Theoretical Population Biology 102 (2015) 3–15

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Genealogical histories in structured populations

Seiji Kumagai, Marcy K. Uyenoyama *

Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338, USA

ARTICLE INFO

Article history:
Received 27 May 2013
Available online 11 March 2015

Keywords:
Population structure
Migration
Coalescence time
MRCA age
Generating functions

ABSTRACT

In genealogies of genes sampled from structured populations, lineages coalesce at rates dependent on the states of the lineages. For migration and coalescence events occurring on comparable time scales, for example, only lineages residing in the same deme of a geographically subdivided population can have descended from a common ancestor in the immediately preceding generation. Here, we explore aspects of genealogical structure in a population comprising two demes, between which migration may occur. We use generating functions to obtain exact densities and moments of coalescence time, number of mutations, total tree length, and age of the most recent common ancestor of the sample. We describe qualitative features of the distribution of gene genealogies, including factors that influence the geographical location of the most recent common ancestor and departures of the distribution of internode lengths from exponential.

1. Introduction

Under population structure, the rate of coalescence among genetic lineages depends upon the states of the lineages. The state may include, for example, whether lineages reside in the same or distinct individuals in a population undergoing inbreeding or in the same or distinct demes of a subdivided population. Crow and Maruyama (1971) addressed the effective number of alleles (inverse of the probability of identity between a pair of genes randomly sampled from the same deme), discussing its relationship to the time to fixation of a neutral mutation and the level of heterozygosity in a structured population. Much of the substantial body of work on structured populations has focused on analytical solutions for small samples (e.g., Nei and Feldman, 1972; Li, 1976; Griffiths, 1981; Strobeck, 1987; Takahata, 1988; Hudson, 1990; Nath and Griffiths, 1993; Wakeley, 1996; Rosenberg and Feldman, 2002; Innan and Watanabe, 2006; Wilkinson-Herbots, 2008).

In the context of an isolation-with-migration model (IM, Nielsen and Wakeley, 2001), Wang and Hey (2010) provided a detailed description of the nature of the coalescence process. Only lineages that reside in the same deme can coalesce in the immediately ancestral generation, with migration between demes inducing changes in the number residing in a given deme. Wang and Hey (2010) computed the likelihood as a convolution over the numbers of the various kinds of migration events and the time spent in various states (configuration of lineages among demes). Using a continuous time Markov chain (CTMC) framework, Hobolth et al. (2011) gave an analytical solution for the density of time since the most recent coalescence in a small sample in an IM model comprising two extant populations between which migration may have occurred since their divergence from an ancestral population. Zhu and Yang (2012) used this solution to develop a prior distribution of genealogical histories in their Markov Chain Monte Carlo (MCMC) sampler involving three extant populations. Mailund et al. (2011) incorporated the CTMC into a Hidden Markov Model for the estimation of divergence time and effective sizes of populations. Andersen et al. (2013) developed a CTMC framework to accommodate the IM model, deriving densities for coalescence time and mutation numbers.

This literature illustrates two related lines of research: developing a qualitative understanding of the effects of population structure on patterns of genetic variation and developing a computationally feasible approach to sampling-based inference frameworks. Our exploration of aspects of a sample derived from two demes falls at an intermediate point along this continuum. Our objectives include enhancing intuition about genealogies within structured populations as well as developing a model-based computational approach that may contribute to the determination of likelihoods of models and their parameters.

Building on previous methods (especially Takahata, 1988; Hudson, 1990; Uyenoyama and Takebayashi, 2004), we describe generating functions for total tree length and the number of segregating sites in a sample of arbitrary size. We explore the effect

* Corresponding author.
E-mail address: marcy@duke.edu (M.K. Uyenoyama).

http://dx.doi.org/10.1016/j.tpb.2015.01.003
0040-5809/

of population structure on the density of time between successive coalescence events (internode length) within gene genealogies. To illustrate the qualitative behavior of the process, we conduct a sensitivity analysis of the location of the most recent common ancestor (MRCA) of a sample derived from a population comprising two demes.

2. Model

For a sample of n genes, the gene genealogy comprises $n-1$ coalescence events, demarcating $n-1$ levels, with level ℓ corresponding to the segment in which exactly ℓ lineages ancestral to the sample exist. At any point in the gene genealogy, each lineage may exist in various states. Within the framework of an IM model, for example, aspects of a configuration may include the type of each lineage (*e.g.*, deme of origin or location of descendants). We describe the spectrum of the states of all lineages at a level boundary as the configuration. We denote the configuration at the more recent boundary of a level as the entrance state and the configuration at the more ancient boundary the exit state. The exit state of a given level is identical to the entrance state of the next level into the past. A gene genealogy corresponds to a list of configurations, corresponding to the observed sample and every level boundary back to the MRCA, together with a list of ages of the level boundaries.

2.1. Continuous time

For any level of the gene genealogy, the process initiated at a given entrance state may visit various transient (non-coalescent) states en route to an absorbing (coalescent) state. We characterize the within-level process as a continuous-time, finite-state Markov chain, from the entrance state to the exit state. This framework appears to be essentially equivalent to that of Andersen et al. (2013).

Neuts (1995, Chapter 5) provides a lucid exposition of phase-type densities obtained from transition matrices of the form described in (5). Here, we place this general framework in the present context in order to facilitate presentation of our approach through generating functions (Section 2.2).

2.1.1. Transition probability matrix

Let $\boldsymbol{P}(t)$ denote the transition probability matrix, of which the ijth element represents the probability that a process exists in state j at time $t+s$ given its existence in state i at time s, for any s. With respect to entrance state i, we denote as transient any state j for which $\lim_{t\to\infty} \boldsymbol{P}_{ij}(t) = 0$. Termination of the level corresponds to absorption in exit state j ($\boldsymbol{P}_{jj}(t) = 1$ for all t). We restrict attention to processes in which all configurations can be classified as either transient or absorbing.

Under the Markov properties, $\boldsymbol{P}(t)$ satisfies the Chapman–Kolmogorov equations:

$$\boldsymbol{P}(t+s) = \boldsymbol{P}(t)\boldsymbol{P}(s). \tag{1}$$

In particular,

$$\boldsymbol{P}(t+dt) - \boldsymbol{P}(t) = \boldsymbol{P}(t)[\boldsymbol{P}(dt) - \boldsymbol{I}] = [\boldsymbol{P}(dt) - \boldsymbol{I}]\boldsymbol{P}(t), \tag{2}$$

for $\boldsymbol{I}$ the identity matrix and dt a small time increment, with instantaneous rates of change given by

$$\lim_{dt\to 0} \frac{\boldsymbol{P}(dt) - \boldsymbol{I}}{dt} = \boldsymbol{P}'(0) = \boldsymbol{A}. \tag{3}$$

From (2) we obtain

$$\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{P}'(0) = \boldsymbol{P}'(0)\boldsymbol{P}(t),$$

the solution of which gives

$$\boldsymbol{P}(t) = e^{\boldsymbol{A}t} = \boldsymbol{I} + \sum_{k=1}^{\infty} \frac{(\boldsymbol{A}t)^k}{k!} \tag{4}$$

(see, for example, Taylor and Karlin, 1998, Chapter VI, Section 6).

Given the entrance state of a level, the process may visit some number of transient states before absorption in an exit state. For a model of multiple demes, for example, the transient states may reflect different arrangements of the lineages among the demes and an exit state coalescence in one of the demes. For α the number of transient states accessible from the entrance state and β the total number of exit states accessible from any of the transient states, the instantaneous rates of change (3) correspond to

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \tag{5}$$

in which $\boldsymbol{U}$ (dimension $\alpha \times \alpha$) gives the instantaneous rates of within-level moves and $\boldsymbol{V}$ ($\alpha \times \beta$) between-level moves, with the lower blocks representing matrices of zeros of appropriate size ($\beta \times \alpha$ and $\beta \times \beta$).

Let λ_i represent the instantaneous rate of change from transient state i, the sum of the off-diagonal elements of row i. As the rows of $\boldsymbol{A}$ must sum to zero, the diagonal elements of $\boldsymbol{U}$ correspond to $-\boldsymbol{L}$, for $\boldsymbol{L}$ a diagonal matrix of the λ_i:

$$\boldsymbol{L} = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_\alpha). \tag{6}$$

From (4) and (5), we obtain the transition probability matrix

$$\boldsymbol{P}(t) = \begin{pmatrix} e^{\boldsymbol{U}t} & (e^{\boldsymbol{U}t} - \boldsymbol{I})\boldsymbol{U}^{-1}\boldsymbol{V} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}. \tag{7}$$

2.1.2. Hitting probabilities

That $\boldsymbol{U}$ represents rates of transition among transient states implies

$$\lim_{t\to\infty} e^{\boldsymbol{U}t} = \boldsymbol{0}.$$

From (7), this implies that the probability of absorption in exit state j from transient state i corresponds to the ijth element of

$$\boldsymbol{Q} = -\boldsymbol{U}^{-1}\boldsymbol{V}. \tag{8}$$

2.1.3. Density of internode length

In (7), the elements of $(e^{\boldsymbol{U}t} - \boldsymbol{I})\boldsymbol{U}^{-1}\boldsymbol{V}$ represent cumulative distribution functions, the ijth element of which gives the probability that the process reaches absorbing state j from transient state i by time t. Taking derivatives, we obtain the density of absorption time:

$$e^{\boldsymbol{U}t}\boldsymbol{V}. \tag{9}$$

A number of authors (*e.g.*, Takahata, 1988; Nath and Griffiths, 1993) have analyzed Laplace transforms of coalescence times, equivalent to moment generating functions (mgfs) of these non-negative variables. The mgf of duration of a level of the sample genealogy corresponds to

$$\boldsymbol{h}(b) = -[\boldsymbol{I}a + \boldsymbol{U}]^{-1}\boldsymbol{V}. \tag{10}$$

In general, the density may be recovered from the mgf using the inversion function

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ibt}\phi(b)db \tag{11}$$

in which $i = \sqrt{-1}$ and $\phi(b)$ represents the corresponding characteristic function,

$$\phi(b) = h(ib).$$

Whether numerical integration of (11) or computation of (9) poses greater challenges may depend on sample size and the process modeled.

2.2. Generating functions

For a pair of genes sampled from a structured population, Hudson (1990) elucidated the relationship between the coefficient of identity, the moment generating function (mgf) of the age of the most recent common ancestor of a pair of genes, and the probability generating function (pgf) of the number of segregating sites. This key insight is easily extended to any number of genes (Appendix A). The probability of a monomorphic sample (P), the pgf of the number of segregating sites (g_Z), and the mgf of tree length (h_T) are related by

$$g_Z(a) = h_T(\theta(a-1)) = P(\theta(1-a)), \tag{12}$$

for θ the scaled rate of mutation under the infinite sites model (16). We incorporate this relationship into the approach of Uyenoyama and Takebayashi (2004) to obtain recursions in generating functions of mutation numbers and coalescence times.

As many previous analyses have addressed the relationship between pairs of genes, Appendix C compares earlier results to ours for samples of size $n = 2$.

2.2.1. Rates of transition

To illustrate the method, we address a generalization of the model explored by Takahata (1988). Lineages reside either in deme 0 or deme 1, which respectively comprise effective numbers of genes N_0 and N_1. Migration corresponds to descent of one of the lineages in deme j from a gene in the other deme in the preceding generation at rate m_j. This quantity is often called the backward migration rate: m_j represents the proportion of genes in deme j that have an immediate ancestor in a different deme.

For level ℓ of the gene genealogy, the entrance state corresponds to the number of lineages residing in deme 0 (i), with the remaining lineages ($\ell - i$) residing in deme 1. The exit state specifies the location of the coalescence event as well as the number of lineages residing in each deme. In general, the entrance state determines the accessible transient states and the possible exit states.

We assume that migration and coalescence represent independent processes, each with an exponentially-distributed waiting time. Coalescence of a pair of lineages in deme j ($j = 0, 1$) occurs at a rate proportional to $1/N_j$. From a state with i lineages residing in deme 0, for example, the probability that the most recent event corresponds to a migration of one of the lineages in deme 0 is

$$\frac{im_0}{im_0 + (\ell - i)m_1 + \binom{i}{2}/N_0 + \binom{\ell-i}{2}/N_1}. \tag{13}$$

The waiting time in generations since the most recent event (migration or coalescence) has an exponential distribution with parameter proportional to the denominator of (13). In units of $2N$ generations, the total rate of departure from state i (6) corresponds to

$$\lambda_i = iM_0 + (\ell - i)M_1 + 2\binom{i}{2}c_0 + 2\binom{\ell - i}{2}c_1, \tag{14}$$

with unmeaningful binomial coefficients (*e.g.*, $\binom{1}{2}$) defined as zero and

$$\begin{aligned} M_0 &= \lim_{\substack{m_0 \to 0 \\ N \to \infty}} 2Nm_0 \\ M_1 &= \lim_{\substack{m_1 \to 0 \\ N \to \infty}} 2Nm_1, \end{aligned} \tag{15a}$$

and the c_j the relative rates of coalescence:

$$\begin{aligned} c_0 &= \lim_{\substack{N_0 \to \infty \\ N \to \infty}} N/N_0 \\ c_1 &= \lim_{\substack{N_1 \to \infty \\ N \to \infty}} N/N_1, \end{aligned} \tag{15b}$$

for N an arbitrary constant. For cases involving mutation, the scaled mutation rate corresponds to

$$\theta = \lim_{\substack{u \to 0 \\ N \to \infty}} 2Nu, \tag{16}$$

for u the per-generation rate of mutation under the infinite sites model.

2.2.2. Number of segregating sites

Uyenoyama and Takebayashi (2004) derived a recursion in probability generating functions of the numbers of segregating sites in a sample drawn from a structured population (see also Lohse et al., 2011). Their recursion generalizes the approach of (Watterson, 1975), who noted that the number of segregating sites accumulated on each level of a gene genealogy has a geometric distribution that is independent of those for other levels of the tree. Here, we begin with a heuristic description of this generalization.

We seek to determine the probability generating function (pgf) of the number of segregating sites in a sample of a given size corresponding to each configuration (partition of lineages across demes). Let $\boldsymbol{g}_\ell(a)$ represent the vector of these pgfs for a sample of size ℓ and $\tilde{\boldsymbol{U}}_\ell$ and $\tilde{\boldsymbol{V}}_\ell$ matrices of probabilities (rather than rates) of transitions:

$$\begin{aligned} \tilde{\boldsymbol{U}}_\ell &= \boldsymbol{I} + \boldsymbol{L}^{-1}\boldsymbol{U}_\ell \\ \tilde{\boldsymbol{V}}_\ell &= \boldsymbol{L}^{-1}\boldsymbol{V}_\ell, \end{aligned} \tag{17}$$

in which $\boldsymbol{L}$ is given in (6), with $\boldsymbol{U}_\ell$ and $\boldsymbol{V}_\ell$ identical to $\boldsymbol{U}$ and $\boldsymbol{V}$ in Section 2, but with subscripts explicitly indicating level. Application of $\boldsymbol{L}^{-1}$ converts the off-diagonal elements of rate matrix $\boldsymbol{A}$ (5) to probabilities of the most recent event back in time, with the zeroed diagonal elements of $\tilde{\boldsymbol{U}}_\ell$ reflecting that any step entails a change in state of the process. For a given configuration, the most recent migration or coalescence event reflects either a within-level transition (with probabilities given by $\boldsymbol{U}_\ell$) or a between-level transition ($\boldsymbol{V}_\ell$). In either case, the total number of mutations corresponds to the number accumulated more recently than the transition plus the number for a sample at the configuration implied by the transition:

$$\begin{aligned} \boldsymbol{g}_\ell(a) &= \tilde{\boldsymbol{F}}_\ell(a)[\tilde{\boldsymbol{U}}_\ell \boldsymbol{g}_\ell(a) + \tilde{\boldsymbol{V}}_\ell \boldsymbol{g}_{\ell-1}(a)] \\ &= [\boldsymbol{I} - \tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{V}}_\ell \boldsymbol{g}_{\ell-1}(a), \end{aligned} \tag{18}$$

in which $\tilde{\boldsymbol{F}}_\ell(a)$ represents a diagonal matrix of pgfs of the numbers of mutations arising since the most recent state transition and $\boldsymbol{I}$ the identity matrix.

For the model described in Section 2.2.1, matrix $\tilde{\boldsymbol{U}}_\ell$ has dimensions $(\ell+1) \times (\ell+1)$, with the element in row i and column j ($i, j = 0, 1, \ldots, \ell$) given by

$$\tilde{\boldsymbol{U}}_{\ell,ij} = \begin{cases} \dfrac{(\ell - i)M_1}{\lambda_i} & \text{for } j = i+1 \\ \dfrac{iM_0}{\lambda_i} & \text{for } j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

Elements of $\tilde{\boldsymbol{V}}_\ell$, of dimensions $(\ell + 1) \times \ell$, correspond to

$$\tilde{\boldsymbol{V}}_{\ell,ij} = \begin{cases} \dfrac{2\binom{\ell-i}{2}c_1}{\lambda_i} & \text{for } j = i \\ \dfrac{2\binom{i}{2}c_0}{\lambda_i} & \text{for } j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

The probability that a process presently in state i exits the level through state j corresponds to the element in row i and column j of

$$\tilde{\boldsymbol{V}} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{V}} + \tilde{\boldsymbol{U}}^2\tilde{\boldsymbol{V}} + \cdots = [\boldsymbol{I} - \tilde{\boldsymbol{U}}]^{-1}\tilde{\boldsymbol{V}} \tag{19}$$

(with subscripts suppressed for clarity). Substitution of (17) shows that this matrix of hitting probabilities corresponds to $\boldsymbol{Q}$ (8).

Vector $\boldsymbol{g}_\ell(a)$ (18) provides pgfs of the number of segregating sites, with the element in row i corresponding to the state in which i of the ℓ lineages derive from deme 0. Matrix $\tilde{\boldsymbol{F}}_\ell(a)$ corresponds to

$$\begin{aligned}\tilde{\boldsymbol{F}}_\ell(a) &= [\boldsymbol{L} + \ell\theta(1-a)\boldsymbol{I}]^{-1}\boldsymbol{L} \\ &= \text{Diagonal}\left[\frac{\lambda_i}{\lambda_i + \ell\theta(1-a)}\right]_{i=0,1,\ldots,\ell}\end{aligned} \tag{20}$$

for θ (16) the scaled mutation rate under the infinite sites model.

From (18), the vector of pgfs of the total number of segregating sites in the sample of size n corresponds to

$$\boldsymbol{g}_{\boldsymbol{n}}(a) = \prod_{\ell=2}^{n}[\boldsymbol{I} - \tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{V}}_\ell\boldsymbol{g}_{\boldsymbol{1}}(a), \tag{21}$$

in which the matrix product starts on the left with $\ell = n$ and ends on the right with $\ell = 2$. In particular,

$$[\boldsymbol{I} - \tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{F}}_\ell(a)\tilde{\boldsymbol{V}}_\ell$$

represents the pgfs of the number of mutations accumulated within level ℓ. On level 2, corresponding to a pair of lineages, we have

$$\tilde{\boldsymbol{U}}_2 = \begin{pmatrix} 0 & \frac{2M_1}{\lambda_0} & 0 \\ \frac{M_0}{\lambda_1} & 0 & \frac{M_1}{\lambda_1} \\ 0 & \frac{2M_0}{\lambda_2} & 0 \end{pmatrix} \tag{22a}$$

$$\tilde{\boldsymbol{V}}_2 = \begin{pmatrix} \frac{2c_1}{\lambda_0} & 0 \\ 0 & 0 \\ 0 & \frac{2c_0}{\lambda_2} \end{pmatrix} \tag{22b}$$

$$\tilde{\boldsymbol{F}}_2(a) = \begin{pmatrix} \frac{\lambda_0}{\lambda_0 + 2\theta(1-a)} & 0 & 0 \\ 0 & \frac{\lambda_1}{\lambda_1 + 2\theta(1-a)} & 0 \\ 0 & 0 & \frac{\lambda_2}{\lambda_2 + 2\theta(1-a)} \end{pmatrix}, \tag{22c}$$

for the λ_i given by (14) with $\ell = 2$, and with

$$\boldsymbol{g}_1(a) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{23}$$

indicating that a sample comprising a single gene is monomorphic with probability 1.

2.2.3. Total tree length

We now address total tree length, the sum of all branches in the gene genealogy.

Using (12), moment generating functions for total tree length are easily obtained from the pgfs for mutation number (18). Total tree length of a sample of size ℓ corresponds to the contribution of level ℓ to total tree length (ℓ times length of level ℓ) plus the total tree length of a sample of size $\ell - 1$. For element i of vector $\boldsymbol{h}_\ell(\boldsymbol{b})$ denoting the mgf of total tree length of a sample in configuration i within level ℓ, we have

$$\boldsymbol{h}_\ell(b) = [\boldsymbol{I} - \tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{V}}_\ell\boldsymbol{h}_{\boldsymbol{\ell-1}}(b) \tag{24}$$

(compare (18)), in which

$$\begin{aligned}\tilde{\boldsymbol{H}}_\ell(b) &= \tilde{\boldsymbol{F}}_\ell(1 + b/\theta) = [\boldsymbol{L} - lb\boldsymbol{I}]^{-1}\boldsymbol{L} \\ &= \text{Diagonal}\left[\frac{\lambda_i}{\lambda_i - \ell b}\right]_{i=0,1,\ldots,\ell}.\end{aligned} \tag{25}$$

Level ℓ of the sample genealogy, comprising ℓ lineages, contributes ℓ times the length of level ℓ to total tree length. This contribution has mgf

$$[\boldsymbol{I} - \tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{V}}_\ell. \tag{26}$$

Use of (17) confirms the correspondence of (26) to (10), the mgf of the duration of level ℓ. We obtain from (24) mgfs for samples of size n in terms of samples of size 1:

$$\boldsymbol{h}_{\boldsymbol{n}}(b) = \prod_{\ell=2}^{n}[\boldsymbol{I} - \tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{U}}_\ell]^{-1}\tilde{\boldsymbol{H}}_\ell(b)\tilde{\boldsymbol{V}}_\ell\boldsymbol{h}_{\boldsymbol{1}}(b), \tag{27}$$

in which $\boldsymbol{h}_{\boldsymbol{1}}(b) = \boldsymbol{g}_{\boldsymbol{1}}(a)$ (23).

From the recursions in generating functions of number of segregating sites (18) or total tree length (24), we obtain recursions in the moments of these distributions. To facilitate taking derivatives, we rewrite (24) as

$$\boldsymbol{h}_\ell(b) = [\boldsymbol{I} - \boldsymbol{H}_\ell(b)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(b)\boldsymbol{V}_\ell\boldsymbol{h}_{\boldsymbol{\ell-1}}(b), \tag{28}$$

in which

$$\boldsymbol{H}_\ell(b) = \tilde{\boldsymbol{H}}_\ell(b)\boldsymbol{L}^{-1} = \text{Diagonal}\left[\frac{1}{\lambda_i - \ell b}\right]_{i=0,1,\ldots,\ell}. \tag{29}$$

Appendix B shows that the kth derivative with respect to b of the mgf $\boldsymbol{h}_\ell(b)$ corresponds to

$$\begin{aligned}\boldsymbol{h}_\ell^{(k)}(b) &= \sum_{j=0}^{k}\binom{k}{j}j!\,\ell^j\{[\boldsymbol{I} - \boldsymbol{H}_\ell(b)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(b)\}^{j+1} \\ &\quad \times \boldsymbol{V}_\ell\boldsymbol{h}_{\boldsymbol{\ell-1}}^{(k-j)}(b).\end{aligned} \tag{30}$$

Accordingly, the expected total tree length for a sample of size ℓ corresponds to

$$\begin{aligned}\boldsymbol{h}_\ell^{(1)}(0) &= [\boldsymbol{I} - \boldsymbol{H}_\ell(0)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(0)\boldsymbol{V}_\ell\boldsymbol{h}_{\boldsymbol{\ell-1}}^{(1)}(0) \\ &\quad + \ell\{[\boldsymbol{I} - \boldsymbol{H}_\ell(0)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(0)\}^2\boldsymbol{V}_\ell\boldsymbol{1},\end{aligned} \tag{31}$$

noting that for any level j,

$$\boldsymbol{h}_{\boldsymbol{j}}(0) = \boldsymbol{1},$$

a vector with all elements equal to unity. Beginning with initial condition

$$\boldsymbol{h}_{\boldsymbol{1}}^{(1)}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which contributes to the recursion for level $\ell = 2$, the expected total tree length can be determined recursively from (31) for samples of arbitrary size by working through successively larger values of ℓ.

Similarly, the recursion in the non-central second moment corresponds to

$$\begin{aligned}\boldsymbol{h}_\ell^{(2)}(0) &= \sum_{j=0}^{2}\binom{2}{j}j!\,\ell^j\{[\boldsymbol{I} - \boldsymbol{H}_\ell(0)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(0)\}^{j+1} \\ &\quad \times \boldsymbol{V}_\ell\boldsymbol{h}_{\boldsymbol{\ell-1}}^{(2-j)}(0).\end{aligned} \tag{32}$$

The variance of total tree length may be obtained from

$$\boldsymbol{h}_\ell^{(2)}(0) - \boldsymbol{h}_1^{(1)}(0) \odot \boldsymbol{h}_1^{(1)}(0),$$

in which $\odot$ denotes the element-wise product.

Table 1
Moments of total tree length for a sample of size 10 under the expansion model.

		Mean		Variance		Skewness		Kurtosis	
		Exp[a]	Sim[b]	Exp	Sim	Exp	Sim	Exp	Sim
	0	2.120	2.128	2.247	2.330	1.993	2.013	8.784	8.793
	1	3.073	3.075	2.979	2.879	1.460	1.393	6.338	6.119
	2	3.705	3.715	3.305	3.314	1.286	1.258	5.712	5.510
	3	4.170	4.149	3.494	3.493	1.199	1.196	5.427	5.221
	4	4.530	4.504	3.619	3.542	1.146	1.167	5.265	5.457
Number from Deme 0[c]	5	4.815	4.836	3.706	3.702	1.111	1.091	5.161	5.248
	6	5.041	5.068	3.768	3.891	1.087	1.091	5.091	4.957
	7	5.216	5.209	3.812	3.779	1.071	1.124	5.043	5.760
	8	5.341	5.342	3.840	3.870	1.060	0.974	5.013	4.440
	9	5.407	5.395	3.853	3.809	1.054	1.034	5.000	4.773
	10	5.372	5.392	3.863	3.918	1.048	1.031	4.989	4.884

[a] Analytical values obtained from (30).
[b] Average of 10,000 independent simulations.
[c] Number of genes in a sample of 10 derived from deme 0.

2.2.4. Age of MRCA

As the age of the most recent common ancestor of a sample and total tree length are determined by internode lengths, results for age of the MRCA follow immediately from results for total tree length.

From (28), we obtain mgfs for the age of the MRCA:

$$\boldsymbol{h}_{\ell}^{*}(b) = [\boldsymbol{I} - \boldsymbol{H}_{\ell}^{*}(b)(\boldsymbol{L} + \boldsymbol{U}_{\ell})]^{-1}\boldsymbol{H}_{\ell}^{*}(b)\boldsymbol{V}_{\ell}\boldsymbol{h}_{\ell-\mathbf{1}}^{*}(b), \tag{33}$$

in which

$$\boldsymbol{H}_{\ell}^{*}(b) = \text{Diagonal}\left[\frac{1}{\lambda_i - b}\right]_{i=0,1,\ldots,\ell}. \tag{34}$$

Similarly, the kth derivative with respect to b of the mgf $\boldsymbol{h}_{\ell}^{*}(b)$ corresponds to

$$\boldsymbol{h}_{\ell}^{*(k)}(b) = \sum_{j=0}^{k}\binom{k}{j} j!\{[\boldsymbol{I} - \boldsymbol{H}_{\ell}^{*}(b)(\boldsymbol{L} + \boldsymbol{U}_{\ell})]^{-1}\boldsymbol{H}_{\ell}^{*}(b)\}^{j+1} \times \boldsymbol{V}_{\ell}\boldsymbol{h}_{\ell-\mathbf{1}}^{*(k-j)}(b), \tag{35}$$

with initial condition

$$\boldsymbol{h}_{\mathbf{1}}^{*(1)}(0) = \begin{pmatrix}0\\0\end{pmatrix}.$$

We use multidimensional generating functions to address the joint distribution of total tree length (X) and MRCA age (Y). For state i on level ℓ, the joint mgf of increments accumulated up to the most recent migration or coalescence event corresponds to

$$E[e^{Xc}e^{Yd}] = \left[\frac{1}{\lambda_i - \ell c - d}\right].$$

Replacement in (33) of $\boldsymbol{h}_{\ell}^{*}(b)$ by $\boldsymbol{h}_{\ell}^{*}(c, d)$ and $\boldsymbol{H}_{\ell}^{*}(b)$ by

$$\boldsymbol{H}_{\ell}^{*}(c, d) = \text{Diagonal}\left[\frac{1}{\lambda_i - \ell c - d}\right]_{i=0,1,\ldots,\ell}$$

produces recursions in joint mgfs.

3. Applications

3.1. Density and moments of total tree length

Here, we describe the results of a numerical simulation study to verify our analysis and explore the qualitative effects of population subdivision.

For illustrative purposes, we consider two scenarios, in both of which deme 0 has a higher fraction of residents derived from migrants ($M_0 > M_1$). Under the expansion scenario, deme 0 has undergone an expansion in population size, with a lower rate of coalescence than deme 1 ($c_0 < c_1$), and under the colonization scenario, deme 0 has a higher rate of coalescence ($c_0 > c_1$).

We first address the expansion scenario ($c_0 < c_1$), under the assignments $M_0 = c_1 = 2.0$ and $M_1 = c_0 = 0.5$. Table 1 presents the analytical values (30) of the first 4 moments of the distribution of total tree length for $n = 10$ genes for each sampling arrangement (0, 1, . . . , 10 genes derived from deme 0). Our analytical values (Exp) show excellent agreement with the results of 10,000 independent simulations for each sample (Sim). These results indicate an increase in the mean and variance of tree length with the proportion of the sample derived from deme 0 (lower coalescence rate). Skewness values for all samples suggest a long right tail, but decline with increases in the mean. Similarly, all distributions show high kurtosis (peakedness), but decline toward the Gaussian level (3) as the mean and variance increase.

Table 2 presents a similar comparison for the colonization scenario ($c_0 > c_1$), under the assignments $M_0 = c_0 = 2.0$ and $M_1 = c_1 = 0.5$. Under the colonization scenario, in which the deme with the higher backward migration rate also has higher rates of coalescence, the variance in tree length is higher than under the expansion scenario (lower rates of coalescence in deme 0). This trend may reflect that lineages sampled from deme 0 may either undergo rapid coalesce in deme 0 or migrate to deme 1, in which coalescence occurs at lower rates.

To explore this trend in more detail, we inverted the mgfs using (11) to obtain the full density of total tree length. Fig. 1 compares the analytical density of total branch length (solid line) to the simulation results obtained using `ms` (Hudson, 1990) under the expansion scenario ($M_0 = 2.0, M_1 = 0.5, c_0 = 0.5, c_1 = 2.0$). In accordance with trends noted in Table 1, derivation of more genes from deme 0 (lower coalescence rate, higher backward migration rate) increases both the mean and variance of total tree length. Fig. 2 makes a similar comparison for the colonization scenario ($M_0 = 2.0, M_1 = 0.5, c_0 = 2.0, c_1 = 0.5$). A bimodal density emerges for samples derived entirely from deme 0, characterized by higher rates of both coalescence and backward migration. The first mode appears to reflect rapid coalescence of the sample in deme 0 and the second to migration to deme 1 and coalescence at a lower rate. This behavior appears to be consistent with the findings of Rosenberg and Feldman (2002), who noted that the time to coalescence of a pair of genes ($n = 2$) can be bimodal.

In the absence of population structure, the age of the MRCA (or time to any node boundary) corresponds to the sum of exponentially distributed internode lengths. Griffiths (1984) showed that the age of MRCA is asymptotically Gaussian for large sample size ($n \to \infty$) in panmictic populations of constant size. Fig. 3 confirms a bimodal distribution for the age of the MRCA for a sample of size $n = 10$, with a shoulder persisting in the density for a sample of size $n = 20$.

Table 2
Moments of total tree length for a sample of size 10 under the colonization model.

		Mean		Variance		Skewness		Kurtosis	
		Exp[a]	Sim[b]	Exp	Sim	Exp	Sim	Exp	Sim
Number from Deme 0[c]	0	7.402	7.409	12.564	12.852	1.279	1.305	5.819	5.818
	1	7.918	7.938	12.519	12.651	1.279	1.250	5.837	5.524
	2	8.055	8.068	12.510	12.935	1.279	1.282	5.841	5.697
	3	8.052	8.041	12.486	12.399	1.282	1.338	5.852	6.215
	4	7.962	7.962	12.446	12.294	1.288	1.305	5.870	6.085
	5	7.800	7.795	12.388	12.353	1.296	1.398	5.897	7.012
	6	7.567	7.615	12.309	12.605	1.307	1.290	5.934	5.627
	7	7.251	7.267	12.203	12.530	1.322	1.410	5.985	6.563
	8	6.820	6.859	12.067	12.482	1.343	1.489	6.053	7.280
	9	6.203	6.159	11.924	11.662	1.366	1.323	6.126	5.797
	10	5.136	5.164	12.208	12.775	1.345	1.379	5.973	6.026

[a] Analytical values obtained from (30).
[b] Average of 10,000 independent simulations.
[c] Number of genes in a sample of 10 derived from deme 0.

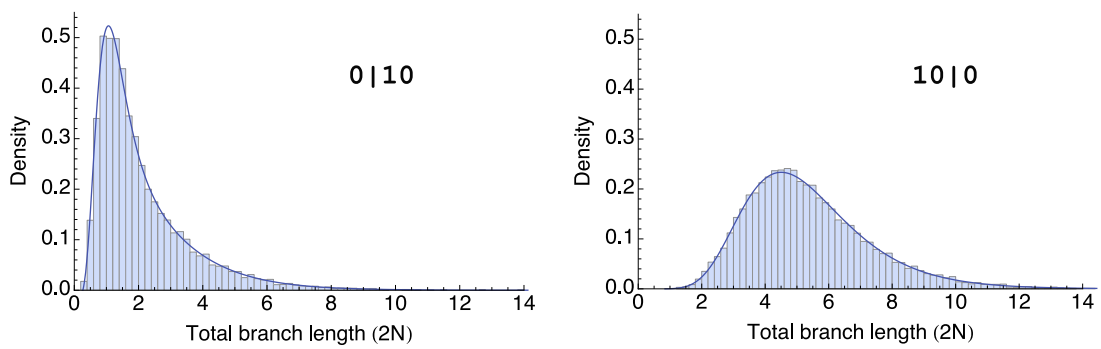

Fig. 1. Distribution of total branch length of a sample of 10 genes, derived entirely from deme 1 (left panel) and entirely from deme 0 (right panel). Each histogram, representing the results of 10,000 independent simulations using ms (Hudson, 1990) under the assignments $M_0 = 2.0$, $M_1 = 0.5$, $c_0 = 0.5$, $c_1 = 2.0$, compares well with the analytical density of total branch length (solid line).

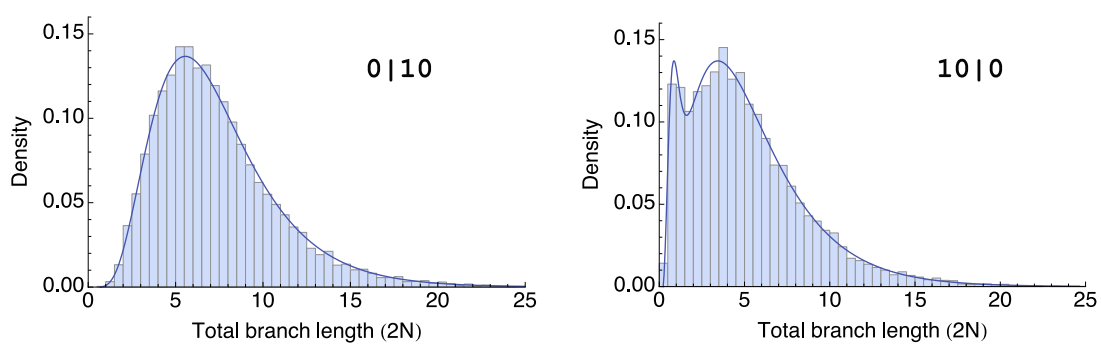

Fig. 2. Distribution of total branch length of a sample of 10 genes, derived entirely from deme 1 (left panel; $M_0 = 2.0$, $c_0 = 2.0$) and entirely from deme 0 (right panel; $M_0 = 2.0$, $c_0 = 2.0$). Each histogram, representing the results of 10,000 independent simulations using ms (Hudson, 1990) under the assignments $M_0 = 2.0$, $M_1 = 0.5$, $c_0 = 2.0$, $c_1 = 0.5$, compares well with the analytical density of total branch length (solid line).

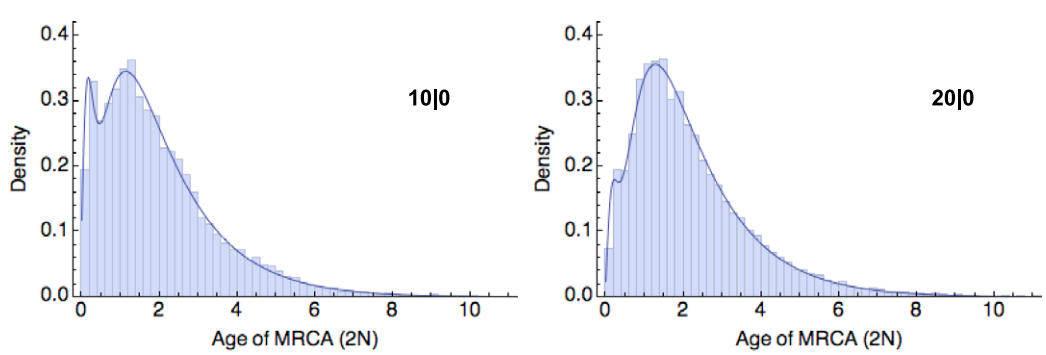

Fig. 3. Distributions of MRCA age of samples derived entirely from deme 0 under the assignments $M_0 = 2.0$, $M_1 = 0.5$, $c_0 = 2.0$, $c_1 = 0.5$. The left panel corresponds to a sample size of $n = 10$, and the right to $n = 20$. Analytical densities are indicated by solid lines and the results of 10,000 independent simulations using ms (Hudson, 1990) by the histograms.

3.2. Internode length across levels of a genealogical tree

While the sum of exponentially distributed internode lengths in panmictic populations is asymptotically Gaussian (Griffiths, 1984), results presented in the preceding section suggest that the distribution of tree length and MRCA age can depart substantially from Gaussian for moderate sample sizes. Here, we further explore the nature and possible causes of departures of the distribution of internode lengths from exponential.

3.2.1. Transition rates

We sharpen the resolution of events by tracking the deme of residence of lineages at more recent boundary of a given level (entrance state) and the demes of origin of the pair of coalescing lineages at the more ancient boundary (exit state). This expansion of the state space of the process facilitates the determination of the minimum number of migration events separating any pair of entrance and exit states.

We refer to lineages residing in deme 0 at the entrance state and their ancestors as blue and the others as yellow. At any point within level ℓ, the state of the process corresponds to $\{b, y\}$, indicating the numbers of blue and yellow lineages currently residing in deme 0. From this state, the probability that the most recent event corresponds to a migration of one of the lineages in deme 0 is

$$\frac{(b+y)m_0}{(b+y)m_0 + (\ell-b-y)m_1 + \binom{b+y}{2}/N_0 + \binom{\ell-b-y}{2}/N_1}. \tag{36}$$

In units of $2N$ generations, the total rate of departure from this state corresponds to

$$(b+y)M_0 + (\ell-b-y)M_1 + \binom{b+y}{2}c_0 + \binom{\ell-b-y}{2}c_1.$$

Upon coalescence, the exit state corresponds to $\{\tilde{b}, \tilde{y}, d, e\}$, for $\tilde{b}$ and $\tilde{y}$ respectively denoting the number in deme 0 of blue and yellow lineages (none involved in the coalescence event), d an indicator of the deme (0 or 1) in which the coalescence event occurred, and e an indicator of the lineages involved in the event (both blue: 0, one of each color: 1, or both yellow: 2).

3.2.2. Departures from exponentially- or gamma-distributed internode lengths

The intervention of multiple migration events between the entrance and exit states can cause the density of internode length to deviate strongly from exponential. Our index of deviation corresponds to the relative squared difference between $F(\cdot)$, the actual cumulative distribution function (cdf), and the cdf of approximating distribution $G(\cdot)$:

$$h = \int_{t_1}^{t_2} \frac{[F(t) - G(t)]^2}{F(t)} dt, \tag{37}$$

for t_1 and t_2 limits of the central 99.8% interval of the actual distribution ($F(t_1) = 0.001$, $F(t_2) = 0.999$). We restricted consideration to $[t_1, t_2]$ to avoid technical problems with the numerical integration.

For example, Fig. 4 shows the actual density (black) of waiting time between the entrance state with 5 lineages in each deme and an exit state that requires a minimum of 3 migration events. Also shown are an exponential distribution (blue) with mean matched to the actual mean and a gamma distribution (red) with both mean and variance matched to the actual distribution. Both distributions deviate strongly from the actual density, the exponential more so than the gamma.

Paths that comprise many migration events tend to be less probable. For the same model and entrance state studied in Fig. 4,

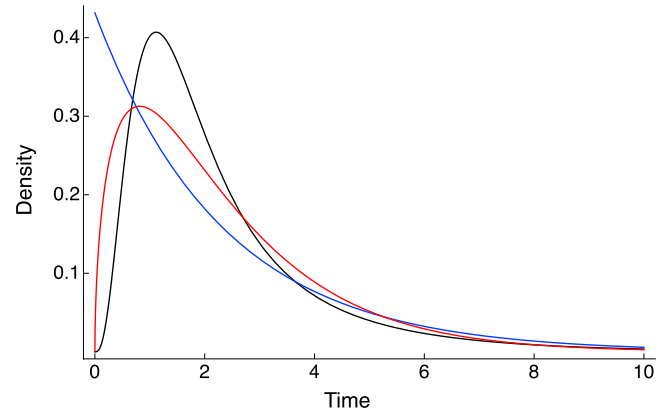

Fig. 4. Density of waiting time between the entrance state comprising 5 lineages in each deme and the exit state characterized by coalescence in deme 0 of a pair of lineages initially residing in deme 0, with all other lineages now residing in deme 1, under the assignment of population parameters $M_0 = M_1 = 0.3$ and $c_0 = c_1 = 0.01$. The black curve represents the analytical solution (9) for the density of internode length, blue an exponential distribution with the same mean, and red a gamma distribution with the same mean and variance. Time is measured in units of $4N$ generations. The indices of deviation (37) are $h = 0.73$ for the exponential and $h = 0.12$ for the gamma. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 5 illustrates that hitting probability (8) declines and the departure of the internode density from exponential increases as the minimum number of migration events separating the entrance and exit states increases. Accordingly, the density of internode length, marginalized over exit states, shows a good correspondence to an exponential distribution with the same mean (Fig. 6).

From an entrance state comprising b lineages in deme 0 and y in deme 1, the relative probability that the most recent event corresponds to migration rather than coalescence is

$$R = \frac{bM_0 + yM_1}{b(b-1)c_0 + y(y-1)c_1}. \tag{38}$$

Small values of R ($R \ll 1$) imply that the probability $1/(1+R)$ that the most recent event corresponds to coalescence is high, suggesting that the distribution of internode length may close to exponential. Alternatively, large R implies a high probability of at least one migration event since the most recent coalescence event. As the denominator of R (38) is of the second order in the number of lineages (b or y) while the numerator is of the first order, larger departures of the internode length distribution from exponential may be exhibited closer to the root of the gene genealogy (small b and y). Fig. 7 shows hitting probability (left) and index of deviation from exponential (right) against the minimum number of migration events for each of the 9 exit states accessible from the entrance state with 2 lineages in each deme ($b = y = 2$). Once again, large numbers of migration events appear to induce large deviations of the actual internode length distribution from exponential. However, hitting probability shows no obvious relationship to the minimum number of migration events for this case ($R = 60$), unlike Fig. 5 ($R = 7.5$), which corresponds to a genealogical level further from the root ($b = y = 5$). Accordingly, Fig. 8 indicates that the distribution of internode length, marginalized over exit state, departs strongly from both exponential and gamma for this case ($b = y = 2$), in contrast with Fig. 6 ($b = y = 5$).

3.3. Sensitivity analysis

For any specified state in the gene genealogy of a sample of arbitrary size, our method provides the probability of each

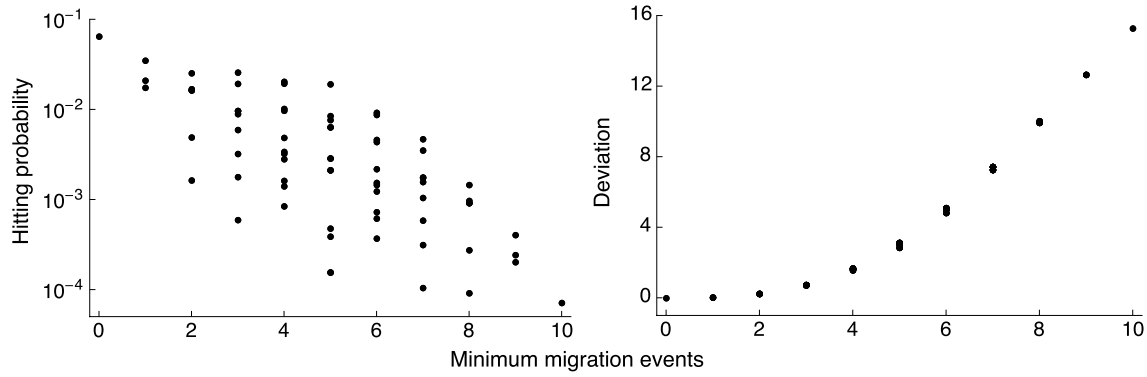

Fig. 5. Hitting probability (8) and the deviation (37) between the actual density (9) and an exponential distribution with the same mean against the minimum number of migration events separating the entrance and exit states for each of the 146 exit states accessible from the entrance state with 5 lineages in each deme, under the parameter assignments for Fig. 4.

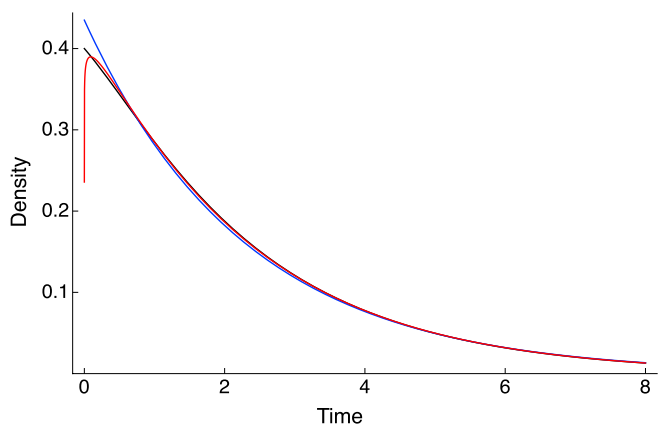

Fig. 6. Density of internode length, marginalized over the 146 exit states accessible from the entrance state with 5 lineages in each deme, under the parameter assignments for Fig. 4. The red curve corresponds to a gamma distribution with mean and variance matched to the actual density (9) (black), and blue an exponential distribution with matched mean. The indices of deviation (37) are $h = 5.1 \times 10^{-4}$ for the exponential and $h = 1.2 \times 10^{-5}$ for the gamma. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exit state (8) and the distribution of waiting time to that state (9). While straightforward, those closed-form expressions can be unattractive. To facilitate the development of an intuitive understanding of the major qualitative trends, we here describe a computational exploration of hitting probabilities, internode length distributions, and location of the MRCA.

We illustrate the contribution of recent patterns of migration and coalescence and of the characteristics of the sample to the location of the MRCA of the sample. We used `ms` (Hudson, 2002) to conduct simulations under a two-deme model with migration, applying a Bayesian regression analysis to summarize the effect of parameters of the model on the location of the MRCA.

For each simulated sample, we noted the location (deme 0 or deme 1) of the MRCA. We assigned independently from uniform distributions the population parameters, including rates of coalescence ($0 < c_0, c_1$, with $c_0 + c_1 = 2$), backward migration rates ($0 < M_0, M_1 < 1$), and the number of genes sampled from deme 0 ($1 \le l_0 \le n$, for n the total sample size). Using code provided by Kruschke (2011), we conducted a Bayesian logistic regression analysis of the dichotomous response variable of the deme containing the MRCA on three predictor variables: the proportion of genes sampled from deme 0,

$$\frac{l_0}{n}, \tag{39a}$$

the relative rate of coalescence in deme 0,

$$\frac{c_0}{c_0 + c_1} = \frac{N_1}{N_0 + N_1}, \tag{39b}$$

and the relative backward migration rate in deme 0,

$$\frac{M_0}{M_0 + M_1}. \tag{39c}$$

After applying a probit transformation to these proportions, we ran the code `MultipleLogisticRegressionBrugs.R` (Kruschke, 2011), which then performed standardization (subtraction of the mean and division by the variance) and specified generic priors (identical diffuse Gaussian distributions) for all predictor variables.

For samples of size $n = 100$, results of the analysis of variance (ANOVA) indicated nonsignificant departures from zero of the intercept and the effect of proportion of sample from deme 1 (39a) and highly significant effects of rates of coalescence ($p < 5 \times 10^{-12}$, (39b)) and backward migration ($p < 2 \times 10^{-16}$, (39c)). Fig. 9 shows the posterior distributions of coefficients of the regression equation estimated from the Bayesian analysis. In agreement with the ANOVA, the posterior distributions indicate that higher rates of coalescence (larger c_0) and lower rates of backward migration (smaller M_0) in deme 0 increase the probability that the MRCA resides in deme 0.

For samples of smaller size ($n = 10$), the ANOVA indicated a marginally significant effect of the proportion of the sample from deme 1 ($p < 8 \times 10^{-3}$, (39a)), in addition to confirming highly significant ($p < 2 \times 10^{-16}$) effects of both coalescence (39b) and backward migration (39c) rates. Fig. 10 shows the Bayesian posterior distributions. These histograms confirm the very strong effects of rates of coalescence and backward migration, and further indicate a weak trend that more intensive sampling in one deme increases the probability that the MRCA occurs in that deme.

4. Discussion

We have addressed the distribution of internode length in gene genealogies of samples of arbitrary size derived from a structured population. In the genealogy of a sample derived from a structured population, specification of the state of the process at any point in time may include the state of each lineage. In accommodating tracking of lineages with respect to their location at the entrance state of each level, for example, this construction includes a generalization of the model explored by Takahata and Slatkin (1990). Their primary interest lay not in coalescence times but rather in the probability distribution of topologies, which can be obtained from the hitting probabilities of successive exit states (8).

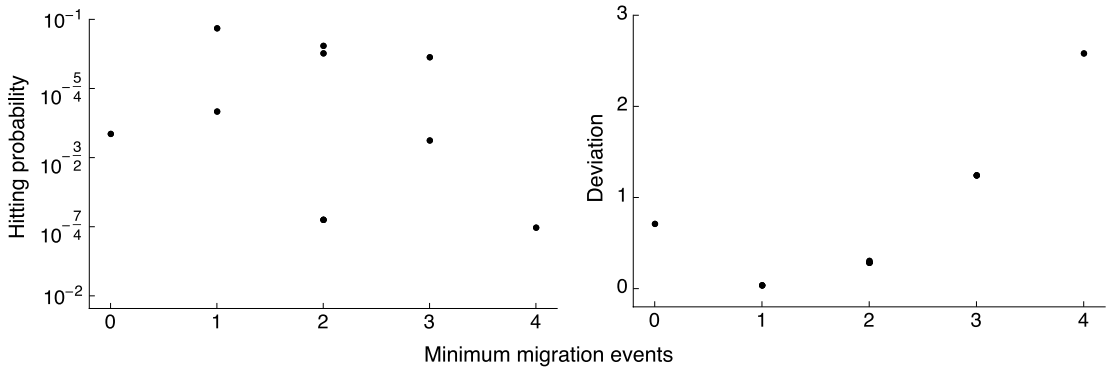

Fig. 7. Hitting probability (8) and the deviation (37) between the actual density (9) and an exponential distribution with the same mean against the minimum number of migration events separating the entrance and exit states for each of the 9 exit states accessible from the entrance state with 2 lineages in each deme, under the same model as for Fig. 5.

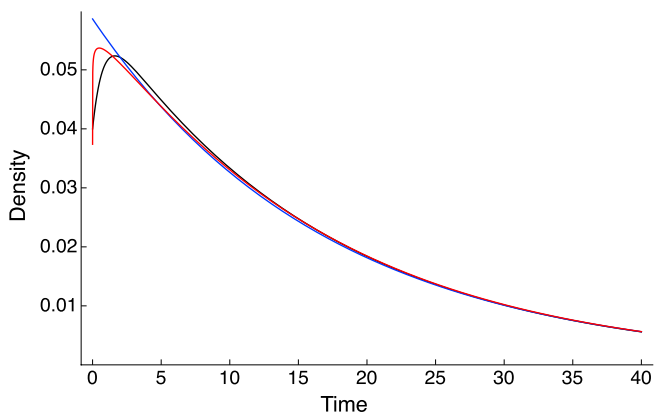

Fig. 8. Density of internode length, marginalized over the 9 exit states accessible from the entrance state with 2 lineages in each deme, under the same model as for Fig. 6. The black curve shows the actual density (9), with blue corresponding to an exponential distribution with the same mean and red to a gamma distribution with the same mean and variance. The indices of deviation (37) are $h = 6.1 \times 10^{-3}$ for the exponential and $h = 1.7 \times 10^{-3}$ for the gamma. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As indicated in the Introduction, our two-fold objective is to promote an intuitive understanding of the implications of population structure as well as to provide a model-based, computational approach to genealogical analysis. A Mathematica document (supplementaryNotebook.nb), which implements many of the tools developed here, is provided as online supplementary material (see Appendix D).

4.1. Qualitative trends

Among the key consequences of population structure is that patterns of variation observed in a sample depend upon the configuration of the sample. In the case of a sample of size $n = 2$, for example, the probabilities of identity and the age of the MRCA of a pair of genes depend on whether the genes are sampled from the same deme (*e.g.,* Hudson, 1990). In general, the location of all lineages at the time points bounding a level in the gene genealogy of a sample affects the hitting probability (8), internode length (9) and the number and pattern of mutations accumulated within the level (18). Our analysis (Section 3.2.1) illustrates the dependence of internode length on the starting and ending configurations of the sample. With regard to the reconstruction of geographical range within the framework of phylogeny–trait associations, for example, the trait of geographical location must be assigned, not only to the nodes of a genealogy, but to every lineage in a horizontal slice across the tree at a level boundary.

The sampling of genes from taxa distributed across a geographical range invites interpretations concerning the phylogeographical history of the taxa. Such interpretations can serve to generate hypotheses about the demographic history or origins of the taxa from which the genes were derived. While the progression to testing such hypotheses is now better appreciated (Knowles, 2004), the geographical location of the root of reconstructed gene trees continues to be regarded as support for the location of an ancestral population (e.g., Wang et al., 2012).

Our sensitivity analysis (Section 3.3) addresses the case in which the genealogy of the sample is contained entirely in the period subsequent to the origin of the demes. Within this period, the shape and location of the root of the genealogies primarily reflect recent patterns of residence and dispersal. Our results suggest that the location of the MRCA depends primarily on the relative rates of backward migration and coalescence (inverse of population size) in the demes. For small samples, the relative level of intensity of sampling (39a) can also influence the location of the MRCA, with the MRCA more likely to reside in the more intensively sampled deme.

Our qualitative findings agree with trends noted by Wakeley (2001) in his study of island-model migration: the MRCA tends to occur in demes with high coalescence rates (small effective sizes) or low backward migration rates. In their exploration of human origins, Takahata et al. (2001) examined genealogical relationships among genes sampled from Africa, Europe, and Asia. They found that the probability that the root of the genealogy lies in Africa increases with the effective size of the African population relative to the sum of the sizes of the non-African populations. This finding is consistent with ours because under their assumption of conservative migration (Strobeck, 1987), demes with higher effective sizes have lower backward migration rates.

4.2. Distribution of times to coalescence

A substantial body of literature on coalescence times in panmictic populations of constant or variable size has addressed various quantities, including the age of the most recent common ancestor of a sample, the distribution of the number of remaining lineages at a given point in the past, and total tree length (*e.g.,* Watterson, 1975; Tavaré, 1984; Griffiths, 1984; Griffiths and Tavaré, 1994, 1998; Takahata, 1988; Polanski et al., 2003; Chen and Chen, 2013). Much of the simplicity and elegance of this work reflects that in panmictic populations, the number of segregating sites, age of the MRCA, total tree length, and other aspects correspond to the sum of independent exponentially-distributed variables.

In structured populations, the distribution of internode length (time between successive coalescence events) departs in general

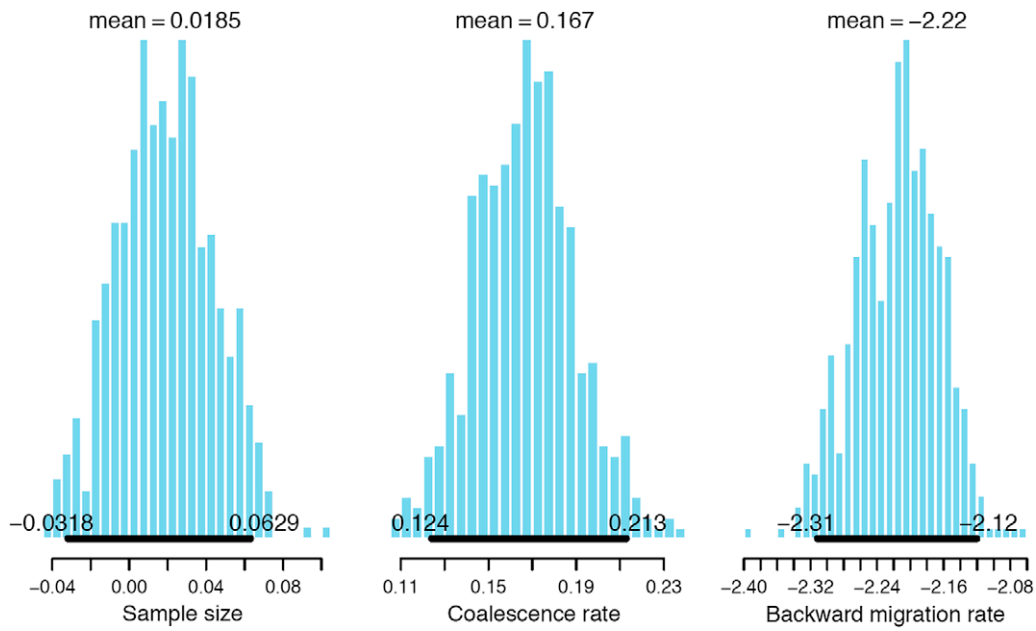

Fig. 9. Posterior distributions of regression coefficients for samples of size 100. Each histogram represents 501 samples from the posterior distribution, 167 recorded at intervals of 100 from each of three MCMC chains. Bars below the histograms indicate 95% credible intervals.

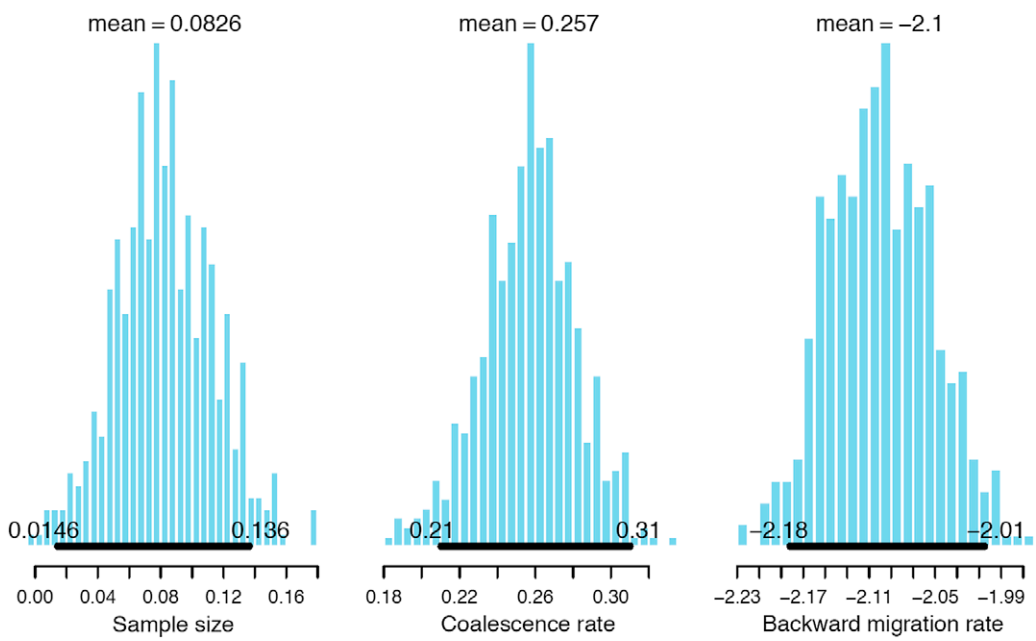

Fig. 10. Posterior distributions of regression coefficients for samples of size 10, with other aspects as given in Fig. 9. The proportion of the Sample size posterior distribution that lies below zero is 0.2%, suggesting a weak positive effect of greater sampling in deme 1 (39a) on the probability that the MRCA resides in that deme.

from exponential. A key property is that internode length depends on the state of the process at both bounding coalescence events. For $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively providing rates of within-level (non-coalescence) and between-level (coalescence) events, the density of waiting time between entrance state i and exit state j corresponds to the element in the ith row and jth column of $e^{\boldsymbol{U}t}\boldsymbol{V}$ (9). An implication of the association between state and branch length is that these aspects are best proposed jointly in Markov chain Monte Carlo (MCMC) algorithms.

Rannala et al. (2012) have explored the effects of misspecification of prior distributions in Bayesian analyses of genealogies. In particular, they demonstrated that adoption of a default prior (identical exponential distributions for all branch lengths) of the widely-used `MrBayes` (Ronquist and Huelsenbeck, 2003) can retard or prevent convergence of the MCMC sampler or even generate an artifactual mode in the posterior distribution.

Our exploration of the qualitative features of the density of internode length (Section 3.2) suggests that the intervention of one or more migration events between the states bounding a level of the genealogy can cause marked departures of the distribution of internode length from exponential. Higher values of R (38), an index of the probability of intervening migration, appear to be associated with greater departures from exponential. For a given set of rates of backward migration (M_i) and coalescence (c_i), R tends to increase as the numbers of lineages decline, suggesting that the departure from exponential of the marginal (over exit state) distribution of internode length may be more pronounced closer to the root (Fig. 8) than the tips (Fig. 6).

Virtually all studies that have addressed population structure have noted that computation of internode lengths very rapidly becomes intractable as sample size grows (*e.g.*, Griffiths, 1984; Takahata, 1988; Polanski et al., 2003; Hobolth et al., 2011). While our approach shares this feature, we have demonstrated the computation of exact densities of MRCA age and total branch length for samples of size $n = 20$ genes (Section 3.1). We suggest that determination of the exact distribution may be feasible and efficient for levels of the tree close to the root (fewer lineages), for which internode length departs strongly from exponential (Fig. 8). An exponential distribution better approximates the marginal distribution further from the root (more lineages, smaller R), a region of the tree for which the hitting probability of an exit state tends to decline with the minimum number of migration events separating it from the entrance state. Exponential or gamma distributions with moments matched to the exact distribution may serve as useful approximations to the internode length distribution for genealogical levels distant from the root, with computation of the analytical density (9) recommended for levels close to the root.

Our approach through generating functions (Section 2.2) can produce full densities of coalescence times or mutation numbers for moderate sample sizes ($n = 20$), rather larger than can currently be accommodated by various other methods. Further, computation of moments of the distributions (*e.g.*, Tables 1 and 2) is quite stable and rapid. In test comparisons with simulated densities, our Mathematica implementation has proved capable of accurately computing the first four moments for samples of 100 genes. Within the context of large genealogies, we suggest that useful approximations to the full densities may be developed from the first k moments (30).

Acknowledgments

We thank John C. Avise and Francisco J. Ayala for their gracious hospitality during a sabbatical leave for MKU at the University of California at Irvine, Benjamin D. Redelings for help with the Bayesian regression, and Editor Noah Rosenberg, Thomas Mailund, and an anonymous reviewer for constructive comments. Public Health Service grant GM 37841 (MKU) provided partial funding for this research.

Appendix A. Monomorphism, tree length, and number of segregating sites

We derive (12), which summarizes relationships among the probability of a monomorphic sample (P), the pgf of the number of segregating sites (g_Z), and the mgf of tree length (h_T).

Let $\hat{T}$ represent the total number of generations in the genealogical tree of a sample of arbitrary size (n). For a sample of size 2, $\hat{T}$ corresponds to twice the age of the MRCA. Under the infinite-sites model, in which all mutations are detectable and distinguishable, the number of segregating sites in a sample corresponds to the number of mutations in the tree (Watterson, 1975). We assume that given tree length, the number of segregating sites (Z) has a Poisson distribution:

$$Z|T \sim \text{Poi}[\theta T],$$

for T corresponding to tree length in units of $2N$ generations,

$$T = \hat{T}/2N,$$

and θ the scaled mutation rate (16). In particular, the probability of a monomorphic sample corresponds to the zero term of the Poisson distribution:

$$\Pr(Z = 0|T = t) = e^{-\theta t}.$$

Taking the expectation over the tree length, we obtain $P(\theta)$, the unconditioned probability of a monomorphic sample:

$$\begin{aligned} P(\theta) &= \int_0^\infty \Pr(Z = 0|T = t) f(t)dt \\ &= E_T[e^{-\theta t}], \end{aligned}$$

for $f(t)$ denoting the density of tree length and E_T the expectation with respect to tree length. Similarly, the unconditioned probability generating function (pgf) of the number of segregating sites corresponds to

$$\begin{aligned} g_Z(a) &= \sum_{k=0}^{\infty} \int_0^\infty \frac{(a\theta t)^k e^{-\theta t}}{k!} f(t)dt \\ &= \int_0^\infty e^{\theta(a-1)t} f(t)dt \\ &= E_T[e^{\theta(a-1)t}]. \end{aligned} \tag{A.1}$$

Hudson (1990) observed that the probability of identity between a pair of genes represents the moment generating function (mgf) of tree length T with parameter $-\theta$:

$$P(\theta) = \mathbb{E}[e^{-\theta T}] = h_T(-\theta). \tag{A.2}$$

This expression, together with (A.1), implies (12).

Appendix B. Derivatives of generating functions

We address the development of a recursion in moments from the recursion in moment generating functions for total tree length (28),

$$\boldsymbol{h}_\ell(b) = [\boldsymbol{I} - \boldsymbol{H}_\ell(b)(\boldsymbol{L} + \boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(b)\boldsymbol{V}_\ell\boldsymbol{h}_{\ell-1}(b).$$

Note that the present level introduces terms involving b only through $\boldsymbol{H}_\ell(b)$ (29), the first derivative of which corresponds to

$$\boldsymbol{H}_\ell^{(1)}(b) = \ell\boldsymbol{H}_\ell^2(b).$$

We first determine the first derivative of $[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}$ for

$$\boldsymbol{X}(b) = \boldsymbol{H}_\ell(b)(\boldsymbol{L} + \boldsymbol{U}_\ell).$$

Using the product rule to determine the derivative of

$$[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1} = \boldsymbol{I} + \boldsymbol{X}(b) + \boldsymbol{X}(b)^2 + \cdots,$$

we take the derivative of the first, second, $\ldots \boldsymbol{X}(b)$ in each power of $\boldsymbol{X}(b)$:

$$\begin{aligned} &\frac{d\left([\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\right)}{db} \\ &= \boldsymbol{X}^{(1)}(b) + \boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b) + \boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b)^2 + \cdots \\ &\quad + \boldsymbol{X}(b)\boldsymbol{X}^{(1)}(b) + \boldsymbol{X}(b)\boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b) \\ &\quad + \boldsymbol{X}(b)\boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b)^2 + \cdots + \boldsymbol{X}(b)^2\boldsymbol{X}^{(1)}(b) \\ &\quad + \boldsymbol{X}(b)^2\boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b) + \boldsymbol{X}(b)^2\boldsymbol{X}^{(1)}(b)\boldsymbol{X}(b)^2 + \cdots \\ &= \{\boldsymbol{X}^{(1)}(b) + \boldsymbol{X}(b)\boldsymbol{X}^{(1)}(b) + \boldsymbol{X}(b)^2\boldsymbol{X}^{(1)}(b)\}[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1} \\ &= [\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{X}^{(1)}(b)[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}, \end{aligned}$$

in which the first derivative of $\boldsymbol{X}(b)$ corresponds to

$$\boldsymbol{X}^{(1)}(b) = \boldsymbol{H}_\ell^{(1)}(b)(\boldsymbol{L} + \boldsymbol{U}_\ell) = \ell\boldsymbol{H}_\ell(b)\boldsymbol{X}(b).$$

We then obtain

$$\begin{aligned} &\frac{d\left([\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\right)}{db} \\ &= [\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{X}^{(1)}(b)[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b) + [\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell^{(1)}(b) \\ &= \ell\{[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\boldsymbol{X}(b)[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b) \\ &\quad + [\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)^2\} \\ &= \ell[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\{\boldsymbol{X}(b)[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b) + \boldsymbol{H}_\ell(b)\} \\ &= \ell[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\{\boldsymbol{X}(b) + \boldsymbol{I} + \boldsymbol{X}(b)\}[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b) \\ &= \ell\{[\boldsymbol{I} - \boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\}^2, \end{aligned}$$

with the jth derivative given by

$$\frac{d^j\left([\boldsymbol{I}-\boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\right)}{db^j}=j!\,\ell^j\{[\boldsymbol{I}-\boldsymbol{X}(b)]^{-1}\boldsymbol{H}_\ell(b)\}^{j+1}.$$

For level ℓ, the first derivative with respect to b of the mgfs of total tree length corresponds to

$$\begin{aligned}\boldsymbol{h}_\ell^{(1)}(b) &= [\boldsymbol{I}-\boldsymbol{H}_\ell(b)(\boldsymbol{L}+\boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(b)\boldsymbol{V}_\ell\boldsymbol{h}_{\ell-1}^{(1)}(b)\\ &\quad+\ell\{[\boldsymbol{I}-\boldsymbol{H}_\ell(b)(\boldsymbol{L}+\boldsymbol{U}_\ell)]^{-1}\boldsymbol{H}_\ell(b)\}^2\boldsymbol{V}_\ell\boldsymbol{h}_{\ell-1}(b),\end{aligned}$$

and the kth derivative to (30). Recursions in derivatives of the pgfs for the number of segregating sites can be obtained by substituting $\boldsymbol{g}_\ell(a)$ for $\boldsymbol{h}_\ell(b)$, $\tilde{\boldsymbol{G}}_\ell(a)$ for $\tilde{\boldsymbol{H}}_\ell(b)$, and $\ell\theta$ for ℓ.

Appendix C. Pairwise comparisons

To link our results for samples of arbitrary size to the many earlier pairwise analyses, we address a sample of $n=2$ genes, sampled either from the same or different demes, which themselves descend from a panmictic ancestral population.

Transition rates

In the post-divergence era (more recent than the divergence of the demes), the scaled rates of migration (M_j) and coalescence (c_j) remain as defined in (15a) and (15b). In addition, we assume an exponential waiting time to the divergence event, with parameter corresponding to a per-generation rate of s or

$$S=\lim_{\substack{s\to 0\\ N\to\infty}}2Ns$$

in units of $2N$ generations. For example, on level ℓ, with i the current number of lineages in deme 0, the probability that the most recent event corresponds to a migration of a lineage in deme 0 is

$$\begin{aligned}&\lim\frac{im_0}{im_0+(\ell-i)m_1+\binom{i}{2}/N_0+\binom{\ell-i}{2}/N_1+s}\\ &=\frac{iM_0}{iM_0+(\ell-i)M_1+\binom{i}{2}2c_0+\binom{\ell-i}{2}2c_1+S}.\end{aligned}$$

A similar construction proved useful in the analysis of interspecific introgression conducted by Leman et al. (2005).

In the pre-divergence era, we now interpret the arbitrary constant N as the effective number of genes in the ancestral population. On level ℓ, only one entrance state and one exit state exist, reflecting the number of ancestral lineages, with the waiting time to the exit state (coalescence) having an exponential distribution with per-generation parameter

$$\binom{\ell}{2}\Big/N,$$

implying an expected internode length of $1/[\ell(\ell-1)]$ in units of $2N$ generations.

Two genes sampled from distinct demes

From the hitting probabilities (8), a pair of lineages residing in distinct demes coalesce in the post-divergence era in deme 0 with probability

$$2c_0M_1(2c_1+2M_1+S)/D, \tag{C.1}$$

in which

$$\begin{aligned}D &= S[S+c_0+c_1+M_0+M_1][S+c_0+c_1+2(M_0+M_1)]\\ &\quad-S(c_0-c_1)(c_0+M_0-c_1-M_1)\\ &\quad+2[c_0M_1(S+2c_1+2M_1)+c_1M_0(S+2c_0+2M_0)].\end{aligned} \tag{C.2}$$

This probability decreases with the rate of speciation (S). It increases with the relative rate of coalescence in deme 0 (c_0) and decreases with the coalescence rate in deme 1 (c_1). Low backward migration rates in deme 0 (small M_0) and high rates in the opposite deme (large M_1) promote coalescence in deme 0. The probability of coalescence in deme 1 is identical to (C.1), but with the indices denoting deme reversed. Because the lineages coalesce either in one of the demes (post-divergence) or in the ancestral species (pre-divergence), the probability that coalescence predates divergence is the complement of the sum of the probabilities of coalescence in deme 0 and deme 1.

Wakeley (1996) studied the mean and variance of the number of substitutions between a pair of genes under the symmetric case, assuming an ancestral population size twofold greater than either descendant deme ($N_0=N_1=N/2$):

$$\begin{aligned}M_0&=M_1=M\\ c_0&=c_1=2.\end{aligned} \tag{C.3}$$

Using his results, Rosenberg and Feldman (2002, their (9.7)) gave the probability of coalescence in the post-speciation era given the speciation time:

$$\begin{aligned}&1-e^{-2(1+M)t}\{\sqrt{1+M^2}\cosh[2t\sqrt{1+M^2}]\\ &\quad+(1+M)\sinh[2t\sqrt{1+M^2}]\}/\sqrt{1+M^2},\end{aligned} \tag{C.4}$$

for t the time since speciation in units of $2N$ generations.

From (C.1), we obtain the probability of coalescence in the post-speciation era under special case (C.3):

$$\frac{8M}{S^2+4S(1+M)+8M}. \tag{C.5}$$

Rather than assuming a time since speciation, our formulation incorporates its estimation into the same framework, characterizing its density as exponential with parameter S. Integrating (C.4) over this density produces (C.5).

Two genes sampled from the same deme

Two lineages sampled from deme 1 coalesce in the pre-divergence era (P), in deme 0 (P_0), and deme 1 (P_1) with probabilities

$$\begin{aligned}P &= S\ \{2[c_1(3M_0+M_1)+(M_0+M_1)^2]\\ &\quad+S[2c_1+3(M_0+M_1)]+S^2\}/D\end{aligned} \tag{C.6a}$$

$$\begin{aligned}P_0 &= 2c_0[2c_1(M_0+M_1)+2M_1^2\\ &\quad+S(2c_1+M_0+3M_1)+S^2]/D\end{aligned} \tag{C.6b}$$

$$P_1=4c_1M_0^2/D, \tag{C.6c}$$

respectively, for D given in (C.2). Comparison between (C.1) and (C.6b) indicates that a pair of lineages sampled from deme 0 are more likely to coalesce in deme 0 than a pair sampled from distinct demes for any finite size of deme 0 ($c_0>0$). Similarly, comparison between (C.1) and (C.6c) indicates that the pair are less likely to coalesce in deme 1 for any finite size of deme 1 and positive rate of backward migration to deme 1 ($c_1M_0>0$). A pair of genes sampled from the same deme are less likely to coalesce in the pre-speciation era than a pair sampled from different demes if

$$S[c_0(2c_1+M_0+M_1+S)+c_0M_1-c_1M_0]>0.$$

This condition is satisfied, for example, under the constraint of equal numbers of migrant lineages in the two demes, corresponding to Strobeck's (1987) criterion for conservative migration:

$$c_0M_1-c_1M_0=\frac{2N^2(N_1m_1-N_0m_0)}{N_0N_1}=0. \tag{C.7}$$

Under some conditions, sampling of the lineages from the same deme can increase the probability that coalescence occurs in the pre-speciation era over the case in which they are sampled from different demes. This case requires that coalescence occurs more slowly in deme 0 than in deme 1 ($c_1 > c_0$) and a sufficiently high backward migration rate in deme 0:

$$M_0 > \frac{c_0(S + 2M_1 + 2c_1)}{c_1 - c_0}.$$

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.tpb.2015.01.003.

References

Andersen, L.N., Mailund, T., Hobolth, A., 2013. Efficient computation in the IM model. J. Math. Biol. 68, 1423–1451.

Chen, H., Chen, K., 2013. Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. Genetics 194, 721–736.

Crow, J.F., Maruyama, T., 1971. The number of neutral alleles maintained in a finite, geographically structured population. Theor. Popul. Biol. 2, 437–453.

Griffiths, R.C., 1981. The number of heterozygous loci between two randomly chosen completely linked sequences of loci in two subdivided population models. J. Math. Biol. 12, 251–261.

Griffiths, R.C., 1984. Asymptotic line-of-descent distributions. J. Math. Biol. 21, 67–75.

Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B 344, 403–410.

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

Hobolth, A., Andersen, L.N., Mailund, T., 2011. On computing coalescence time density in an isolation-with-migration model with few samples. Genetics 187, 1241–1243.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D., Antonovics, J. (Eds.), Oxford Surveys in Evolutionary Biology, Vol. 7. Oxford Univ. Press, New York, pp. 1–44.

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

Innan, H., Watanabe, H., 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol. Biol. Evol. 23, 1040–1047.

Knowles, L.L., 2004. The burgeoning field of statistical phylogeography. J. Evol. Biol. 17, 1–10.

Kruschke, J.K., 2011. Doing Bayesian Data Analysis. Academic Press, Burlington, MA.

Leman, S.C., Chen, Y., Stajich, J.E., Noor, M.A.F., Uyenoyama, M.K., 2005. Likelihoods from summary statistics: recent divergence between species. Genetics 171, 1419–1436.

Li, W.-H., 1976. Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. 10, 303–308.

Lohse, K., Harrison, R.J., Bartoni, N.H., 2011. A general method for calculating likelihoods under the coalescent process. Genetics 189, 977–987.

Mailund, T., Dutheil, J.Y., Hobolth, A., Lunter, G., Schierup, M.H., 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genet. 7, e1001319.

Nath, H.B., Griffiths, R.C., 1993. The coalescent in two colonies with symmetric migration. J. Math. Biol. 31, 841–852.

Nei, M., Feldman, M.W., 1972. Identity of genes by descent within and between populations under mutation and migration pressures. Theor. Popul. Biol. 3, 460–465.

Neuts, M.F., 1995. Algorithmic Probability: A Collection of Problems. Chapman & Hall, London.

Nielsen, R., Wakeley, J., 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158, 885–896.

Polanski, A., Bobrowski, A., Kimmel, M., 2003. A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63, 33–40.

Rannala, B., Zhu, T., Yang, Z., 2012. Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. Mol. Biol. Evol. 29, 325–335.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Rosenberg, N.A., Feldman, M.W., 2002. The relationship between coalescence times and population divergence times. In: Slatkin, M., Veuille, M. (Eds.), Modern Developments in Theoretical Population Genetics—The Legacy of Gustave Malécot. Oxford University Press, Oxford, pp. 130–164.

Strobeck, C., 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics 117, 149–155.

Takahata, N., 1988. The coalescent in two partially isolated diffusion populations. Genet. Res. 52, 213–222.

Takahata, N., Lee, S.-H., Satta, Y., 2001. Testing multiregionality of modern human origins. Mol. Biol. Evol. 18, 172–183.

Takahata, N., Slatkin, M., 1990. Genealogy of neutral genes in two partially isolated populations. Theor. Popul. Biol. 38, 331–350.

Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164.

Taylor, H.M., Karlin, S., 1998. An Introduction to Stochastic Modeling, third ed. Academic Press, New York.

Uyenoyama, M.K., Takebayashi, N., 2004. A simple method for computing exact probabilities of mutation numbers. Theor. Popul. Biol. 65, 271–284.

Wakeley, J., 1996. Pairwise differences under a general a model of population subdivision. J. Genet. 75, 81–89.

Wakeley, J., 2001. The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. 59, 133–144.

Wang, Y., Hey, J., 2010. Estimating divergence parameters with small samples from a large number of loci. Genetics 184, 363–379.

Wang, B., Jiang, J., Xie, F., Li, C., 2012. Postglacial colonization of the Qinling Mountains: phylogeography of the swelled vent frog (*Feirana quadranus*). PLoS One 7, e41579.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276.

Wilkinson-Herbots, H.M., 2008. The distribution of the coalescence time and the number of pairwise nucleotide differences in the isolation with migration model. Theor. Popul. Biol. 73, 277–288.

Zhu, T., Yang, Z., 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. Mol. Biol. Evol. 29, 3131–3142.