A compendium of covariances and correlation coefficients of coalescent tree properties

Egor Alimpiev*, Noah A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, United States of America

ARTICLE INFO

Article history:
Received 2 March 2021
Available online 30 October 2021

Keywords:
Coalescent
External branches
Genealogies
Internal branches
Time to the most recent common ancestor

ABSTRACT

Gene genealogies are frequently studied by measuring properties such as their height ($H_n$), length ($L_n$), sum of external branches ($E_n$), sum of internal branches ($I_n$), and mean of their two basal branches ($B_n$), and the coalescence times that contribute to the other genealogical features ($T_n$). These tree properties and their relationships can provide insight into the effects of population-genetic processes on genealogies and genetic sequences. Here, under the coalescent model, we study the 15 correlations among pairs of features of genealogical trees: $H_n$, $L_n$, $E_n$, $I_n$, $B_n$, and $T_k$ for a sample of size $n$, with $2 \leq k \leq n$. We report high correlations among $H_n$, $L_n$, $I_n$, and $B_n$, with all pairwise correlations of these quantities having values greater than or equal to $\sqrt{6(3 + 6 - n^2)}/(\pi \sqrt{18 + 9n^2} - n^2) \approx 0.84930$ in the limit as $n \to \infty$, where $\zeta$ is the Riemann zeta function. Although $E_n$ has expectation 2 for all $n$ and $H_n$ has expectation 2 in the $n \to \infty$ limit, their limiting correlation is 0. The results contribute toward understanding features of the shapes of coalescent trees.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

In coalescent theory, features of gene genealogies are investigated in relation to the evolutionary processes that are included in population-genetic models (Hein et al., 2005; Wakeley, 2009). For example, comparing a constant-sized and an exponentially growing population, exponential growth increases the total length of the branches of a gene genealogy in relation to its height (Slatkin and Hudson, 1991; Slatkin, 1996; Sano and Tachida, 2005). Coalescences are rare in recent generations, when the population is large, and they occur primarily in the period deep in the past when the population was small.

Several tree features have been used for measuring effects of population-genetic processes on gene genealogies (Slatkin, 1996; Uyenoyama, 1997; Schierup and Hein, 2000; Rosenberg, 2006). For a binary ultrametric tree of $n$ lineages, these features (Fig. 1) include the tree height from the tips to the root ($H_n$), the total length of all the branches ($L_n$), the total length of external branches connecting tips to the nearest internal node ($E_n$), the total length of internal branches connecting internal nodes to other internal nodes ($I_n$), and the mean length of the two basal branches incident to the root node ($B_n$).

These tree features can all be expressed as linear combinations, random linear combinations in some cases, of the same underlying random variables — the coalescence times $T_k$ for coalescence of $k$ to $k - 1$ lineages, with $2 \leq k \leq n$. Hence, the quantities are correlated. For example, the tree height $H_n$ includes the mean basal branch length $B_n$, and the total length $L_n$ is the sum of the length $E_n$ of the external branches and the length $I_n$ of the internal branches; an increase in $L_n$ necessarily increases $E_n$, $I_n$, or both.

Analyses of coalescent models have examined some of the correlations between tree features, notably the relationship between $H_n$ and $L_n$ (Fu, 1996; Griffiths and Tavaré, 1996; Rosenberg and Hirsh, 2003; Arbisser et al., 2018). Here, we perform a detailed investigation of correlations among $H_n$, $L_n$, $E_n$, $I_n$, and $B_n$. For each pair, under the coalescent, assuming a constant-sized population,
we evaluate their covariance and correlation. We explore limiting values as \( n \to \infty \). The approach follows Arbibser et al. (2018), who obtained the covariance and correlation of \( H_n \) and \( L_n \); we perform analogous calculations for all 10 pairs among \((H_n, L_n, E_n, I_n, B_n)\), as well as for the five pairs involving one of \((H_n, I_n, E_n, I_n, B_n)\) and \( T_k \).

2. Tree properties

We consider the standard coalescent model of a constant-sized population of size \( N \) haploids. Time is measured in units of the population size, with one unit representing \( N \) generations. For sample size \( n \geq 2 \), we examine tree properties \( H_n, L_n, E_n, I_n, \) and \( B_n \), as well as the coalescence time \( T_k \), \( 2 \leq k \leq n \). In this section, we recall basic features of the various quantities.

For convenience, for a mathematical expression we will use frequently, we write

\[
S_{p,n} = \sum_{k=1}^{n} \frac{1}{k^p}.
\]  

The limit \( S_{p,\infty} = \lim_{n \to \infty} S_{p,n} \) is the Riemann zeta function \( \zeta(p) \). The harmonic sum \( S_{1,\infty} \) diverges, and the sum of the reciprocals of squares is \( S_{2,\infty} = \pi^2/6 \approx 1.64493. \) The sum of the reciprocals of cubes is Apéry’s constant, \( S_{3,\infty} = \zeta(3) \approx 1.20206. \)

2.1. \( T_k \)

\( T_k \) is a random variable representing the time during which \( k \) lineages coalesce to \( k-1 \) lineages. The \( T_k, 2 \leq k \leq n \), are independent and exponentially distributed with probability density function \( f_{T_k}(t_k) = \frac{1}{k}e^{-t_k/k} \) (Wakeley, 2009, p. 60). The expectation and variance of \( T_k \) are then

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

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

As \( n, k \to \infty \) with \( k \leq n \), both \( \mathbb{E}[T_k] \) and \( \text{Var}[T_k] \) have limit 0.

2.2. \( H_n \)

For \( n \geq 2 \), the height \( H_n \) of a tree from root to leaves can be written

\[
H_n = \sum_{k=2}^{n} T_k.
\]

The expectation and variance of \( H_n \) are then found using the expectation and variance of \( T_k \) (Eqs. (2) and (3)), noting that the \( T_k \) are independent:

\[
\mathbb{E}[H_n] = \sum_{k=2}^{n} \mathbb{E}[T_k] = \frac{2(n-1)}{n},
\]

\[
\text{Var}[H_n] = 8 \left( \sum_{k=2}^{n} \frac{1}{k^2} \right) - 4 \left( \frac{n-1}{n} \right)^2.
\]

The variance can be written \( \text{Var}[H_n] = 4(2S_{2,n}n^2 - 3n^2 + 2n - 1)/n^2. \) The limits are \( \lim_{n \to \infty} \mathbb{E}[H_n] = 2 \) and \( \lim_{n \to \infty} \text{Var}[H_n] = 4\pi^2/3 \approx 11.5947 \) (Wakeley, 2009, p. 76).

2.3. \( L_n \)

For \( n \geq 2 \), the total tree length, summing the lengths of all branches of a tree, is

\[
L_n = \sum_{k=2}^{n} k T_k.
\]

By Eqs. (2) and (3) and the independence of the \( T_k \), we have

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

\[
\text{Var}[L_n] = 4 \sum_{k=1}^{n-1} \frac{1}{k^2}.
\]

In terms of \( S_{p,n} \) (Eq. (1)), these expressions are \( \mathbb{E}[L_n] = 2S_{1,n-1} \) and \( \text{Var}[L_n] = 4S_{2,n-1}. \) The limits are \( \lim_{n \to \infty} \mathbb{E}[L_n] = \infty \) and \( \lim_{n \to \infty} \text{Var}[L_n] = 2\pi^2/3 \approx 6.57974 \) (Wakeley, 2009, p. 76).

2.4. \( E_n \)

The external branches of a tree are the branches that connect leaves to their nearest internal nodes. Denoting the individual external branch lengths \( e_1^{(n)}, e_2^{(n)}, \ldots, e_n^{(n)} \), the sum of external branch lengths is

\[
E_n = e_1^{(n)} + e_2^{(n)} + \cdots + e_n^{(n)}.
\]

The \( e_k^{(n)} \) are identically distributed, and we write \( e_n \) for the length of a randomly chosen external branch of a tree of \( n \) lineages. The sum of the external branches has expectation

\[
\mathbb{E}[E_n] = n \mathbb{E}[e_n].
\]

The random variable \( e_n \) can be written recursively as (Fu and Li, 1993, eq. 7)

\[
e_n = \begin{cases} 
    e_{n-1} + T_n, & \text{with probability } \frac{n-2}{n}, \\
    T_n, & \text{with probability } \frac{2}{n}.
\end{cases}
\]

Expressions for \( \mathbb{E}[e_n], \mathbb{E}[E_n], \) and \( \text{Var}[E_n] \) can then be obtained by solving recurrence equations (Fu and Li, 1993). We have

\[
\mathbb{E}[e_n] = \frac{2}{n},
\]

\[
\mathbb{E}[E_n] = 2,
\]

\[
\text{Var}[E_n] = \begin{cases} 
    4, & n = 2, \\
    \frac{8}{(n-1)(n-2)} \left[ n \left( \sum_{k=1}^{n-1} \frac{1}{k} \right) - 2(n-1) \right], & n > 2.
\end{cases}
\]

\( \mathbb{E}[E_n] \) is equal to 2 irrespective of the choice of \( n \), so that \( \lim_{n \to \infty} \mathbb{E}[E_n] = 2. \) The limit of the variance is \( \lim_{n \to \infty} \text{Var}[E_n] = 0 \) (Fu and Li, 1993).

2.5. \( I_n \)

The internal branches connect internal nodes to other internal nodes. Their total length is \( I_n \), with

\[
I_n = L_n - E_n.
\]

The mean and variance of the sum of internal branches are (Fu and Li, 1993, Eqs. 12 and 17)

\[
\mathbb{E}[I_n] = \mathbb{E}[L_n] - \mathbb{E}[E_n] = 2 \left( \sum_{k=1}^{n} \frac{1}{k} \right) - 2, 
\]

\[
\text{Var}[I_n] = 4 \left[ 2S_{1,n-1}n - 2(n-1) \left( \frac{1}{n-1} - \frac{1}{n} \right) - 2S_{1,n-1} + S_{2,n-1} \right].
\]

The limits are \( \lim_{n \to \infty} \mathbb{E}[I_n] = \infty \) and \( \lim_{n \to \infty} \text{Var}[I_n] = 2\pi^2/3 \approx 6.57974, \) the same as for \( L_n \) (Section 2.3).
2.6. \( B_n \)

Finally, we consider the basal branches, the two branches that extend from the root. We define \( B_n \) as the mean of the two branch lengths. One of the branches has length \( T_2 \), and we denote the other length \( B_n \). We assume here that \( n \geq 4 \) for calculations involving \( B_n \). The appendix of \textcite{Uyenoyama1997} gives

\[
B_n = \frac{T_2 + b_n}{2},
\]

with

\[
b_n = \left[ \sum_{j=1}^{n-1} \sum_{k=2}^{j-1} T_j \prod_{i=3}^{j} \left(1 - \frac{1}{i}\right) \right] + \left[ \sum_{k=2}^{n} T_k \prod_{i=3}^{n} \left(1 - \frac{1}{i}\right) \right]
\]

for \( n \geq 4 \). A convenient form for \( b_n \) encodes the fact that with probability \( 2/(j(j-1)) \), \( b_n = H_j \) for \( j = 3, 4, \ldots, n - 1 \), and with probability \( 2/(n-1) \), \( b_n = H_n \):

\[
b_n = \left[ \sum_{j=1}^{n-1} \sum_{k=2}^{j-1} \frac{2}{(j-1)T_j} \right] + \left[ \sum_{k=2}^{n} \frac{2}{n-1} - T_k \right].
\]

Assuming \( n \geq 4 \), the branch length \( b_n \) has expectation \cite{Uyenoyama1997}:

\[
\mathbb{E}[b_n] = \frac{4}{n} + 4 \sum_{k=3}^{n-1} \frac{1}{k^2}.
\]

The expectation and variance of \( B_n \) then equal

\[
\mathbb{E}[B_n] = \frac{2}{n} + 2 \sum_{k=3}^{n-1} \frac{1}{k^2},
\]

\[
\text{Var}[B_n] = \frac{2(3S_{2,n} - n^2T_2^2 - 2S_{2,n-1}n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}{n^2}.
\]

The expectation appears in the appendix of \textcite{Uyenoyama1997}. We calculate the expression for the variance in Section 3.11. Taking limits of these equations, we obtain \( \lim_{n \to \infty} \mathbb{E}[B_n] = \pi^2/3 - 2 \approx 1.28987 \) and \( \lim_{n \to \infty} \text{Var}[B_n] = 2 + \pi^2 - \pi^2/9 \approx 1.04637 \).

3. Theoretical results

For pairs of variables among \( \{H_n, L_n, E_n, I_n, B_n, T_k\} \), we apply results from Section 2 to compute covariances and correlations. First, for each pair, we compute their covariance. The covariance together with the variances of the two quantities from Section 2 provides their correlation. We obtain the limiting correlation for large trees by taking \( n \to \infty \). Among the 15 pairs, our analyses for 13 are exact; for \( (E_n, B_n) \) and \( (I_n, B_n) \), we offer approximate covariances and correlations. We also provide the derivation of Eq. (22) for \( \text{Var}[b_n] \). Note that correlations in pairs involving \( E_n \) have distinct forms for \( n = 2 \) and \( n \geq 3 \), owing to the piecewise definition of \( \text{Var}[E_n] \) in Eq. (14). We exclude the case of \( n = 2 \) for pairs involving \( I_2 \), as \( L_2 = 0 \) with \( \text{Var}[I_2] = 0 \). We also assume that \( b_n \) is defined only for \( n \geq 4 \).

We present a summary of our mathematical results in Tables 1 and 2. Table 1 shows covariances of pairs of variables and their limits as \( n \to \infty \). Table 2 shows correlations and their \( n \to \infty \) limits.

3.1. \( H_n \) and \( T_k \)

We calculate the covariance of \( H_n \) and \( T_k \) using \( \text{Cov}[H_n, T_k] = \mathbb{E}[H_n T_k] - \mathbb{E}[H_n] \mathbb{E}[T_k] \). Recalling that \( T_i \) and \( T_j \) are independent for \( i \neq j \) (Section 2.1), we have \( \mathbb{E}[T_i T_j] = \mathbb{E}[T_i] \mathbb{E}[T_j] \) for \( i \neq j \). Hence, inserting Eq. (4) for \( H_n \) and Eq. (2) for \( \mathbb{E}[T_k] \), we have

\[
\text{Cov}[H_n, T_k] = \mathbb{E} \left[ T_k \sum_{i=2}^{n} T_i \right] - \mathbb{E} \left[ T_k \right] \mathbb{E} \left[ T_i \right] \]

\[
= \sum_{i=2}^{n} \mathbb{E} \left[ T_i T_k \right] - \mathbb{E} \left[ T_i \right] \mathbb{E} \left[ T_k \right] \]

\[
= \text{Var} \left[ T_k \right] + \sum_{i=2}^{n} \mathbb{E} \left[ T_i \right] \mathbb{E} \left[ T_k \right] \]

\[
- \sum_{i=2}^{n} \mathbb{E} \left[ T_i T_k \right] \mathbb{E} \left[ T_k \right] = \text{Var} \left[ T_k \right] = \frac{4}{k(k-1)^2},
\]

where the last step uses \( \text{Var}[T_k] \) from Eq. (3). We observe that the covariance is independent of \( n \).

For the correlation coefficient \( \text{Corr}[H_n, T_k] = \text{Cov}[H_n, T_k] / \sqrt{\text{Var}[H_n] \text{Var}[T_k]} \), applying Eq. (6) for \( \text{Var}[H_n] \), Eq. (3) for \( \text{Var}[T_k] \), and Eq. (23) for \( \text{Cov}[H_n, T_k] \), we have

\[
\text{Corr}[H_n, T_k] = \frac{n}{\sqrt{2S_{2,n}n^2 - 3n^2 + 2n - 1}} \frac{1}{k(k-1)^2}.
\]

Taking a limit as \( n \to \infty \), we obtain

\[
\lim_{n \to \infty} \text{Corr}[H_n, T_k] = \frac{\sqrt{3}}{\sqrt{\pi^2 - 9}} \frac{1}{k(k-1)^2}.
\]

The limiting correlation decreases to 0 with \( k \) from an initial value of \( \frac{1}{\sqrt{2(\pi^2 - 9)}} \approx 0.92869 \) at \( k = 2 \).

3.2. \( L_n \) and \( T_k \)

For \( L_n \) and \( T_k \), applying Eqs. (7), (8) and (2), we have for the covariance

\[
\text{Cov}[L_n, T_k] = \mathbb{E} \left[ L_n T_k \right] - \mathbb{E} \left[ L_n \right] \mathbb{E} \left[ T_k \right] \]

\[
= \mathbb{E} \left[ T_k \sum_{i=2}^{n} i T_i \right] - \mathbb{E} \left[ \sum_{i=2}^{n} i T_i \right] \mathbb{E} \left[ T_k \right] \]

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

where the last step uses Eq. (3). The covariance of \( L_n \) and \( T_k \), like \( \text{Cov}[H_n, T_k] \) (Eq. (23)), is independent of \( n \).

Now we calculate the correlation coefficient from Eqs. (26), (9) and (3):

\[
\text{Corr}[L_n, T_k] = \frac{1}{\sqrt{2S_{2,n+1}n^2 - 3n^2 - 2n + 1}} \frac{1}{k(k-1)^2}.
\]

If we let \( n \to \infty \), then this quantity becomes

\[
\lim_{n \to \infty} \text{Corr}[L_n, T_k] = \frac{6}{\pi} \frac{1}{k(k-1)}.
\]

The limiting correlation decreases to 0 with \( k \), starting for \( k = 2 \) at \( \sqrt{6}/\pi \approx 0.77970 \).

3.3. \( H_n \) and \( L_n \)

\textcite{Arbiser2018} reported the covariance and correlation of \( H_n \) and \( L_n \). By Eq. (4) and the linearity of the covariance,

\[
\text{Cov}[H_n, L_n] = \sum_{k=2}^{n} \text{Cov}[L_n, T_k].
\]
Applying Eq. (26), we obtain

\[
\text{Cov}[H_n, L_0] = 4S_{2,n-1} - 4 + \frac{4}{n}. \tag{29}
\]

The limit of the covariance is

\[
\lim_{n \to \infty} \text{Cov}[H_n, L_0] = \frac{2\pi^2}{3} - 4 \approx 2.57974. \tag{30}
\]

Dividing the covariance in Eq. (29) by the square root of the product of Eqs. (6) and (9), we obtain

\[
\text{Corr}[H_n, L_0] = \frac{S_{2,n-1} - n + 1}{\sqrt{S_{2,n-1}(2S_{2,n-1} - 2n^2 + 2n - 1)}}. \tag{31}
\]

The limit is

\[
\lim_{n \to \infty} \text{Corr}[H_n, L_0] = \frac{\pi^2 - 6}{\pi \sqrt{2\pi^2 - 9}} \approx 0.93399. \tag{32}
\]

### 3.4. \( H_n \) and \( E_n \)

For the covariance \( \text{Cov}[H_n, E_n] = \text{Var}[H_n] - \text{Var}[E_n] \), we first note that by Eqs. (5) and (13), the second term is simply \( 4 \left( 1 - \frac{1}{n} \right) \). Expanding \( \text{Cov}[H_n, E_n] \) by using the definition of \( H_n \) (Eq. (4)) gives us

\[
\text{Cov}[H_n, E_n] = \sum_{i=2}^{n} \text{Cov}[E_i T_i] = n \sum_{i=2}^{n} \text{Cov}[e_i T_i],
\]

as all external branch lengths are identically distributed (Eq. (10)).

For integers \( k, i \) with \( 2 \leq k, i \leq n \), the external branch length \( e_i \), representing the length of a randomly chosen external branch for a tree with \( k \) leaves, and the coalescence time \( T_i \), satisfy (Eq. (11))

\[
e_i T_i = \frac{e_i - 1}{T_i} + T_i, \quad \text{with probability } \frac{k^2}{k}, \quad \text{with probability } \frac{2}{k}. \tag{33}
\]

where for convenience, we write \( e_1 = 0 \).

Note that \( e_i \) and \( T_i \) are independent for \( i > k \), as the recurrence for \( e_i \) is constructed only using coalescence times \( T_2, T_3, \ldots, T_k \) (Eq. (11)); each of these times is independent of \( T_i \) for \( i > k \) (Section 2.1). We solve to find \( \text{E}[e_i T_i] \) by computing \( \text{E}[e_i T_i] \), incrementing \( k \) from 2 to \( n \). The calculations are similar to those of the Appendix of Fu and Li (1993).

\[
\text{E}[e_i T_i] = \text{Var}[T_i] + \text{E}[T_i]^2 = 2\text{E}[T_i]^2 \quad \text{by Eqs. (2) and (3),}
\]

By Eqs. (12) and (2) and the independence of \( e_i \) and \( T_i \) for \( i > k \), for \( i \geq 3 \),

\[
\text{E}[e_i T_i] = \text{E}[e_i - 1] + \text{E}[T_i] = \frac{4}{i(i - 1)^2}. \tag{34}
\]

Noting \( \text{E}[T_i^2] = \text{Var}[T_i] + \text{E}[T_i]^2 = 2\text{E}[T_i]^2 \) by Eqs. (2) and (3), we use Eq. (33) to write an expression for \( \text{E}[e_i T_i] \):

\[
\text{E}[e_i T_i] = \frac{i - 1}{i} \text{E}[e_i - 1] + \text{E}[T_i]^2 = \frac{4}{i(i - 1)^2}. \tag{35}
\]

Next, incrementing Eq. (34), we have

\[
\text{E}[e_i T_i] = \frac{i - 1}{i} \text{E}[e_i - 1] + \text{E}[T_i + 1] + \text{E}[T_i] = \frac{4}{i(i - 1)^2}, \tag{35}
\]

by Eqs. (2) and (12).

The final step is to solve the recurrence equation

\[
\text{E}[e_i T_i] = \frac{n - 2}{n} \text{E}[e_{n-1} T_i] + \text{E}[T_i]. \tag{35}
\]
with initial condition Eq. (35). Recalling the case of \( i = n = 2 \), with 2 \( \leq i \leq n \), we obtain solution
\[
\mathbb{E}[e_{nT}] = \frac{4}{i(i-1)(n-1)}. \tag{36}
\]

Applying Eq. (10), the expression for \( \text{Cov}[H_n, E_n] \) becomes
\[
\text{Cov}[H_n, E_n] = n \sum_{i=2}^{n} \frac{4}{i(i-1)(n-1)} (1 - \frac{1}{i}) = \frac{4}{n}. \tag{37}
\]

The limit of the covariance as \( n \to \infty \) is
\[
\lim_{n \to \infty} \text{Cov}[H_n, E_n] = 0. \tag{38}
\]

Dividing Eq. (37) by the square root of the product of variances from Eqs. (6) and (14), the correlation is
\[
\text{Corr}[H_n, E_n] = \frac{\sqrt{n}}{\sqrt{2(2S_{2,n}^2 + 3n^2 + 2n - 1) (S_{n,t-1} - n - 1)}} = \frac{n-1}{n} \tag{39}
\]

The limit of the correlation is
\[
\lim_{n \to \infty} \text{Corr}[H_n, E_n] = 0. \tag{40}
\]

3.5. \( E_n \) and \( T_k \)

In the process of computing \( \text{Cov}[H_n, E_n] \), we have obtained an expression for \( \mathbb{E}[e_{nT}] \) (Eq. (36)), from which we can obtain
\[
\text{Cov}[E_n, T_k] = n \mathbb{E}[e_{nT}] = \mathbb{E}[E_n] \mathbb{E}[T_k]. \tag{41}
\]

Applying Eqs. (13) and (2), we have
\[
\text{Cov}[E_n, T_k] = \frac{4}{k(k-1)}. \tag{42}
\]

Irrespective of the value of \( k \), we have
\[
\lim_{n \to \infty} \text{Cov}[E_n, T_k] = 0. \tag{43}
\]

3.6. \( L_n \) and \( E_n \)

Fu and Li (1993) provided the expression for \( \mathbb{E}[L_n E_n] \) (see also p. 167 of Durrett (2008)), with all values scaled by \( 2N_e \). The main result is the following expression, obtained by solving recurrence equations:
\[
\mathbb{E}[L_n E_n] = \frac{4S_{1,n-1}n}{n-1}. \tag{44}
\]
We can use this result to calculate the covariance of \( L_n \) and \( E_n \) by
\[
\text{Cov}[L_n, E_n] = \mathbb{E}[L_n E_n] - \mathbb{E}[L_n] \mathbb{E}[E_n]
\]
with Eqs. (8) and (13). The covariance can also be quickly obtained from Eqs. (7) and (41),
\[
\text{Cov}[L_n, E_n] = \sum_{k=2}^{n} k \text{Cov}[E_n, T_{k}] = \frac{4S_{1,n-1}}{n - 1}.
\]
(45)
The limit is
\[
\lim_{n \to \infty} \text{Cov}[L_n, E_n] = 0.
\]
(46)
Applying Eqs. (45), (9) and (14), the correlation coefficient of \( L_n \) and \( E_n \) is
\[
\text{Corr}[L_n, E_n] = \frac{1}{\sqrt{2S_{2,n-1}(n-1)(n-2) + 4(n-2)}} = \frac{1}{\sqrt{2S_{2,n-1}(n-1)(n-2) + 4(n-2)}}
\]
(47)
with the limit
\[
\lim_{n \to \infty} \text{Corr}[L_n, E_n] = 0.
\]
(48)

3.7. \( H_n \) and \( I_n \)

For the pair \( H_n \) and \( I_n \), we exploit results obtained for other pairs to quickly obtain the covariance. Remembering that \( I_n = L_n - E_n \) (Eq. (15)), we use Eqs. (29) and (37) to obtain for \( n \geq 3 \)
\[
\text{Cov}[H_n, I_n] = \text{Cov}[H_n, I_n] - \text{Cov}[H_n, E_n] = 4S_{2,n-1} - 4.
\]
(49)
For this covariance, we have
\[
\lim_{n \to \infty} \text{Cov}[H_n, I_n] = \frac{2n^2}{3} - 4 \approx 2.57974.
\]
(50)
From the covariance in Eq. (49) and variances in Eqs. (6) and (17), we compute the correlation coefficient:
\[
\text{Corr}[H_n, I_n] = \frac{\pi^2 - 6}{\pi \sqrt{2(n^2 - 9)}} \approx 0.93399.
\]
(52)

3.8. \( I_n \) and \( T_k \)

By Eqs. (15), (26) and (41), assuming \( n \geq 3 \), we have
\[
\text{Cov}[I_n, T_k] = \text{Cov}[L_n, T_k] - \text{Cov}[E_n, T_k] = \frac{4(n-k)}{k(k-1)^2}.
\]
(53)
The limit of this expression is a rapidly decreasing function of \( k \),
\[
\lim_{n \to \infty} \text{Cov}[I_n, T_k] = \frac{4}{k(k-1)^2}.
\]
(54)
Using the variances in Eqs. (17) and (3), the correlation coefficient is
\[
\text{Corr}[I_n, T_k] = \frac{(n-k)\sqrt{n-2}}{(k-1)\sqrt{n-1}} \left( 4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1) \right)
\]
(55)
with limit
\[
\lim_{n \to \infty} \text{Corr}[I_n, T_k] = \frac{\sqrt{6}}{\pi} \frac{1}{k-1}.
\]
(56)
The limit of \( \text{Corr}[I_n, T_k] \) as \( n \to \infty \) is equal to that of \( \text{Corr}[L_n, T_k] \) (Eq. (28)).

3.9. \( L_n \) and \( I_n \)

By Eq. (15), we can apply Eqs. (9) and (45) to obtain for \( n \geq 3 \)
\[
\text{Cov}[I_n, I_n] = \text{Var}[I_n] - \text{Cov}[I_n, E_n] = 4S_{2,n-1} - 4S_{1,n-1}.
\]
(57)
The limit as \( n \to \infty \) is
\[
\lim_{n \to \infty} \text{Cov}[I_n, I_n] = \frac{2n^2}{3} \approx 6.57974.
\]
(58)
For the correlation coefficient, applying Eqs. (57), (9) and (17), we get
\[
\text{Corr}[I_n, I_n] = \frac{[S_{2,n-1}(n-1) - S_{1,n-1}]\sqrt{n-2}}{(n-1)\sqrt{S_{2,n-1}(n-1)(n-2) - 4(n-1)}}.
\]
(59)
with
\[
\lim_{n \to \infty} \text{Corr}[I_n, I_n] = 1.
\]
(60)

3.10. \( E_n \) and \( I_n \)

For this pair, with \( n \geq 3 \), the covariance was reported by Fu and Li (1993):
\[
\text{Cov}[E_n, I_n] = \frac{4S_{1,n-1}}{n-1} - \frac{8S_{1,n-1}}{(n-1)(n-2)} + \frac{16}{n-2}.
\]
(61)
We can also obtain this result quickly from Eqs. (15), (57) and (17), as \( \text{Cov}[E_n, I_n] = \text{Cov}[L_n - E_n, I_n] = \text{Cov}[L_n, I_n] - \text{Var}[I_n] \). In the limit, we have
\[
\lim_{n \to \infty} \text{Cov}[E_n, I_n] = 0.
\]
(62)
For the correlation coefficient, we divide Eq. (61) by the product of the square roots of Eqs. (14) and (17):
\[
\text{Corr}[E_n, I_n] = \frac{4(n-1) - S_{1,n-1}(n-1)}{\sqrt{2(S_{1,n-1}(n-2) + 2)(4S_{1,n-1} + S_{2,n-1}(n-1)(n-2) - 4(n-1))}}.
\]
(63)
with the limit
\[
\lim_{n \to \infty} \text{Corr}[E_n, I_n] = 0.
\]
(64)
This result is equal to the limit for \( \text{Corr}[L_n, E_n] \) (Eq. (48)).

3.11. \( \text{Var} [B_n] \)

To obtain correlation coefficients involving \( B_n \), assuming \( n \geq 4 \), we first verify the expression for \( \text{Var} [B_n] \) in Eq. (22). By definition of \( B_n \) in Eq. (18), we have
\[
\text{Var}[B_n] = \mathbb{E}\left[ \frac{1}{4}(T_2 + b_n)^2 - \mathbb{E}\left( \frac{1}{2}(T_2 + b_n)^2 \right)^2 \right] = \frac{1}{4} \text{Var}[b_n] + \frac{1}{2} \text{Cov}[b_n, T_2] + \frac{1}{4},
\]
(65)
where we have used \( \text{Var}[T_2] = 1 \) (Eq. (3)).
By Eq. (67), recalling \( n \), the covariance is independent of \( \mathbb{E} \). Aliyev and N.A. Rosenberg \ Theoretical Population Biology 143 (2022) 1–13

Next, we compute \( \text{Cov} \{ b_n, T_1 \} \) and insert \( k = 2 \). By Eq. (19), applying the independence of the \( T_i \) (Section 2.1) and inserting Eq. (3), we have

\[
\text{Cov} \{ b_n, T_k \} = \sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \text{Var} \{ T_i, T_k \} + \left( \sum_{j=3}^{n} \frac{2}{n-1} \text{Cov} \{ T_i, T_k \} \right) = \sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \text{Cov} \{ T_i, T_k \} + \left( \sum_{j=3}^{n} \frac{2}{n-1} \text{Var} \{ T_i \} \right)
\]

Inserting \( \text{Var} \{ b_n \} \) from Eq. (66) and \( \text{Cov} \{ b_n, T_k \} \) from Eq. (67) into Eq. (65), we confirm Eq. (22).

**Box 1.**

To calculate \( \text{Var} \{ b_n \} \), we first recall that for \( j = 3, 4, \ldots, n-1 \), with probability \( 2/(j(j-1)) \), we have \( b_n = H_j \); with probability \( 2/(n-1) \) we have \( b_n = H_n \). Hence, applying Eq. (19) and \( \mathbb{E} \{ H_j \} = \text{Var} \{ H_j \} + \mathbb{E} \{ H_j \}^2 \) with Eqs. (5) and (6), we have

\[
\mathbb{E} \{ b_n \}^2 = \sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \mathbb{E} \{ H_j \}^2 + \frac{2}{n-2} \mathbb{E} \{ H_n \}^2 = -16 \mathbb{E} \{ b_n \}^2 + 30n^2 - 16n - 16
\]

Using the expression for \( \mathbb{E} \{ b_n \} \) from Eq. (20), we obtain

\[
\text{Var} \{ b_n \} = \frac{24S_{2,n-1} - 16S_{2,n-1}^2 + 5n^2 - 32S_{2,n-1}n + 24n - 32}{n^2}.
\]

(66)

The covariance is independent of \( n \). For the correlation coefficient, using Eqs. (68), (22) and (3), we have

\[
\text{Cov} \{ b_n, T_k \} = \sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \text{Var} \{ T_i, T_k \} + \left( \sum_{j=3}^{n} \frac{2}{n-1} \text{Cov} \{ T_i, T_k \} \right) = \sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \text{Var} \{ T_i \} \quad k = 2,
\]

\[
\sum_{j=3}^{n-1} \sum_{l=2}^{n} \frac{2}{j(j-1)} \text{Cov} \{ T_i \}, \quad k = 3, 4, \ldots, n-1,
\]

\[
\frac{2}{n-1} \text{Var} \{ T_i \} \quad k = n,
\]

\[
1, \quad k = 2, \quad \frac{\pi}{\pi + 1}, \quad k = 3, 4, \ldots, n.
\]

(67)

3.12. \( b_n \) and \( T_k \)

We extract \( \text{Cov} \{ b_n, T_k \} \) from Section 3.11, as \( \text{Cov} \{ b_n, T_k \} = \text{Cov} \{ b_n, T_k \} \times \text{Cov} \{ T_2, T_k \}/2 \times \text{Cov} \{ T_2, T_k \} \), where \( \delta \) is the Kronecker delta (Section 2.1). By Eq. (67), recalling \( n \geq 4 \),

\[
\text{Cov} \{ b_n, T_k \} = \frac{4}{k(k-1)}.
\]

(68)

The asymptotic limit of \( \text{Cov} \{ b_n, T_k \} \) is

\[
\lim_{n \to \infty} \text{Cov} \{ b_n, T_k \} = \frac{6}{18 + 9\pi^2 - \pi^4} \frac{1}{k(k-1)}.
\]

(70)

The second term was computed in Eq. (23). For the first term, \( \text{Cov} \{ b_n, T_k \} \) decomposes into Eq. (67). The limit begins as \( 3\sqrt{18 + 9\pi^2 - \pi^4} \approx 0.97759 \) for \( k = 2 \) and quickly decreases to 0 as \( k \) increases.

3.13. \( H_n \) and \( B_n \)

To obtain \( \text{Cov} \{ H_n, B_n \} \) with \( n \geq 4 \), we begin from Eq. (18):

\[
\text{Cov} \{ H_n, B_n \} = \frac{1}{2} \text{Cov} \{ H_n, b_n \} + \frac{1}{2} \text{Cov} \{ H_n, T_2 \}.
\]

The asymptotic limit of \( \text{Cov} \{ H_n, B_n \} \) is

\[
\lim_{n \to \infty} \text{Cov} \{ H_n, B_n \} = 4 \psi(3) + 16 - 2\pi^2 \approx 1.06902.
\]

The correlation coefficient is then equal to Eq. (73) given in Box 1.

The limit of the correlation coefficient is:

\[
\lim_{n \to \infty} \frac{3\sqrt{3}[2\psi(3) + 8 - \pi^2]}{\sqrt{(\pi^2 - 9)(18 + 9\pi^2 - \pi^4)}} \approx 0.97054.
\]

(74)

3.14. \( L_n \) and \( B_n \)

In a manner similar to that used in Section 3.13, with \( n \geq 4 \), we expand \( \text{Cov} \{ L_n, B_n \} \) using Eq. (18):

\[
\text{Cov} \{ L_n, B_n \} = \frac{1}{2} \text{Cov} \{ L_n, b_n \} + \frac{1}{2} \text{Cov} \{ L_n, T_2 \}.
\]

The first term is \( \text{Cov} \{ L_n, T_2 \} = 2 \) by Eq. (26). The second term is decomposable by Eq. (7); applying Eq. (67),

\[
\text{Cov} \{ L_n, b_n \} = \sum_{k=2}^{n} k \text{Cov} \{ b_n, T_k \} = 2 + \sum_{k=3}^{n} \frac{8}{k(k-1)}.
\]

(71)
\[ \tilde{\text{Corr}}[E_n, B_n] = \frac{(S_{2,n-1}n - n + 1)\sqrt{n - 2}}{\sqrt{(n - 1)(S_{1,n-1}n - 2n + 2)(3S_{2,n-1}n^2 - 2S_{2,n-1}n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \]  

Box II.

**3.15. \( E_n \) and \( B_n \)**

For \( \text{Cov}[E_n, B_n] \), we obtain an approximate ratio rather than exact answer. Decomposing \( B_n \) by Eq. (18), we have

\[ \text{Cov}[E_n, B_n] = \frac{1}{2} \text{Cov}[E_n, T_2] + \frac{1}{2} \text{Cov}[E_n, \text{b}_n]. \]  

We next perform an approximation by ignoring the second term in the covariance decomposition. Noting that \( \text{Cov}[E_n, T_2] = \frac{n}{n-1} \) by Eq. (41), we use Eq. (79) together with Eq. (80) to write approximations

\[ \tilde{\text{Cov}}[E_n, b_n] = \frac{1}{n-1} + \frac{2}{n} \text{Cov}[E_n, b_n]. \]

Weigthing each \( \text{Cov}[H_j, E_n] \) by the associated probability \( p_j \), and decomposing \( H_j \) by Eq. (4), Eq. (81) gives

\[ \mathbb{E}\left[ \text{Cov}[E_n, b_n|J] \right] = \sum_{j=1}^n p_j \text{Cov}[E_n, b_n|J = j] \]

\[ = \frac{2}{n-1} \sum_{j=1}^n \text{Cov}[H_j, E_n]. \]

Summing the series, we have

\[ \text{Cov}[L_n, B_n] = 2 + \sum_{k=3}^{n} \frac{4}{k(k-1)^3} \sqrt{\frac{n}{n - 1}} \]

\[ = \frac{4(S_{3,n-1}n - S_{2,n-1}n + n - 1)}{\sqrt{S_{2,n-1}(3S_{2,n-1}n^2 - 2S_{2,n-1}n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}. \]

The limit is

\[ \lim_{n \to \infty} \text{Cov}[L_n, B_n] = \frac{\sqrt{6}(3\pi + 6 - \pi^2)}{\pi \sqrt{18 + 9\pi^2 - \pi^4}} \approx 0.84930. \]

3.16. \( I_n \) and \( B_n \)

We use Eq. (15) and results involving \( L_n \) (Eq. (75)) and \( E_n \) (Eq. (84)) to obtain

\[ \text{Cov}[I_n, B_n] = \text{Cov}[L_n, B_n] - \text{Cov}[E_n, B_n] \]

\[ = \frac{4(S_{3,n-1}n - S_{2,n-1}n + n - 5S_{3,n-1} - 1)}{n-1}. \]

Finally, inserting Eq. (83) into Eq. (82), we have

\[ \tilde{\text{Corr}}[E_n, B_n] = \frac{2(4S_{2,n-1}n - 5n + 4)}{n(n - 1)}. \]

The limit is

\[ \lim_{n \to \infty} \tilde{\text{Corr}}[E_n, B_n] = 0. \]

For the approximate correlation coefficient \( \tilde{\text{Corr}}[E_n, B_n] \), we use Eqs. (84), (14) and (22) to obtain Eq. (86) given in Box II, with limit

\[ \lim_{n \to \infty} \tilde{\text{Corr}}[E_n, B_n] = 0. \]

4. Numerical and simulation-based analysis

4.1. Analysis methods

We examine the results of Section 3 summarized in Tables 1 and 2 numerically and by coalescent simulation. For 13 of 15 covariances and correlations, the theoretical results are exact, and simulations merely verify that the mathematics has proceeded without error. For the covariances and correlations involving \( (E_n, B_n) \) and \( (I_n, B_n) \), the theoretical results are approximate, and the simulations assess the accuracy of the approximations.

We simulated the coalescent process for a series of values of \( n \) beginning with \( n = 2 \), at each value of \( n \) performing 100,000
replicate simulations. To generate the simulated replicates, we employed ms (Hudson, 2002), using the command `ms n 100000 -T`, with n taken from \{2, 3, \ldots, 50\}. In the set of simulated replicates, we evaluated simulated covariances and correlation coefficients for pairs of quantities.

We plot the mathematical results of Tables 1 and 2 together with simulation values in Figs. 2–5. Figs. 2 and 3 show covariances of pairs of variables; Figs. 4 and 5 show correlations.

4.2. Accuracy of approximations

Fig. 2 shows the analytical and simulated covariances, and Fig. 4 shows the analytical and simulated correlations, for pairs of variables among \{H_n, L_n, I_n, E_n, B_n\}. For pairs of variables for which no approximations were needed in obtaining covariances— all except \((E_n, B_n)\) and \((I_n, B_n)\)—the simulated and analytical values produce plots that are nearly indistinguishable.

\[
\widetilde{\text{Corr}}[I_n, B_n] = \frac{\sqrt{2}(S_{1,n-1}n - S_{2,n-1}n + n - S_{3,n-1} - 1)n\sqrt{n - 2}}{\sqrt{(n - 1)(4S_{1,n-1} + S_{2,n-1}(n - 1)(n - 2) - 4(n - 1))(3S_{2,n-1}n^2 - 2S_{2,n-1}n^2 + n^2 - 4S_{2,n-1}n + 3n - 4)}}.
\]
For \((E_n, B_n)\) and \((I_n, B_n)\), the approximate and simulated correlations are close, but noticeably different (Fig. 4); the mean absolute difference between the analytical and simulated values across choices of \(n\) from 4 to 30 is 0.02458 for \((E_n, B_n)\) and 0.01089 for \((I_n, B_n)\). For covariance, which unlike the correlation coefficient is not standardized to lie in \([-1, 1]\), the approximate and simulated values are quite close, with corresponding mean absolute deviations of 0.02372 for \((E_n, B_n)\) and 0.03101 for \((I_n, B_n)\).

4.3. Properties of correlations

We observe that \(H_n, L_n, I_n\), and \(B_n\) all remain strongly correlated as \(n\) increases, with the six limiting correlations among these four quantities lying between 0.84930 for \(\text{Corr}[L_n, B_n]\) and \(\text{Corr}[I_n, B_n]\) (Table 2). The high limiting \(\text{Corr}[H_n, L_n]\) of approximately 0.93399 reflects the strong influence of times \(T_2\) with small \(k\) on both \(H_n\) and \(L_n\) (Figs. 3 and 5). As \(n\) increases, \(\text{Var}[I_n]\) increases without bound (Eq. (16)), whereas \(\text{Var}[E_n]\) remains constant (Eq. (13)); the contribution of \(E_n\) to the total tree length \(L_n\) becomes negligible, and \(\text{Corr}[L_n, I_n]\) approaches 1. \(\text{Corr}[H_n, I_n]\) has the same limiting value as \(\text{Corr}[H_n, L_n]\), and \(H_n, L_n\), and \(I_n\) all have limiting correlation 0 with \(E_n\). Interestingly, although \(H_n\) and \(E_n\) have the same limiting expectation of 2, the limit of their correlation \(\text{Corr}[H_n, E_n]\) is 0.

The correlations of \(H_n, L_n, I_n\) with \(B_n\) like their correlations with each other, are relatively high. \(\text{Corr}[H_n, B_n]\) is nearly constant in \(n\), with limit approximately 0.97054; both \(H_n\) and \(B_n\) are determined in large part by the \(T_k\) with small \(k\) [Eqs. (5) and (21)], so that little change occurs in the correlation as \(n\) increases. Because \(\text{Corr}[H_n, B_n]\) is high and \(\text{Corr}[H_n, L_n]\) is also high, the constraint on a correlation \(\text{Corr}[Y, Z]\) given \(\text{Corr}[X, Y]\) and \(\text{Corr}[X, Z]\), or (Wickens, 2014, eq. 7.1)

\[
\text{Corr}[Y, Z] \geq \text{Corr}[X, Y] \text{Corr}[X, Z] - \sqrt{1 - \text{Corr}[X, Y]^2} \sqrt{1 - \text{Corr}[X, Z]^2}
\]

(92)

\[
\text{Corr}[Y, Z] \leq \text{Corr}[X, Y] \text{Corr}[X, Z] + \sqrt{1 - \text{Corr}[X, Y]^2} \sqrt{1 - \text{Corr}[X, Z]^2},
\]

(93)

forces a high value for \(\text{Corr}[L_n, B_n]\) as well. In particular, placing \(H_n, L_n, B_n\) in the roles of \(X, Y, Z\), with \(\lim_{n \to \infty} \text{Corr}[H_n, L_n] \approx 0.93399\) and \(\lim_{n \to \infty} \text{Corr}[H_n, B_n] \approx 0.97054\), we obtain an interval \(0.82037 \leq \lim_{n \to \infty} \text{Corr}[L_n, B_n] \leq 0.99256\) from Eqs. (92) and (93); \(\lim_{n \to \infty} \text{Corr}[L_n, B_n] \approx 0.84930\) lies near its lower end. Eqs. (92) and (93) similarly force a high value for \(\lim_{n \to \infty} \text{Corr}[L_n, B_n]\) using \(H_n, I_n, B_n\) as \(X, Y, Z\).

Next, for correlations involving the \(T_k\), we observe that for fixed \(n\), as \(k\) increases from 2 to \(n\), \(\text{Corr}[H_n, T_k]\) decreases (Fig. 5). At fixed \(n\) and \(k\), \(\text{Corr}[L_n, T_k]\) generally exceeds \(\text{Corr}[H_n, T_k]\); \(k\) copies of the branch length \(T_k\) contribute to tree length \(L_n\) (Eq. (7)), whereas only one copy contributes to the tree height \(H_n\) (Eq. (4)), giving rise to a greater value for the correlation of \(T_k\) with \(L_n\) than with \(H_n\). For \(k > 2\), \(\text{Corr}[B_n, T_k]\) is generally smaller than \(\text{Corr}[H_n, T_k]\); because \(B_n\) is determined to a larger extent by \(T_2\) than is \(H_n\), the correlations of \(B_n\) with \(T_k\) for \(k > 2\) are generally smaller. Finally, because tree length \(L_n\) consists primarily of internal branches for large \(n\), the correlation \(\text{Corr}[L_n, T_k]\) is similar to \(\text{Corr}[L_n, T_1]\) (Fig. 5), approaching the same limit as \(n \to \infty\) (Table 2); the correlation of \(E_n\) and \(T_k\) is a constant that does not depend on \(k\).

5. Discussion

We have examined relationships between pairs of tree features under the coalescent model by deriving expressions for their covariances and correlation coefficients (Tables 1 and 2). For 13 of 15 pairs examined, we obtained exact expressions for the covariances and correlation coefficients, and for the remaining two pairs, we obtained quantities observed in simulations to closely approximate the desired quantities (Figs. 2 and 4). The results provide a compendium of basic relationships among
coalescent tree features, contributing to a more precise understanding of the way in which the properties of coalescent trees relate to each other.

In most cases, the covariances have relatively simple expressions, comparable to the simplicity of most expressions for expectations and variances (Table 1). Expressions for the correlation coefficients are somewhat more complex, in many cases with \( n \to \infty \) limits that contain terms resulting from the limit \( \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \).

Numerically, we obtain tight correlations between \( H_n, L_n, I_n, \) and \( B_n \) as \( n \) grows large, with all of these quantities possessing limiting correlations of 0.84930 or greater (Table 2). In the limit, \( L_n \) and \( I_n \) are perfectly correlated, and all limiting correlations of other quantities with \( L_n \) are equal to their corresponding correlations with \( I_n \). Decreasing correlations are observed for \( H_n, L_n, I_n, \) and \( B_n \) with \( E_n \), with limits of 0 observed in all cases (Table 2). Although \( H_n \) and \( E_n \) both have limiting expectation 2 (Eqs. (5) and (13)), their limiting correlation coefficient is 0. The correlations among \( H_n, L_n, \) and \( B_n \) are all large; however, the limiting correlation for \( (L_n, B_n) \) is near the lower end of the interval suggested by the larger limiting correlations for \( (H_n, L_n) \) and \( (H_n, B_n) \) (Eqs. (92) and (93)). This result suggests that \( L_n \) and \( B_n \) capture relatively distinct features of coalescent trees in relation to the constraints placed on a pair of correlated variables that are each highly correlated with a third variable \( (H_n) \). A similar observation can be made concerning \( I_n \) and \( B_n \), as \( L_n \) and \( I_n \) are asymptotically fully correlated.

Although tree properties such as \( H_n, L_n, E_n, I_n, \) and \( B_n \) are not themselves observable in genetic sequences, interest in these quantities arises in part from their relationship to statistical tests that assess the fit of the coalescent model to data on genetic variation. Features of tree shape underlie predictions of the coalescent regarding allele frequencies; in particular, tree properties contribute to predictions for the unfolded site-frequency spectrum (SFS) of a genomic region, the vector that for a sample of size \( n \) tabulates how many variable (biallelic) sites in the region possess allele frequencies \( 1/n, 2/n, \ldots, (n-1)/n \) for the derived allele (e.g. Fu, 1995; Ferretti et al., 2017). Test statistics then assess agreement of site-frequency spectra with the predictions (e.g. Zeng et al., 2006; Achaz, 2009; Ferretti et al.,
Fig. 5. Theoretical values of correlation coefficients Corr[\(X, Y\)] 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 \(n = 50\).

Our computations augment earlier calculations concerning quantities associated with coalescent trees. The pairs \((H_n, L_n)\) (Arbisser et al., 2018) and \((I_n, E_n)\) and \((E_n, I_n)\) (Fu and Li, 1993) have been studied in detail. Results for pairs \((H_n, T_k), (L_n, T_k),\) and \((I_n, L_n)\) follow trivially from the derivations and results of Arbisser et al. (2018) and Fu and Li (1993), but were not highlighted in those studies. Results for pairs \((H_n, E_n), (H_n, I_n), (E_n, T_k),\) and \((I_n, T_k)\) follow from derivations similar to those of Fu and Li (1993), but to our knowledge, they have not been previously reported.

The least-studied of the variables we consider, \(B_n\), was introduced by Uyenoyama (1997) in the context of balancing selection and self-incompatibility alleles in plants. Under balancing selection, the mean \(B_n\) of the two basal branches is expected to be long in relation to the tree length \(L_n\), so that \(2B_n/L_n\) predicts the fraction of segregating sites that distinguish two long-separated sets of lineages. For \(B_n\), which gives a portion of the height \(H_n\)—but which, unlike \(H_n\), is obtained from a sum with a random length—we derived the variance (Eq. (22)), as well as exact covariances and correlations with \(H_n, L_n,\) and \(T_k\) and approximate covariances and correlations with \(E_n\) and \(I_n\). Several studies have extended the work of Fu and Li (1993) on features of the external and internal branch lengths (Blum and François, 2005; Caliebe et al., 2007; Janssen and Kesting, 2011; Dahmer and Kesting, 2015, 2017; Disanto and Wiehe, 2020); it may be possible to seek exact rather than approximate covariances and correlations for \((E_n, B_n)\) and \((I_n, B_n)\) by building on these studies.

When examining joint distributions of \(H_n\) and \(L_n\), Arbisser et al. (2018) used computations of the expectations and variances of \(H_n\) and \(L_n\) and the covariance of \(H_n\) and \(L_n\) to obtain approximations for the expectation and variance of \(H_n/L_n\). Following the approach of Arbisser et al. (2018), our results could be used to obtain similar approximate expressions for expectations and variances of ratios of additional pairs.

Acknowledgments

We thank Yun-Xin Fu and John Wakeley for comments on the manuscript. We acknowledge support from NIH, United States grants R01 GM131404 and R01 HG005855 and NSF, United States grant BCS-2116322.

References


