Theoretical Population Biology 75 (2009) 260–265

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

The genealogy, site frequency spectrum and ages of two nested mutant alleles

Asger Hobolth, Carsten Wiuf*

Bioinformatics Research Center, Aarhus University, Denmark

ARTICLE INFO

Article history:
Received 31 December 2008
Available online 26 February 2009

Keywords:
Age of a mutation
Coalescent
Genealogy
Jump chain
Nested mutations

ABSTRACT

In this paper we consider the genealogy of two nested mutant alleles, assuming the constant-size neutral coalescent model with infinite sites mutation. We study the conditional genealogy and derive explicit formulas for the joint and marginal site frequency spectra for the double, single and zero mutant allele. In addition, we find the mean ages of the two mutations. We show that the age of the youngest mutation does not depend on the frequency of the single mutant allele and that the frequency spectra for the single mutant allele and the zero mutant allele are the same.

1. Introduction

Theory for a single segregating site in a population goes back a long time. Kimura and Ohta (1973) found the expected age of a mutant allele,

$$-\frac{2f}{1-f}\log f, \tag{1}$$

where f is the mutant's frequency in the population. Two years later, Watterson (1975) showed that the mutant's frequency spectrum in a sample of size n is

$$\frac{1/d}{\sum_{j=1}^{n-1} 1/j}, \tag{2}$$

where d is the number of mutants in the sample.

Later authors have extensively made use of coalescent arguments and re-derived old as well as new results about the age, frequency spectrum and genealogical structure of the allele, in addition to relaxing the assumption of a constant size population; see e.g. Innan and Tajima (1997), Griffiths and Tavaré (1998), Wiuf and Donnelly (1999), and Stephens (2000). More recently, diffusion arguments, similar to those of Kimura and Ohta (1973), have been used to provide population-based statements about the mutant allele; see e.g. van Herwaarden and van der Wal (2002).

In this paper we extend the setting in a new direction. We consider two completely linked and nested mutant alleles and their genealogical history (see Fig. 1). We assume the infinite sites model with scaled mutation rate $\theta = 4Nu$ (Watterson, 1975), where u is the mutation rate per gene per generation and $2N$ the effective population size. In particular, we are interested in the situation as $\theta \to 0$ and/or the situation as the sample size goes to infinity while the frequencies of the two alleles are kept fixed. We provide analogues of Eqs. (1) and (2) and compare our results to the classical setting of one mutation. Griffiths and Tavaré (2003) discuss properties of samples subtending a mutation and we compare one of their results to ours.

We assume the standard coalescent (Kingman, 1982), but note that some of our results can be stated more generally in terms of binary coalescents (Griffiths and Tavaré, 1998) . The standard coalescent with infinite sites mutation is characterized by:

(S1) The waiting times between coalescent events in a sample of size n are given by $W_2, W_3, \ldots, W_n$, where W_i is the time while there are i ancestors to the sample.
(S2) The W_i's are independent and exponentially distributed $\mathrm{Exp}(i(i-1)/2)$.
(S3) At each coalescent event, a pair of genes is chosen at random to coalesce.
(S4) Mutations occur according to a Poisson Process $Po(\theta L_n/2)$, where $\theta = 4Nu$ is the scaled mutation rate and $L_n = \sum_i iW_i$ is the sum of all branches in the genealogy.

A general binary coalescent fulfills S1, S3 and S4, but puts no constraints on the distribution of waiting times (S2). Note that S3 implies that the jump chain, the process specifying the coalescing genes, is independent of the times between events (Kingman, 1982).

In Section 2, we consider the jump chain for nested groups of genes. We assume exchangeability (S3) and extend a result derived by Wiuf and Donnelly (1999) for two groups. In the remaining

* Corresponding address: Bioinformatics Research Center, Aarhus University, Hoegh-Guldbergsgade 10, Bldg 1110, 8000 Aarhus C, Denmark.
E-mail addresses: asger@birc.au.dk (A. Hobolth), wiuf@birc.au.dk (C. Wiuf).

0040-5809/$ – see front matter
doi:10.1016/j.tpb.2009.02.001

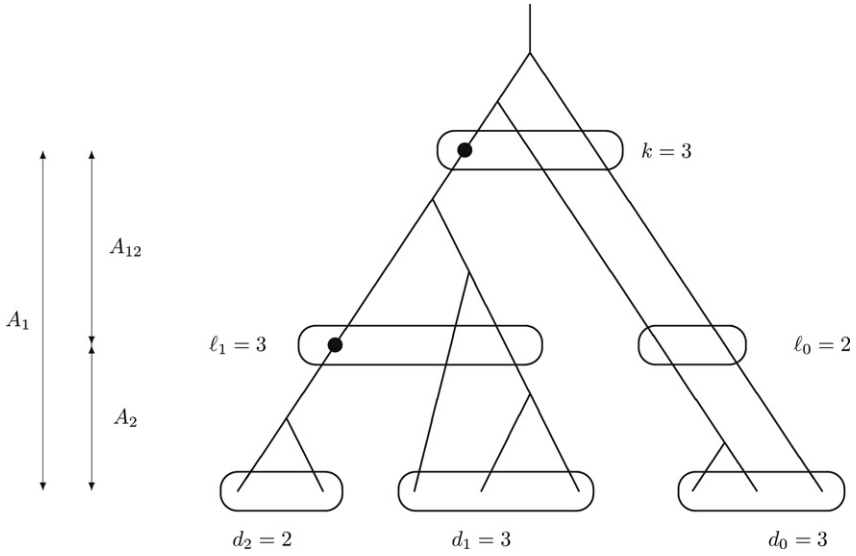

Fig. 1. A possible genealogical history for two nested mutant alleles. From the top down, the first mutation happens at level $k = 3$ while the second mutation happens at level $\ell = \ell_0 + \ell_1 = 5$. The second mutation occurs in a lineage carrying the first mutation which is in frequency $\ell_1/\ell = 3/5$. The sample size is $n = 8$, with $d_2 = 2$ alleles having both mutations (the double mutant allele), $d_1 = 3$ alleles having the first mutation only (the single mutant allele), and $d_0 = n - d_1 - d_2 = 3$ alleles not having any mutations (the zero mutant allele). The age of the youngest mutation is A_2 and the age of the oldest mutation is A_1; A_{12} is the time from the youngest mutation to the oldest.

parts of the paper, we assume the infinite sites coalescent model. Section 3 outlines and derives the classical results for a single mutant allele; the derivations are extended in Section 4 to two nested mutant alleles. In Section 5 we consider the limit $\theta \to 0$. In particular, we provide analytical expressions for the site frequency spectra for the double, single and zero mutant allele. Finally, in Section 6, we consider the ages of the mutations. The paper ends with a discussion on applications and potential future directions.

2. Topological characterization

We start by recapitulating two statements about the number of descendants of $G + 1$ (ordered) lineage groups with $\mathbf{m} = (m_0, m_1, \ldots, m_G)$ members. The statements are independent of the waiting times between events and depend on the exchangeability assumption S3 of neutral models only; i.e. they are true for binary coalescents and, in fact, relate to urn models. Lemma 2 is a corollary of Lemma 1; a proof of Lemma 1 can be found in Griffiths (1980) and Kingman (1982).

Lemma 1. *The probability $P(\mathbf{d}|\mathbf{m})$ that $\mathbf{m} = (m_0, m_1, \ldots, m_G)$ lineages leave $\mathbf{d} = (d_0, d_1, \ldots, d_G)$ descendants is given by*

$$P(\mathbf{d}|\mathbf{m}) = \binom{d_0 - 1}{m_0 - 1} \cdots \binom{d_G - 1}{m_G - 1} \binom{n - 1}{m - 1}^{-1}, \tag{3}$$

where $n = \sum_{i=0}^{G} d_i$ and $m = \sum_{i=0}^{G} m_i$.

Lemma 2. *Given the configurations $\mathbf{m}$ and $\mathbf{d}$, the probability $P(\mathbf{d}_i|\mathbf{d}, \mathbf{m})$ that the last event duplicated a gene in group i, is*

$$P(\mathbf{d}_i|\mathbf{d}, \mathbf{m}) = \frac{d_i - m_i}{n - m} \tag{4}$$

where $\mathbf{d}_i = (d_0, \ldots, d_{i-1}, d_i - 1, d_{i+1}, \ldots, d_G)$.

The probability in Lemma 1 is termed a 'forward' probability as it relates a configuration, $\mathbf{m}$, to future configurations, $\mathbf{d}$. In contrast, the probability in Lemma 2 is termed a 'backward' probability, because it relates the current configuration, $\mathbf{d}$, to its history. If the ancestral configuration, $\mathbf{m}$, is known, then the jump chain transition probabilities are given by Lemma 2.

A mutation implies a topological constraint on the sample genealogy; all genes sharing the mutation must coalesce before coalescing with any other gene (Fig. 1). Wiuf and Donnelly (1999) showed how the topological constraint alone (i.e. without assuming it is caused by a mutation) affects the jump chain. Here the result is generalized to cover a series of nested groups. Consider a series of $G + 1$ nested groups with members $\mathbf{d} = (d_0, d_1, \ldots, d_G)$, $d_i \geq 1$, such that the lowest group has d_G members and the ith group has $d_{i\cdot} = d_i + \cdots + d_G$ members. All members/genes of group i must coalesce with each other before coalescing with any gene in group $j < i$. When there is only one member of group G ($d_G = 1$), it is allowed to coalesce with genes in group $G - 1$, whereby group G ceases to exist. We denote the topological constraint by E.

The case $G = 1$ corresponds to Wiuf and Donnelly's case and the case $G = 2$ is illustrated in Fig. 1 (ignoring the mutations).

Theorem 3. *The probability that the last coalescent event is among genes in group $i = 0, 1, \ldots, G$, is*

$$P(\mathbf{d} - \mathbf{e}_i|E, \mathbf{d}) = \begin{cases} \dfrac{d_0 - 1}{d_{0\cdot}} & \text{if } i = 0 \\ \dfrac{d_i - 1}{d_{i\cdot}} \displaystyle\prod_{j=0}^{i-1} \frac{d_{(j+1)\cdot} + 1}{d_{j\cdot}} & \text{if } i \leq G - 1 \\ \displaystyle\prod_{j=0}^{G-1} \frac{d_{(j+1)\cdot} + 1}{d_{j\cdot}} & \text{if } i = G \end{cases} \tag{5}$$

where $\mathbf{e}_i$ is the $(i + 1)$th unit vector, $d_{i\cdot} = d_i + \cdots + d_G$ and $d_i \geq 1$. If $d_G = 1$, $P(\mathbf{d} - \mathbf{e}_G|E, \mathbf{d})$ is the probability that the only member of group G coalesces with a member in group $G - 1$, and hence group G ceases to exist.

In particular, for 2 groups ($G = 1$) we retrieve Wiuf and Donnelly's result

$$P(\mathbf{d} - \mathbf{e}_0|E, \mathbf{d}) = \frac{d_0 - 1}{n} \quad \text{and} \quad P(\mathbf{d} - \mathbf{e}_1|E, \mathbf{d}) = \frac{d_1 + 1}{n}, \tag{6}$$

where $n = d_0 + d_1$. For 3 groups ($G = 2$) we find

$$P(\mathbf{d} - \mathbf{e}_0|E, \mathbf{d}) = \frac{d_0 - 1}{n},$$
$$P(\mathbf{d} - \mathbf{e}_1|E, \mathbf{d}) = \frac{d_1 - 1}{d_1 + d_2}\frac{d_1 + d_2 + 1}{n} \quad \textit{and} \tag{7}$$
$$P(\mathbf{d} - \mathbf{e}_2|E, \mathbf{d}) = \frac{d_1 + d_2 + 1}{n}\frac{d_2 + 1}{d_1 + d_2},$$

where $n = d_0 + d_1 + d_2$.

Proof. See Appendix. □

A main difference between Lemma 2 and Theorem 3 is that Lemma 2 conditions on the ancestral configuration **m**. Despite the simple structure in Theorem 3, the structure becomes much more complicated when mutations are imposed (see Section 4).

3. A single mutant allele

To set the stage, we start by re-deriving a few result for a single mutation. The results for two nested mutations use the same line of arguments. Assume a mutation happens at level k and that no further mutations happen in the sample history. Under the infinite sites model, the probability of this event is proportional to

$$p_n(k|M, \theta) \propto k\,\mathrm{E}\left(W_k e^{-\theta L_n/2}\right) \propto \frac{1}{k - 1 + \theta}, \tag{8}$$

where M is the event that exactly one mutation has occurred. Eq. (8) follows from the Poisson nature of the mutation process, *cf.* S4. One mutation must happen while at level k (branch length kW_k) and no other mutations at any other level (branch length $L_n - kW_k$). From Lemma 1, we know the probability that $(k - 1, 1)$ ancestors leave $(n - d, d)$ descendants and we find

$$p_n(k, d|M, \theta) \propto \frac{1}{k - 1 + \theta}\binom{n - d - 1}{k - 2}\binom{n - 1}{k - 1}^{-1}, \tag{9}$$

with $k = 2, \ldots, n - d + 1$. Consequently, the probability that d alleles carry the mutation is given by

$$p_n(d|M, \theta) \propto \sum_{k=2}^{n-d+1} \frac{1}{k - 1 + \theta}\binom{n - d - 1}{k - 2}\binom{n - 1}{k - 1}^{-1}$$
$$= \sum_{k=2}^{n-d+1} \frac{k - 1}{d(k - 1 + \theta)}\binom{n - k}{d - 1}\binom{n - 1}{d}^{-1}. \tag{10}$$

Here, and elsewhere, we abuse notation slightly: k always refers to the (first) mutation, ℓ to the second (if applicable), and d to the partition imposed by the mutation(s) in the sample. Thus, $p_n(k|M, \theta)$ is the probablity that the mutation happens at level k and $p_n(d|M, \theta)$ is the probability of observing d mutant alleles in the sample.

In the limit as $\theta \to 0$, Eq. (10) reduces to

$$p_n(d|M) \propto \sum_{k=2}^{n-d+1} \frac{1}{d}\binom{n - k}{d - 1}\binom{n - 1}{d}^{-1} = \frac{1}{d}, \tag{11}$$

which is Watterson's (1975) frequency spectrum. The argument is essentially the argument given in Griffiths and Tavaré (1998) and Stephens (2000).

Finally, the jump chain, conditional on the mutation, has transition probabilities given by

$$p_n(d - 1, n - d|M, d, n - d)$$
$$= \frac{p_{n-1}(d - 1|M)q_{d-1,n-1}}{p_{n-1}(d - 1|M)q_{d-1,n-1} + p_{n-1}(d|M)q_{d,n-1}}$$
$$= \frac{1}{d - 1}\frac{d - 1}{n - 1}\Bigg/\left\{\frac{1}{d - 1}\frac{d - 1}{n - 1} + \frac{1}{d}\frac{n - d - 1}{n - 1}\right\}$$
$$= \frac{d}{n - 1}, \tag{12}$$

where $q_{d-1,n-1} = P(d, n-d|d-1, n-d)$ and $q_{d,n-1} = P(d, n-d|d, n - d - 1)$ are the forward probabilities in Lemma 1.

For a general binary coalescent similar results follow by replacing $1/(k - 1 + \theta)$ with $k\mathrm{E}(W_k e^{-\theta L_n/2})$ (see (Griffiths and Tavaré, 1998)). We note that the results depend only on the form of this expectation and Lemma 1.

4. Two nested mutant alleles

In the remaining sections we consider the situation depicted in Fig. 1 in which three alleles are the result of two nested mutations. The first allele does not bear any mutations and is observed in $d_0 \geq 1$ copies, the second has one mutation and is observed in $d_1 \geq 1$ copies, and finally the last allele is of multiplicity $d_2 \geq 1$ and has both the first as well as the second mutation. Thus, the total number of genes carrying the oldest mutation is $d_1 + d_2$.

Under the infinite sites model the probability that the first mutation happens at level k, the second at level $\ell > k$ and no further mutations happen at any other level is proportional to

$$p_n(k, \ell|\tilde{M}_2, \theta) \propto k\ell\,\mathrm{E}(W_k W_\ell e^{-\theta L_n/2}), \tag{13}$$

where $\tilde{M}_2$ denotes that exactly two mutations occur. In the standard coalescent, the W_i's are independent and the equation reduces to

$$p_n(k, \ell|\tilde{M}_2, \theta) \propto \frac{1}{(k - 1 + \theta)(\ell - 1 + \theta)}. \tag{14}$$

Suppose the oldest mutation is of multiplicity $\ell_1 < \ell$ when the youngest arrives (Fig. 1). We can apply Lemma 1 to determine the probability that $(k - 1, 1)$ ancestors leave (ℓ_0, ℓ_1) descendants $(\ell_0 + \ell_1 = \ell)$ and likewise the probability that $(\ell_0, \ell_1 - 1, 1)$ ancestors leave (d_0, d_1, d_2) descendants. These probabilities are given by

$$P(\ell_0, \ell_1|k - 1, 1) = \binom{\ell_0 - 1}{k - 2}\binom{\ell - 1}{k - 1}^{-1} \tag{15}$$

and

$$P(d_0, d_1, d_2|\ell_0, \ell_1 - 1, 1) = \binom{d_0 - 1}{\ell_0 - 1}\binom{d_1 - 1}{\ell_1 - 2}\binom{n - 1}{\ell - 1}^{-1}, \tag{16}$$

where $n = d_0 + d_1 + d_2$.

When the second mutation occurs it must hit one of the ℓ_1 lineages; hence the probability of a sample with mutations at given levels can be found by combining Eqs. (13), (15) and (16) with the probability ℓ_1/ℓ of hitting one of the ℓ_1 lineages:

$$p_n(k, \ell_0, \ell_1, d_0, d_1|M_2, \theta)$$
$$= p_n(k, \ell|\tilde{M}_2, \theta)\binom{\ell_0 - 1}{k - 2}\binom{\ell - 1}{k - 1}^{-1}$$
$$\times \frac{\ell_1}{\ell}\binom{d_0 - 1}{\ell_0 - 1}\binom{d_1 - 1}{\ell_1 - 2}\binom{n - 1}{\ell - 1}^{-1}$$
$$= p_n(k, \ell|\tilde{M}_2, \theta)\frac{k - 1}{\ell}\binom{\ell - k}{\ell_1 - 1}\binom{\ell - 1}{\ell_1}^{-1}$$
$$\times \binom{d_0 - 1}{\ell_0 - 1}\binom{d_1 - 1}{\ell_1 - 2}\binom{n - 1}{\ell - 1}^{-1}, \tag{17}$$

where $\ell = \ell_0 + \ell_1$ and M_2 denotes the event of exactly two nested mutations. We note that d_0 can take the values $d_0 = 1, \ldots, n - 2$; d_1 the values $d_1 = 1, \ldots, n - d_0 - 1$; k the values $k = 2, \ldots, d_0 + 1$; ℓ the values $\ell = k + 1, \ldots, n - d_1 + 1$; and finally ℓ_1 the values $\ell_1 = \max(2, \ell - d_0), \ldots, \min(d_1 + 1, \ell - 1)$.

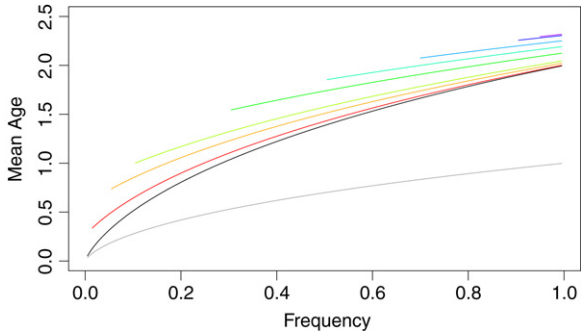

Fig. 2. Shown is the mean age, $E(A_2|M_2)$, of the youngest mutation (grey line, Eq. (32)) and the mean age of a single mutation (black, Eq. (29)), as a function of frequency. Also shown is the mean age of the oldest mutation, $E(A_1|M_2)$. It depends on the frequency of the youngest; from the black line upwards, $f_2 = 0.01$ (red), 0.05, 0.10, 0.30, 0.50, 0.70, 0.90 and 0.95 (purple).

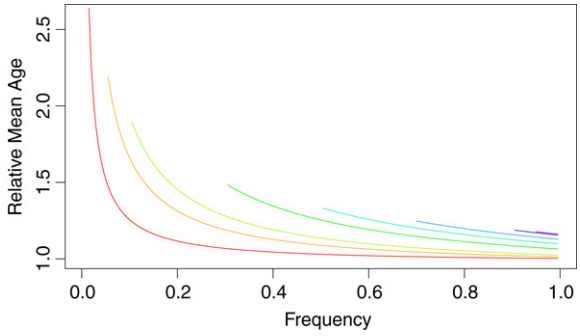

Fig. 3. Shown is the mean age, $E(A_1|M_2)$, of the oldest mutation relatively to the mean age of a single mutation (Eq. (29)), as function of frequency. The frequency of the youngest mutation affects $E(A_1|M_2)$; from the black line upwards, $f_2 = 0.01$ (red), 0.05, 0.10, 0.30, 0.50, 0.70, 0.90 and 0.95 (purple).

The calculations are tedious and not very informative and therefore left out. The result in Eq. (32) has previously been derived by Griffiths and Tavaré (2003) using different arguments (their expression has a factor of 4 in the numerator which should be 2). In Fig. 2 we compare Eqs. (29) and (32). The mean age of the youngest mutation is approximately half the mean age of a single mutation, when they both are in the same frequency in the population. However, there does not seem to be any particular reason for this observation.

The age of the oldest mutation is more difficult to find as it requires knowledge of both levels, k and ℓ; or of ℓ and the number of the oldest mutants, ℓ_1, at level ℓ. Given the level ℓ of the youngest mutation, Lemma 4 provides a means to find ℓ_1. Under the stated conditions,

$$\lim_{\theta\to 0} p_n(\ell_1|M_2,\ell,d_0,d_1) = p(\ell_1|M_2,\ell,g_1) = \binom{\ell-3}{\ell_1-2} g_1^{\ell_1-2}(1-g_1)^{\ell-\ell_1-1}, \quad (33)$$

where $\ell_1 = 2,\ldots,\ell-1$ and $g_1 = f_1/(f_0+f_1)$. Knowing ℓ and ℓ_1, the mean age of the oldest mutant, counted from level ℓ, is

$$A(\ell,\ell_1) = -\frac{2}{\ell} + 2\binom{\ell-1}{\ell_1}^{-1}\sum_{i=1}^{\ell-\ell_1}\frac{1}{i}\binom{\ell-i-1}{\ell_1-1}. \quad (34)$$

(Griffiths and Tavaré, 1998; Wiuf and Donnelly, 1999). It can also be derived from Eqs. (9) and (11) directly.

Writing the age of the oldest mutation as $A_1 = A_2 + A_{12}$, where A_{12} is the age from the youngest to the oldest, we find

$$\begin{aligned} E(A_1|M_2) &= E(A_2|M_2) + \sum_{\ell=3}^{\infty}\sum_{\ell_1=2}^{\ell-1} A(\ell,\ell_1)\, p(\ell_1|M_2,\ell,g_1)\, p(\ell|M_2,f_2) \\ &= E(A_2|M_2) + \sum_{\ell=3}^{\infty}\sum_{i=1}^{\ell-2}\frac{2(\ell-i-1)}{(\ell-1)(\ell-2)i}\{2+(\ell-i-2)g_1\} \\ &\quad\times(1-g_1)^{i-1}\, p(\ell|M_2,f_2). \end{aligned} \quad (35)$$

The latter sum can easily be approximated numerically for given values of f_2 and g_1.

Let the frequency of the oldest mutation be $f = f_1 + f_2$. In Fig. 2, we compare $E(A_1|M_2)$ to the expectation in Eq. (29) for fixed f and varying f_2, i.e. we consider pairs (f_2, g_1) such that $g_1 = (f-f_2)/(1-f_2)$. In Fig. 3, we show $E(A_1|M_2)$ relatively to the expectation in Eq. (29), again for fixed f and varying f_2. We note that the frequency (f_2) of the youngest mutation affects the age of the oldest mutation severely when they both are in low frequency, i.e. when they both are rare. In that case the relative difference can be manyfold.

When the oldest mutation is in high frequency, $f \approx 1$, and the youngest in low frequency, $f_2 \approx 0$, the mean age of the oldest is close to 2. This is expected: Consider the situation of a single mutation in frequency $n-1$ in a large sample of size n. The mutation must happen while there are two ancestral lineages; hence the mean age is $1+1=2$ (time until two ancestral genes + time until a mutation in two genes).

When both mutations are in high frequency, $f \approx 1$ and $f_2 \approx 1$, the mean age of the oldest is close to $2 + 1/3 \approx 2.33$. This is also as expected: Consider the situation of one mutation in frequency $n-2$ and the other in frequency $n-1$ in a large sample of size n. The first mutation must happen while there are three ancestral lineages, while the second must happen while there are two. This implies that the mean age is $(1-1/3)+1/3+1/3+1 = 2+1/3$ (time until three ancestral genes + time until a mutation in three genes + time until two ancestral genes + time until a mutation in two genes).

7. Discussion

We have provided a rigorous theoretical study of conditional genealogies, site frequency spectra and ages of two nested mutations. We obtain nice analytical results that allows us to gain insight into the complex genealogical structure of the coalescent model. Our methodology builds upon previous work in the case of a single mutation; Griffiths and Tavaré (1998), Wiuf and Donnelly (1999) and Stephens (2000). Of particular interest is the fact that the age of the youngest mutation does not depend on the frequency of the single mutant allele and that the frequency spectra for the single mutant allele and the zero mutant allele are the same.

In Hobolth et al. (2008), we applied results from the one mutation case to formulate an improved Importance Sampling proposal distribution for inference on the scaled mutation rate. The results in this paper could potentially be used to further improve the proposal distribution for inference in coalescent models. However, our results also indicate that for two or more mutations analytical results become less tractable, compared to the case of one mutation.

We also considered the case of two non-nested, but completely linked mutant alleles (results not shown). It is straightforward to derive a formula similar to Eq. (17), but unfortunately the distribution is not analytically tractable. Conditional genealogies for non-nested mutations are also not analytically tractable.

Acknowledgements

CW is supported by the Danish Cancer Society and the Danish Research Council. AH is supported by the Danish Research Council. We thank Jens Ledet Jensen for discussions.

Appendix

Proof of Theorem 3. Let a partition of nested groups be given, $\mathbf{d} = (d_0, \ldots, d_G)$, where d_G are the members of the lowest group. The case $G = 1$ is studied in Wiuf and Donnelly (1999). They found the probability $Q_1(d_0, d_1)$ that the d_1 genes find a common ancestor before coalescing with any ancestor of the remaining d_0 genes:

$$Q_1(d_0, d_1) = \frac{2}{d_1 + 1} \binom{d_0 + d_1 - 1}{d_1 - 1}^{-1}. \tag{36}$$

(Eq. (2) in Wiuf and Donnelly (1999)). For $G > 1$, the probability $Q_G(d_0, \ldots, d_G)$ that for all i, the d_i genes find a common ancestor before coalescing with any ancestor of the $d_{i-1} + \cdots + d_0$ genes is given by

$$\begin{aligned} Q_G(d_0, \ldots, d_G) &= Q_{G-1}(d_1, \ldots, d_G) Q_1(d_0, d_{1\cdot}) \\ &= \prod_{i=0}^{G-1} Q_1(d_i, d_{(i+1)\cdot}). \end{aligned} \tag{37}$$

Thus, the probability that the last event is a coalescence event among genes in group i is

$$P(\mathbf{d} - \mathbf{e}_i | E, \mathbf{d}) = \frac{d_i(d_i - 1)}{n(n-1)} \frac{Q_G(\mathbf{d} - \mathbf{e}_i)}{Q_G(\mathbf{d})}. \tag{38}$$

(Wiuf and Donnelly, 1999). By insertion of Eq. (36) into Eq. (38) we obtain the expression in Theorem 3.

Proof of Theorem 5. Eq. (23) is a consequence of Eq. (20) by summing over all instances such that $d_0 + d_1 = n - d_2$ is constant. Since Eq. (20) depends on d_1 only through $d_0 + d_1$, it follows that the distribution of d_1 is uniform conditional on d_2 and that $p_n(d_1|M_2) = p_n(d_0|M_2)$. The expression for $p_n(d_1|M_2)$ follows by first summing over d_0 in Eq. (20), then over ℓ. Finally, the constant is obtained by summing over d_0, d_1 and d_2, then ℓ in Eq. (20).

Proof of Theorem 6. The transition probabilities are a consequence of Eq. (20). For example, $p_n(\mathbf{d} - \mathbf{e}_2|M_2, \ell, \mathbf{d})$ can be rewritten using Bayes' formula

$$p_n(\mathbf{d} - \mathbf{e}_2|M_2, \ell, \mathbf{d}) = P(\mathbf{d}|\mathbf{d} - \mathbf{e}_2) \frac{p_{n-1}(\ell, d_0, d_1|M_2)}{p_n(\ell, d_0, d_1|M_2)}, \tag{39}$$

where $P(\mathbf{d}|\mathbf{d} - \mathbf{e}_2)$ is given in Lemma 1. Insertion of Eq. (20) gives the desired result.

References

Griffiths, R.C., 1980. Lines of descent in the diffusion approximation of neutral Wright–Fisher models. Theor. Popul. Biol. 17, 37–50.

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

Griffiths, R.C., Tavaré, S., 2003. The genealogy of a neutral mutation. In: Green, P.J., Hjort, N.L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press, Oxford, pp. 393–413.

Hobolth, A., Uyenoyama, M., Wiuf, C., 2008. Importance sampling for the infinite sites model. Stat. Appl. Genet. Mol. Biol. 7, 32.

Innan, H., Tajima, F., 1997. The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. Genetics 147, 1431–1444.

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

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13, 235–248.

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

van Herwaarden, O.A., van der Wal, N.J., 2002. Extinction time and age of an allele in a large finite population. Theor. Popul. Biol. 61, 311–318.

Watterson, G.A., 1975. On the number of segregation sites. Theor. Popul. Biol. 7, 256–276.

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