Genetics: Published Articles Ahead of Print, published on May 17, 2010 as 10.1534/genetics.109.112847

Topologies of the Conditional Ancestral Trees and Full Likelihood-based Inference in the General Coalescent Tree Framework

Ori Sargsyan

Department of Organismic and Evolutionary Biology

Harvard University, Cambridge, MA 02138

Running Head: Inference in the general coalescent tree framework

Keywords: General coalescent tree framework, full likelihood-based inference, ages of mutations, topologies of conditional ancestral trees, Monte Carlo method

Corresponding author:

Ori Sargsyan

Biological Laboratories Room 4100

16 Divinity Ave.

Cambridge, MA 02138

Telephone: (617) 335-7854

FAX: (617) 384-5874

Email: sargsyan@fas.harvard.edu

Manuscript submitted to *Genetics*, April 22, 2010

Abstract

The general coalescent tree framework is a family of models for determining ancestries among random samples of DNA sequences at a non-recombining locus. The ancestral models included in this framework can be derived under various evolutionary scenarios. Here, a computationally tractable full likelihood-based inference method for neutral polymorphisms is presented, using the general coalescent tree framework and the infinite-sites model for mutations in DNA sequences. First, an exact sampling scheme is developed to determine the topologies of conditional ancestral trees. However, this scheme has some computational limitations and to overcome these limitations a second scheme based on importance sampling is provided. Next, these schemes are combined with Monte Carlo integrations to estimate the likelihood of full polymorphism data, the ages of mutations in the sample, and the time of the most recent common ancestor. In addition, this paper shows how to apply this method for estimating the likelihood of neutral polymorphism data in a sample of DNA sequences completely linked to a mutant allele of interest. This method is illustrated using the data in a sample of DNA sequences at the APOE gene locus.

INTRODUCTION

The interest in analyzing polymorphism data in contemporary samples of DNA sequences under various evolutionary scenarios creates a demand to design computationally tractable full likelihood-based inference methods. For an evolutionary scenario of interest, an ancestral-mutation model can be used to design such a method. The ancestral-mutation model for a sample of DNA sequences at a non-recombining locus is a combination of two processes: one is an ancestral process that traces the lineages of the sample back in time until the most recent common ancestor, constructing an ancestral tree for the sample. The second is a mutation process that is superimposed on the ancestral tree. The complexities of ancestral-mutation models make the design of such methods challenging. Full data is used instead of summary statistics, which can result in loss of important information in the data (see FELSENSTEIN 1992, DONNELLY and TAVARÉ 1995). In addition, current methods use specific features of the underlying ancestral-mutation models, so they lose flexibility to be applicable to other ancestral-mutation models.

More specifically, GRIFFITHS and TAVARÉ (1994c, 1995) and KUHNER *et al.* (1995) developed full likelihood-based inference methods for neutral polymorphisms at non-recombining locus. They used the combinations of the standard coalescent (KINGMAN 1982a,b,c, HUDSON 1983, TAJIMA 1983) with the finite-sites or infinite-sites (WATTERSON 1975) models as ancestral-mutation models. STEPHENS and DONNELLY (2000) designed an importance sampling method to estimate the full likelihood of the data using the same settings for the ancestral-mutation models. HOBOLTH *et al.* (2008) provided another importance sampling scheme restricted to the infinite-sites model. The last two methods are computationally more efficient than the first two methods, but they lose flexibility to be applicable to ancestral models without standard coalescent features with independent coalescence waiting times, such as the coalescent processes with exponential growth (SLATKIN and HUDSON 1991,GRIFFITHS and TAVARÉ 1994b).

To incorporate the coalescent processes with exponential growth, KUHNER *et al.* (1998) and

GRIFFITHS and TAVARÉ (1994a, 1999) extended their previous methods. For example, the method of GRIFFITHS and TAVARÉ (1994a, 1999) allows to consider ancestral models based on coalescent processes with variable population sizes. COOP and GRIFFITHS (2004) modified this inference method and made it applicable for analyzing full polymorphism data in a sample of DNA sequences from a non-recombining locus completely linked to a mutant allele of interest, either neutral or under selection. Additionally, ancestral models have been developed for this type of samples, where the mutant allele is either neutral (GRIFFITHS and TAVARÉ 1998, 2003, WIUF and DONNELLY 1999, STEPHENS 2000) or under selection (STEPHENS and DONNELLY 2003,SLATKIN and RANNALA 1997). The ancestral model of SLATKIN and RANNALA (1997) is part of a family of ancestral models derived by THOMPSON (1975), NEE *et al.* (1994), and RANNALA (1997), using a linear birth-death process as an evolutionary process in a population. Although all the ancestral models mentioned above differ in their properties and evolutionary scenarios, they are part of the general coalescent tree framework (GRIFFITHS and TAVARÉ 1998). Therefore, a computationally tractable full likelihood-based inference method based on this general framework is of great interest.

For a sample of n sequences, an ancestral model in the general coalescent tree framework is described as a bifurcating rooted tree with $n-1$ internal nodes and n leaves, where the internal nodes are coalescent events that happen one at a time. The tree is a combination of two independent components: the topology and the branch lengths. The topology of the tree is constructed going backwards in time by combining two randomly chosen ancestral lineages of the sample at each node; the branch lengths of the tree are defined by the joint distribution function of the coalescence waiting times. Note that any density function for coalescence waiting times can define an ancestral model in the general coalescent tree framework.

The n leaves (and the sequences in the sample) are labeled from 1 to n; and the $n-1$ internal nodes of the ancestral tree are labeled from 1 to $n-1$ (in order of occurrence of the coalescent events backwards in time). Thus, the topology of an ancestral tree is a leaf-labeled bifurcating rooted tree with totally ordered interior vertices. These trees are called topological trees.

When using the general coalescent tree framework and the infinite-sites model, an evolutionary

process that generates polymorphism data in a sample of DNA sequences can be described in the following way. An ancestral tree is constructed, as described above, and mutations are added independently on different branches of the ancestral tree as Poisson processes with equal rates, $\theta/2$, in which θ is the mutation rate at the locus. Then, at the mutation events, the ancestral sequences of the sample are changed according to the infinite-sites model; that is, each mutation occurs at a site of an ancestral sequence at which no previous mutations occurred. Thus, these changes define polymorphism data.

Naively, this probabilistic framework can be used to estimate the likelihood of the full observed data in a sample of n sequences. That is, data sets are simulated independently as described above and each simulated data set is compared to the observed data. The proportion of the simulated data sets that match the observed data is an estimate of the likelihood of the observed data. Although this approach provides an estimate for the likelihood of the observed data, this method is computationally infeasible, because the topologies of the ancestral trees of the generated data sets are sampled from the space of all the possible topological trees with n leaves. This space has size $n!(n-1)!/2^{n-1}$ (EDWARDS 1970), which is huge for moderate values of n. The topologies of the ancestral trees of the generated data sets that match the observed data represent a small portion of that space. Thus, designing a method that samples topologies of the ancestral trees from this subspace can make the method computationally tractable.

Based on this idea, I use the general coalescent tree framework with the infinite-sites model to develop a computationally tractable full likelihood-based inference method for polymorphisms in DNA sequences at a non-recombining locus. First, an exact sampling scheme for topologies of the conditional ancestral trees is developed. This method has some computational limitations, so to overcome these limitations a second scheme based on an importance sampling is provided. These sampling schemes are combined with Monte Carlo integrations to estimate the likelihood of the full data, the ages of the mutations in the sample, and the time of the most recent common ancestor of the sample. I describe an application of this method for neutral polymorphism data in a sample of DNA sequences at a non-recombining locus that is completely linked to a mutant allele

of interest, either neutral or under selection. The method is illustrated using the data in a sample of DNA sequences at the APOE gene locus from FULLERTON *et al.* (2000).

ESTIMATION OF THE LIKELIHOOD FUNCTION

The likelihood function: The naive approach described in the previous section is formalized and modified below to develop an efficient method for estimating the likelihood function $L(\boldsymbol{\alpha} \mid D)$, where D is the polymorphism data in the sample of n DNA sequences; and $\boldsymbol{\alpha}$ is a vector of parameters of the ancestral-mutation model in the general coalescent tree framework with the infinite-sites model. So, the elements of this vector can be the mutation rate θ and parameters in the distribution function of the coalescence waiting times, such as the growth rate in the exponential population growth model.

For estimating the likelihood function $L(\boldsymbol{\alpha} \mid D)$, the probability of the data $\mathbb{P}(D \mid \boldsymbol{\alpha})$ is estimated up to a multiplicative constant C, which does not depend on $\boldsymbol{\alpha}$. To estimate $\mathbb{P}(D \mid \boldsymbol{\alpha})$, first an expression is derived for it. Based on the naive approach, the equation

$$\mathbb{P}(D \mid \boldsymbol{\alpha}) = \mathbb{E}\mathbf{I}\{\Delta \cong D\} \tag{1}$$

holds, where $\mathbf{I}\{x\}$ is an indicator function that is 1 if x is true, or 0 if x is false. An ancestral tree with mutations Δ is constructed by combining the infinite-sites model with an ancestral model in the general coalescent tree framework. For each mutation in the sample, the set of labels of sequences (leaves) that carry that mutation is called a mutant group. The whole sample is considered as a mutant group created by a hypothetical mutation. The notation $\Delta \cong D$ means that Δ is consistent with D; that is, there is a one-to-one mapping between mutations in the sample and those in Δ such that the mutant groups of the corresponding mutations are the same. For ease of presentation, this paper assume that the ancestral nucleotide at each polymorphic site in the sample is known. (See DISCUSSION for other possibilities.)

Now some notations are needed. Let $\boldsymbol{T} = (T_2, \ldots, T_n)$ denote the waiting times between consecutive coalescent events in an ancestral model, where T_i is the time between coalescent events when the number of distinct ancestral lineages of the sample changes from i to $i-1$. Let $f(t_2, \ldots, t_n)$ be the joint density function of $\boldsymbol{T}$ and let Γ be the topology of the ancestral tree in Δ. The expectation in (1) is conditiond on the ancestral tree $(\Gamma, \boldsymbol{T})$ of Δ. That is, the equations

$$\mathbb{P}(D \mid \boldsymbol{\alpha}) = \mathbb{E}\mathbb{E}(\mathbf{I}\{\Delta \cong D\} \mid \Gamma, \boldsymbol{T}, \boldsymbol{\alpha}) = \mathbb{E}\mathbb{P}(\Delta \cong D \mid \Gamma, \boldsymbol{T}, \boldsymbol{\alpha}), \tag{2}$$

hold, where Γ and $\boldsymbol{T}$ are jointly distributed as $\mathbb{P}(\Gamma, \boldsymbol{T}) = f(t_2, \ldots, t_n)\mathbb{P}_n(\Gamma)$, and

$$\mathbb{P}_n(\Gamma) = \frac{2^{n-1}}{n!(n-1)!},\ \Gamma \in \Omega(n),$$

$\Omega(n)$ is the space of all the possible topological trees with n leaves.

Before modifying further the right side of (2), it is easy to see that if Δ is consistent with D then each mutation in D constrains the topology of the ancestral tree in Δ. The ancestral lineages associated with each mutant group coalesce within their group before they coalesce with other ancestral lineages of the sample. Thus, a topological tree is consistent with D if it satisfies the constraints created by all mutant groups in D. Let Γ_D denote such a topological tree and let $\Omega(D)$ be the space of this type of topological trees. Note that $\Omega(D)$ is a subspace of $\Omega(n)$.

From this observation, it follows that if a topological tree Γ is not consistent with D then the equation $\mathbb{P}(\Delta \cong D \mid \Gamma, \boldsymbol{T}, \boldsymbol{\alpha}) = 0$ holds. Hence, using the equation

$$\mathbb{P}_n(\Gamma_D) = \mathbb{P}_n(\Gamma_D \mid \Omega(D))\mathbb{P}_n(\Omega(D)),$$

and narrowing the space $\Omega(n)$ to $\Omega(D)$, equation (2) can be written as

$$\mathbb{P}(D \mid \boldsymbol{\alpha}) = \mathbb{P}_n(\Omega(D))\mathbb{E}\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha}), \tag{3}$$

where $(\Gamma_D, \boldsymbol{T})$ is distributed according to $f(t_2, \ldots, t_n)\mathbb{P}_D(\Gamma_D)$, and $\mathbb{P}_D(\cdot) \equiv \mathbb{P}_n(\cdot \mid \Omega(D))$.

Finally, the probability $\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha})$ can be simplified because for a given ancestral tree the mutations along different branches of the ancestral tree are the result of independent Poisson processes with equal rates, $\theta/2$. Thus, it is easy to see that the following equation holds:

$$\mathbb{P}(\Delta \cong D|\Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha}) = \prod_{x \in S'} \frac{(\frac{\theta}{2} L(x))^{m_x} e^{-\frac{\theta}{2} L_n}}{m_x!}, \tag{4}$$

where $L_n = \sum_{i=2}^{n} iT_i$ is the total length of the ancestral tree. S' is the set of mutation classes defined by consideration that two mutations in the sample are equivalent if they represent the same mutant group and m_x is the size of mutation class x. Let $L(x)$ be the time interval between two coalescent events in the ancestral tree $(\Gamma_D, \boldsymbol{T})$. In the first event, the mutant group corresponding to x has the most recent common ancestor. In the second event this mutant group shares, for the first time, an ancestor with other sequences in the sample. Let $J_1(x)$ and $J_0(x)$ denote the numbers of distinct ancestral lineages of the sample at the first and second events, respectively. So, $L(x)$ is equal to the sum $\sum_{i=J_1(x)}^{J_0(x)+1} T_i$.

Note that from (3) it follows that $\mathbb{EP}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha})$ is also the likelihood function because $\mathbb{P}_n(\Omega(D))$ depends only on the constraints of the data and not on θ and $\boldsymbol{T}$, hence, not on $\boldsymbol{\alpha}$. The formula in (4) is derived assuming that the order of the ages of mutations within the mutation classes is not known. However, when the order of the ages of mutations in the class x is known, the right side of (4) must be multiplied by $1/m_x!$ for each mutation class x in the sample. Because these multipliers depend only on data D, they do not change the likelihood function.

Algorithms for estimating the likelihood function: Below, I provided two algorithms for estimating the likelihood function using Monte Carlo integration based on (4) and (3).

Algorithm 1:

1. Sample a topological tree Γ_D from the probability space $(\Omega(D), \mathbb{P}_D(\cdot))$. Sample coalescence waiting times $\boldsymbol{T}$ using the joint density function $f(t_2, \ldots, t_n)$.

2. Compute the expression in the right side of (4).

3. Repeat Steps 1 and 2 many times and then average the computed values in Step 2. (According

to the Law of Large Numbers, the average will converge to the likelihood function as the number of the repetitions goes to infinity.)

The second algorithm is a modification of the first algorithm. Let $\mathbf{Q}_D(\cdot)$ be a probability distribution on the space $\Omega(D)$, such that it is positive for all topological trees Γ_D. Then equation (3) can be rewritten as

$$\mathbb{P}(D \mid \boldsymbol{\alpha}) = \mathbb{E}\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha}) W(\Gamma_D), \tag{5}$$

where the joint distribution of $(\Gamma_D, \boldsymbol{T})$ is $f(t_2, \ldots, t_n)\mathbf{Q}_D(\Gamma_D)$ and $W(\Gamma_D) \equiv \mathbb{P}_n(\Omega(D))\frac{\mathbb{P}_D(\Gamma_D)}{\mathbf{Q}_D(\Gamma_D)}$. Based on these modifications, I present the second algorithm.

Algorithm 2:

1. Sample a topological tree Γ_D from the probability space $(\Omega(D), \mathbf{Q}_D(\cdot))$, and sample coalescence waiting times $\boldsymbol{T}$ using the density function $f(t_2, \ldots, t_n)$.

2. Compute the expression $\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha}) W(\Gamma_D)$.

3. Repeat Steps 1 and 2 many times and then average the computed values in Step 2. (As in Algorithm 1, the average will converge to the likelihood function as the number of the repetitions goes to infinity.)

It is important to keep in mind some notes regarding the two algorithms. Algorithm 2 is an importance sampling method and similar algorithms are derived for estimating the ages of the mutations in the sample (see in a below section). The definition of the proposal distribution $\mathbf{Q}_D(\cdot)$ and computation of $W(\Gamma_D)$ is provided in the next sections. Two sampling methods are developed for topological trees Γ_D based on the distributions $\mathbb{P}_D(\cdot)$ and $\mathbf{Q}_D(\cdot)$. It follows from (4) that for the Monte Carlo methods described in the algorithms, only the values of $(J_1(\cdot), J_0(\cdot))$, not the whole topology of the ancestral tree, are necessary to know for each mutation class. In Algorithm 2, it might seem that to compute $W(\Gamma_D)$ the probability $\mathbb{P}_n(\Omega(D))$ needs to be computed, but that is not the case (see a below section).

TOPOLOGICAL STRUCTURE OF THE CONDITIONAL ANCESTRAL TREES

Representation of the polymorphism data in a tree structure: To understand the structure of the topological trees conditioned on the data D and how to sample them from the space $\Omega(D)$, the data D is represented in a tree structure form called mutation tree. The root of the mutation tree corresponds to the whole sample. The other nodes of the tree correspond to mutant groups (or mutation classes, because there is a one-to-one mapping between the mutant groups and the mutation classes) in the sample. The nodes are connected based on parent-child relationship inherited from corresponding mutant groups. For example, a mutant group A is the parent of a mutant group B (or B is a child of A) if B is nested in A and B is the biggest one nested in A. In other words, there is no other mutant group C in the sample, distinct from A and B, such that C is nested in A, and B is nested in C. A recursive procedure can be used to construct this tree by linking parent node to its child nodes, starting from the root node. An algorithm for building a mutation tree from the data is given in Supplementary Material; the algorithm is a slight modification of the algorithm of GUSFIELD (1991). For example, Figure 2 shows the mutation tree corresponding to the polymorphism data in Figure 1.

Mutation trees consisting of only parent (root) node and its children are called simple mutation trees. An example of a simple mutation tree is shown in Figure 3. Any mutation tree can be represented as a combination of the simple mutation trees associated with the nodes of the mutation tree. This representation is used to construct a topological tree consistent with the mutation tree. For example, Figure 4 shows the simple mutation trees embedded in the mutation tree in Figure 2.

A topological tree is consistent with a mutation tree if it satisfies the constraints created by the mutant groups corresponding to the nodes of the mutation tree. To construct a topological tree consistent with a mutation tree of D, first, topological trees are constructed consistent with each simple mutation tree embedded in the mutation tree. Second, all of these topological trees are then combined. I explain the first part of the procedure in a following section. In the second part of the procedure, the topological trees consistent with simple mutation trees of the mutation

tree are changed recursively. This procedure starts with the topological tree that is consistent with the simple mutation tree of the root node of the mutation tree. In this tree, topological subtrees associated with mutant groups of child nodes of the root are cut and respectively replaced by the topological trees consistent with simple mutation trees of the child nodes of the root. The order of the coalescent events do not change during these cut-and-paste procedures. That is, the topological tree that is pasted inherits the labels of the coalescent events in the topological tree that was cut, and the coalescent events stay in the same order as their labels on the updated topological tree. As a result of this sampling procedure, the labeled coalescent events get assigned to the mutant groups. A realization of the procedure for the mutation tree in Figure 2 is shown in Figure 5. In the next section, I describe a probabilistic approach for sampling a topological tree consistent with a simple mutation tree.

THE PROBABILITIES OF THE TOPOLOGIES OF THE CONDITIONAL ANCESTRAL TREES

The probability of a simple mutation tree: The topological trees with n leaves and consistent with a mutation tree constitute a subset of the set $\Omega(n)$. The probability of that mutation tree is defined as the probability of that subset in the probability space $(\Omega(n), \mathbb{P}_n(\cdot))$. The probability of a mutation tree can be computed using the following lemma, which is based on the computation of the probabilities of the simple mutation trees.

Lemma 1. *The probability of a mutation tree is the product of the probabilities of the simple mutation trees embedded in that mutation tree.*

Lemma 1 easily follows from the sampling procedure described in the previous section.

To compute the probability of a simple mutation tree the following notations are needed. A simple mutation tree can be represented as a vector $(\boldsymbol{A}, C) = (A_1, \ldots, A_k, C)$, where in this mutation tree the child nodes corresponding to mutant groups of size greater than 1 are labeled as

$A_1, \ldots, A_k$, C_1 is the set of the child nodes corresponding to mutant groups of size 1, C_0 is the set of the labels in the mutant group of the root node that do not belong to any of the mutant groups corresponding to the child nodes of the simple mutation tree. Note that the mutant groups of size 1 in a simple mutation tree do not constraint the topological trees consistent with the simple mutation tree and k is defined as the number of mutation constraints for the simple mutation tree $(\boldsymbol{A}, C)$. When there is no confusion, instead of $(\boldsymbol{A}, C)$ the numeric vector $(\boldsymbol{a}, c) = (a_1, \ldots, a_k, c)$ is used, where a_i is the size of the mutant group corresponding to A_i; c_1 and c_0 are the sizes of the sets C_1 and C_0, respectively, and c is the sum $c_1 + c_0$. Let m be the sum $c + \sum_{i=1}^{k} a_i$.

Let $N_m(\boldsymbol{a}, c)$ be the total number of topological trees with m leaves and consistent with $(\boldsymbol{A}, C)$. These topological trees can be classified into disjoint groups according to the first coalescent event, backwards in time. That is, the coalescent event could happen with two ancestral lineages associated with a mutant group corresponding to one of the A_i, $i = 1, \ldots, k$, or with C. So, the following equation holds:

$$N_m(\boldsymbol{a}, c) = \sum_{i=1}^{k} \binom{a_i}{2} N_{m-1}(\boldsymbol{a} - \boldsymbol{e}_i, c) + \binom{c}{2} N_{m-1}(\boldsymbol{a}, c-1), \tag{6}$$

where $\boldsymbol{e}_i$ is a vector with k elements, the ith element is 1, and the others are zero.

According to the above notation, $N_m(m)$ is the number of all possible topological trees with m leaves, so

$$N_m(m) = \frac{m!(m-1)!}{2^{m-1}}$$

and the probability of having a topological tree consistent with the simple mutation tree $(\boldsymbol{A}, C)$ is

$$\mathbb{P}_m(\boldsymbol{a}, c) = \frac{N_m(\boldsymbol{a}, c)}{N_m(c + \sum_{i=1}^{k} a_i)}.$$

So, from (6) it follows that the probabilities of the simple mutation trees satisfy the following

recursion formula:

$$\binom{\sum_{i=1}^{k} a_i + c}{2} \mathbb{P}_m(\boldsymbol{a}, c) = \sum_{i=1}^{k} \binom{a_i}{2} \mathbb{P}_{m-1}(\boldsymbol{a} - \boldsymbol{e}_i, c) + \binom{c}{2} \mathbb{P}_{m-1}(\boldsymbol{a}, c-1), \tag{7}$$

with the initial conditions

$$\mathbb{P}_c(c) = 1, \text{ for any positive integer } c.$$

For consistency with notations above, it is important to note that if on the right side of (6) and (7) the size of one of the mutant groups corresponding to A_j $(j = 1, \ldots, k)$ becomes 1 the first time, let it happen with A_i, then the expressions $N_{m-1}(\boldsymbol{a} - \boldsymbol{e}_i, c)$ and $\mathbb{P}_{m-1}(\boldsymbol{a} - \boldsymbol{e}_i, c)$ should be replaced by $N_{m-1}(\boldsymbol{a_{(-i)}}, c+1)$ and $\mathbb{P}_{m-1}(\boldsymbol{a_{(-i)}}, c+1)$, respectively, where $\boldsymbol{a_{(-i)}}$ is as vector $\boldsymbol{a}$ with deleted i element.

Although the probability of a simple mutation tree can be computed using (7), this will be inefficient when the number of mutation constraints of the simple mutation tree or the sizes of the mutant groups corresponding to the child nodes of the simple mutation tree are large. In general, it appears to be a difficult problem to find explicit formulas for the probabilities of the simple mutation trees. For simple mutation trees of the form (a_1, c), the following formula (WIUF and DONNELLY 1999) can be used:

$$\mathbb{P}_m(a_1, c) = \frac{2}{(a_1+1)\binom{c+a_1-1}{a_1-1}}, \; m = a_1 + c. \tag{8}$$

Below, I derive explicate formulas for the probability of a simple mutation tree of the form (a_1, a_2, c).

Lemma 2. *The number of the topological trees consistent with simple mutation tree of the form* (a_1, a_2, c) *can be computed using the following formulas: If* $c \neq 0$*, then*

$$a) \;\; N_m(a_1, a_2, c) = \frac{Q(a_1, a_2, c)(a_1 + a_2 + c)!}{2(a_1+1)!(c-1)!(a_2+1)!} \Phi(a_1, a_2, c), \; m = a_1 + a_2 + c,$$

otherwise,

$$b)\ \ N_m(a_1,a_2,0)=\frac{a_1!(a_1-1)!}{2^{a_1-1}}\frac{a_2!(a_2-1)!}{2^{a_2-1}}\binom{a_1+a_2-2}{a_1-1},\ m=a_1+a_2.$$

The expressions for $\Phi(a,b,c)$ *and* $Q(a,b,c)$ *are defined as*

$$\Phi(a,b,c)\equiv 2\left(\frac{2ab+a^2b^2}{(a+b-1)(a+b)(a+b+1)}+\frac{ab}{a+b}+c\right)$$

and

$$Q(a,b,c)\equiv\frac{a!(a-1)!}{2^{a-1}}\frac{b!(b-1)!}{2^{b-1}}\frac{c!(c-1)!}{2^{c-1}}.$$

The proof of Lemma 2 is given in Supplementary Material.

Remark: From Lemma 2 it follows that if $c\neq 0$, then

$$\mathbb{P}_m(a_1,a_2,c)=\frac{2(a_1-1)!(a_2-1)!c!\Phi(a_1,a_2,c)}{(a_1+a_2+c-1)!(a_1+1)(a_2+1)},\ m=a_1+a_2+c, \tag{9}$$

otherwise,

$$\mathbb{P}_m(a_1,a_2,0)=\frac{2a_1!a_2!}{(a_1+a_2-1)(a_1+a_2)!},\ m=a_1+a_2.$$

SAMPLING ALGORITHMS FOR THE TOPOLOGIES OF THE CONDITIONAL ANCESTRAL TREES

A sampling method from $\Omega(D)$ **based on** $\mathbb{P}_D(\cdot)$: An algorithm for sampling topological trees consistent with simple mutation trees from the probability space $(\Omega(D),\mathbb{P}_D(\cdot))$ is developed, recall that $\mathbb{P}_D(\cdot)$ is defined as $\mathbb{P}_n(\cdot\mid\Omega(D))$. It is then extended for the general case.

Let a simple mutation tree have the form $(\boldsymbol{A},C)=(A_1,\ldots,A_k,C)$. To sample a topological tree from the probability space $(\Omega(\boldsymbol{A},C),\mathbb{P}_{(\boldsymbol{A},C)}(\cdot))$, where $\mathbb{P}_{(\boldsymbol{A},C)}(\cdot)\equiv\mathbb{P}_m(\cdot\mid\Omega(\boldsymbol{A},C))$,

$m = c + \sum_{i=1}^{k} a_i$, equation (7) is used. The sampling procedure is based on recursively constructing coalescent events, backwards in time, by finding which two lineages should coalesce in each coalescent event. So two ancestral lineages that coalesce at the first coalescent event belong to one of the mutant groups of $A_1, \ldots, A_k$ or in group C. To find the group in which they are a decision based on (7) is made using the following probabilities: the coalescent event occurs with two ancestral lineages in the mutant group corresponding to A_j with probability

$$\frac{\binom{a_j}{2}\mathbb{P}_{m-1}(\boldsymbol{a} - \boldsymbol{e}_j, c)}{\binom{\sum_{i=1}^{k} a_i + c}{2}\mathbb{P}_m(\boldsymbol{a}, c)}$$

or in group C with probability

$$\frac{\binom{c}{2}\mathbb{P}_{m-1}(\boldsymbol{a}, c-1)}{\binom{\sum_{i=1}^{k} a_i + c}{2}\mathbb{P}_m(\boldsymbol{a}, c)}.$$

If the two coalescing lineages are in group C, then both lineages are from group C_1 or from C_0; otherwise, one of them is from group C_1, and the other one is from C_0. These three possibilities have the following probabilities:

$$\frac{c_1(c_1-1)}{c(c-1)}, \ \frac{c_0(c_0-1)}{c(c-1)}, \ \frac{2c_1c_0}{c(c-1)}. \tag{10}$$

If the coalescent event is in the mutant group of A_i, then randomly choosing two ancestral lineages associated with this group and combining them in one lineage. The two lineages are replaced by the new lineage and a_i is decreased by 1. If a_i is 1, then the node corresponding to this mutant group is moved to group C_1; c_1 and c are increased by 1 and the value of k is decreased by 1. If the coalescent event is in group C_0, then two ancestral lineages from this group are chosen randomly and combined into a new lineage. They are replaced by the new lineage and c_0 and c are decreased by 1. If the coalescent event is in group C_1, then two ancestral lineages associated to the group C_1 are chosen randomly and combined together. The new lineage is moved to group C_0

and c_1 is decreased by 2, c_0 and c are increased by 1. The last case involves randomly choosing a lineage from C_1 and C_0. After combining them, c_1 and c are decreased by 1. So, after changing the two vectors $(A_1, \ldots, A_k, C)$ and $(a_1, \ldots, a_k, c)$, the procedure is applied to this new simple mutation tree. Recursively repeating this procedure $\sum_{i=1}^{k} a_i + c - 1$ times, the construction of the topological tree will be completed.

In this sampling method, it is necessary for the probabilities of the simple mutation trees in the decision-making steps to have already been computed. Although these probabilities can be computed recursively using (7), doing so can be very inefficient. However, this sampling method can be applied to simple mutation trees of the forms (C), (A_1, C), and (A_1, A_2, C) because explicit formulas are given for the probabilities of these types of simple mutation trees (see above). To overcome this limitation, a second sampling method is presented below, which does not require knowing the values of the probabilities.

The second sampling method based on $\mathbf{Q}_D(\cdot)$: First, a proposal distribution $\mathbf{Q}_{(\boldsymbol{A},C)}(\cdot)$ is defined for the simple mutation tree $(\boldsymbol{A}, C) = (A_1, \ldots, A_k, C)$ and then extended for $\mathbf{Q}_D(\cdot)$. To define proposal distribution $\mathbf{Q}_{(\boldsymbol{A},C)}(\cdot)$, below equation (7) is rewritten in another form (see (11)). This step is similar to the approach used by GRIFFITHS and TAVARÉ 1994c.

After rewriting (7) as

$$\mathbb{P}_m(\boldsymbol{a}, c) = \frac{\sum_{i=1}^{k} \binom{a_i}{2} + \binom{c}{2}}{\binom{\sum_{i=1}^{k} a_i + c}{2}} \left(\sum_{j=1}^{k} \frac{\binom{a_j}{2}}{\sum_{i=1}^{k} \binom{a_i}{2} + \binom{c}{2}} \mathbb{P}_{m-1}(\boldsymbol{a} - \boldsymbol{e}_j, c) + \frac{\binom{c}{2}}{\sum_{i=1}^{k} \binom{a_i}{2} + \binom{c}{2}} \mathbb{P}_{m-1}(\boldsymbol{a}, c-1) \right), \tag{11}$$

a new decision-making rule is defined as follows: the coalescent event occurs with two ancestral

lineages associated with the mutant group of A_j with probability

$$\frac{\binom{a_j}{2}}{\sum_{i=1}^{k}\binom{a_i}{2}+\binom{c}{2}}$$

or with group C with probability

$$\frac{\binom{c}{2}}{\sum_{i=1}^{k}\binom{a_i}{2}+\binom{c}{2}}.$$

When the two lineages are from group C, the coalescent probabilities of the three possibilities for the two lineages are the same as in (10).

So, using the previous sampling procedure with this decision-making rule, a topological tree, $\Gamma_{(\boldsymbol{A},C)}$, will be sampled consistent with $(\boldsymbol{A}, C)$. This sampling defines the proposal distribution $\mathbf{Q}_{(\boldsymbol{A},C)}(\cdot)$ on the space $\Omega(\boldsymbol{A}, C)$. As shown below, several quantities must also be computed during this sampling method. For the simple mutation tree, $(a_1, \ldots, a_k, c)$, the expression $W(a_1, \ldots, a_k, c)$ is computed before each decision-making step using the following formula:

$$W(a_1, \ldots, a_k, c) = \frac{\sum_{i=1}^{k}\binom{a_k}{2}+\binom{c}{2}}{\binom{\sum_{i=1}^{k} a_i + c}{2}}.$$

At the end of the procedure for sampling topological tree $\Gamma_{(\boldsymbol{A},C)}$, these quantities, $W(a_1, \ldots, a_k, c)$, are multiplied together and the result is denoted by $W(\Gamma_{(\boldsymbol{A},C)})$. So, from equation (11) and definition of $\mathbf{Q}_{(\boldsymbol{A},C)}(\cdot)$ it easily follows that the following relation holds for the probabilities of $\Gamma_{(\boldsymbol{A},C)}$:

$$\mathbb{P}_m(\Gamma_{(\boldsymbol{A},C)} \mid \Omega(\boldsymbol{A}, C)) = \frac{W(\Gamma_{(\boldsymbol{A},C)})}{\mathbb{P}_m(\boldsymbol{a}, c)}\mathbf{Q}_{(\boldsymbol{A},C)}(\Gamma_{(\boldsymbol{A},C)}). \tag{12}$$

In the previous section, a procedure was described for sampling a topological tree consistent with a mutation tree. This method allows one to define both a proposal distribution $\mathbf{Q}_D(\cdot)$ using

$\mathbf{Q}_{(\boldsymbol{A},C)}(\cdot)$ and a sampling scheme from the space $(\Omega(D), \mathbf{Q}_D(\cdot))$. Another form of this sampling scheme is in Algorithm 3 listed below. This algorithm can be modified to make it applicable for sampling from the space $(\Omega(D), \mathbb{P}_n(\cdot \mid D))$.

Let i be a variable with the values of the labels of the coalescent events, let p be a simple mutation tree associated with a node of the mutation tree, and let p have the form $(A_1, \ldots, A_k, C)$. It is easy to see that the following equations hold for each mutant group x. In a sampled topological tree consistent with D, the labels of the two coalescent events (when mutant group x has the most recent common ancestor and when, for the first time, shares an ancestor with others in the sample) are respectively equal to $n - J_1(x)$ and $n - J_0(x)$.

Algorithm 3:

1. Assign 1 to i, the simple mutation tree of the root of the mutation tree to p, 1 to $W(\Gamma_D)$, n to $J_1(\cdot)$ and 1 to $J_0(\cdot)$ (for all mutant groups).

2. Compute the following expressions for p:

$$W(a_1, \ldots, a_k, c) = \frac{\sum_{i=1}^{k} \binom{a_k}{2} + \binom{c}{2}}{\binom{\sum_{i=1}^{k} a_i + c}{2}}$$

   and assign $W(\Gamma_D)W(a_1, \ldots, a_k, c)$ to $W(\Gamma_D)$.

3. Apply the decision-making rule in the second sampling method (based on (7)) to p to find in which group the i coalescent event occurs.

4. If the i coalescent event is in group C, then there are the following three possibilities: the two coalescing lineages are (1) from group C_1, or (2) from C_0, or (3) from group C_0 and from group C_1. For the first possibility, randomly choose two lineages (without replacement) from the mutant groups of the nodes in C_1 and combine the two lineages together. Assign $n - i$ to $J_0(\cdot)$ for these two mutant groups, and move the new lineage to C_0. Delete the mutant groups of the two lineages from the group C_1, and update the sizes of the groups. For the second

possibility, randomly choose two lineages from group C_0 and combine them to make a new lineage in C_0. Update the size of C_0 decreasing it by 1. For the third possibility, randomly choose one lineage from group C_1 and one from group C_0, and combine them in one lineage. Assign $n-i$ to $J_0(\cdot)$ for the mutant group from C_1, delete this node from C_1 and decrease c_1 by 1.

5. If the i coalescent event occurs in group A_j, then check the size of the mutant group corresponding to A_j. If the size of that group is greater than 2, assign the simple mutation tree of node A_j to p, keep the value of i and go to Step 2. Otherwise, assign $n-i$ to $J_1(A_j)$ (this coalescent event defines $J_1(\cdot)$ for this mutant group). The simple mutation tree associated with the node of this mutant group has zero mutation constraint, $k=0$, and it has the form (c_1, c_0), for which it is $(2,0)$ or $(1,1)$ or $(0,2)$. So, in all these cases, the two lineages coalesce and the mutation tree updates according to the three cases.

6. Increase the value of i by 1 (that is, move to the next coalescent event) and assign the simple mutation tree of the root of the updated mutation tree to p. Then go to Step 2.

Remark: An equation similar to (12) holds for this general case:

$$\mathbb{P}_n(\Gamma_D \mid \Omega(D)) = \frac{W(\Gamma_D)}{\mathbb{P}_n(\Omega(D))}\mathbf{Q}_D(\Gamma_D). \tag{13}$$

This formula results from the following steps: write equation (11) for each node of the mutation tree and expand the right sides of these equations recursively, then multiply all these equations together. The left side of the new equation will be the product of the probabilities of the simple mutation trees associated with each node of the mutation tree. However, this product according to Lemma 1 is the probability of Γ_D in the space $\Omega(n)$. Thus, after expanding the products on the right side of the new equation and dividing both sides of the equation on $\mathbb{P}_n(\Omega(D))$ and using the definitions of $\mathbf{Q}_D(\cdot)$ and $W(\Gamma_D)$, the proof is complete.

AGES OF MUTATIONS CONDITIONAL ON THE DATA

The order of the ages of mutations in a mutation class is known: This section describes how to infer the ages of mutations in the sample conditional on full polymorphism data D. Let η be the age of a mutation in a mutation class x. Using similar arguments as in the derivation of equation (3), the mean of the functional $\lambda(\eta)$ conditional on the data D can be expressed as

$$\mathbb{E}(\lambda(\eta) \mid D) = \frac{\mathbb{E}\mathbb{E}(\lambda(\eta) \mid \Delta \cong D, \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha})\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha})}{\mathbb{P}(D \mid \boldsymbol{\alpha})}. \tag{14}$$

To estimate the first and second moments of η, and its distribution function at the point t, the functional $\lambda(\eta)$ is considered to be equal to η, η^2 and $\mathbf{I}\{\eta \leq t\}$, respectively. Because mutations are superimposed independently on different branches of the ancestral tree of the sample as Poisson processes with equal rates, $\theta/2$, it follows that conditional on $\Delta \cong D, \Gamma_D, \boldsymbol{T}$ the age η of a mutation in mutation class x is uniformly distributed on the interval $(S_{J_1(x)+1}, S_{J_0(x)+1})$, where $S_j = \sum_{i=j}^{n} T_i$. Based on this fact, $\eta \mid \Delta \cong D, \Gamma_D, \boldsymbol{T}$ is equal to $S_{J_1(x)+1} + U(S_{J_0(x)+1} - S_{J_1(x)+1})$, in distribution, where U is a uniform random variable on the interval $(0, 1)$. From this representation, it follows that if there is no extra information about the order of the ages of mutations in mutation class x, the ages of these mutations are conditionally independent. But, if the order of the ages of the mutations is known, then after adding this condition, the joint conditional distribution of the ages of mutations in mutation class x (ranked in increasing order) is distributed as the vector $S_{J_1(x)+1} + \boldsymbol{U}_o(S_{J_0(x)+1} - S_{J_1(x)+1})$, where $\boldsymbol{U}_o = (U_{(1)}, \ldots, U_{(m_x)})$ is the ordered statistics of m_x independent uniform random variables on the interval $(0, 1)$. Note that $U_{(k)}$ has beta distribution $B(\alpha, \beta)$, where $\alpha = k$ and $\beta = m_x + 1 - k$ (see, e.g., DEVROYE 1986). Thus, the following Monte Carlo method can be used to estimate the first and the second moments of the age of the kth mutation in the mutation class x and its distribution at point t: Sample $(\Gamma_D, \boldsymbol{T})$ many times and compute the following expressions:

$$\frac{\alpha S_{J_0(x)+1} + \beta S_{J_1(x)+1}}{\alpha + \beta}\mathbb{P}(\Delta \cong D \mid \Gamma_D, \boldsymbol{T}, \boldsymbol{\alpha}), \tag{15}$$

$$\frac{\alpha(\alpha+1)S^2_{J_0(x)+1}+2\alpha\beta S_{J_0(x)+1}S_{J_1(x)+1}+\beta(\beta+1)S^2_{J_1(x)+1}}{(\alpha+\beta)(1+\alpha+\beta)}\mathbb{P}(\Delta\cong D\mid\Gamma_D,\boldsymbol{T},\boldsymbol{\alpha}),\qquad(16)$$

and

$$\mathbb{P}\left(U_{(k)}\le\frac{t-S_{J_1(x)+1}}{S_{J_0(x)+1}-S_{J_1(x)+1}}\right)\mathbb{P}(\Delta\cong D\mid\Gamma_D,\boldsymbol{T},\boldsymbol{\alpha}).\qquad(17)$$

Average the computed values, respectively, and divide by the estimate of $\mathbb{P}(D\mid\boldsymbol{\alpha})$. Note that for evaluating the probabilities above, use the following formula

$$\mathbb{P}(U_{(k)}\le u)=\frac{m_y!}{(k-1)!(m_y-k)!}\int_0^u v^{\alpha-1}(1-v)^{\beta-1}dv,\ 0<u<1,$$

where $\mathbb{P}(U_{(k)}\le u)$ is equal to 0 and 1 when $u\le 0$ and $u\ge 1$, respectively.

Note that if the order of the ages of mutations in mutation class x is unknown, then the a similar approach as above can be used to estimate the ages of the mutations.

Time to the most recent common ancestor: Similarly, as in the method of the previous section, a Mote Carlo approach can be applied to infer the time of the most recent common ancestor of the sample, T_{MRCA}, conditional on D. To estimate the first and second moments, and the distribution of T_{MRCA}, the method above should be modified by using the following expressions instead of (15)-(17), respectively:

$$S_2\mathbb{P}(\Delta\cong D\mid\Gamma_D,\boldsymbol{T},\boldsymbol{\alpha}),$$

$$S_2^2\mathbb{P}(\Delta\cong D\mid\Gamma_D,\boldsymbol{T},\boldsymbol{\alpha}),$$

$$\mathbf{I}\{S_2\le t\}\mathbb{P}(\Delta\cong D\mid\Gamma_D,\boldsymbol{T},\boldsymbol{\alpha}).$$

APPLICATIONS

Inference for full polymorphism data completely linked to a mutant allele: To show flexibility and applicability of the inference method developed in the previous sections, below I describe applications of the method for neutral polymorphism data in a sample of DNA sequences completely

linked to a mutant allele, either neutral or under selection.

To formalize the problem the following notations are needed. Assume that n DNA sequences at a non-recombinant locus are drawn at random from a population and b sequences in this sample carry an allele, which is the result of a unique mutation event. Let D_b denote the polymorphism data just among the b sequences and assume that the polymorphism data is neutral and consistent with the infinite-sites model. Here the data is the pair (b, D_b). Using this method, the probability of the data (b, D_b), and the distributions of the ages of the mutant allele and mutations in b sequences are inferred.

In the literature, many probabilistic frameworks have been developed for modeling the ancestral tree of the b sequences. GRIFFITHS and TAVARÉ (1998, 2003) developed a theory for extracting the ancestral tree of the b sequences from the ancestral tree of the n sequences for which the ancestral tree is modeled in the general coalescent tree framework when the mutant allele is neutral. (For special cases of this framework such as the standard coalescent and the coalescent with exponential growth, also see WIUF and DONNELLY (1999) and STEPHENS (2000).)

STEPHENS and DONNELLY (2003) and COOP and GRIFFITHS (2004) developed ancestral models for the b sequences, where the n sequences are drawn at random from an equilibrium population with a mutant allele under selection. In all of these models, the ancestral tree of the b sequences is extracted from the conditional ancestral tree of the n sequences. SLATKIN and RANNALA (2000, 1997) used a linear birth-death process to model the ancestral tree of the b sequences. This model is different from the previous two models in a few aspects: this model assumes that the mutant allele is in low frequency in the population and under additive selection, the implementation of this model is computationally less complex, and it allows population growth. Although the above models are different in their properties, they all fit to the general coalescent tree framework.

To estimate the probability of the data (b, D_b), this probability is represented as a product of two probabilities:

$$\mathbb{P}(b, D_b) = \mathbb{P}(D_b \mid b)\mathbb{P}(b).$$

Thus, estimation of the probabilities $\mathbb{P}(D_b \mid b)$ and $\mathbb{P}(b)$ is explained below.

The neutral case: For the case of the neutral mutant allele, the probability $\mathbb{P}(b)$ can be estimated using the formula (3.3) derived by GRIFFITHS and TAVARÉ (1998):

$$\mathbb{P}(b) = \frac{\mathbb{E}\sum_{k=2}^{n-b+1} kp_{n,k}(b)T_k}{\mathbb{E}\sum_{k=2}^{n} kT_k}.$$

The theory developed by GRIFFITHS and TAVARÉ (1998) can be used to derive the following formula:

$$\mathbb{P}(D_b \mid b) = \frac{\mathbb{E}\sum_{k=2}^{n-b+1} kp_{n,k}(b)T_k\mathbb{P}(D_b|b, \Gamma_{D_b}, \boldsymbol{J}(k), \boldsymbol{T}, k)}{\mathbb{E}\sum_{k=2}^{n-b+1} kp_{n,k}(b)T_k}, \tag{18}$$

where $\boldsymbol{T} = (T_2, \ldots, T_n)$ represents the coalescence waiting times on the ancestral tree of n sequences, and the joint density function $f(t_2, \ldots, t_n)$ of $\boldsymbol{T}$ is not conditioned on the existence of the unique mutant allele. Here $\boldsymbol{J}(k) = (j_1, \ldots, j_{b-1})$ is a random vector uniformly distributed on

$$\{(j_1, \ldots, j_{b-1}) : k \leq j_1 < j_2 \cdots < j_{b-1} \leq n-1\}.$$

The quantity $\mathbb{P}(D_b|b, \Gamma_{D_b}, K, \boldsymbol{J}(k), \boldsymbol{T})$ is computed as the expression in equation (4), but, instead of n, D, and T, the following quantities are used, respectively: b, D_b, and $\tilde{\boldsymbol{T}} = (\tilde{T}_2, \ldots, \tilde{T}_b)$; $\tilde{T}_i$ is defined as follows

$$\tilde{T}_i = S_{j_{i-1}+1} - S_{j_i+1},\ i = 2, \ldots, b, \text{and } j_b = n,$$

where $S_i = \sum_{j=i}^{n} T_j$.

Similar expressions can be derived for estimating the ages of the mutant allele and mutations in b sequences conditional on (b, D_b). See the Supplementary Material for more details about these expressions, and algorithms based on Monte Carlo integrations for estimating these quantities.

The selection case: SLATKIN and RANNALA (2000, 1997), STEPHENS and DONNELLY (2003), and COOP and GRIFFITHS (2004) developed ancestral models for the b sequences when the mutant allele is under selection. So, to estimate the probability of $\mathbb{P}(D_b \mid b)$ when using one of these models, only the coalescence waiting times of the ancestral tree of the b sequences are needed for application of the estimation method developed in the previous section. When using the model by

SLATKIN and RANNALA (1997) the probability $\mathbb{P}(b)$ can be estimated using their equation (7). For the other models, this probability can be evaluated using the equation (24) by GRIFFITHS (2003), or the approach suggested by STEPHENS and DONNELLY (2003).

APOE data analysis: In humans, the APOE gene is located on chromosome 19. Many studies have suggested that some variations of this gene are associated with coronary artery disease and Alzheimer disease. Thus, it is important to understand the impact of evolutionary processes on genetic variation of this gene. FULLERTON *et al.* (2000) sequenced the APOE gene locus (5.5kbp long) from 96 individuals (192 sequences) from four different populations world wide; 24 individuals (48 chromosomes) from each population. One of the nucleotide sites in this gene (at position 3937, see Table 1 by FULLERTON *et al.* 2000) is polymorphic in all human populations, but not in nonhuman primate relatives that carry the ancestral haplotype (see FULLERTON *et al.* 2000 and references therein). Out of 192 sequences, 165 carry the mutant allele at this position. To illustrate, the inference methods developed in this paper are used to analyze the polymorphism data in the 165 sequences that carry the mutant allele. The results of this analysis are contrasted with the inference based on the full polymorphism data in the 192 sequences.

To analyze to full polymorphism data in the 192 sequences, FULLERTON *et al.* (2000) used the Genetree 9.0 program, written by R.C. Griffiths. The program uses the methods developed by GRIFFITHS and TAVARÉ (1994a,b,c, 1999). In this analysis, the population exponential growth model is considered and the maximum-likelihood is used to estimate the underlying parameters in the model: that is, the mutation rate θ at the locus and the exponential growth rate β. Using the likelihood-ratio test, the null hypothesis ($\beta = 0$) was not rejected. They did not find a significant signal for population exponential growth. When $\beta = 0$, their maximum-likelihood estimate of the mutation rate θ is 3.66.

This paper also considers the population exponential growth model as the underlying evolutionary scenario for the data in the sample. For consistency with the infinite-sites model and non-recombining locus conditions, 4 sequences were removed from the whole sample and the polymorphism data was restricted to a 4528bp long region between positions 833 and 5360. First, the

maximum-likelihood estimates of the mutation rate θ at the locus and the exponential growth rate β are derived based on the full polymorphism data in the 165 sequences that are completely linked to the mutant allele. The estimates are $\theta = 2.75$ and $\beta = 0$. The inference is done using software that implements the method developed in this paper (the software is written in the C programing language and available from the author upon request). Ten independent runs were used with 7×10^7 iterations each (as in Algorithm 5 in Supplementary Material); the likelihood surface for the parameters (θ, β) is estimated at the points (θ, β), where θ is in the range of 0.5 to 7.25 (10 equidistant points) and β is in the range of 0 to 4 (5 equidistant points). The contour plot of the likelihood surface is shown in Figure 6. The time required for one run is 58 hours. Note that the likelihood function at the 50 points (θ, β) are estimated simultaneously in one run. This estimation approach is appropriate because in this method the topological trees and coalescence times are sampled independently. In addition, the coalescence times in the exponential growth model can be presented as a transformation from coalescence times in the standard coalescent, hence the sampling of the coalescence times can be done by sampling coalescence times under the standard coalescent and then using the transformation for each β to convert them, respectively. For these estimated parameter values, the mean age of the mutant allele is 1.66 with a standard deviation of 0.89. Time is measured in $2N_e$ generations, as a human effective population size, N_e, can be considered 10000, with 20 years for a generation time.

The inference methods have also been applied to full polymorphism data in the 188 sequences and the results are contrasted with above estimates. The contour plot of the likelihood surface of (θ, β) is shown in Figure 7. The range for θ is from 0.5 to 9.4, and it is from 0 to 9.3 for β; for each parameter 16 equidistant points are considered. To compute the likelihood function for these parameter values one simulation run was done with 6×10^7 iterations and 10 independent runs. The time required for one run is 58 hours. The maximum-likelihood estimates of the parameters are $\theta = 4.0$ and $\beta = 2.5$. To test for the population exponential growth, mutation rate θ is also estimated when $\beta = 0$. In this case, the maximum-likelihood estimate of the mutation rate θ is 2.87. The likelihood-ratio statistic is 2.3, which is not significant. The mean time to the most

recent common ancestor is 1.26 with a 0.3 standard deviation. The mean age of the mutant allele is 0.83 with a standard deviation of 0.25.

Note that the mean age of the mutant allele based only on its frequency in the sample is 1.85 (using the formula by KIMURA and OHTA (1973) or by GRIFFITHS and TAVARÉ 1999). The WATTERSON (1975) estimate of the mutation rate at this locus is 2.4. All of these estimates of the mutation rate at the locus and the age of the mutant allele derived in this section do not provide evidence that the mutant allele increased its frequency because of selection.

DISCUSSION

This paper developed a computationally tractable full likelihood-based inference method using the general coalescent tree framework with the infinite-sites model for polymorphism data in a sample of DNA sequences at a non-recombining locus. This method is different from existing full likelihood-based inference methods in several ways. First, it is applicable to a wide range of ancestral models that fit into the general coalescent tree framework, which includes various demographic scenarios for populations. Second, the method was developed given the assumption of the infinite-sites model for mutation scheme in DNA sequences, a non-recombining locus, and no population structure. In contrast, full-likelihood based inference methods were developed for evolutionary scenarios such as population subdivision (BAHLO and GRIFFITHS 2000, DE IORIO and GRIFFITHS 2004, DE IORIO *et al.* 2005), or recombination at the locus (GRIFFITHS and MARJORAM 1996, FEARNHEAD and DONNELLY 2001), or other mutation schemes such as stepwise mutation models (NIELSEN 1997, DRUMMOND *et al.* 2002, WILSON and BALDING 1998, STEPHENS and DONNELLY 2000), assuming large constant size for a population or subpopulations in Wright-Fisher models.

The construction of the importance sampling method developed in this paper differentiates this method from other methods. In the other importance sampling methods, ancestral and mutation processes are combined together, whereas, in this method, an ancestral tree is sampled that is

consistent with data and then it is combined with the mutation process. Based on this sampling approach, the likelihood curve of the mutation rate can be estimated for a set of values of the mutation rate in one simulation run; it does not require choosing a particular value as a starting value. The reason for this is that the sampling of the topologies of the ancestral trees and coalescent waiting times do not depend on mutation rate. The same advantage also holds for estimating the likelihood surface of the mutation rate and any parameters in the distribution function of the coalescence waiting times for variable population size model. The coalescence times in this model can be derived by a transformation of the coalescence times in the standard coalescent (see GRIFFITHS and TAVARÉ 1994a and GRIFFITHS and TAVARÉ 1999). So, it is enough to sample coalescence times from the standard coalescent and then to apply the transformation to them for different values of the parameters. Using the same idea, Bayesian inference can be developed for the parameters in this framework.

In this paper, two sampling methods have been developed for topologies of conditional ancestral trees. The first method samples the topologies from the exact distribution, but the second method samples the topologies from a proposal distribution. The first method has some limitations for applicability because the probabilities of the simple mutation trees must be known in decision making steps of this method and no closed-form formula is known for these probabilities for general case. However, This method can be applied to data sets for which all the simple mutation trees embedded in the mutation tree of the data have no more than a 2-mutation constraint. Although this is a restriction, there can be many real data sets that satisfy this condition. For example, the data sets from FULLERTON *et al.* (1994) and GARRIGAN *et al.* (2005) satisfy this condition when the polymorphic sites of the data sets are phased based on the reference sequence from chimpanzee genome. In addition, data sets are generated for a sample size of 20 under the standard coalescent for a wide range of mutation rates and for each data set the corresponding mutation tree is constructed and the maximal number of mutation constraints of simple mutation trees embedded in the mutation tree is computed. Then the distribution of this statistic for each θ is estimated. For each θ, the value of this statistic with highest probability is at 0, 1 or 2. Note that the same thing is

true for sample sizes less than 20 because coalescence waiting times in the standard coalescent are independent.

The software developed for the inference method in this paper was compared with Genetree9.0 on many simulated data sets. The estimates from these programs agree on all data sets (data not shown), and these programs show some run-time advantage over Genetree 9.0. The run-time differences between the methods are shown in Table 1, in which data sets from FULLERTON *et al.* (1994) and GARRIGAN *et al.* (2005) are used. The programs are different based on the sampling methods for the topologies of the conditional ancestral trees. The two sampling methods are described in the previous sections, the third one is explained below.

Modifications and extensions: This paper assumed that the sequences in the sample are labeled and ordered. Considering the labeled and unordered case relaxes this condition. In this case, similar inference methods can be developed by changing decision-making rules and the equations on which they are based. The details of these changes are given in the Supplementary Material.

For purposes of this research, I assume that the ancestral nucleotide at each polymorphic site is known. Note that the ancestral nucleotide at each polymorphic site in the sample of DNA sequences consistent with the infinite-sites model can be phased based on a reference sequence from an outgroup species. If it is not available, the more frequent allele at each polymorphic site can be considered as the ancestral allele because this fact holds in the general coalescent tree framework (see SARGSYAN and WAKELEY 2008). Otherwise, all possible phasing should be considered (for more details see GRIFFITHS and TAVARÉ 1995).

An ancestral model based on the model-free approach has also been considered in population genetics. (More about model-free approach, see, e.g., STUMPF and GOLDSTEIN 2001 and MELIGKOTSIDOU and FEARNHEAD 2005, and references therein.) In this model the ancestral tree of a sample is a rooted-bifurcating tree with random-joining topological structure, but the time intervals between consecutive coalescent events are parameters. This model does not fit to the general coalescent tree framework, where coalescence waiting times are random variables. However, the sampling methods developed in this paper for the topologies of the conditional ancestral trees can

be applied to design a full likelihood-based inference method, using this ancestral model with the infinite-sites model.

I thank John Wakeley and Simon Tavaré for stimulating discussions and helpful comments on the drafts of the manuscript.

Datasets \ Program	Genetree9.0	LNO	LO	ES
1	3800	1356	1356	1150
2	3383	1381	1360	1056

Table 1: Run times comparison of Genetree 9.0 with three programs based on the sampling schemes developed in this paper. The three programs are the following: In the inference method of the first program (LNO) the topologies of conditional ancestral trees are sampled using the proposal distribution when the sequences in the sample are labeled but not ordered. In the method of the second program (LO) the proposal distribution is used based on assumption that the sequences in the sample are labeled and ordered. In the method of the third program (ES) it is assumed that the sequences are labeled and ordered but topologies of the conditional ancestral trees are sampled from exact distribution. The first data is from β-globin gene locus, sampled by FULLERTON *et al.* (1994), the sample size is 57 chromosomes and there are 13 polymorphic sites; the second data is from the human RRM2P4 pseudogene, sampled by GARRIGAN *et al.* (2005), the sample size is 41 and there are 13 polymorphic sites. The number of the iterations is 10^8, time is measured in seconds. The outputs of these programs are the estimates of the likelihood function and the ages of the mutations in the sample. The standard coalescent is assumed as the underlying ancestral model and mutation rate equal to 3.

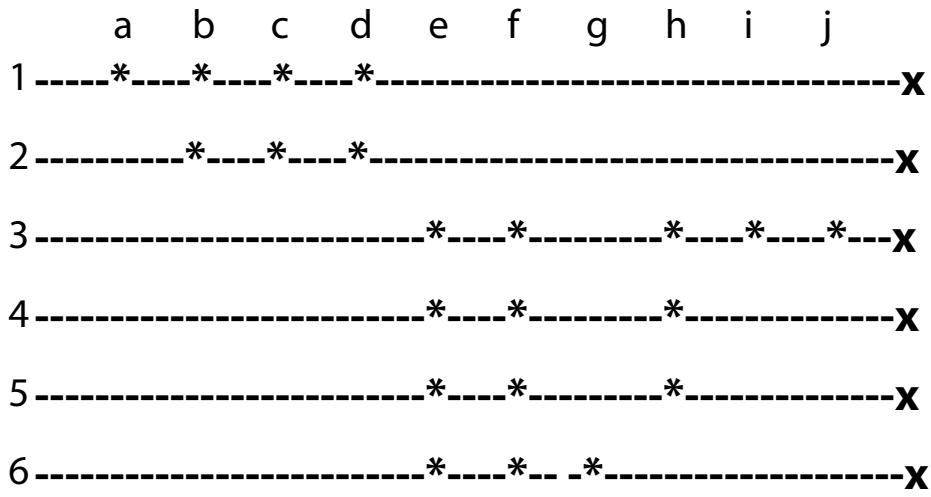

Figure 1: An sample of six aligned DNA sequences, where the mutant allele at each polymorphic site is denoted by '*', and the ancestral allele by '-'. 'x' is used for the hypothetical mutation. The mutant groups in this sample are {1}, {1,2}, {3,4,5,6}, {6}, {3,4,5}, {3}, {1,2,3,4,5,6}.

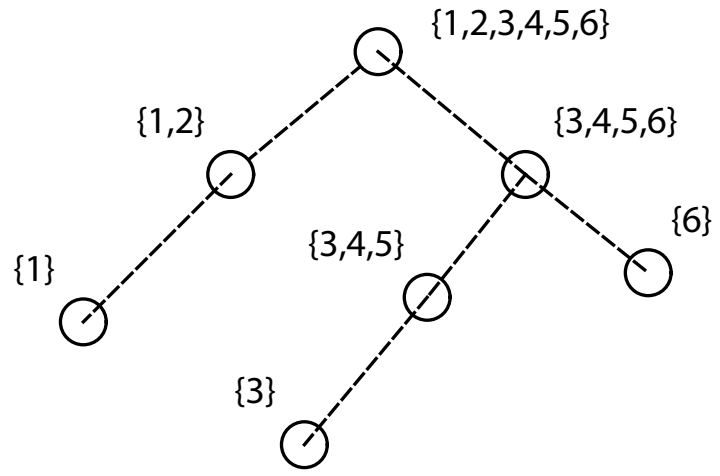

Figure 2: The mutation tree corresponding to the polymorphism data in Figure 1.

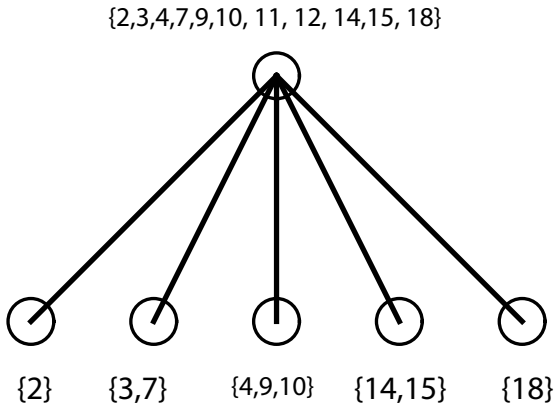

Figure 3: A simple mutation tree in the form (A_1, A_2, A_3, C), where A_1, A_2, and A_3 are the nodes of the simple mutation tree that correspond to the mutant groups {3,7}, {4,9,10}, and {14,15}, respectively; C_0 is {11,12}; C_1 consists the nods corresponding to the mutant groups {2} and {18}.

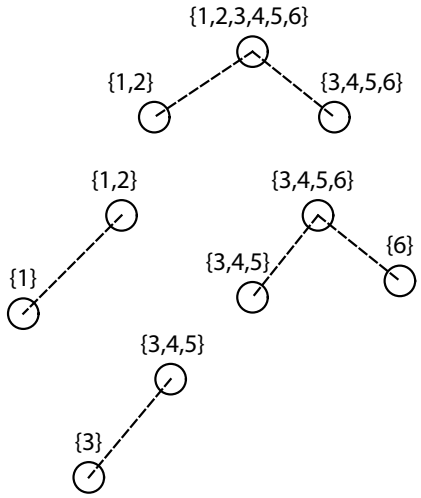

Figure 4: The simple mutation trees embedded in the mutation tree in Figure 2.

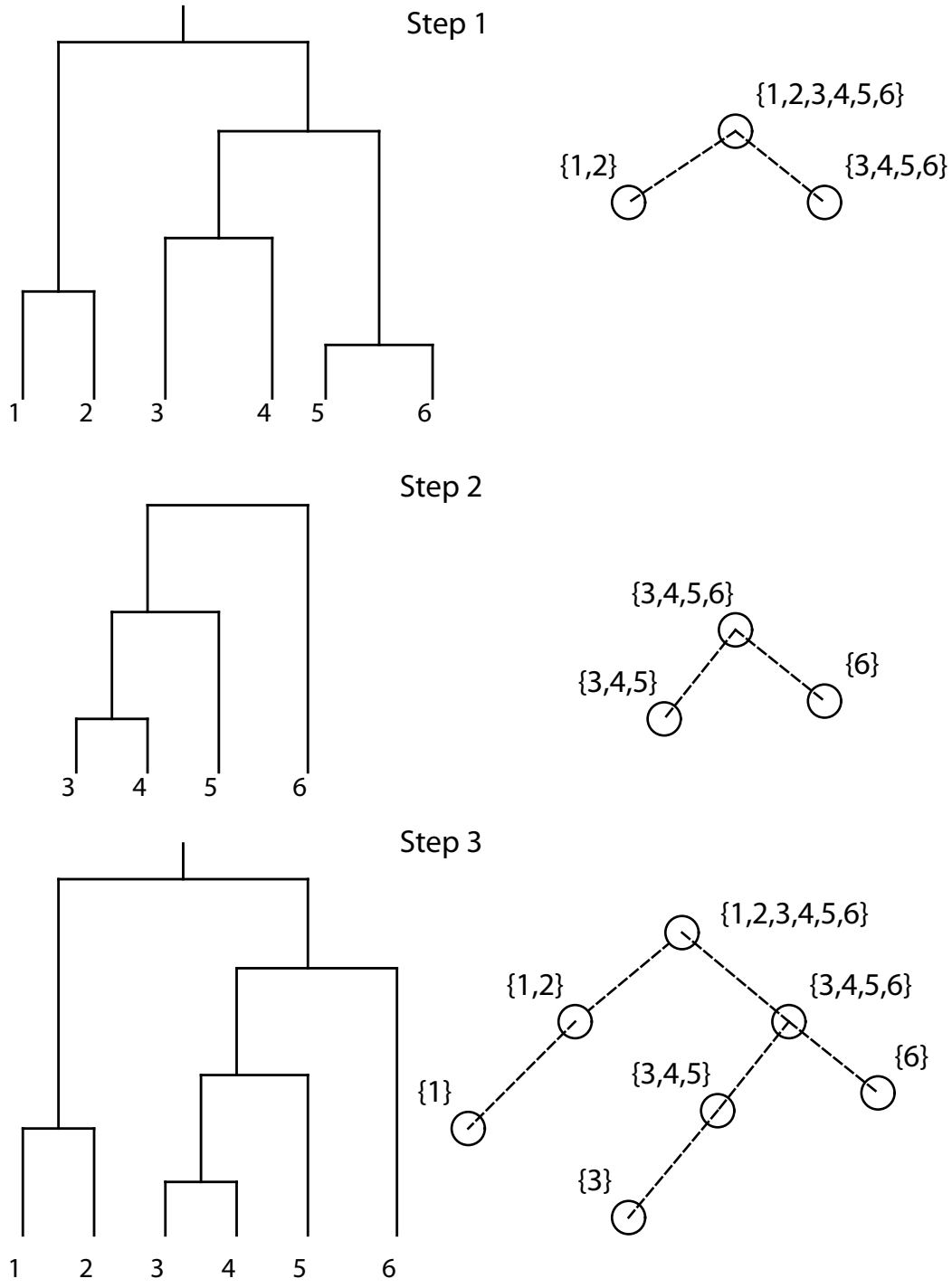

Figure 5: A realization of the sampling procedure for a topological tree consistent with the mutation tree in Figure 2 is presented. In Step 1 a topological tree is sampled consistent with the simple mutation tree shown next to it. In Step 2 a topological tree is sampled that is consistent with the simple mutation tree shown next to it. In Step 3, the topological tree of the ancestry of the sample { 3,4,5,6}, embedded in the topological tree constructed in Step 1, is cut and replaced by the topological tree sampled in Step 2.

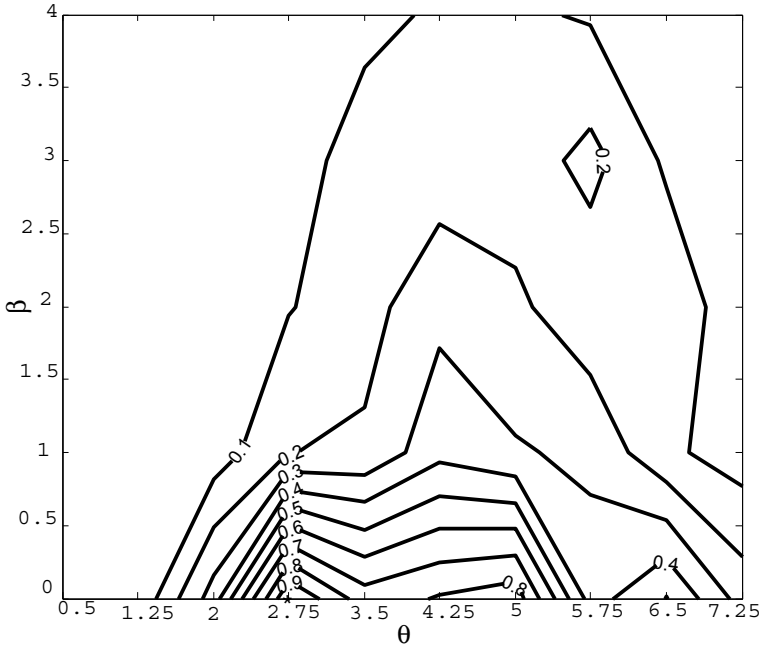

Figure 6: The contour plot of the likelihood function of (θ, β) for the full polymorphism data in the sample of 165 sequences completely linked to the mutant allele at the APOE gene locus. In the contour plot the estimate of the likelihood function is scaled by the maximum of the estimated likelihood function.

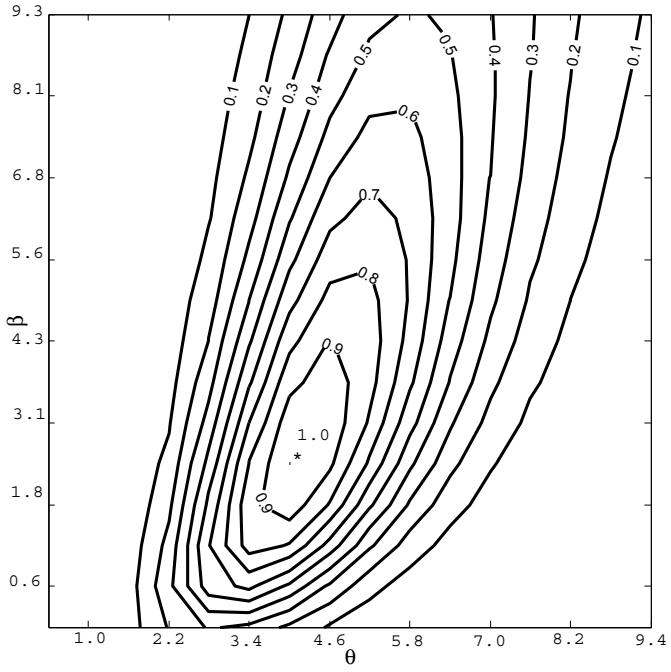

Figure 7: The contour plot of the likelihood function of (θ, β) for the full polymorphism data in the sample of 188 sequences at the APOE gene locus. In the contour plot the likelihood function is scaled by the maximum of the estimated likelihood function.

LITERATURE CITED

BAHLO, M. and R. C. GRIFFITHS, 2000 Inference from gene trees in a subdivided population. Theoretical Population Biology **57**: 79–95.

COOP, G. and R. C. GRIFFITHS, 2004 Ancestral inference on gene trees under selection. Theoretical Population Biology **66**: 219–232.

DE IORIO, M. and R. GRIFFITHS, 2004 Importance sampling on coalescent histories. II: Subdivided population models. Adv. in Appl. Probab. **36**: 434–454.

DE IORIO, M., R. GRIFFITHS, R. LEBLOIS, and F. ROUSSET, 2005 Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theoretical Population Biology **68**: 41–53.

DEVROYE, L., 1986 *Nonuniform random variate generation*. Springer-Verlag, New York.

DONNELLY, P. and S. TAVARÉ, 1995 Coalescents and genealogical structure under neutrality. Annual Rev. Genet. **29**: 401–421.

DRUMMOND, A., G. NICHOLLS, A. RODRIGO, and W. SOLOMON, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**: 1307–20.

EDWARDS, A. W. F., 1970 Estimation of the branch points of a branching diffusion process. Journal of the Royal Statistical Society, Series B **32**: 154–174.

FEARNHEAD, P. and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. Genetics **159**: 1299–318.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet. Res. **60**: 209–220.

FULLERTON, S., A. CLARK, K. WEISS, D. NICKERSON, S. TAYLOR, J. STENGÅRD, V. SALOMAA, E. VARTIAINEN, M. PEROLA, E. BOERWINKLE, and C. SING, 2000 Apolipoprotein E variation at the sequence haplotype level: Implications for the origin and maintenance of a major human polymorphism. American Journal of Human Genetics **67**: 881–900.

FULLERTON, S., R. HARDING, A. BOYCE, and J. CLEGG, 1994 Molecular and population genetic

analysis sequence diversity at the human β-globin locus. Proc. Natl. Acad. Sci. **91**: 1805–09.

GARRIGAN, D., Z. MOBASHER, T. SEVERSON, J. WILDER, and M. HAMMER, 2005 Evidence for archaic Asian ancestry on the human X chromosome. Molecular biology and evolution **22**: 189–92.

GRIFFITHS, R., 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 P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput Biol. **3**: 479–502.

GRIFFITHS, R. C. and S. TAVARÉ, 1994a Ancestral inference in population genetics. Statistical Science **9**: 307–319.

GRIFFITHS, R. C. and S. TAVARÉ, 1994b Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B **344**: 403–410.

GRIFFITHS, R. C. and S. TAVARÉ, 1994c Simulating probability distributions in the coalescent. Theoretical Population Biology **46**: 131–159.

GRIFFITHS, R. C. and S. TAVARÉ, 1995 Unrooted genealogical tree probabilities in the infinity-many-sites model. Mathematical Biosciences **127**: 77–98.

GRIFFITHS, R. C. and S. TAVARÉ, 1998 The age of a mutation in a general coalescent tree. Stochastic Models **14**: 273–295.

GRIFFITHS, R. C. and S. TAVARÉ, 1999 The ages of mutations in gene trees. The Annals of Applied Probability **9**: 567–590.

GRIFFITHS, R. C. and S. TAVARÉ, 2003 The genealogy of a neutral mutation. In *Highly Structured Stochastic Systems*, edited by P. J. Green, N. L. Hjort, and S. Richardson, Oxford Statistical Science, pp. 393–413, Oxford University Press, Oxford.

GUSFIELD, D., 1991 Efficient algorithms for inferring evolutionary trees. Networks **21**: 19–28.

HOBOLTH, A., M. UYENOYAMA, and C. WIUF, 2008 Importance sampling for the infinite sites model. Stat. Appl. Genet. Mol. Biol. **7**: 32.

HUDSON, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data.

Evolution **37**: 203–217.

KIMURA, M. and T. OHTA, 1973 The age of a neutral mutant persisting in a finite population. Genetics **75**: 199–5312.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. Journal of Applied Probability **19A**: 27–43.

KINGMAN, J. F. C., 1982b The coalescent. Stochastic Processes and their Applications **13**: 235–248.

KINGMAN, J. F. C., 1982c Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino, pp. 97–112, North Holland Publishing Company.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140**: 1421–1430.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149**: 429–434.

MELIGKOTSIDOU, L. and P. FEARNHEAD, 2005 Maximum-likelihood estimation of coalescence times in genealogical trees. Genetics **171**: 2073–2084.

NEE, S., R. M. MAY, and P. H. HARVEY, 1994 The reconstructed evolutionary process. Phil. Trans. R. Soc. B **344**: 305–311.

NIELSEN, R., 1997 A likelihood approach to populations samples of microsatellite alleles. Genetics **146**: 711–6.

RANNALA, B., 1997 Gene genealogy in a population of variable size. Heredity **78**: 417–423.

SARGSYAN, O. and J. WAKELEY, 2008 A coalescent process with simulaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical Population Biology **74**: 104–114.

SLATKIN, M. and R. R. HUDSON, 1991 Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129**: 555–562.

SLATKIN, M. and B. RANNALA, 1997 Estimating the age of alleles by use of interaallelic variability. American Journal of Human Genetics **60**: 447–458.

SLATKIN, M. and B. RANNALA, 2000 Estimating allele age. Annu. Rev. Genomics Hum. Genet. **01**: 225–249.

STEPHENS, M., 2000 Times on Trees, and the Age of an Allele. Theoretical Population Biology **57**: 109–119.

STEPHENS, M. and P. DONNELLY, 2000 Inference in molecular population genetics. Journal of the Royal Statistical Society, Series B **62**: 605–635.

STEPHENS, M. and P. DONNELLY, 2003 Ancestral inference in population genetics models with selection (with discussion). Aust. N.Z. J. of Stats. **45**: 395–430.

STUMPF, M. P. H. and D. B. GOLDSTEIN, 2001 Genealogical and evolutionary inference with the human Y chromosome. Science **291**: 1738–1742.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.

THOMPSON, E. A., 1975 *Human evolutionary trees*. Cambridge University Press, Cambridge.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**: 256–276.

WILSON, I. and D. BALDING, 1998 Genealogical Inference From Microsatellite Data. Genetics **150**: 499–510.

WIUF, C. and P. DONNELLY, 1999 Conditional genealogies and the age of a neutral mutant. Theoretical Population Biology **56**: 183–201.