G3, 2022, **12(10)**, jkac205
https://doi.org/10.1093/g3journal/jkac205
Advance Access Publication Date: 11 August 2022
Investigation

Approximations to the expectations and variances of ratios of tree properties under the coalescent

Egor Lappo ,* Noah A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, USA

*Corresponding author: Department of Biology, Stanford University, Stanford, CA 94305, USA. Email: alimpiev@stanford.edu

Abstract

Properties of gene genealogies such as tree height (*H*), total branch length (*L*), total lengths of external (*E*) and internal (*I*) branches, mean length of basal branches (*B*), and the underlying coalescence times (*T*) can be used to study population-genetic processes and to develop statistical tests of population-genetic models. Uses of tree features in statistical tests often rely on predictions that depend on pairwise relationships among such features. For genealogies under the coalescent, we provide exact expressions for Taylor approximations to expected values and variances of ratios X_n/Y_n, for all 15 pairs among the variables $\{H_n, L_n, E_n, I_n, B_n, T_k\}$, considering *n* leaves and $2 \leq k \leq n$. For expected values of the ratios, the approximations match closely with empirical simulation-based values. The approximations to the variances are not as accurate, but they generally match simulations in their trends as *n* increases. Although E_n has expectation 2 and H_n has expectation 2 in the limit as $n \to \infty$, the approximation to the limiting expectation for E_n/H_n is not 1, instead equaling $\pi^2/3 - 2 \approx 1.28987$. The new approximations augment fundamental results in coalescent theory on the shapes of genealogical trees.

Keywords: coalescent theory; external branches; internal branches; time to the most recent common ancestor

Introduction

Coalescent theory models random genealogies conditional on assumptions about the evolutionary process (Hein *et al.* 2005; Wakeley 2009). In coalescent theory, a gene genealogy is a tree or network structure that represents a random draw from a coalescent model.

Genealogies in coalescent theory can be summarized using a variety of quantities. For example, for random tree-like genealogies with n lineages, the tree height H_n records the sum of branch lengths on a path from a leaf to the root, and the tree length L_n sums all branch lengths in the tree. The total length E_n of external branches sums over leaves the lengths of paths from leaves to their nearest internal nodes, and the total length of internal branches, $I_n = L_n - E_n$, sums the lengths of all remaining branches.

Studies in coalescent theory have often investigated the properties of tree summaries conditional on assumptions of coalescent models, with the goal of understanding how shapes of the genealogies relate to processes such as population growth and migration (e.g. Slatkin 1996; Rosenberg and Feldman 2002). Because mutations can be viewed as occurring conditionally on underlying genealogies (Hudson 1990), features of genealogical shape affect the patterns of genetic variation produced by coalescent models that permit mutation. Thus, the understanding of summaries of tree shape predicted by coalescent models is a component of the interpretation of patterns of genetic variation in relation to evolutionary processes.

Initial results concerning summaries of genealogical shape focused on single quantities, producing results on quantities such as H_n and L_n (Kingman 1982; Hudson 1983, 1990; Tajima 1983). Studies soon examined the information that resides in the relationships between pairs of summaries; genetic variation statistics such as those of Tajima (1989) and Fu and Li (1993) can be viewed as assessing whether or not one aspect of a tree contains long branches in relation to another.

Recently, Arbisser *et al.* (2018) performed a detailed investigation of the relationship between H_n and L_n under coalescent models. They studied the mathematical relationship between these two quantities, computing under a standard coalescent model with a constant-sized population the covariance and correlation coefficient of H_n and L_n. Extending the work of Arbisser *et al.* (2018) on H_n and L_n, we (Alimpiev and Rosenberg 2022) reported covariances and correlations for all pairs of variables among $\{H_n, L_n, E_n, I_n, B_n, T_k\}$, where B_n is the mean of the lengths of the two basal branches of a genealogy and T_k is the coalescence time from k to k—1 lineages, $2 \leq k \leq n$. Our compendium in Tables 1 and 2 of Alimpiev and Rosenberg (2022) summarizes pairwise relationships for several of the most commonly used features of coalescent tree shape, recording both new and previously known results.

In addition to computing the covariance and correlation coefficient of H_n and L_n, Arbisser *et al.* (2018) also found approximations to the expectation and variance of the ratio H_n/L_n under the coalescent model. This ratio gives a summary of the joint distribution of H_n and L_n that characterizes the relative magnitudes of the variables—a feature not captured by their covariance or correlation. Arbisser *et al.* (2018) found that although the approximation to $\mathrm{Var}[H_n/L_n]$ differed noticeably from the exact value,

Received: March 30, 2022. **Accepted:** August 1, 2022

as obtained by numerical integration and simulations of the coalescent model, the approximation to $\mathbb{E}[H_n/L_n]$ was quite accurate.

In this article, we extend the work of Arbisser et al. (2018) to compute approximations to the expectations and variances for ratios of the 14 remaining pairs among $\{H_n, L_n, E_n, I_n, B_n, T_k\}$. The study performs for the expectation and variance of coalescent ratios an analogous extension of Arbisser et al. (2018) to that performed by Alimpiev and Rosenberg (2022) for the covariance and correlation coefficient.

Materials and methods

Tree variables

We work with a haploid population of constant size N that follows a standard coalescent model. Time is measured in units of N generations. In this section, we recall the definitions of the coalescence time T_k and tree properties H_n, L_n, E_n, I_n, and B_n for sample size $n \geq 2$ and $2 \leq k \leq n$.

T_k is defined to be a random variable representing the time to coalescence of k to k—1 lineages, for $2 \leq k \leq n$. Variable T_k has exponential probability density function

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

The expectation and variance of T_k are

$$\mathbb{E}[T_k] = \frac{2}{k(k-1)}, \tag{1}$$

$$\mathrm{Var}[T_k] = \frac{4}{k^2(k-1)^2}. \tag{2}$$

Table 1. Definitions of random variables associated with various tree summaries.

Variable	Definition
H_n	$\sum_{k=2}^{n} T_k$
L_n	$\sum_{k=2}^{n} kT_k$
E_n	$\sum_{i=1}^{n} e_i^{(n)}$
I_n	$L_n - E_n$
B_n	$\frac{1}{2}T_2 + \left[\sum_{j=3}^{n-1} \sum_{k=2}^{j} \frac{1}{j(j-1)} T_k\right] + \left(\sum_{k=2}^{n} \frac{1}{n-1} T_k\right)$

Here, T_k is the random variable representing the coalescence time from k to k—1 lineages, and $e_i^{(n)}$ is the (random) length of the ith external branch of a tree with n leaves. We define H_n, L_n, and E_n for $n \geq 2$, I_n for $n \geq 3$, and B_n for $n \geq 4$. The expression for B_n follows a form that incorporates terms associated with all of its contributing branches, following p. 1400 of Uyenoyama (1997) and Section 2.6 of Alimpiev and Rosenberg (2022), and it can be simplified to $B_n = \sum_{k=2}^{n} \frac{1}{k-1} T_k$.

The tree properties H_n, L_n, E_n, I_n, and B_n are defined in terms of the T_k. Visual depictions of these properties appear in Fig. 1, and mathematical definitions of these quantities appear in Table 1.

We define $S_{p,n} = \sum_{k=1}^{n} k^{-p}$ as a useful shorthand. The limit $\lim_{n\to\infty} S_{p,n} = S_{p,\infty}$ is the Riemann zeta function, usually denoted $\zeta(p)$. In particular, $S_{1,\infty}$ diverges, $S_{2,\infty} = \pi^2/6 \approx 1.64493$, and $S_{3,\infty}$ is Apéry's constant, approximately 1.20206.

Taylor approximations to expectations and variances of ratios

To compute approximate expressions for expected values and variances of the ratios of various tree properties, we rely on Taylor approximations. In particular, consider random variables X and Y with $\mathbb{E}[X], \mathbb{E}[Y] \neq 0$. For the expectation, we have (second-order) approximation (Elandt-Johnson and Johnson 1999, eq. 3.88):

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

For the variance, we have (first-order) approximation (Stuart and Ord 1994, eq. 10.17):

$$\mathrm{Var}\left[\frac{X}{Y}\right] \approx \left(\frac{\mathbb{E}[X]}{\mathbb{E}[Y]}\right)^2 \left(\frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} - \frac{2\mathrm{Cov}[X,Y]}{\mathbb{E}[X]\mathbb{E}[Y]} + \frac{\mathrm{Var}[Y]}{\mathbb{E}[Y]^2}\right). \tag{4}$$

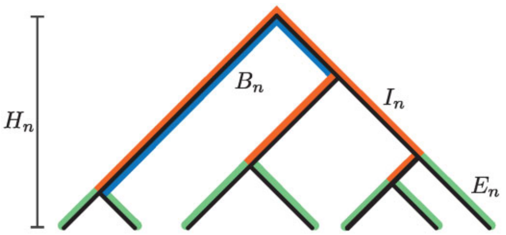

Fig. 1. Properties of genealogical trees. The tree height is H_n. The sum of the lengths of all branches is L_n. External branches have total length E_n (green). Internal branches have total length I_n (orange). Basal branches have mean length B_n (blue).

Table 2. Expectations and variances of properties of tree branch lengths.

X_n	$\mathbb{E}[X_n]$	$\lim_{n\to\infty} \mathbb{E}[X_n]$	$\mathrm{Var}[X_n]$	$\lim_{n\to\infty} \mathrm{Var}[X_n]$
H_n	$\frac{2(n-1)}{n}$	2	$8(S_{2,n}-1) - 4\left(\frac{n-1}{n}\right)^2$	$\frac{4\pi^2}{3} - 12 \approx 1.15947$
L_n	$2S_{1,n-1}$	∞	$4S_{2,n-1}$	$\frac{2\pi^2}{3} \approx 6.57974$
E_n	2	2	$\begin{cases} 4, & n=2, \\ \frac{8}{(n-1)(n-2)}[S_{1,n-1}n - 2(n-1)], & n>2. \end{cases}$	0
I_n	$2S_{1,n-1} - 2$	∞	$4\left[\frac{2[S_{1,n-1}n-2(n-1)]}{(n-1)(n-2)} - \frac{2S_{1,n-1}}{n-1} + S_{2,n-1}\right]$	$\frac{2\pi^2}{3} \approx 6.57974$
B_n	$2S_{2,n-1} - 2 + \frac{2}{n}$	$\frac{\pi^2}{3} - 2 \approx 1.28987$	$\frac{2(3S_{2,n-1}n^2 - 2S_{2,n-1}^2 n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}{n^2}$	$-\frac{\pi^4}{9} + \pi^2 + 2 \approx 1.04637$
T_k	$\frac{2}{k(k-1)}$	$\frac{2}{k(k-1)}$	$\frac{4}{k^2(k-1)^2}$	$\frac{4}{k^2(k-1)^2}$

These expressions can be found in Alimpiev and Rosenberg (2022). Note that for L_n and I_n, although the limiting variance is finite, the expectation is infinite (Tavaré et al. 1997; Wakeley 2009, p. 76).

Table 3. Covariances of pairs of variables that summarize genealogical trees.

(X_n, Y_n)	$\mathrm{Cov}[X_n, Y_n]$	$\lim_{n\to\infty} \mathrm{Cov}[X_n, Y_n]$
H_n, T_k	$\frac{4}{k^2(k-1)^2}$	$\frac{4}{k^2(k-1)^2}$
H_n, L_n	$4S_{2,n-1} - 4 + \frac{4}{n}$	$\frac{2\pi^2}{3} - 4 \approx 2.57974$
H_n, E_n	$\frac{4}{n}$	0
H_n, I_n	$4S_{2,n-1} - 4$	$\frac{2\pi^2}{3} - 4 \approx 2.57974$
H_n, B_n	$\frac{4[S_{3,n-1}n^2 - 3S_{2,n-1}n^2 + (4n+1)(n-1)]}{n^2}$	$-2\pi^2 + 4\zeta(3) + 16 \approx 1.06902$
L_n, T_k	$\frac{4}{k(k-1)^2}$	$\frac{4}{k(k-1)^2}$
L_n, E_n	$\frac{4S_{1,n-1}}{n-1}$	0
L_n, I_n	$4S_{2,n-1} - \frac{4S_{1,n-1}}{n-1}$	$\frac{2\pi^2}{3} \approx 6.57974$
L_n, B_n	$\frac{4[S_{3,n-1}n - S_{2,n-1}n + n - 1]}{n}$	$-\frac{2\pi^2}{3} + 4\zeta(3) + 4 \approx 2.22849$
E_n, T_k	$\frac{4}{k(k-1)(n-1)}$	0
E_n, I_n	$\frac{4S_{1,n-1}}{n-1} - \frac{8S_{1,n-1}n}{(n-1)(n-2)} + \frac{16}{n-2}$	0
E_n, B_n	$\frac{4(S_{2,n-1}n - n + 1)}{n(n-1)}$	0
I_n, T_k	$\frac{4(n-k)}{k(k-1)^2(n-1)}$	$\frac{4}{k(k-1)^2}$
I_n, B_n	$\frac{4(S_{3,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)}{n-1}$	$-\frac{2\pi^2}{3} + 4\zeta(3) + 4 \approx 2.22849$
B_n, T_k	$\frac{4}{k^2(k-1)^3}$	$\frac{4}{k^2(k-1)^3}$

For pairs involving E_n or I_n, expressions apply for $n \geq 3$; expressions involving B_n apply for $n \geq 4$. The expressions can be found in Alimpiev and Rosenberg (2022).

We use $\widetilde{\mathbb{E}}[X/Y]$ and $\widetilde{\mathrm{Var}}[X/Y]$ to denote approximations from equations (3) and (4). For both the expectation and the variance, we also take the $n \to \infty$ limit of the approximations.

Exact expectations, variances, and covariances of tree properties

Expected values and variances of variables H_n, L_n, E_n, I_n, B_n, and T_k that are used in equations (3) and (4) are known, in many cases, from early studies in coalescent theory (Fu and Li 1993; Tavaré *et al.* 1997; Wakeley 2009). We summarize these expectations and variances in Table 2.

The covariances compiled by Alimpiev and Rosenberg (2022) appear in Table 3. In the case of pairs (E_n, B_n) and (I_n, B_n), the covariances are approximate, as described by Alimpiev and Rosenberg (2022).

Evaluating the approximations

For each of 15 pairs of random variables, considering H_n, L_n, E_n, I_n, and B_n as well as T_k, we substitute expressions from Tables 2 and 3 into equations (3) and (4) to obtain approximate expectations and variances for ratios of pairs of variables. For each pair, we choose one variable for the numerator and the other for the denominator; approximate expectations and variances for the reciprocals can be obtained similarly. We present the approximations in Tables 4 and 5, and we plot them in Figs. 2–5.

For pairs (X_n, Y_n), we simulate the values of $\mathbb{E}[X_n/Y_n]$ and $\mathrm{Var}[X_n/Y_n]$ under the coalescent model using ms (Hudson 2002), performing 100,000 replicate simulations for each tree size

Table 4. Approximations to expectations of ratios of pairs of variables.

(X_n, Y_n)	$\widetilde{\mathbb{E}}[X_n/Y_n]$	$\lim_{n\to\infty} \widetilde{\mathbb{E}}[X_n/Y_n]$
H_n, T_k	$\frac{(2k^2-2k-1)n - 2k(k-1)}{n}$	$2k^2 - 2k - 1$
H_n, L_n	$\frac{n-1}{S_{1,n-1}n} - \frac{S_{2,n-1}n - n + 1}{S_{1,n-1}^2 n} + \frac{S_{2,n-1}(n-1)}{S_{1,n-1}^3 n}$	0
E_n, H_n	$\frac{n(2S_{2,n}n^2 - 2n^2 - n + 1)}{(n-1)^3}$	$\frac{\pi^2}{3} - 2 \approx 1.28987$
H_n, I_n	$\frac{n-1}{(S_{1,n-1}-1)n} - \frac{S_{2,n-1}-1}{(S_{1,n-1}-1)^2} + \frac{S_{2,n-1}(n-1)(n-2) - 4n + 4S_{1,n-1} + 4}{(S_{1,n-1}-1)^3 n(n-2)}$	0
B_n, H_n	$\frac{S_{2,n-1}n - n + 1}{n-1} + \frac{3S_{2,n-1}n^2 - S_{3,n-1}n^2 - 4n^2 + 3n + 1}{(n-1)^2} + \frac{(S_{2,n-1}n - n + 1)(2S_{2,n}n^2 - 3n^2 + 2n - 1)}{(n-1)^3}$	$\frac{\pi^4}{18} - \frac{\pi^2}{6} - \zeta(3) - 2 \approx 0.56463$
L_n, T_k	$2S_{1,n-1}k^2 - (2S_{1,n-1} + 1)k$	∞
E_n, L_n	$\frac{(S_{1,n-1}^2 + S_{2,n-1})n - 2S_{1,n-1}^2 - S_{2,n-1}}{S_{1,n-1}^3(n-1)}$	0
L_n, I_n	$\frac{(S_{1,n-1}^3 + S_{2,n-1})(n-1)(n-2) - S_{1,n-1}^2(2n^2 - 7n + 2) + S_{1,n-1}(n^2 - 8n + 8)}{(S_{1,n-1}-1)^3(n-1)(n-2)}$	1
B_n, L_n	$\frac{S_{2,n-1}n - n + 1}{S_{1,n-1}n} + \frac{S_{2,n-1}n - S_{3,n-1}n - n + 1}{S_{1,n-1}^2 n} + \frac{S_{2,n-1}(S_{2,n-1}n - n + 1)}{S_{1,n-1}^3 n}$	0
E_n, T_k	$\frac{k(k-1)(2n-3)}{n-1}$	$2k(k-1)$
E_n, I_n	$\frac{S_{1,n-1}^2(n^2 - 2n + 4) - S_{1,n-1}(2n^2 - n - 2) + (S_{2,n-1} + 1)(n-1)(n-2)}{(S_{1,n-1}-1)^3(n-1)(n-2)}$	0
B_n, E_n	$\frac{(n^2 + 2S_{1,n-1}n - 8n + 8)(S_{2,n-1}n - n + 1)}{n(n-1)(n-2)}$	$\frac{\pi^2}{6} - 1 \approx 0.64493$
I_n, T_k	$2k(k-1)(S_{1,n-1} - 1) - \frac{k(n-k)}{n-1}$	∞
B_n, I_n	$\frac{S_{2,n-1}n - n + 1}{(S_{1,n-1}-1)n} + \frac{(S_{2,n-1}n - n + 1)[S_{2,n-1}(n-1)(n-2) - 4n + 4S_{1,n-1} + 4]}{(S_{1,n-1}-1)^3 n(n-1)(n-2)} + \frac{S_{2,n-1}n - (S_{3,n-1} + 1)(n-1)}{(S_{1,n-1}-1)^2(n-1)}$	0
B_n, T_k	$\frac{2k(k-1)(S_{2,n-1}n - n + 1)}{n} - \frac{1}{k-1}$	$\frac{1}{3}(\pi^2 - 6)k(k-1) - \frac{1}{k-1}$

Expressions involving E_n or I_n apply for $n \geq 3$; expressions involving B_n apply for $n \geq 4$. The value for (H_n, L_n) follows equation 15 of Arbisser *et al.* (2018). The expressions are obtained using equation 3 and Tables 2 and 3.

Table 5. Approximations to variances of ratios of pairs of variables.

(X_n, Y_n)	$\widetilde{\mathrm{Var}}[X_n/Y_n]$	$\lim_{n\to\infty} \widetilde{\mathrm{Var}}[X_n/Y_n]$
H_n, T_k	$\frac{2k(k-1)[k(k-1)S_{2,n}n-(k^2-k+1)n+1]}{n}$	$\frac{1}{3}k(k-1)[(\pi^2-6)k^2-(\pi^2-6)k-6]$
H_n, L_n	$\left(\frac{n-1}{S_{1,n-1}n}\right)^2\left[\frac{2(S_{2,n}-1)n^2-(n-1)^2}{(n-1)^2}-\frac{2[S_{2,n-1}n-(n-1)]}{S_{1,n-1}(n-1)}+\frac{S_{2,n-1}}{S_{1,n-1}^2}\right]$	0
E_n, H_n	$\frac{[2S_{1,n-1}n(n-1)+2S_{2,n}n^2(n-2)-(n^2-3)(3n-2)]n^2}{(n-1)^4(n-2)}$	$\frac{\pi^2}{3}-3\approx 0.28987$
H_n, I_n	$\frac{2S_{2,n}n^2-3n^2+2n-1}{(S_{1,n-1}-1)^2n^2}+\frac{1}{(S_{1,n-1}-1)^4}\left[\frac{[[S_{2,n-1}(n-2)-4](n-1)+4S_{1,n-1}](n-1)}{n^2(n-2)}-\frac{2(S_{1,n-1}-1)(S_{2,n-1}-1)(n-1)}{n}\right]$	0
B_n, H_n	$\frac{(4S_{3,n-1}n^2+4S_{2,n}n^2+11n^2-5n-10)(n-1)^2-S_{2,n-1}(4S_{3,n-1}n^2+8S_{2,n}n^2+13n^2-9n-12)n(n-1)+4S_{2,n-1}^2(S_{2,n}n^2+n^2-n-1)n^2}{2(n-1)^4}$	$\frac{\pi^6}{108}-\frac{\pi^4}{18}-\frac{3\pi^2}{4}+\frac{11}{2}+2\zeta(3)$ $-\frac{\pi^2\zeta(3)}{3}\approx 0.03744$
L_n, T_k	$k^2(k-1)^2S_{1,n-1}^2\left[\frac{S_{2,n-1}}{S_{1,n-1}^2}-\frac{2}{(k-1)S_{1,n-1}}+1\right]$	∞
E_n, L_n	$\frac{2S_{1,n-1}^3n-S_{1,n-1}^2(6n-8)+S_{2,n-1}(n-1)(n-2)}{S_{1,n-1}^4(n-1)(n-2)}$	0
L_n, I_n	$\frac{2S_{1,n-1}^3n-S_{1,n-1}^2(6n-8)+S_{2,n-1}(n-1)(n-2)}{(S_{1,n-1}-1)^4(n-1)(n-2)}$	0
B_n, L_n	$\frac{S_{1,n-1}^2[-2S_{2,n-1}^2n^2+S_{2,n-1}(3n-4)n+n^2+3n-4]+4S_{1,n-1}(S_{2,n-1}n-n+1)(S_{2,n-1}n-S_{3,n-1}n-n+1)+2S_{2,n-1}(S_{2,n-1}n-n+1)^2}{2S_{1,n-1}^4n^2}$	0
E_n, T_k	$\frac{k^2(k-1)^2(n^2+2S_{1,n-1}n-9n+10)}{(n-1)(n-2)}$	$k^2(k-1)^2$
E_n, I_n	$\frac{2S_{1,n-1}^3n-S_{1,n-1}^2(6n-8)+S_{2,n-1}(n-1)(n-2)}{(S_{1,n-1}-1)^4(n-1)(n-2)}$	0
B_n, E_n	$\frac{4S_{1,n-1}n(S_{2,n-1}n-n+1)^2-2S_{2,n-1}^2(n^2+3n-6)n^2+S_{2,n-1}(3n-4)(n+6)n(n-1)+(n^2-10n+8)(n-1)^2}{2n^2(n-1)(n-2)}$	$-\frac{\pi^4}{30}+\frac{\pi^2}{4}+\frac{1}{2}\approx 0.26159$
I_n, T_k	$\frac{k^2(k-1)[(k-1)S_{1,n-1}^2(n-1)(n-2)-2S_{1,n-1}(kn^2-4kn+n+2k)+(k-1)S_{2,n-1}(n-1)(n-2)+kn^2+n^2-9kn+3n+10k-6]}{(n-1)(n-2)}$	∞
B_n, I_n	$\frac{[S_{2,n-1}(n-1)(n-2)-4n+4S_{1,n-1}+4](S_{2,n-1}n-n+1)^2}{(S_{1,n-1}-1)^4n^2(n-1)(n-2)}+\frac{2[S_{2,n-1}n-(S_{3,n-1}+1)(n-1)](S_{2,n-1}n-n+1)}{(S_{1,n-1}-1)^3n(n-1)}-\frac{2S_{2,n-1}^2n^2-S_{2,n-1}(3n-4)n-(n+4)(n-1)}{2(S_{1,n-1}-1)^2n^2}$	0
B_n, T_k	$\frac{k}{2}\left[\frac{[k(k-1)^2(3n+2)+4n](n-1)}{n^2}-(k+1)(k^2-3k+4)S_{2,n-1}\right]$	$\frac{1}{12}k[(18-\pi^2)k^3-2(18-\pi^2)k^2$ $+(18-\pi^2)k-4(\pi^2-6)]$

Expressions involving E_n or I_n apply for $n \geq 3$; expressions involving B_n apply for $n \geq 4$. The value for (H_n, L_n) follows equation 18 of Arbisser *et al.* (2018). The expressions are obtained using equation 4 and Tables 2 and 3.

$n = 2, 3, \ldots, 50$. We plot the simulated values alongside the approximate values from Tables 4 and 5 in Figs. 2 and 4.

Results

Expectations of the ratios

The approximate expected values in Table 4, as approximations of ratios, have the form of rational functions. As n grows, the approximate expectations of H_n/L_n, H_n/I_n, E_n/L_n, E_n/I_n, B_n/L_n, and B_n/I_n approach 0. This behavior is sensible when considering the properties of the coalescent model: in the numerators, E_n has expectation 2 and $\mathbb{E}[H_n]$ and $\mathbb{E}[B_n]$ have bounded expectation in the limit as $n \to \infty$; in the denominators, L_n and I_n have expectations that grow without bound (Table 2). Similarly, approximate expectations of ratios L_n/T_k or I_n/T_k with L_n and I_n in the numerator and T_k in the denominator grow to infinity as n increases. The approximation to $\mathbb{E}[L_n/I_n]$ approaches 1 in the limit as $n \to \infty$: as the number of leaves in the tree grows, internal branches occupy an increasingly large fraction of the total branch length.

For pairs of variables that both have finite expectation, the approximate expectations of their associated ratios—H_n/T_k, E_n/H_n, E_n/T_k, B_n/H_n, B_n/E_n, and B_n/T_k—also approach finite values in the limit as $n \to \infty$. It is interesting to observe that although $\lim_{n\to\infty}\mathbb{E}[E_n] = \lim_{n\to\infty}\mathbb{E}[H_n] = 2$ (Table 2), $\widetilde{\mathbb{E}}[E_n/H_n] = \pi^2/3 - 2 \approx 1.28987 \neq 1$. In other words, although expectations of the individual variables approach the same value, we expect E_n/H_n to be somewhat larger than 1 on average.

For each of the 10 pairs of variables among $\{H_n, L_n, E_n, I_n, B_n\}$, the approximate expectations from Table 4 are plotted in Fig. 2 together with the simulated values. Although some divergences are present for small n, the approximate and simulated values match closely.

The approximate ratios involving T_k are shown in Fig. 3 as functions of k for each of three values of n. L_n is the fastest-growing variable according to the expression for its expectation (Table 2), and the graph for $\widetilde{\mathbb{E}}[L_n/T_k]$ is topmost in all three plots. As expectations of H_n and E_n are close (Table 2), the graphs for $\widetilde{\mathbb{E}}[H_n/T_k]$ and $\widetilde{\mathbb{E}}[E_n/T_k]$ are close in Fig. 3.

Variances of the ratios

The limits of approximations of variances of ratios are presented in Table 5. They behave similarly to the expectations in Table 4. Because L_n and I_n have expectations that grow without bound, for

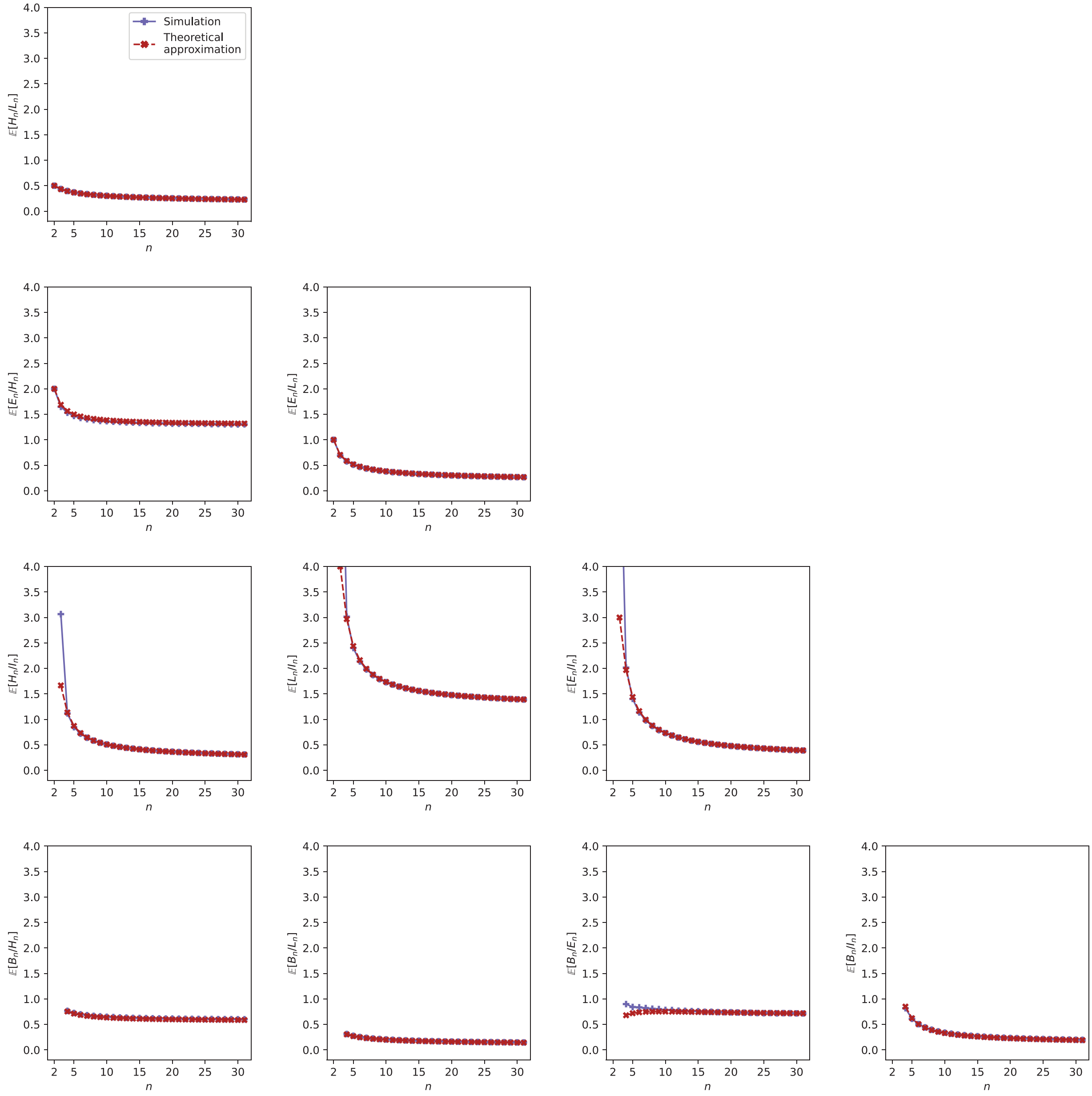

Fig. 2. Simulated and theoretical approximations of expectations of ratios of pairs of variables, plotted as functions of sample size n. Expressions for theoretical values are taken from Table 4.

ratios H_n/L_n, H_n/I_n, E_n/L_n, B_n/L_n, E_n/I_n, B_n/I_n—with L_n or I_n in the denominator—the limits of the variance approximations are 0. As n grows, the denominators grow much faster than the numerators, and the values are therefore increasingly concentrated around 0. Hence, the variances also approach 0.

Because L_n and I_n are much larger than the coalescence times T_k, approximations to variances of L_n/T_k and I_n/T_k diverge to infinity as n increases. Interestingly, however, the approximate variance of L_n/I_n, a ratio of two quantities with diverging expectations, approaches 0.

The variance approximations with finite nonzero limits are those for H_n/T_k, E_n/H_n, E_n/T_k, B_n/H_n, B_n/E_n, and B_n/T_k. All give ratios of two variables with finite expectation and variance as $n \to \infty$ (Table 2).

Figure 4 shows the expressions from Table 5 together with the simulated values. Compared to the plots of expectations of ratios (Fig. 2), differences between the simulated and approximate variances are prominent at small n. For the variances of H_n/L_n, B_n/H_n, and B_n/L_n, the simulated and approximate values differ substantially even as n increases. Because the theoretical value of $\mathrm{Cov}[E_n, B_n]$ that contributes to the approximate variance of B_n/E_n is itself an approximation, one of the larger differences between simulation and approximation occurs for the plot for $\widetilde{\mathrm{Var}}[B_n/E_n]$.

Figure 5 shows variances of ratios involving T_k for varying k, for each of three values of n. Qualitatively, the values for approximate variances behave similarly to expectations in Fig. 3: in particular, the vertical placement of the curves follows the same order. Our approximations to the variances of L_n/T_k and I_n/T_k grow fastest, as the numerators are typically large and the expected value of the denominator T_k decreases as k grows. Approximations to variances of H_n/T_k, E_n/T_k, and B_n/T_k all display

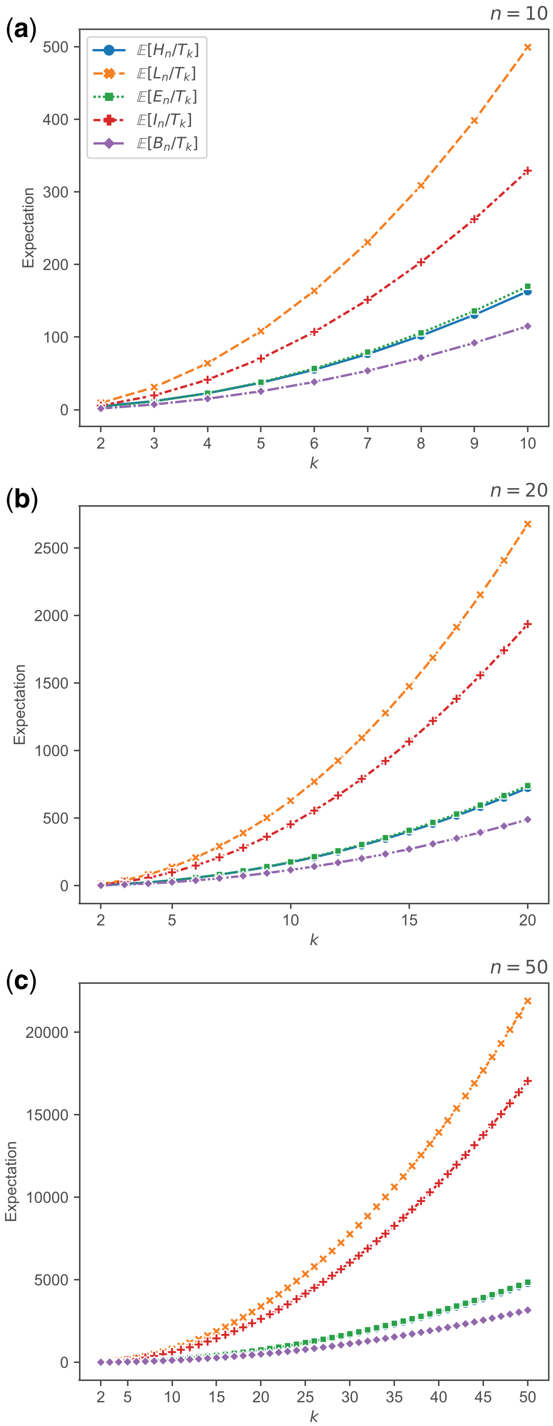

Fig. 3. Theoretical approximations $\widetilde{\mathbb{E}}[X/T_k]$ for variables X in $\{H_n, L_n, E_n, I_n, B_n\}$, plotted as functions of k for $n = 10$, $n = 20$, and $n = 50$. The expressions plotted are taken from Table 4.

much slower growth; for these quantities, the expectations of numerators of the ratios are bounded above by 2 for all n.

Discussion

In this article, we have computed approximations to expected values and variances of ratios of various branch lengths under the standard coalescent model. We have considered all 15 possible pairs of variables in $\{H_n, L_n, E_n, I_n, B_n, T_k\}$, a set of variables whose properties have been studied in detail individually. We have also assessed the accuracy of approximations to the expectation and variance by comparing them with values computed by simulation. We have observed that the approximate expressions behave in a way that matches mathematical intuition about the behavior of random variables associated with the branch lengths.

In plots of the various approximations, we have illustrated how the random variables relate to each other, both among $\{H_n, L_n, E_n, I_n, B_n\}$ (Figs. 2 and 4) as well as between pairs including one of $\{H_n, L_n, E_n, I_n, B_n\}$ along with T_k (Figs. 3 and 5). As n grows large, the ratios involving L_n and I_n have nearly identical behavior in the plots, an observation that is explained by the fact that internal branches take up increasingly large fractions of the total branch length. In the limit as $n \to \infty$, expectations of both H_n and E_n approach a constant value of 2 (Table 2), and $\mathrm{Cov}[H_n, E_n]$ approaches 0 (Table 3). However, we observed that $\lim_{n\to\infty} \widetilde{\mathbb{E}}[E_n/H_n] = \pi^2/3 - 2 \approx 1.28987$ is not equal to $\lim_{n\to\infty} \mathbb{E}[E_n]/\lim_{n\to\infty} \mathbb{E}[H_n] = 1$. For the ratio B_n/E_n, the approximation aligns with the naive prediction, $\lim_{n\to\infty} \widetilde{\mathbb{E}}[B_n/E_n] = \pi^2/6 - 1 = \lim_{n\to\infty} \mathbb{E}[B_n]/\lim_{n\to\infty} \mathbb{E}[E_n]$, even though $\mathrm{Cov}[E_n, B_n]$ is also zero in the limit (Table 2). For B_n and H_n, which possess a high correlation, $\lim_{n\to\infty} \widetilde{\mathbb{E}}[B_n/H_n] = \pi^4/18 - \pi^2/6 - \zeta(3) - 2 \approx 0.56463$, whereas $\lim_{n\to\infty} \mathbb{E}[B_n]/\lim_{n\to\infty} \mathbb{E}[H_n] = \pi^2/6 - 1 \approx 0.64493$.

Previously, we evaluated covariances and correlation coefficients under the coalescent model for the pairs of variables that we consider here, obtaining exact covariances and correlations for 13 of 15 pairs and approximations for the other two. We obtained limiting expressions for these covariances and correlations as $n \to \infty$. The approximate values that we have provided here for expectations and variances of ratios make use of these previous results concerning covariances, adding to the understanding of the properties of joint distributions of pairs of genealogical variables in coalescent theory.

Many statistical tests of population-genetic models rely on a model prediction of an equivalence between two quantities, framed as a null hypothesis that a test statistic equals a particular value. The prediction is often formulated as a null hypothesis that a difference between two quantities equals 0 or that their ratio equals a null value such as 1. In coalescent theory, tests that evaluate site-frequency spectra for agreement with predictions of coalescent models tend to use differences or other linear combinations (Zeng *et al.* 2006; Achaz 2009; Ferretti *et al.* 2010, 2017; Ronen *et al.* 2013; Fu 2022). However, several modeling studies and inference procedures in coalescent theory do emphasize ratios (Slatkin 1996; Uyenoyama 1997; Schierup and Hein 2000; Rosenberg and Hirsh 2003; Eldon 2011; Arbisser *et al.* 2018), as do some test statistics (Schlötterer 2002; Lohse and Kelleher 2009). Widely used tests in the area of molecular evolution, such as tests of the relative count of nonsynonymous and synonymous substitutions and the McDonald–Kreitman test of polymorphism and divergence, also make use of ratios (Yang 2014).

The choice of a difference or a ratio in formulating a test statistic can rely on several factors. Ratios are unitless, so that their values do not depend on conventions chosen during computation (e.g. scaling time in units of N or $2N$). Ratios might take values in a prescribed range that can be simply interpreted, such as the range of the coalescent ratio H_n/L_n from $\frac{1}{n}$ to $\frac{1}{2}$ (Arbisser *et al.* 2018). However, the statistical properties of random variables formulated as differences are generally easier to

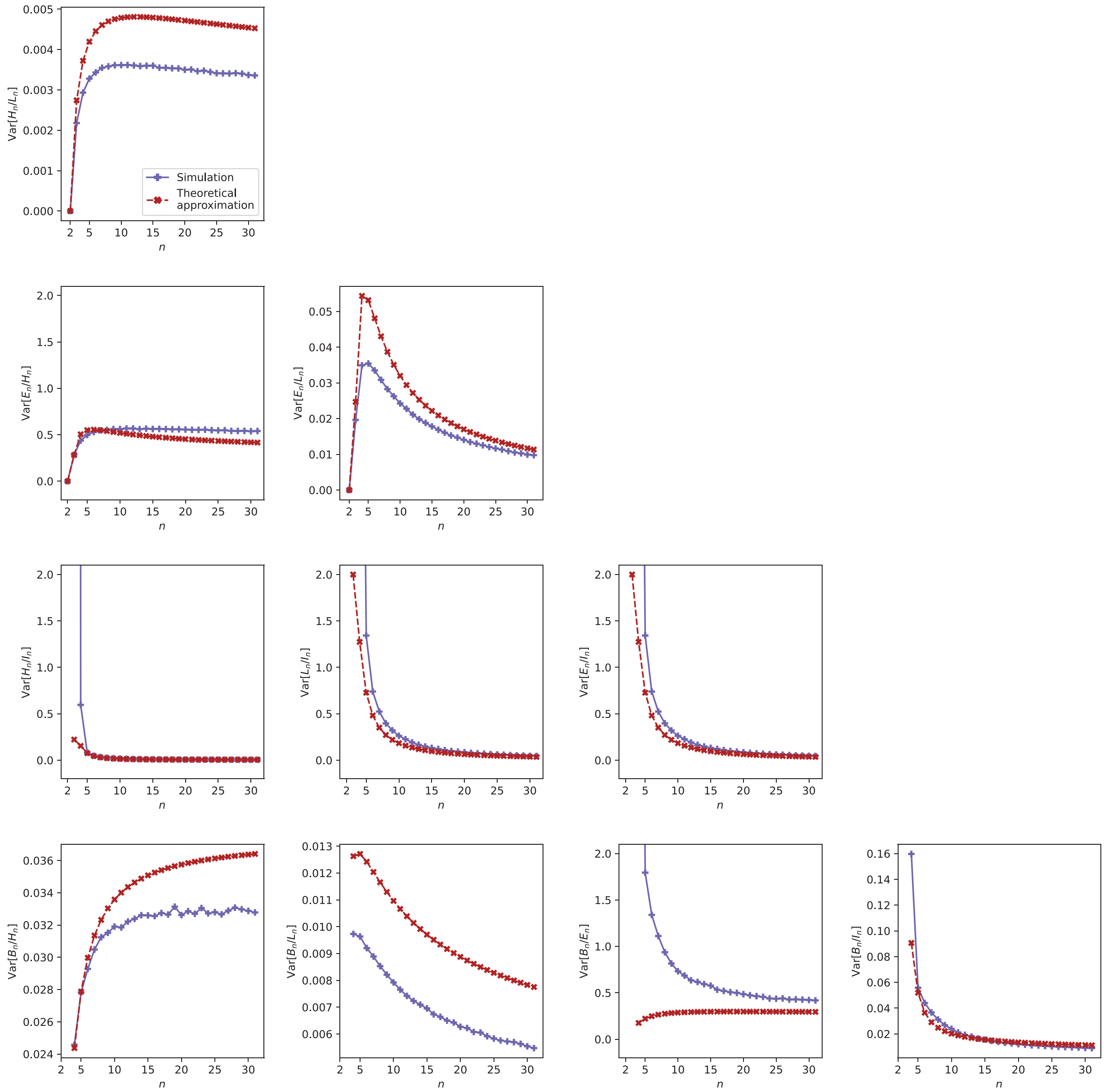

Fig. 4. Simulated and theoretical approximations of variances of ratios of pairs of variables, plotted as functions of sample size n. Expressions for theoretical values are taken from Table 5.

compute from the properties of the separate random variables whose difference is taken than are the properties of corresponding statistics formulated as ratios. In general, corresponding differences and ratios in coalescent theory have not been formally compared for features such as their power to reject the null hypothesis when processes such as natural selection or population or species divergence affect the shapes of evolutionary trees. Our work to obtain approximate expectations and variances of ratios can augment understanding of scenarios in which coalescent ratios are considered, and it can assist in evaluating the relative utility of difference-based and ratio-based statistics.

We have found that approximations for fixed n and in the limit as $n \to \infty$ are quite accurate in predicting the expected values seen in coalescent simulations of the ratios (Fig. 2). For the variances, the approximations are generally less accurate, although in most cases, graphs of the approximations and simulated values have similar shape (Fig. 4). These approximations are obtained from a Taylor approximation for the variance of a ratio (equation 4), and higher-order approximations of this variance could potentially be applied by use of Taylor's theorem; as the order of the approximation increases, however, the complexity of the resulting formula also increases. For those variances for which the approximation and simulation are not close in Fig. 4, we advise caution in using the variances in settings in which a precise approximation is needed.

Data availability

The ms command for simulations is ms n 100,000 -T, where n is taken from $\{2, 3, \ldots, 50\}$ and gives the number of leaves of simulated trees.

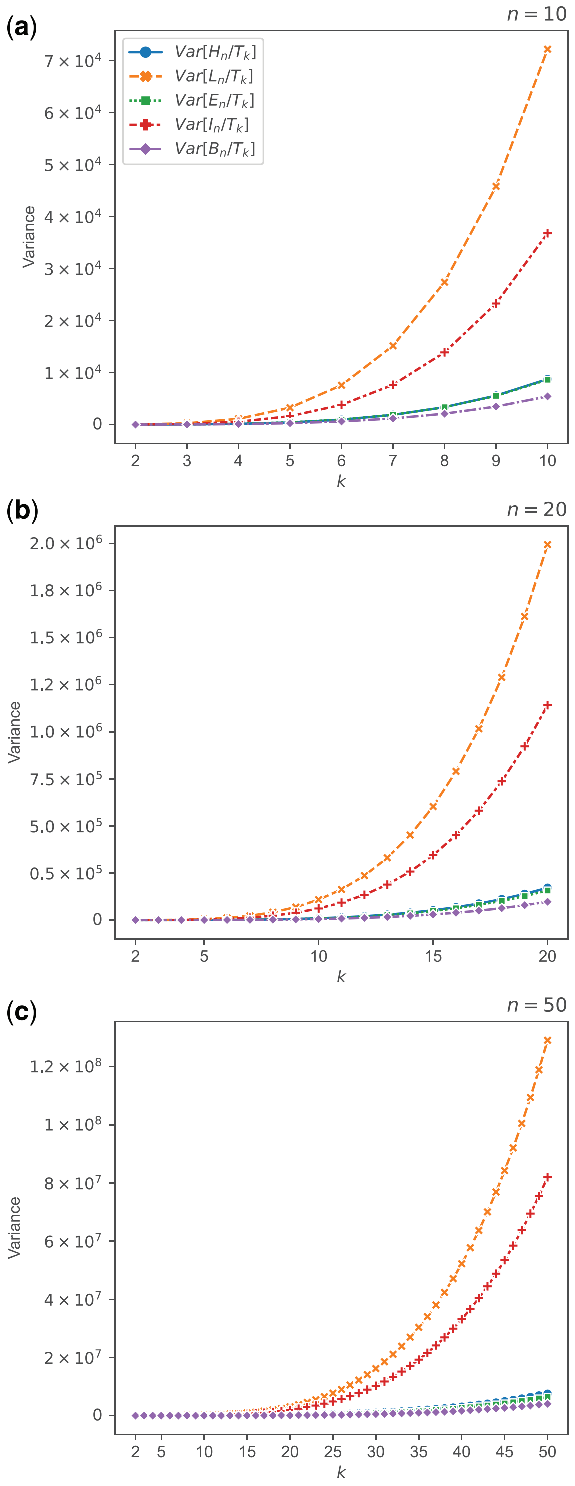

Fig. 5. Theoretical approximations $\widetilde{\mathrm{Var}}[X/T_k]$ for variables X in $\{H_n, L_n, E_n, I_n, B_n\}$, plotted as functions of k for $n=10$, $n=20$, and $n=50$. The expressions plotted are taken from Table 5.

Funding

The authors acknowledge support from NIH grants R01 GM131404 and R01 HG005855 and NSF grant BCS-2116322.

Conflicts of interest

None declared.

Literature cited

Achaz G. Frequency spectrum neutrality tests: one for all and all for one. Genetics. 2009;183(1):249–258.

Alimpiev E, Rosenberg NA. A compendium of covariances and correlation coefficients of coalescent tree properties. Theor Popul Biol. 2022;143:1–13.

Arbisser IM, Jewett EM, Rosenberg NA. On the joint distribution of tree height and tree length under the coalescent. Theor Popul Biol. 2018;122:46–56.

Elandt-Johnson RC, Johnson NL. Survival Models and Data Analysis. New York (NY): Wiley; 1999.

Eldon B. Estimation of parameters in large offspring number models and ratios of coalescence times. Theor Popul Biol. 2011;80(1): 16–28.

Ferretti L, Ledda A, Wiehe T, Achaz G, Ramos-Onsins SE. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. Genetics. 2017;207(1):229–240.

Ferretti L, Perez-Enciso M, Ramos-Onsins S. Optimal neutrality tests based on the frequency spectrum. Genetics. 2010;186(1): 353–365.

Fu YX. Variances and covariances of linar summary statistics of segregating sites. Theor Popul Biol. 2022;145:95–108.

Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133(3):693–709.

Hein J, Schierup M, Wiuf C. Gene Genealogies, Variation and Evolution. Oxford: Oxford University Press; 2005.

Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 1983;37(1):203–217.

Hudson RR. Gene genealogies and the coalescent process. Oxford Surv Evol Biol. 1990;7:1–44.

Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2): 337–338.

Kingman JFC. On the genealogy of large populations. J Appl Probab. 1982;19(A):27–43.

Lohse K, Kelleher J. Measuring the degree of starshape in genealogies—summary statistics and demographic inference. Genet Res (Camb). 2009;91(4):281–292.

Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. Genetics. 2013;195(1):181–193.

Rosenberg NA, Feldman MW. The relationship between coalescence times and population divergence times. In: Slatkin M, Veuille M, editors. Modern Developments in Theoretical Population Genetics. Oxford: Oxford University Press; 2002. Chapter 9, p. 130–164.

Rosenberg NA, Hirsh AE. On the use of star-shaped genealogies in inference of coalescence times. Genetics. 2003;164(4):1677–1682.

Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics. 2000;156(2):879–891.

Schlötterer C. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics. 2002;160(2): 753–763.

Slatkin M. Gene genealogies within mutant allelic classes. Genetics. 1996;143(1):579–587.

Stuart A, Ord JK. Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory. 6th ed. Chichester: Wiley; 1994.

Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437–460.

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3): 585–595.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. Genetics. 1997;145(2):505–518.

Uyenoyama MK. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics. 1997;147(3):1389–1400.

Wakeley J. Coalescent Theory. Greenwood Village (CO): Roberts & Company; 2009.

Yang Z. Molecular Evolution: A Statistical Approach. Oxford: Oxford University Press; 2014.

Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006; 174(3):1431–1439.

Communicating editor: R. Hernandez