$ and $|\lambda'(0)| < \infty$. As $l \to \infty$ the general term of the product satisfies

$$\begin{aligned}
\frac{\theta\lambda(T_l) + l}{\theta\lambda(T_{l+1}) + l} &\approx \frac{\theta + l + 1 + T_l^\circ \lambda'(0)}{\theta + l + 1 + T_{l+1}^\circ \lambda'(0)} \\
&= \frac{\theta + l + 1 + (\tau_l + T_{l+1}^\circ)\lambda'(0)}{\theta + l + 1 + T_{l+1}^\circ \lambda'(0)} \\
&\approx 1 + \frac{\tau_l \lambda'(0)}{l} \\
&\approx 1 + \frac{2Y_l \lambda'(0)}{l^3}, \qquad (33)
\end{aligned}$$

where $\{Y_l, l \geq 1\}$ are *i.i.d.* exponential random variables with unit rates. The product converges because of the cubic term in the denominator in (33).

5.4 Laplace transforms

The Laplace transform of $-\log(X_1)$, conditional on $\{T_l, l > 1\}$, is

$$\begin{aligned}
\mathbb{E}\Big[e^{\phi \log(X_1)}\Big] &= \mathbb{E}\Big[X_1^\phi\Big] \\
&= \mathbb{E}\Big[\prod_{l=2}^\infty \frac{i_l - 1}{i_l - 1 + \phi}\Big] \\
&= \mathbb{E}\Big[\prod_{j=2}^\infty \Big(1 - \frac{\phi\chi_j}{j - 1 + \phi\chi_j}\Big)\Big] \\
&= \prod_{j=2}^\infty \Big(1 - \frac{\phi}{j - 1 + \phi} \cdot \frac{\theta\lambda(T_j)}{\theta\lambda(T_j) + j - 1}\Big) \qquad (34) \\
&= \prod_{j=2}^\infty \Big[1 - \rho_j\Big(\omega_j - 1\Big)\Big]^{-1}, \qquad (35)
\end{aligned}$$

with notation $\beta_l = \theta\lambda(T_l) + l - 1$, $\rho_l = \theta\lambda(T_l)/(l-1)$, and $\omega_l = (1 + \phi/\beta_l)^{-1}$ for $l > 1$.

The moments of X_1 can be found from (34), for $k = 0, 1, \ldots$, by setting $\phi = k$

to obtain

$$\mathbb{E}(X_1^k) = \mathbb{E}\left[\prod_{j=2}^{\infty}\Big(1 - \frac{k}{k+j-1}\cdot\frac{\theta\lambda(T_j)}{\theta\lambda(T_j)+j-1}\Big)\right]. \tag{36}$$

A representation shown by (35) is

$$-\log(X_1) = \sum_{j=2}^{\infty}\gamma_j, \tag{37}$$

where $\{\gamma_j,\ j>1\}$ are mutually independent random variables with Laplace transforms

$$\mathbb{E}\Big(e^{-\phi\gamma_j}\Big) = \Big[1-\rho_j\Big(\omega_j - 1\Big)\Big]^{-1},\ j>1.$$

The random variable γ_j has an atom at zero with probability $(1+\rho_j)^{-1}$, and a continuous density of

$$\begin{aligned}
&\sum_{l=1}^{\infty}\Big(\frac{\rho_j}{1+\rho_j}\Big)^l\cdot\frac{1}{1+\rho_j}\cdot\frac{\beta_j^l\gamma^{l-1}}{(l-1)!}e^{-\beta_j\gamma}\\
=&\frac{\rho_j}{1+\rho_j}\cdot\frac{\beta_j}{1+\rho_j}\cdot e^{-\frac{\beta_j}{1+\rho_j}\gamma}\\
=&\frac{\rho_j}{1+\rho_j}\cdot(j-1)\cdot e^{-(j-1)\gamma},\ \gamma>0.
\end{aligned}\tag{38}$$

Of course $-\log(X_1)$ is continuous, which agrees with

$$P\Big(\sum_{j=2}^{\infty}\gamma_j = 0\Big) = \prod_{j=2}^{\infty}(1+\rho_j)^{-1} = 0,$$

since the series diverges to zero, because ρ_j is asymptotic to j^{-1}. Note that directly from (34) $\gamma_j = \chi_j\kappa_j$, for $j>1$, where $\{\chi_j,\ j>1\}$ and $\{\kappa_j,\ j>1\}$ are independent with $\{\kappa_j,\ j>1\}$ mutually independent exponential random variables with rates $\{j-1,\ j>1\}$. The Laplace transform of $-\log(X_m)$, conditional on i_m and $\{T_l,\ l>1\}$ is

$$\begin{aligned}
&\mathbb{E}\Big[e^{\phi\log(X_m)}\Big]\\
=&\mathbb{E}\Big[\Big(1-e^{-V_{m-1}}\Big)^{\phi}\prod_{j=m}^{\infty}e^{-\phi V_j}\Big]\\
=&\prod_{j=2}^{i_m}\Big(1+\frac{\phi}{j-1}\Big)^{-1}\cdot\prod_{j=i_m+1}^{\infty}\Big[1-\rho_j\Big(\omega_j-1\Big)\Big]^{-1}.
\end{aligned}\tag{39}$$

The first product in (39) is obtained from

$$\begin{aligned}\mathbb{E}\Big[\Big(1-e^{-V_{m-1}}\Big)^{\phi}\Big] &= (i_m-1)\int_0^{\infty} e^{-(i_m-1)v}\Big(1-e^{-v}\Big)^{\phi}dv \\ &= (i_m-1)\int_0^1 y^{i_m-2}(1-y)^{\phi}dy \\ &= (i_m-1)B(i_m-1,\phi+1) \\ &= \prod_{j=2}^{i_m}\frac{j-1}{j-1+\phi}.\end{aligned}$$

The structure of (39) clearly implies that

$$-\log(X_m) = \sum_{k=2}^{i_m}\delta_k + \sum_{j=i_m+1}^{\infty}\gamma_j, \tag{40}$$

where $\{\delta_j,\ j>1\}$ are independent exponential random variables such that δ_j has rate $j-1$. Trying to simplify (39) furthur by taking expectation with respect to i_m seems complicated.

As an application, it is of interest to calculate the probability p_O that the oldest type in a sample of genes is the oldest type in the population. In a constant size population, from the GEM distribution

$$p_O = 1-\mathbb{E}\Big((1-X_1)^n\Big) = 1-\frac{\theta}{\theta+n} = \frac{n}{\theta+n}.$$

In a variable-sized population model, using (36)

$$p_O = \sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}\mathbb{E}\left[\prod_{j=2}^{\infty}\left(1-\frac{k}{k+j-1}\cdot\frac{\theta\lambda(T_j)}{\theta\lambda(T_j)+j-1}\right)\right]. \tag{41}$$

5.5 GEM distribution equivalence with a constant population size

The consistency of the representation (27) with the GEM distribution in a constant sized population model where $\lambda(t)=1, t\geq 0$, is shown in this section. A long calculation shows that moments in the finite-dimensional distributions of (27) coincide with moments in the GEM representation in the constant population size model. Let $r_1,\ldots,r_m$ be non-negative integers. In the GEM distribution

$$\begin{aligned}\mathbb{E}\Big[\prod_{l=1}^{m} X_l^{r_l}\Big] &= \prod_{l=1}^{m} \mathbb{E}\Big[Z_l^{r_l}\Big(1-Z_l\Big)^{\sum_{\nu=l+1}^{m} r_\nu}\Big] \\ &= \theta^m \prod_{l=1}^{m} B\Big(r_l+1, \sum_{\nu=l+1}^{m} r_\nu + \theta\Big) \\ &= \theta^m \prod_{l=1}^{m} \frac{r_l!\Gamma\Big(\sum_{\nu=l+1}^{m} r_\nu + \theta\Big)}{\Gamma\Big(\sum_{\nu=l}^{m} r_\nu + \theta + 1\Big)} \\ &= \frac{\Gamma(\theta)}{\Gamma\Big(|r|+\theta\Big)} \cdot \theta^m \prod_{l=1}^{m} \frac{r_l!}{\sum_{\nu=l}^{m} r_\nu + \theta}. \end{aligned} \tag{42}$$

Notation used is $|r| = \sum_{\nu=1}^{m} r_\nu$. To show the equivalence of the GEM distribution and (27), moments in (27) are calculated and shown to agree with (42). An identity that establishes the equivalence is that for $m > 1$

$$\begin{aligned}\mathbb{E}\Big(\prod_{l=1}^{m} X_l^{r_l}\Big) = \frac{\theta}{\theta + r_m} \cdot \frac{r_{m-1}! r_m!}{(r_{m-1}+r_m)!} \\ \cdot \mathbb{E}\Big(\Big[\prod_{l=1}^{m-2} X_l^{r_l}\Big] \cdot X_{m-1}^{r_{m-1}+r_m}\Big).\end{aligned} \tag{43}$$

It then follows by induction on m that (42) is satisfied. Note that for $m = 1$, directly from (27),

$$\mathbb{E}\Big(X_1^{r_1}\Big) = \mathbb{E}\Big[\prod_{l=1}^{\infty} \frac{i_{l+1}-1}{i_{l+1}+r_1-1}\Big],$$

and for $m > 1$,

$$\begin{aligned}\mathbb{E}\Big[\prod_{l=1}^{m} X_l^{r_l}\Big] &= \mathbb{E}\Big[\prod_{\alpha=1}^{m-1} \xi_\alpha^{r_{\alpha+1}}(1-\xi_\alpha)^{\sum_{\nu=1}^{\alpha} r_\nu} \cdot \prod_{l=m}^{\infty}(1-\xi_l)^{\sum_{\nu=1}^{m} r_\nu}\Big] \\ &= \mathbb{E}\Big[\prod_{\alpha=1}^{m-1}(i_{\alpha+1}-1)B(r_{\alpha+1}+1, i_{\alpha+1}+\sum_{\nu=1}^{\alpha} r_\nu - 1) \\ &\qquad \cdot \prod_{l=m}^{\infty} \frac{i_{l+1}-1}{i_{l+1}+|r|-1}\Big].\end{aligned} \tag{44}$$

The expectation of the last product in (44) conditional on i_m is

$$
\begin{aligned}
&\mathbb{E}\Big[\prod_{l=m}^{\infty}\frac{i_{l+1}-1}{i_{l+1}+|r|-1}\Big]\\
&\quad=\mathbb{E}\Big[\prod_{j=i_m+1}^{\infty}\Big(1-\frac{\chi_j|r|}{j+|r|-1}\Big)\Big]\\
&\quad=\prod_{j=i_m+1}^{\infty}\Big(1-\frac{\theta|r|}{(j+|r|-1)(j+\theta-1)}\Big)\\
&\quad=\prod_{j=i_m+1}^{\infty}\Big[\frac{(j-1)(j+\theta+|r|-1)}{(j+|r|-1)(j+\theta-1)}\Big]\\
&\quad=\frac{1}{(i_m-1)!}\cdot\frac{\Gamma(\theta+i_m)\Gamma(|r|+i_m)}{\Gamma(\theta+|r|+i_m)}.
\end{aligned}
\tag{45}
$$

Simplification of the second last line in (45) to the last line follows by taking the limit of the product from i_m+1 to n as $n\to\infty$ and applying Euler's formula

$$
\Gamma(z)=\lim_{n\to\infty}\frac{n!n^z}{\prod_{j=0}^{n}(z+j)}.
$$

If $m=1$, $i_m=1$, this shows that

$$
\mathbb{E}\Big(X_1^{r_1}\Big)=\frac{\Gamma(\theta+1)\Gamma(r_1+1)}{\Gamma(\theta+r_1+1)},
$$

in agreement with (42). The term containing i_m in (44) when $m>1$ is thus

$$
\begin{aligned}
&(i_m-1)B(r_m+1,i_m+|r|-r_m-1)\\
&\quad\cdot\frac{1}{(i_m-1)!}\cdot\frac{\Gamma(\theta+i_m)\Gamma(|r|+i_m)}{\Gamma(\theta+|r|+i_m)}\\
&\quad=\frac{\Gamma(r_m+1)\Gamma(i_m+|r|-r_m-1)\Gamma(\theta+i_m)}{(i_m-2)!\Gamma(\theta+|r|+i_m)}.
\end{aligned}
\tag{46}
$$

The probability that $i_m=i_{m-1}+j$, for $j\geq 1$, conditional on i_{m-1} is

$$
\frac{\theta(i_{m-1}+j-2)!\Gamma(\theta+i_{m-1})}{(i_{m-1}-1)!\Gamma(i_{m-1}+j+\theta)}.
\tag{47}
$$

The expected value of the expression (46) conditional on i_{m-1} is obtained by multiplying (46) by (47) and summing; that is

$$
\begin{aligned}
&\frac{\theta\Gamma(r_m+1)\Gamma(\theta+i_{m-1})}{(i_{m-1}-1)!\Gamma(\theta+r_m+1)}\sum_{j=1}^{\infty}B(\theta+r_m+1,j-1+i_{m-1}+|r|-r_m)\\
&\quad=\frac{\theta r_m!\Gamma(\theta+i_{m-1})}{(i_{m-1}-1)!\Gamma(\theta+r_m+1)}B(\theta+r_m,i_{m-1}+|r|-r_m).
\end{aligned}
\tag{48}
$$

Simplification from the second last line in (48) follows by expressing the Beta function as an integral, then summing in the integrand. The identity used is that for $a > 0, b > 0$,

$$\sum_{j=1}^{\infty} B(a+1, b+j-1) = B(a, b).$$

Multiplying the last line in (48) by the term containing i_{m-1} in (44)

$$(i_{m-1} - 1)B(r_{m-1} + 1, i_{m-1} + |r| - r_m - r_{m-1} - 1)$$

results in the expression

$$\begin{aligned} &\frac{\theta}{\theta + r_m} \cdot \frac{r_m! r_{m-1}!}{(r_m + r_{m-1})!} \\ &\quad \cdot \frac{\Gamma(r_{m-1} + r_m + 1)\Gamma(i_{m-1} + |r| - r_{m-1} - r_m - 1)\Gamma(\theta + i_{m-1})}{(i_{m-1} - 2)!\Gamma(\theta + |r| + i_{m-1})}. \end{aligned} \tag{49}$$

Comparing (46) and (49) establishes the identity (43) and therefore completes the proof of the equivalence of the GEM representation and (27).

6 Genealogy of a derived type in a population of constant size

In the case of a constant population size, the probability that a sample of n genes contains n_l genes of type l for $l = 1, \ldots, k$ with $\sum_l n_l = n$ does not depend on the order in which the sampled genes are lost backward in time either by mutation or coalescence. Therefore, if a given type is represented r times and known to have been derived from another type in the sample, the probability for the last gene of this type to be lost when there remain $m + 1$ genes for $m = 1, \ldots, n - r$ is given by

$$\frac{\begin{pmatrix} n-m-1 \\ r-1 \end{pmatrix}}{\begin{pmatrix} n-1 \\ r \end{pmatrix}}, \tag{50}$$

which converges to $q(1-q)^{m-1}$ as n and r tend to infinity such that r/n converges to q. The time of occurrence of this event, represented by T_{m+1}, is distributed as the sum of independent exponential variables of parameters

$i(\theta+i-1)/2$ for $i=m+1,\ldots,n$, whose expectation is

$$\sum_{i=m+1}^{n}\frac{2}{i(\theta+i-1)}. \tag{51}$$

Multiplying and summing over m, the mean age of the mutation that has given rise to the family of size r is

$$\sum_{m=1}^{n-r}\frac{\begin{pmatrix} n-m-1 \\ r-1 \end{pmatrix}}{\begin{pmatrix} n-1 \\ r \end{pmatrix}}\cdot\sum_{i=m+1}^{n}\frac{2}{i(\theta+i-1)}. \tag{52}$$

The limit of (52) as $n\to\infty$ is

$$\frac{2q}{\theta-1}\int_0^1\frac{1-v^{\theta-1}}{1-v}\cdot\frac{v}{1-(1-q)v}dv. \tag{53}$$

If $\theta\to 0$ and $n\to\infty$, then the mean age, calculated directly from (52), is

$$\sum_{m=1}^{\infty}q(1-q)^{m-1}\cdot\frac{2}{m}=-\frac{2q}{1-q}\log(q). \tag{54}$$

Kimura and Ohta (1973) derived the classical formula (54). Griffiths and Marjoram (1996), Griffiths and Tavaré (1998), Wiuf and Donnelly (1999) and Stephens (2000) show that the mean age of a mutation that gave rise to a family of size r is (52) when $\theta=0$. Griffiths (2003) shows that there is a simplification to

$$2r(n-r)^{-1}\sum_{j=r+1}^{n}j^{-1}. \tag{55}$$

In the treatment of the above authors only the lineages containing a given mutation are considered, with other mutations not affecting lineages.

Similarly to the the derivation of (50), the probability for a gene of the derived type to be lost when there remain $m+1$ genes among which j of the derived type for $j=2,\ldots,r$ and $m=j,\ldots,n-r+j-1$ is

$$\frac{\begin{pmatrix} m-1 \\ j-1 \end{pmatrix}\cdot\begin{pmatrix} n-m-1 \\ r-j \end{pmatrix}}{\begin{pmatrix} n-1 \\ r \end{pmatrix}},$$

whose limit is

$$\binom{m-1}{j-1} q^j (1-q)^{m-j},$$

and the time of occurrence of the coalescence event responsible for this loss has expectation

$$\sum_{m=j}^{n-r+j-1} \frac{\binom{m-1}{j-1} \cdot \binom{n-m-1}{r-j}}{\binom{n-1}{r}} \cdot \sum_{i=m+1}^{n} \frac{2}{i(\theta+i-1)}, \tag{56}$$

whose limit as $n \to \infty$ is

$$\frac{2q^j}{\theta-1} \cdot \int_0^1 \frac{(1-v^{\theta-1})v^j}{(1-v)(1-(1-q)v)^j} dv. \tag{57}$$

Additionally, as $\theta \to 0$ the limit is

$$\sum_{m=j}^{\infty} \binom{m-1}{j-1} q^j (1-q)^{m-j} \cdot \frac{2}{m} = 2 \left(\frac{q}{1-q}\right)^j \int_q^1 \frac{(1-y)^{j-1}}{y^j} dy. \tag{58}$$

In the case $j = 2$, (58) evaluates to

$$\frac{2q}{1-q} + \frac{2q^2}{(1-q)^2} \log(q), \tag{59}$$

which corresponds to the expected time it takes for all genes of the derived type to coalesce.

The above treatment shows that, in the limit, the total number of genes remaining the first time there remain $j-1$ genes of the derived type, denoted by $M(j)$, is distributed as a sum of j independent geometric variables of parameter q, and therefore the distribution of $qM(j)$ as q tends to 0 converges to the distribution of a sum of j independent exponential variables of parameter 1. Moreover, assuming $qM(j) = x$ fixed and multiplying the unit of time by q, the last time there remain j genes of the derived type converges in distribution to its mean, which is $2/x$, as θ and q tend to 0, since its variance, which is given by

$$\sum_{i=\frac{x}{q}+1}^{n} \frac{4}{q^2 i^2 (\theta+i-1)^2},$$

is bounded by

$$\int_{\frac{x}{q}-1}^{\infty} \left(\frac{4}{q^2 y^4}\right) dy = \frac{4q}{3(x-q)^3},$$

which converges to 0 as q tends to 0, in agreement with Wiuf (2000). This means that the last time there remain j genes of the derived type is distributed, in the limit, as twice the inverse of a gamma distribution.

7 Genealogy of a derived type with variable population size

When the population size is variable, the probability of having r genes of a derived type and $n-r$ genes of an ancestral type in a sample of size n for $r = 1, \ldots, n-1$ is

$$\frac{(n-1)! \cdot \theta}{r} \sum_{m=1}^{n-r} \frac{\binom{n-m-1}{r-1}}{\binom{n-1}{r}} E\left\{ \frac{\lambda(T_{m+1})}{\prod_{i=2}^{n}[\theta\lambda(T_i) + i - 1]} \right\}. \tag{60}$$

The probability of this event, given that there are two types in the sample, is proportional, as θ tends to 0, to

$$\frac{n}{r} \cdot \sum_{m=1}^{n-r} \frac{\binom{n-m-1}{r-1}}{\binom{n-1}{r}} E(\lambda(T_{m+1})), \tag{61}$$

which converges, as n and r tend to 0 such that r/n converges to q, to

$$L(q) = \sum_{m=1}^{\infty} (1-q)^{m-1} E(\lambda(T_{m+1})). \tag{62}$$

Moreover, given a frequency q of the derived type, the last gene of this type is lost by mutation when there remain $m+1$ genes with probability

$$\frac{(1-q)^{m-1}}{L(q)} E(\lambda(T_{m+1})), \tag{63}$$

for $m \geq 1$, and the time of occurrence of this event is T_{m+1} with this probability.

The coalescent process in a variable sized population can be coupled with a process in a population of constant size $N(0)$ by measuring time backwards in units of $\tau = \int_0^t \lambda(s)^{-1} ds$. In a population which decreases in size exponentially

back in time $N(t) = N(0)e^{-\beta t}$, that is $\lambda(t) = e^{-\beta t}$, and $\beta t = \log(1 + \beta\tau)$. In such a case

$$E(\lambda(T_{m+1})) = E\left\{\frac{1}{1+\beta\mathcal{T}_{m+1}}\right\}, \tag{64}$$

where $\mathcal{T}_{m+1}$ is distributed as a sum of independent exponential variables of parameters $i(i-1)/2$ for $i = m+1, \ldots, n$. Keeping $qm = x$ and $q\beta = b$ fixed as q tends to 0, the variable $\beta\mathcal{T}_{m+1}$ converges in distribution to $2b/x$. Then the variable $qM = X$, where M represents the number of genes remaining just after the loss of the last gene of the derived type has a limiting density function, as q tends to 0, that is proportional to

$$f(x) = \frac{e^{-x}}{1+2b/x}, \tag{65}$$

for $x > 0$. Moreover, the time of occurrence of this event in time units of $qN(0)$ generations is distributed, as q tends to 0, as

$$(1/b)\log(1+2b/X). \tag{66}$$

Similarly, since the probability that the sampled genes are lost in a given order depends only on the position of the last gene of the derived type, the variable $qM(j) = X(j)$, where $M(j)$ represents the number of genes remaining the first time there remain $j-1$ genes of the derived type, will be distributed, as q tends to 0, as a sum of j independent random variables, one of which has a density function proportional to $f(x)$ and the other $j-1$ have an exponential distribution with parameter 1. Moreover, the time of occurrence of this event in time units of $qN(0)$ generations will be distributed, as q tends to 0, as

$$(1/b)\log(1+2b/X(j)). \tag{67}$$

Again, this is in agreement with Wiuf (2000, 2001).

References

Cannings, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models, *Advances in Applied Probability* **6**, 260–290.

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

Donnelly, P., and Kurtz, T. G. (1996). A countable representation of the Fleming-Viot measure-valued diffusion, *Annals of Probability* **24**, 698–742.

Donnelly, P., and Tavaré, S. (1986). The ages of alleles and a coalescent, *Advances in Applied Probability* **18**, 1–19.

Donnelly, P., and Tavaré, S. (1987). The population genealogy of the infinitely-many neutral alleles model, *Journal of Mathematical Biology* **251**, 381–391.

Ethier, S.N., and Kurtz, T.G. (1981). The infinitely-many-neutral-alleles diffusion model, *Advances in Applied Probability* **13**, 429–452.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology* **3**, 87–112.

Ewens, W. J. (1990). Population genetics theory - The past and the future, *in* "Mathematical and Statistical Developments of Evolutionary Theory" (S. Lessard, Ed.), NATO ASI Series C: Mathematical and Physical Sciences, Vol. 299, pp. 177–227, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models, *Theoretical Population Biology* **17**, 37–50.

Griffiths, R. C. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model, *Theoretical Population Biology* **64**, 241–251.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination, *Journal of Computational Biology* **3**, 479–502.

Griffiths, R.C., and Tavaré, S. (1998). The age of a mutation in a general coalescent tree, *Stochastic Models* **14**, 273–295.

Griffiths, R. C. and Tavaré, S. (2003). The genealogy of a neutral mutation, *in* "Highly Structured Stochastics Systems" (P. J. Green, N. L. Hjort and S. Richardson, Eds.), Oxford Statistical Science Series **27**, Oxford University Press, Oxford, United Kingdom.

Hoppe, F. M. (1984). Polya-like urns and the Ewens' sampling formula, *Journal of Mathematical Biology* **20**, 91–94.

Joyce, P., and Tavaré, S. (1987). Cycles, permutations and the structures of the Yule process with immigration, *Stochastic Processes and Their Applications* **25**, 309–314.

Karlin, S. and McGregor, J.L. (1972). Addendum to a paper of W. Ewens, *Theoretical Population Biology* **3**, 113–116.

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

Kingman, J.F.C. (1978). Random partitions in population genetics, *Proceedings of the Royal Society of London Series A* **361**, 1–20.

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

Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation, *Advances of Applied Probability* **28**, 525–539.

Rannala, B., and Slatkin, M. (1998). Likelihood analysis of disequilibrium mapping, and related problems, *American Journal of Human Genetics* **62**, 459–473.

Slatkin, M., and Rannala, B. (1997). Estimating the age of alleles by use of intra-allelic variability, *American Journal of Human Genetics* **60**, 447–458.

Stephens, M. (2000). Times on trees, and the age of an allele, *Theoretical Population Biology* **57**, 109–119.

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

Watterson, G.A. (1984). Lines of descent and the coalescent, *Theoretical Population Biology* **10**, 239–253.

Wiuf, C. (2000). On the genealogy of a sample of neutral rare alleles, *Theoretical Population Biology* **58**, 61–75.

Wiuf, C. (2001). Rare alleles and selection, *Theoretical Population Biology* **59**, 287–296.

Wiuf, C., and Donnelly, P. (1999). Conditional genealogies and the age of a neutral mutant, *Theoretical Population Biology* **56**, 183-201.