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 $I_n$ (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 $\text{Var}[H_n/L_n]$ differed noticeably from the exact value,
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 Alimpiiev 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

\[
 f_{T_k}(t) = \left( \frac{k}{2} \right) e^{-\left( \frac{k}{2} \right) t}.
\]

The expectation and variance of \( T_k \) are

\[
 \mathbb{E}[T_k] = \frac{2}{k(k-1)}, \quad \text{Var}[T_k] = \frac{4}{k^2(k-1)^2}.
\]

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 \( S_{p,n} = \sum_{i=1}^{n} k^{-p} \) 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):

\[
 \mathbb{E}\left[ \frac{X}{Y} \right] \approx \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} + \frac{\mathbb{E}[X]\text{Var}[Y]}{\mathbb{E}[Y]^3}.
\]

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

\[
 \text{Var}\left[ \frac{X}{Y} \right] \approx \left( \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} \right)^2 \left( \frac{\text{Var}[X]}{\mathbb{E}[X]^2} + \frac{2\text{Cov}[X,Y]}{\mathbb{E}[X]\mathbb{E}[Y]} + \frac{\text{Var}[Y]}{\mathbb{E}[Y]^2} \right).
\]

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_n )</td>
<td>( \sum_{k=2}^{n} T_k )</td>
</tr>
<tr>
<td>( L_n )</td>
<td>( \sum_{k=2}^{n} kT_k )</td>
</tr>
<tr>
<td>( E_n )</td>
<td>( \sum_{k=2}^{n} \epsilon^{(k)} )</td>
</tr>
<tr>
<td>( I_n )</td>
<td>( \frac{1}{2} T_2 + \left( \sum_{k=3}^{n-1} \sum_{j=k}^{n} \frac{1}{j-1} T_j \right) + \left( \sum_{k=2}^{n-2} \frac{1}{k-1} T_k \right) )</td>
</tr>
<tr>
<td>( B_n )</td>
<td>( \frac{1}{2} \sum_{k=2}^{n-1} \left( \sum_{j=k+1}^{n} \frac{1}{j-1} T_j \right) )</td>
</tr>
</tbody>
</table>

Here, \( T_k \) is the random variable representing the coalescence time from \( k \) to \( k-1 \) lineages, and \( \epsilon^{(k)} \) is the (random) length of the \( k \)th 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 \( n \geq 4 \).

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

<table>
<thead>
<tr>
<th>( X_n )</th>
<th>( \mathbb{E}[X_n] )</th>
<th>( \lim_{n \to \infty} \mathbb{E}[X_n] )</th>
<th>( \text{Var}[X_n] )</th>
<th>( \lim_{n \to \infty} \text{Var}[X_n] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_n )</td>
<td>( \frac{2(n^2-1)}{n^2} )</td>
<td>2</td>
<td>( 8(S_{2,n-1} - 4 \epsilon_{n}/n) )</td>
<td>( 4S_{2,n-1} )</td>
</tr>
<tr>
<td>( L_n )</td>
<td>2 ( S_{1,n-1} )</td>
<td>( \infty )</td>
<td>( 4 \left( \frac{8}{(n-1)(n-2)} S_{1,n-1} \right) )</td>
<td>( 6.57974 )</td>
</tr>
<tr>
<td>( E_n )</td>
<td>2</td>
<td>2</td>
<td>( n \geq 2 )</td>
<td>( \frac{6}{n} )</td>
</tr>
<tr>
<td>( I_n )</td>
<td>2 ( S_{1,n-1} )</td>
<td>( \infty )</td>
<td>( 4 \left( \frac{2 \left( S_{1,n-1} \right) - 2(n-1)}{n^2(n-1)} \right) )</td>
<td>( 6.57974 )</td>
</tr>
<tr>
<td>( B_n )</td>
<td>2 ( S_{2,n-1} )</td>
<td>( \frac{1}{2} )</td>
<td>( \frac{n^2}{2} - 2 \approx 1.28987 )</td>
<td>( \frac{1}{2} n^2 + 2 \approx 1.04637 )</td>
</tr>
<tr>
<td>( T_k )</td>
<td>( \frac{2}{k^2(k-1)} )</td>
<td>( \frac{2}{k^2(k-1)} )</td>
<td>( \frac{4}{k^3(k-1)} )</td>
<td>( \frac{4}{k^3(k-1)} )</td>
</tr>
</tbody>
</table>

These expressions can be found in Alimpiiev 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).
Covariances of pairs of variables that summarize genealogical trees:

<table>
<thead>
<tr>
<th>(X_{n}, Y_{n})</th>
<th>(\text{Cov}[X_{n}, Y_{n}])</th>
<th>(\text{lim}<em>{n \to \infty} \text{Cov}[X</em>{n}, Y_{n}])</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H_{n}, T_{k})</td>
<td>(\frac{6}{n} - \frac{k}{n(n-1)})</td>
<td>(\frac{6}{n(n-1)})</td>
</tr>
<tr>
<td>(H_{n}, L_{n})</td>
<td>(4S_{2,n-1} - 4 + \frac{4}{n})</td>
<td>(2\pi^2 - 4 \approx 2.57974)</td>
</tr>
<tr>
<td>(H_{n}, E_{n})</td>
<td>(\frac{4}{n})</td>
<td>(0)</td>
</tr>
<tr>
<td>(H_{n}, I_{n})</td>
<td>(4S_{2,n-1} - 4)</td>
<td>(\frac{4\pi^2}{n(n-1)} - 4 \approx 2.57974)</td>
</tr>
<tr>
<td>(H_{n}, B_{n})</td>
<td>(4S_{2,n-1} - 4 + \frac{4}{n})</td>
<td>(\frac{4\pi^2}{n(n-1)})</td>
</tr>
<tr>
<td>(E_{n}, T_{k})</td>
<td>(\frac{4}{n})</td>
<td>(0)</td>
</tr>
<tr>
<td>(E_{n}, L_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(E_{n}, E_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(E_{n}, I_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(E_{n}, B_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(B_{n}, T_{k})</td>
<td>(\frac{4}{n})</td>
<td>(0)</td>
</tr>
<tr>
<td>(B_{n}, L_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(B_{n}, E_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(B_{n}, I_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
<tr>
<td>(B_{n}, B_{n})</td>
<td>(\frac{4S_{2,n-1} - 4}{n} + \frac{4}{n(n-1)})</td>
<td>(0)</td>
</tr>
</tbody>
</table>

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 \(\mathbb{E}[X/Y]\) and \(\text{Var}[X/Y]\) 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 \(H_{n}, L_{n}, E_{n}, I_{n}, B_{n}\), and \(T_{k}\) as well as \(S_{n}\), 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 \(\text{Var}[X_{n}/Y_{n}]\) under the coalescent model using ms (Hudson 2002), performing 100,000 replicate simulations for each tree size.
Table 5. Approximations to variances of ratios of pairs of variables.

<table>
<thead>
<tr>
<th>(Xn, Yn)</th>
<th>Var[Xn/Yn]</th>
<th>limn→∞ Var[Xn/Yn]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hn, Tk</td>
<td>$\frac{2k(k-1)p(n-1)}{n(n-2)} + \frac{2(k^2-6)(k-6)}{n(n-2)}$</td>
<td>$\frac{1}{2}k(k-1)((n^2 - 6)k^2 - (n^2 - 6)k - 6)$</td>
</tr>
<tr>
<td>Hn, I0</td>
<td>$\frac{2k^2}{(n-1)^3}$</td>
<td>0</td>
</tr>
<tr>
<td>En, Hn</td>
<td>$\frac{2k^2}{(n-1)^2}$</td>
<td>$\frac{n^2}{3} - \frac{k}{n} = 0.2897$</td>
</tr>
<tr>
<td>Hn, I0</td>
<td>$\frac{2k^2}{(n-1)^2}$</td>
<td>0</td>
</tr>
<tr>
<td>Bn, Hn</td>
<td>$\frac{2(k^2-6)(k-6)}{n(n-2)}$</td>
<td>$\frac{n^2}{3} + \frac{k}{n} = 0.26159$</td>
</tr>
<tr>
<td>En, Tk</td>
<td>$\frac{k^2(k-1)^2}{n(n-2)}$</td>
<td>$\frac{1}{2}k^2(k-1)^2$</td>
</tr>
<tr>
<td>En, I0</td>
<td>$\frac{k^2}{n(n-2)}$</td>
<td>0</td>
</tr>
<tr>
<td>Bn, En</td>
<td>$\frac{2k^2}{(n-1)^2}$</td>
<td>$\frac{1}{2}k^2(k-1)^2 + \frac{2k^2}{n(n-2)}$</td>
</tr>
<tr>
<td>I0, Tk</td>
<td>$\frac{k^2(k-1)^2}{n(n-2)}$</td>
<td>$\frac{1}{2}k^2(k-1)^2 + \frac{2k^2}{n(n-2)}$</td>
</tr>
<tr>
<td>Bn, I0</td>
<td>$\frac{2k^2}{(n-1)^2}$</td>
<td>$\frac{1}{2}k^2(k-1)^2 + \frac{2k^2}{n(n-2)}$</td>
</tr>
</tbody>
</table>

Expressions involving En or I0 apply for n ≥ 3; expressions involving Bn apply for n ≥ 4. The value for (Hn, L0) follows equation 18 of Arbisser et al. (2018). The expressions are obtained using equation 4 and Tables 2 and 3.

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 Hn/I0, Hn/I0, En/I0, En/I0, Bn/I0, and Bn/I0 approach 0. This behavior is sensible when considering the properties of the coalescent model: in the numerators, En has expectation 2 and E[H0] and E[Bn] have bounded expectation in the limit as n → ∞; in the denominators, L0 and I0 have expectations that grow without bound (Table 2). Similarly, approximate expectations of ratios I0/Tk or I0/Tk with I0 and Tk in the numerator and Tk in the denominator grow to infinity as n increases. The approximation to E[L0/I0] approaches 1 in the limit as n → ∞: 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—Hn/Tk, En/Tk, Bn/Hn, Bn/E0, and Bn/Tk—also approach finite values in the limit as n → ∞. It is interesting to observe that although limn→∞ E[En] = limn→∞ E[H0] = 2 (Table 2), E[H0/E0] = π2/3 – 2 ≈ 1.2897 ≠ 1. In other words, although expectations of the individual variables approach the same value, we expect En/H0 to be somewhat larger than 1 on average.

For each of the 10 pairs of variables among {Hn, L0, En, I0, Bn}, 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 Tk are shown in Fig. 3 as functions of k for each of three values of n. Ln is the fastest-growing variable according to the expression for its expectation (Table 2), and the graph for E[L0/Tk] is topmost in all three plots. As expectations of Hn and En are close (Table 2), the graphs for E[H0/Tk] and E[En/Tk] 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 Ln and I0 have expectations that grow without bound, for
ratios $H_n/L_n$, $E_n/L_n$, $E_n/H_n$, $B_n/L_n$, $E_n/L_n$, $B_n/L_n$—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 $\text{Cov}(E_n, B_n)$ 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 $\text{Var}(B_n/E_n)$.

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
The expressions plotted are taken from Table 4. The standard coalescent model. We have considered all 15 values and variances of ratios of various branch lengths under this model. In this article, we have computed approximations to expected 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 $\{H_n, L_n, E_n, I_n, B_n, T_n\}$, 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 $\{H_n, I_n, E_n, I_n, B_n\}$ (Figs. 2 and 4) as well as between pairs including one of $\{H_n, I_n, E_n, I_n, B_n\}$ along with $T_n$ (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 $n \rightarrow \infty$, expectations of both $H_n$ and $E_n$ approach a constant value of 2 (Table 2), and $\text{Cov}[H_n, E_n]$ approaches 0 (Table 3). However, we observed that $\lim_{n \rightarrow \infty} \frac{E[H_n]}{E[1]} = \frac{\pi^2}{3} - 2 \approx 1.28987$ is not equal to $\lim_{n \rightarrow \infty} \frac{E[H_n]}{ \text{lim}_{n \rightarrow \infty} E[1]} = 1$. For the ratio $B_n/E_n$, the approximation aligns with the naive prediction, $\lim_{n \rightarrow \infty} \frac{E[B_n/E_n]}{E[1]} = \frac{\pi^2}{6} - 1 = \frac{\pi^2}{6} - 2 \approx 0.56463$, whereas $\lim_{n \rightarrow \infty} \frac{E[B_n]}{\text{lim}_{n \rightarrow \infty} E[1]} = \frac{\pi^2}{6} - 1 \approx 0.64493$.

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 $n \rightarrow \infty$. 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 (Schlotterer 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 $H_n/I_n$ from $\frac{1}{2}$ to $\frac{3}{2}$ (Arbisser et al. 2018). However, the statistical properties of random variables formulated as differences are generally easier to
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.
The authors acknowledge support from NIH grants R01 GM131404 and R01 HG005855 and NSF grant BCS-2116322.

Conflicts of interest
None declared.

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

Literature cited


Communicating editor: R. Hernandez