Theoretical Population Biology 122 (2018) 46–56

Contents lists available at ScienceDirect

Theoretical Population Biology

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

On the joint distribution of tree height and tree length under the coalescent

Ilana M. Arbisser [a,*], Ethan M. Jewett [b], Noah A. Rosenberg [a]

[a] Department of Biology, Stanford University, Stanford, CA 94305, USA
[b] Departments of Electrical Engineering & Computer Science and Statistics, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 3 April 2017
Available online 10 November 2017

Keywords:
Migration
Population growth
Time to the most recent common ancestor

ABSTRACT

Many statistics that examine genetic variation depend on the underlying shapes of genealogical trees. Under the coalescent model, we investigate the joint distribution of two quantities that describe genealogical tree shape: tree height and tree length. We derive a recursive formula for their exact joint distribution under a demographic model of a constant-sized population. We obtain approximations for the mean and variance of the ratio of tree height to tree length, using them to show that this ratio converges in probability to 0 as the sample size increases. We find that as the sample size increases, the correlation coefficient for tree height and length approaches $(\pi^2 - 6)/[\pi\sqrt{2\pi^2 - 18}] \approx 0.9340$. Using simulations, we examine the joint distribution of height and length under demographic models with population growth and population subdivision. We interpret the joint distribution in relation to problems of interest in data analysis, including inference of the time to the most recent common ancestor. The results assist in understanding the influences of demographic histories on two fundamental features of tree shape.

1. Introduction

Coalescent models that consider different demographic histories for a population often produce genealogies with distinct properties. Differing shapes for genealogies generated under specific coalescent models provide information about the demographic processes that have given rise to sampled DNA sequences.

Efforts to use the shapes of coalescent trees to investigate the effects of population-genetic processes have examined a variety of summaries of tree shape (e.g. Slatkin, 1996; Uyenoyama, 1997; Schierup and Hein, 2000; Rosenberg, 2006; Disanto and Wiehe, 2016). These summaries have included both discrete properties, including counts of nodes with particular numbers of descendants, as well as continuous properties, such as the mean pairwise coalescence time, external branch lengths, internal branch lengths, and lengths of the two branches descended immediately from the root.

Perhaps the two most frequently used summaries of coalescent trees are the tree height, H – also known as the time to the most recent common ancestor (T_{MRCA}) – and the tree length, L, which sums the total length of all branches in a genealogy. The tree height has been of great interest for its role in understanding the age of a genealogy and its relationship to the origin of a population—for example, in estimating the times of the most recent common ancestors of all human Y chromosomes or mitochondrial DNA sequences (e.g. Thomson et al., 2000; Tang et al., 2002; Mendez et al., 2013). The tree length is important because of its direct relationship with the count of derived mutations in a sample of DNA sequences resulting from the genealogy, a basic statistic of population genetics that is informative about population parameters (e.g. Tavaré et al., 1997; Joyce, 1999).

Previous theoretical studies have examined properties of genealogical summary measures, computing features of their means, variances, and distributions under various coalescent models. Consider a constant-population-size coalescent model in which n lineages are sampled in a haploid population of size N. Let T_k denote the random time during which exactly k lineages are ancestral to the sample. Measuring time in units of N generations, T_k, for $t_k \geq 0$, is exponentially distributed with probability density (Kingman, 1982; Wakeley, 2009)

$$f_{T_k}(t_k) = \binom{k}{2} e^{-\binom{k}{2} t_k}. \tag{1}$$

For each $n \geq 2$, the tree height, H_n, satisfies

$$H_n = \sum_{k=2}^{n} T_k. \tag{2}$$

* Corresponding author.
E-mail address: ilanama@stanford.edu (I.M. Arbisser).

https://doi.org/10.1016/j.tpb.2017.10.008
0040-5809/

Under a constant-sized model, H_n has mean and variance (e.g. Tavaré et al., 1997; Wakeley, 2009)

$$\mathbb{E}[H_n] = \sum_{k=2}^{n} \mathbb{E}[T_k] = 2\left(1 - \frac{1}{n}\right) \tag{3}$$

$$\mathrm{Var}[H_n] = \left(8\sum_{k=2}^{n} \frac{1}{k^2}\right) - 4\left(1 - \frac{1}{n}\right)^2, \tag{4}$$

with limits $\lim_{n\to\infty}\mathbb{E}[H_n] = 2$ and $\lim_{n\to\infty}\mathrm{Var}[H_n] = \frac{4\pi^2}{3} - 12$.

For each $n \geq 2$, the corresponding tree length, L_n, satisfies

$$L_n = \sum_{k=2}^{n} kT_k, \tag{5}$$

and its mean and variance under the constant-sized model are (e.g. Tavaré et al., 1997; Wakeley, 2009)

$$\mathbb{E}[L_n] = \sum_{k=2}^{n} k\mathbb{E}[T_k] = 2\sum_{k=1}^{n-1} \frac{1}{k} \tag{6}$$

$$\mathrm{Var}[L_n] = 4\sum_{k=1}^{n-1} \frac{1}{k^2}, \tag{7}$$

with limits $\lim_{n\to\infty}\mathbb{E}[L_n] = \infty$ and $\lim_{n\to\infty}\mathrm{Var}[L_n] = \frac{2\pi^2}{3}$.

Relationships among two or more summaries potentially provide more information about genealogical phenomena than can be obtained from studying the distribution of a single quantity alone. This principle underlies many coalescent-based statistics used in population-genetic data analysis. For example, Tajima's D statistic for testing the agreement of DNA sequence data to the selectively neutral constant-sized coalescent model can be interpreted as a comparison of mean pairwise coalescence time and total tree length to determine if they fit the same underlying model (Tajima, 1989). Similarly, the related D statistics of Fu and Li (1993) can be interpreted as a comparison of the length of the external branches of a genealogy to that of the internal branches. In fact, many neutrality tests that rely on counts of the number of sequence positions with different allele frequencies – the site-frequency spectrum – can be viewed as comparisons between summary statistics representing different aspects of tree shape (Achaz, 2009; Ferretti et al., 2017).

Here, to deepen an understanding of the effects of demographic phenomena on the shapes of genealogical trees, we investigate the joint distribution of tree height and tree length in a series of coalescent models. Both H and L have appeared in studies focusing on tree shape. Rosenberg and Hirsh (2003) explored the bias that occurs in estimating H for a genealogy using L-based estimators. Uyenoyama (1997) used four ratios involving H and L with pairwise coalescence time, the sum of external branch lengths, and the average length of the two base branches in developing a method for characterizing genealogical structure. Fu (1996) and Griffiths and Tavaré (1996) studied the joint distribution of H together with the number of segregating sites, S, whose expectation is proportional to L. Our goal is different: S is a property of a sample of sequences, whereas we aim to study properties of the genealogies themselves.

We consider this joint distribution both theoretically under the standard neutral coalescent with constant population size as well as by simulation under exponential growth and population structure models. The results illustrate how the distribution is concentrated near either the upper or the lower bound on H_n in terms of L_n as values of the demographic parameters change. We interpret the behavior of H_n and L_n in relation to problems of interest in data analysis.

2. Theory

2.1. Upper and lower bounds

To evaluate the joint distribution of (H_n, L_n), we first determine the space of possible values for the pair of variables (H_n, L_n). Suppose $H_n \geq 0$ is fixed. Then L_n is maximized when T_n contributes a large proportion of the height of a genealogy (Fig. 1A) and minimized when, instead, T_2 contributes a large proportion of the height (Fig. 1B). In the former case, L_n approaches a maximum of nH_n, whereas in the latter, L_n approaches a minimum of $2H_n$. Thus, under any demographic model, the joint probability density of H_n and L_n must be bounded between the lines $H_n = \frac{1}{n}L_n$ and $H_n = \frac{1}{2}L_n$.

For $n \geq 2$, recalling the expressions for H_n and L_n from Eqs. (2) and (5), respectively, when T_n is large in relation to $T_2, T_3, \ldots, T_{n-1}$,

$$\frac{H_n}{L_n} = \frac{\sum_{k=2}^{n} T_k}{\sum_{k=2}^{n} kT_k} = \frac{\left(\sum_{k=2}^{n-1} T_k\right) + T_n}{\left(\sum_{k=2}^{n-1} kT_k\right) + nT_n} \approx \frac{T_n}{nT_n} = \frac{1}{n}. \tag{8}$$

Alternatively, when T_2 is large in relation to $T_3, T_4, \ldots, T_n$,

$$\frac{H_n}{L_n} = \frac{\sum_{k=2}^{n} T_k}{\sum_{k=2}^{n} kT_k} = \frac{\left(\sum_{k=3}^{n} T_k\right) + T_2}{\left(\sum_{k=3}^{n} kT_k\right) + 2T_2} \approx \frac{T_2}{2T_2} = \frac{1}{2}. \tag{9}$$

2.2. Recursion for the joint density of H and L under a constant-sized model

With the domain for (H_n, L_n) established, we now obtain a recursive formula for the joint probability density of H_n and L_n. Let $f_{H_n,L_n}(h, \ell)$ be the exact joint density function for H and L at sample size n. For $n \geq 4$, under a constant-sized coalescent model, $f_{H_n,L_n}(h, \ell)$ can be determined recursively.

Let $f_{T_k}(t_k)$ be the probability density of the waiting time until the coalescence of k lineages to $k-1$ lineages, as given by Eq. (1). Then for $n \geq 4$,

$$f_{H_n,L_n}(h, \ell) = \int_{\alpha}^{\beta} f_{H_{n-1},L_{n-1}}(h - t_n, \ell - nt_n) f_{T_n}(t_n)\, dt_n \tag{10}$$

$$\alpha = \max\{0, \ell - (n-1)h\} \tag{11}$$

$$\beta = \frac{\ell - 2h}{n-2}. \tag{12}$$

We set the base case of the recursion to be the simplest nontrivial case, $n = 3$; we will see that $f_{H_3,L_3}(h, \ell) = 3e^{-2\ell+3h}$ for $h \geq 0$ and $\frac{1}{3}\ell \leq h \leq \frac{1}{2}\ell$. Note that for $n = 2$, because $H_2 = T_2$ and $L_2 = 2T_2$, $f_{H_2,L_2}(h, \ell) = f_{T_2}(t_2) = e^{-t_2} = e^{-h} = e^{-\frac{\ell}{2}}$, with nonzero values only when $h = \frac{1}{2}\ell$.

The recursive formula makes use of the fact that $T_n, T_{n-1}, \ldots, T_2$ are independent random variables. Therefore, for $n \geq 4$, $f_{H_n,L_n}(h, \ell|t_n) = f_{H_{n-1},L_{n-1}}(h - t_n, \ell - nt_n)$. Thus, the joint probability density of H_n and L_n for a genealogy with n lineages can be found using the joint density of H_{n-1} and L_{n-1} for a genealogy of $n-1$ lineages, marginalizing over all possible values of t_n.

To determine the bounds of integration, Eqs. (11) and (12), in Eq. (10), we separately consider the upper bound, β, and lower bound, α. Because the density is calculated for a given height, h, and length, ℓ, the bounds for t_n are considered for (h, ℓ) with $h \geq 0$ and $\frac{1}{n}\ell \leq h \leq \frac{1}{2}\ell$.

For the upper bound, β, for t_n given h and ℓ, we minimize all other waiting times between subsequent coalescences. In this scenario, as t_n approaches its maximum given h and ℓ, ℓ approaches $2(h - t_n) + nt_n$ from above, where the first term $2(h - t_n)$ is the minimal length for the parts of the tree subsequent to the interval

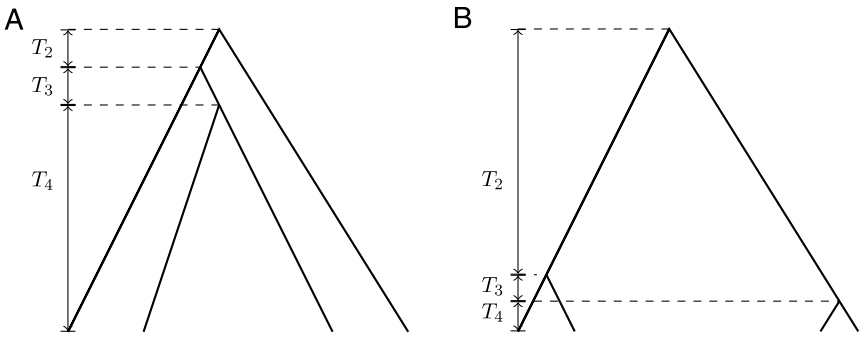

Fig. 1. Two extreme tree topologies that illustrate the bounds on H as a function of L. (A) Topology for which $\frac{H_n}{L_n} \approx \frac{1}{n}$, as in Eq. (8). (B) Topology for which $\frac{H_n}{L_n} \approx \frac{1}{2}$, as in Eq. (9). In both panels, $n = 4$.

Table 1
Joint density of tree height and length in a population of constant size, $f_{H_n,L_n}(h, \ell)$, for sample sizes $n = 3$, 4, and 5.

n	Domain for ℓ	$f_{H_n,L_n}(h, \ell)$	
		Computation	Result
3	$[2h, 3h]$	Base case	$3e^{-2\ell+3h}$
4	$[2h, 3h]$	$\int_0^{(\ell-2h)/2} 3e^{-2(\ell-4t_4)+3(h-t_4)}\, 6e^{-6t_4}\, dt_4$	$18(e^{-2\ell+3h} - e^{-\frac{5}{2}\ell+4h})$
	$[3h, 4h]$	$\int_{\ell-3h}^{(\ell-2h)/2} 3e^{-2(\ell-4t_4)+3(h-t_4)}\, 6e^{-6t_4}\, dt_4$	$18(e^{-3\ell+6h} - e^{-\frac{5}{2}\ell+4h})$
5	$[2h, 3h]$	$\int_0^{(\ell-2h)/3} 18[e^{-2(\ell-5t_5)+3(h-t_5)} - e^{-\frac{5}{2}(\ell-5t_5)+4(h-t_5)}]\, 10e^{-10t_5}\, dt_5$	$60(e^{-3\ell+5h} - 2e^{-\frac{5}{2}\ell+4h} + e^{-2\ell+3h})$
	$[3h, 4h]$	$\int_0^{(\ell-3h)/2} 18[e^{-3(\ell-5t_5)+6(h-t_5)} - e^{-\frac{5}{2}(\ell-5t_5)+4(h-t_5)}]\, 10e^{-10t_5}\, dt_5 +$ $\int_{(\ell-3h)/2}^{(\ell-2h)/3} 18[e^{-2(\ell-5t_5)+3(h-t_5)} - e^{-\frac{5}{2}(\ell-5t_5)+4(h-t_5)}]\, 10e^{-10t_5}\, dt_5$	$60(e^{-3\ell+5h} - 2e^{-\frac{7}{2}\ell+\frac{15}{2}h} - 2e^{-\frac{5}{2}\ell+4h} + 3e^{-3\ell+6h})$
	$[4h, 5h]$	$\int_{\ell-4h}^{(\ell-3h)/2} 18[e^{-3(\ell-5t_5)+6(h-t_5)} -$ $e^{-\frac{5}{2}(\ell-5t_5)+4(h-t_5)}]\, 10e^{-10t_5}\, dt_5 + \int_{(\ell-3h)/2}^{(\ell-2h)/3} 18[e^{-2(\ell-5t_5)+3(h-t_5)} -$ $e^{-\frac{5}{2}(\ell-5t_5)+4(h-t_5)}]\, 10e^{-10t_5}\, dt_5$	$60(e^{-3\ell+5h} - 2e^{-\frac{7}{2}\ell+\frac{15}{2}h} + e^{-4\ell+10h})$

The function $f_{H_n,L_n}(h, \ell)$ in Eq. (10) is defined for (h, ℓ) satisfying $h \geq 0$ and $2h \leq \ell \leq nh$. Owing to the maximum that appears in Eq. (11), $f_{H_n,L_n}(h, \ell)$ has a piecewise structure, with transitions at $\ell = jh$ for $j = 3, 4, \ldots, n-1$. For $f_{H_5,L_5}(h, \ell)$, for $3h \leq \ell \leq 4h$ and $4h \leq \ell \leq 5h$, two integrals are summed because the piecewise definition of the integrand $f_{H_4,L_4}(h - t_n, \ell - nt_n)$ transitions between the two pieces in the definition of $f_{H_4,L_4}(h, \ell)$ at $\ell - 5t_5 = 3(h - t_5)$, or $t_5 = (\ell - 3h)/2$. Only a single integral appears for $2h \leq \ell \leq 3h$ because in this region for h and ℓ, the transition point $t_5 = (\ell - 3h)/2$ is negative. It is straightforward to show that the piecewise definitions for $f_{H_n,L_n}(h, \ell)$ agree at domain boundaries.

represented by t_n, and the second term nt_n is the contribution to the tree length from the most recent interval, t_n. Rearranging $\ell \geq 2(h - t_n) + nt_n$ yields $t_n \leq \frac{\ell-2h}{n-2}$, and therefore $\beta = \frac{\ell-2h}{n-2}$.

Given h and ℓ, the minimal t_n, α, is either 0 or achieved when t_{n-1} approaches h from below. In the latter case, $t_n = (\ell - \sum_{k=2}^{n-1} kt_k)/n$, so t_n is minimized when $\sum_{k=2}^{n-1} kt_k$ is maximized. The minimal t_n occurs when $t_{n-1} = h - t_n$ and $\sum_{k=2}^{n-1} kt_k = (n-1)t_{n-1}$. Therefore, in the case that minimizes t_n, $\ell \leq (n-1)(h - t_n) + nt_n$. Noting that t_n must be non-negative yields $\alpha = \max\{0, \ell - (n-1)h\}$.

We derive the base case for the recursion, $f_{H_3,L_3}(h, \ell)$, using the fact that H_3 and L_3 are determined by T_2 and T_3, noting that $H_3 = T_2 + T_3$ and $L_3 = 2T_2 + 3T_3$. Given h and ℓ, we have:

$$t_3 = \ell - 2h$$
$$t_2 = 3h - \ell.$$

Because H_3 and L_3 are determined exactly by the two independent quantities T_2 and T_3, $f_{H_3,L_3}(h, \ell) = f_{H_3,L_3}(t_2 + t_3, 2t_2 + 3t_3) = f_{T_2,T_3}(t_2, t_3) = f_{T_3}(t_3) f_{T_2}(t_2)$. Consequently, by Eq. (1),

$$\begin{aligned} f_{H_3,L_3}(h, \ell) &= f_{T_3}(\ell - 2h) f_{T_2}(3h - \ell) \\ &= 3e^{-2\ell+3h}. \end{aligned}$$

For $n \geq 4$, the joint density of (H_n, L_n) can be calculated analytically using the recursive formula in Eqs. (10)–(12). Examples for the analytical formulas for the joint density function of H_n and L_n at $n = 3$, 4, and 5 appear in Table 1, and plots for $n = 3$, 4, 10, and 20 appear in Fig. 2. The figures show the density between $h_n \in [0, 10]$ and $\ell_n \in [0, 24]$, a domain chosen to represent most of the probability density (Appendix A).

The joint probability density for each n lies between the lower bound, $H_n = \frac{1}{n} L_n$, and the upper bound, $H_n = \frac{1}{2} L_n$. For small values of n, such as $n = 3$ and $n = 4$, much of the density lies close to the upper bound (Fig. 2A and B). As can be seen in the density plots for $n = 10$ and $n = 20$, for larger n, the density is more centered between the bounds (Fig. 2C and D).

We can interpret the plots in Fig. 2 of the joint density of H_n and L_n based on the equations in Section 2.1. As n increases, the distribution shifts away from the upper bound, $H_n = \frac{1}{2} L_n$. The waiting times are independent, so T_2 has the same distribution regardless of n. A scenario with large n, however, has more waiting times T_k, for $2 \leq k \leq n$, in addition to T_2. These additional times make a relatively small contribution to H_n, which has a limiting expectation of 2 as $n \to \infty$. However, the contributions of the additional terms kT_k in L_n have a greater impact. Because $\mathbb{E}[H_n]$ has a finite limit (Eq. (3)) and $\mathbb{E}[L_n]$ increases without bound (Eq. (6)), but both have finite variances (Eqs. (4) and (7)), it is increasingly likely that H_n will be relatively small compared to L_n. Thus, as n increases, the joint density shifts away from the upper bound $H_n = \frac{1}{2} L_n$ and moves closer to the lower bound, $H_n = \frac{1}{n} L_n$.

2.3. The ratio $\frac{H_n}{L_n}$

The ratio $\frac{H_n}{L_n}$ can be viewed as a single summary of the joint density of (H_n, L_n). Because the joint density is bounded below by $H_n = \frac{1}{n} L_n$ and above by $H_n = \frac{1}{2} L_n$, the lower and upper bounds on $\frac{H_n}{L_n}$ are $\frac{1}{n}$ and $\frac{1}{2}$, respectively.

Fig. 3A shows $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of n. The expectation is calculated numerically with the joint probability density for (h_n, ℓ_n)

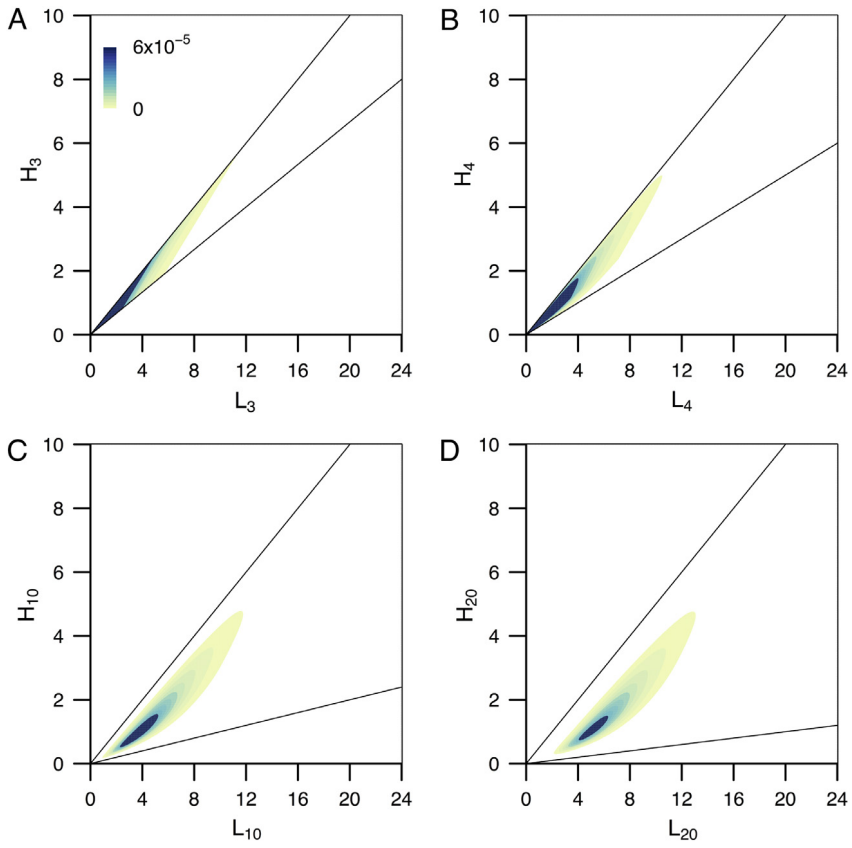

Fig. 2. Joint distribution of H_n and L_n, measured in units of N generations, under a constant-sized coalescent model. The distribution is calculated analytically from Eqs. (10)–(12), in increments of 0.01 for each variable, for h_n in [0, 10] and ℓ_n in [0, 24]. (A) $n = 3$. (B) $n = 4$. (C) $n = 10$. (D) $n = 20$.

obtained from the recursive formula in Eqs. (10)–(12). This expectation, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$, decreases toward the lower bound as n increases, reflecting the shift of the joint density of H_n and L_n away from the upper bound, $H_n = \frac{1}{2}L_n$, for large n, as seen in Fig. 2. $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ decays slowly as n grows, because as n increases, the additional terms in H_n and L_n are increasingly small and thus have lesser effects on the ratio $\frac{H_n}{L_n}$.

We can approximate $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ via the Taylor approximation for the function $f(x, y) = \frac{x}{y}$ around the point $(\mathbb{E}[H_n], \mathbb{E}[L_n])$. We use the second-order Taylor approximation for $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and the first-order Taylor approximation for $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$.

For the mean, applying eq 3.88 of Elandt-Johnson and Johnson (1999) to obtain the expectation of the second-order Taylor expansion of $f(x, y) = \frac{x}{y}$ around $(\mathbb{E}[H_n], \mathbb{E}[L_n])$, we have

$$\mathbb{E}\left[\frac{H_n}{L_n}\right] \approx \frac{\mathbb{E}[H_n]}{\mathbb{E}[L_n]} - \frac{\mathrm{Cov}[H_n, L_n]}{\mathbb{E}[L_n]^2} + \frac{\mathbb{E}[H_n]\,\mathrm{Var}[L_n]}{\mathbb{E}[L_n]^3}. \tag{13}$$

We can write the approximation for $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ using expressions for $\mathbb{E}[H_n]$, $\mathbb{E}[L_n]$, $\mathrm{Var}[L_n]$, and $\mathrm{Cov}[H_n, L_n]$. To find the expression for $\mathrm{Cov}[H_n, L_n]$, we note that because the waiting times between coalescent events are independent, $\mathrm{Cov}[T_k, jT_j] = 0$ when $k \neq j$. We then have

$$\begin{aligned}\mathrm{Cov}[H_n, L_n] &= \mathrm{Cov}\left[\sum_{k=2}^{n} T_k, \sum_{k=2}^{n} kT_k\right] \\ &= \sum_{k=2}^{n} k\mathrm{Var}[T_k] \\ &= \sum_{k=2}^{n} \frac{4}{k(k-1)^2},\end{aligned}$$

where the last step follows from Eq. (1) and the fact that an exponential distribution with rate λ has variance $1/\lambda^2$. The expression for the covariance can be simplified by a partial fraction decomposition. Simplifying the notation by writing $S_{p,n} = \sum_{k=1}^{n} \frac{1}{k^p}$, we obtain

$$\begin{aligned}\mathrm{Cov}[H_n, L_n] &= 4\left[\sum_{k=1}^{n-1} \frac{1}{k^2} - \left(1 - \frac{1}{n}\right)\right] \\ &= 4\left[S_{2,n-1} - \left(1 - \frac{1}{n}\right)\right].\end{aligned} \tag{14}$$

Applying Eqs. (3), (6), (7), and (14) in Eq. (13), we can approximate $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as

$$\begin{aligned}\mathbb{E}\left[\frac{H_n}{L_n}\right] &\approx \frac{2\left(1 - \frac{1}{n}\right)}{\left(2\sum_{k=1}^{n-1} \frac{1}{k}\right)} - \frac{4[\sum_{k=1}^{n-1} \frac{1}{k^2} - (1 - \frac{1}{n})]}{\left(2\sum_{k=1}^{n-1} \frac{1}{k}\right)^2} \\ &\quad + \frac{8\sum_{k=1}^{n-1} \frac{1}{k^2}(1 - \frac{1}{n})}{\left(2\sum_{k=1}^{n-1} \frac{1}{k}\right)^3} \\ &= \frac{1 - \frac{1}{n}}{S_{1,n-1}} - \frac{S_{2,n-1} - (1 - \frac{1}{n})}{S_{1,n-1}^2} + \frac{S_{2,n-1}(1 - \frac{1}{n})}{S_{1,n-1}^3}.\end{aligned} \tag{15}$$

From the form of the approximation for $\mathbb{E}\left[\frac{H_n}{L_n}\right]$, because $S_{2,n}$ converges to $\frac{\pi^2}{6}$ as $n \to \infty$ whereas $S_{1,n}$ diverges, we can see that the

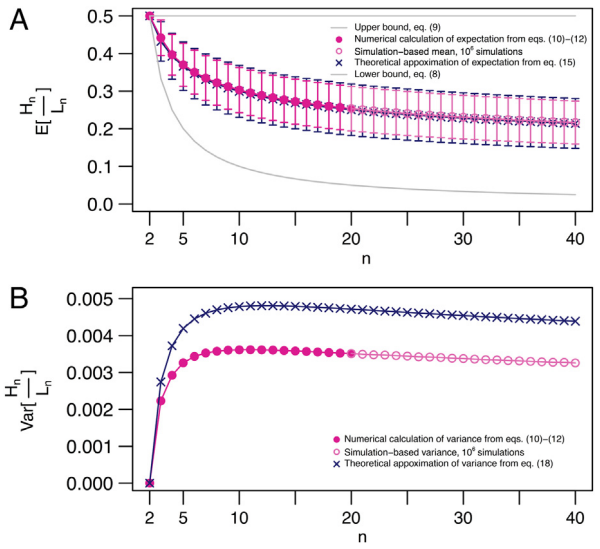

Fig. 3. Mean and variance of $\frac{H_n}{L_n}$ as functions of n under a constant-sized coalescent model. (A) $\mathbb{E}\left[\frac{H_n}{L_n}\right]$. The numerical calculation uses Eqs. (10)–(12), with values of h_n in [0, 10] and values of ℓ_n in [0, 24], both in increments of 0.01 for $n = 2$ to 20. Owing to the computation time required for the analytical joint density of H_n and L_n for large n, for $n = 20$ to 40, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ is calculated from 10^6 ms simulations (the numerical and simulated values are superimposed at $n = 20$). Error bars for $n = 3$ to 20 represent the numerical calculation for the standard deviation, and for $n = 20$ to 40, error bars are taken from 10^6 ms simulations (two sets of error bars are superimposed at $n = 20$). The theoretical approximation is calculated from Eq. (15), and the lower and upper bounds of $\frac{H_n}{L_n}$ are taken from Eqs. (8) and (9), respectively. (B) $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$. The numerical and simulated variances are calculated as in panel A. The theoretical approximation of the variance is taken from Eq. (18). Note that the standard error of the mean of $\frac{H_n}{L_n}$ for 10^6 simulations, which is equal to the simulation-based standard deviation of $\frac{H_n}{L_n}$ divided by the square root of the number of simulations, is for most n quite small, near $\sqrt{0.003}/10^3 \approx 5 \times 10^{-5}$; this small value in relation to the simulation-based mean values of $\frac{H_n}{L_n}$ indicates that 10^6 simulations suffice to provide reasonable estimate of $\mathbb{E}[\frac{H_n}{L_n}]$.

limit of the approximation as the sample size $n \to \infty$ is

$$\lim_{n\to\infty} \frac{1-\frac{1}{n}}{S_{1,n-1}} - \lim_{n\to\infty} \frac{S_{2,n-1}-(1-\frac{1}{n})}{S_{1,n-1}^2} + \lim_{n\to\infty} \frac{S_{2,n-1}(1-\frac{1}{n})}{S_{1,n-1}^3} = 0. \tag{16}$$

For the variance $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$, the first-order Taylor approximation for $f(x, y) = \frac{x}{y}$ around $(\mathbb{E}[H_n], \mathbb{E}[L_n])$ gives rise to Stuart and Ord (1994, eq. 10.17)

$$\mathrm{Var}\left[\frac{H_n}{L_n}\right] \approx \left(\frac{\mathbb{E}[H_n]}{\mathbb{E}[L_n]}\right)^2 \left(\frac{\mathrm{Var}[H_n]}{\mathbb{E}[H_n]^2} - \frac{2\mathrm{Cov}[H_n, L_n]}{\mathbb{E}[H_n]\,\mathbb{E}[L_n]} + \frac{\mathrm{Var}[L_n]}{\mathbb{E}[L_n]^2}\right). \tag{17}$$

Using the expressions for $\mathbb{E}[H_n]$, $\mathbb{E}[L_n]$, $\mathrm{Var}[H_n]$, $\mathrm{Var}[L_n]$, and $\mathrm{Cov}[H_n, L_n]$ from Eqs. (3), (6), (4), (7), and (14) respectively, Eq. (17) gives

$$\begin{aligned}\mathrm{Var}\left[\frac{H_n}{L_n}\right] &\approx \left[\frac{2(1-\frac{1}{n})}{2\sum_{k=1}^{n-1}\frac{1}{k}}\right]^2 \left[\frac{\left(8\sum_{k=2}^{n}\frac{1}{k^2}\right) - 4(1-\frac{1}{n})^2}{4(1-\frac{1}{n})^2} - \frac{8[\sum_{k=1}^{n-1}\frac{1}{k^2} - (1-\frac{1}{n})]}{2(1-\frac{1}{n})\left(2\sum_{k=1}^{n-1}\frac{1}{k}\right)} + \frac{4\sum_{k=1}^{n-1}\frac{1}{k^2}}{\left(2\sum_{k=1}^{n-1}\frac{1}{k}\right)^2}\right] \\ &= \left(\frac{1-\frac{1}{n}}{S_{1,n-1}}\right)^2 \left[\frac{2(S_{2,n}-1) - (1-\frac{1}{n})^2}{(1-\frac{1}{n})^2} - \frac{2\left(S_{2,n-1} - (1-\frac{1}{n})\right)}{(1-\frac{1}{n})S_{1,n-1}} + \frac{S_{2,n-1}}{S_{1,n-1}^2}\right]. \end{aligned} \tag{18}$$

We have seen in Fig. 3A that $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ decreases with increasing n, and in Eq. (16) that the approximation for $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ has limit 0 as $n \to \infty$. From Eq. (18), we see that the approximation of $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ has limit 0 as $n \to \infty$.

Despite the fact that $\frac{H_n}{L_n}$ is bounded below by $\frac{1}{n}$, as n increases, for any $\epsilon > 0$, the probability approaches 0 that $\frac{H_n}{L_n}$ is greater than ϵ (Appendix B). By definition of convergence in probability, $\frac{H_n}{L_n}$ converges in probability to 0. Furthermore, not only do the Taylor approximations for $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ converge to 0, the true values of $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ also converge to 0 and, in fact, all moments of $\frac{H_n}{L_n}$ converge to 0 (Appendix C).

To examine the accuracy of the approximation for the variance in Eq. (18), we calculated $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ numerically using Eqs. (10)–(12). Fig. 3B shows that the theoretical approximation of the variance calculated in Eq. (18) has similar behavior to the variance numerically calculated using Eqs. (10)–(12). Both quantities decrease slowly toward 0. The approximation from Eq. (18) is larger than the variance calculated from Eqs. (10)–(12).

We note that $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ has a maximum that occurs at intermediate values of n. As $n \to \infty$, $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ approaches 0. $\mathrm{Var}\left[\frac{H_2}{L_2}\right]$ is also 0; because $H_2 = \frac{1}{2}L_2$, $\frac{H_2}{L_2}$ is the constant $\frac{1}{2}$ and has no variance. Because $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ is nonnegative and it is 0 both at $n = 2$ and in the limit as n approaches ∞, it must reach a maximum at some $n > 2$. The variance calculated numerically using Eqs. (10)–(12) has its maximum value at $n = 11$, as shown in Fig. 3B. The theoretical approximation for the variance, as calculated by Eq. (18), has its maximum at $n = 12$ (Fig. 3B).

Summarizing this section, the main conclusion is that as $n \to \infty$, the fact that $\mathbb{E}[H_n] \to 2$ whereas $\mathbb{E}[L_n]$ continues to increase leads to the limit $\lim_{n\to\infty}\mathbb{E}\left[\frac{H_n}{L_n}\right] = 0$. This limit is compatible with the observation that the value of $\frac{H_n}{L_n}$ approaches its lower bound as the sample size increases.

2.4. Correlation coefficient

Another summary of the joint distribution of H_n and L_n is the correlation coefficient $\mathrm{Corr}[H_n, L_n]$. To find the correlation coefficient of H_n and L_n, we use Eqs. (14) for the covariance of H_n and L_n, (4) for the variance of H_n, and (7) for the variance L_n:

$$\begin{aligned}\mathrm{Corr}[H_n, L_n] &= \frac{\mathrm{Cov}[H_n, L_n]}{\sqrt{\mathrm{Var}[H_n]\mathrm{Var}[L_n]}} \\ &= \frac{S_{2,n-1} - \left(1-\frac{1}{n}\right)}{\sqrt{S_{2,n-1}\left[2(S_{2,n}-1) - \left(1-\frac{1}{n}\right)^2\right]}}. \end{aligned} \tag{19}$$

As can be seen in Fig. 4, $\mathrm{Corr}[H_n, L_n]$ begins at 1, for $n = 2$, and it decreases as n increases. For $n = 2$, the correlation is 1 because L_2 is linearly determined by H_n. As n increases, additional terms allow a large range of possible H_n values to be associated with a specific L_n. However, $\mathrm{Corr}[H_n, L_n]$ does not decrease to 0 and in fact remains quite large.

We can obtain this result using the limiting value of $\frac{\pi^2}{6}$ for $S_{2,n}$:

$$\lim_{n\to\infty} \mathrm{Corr}[H_n, L_n] = \frac{\pi^2 - 6}{\pi\sqrt{2\pi^2 - 18}} \approx 0.9340. \tag{20}$$

The convergence of the correlation coefficient to a positive constant can be viewed as a consequence of two factors. For any L_n, any achievable H_n is likely to share corresponding largest terms, the

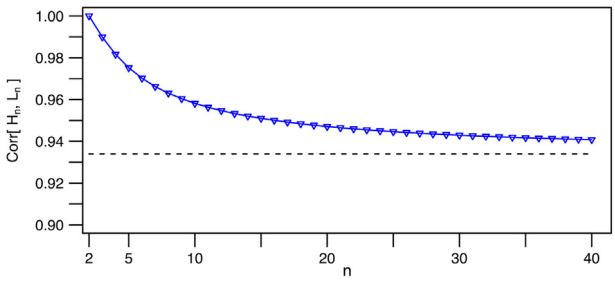

Fig. 4. Correlation coefficient of H_n and L_n as a function of n under a constant-sized coalescent model. The correlation coefficient is calculated analytically from Eq. (19). The asymptote, represented as a dashed line, is shown at $(\pi^2-6)/(\pi\sqrt{2\pi^2-18}) \approx 0.9340$ (Eq. (20)).

coalescence times T_k for small k in H_n, and kT_k for small k in L_n. Thus even as n increases, $\mathrm{Cov}[H_n, L_n]$ remains high. Second, despite the divergence of $\mathbb{E}[L_n]$, owing to the convergence of $\mathrm{Var}[L_n]$, the denominator in $\mathrm{Corr}[H_n, L_n]$ also remains positive and converges, so that the ratio represented in the correlation coefficient converges.

3. Simulations

3.1. Overview

For the constant-population-size model, we have described the exact joint distribution for H_n and L_n using the recursive formula in Eqs. (10)–(12). For more complex models, we use coalescent simulation. We consider a model with exponential growth and a model with two constant-sized populations with a constant migration rate between them.

For each demographic model and each fixed sample size, we simulated 10^6 coalescent trees using `ms` (Hudson, 2002). From these simulated coalescent trees, we measured H_n and L_n and also calculated the mean and variance of $\frac{H_n}{L_n}$ across the simulated trees. In analyzing the results of these simulations, we focus on the relationship of the joint distribution of H_n and L_n to the upper bound, $H_n = \frac{1}{2}L_n$, and the lower bound, $H_n = \frac{1}{n}L_n$, given the different demographic models, model parameters such as growth rate and migration rate, and the sample size, n. We also consider the correlation coefficient of H_n and L_n.

3.2. Exponential growth

First we explore the joint distribution of H_n and L_n under a coalescent model with exponential growth (Slatkin and Hudson, 1991; Donnelly and Tavaré, 1995). We consider a growth parameter, r, such that t units of N generations in the past, the size of a haploid population, currently at size N, was Ne^{-rt}. We simulated 10^6 coalescent genealogies for different values of n and r (including constant-sized cases with $r = 0$).

Fig. 5 shows the joint distribution of H_n and L_n with different values of r at the fixed value of $n = 10$. Note that the simulation with $r = 0$ that appears in Fig. 5A corresponds to the analytical calculation from Eqs. (10)–(12) depicted in Fig. 2C and closely resembles it. As r increases, the joint density of H_{10} and L_{10} is shifted toward smaller values. The joint density of H_{10} and L_{10} also shifts toward the lower bound $H_{10} = \frac{1}{10}L_{10}$.

These observations are also reflected in the behavior of $\frac{H_n}{L_n}$. For fixed values of n, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ decreases as r increases (Fig. 6C), as does $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ (Fig. 6D). Fig. 6D also shows that for increasing r, $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ is smaller with large n than with small n.

At fixed values of r, increasing n leads to a decrease in $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ (Fig. 6A), as was seen in Fig. 3A for constant-sized populations. With increasing values of n for a given value of r, $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ reaches a maximum before decreasing (Fig. 6B). As r increases, $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ reaches its maximum at smaller values of n compared to the case with no growth, where the maximum occurs at $n = 11$. For large $r = 100$, the maximum is at $n = 3$, the smallest n for which $\frac{H_n}{L_n}$ has nonzero variance.

For fixed values of r, $\mathrm{Corr}[H_n, L_n]$ decreases with increasing n (Fig. 7). As was seen in Fig. 4 for $r = 0$ and demonstrated in Section 2.4, $\mathrm{Corr}[H_n, L_n]$ decreases to an asymptote as $n \to \infty$. For large values of r, $\mathrm{Corr}[H_n, L_n]$ decreases more dramatically with increasing n. Fig. 7 suggests that $\mathrm{Corr}[H_n, L_n]$ appears to approach an asymptote at a lower value.

We can explain the difference of the exponential model from the constant-population-size model by considering coalescence times in both models. Under exponential growth, the population size in the present exceeds that in the past. Both H_n and L_n decrease in the exponential growth model compared to the constant-sized model due to the scaling by the population size. In the model considered here, the contemporary population size is N for all r, and decreases further into the past. The external branches – which reflect the larger population size of the recent population – are longer in comparison to the internal branches, which reflect coalescences in the smaller ancestral population. In other words, the coalescent trees generated under a model with exponential growth appear more "star-like": more similar to Fig. 1A than to Fig. 1B (Slatkin and Hudson, 1991; Slatkin, 1996; Sano and Tachida, 2005). Thus, under the growth model, as r increases, T_n becomes comparatively large in relation to H_n; $\frac{H_n}{T_n} \to 1$ and $\frac{L_n}{nT_n} \to 1$, so that $\frac{H_n}{L_n} \to \frac{1}{n}$. Considering Eq. (8), the joint density of H_n and L_n shifts closer to the lower bound, $H_n = \frac{1}{n}L_n$.

3.3. Two populations with symmetric migration

We next study the joint distribution of H_n and L_n under a coalescent model with two equal-sized populations of haploid size N with a constant, symmetric migration rate between them (Nath and Griffiths, 1993; Wakeley, 1998). We denote the rate of migration by m, such that the expected number of migrants per generation is Nm in each direction. We simulated 10^6 coalescent genealogies at each of a series of choices for the total sample size n ($n/2$ per population) and migration rate $2Nm$.

Fig. 8 shows the joint probability density for H_n and L_n in scenarios with four different migration rates. The density is shifted toward larger values of H_n and L_n as $2Nm$ decreases; for the smallest value of $2Nm$ considered, the density is not visible in the region plotted.

Fig. 9A shows the behavior of $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of n for different values of the migration rate $2Nm$, and Fig. 9C shows its behavior as a function of $2Nm$ with different values of n. As $2Nm$ decreases, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ increases toward $\frac{1}{2}$. For large $2Nm$, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ is close to the corresponding values of $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ for a given n in the case with no population structure.

$\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ appears in Fig. 9B as a function of n for different values of $2Nm$, and Fig. 9D shows $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ as a function of $2Nm$ for different n. For a fixed $2Nm$, as seen in the model with no population structure, as n increases, the variance reaches a maximum before decreasing (Fig. 9B). For a fixed n, the variance is small at small values of $2Nm$. It then increases as $2Nm$ increases, before reaching a maximum and then decreasing toward the variance in a model with no population structure (Fig. 9D). Compared to the model with no population structure, $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ reaches its maximum at a larger value of n for small $2Nm$.

$\mathrm{Corr}[H_n, L_n]$ is high for all $2Nm$, and for a fixed n, it decreases as $2Nm$ increases (Fig. 10). For large $2Nm$, $\mathrm{Corr}[H_n, L_n]$

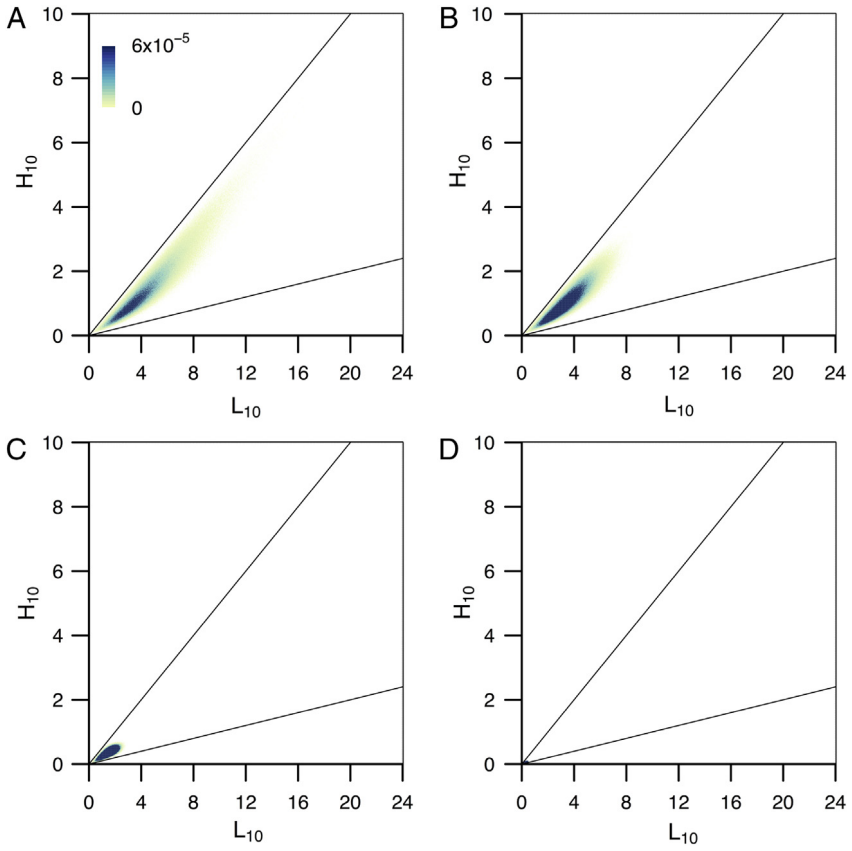

Fig. 5. Joint distribution of H_n and L_n, measured in units of N generations, under an exponential growth model. Each panel is calculated from 10^6 ms simulations for $n = 10$, with exponential growth rate r. N represents the population size at time 0. (A) $r = 0$ (no growth). (B) $r = 1$. (C) $r = 10$. (D) $r = 100$.

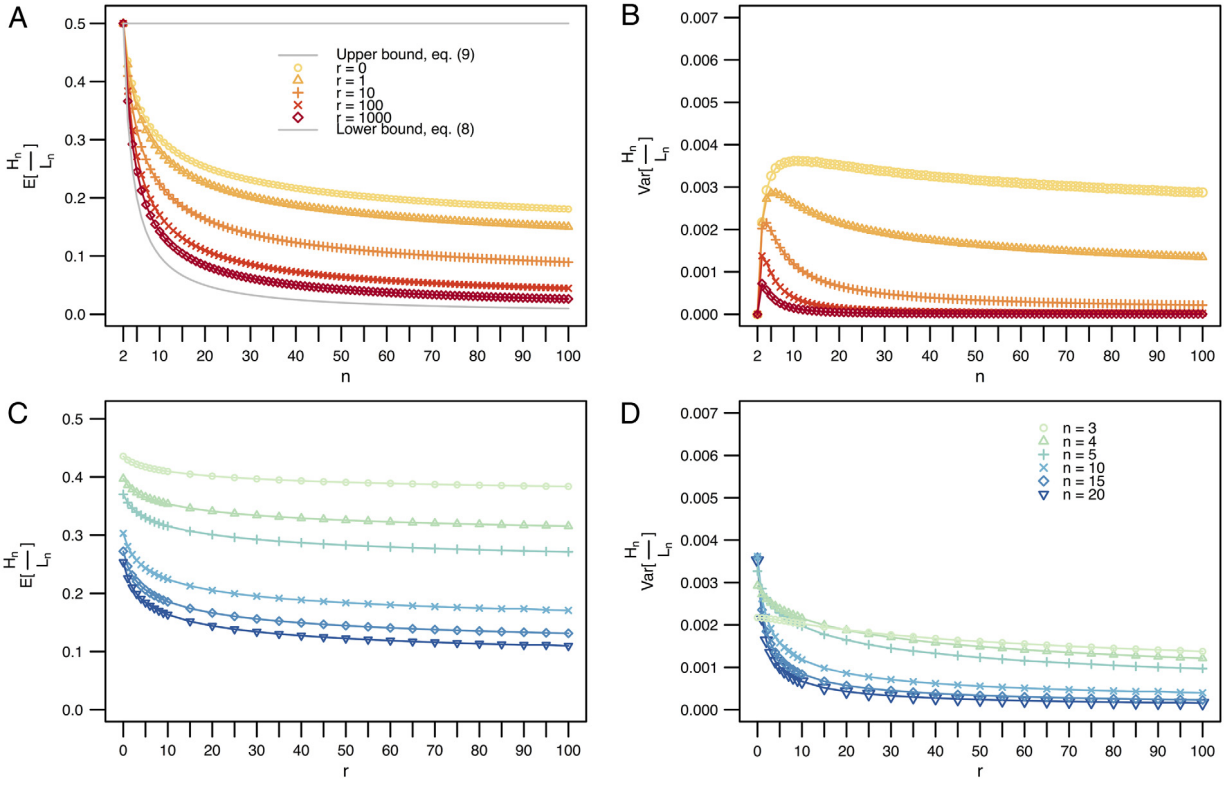

Fig. 6. Mean and variance of $\frac{H_n}{L_n}$ calculated from 10^6 ms simulations, under an exponential growth model. (A) $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of sample size n for fixed growth rate r. (B) $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ as a function of n for fixed r. (C) $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of r for fixed n. (D) $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ as a function of r for fixed n.

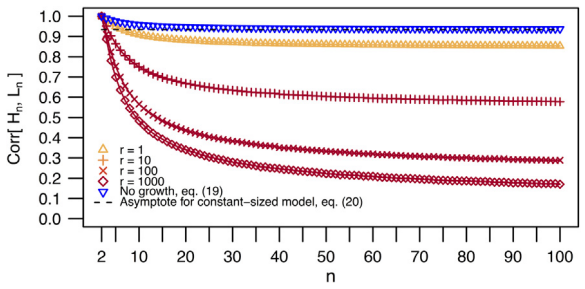

Fig. 7. Correlation coefficient of H_n and L_n as a function of n under an exponential growth model. The correlation coefficient is calculated from 10^6 ms coalescent simulations. For comparison, the correlation coefficient calculated from Eq. (19) in a model with no growth is also shown. The dashed line gives the asymptote for the model with no growth (Eq. (20)).

is close to the value obtained in the model with no population structure.

The patterns in the joint distribution accord with theory for the model with two populations with constant symmetric migration. A coalescence can only occur between two lineages if those lineages are in the same population. For large $2Nm$, the time required for lineages to migrate to the same population is low and does not delay coalescence; H_n approaches the value seen in a model with no population structure (Nath and Griffiths, 1993). As $2Nm$ decreases, both H_n and L_n increase (Fig. 8), reflecting an increase in the waiting time for the final two lineages – one ancestral to the lineages of one population and one ancestral to the lineages of the other – to coalesce. The shape of the genealogies is more likely to resemble Fig. 1B, with $\frac{T_2}{H_n}$ and $\frac{2T_2}{L_n}$ near 1. This result explains why we see that $\mathbb{E}\left[\frac{H_n}{L_n}\right] \to \frac{1}{2}$ as $2Nm$ decreases (Fig. 9A and C). As $2Nm$ decreases, the increased contribution of T_2 to H_n and $2T_2$ to L_n have the consequence that H_n and L_n become increasingly determined by a single random variable, T_2. This phenomenon also explains the decrease in $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ to 0 and the increase in $\mathrm{Corr}[H_n, L_n]$ to 1.

4. Discussion

We have considered the joint distribution of the height, H_n, and length, L_n, of coalescent trees. In a constant-sized population, we studied this distribution, as well as the ratio $\frac{H_n}{L_n}$ and the correlation coefficient $\mathrm{Corr}[H_n, L_n]$. We obtained a recursive formula for the joint probability density of H_n and L_n (Eq. (10)). We found that $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ decreases to 0 as $n \to \infty$, and that $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ is orders of magnitude smaller than $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and also decreases to 0 with increasing n (Section 2.3). Finally, we showed that $\mathrm{Corr}[H_n, L_n]$ is large for all n and decreases to an asymptote, ∼0.9340 (Section 2.4).

We also considered the joint distribution of H_n and L_n under different demographic models. Under our model for exponential growth (Section 3.2), the genealogies become more star-like with increasing r. The joint distribution of H_n and L_n shifts to smaller values of H_n and L_n and closer to the lower bound $H_n = \frac{1}{n}L_n$ compared to the constant-sized model. The plot of $\mathrm{Corr}[H_n, L_n]$ suggests a lower asymptotic value than in the constant model, decreasing with increasing r.

With constant symmetric migration (Section 3.3), the joint distribution of H_n and L_n shifts to larger values of H_n and L_n and closer to the upper bound $H_n = \frac{1}{2}L_n$ than in the constant-sized model with no population structure. In accord with the shift in

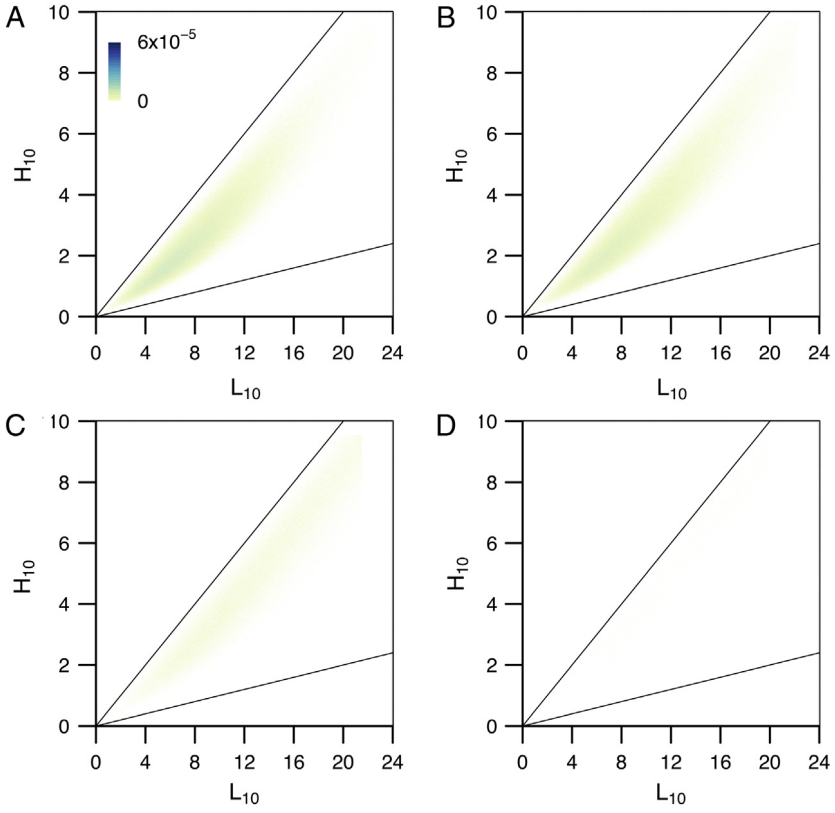

Fig. 8. Joint distribution of H_n and L_n, measured in units of N generations, under an island migration model. Each panel is calculated from 10^6 ms simulations for $n = 10$ (5 lineages per population), with migration rate Nm in each direction. N represents the population size in each of two populations. (A) $2Nm = 10$. (B) $2Nm = 1$. (C) $2Nm = 0.1$. (D) $2Nm = 0.01$.

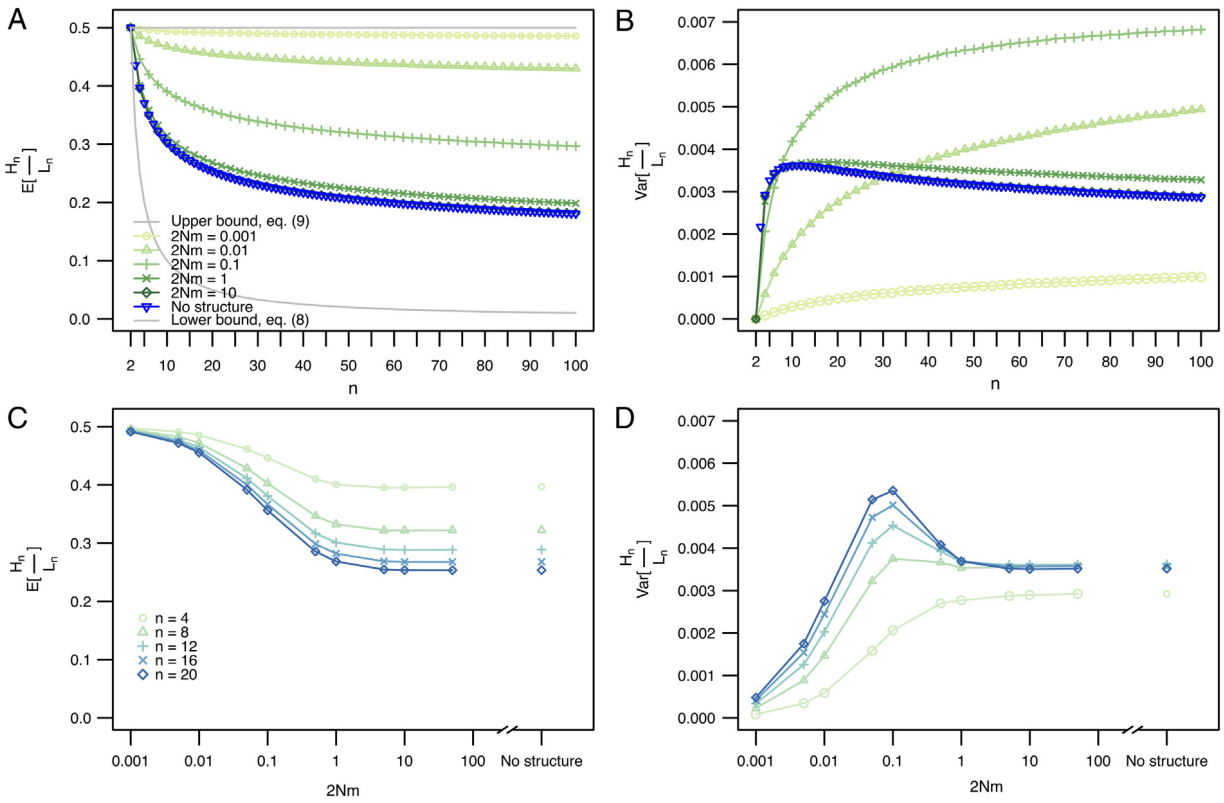

Fig. 9. Mean and variance of $\frac{H_n}{L_n}$ calculated from 10^6 ms simulations, under an island migration model. (A) $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of sample size n for fixed migration rate $2Nm$. (B) $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ as a function of n for fixed $2Nm$. (C) $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ as a function of $2Nm$ for fixed n. (D) $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ as a function of $2Nm$ for fixed n. For comparison, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ and $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ in 10^6 ms simulations of a model with no population structure are also shown. Only even values of n are considered (sample size $n/2$ in each population).

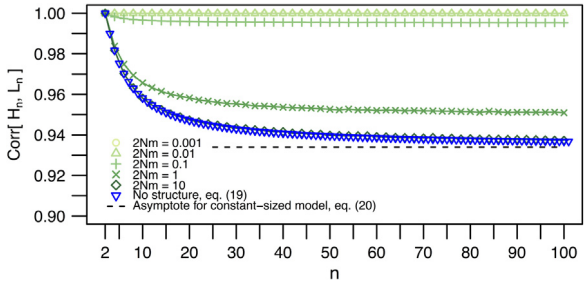

Fig. 10. Correlation coefficient of H_n and L_n as a function of n under an island migration model. The correlation coefficient is calculated from 10^6 ms coalescent simulations. For comparison, the correlation coefficient calculated from Eq. (19) in a model with no population structure is also shown. Only even values of n are considered (sample size $n/2$ in each population). The dashed line gives the asymptote for the model with no structure (Eq. (20)).

the joint distribution toward its upper bound, $\mathbb{E}\left[\frac{H_n}{L_n}\right]$ increases to $\frac{1}{2}$ with decreasing migration rate $2Nm$. H_n approaches T_2 and L_n approaches $2T_2$ as $2Nm$ decreases; consequently, $\mathrm{Corr}[H_n, L_n]$ increases to 1, and $\mathrm{Var}\left[\frac{H_n}{L_n}\right]$ decreases to 0 faster than in the model with no population structure.

Our work contributes to further understanding of tree shapes and joint distributions of tree properties under the coalescent, as represented in several recent studies. Dahmer and Kersting (2015, 2017) explored the distribution and moments of various aspects of internal and external branch lengths in coalescent trees as $n \to \infty$. Ferretti et al. (2017) focused on the relationship of a measure of tree balance for coalescent trees to properties of the site frequency spectrum. Similarly to our work, Ferretti et al. (2017) considered theoretical properties of features of coalescent genealogies at a single locus, examining moments of tree balance measures and identifying extreme tree topologies with respect to the measures of interest. Miroshnikov and Steinrücken (2017) numerically evaluated a cumulative joint distribution for tree height and length as part of an investigation of the marginal distribution of tree length at a single locus in the context of studying the joint distribution of tree length at neighboring loci; their numerical joint distribution is the cumulative distribution function corresponding to the density in Eqs. (10)–(12). As Miroshnikov and Steinrücken (2017) describe, investigating properties of tree length informs coalescent hidden Markov models that rely on accurately modeling coalescent genealogies along the genome.

This study also provides information useful in understanding the properties of methods for estimating T_{MRCA}, or H_n for a sample. Previous work has explored the joint distribution of H_n and S_n, the number of segregating sites (Fu, 1996; Griffiths and Tavaré, 1996), which has expectation proportional to L_n in the infinitely-many-sites model. In estimating H_n from S_n, Fu (1996) studied the dependence of S_n on H_n. Like H_n and L_n, H_n and S_n are positively correlated. The size of the correlation, however, depends on θ, the compound parameter describing the population mutation rate. Fu (1996) noted that as θ increases, the correlation between H_n and S_n also increases. Because segregating sites arise via a random process conditional on L_n rather than H_n, however, the correlation of H_n with S_n is limited by the correlation of H_n with L_n. As we have shown in Section 2.4, $\mathrm{Corr}[H_n, L_n]$ is high in the standard constant-sized coalescent model. Some demographic models, such as constant symmetric migration, also lead to a high correlation (Section 3.3). For a demographic model with extreme growth, however, $\mathrm{Corr}[H_n, L_n]$ is much lower (Section 3.2). Because the demographic scenario can limit the correlation between H_n and L_n, the informativeness of S_n regarding H_n will also be limited in some demographic scenarios.

Our approach of examining joint distributions of tree properties can potentially also be extended to provide information useful in studying more recent methods for estimating H_n. One such method utilizes pairwise coalescence times estimated from numbers of mutations in pairs of lineages, one from one side of the root of a genealogy constructed from data and one from the other side (Tang et al., 2002). This method has recently been expanded from its original single-locus application to incorporate additional data from many unlinked loci (King and Wakeley, 2016). To better understand the behavior of such approaches, it would be of interest to use methods similar to our analysis to examine the joint distribution, ratio, and correlation of H_n and the pairwise coalescence time of randomly chosen samples.

We note that for models more complex than the standard coalescent, we relied on simulation. In the exponential growth model, marginal probability densities for tree height and tree length have previously been studied (Wiuf and Hein, 1999; Polanski et al., 2003). We expect that for the exponential growth model, and potentially for other models as well, further results on the relationship between tree height and length might be possible to obtain analytically.

Acknowledgments

Coalescent modeling and inference were topics of many conversations NAR had with Paul Joyce. We hope this study contributes to further understanding of aspects of coalescent theory that were of interest to Paul. We thank a reviewer for comments and Fabian Freund for pointing us to the work of Drmota et al. (2007). We acknowledge grant support from NIH R01 GM117590, NIH R01 HG005855, and National Science Foundation DBI-1458059.

Appendix A

This appendix proves that the window with $0 \le h_n \le 10$ and $0 \le \ell_n \le 24$ contains most of the probability density of the joint distribution of H_n and L_n for the values of n we consider — more than 0.973, 0.971, 0.964, and 0.960 of the density for $n = 3$, 4, 10, and 20, respectively.

H_n and L_n are nonnegative. We bound from below the probability $\mathbb{P}[H_n < 10 \cap L_n < 24]$.

$$\begin{aligned}\mathbb{P}[H_n < 10 \cap L_n < 24] &= 1 - \mathbb{P}[H_n \ge 10] - \mathbb{P}[L_n \ge 24] \\ &\quad + \mathbb{P}[H_n \ge 10 \cap L_n \ge 24] \\ &\ge 1 - \mathbb{P}[H_n \ge 10] - \mathbb{P}[L_n \ge 24]. \end{aligned} \tag{21}$$

It remains to bound $\mathbb{P}[H_n \ge 10]$ and $\mathbb{P}[L_n \ge 24]$ from above.

Chebyshev's inequality states that $\mathbb{P}[|X - \mathbb{E}[X]| \ge k\sqrt{\mathrm{Var}[X]}] \le \frac{1}{k^2}$. Then

$$\begin{aligned}\frac{1}{k^2} &\ge \mathbb{P}[|H_n - \mathbb{E}[H_n]| \ge k\sqrt{\mathrm{Var}[H_n]}] \\ &\ge \mathbb{P}[H_n - \mathbb{E}[H_n] \ge k\sqrt{\mathrm{Var}[H_n]}] \\ &= \mathbb{P}[H_n \ge k\sqrt{\mathrm{Var}[H_n]} + \mathbb{E}[H_n]]. \end{aligned} \tag{22}$$

To bound $\mathbb{P}[H_n \ge 10]$, we set k as the solution to $k\sqrt{\mathrm{Var}[H_n]} + \mathbb{E}[H_n] = 10$, obtaining $\mathbb{P}[H_n \ge 10] \le \mathrm{Var}[H_n]/(10 - \mathbb{E}[H_n])^2$ from Eq. (22). Analogously, $\mathbb{P}[L_n \ge 24] \le \mathrm{Var}[L_n]/(24 - \mathbb{E}[L_n])^2$.

From Eq. (21), we then conclude

$$\begin{aligned}&\mathbb{P}[H_n < 10 \cap L_n < 24] \\ &\ge 1 - \frac{\mathrm{Var}[H_n]}{(10 - \mathbb{E}[H_n])^2} - \frac{\mathrm{Var}[L_n]}{(24 - \mathbb{E}[L_n])^2}. \end{aligned} \tag{23}$$

Using Eqs. (3), (4), (6), and (7) to evaluate expression (23) for $n = 3$, 4, 10, and 20, we obtain the desired results.

Appendix B

In this appendix, we show that $\frac{H_n}{L_n}$ converges in probability to 0 as $n \to \infty$. By definition of convergence in probability, because $\frac{H_n}{L_n}$ is positive ($H_n = 0$ and $L_n = 0$ both have probability 0), we must prove that, for fixed $\epsilon > 0$, $\lim_{n\to\infty}\mathbb{P}[\frac{H_n}{L_n} > \epsilon] = 0$. We divide the numerator and denominator by the positive quantity $\mu_{L_n} = \mathbb{E}[L_n]$ (Eq. (6)), so that we must show

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{H_n/\mu_{L_n}}{L_n/\mu_{L_n}} > \epsilon\right] = 0. \tag{24}$$

Adapting arguments from Drmota et al. (2007), we prove Eq. (24) in three steps, showing that (i) H_n/μ_{L_n} converges in probability to 0, (ii) L_n/μ_{L_n} converges in probability to 1, and then that (iii) $\frac{H_n}{\mu_{L_n}}/\frac{L_n}{\mu_{L_n}}$ converges in probability to 0.

(i) For a fixed $\epsilon_1 > 0$, we consider

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{H_n}{\mu_{L_n}} - \frac{\mu_{H_n}}{\mu_{L_n}}\right| > \epsilon_1\right] &= \lim_{n\to\infty} \mathbb{P}\left[\left|H_n - \mu_{H_n}\right| > \epsilon_1\mu_{L_n}\right] \\ &\le \lim_{n\to\infty} \frac{\mathrm{Var}[H_n]}{\epsilon_1^2\mu_{L_n}^2}. \end{aligned} \tag{25}$$

The last expression comes from Chebyshev's inequality applied to H_n, which has finite expectation and finite nonzero variance (Eqs. (3) and (4)). Because $\mathrm{Var}[H_n]$ has a finite limit whereas μ_{L_n} diverges, for a fixed ϵ_1, $\lim_{n\to\infty}\mathrm{Var}[H_n]/(\epsilon_1^2\mu_{L_n}^2) = 0$.

Noting that

$$\begin{aligned}\mathbb{P}\left[\frac{H_n}{\mu_{L_n}} > \epsilon_1\right] &= \mathbb{P}\left[\left|\frac{H_n}{\mu_{L_n}}\right| > \epsilon_1\right] \\ &\le \mathbb{P}\left[\left|\frac{H_n}{\mu_{L_n}} - \frac{\mu_{H_n}}{\mu_{L_n}}\right| + \left|\frac{\mu_{H_n}}{\mu_{L_n}}\right| > \epsilon_1\right] \\ &= \mathbb{P}\left[\left|\frac{H_n}{\mu_{L_n}} - \frac{\mu_{H_n}}{\mu_{L_n}}\right| > \epsilon_1 - \frac{\mu_{H_n}}{\mu_{L_n}}\right], \end{aligned}$$

and $\lim_{n\to\infty}\frac{\mu_{H_n}}{\mu_{L_n}} = 0$, we then have $\lim_{n\to\infty}\mathbb{P}[\frac{H_n}{\mu_{L_n}} > \epsilon_1] = 0$, and $\frac{H_n}{\mu_{L_n}}$ converges to 0 in probability.

(ii) For a fixed $\epsilon_2 > 0$,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{L_n}{\mu_{L_n}} - 1\right| > \epsilon_2\right] &= \lim_{n\to\infty} \mathbb{P}\left[\left|L_n - \mu_{L_n}\right| > \epsilon_2\mu_{L_n}\right] \\ &\le \lim_{n\to\infty} \frac{\mathrm{Var}[L_n]}{\epsilon_2^2\mu_{L_n}^2}, \end{aligned} \tag{26}$$

again using Chebyshev's inequality. Because $\mathrm{Var}[L_n]$ has a finite limit (Eq. (7)) whereas μ_{L_n} diverges, for a fixed ϵ_2, $\lim_{n\to\infty}\frac{\mathrm{Var}[L_n]}{\epsilon_2^2\mu_{L_n}^2} = 0$. Therefore $\frac{L_n}{\mu_{L_n}}$ converges to 1 in probability.

(iii) Slutsky's theorem states that if $X_n \to X$ in distribution and $Y_n \to c$ in probability, where c is a nonzero constant, then $X_n/Y_n \to X/c$ in distribution (Serfling, 1980, Theorem 1.5.4). We have shown that $\frac{H_n}{\mu_{L_n}}$ converges to 0 in probability and $\frac{L_n}{\mu_{L_n}}$ converges to 1 in probability. Because convergence in probability implies convergence in distribution (Serfling, 1980, Corollary 1.5.4A), applying Slutsky's theorem, $\frac{H_n/\mu_{L_n}}{L_n/\mu_{L_n}}$ converges to $\frac{0}{1}$ in distribution, and hence $\frac{H_n}{L_n}$ converges in distribution to 0. Furthermore, because convergence in distribution to a constant also implies convergence

in probability (Serfling, 1980, Corollary 1.5.4B), we have shown the stronger statement that $\frac{H_n}{L_n}$ converges in probability to 0.

Appendix C

This appendix shows that $\frac{H_n}{L_n}$ converges in rth mean to 0 for $r > 0$, so that all moments of $\frac{H_n}{L_n}$ have limit 0 as $n \to \infty$, and in particular, both $\mathbb{E}[\frac{H_n}{L_n}]$ and $\mathrm{Var}[\frac{H_n}{L_n}]$ converge to 0. We use the fact that dominated convergence in probability implies convergence in mean (Serfling, 1980, Theorem 1.3.6): that is, if X_n converges in probability to random variable X, $|X_n| \leq |Y|$ with probability 1 for random variable Y and all values of n, and $\mathbb{E}[|Y|^r] < \infty$, then X_n converges in rth mean to X.

We take Y to be the constant $\frac{1}{2}$. Then for all n, $|\frac{H_n}{L_n}| \leq |Y|$ with probability 1, because $\frac{H_n}{L_n}$ is positive and is always bounded above by $\frac{1}{2}$ (Section 2.1). For $r > 0$, $\mathbb{E}[|Y|^r] = \left(\frac{1}{2}\right)^r < \infty$. Because $\frac{H_n}{L_n}$ converges to 0 in probability (Appendix B), Theorem 1.3.6 of Serfling (1980) applies, and $\frac{H_n}{L_n}$ converges to 0 in rth mean.

By definition of convergence in rth mean, $\lim_{n\to\infty}\mathbb{E}[|\frac{H_n}{L_n} - 0|^r] = 0$ for $r > 0$. Therefore $\lim_{n\to\infty}\mathbb{E}[\frac{H_n}{L_n}] = 0$, $\lim_{n\to\infty}\mathbb{E}[(\frac{H_n}{L_n})^2] = 0$, and $\lim_{n\to\infty}\mathrm{Var}[\frac{H_n}{L_n}] = \lim_{n\to\infty}\mathbb{E}[(\frac{H_n}{L_n})^2] - \lim_{n\to\infty}(\mathbb{E}[\frac{H_n}{L_n}])^2 = 0$.

References

Achaz, G., 2009. Frequency spectrum neutrality tests: One for all and all for one. Genetics 183, 249–258.

Dahmer, I., Kersting, G., 2015. The internal branch lengths of the Kingman coalescent. Ann. Appl. Probab. 25, 1325–1348.

Dahmer, I., Kersting, G., 2017. The total external length of the evolving Kingman coalescent. Probab. Theory Related Fields 167, 1165–1214.

Disanto, F., Wiehe, T., 2016. Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. J. Math. Biol. 242, 195–200.

Donnelly, P., Tavaré, S., 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29, 401–421.

Drmota, M., Iksanov, A., Moehle, M., Roseler, U., 2007. Asymptotic results concerning the total branch length of the Bolthausen–Sznitman coalescent. Stochastic Process. Appl. 117, 1404–1421.

Elandt-Johnson, R.C., Johnson, N.L., 1999. Survival Models and Data Analysis. Wiley, New York.

Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., Ramos-Onsins, S., 2017. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics 207, 229–240.

Fu, Y.-X., 1996. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics 144, 829–838.

Fu, Y.-X., Li, W.-H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Griffiths, R., Tavaré, S., 1996. Monte Carlo inference methods in population genetics. Math. Comput. Modelling 39, 141–158.

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

Joyce, P., 1999. No BLUE among phylogenetic estimators. J. Math. Biol. 33, 421–438.

King, L., Wakeley, J., 2016. Empirical Bayes estimation of coalescence times from nucleotide sequence data. Genetics 204, 249–257.

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

Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.-M., Veeramah, K.R., Woerner, A.E., Fomine, F.L.M., Bradman, N., Thomas, M.G., Karafet, T.M., Hammer, M.F., 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am. J. Hum. Genet. 92, 454–459.

Miroshnikov, A., Steinrücken, M., 2017. Computing the joint distribution of the total tree length across loci in populations with variable size. Theor. Popul. Biol. 118, 1–19.

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

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

Rosenberg, N.A., 2006. Gene genealogies. In: Fox, C.W., Wolf, J.B. (Eds.), Evolutionary Genetics: Concepts and Case Studies. Oxford University Press, Oxford, pp. 173–189.

Rosenberg, N.A., Hirsh, A.E., 2003. On the use of star-shaped genealogies in inference of coalescence times. Genetics 164, 1677–1682.

Sano, A., Tachida, H., 2005. Gene genealogy and properties of test statistics of neutrality under population growth. Genetics 169, 1687–1697.

Schierup, M.H., Hein, J., 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156, 879–891.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Slatkin, M., 1996. Gene genealogies within mutant allelic classes. Genetics 143, 579–587.

Slatkin, M., Hudson, R.R., 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562.

Stuart, A., Ord, J.K., 1994. Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory, sixth ed. Wiley, Chichester.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tang, H., Siegmund, D.O., Shen, P., Oefner, P.J., Feldman, M.W., 2002. Frequentist estimation of coalescent times from nucleotide sequence data using a tree-based partition. Genetics 161, 447–459.

Tavaré, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring coalescence times from DNA sequence data. Genetics 145, 505–518.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., Feldman, M.W., 2000. Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc. Natl. Acad. Sci. 97, 7360–7365.

Uyenoyama, M.K., 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 147, 1389–1400.

Wakeley, J., 1998. Segregating sites in Wright's Island model. Theor. Popul. Biol. 53, 166–174.

Wakeley, J., 2009. Coalescent Theory: An Introduction. Roberts & Company, Greenwood Village, CO.

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259.