Supporting Information

Mehta et al. 10.1073/pnas.1601074113

Reduction to the Two-Species Case from Ref. 4

This appendix shows that our recursive Eq. **2** reduces properly to the two-taxon results of Rosenberg (4). Rosenberg (4) studied four monophyly events $C1$, $C2$, $C3$, and $C4$, which in our notation satisfy $C1 \cup C2 = E_S$ and $C1 = E_{SC}$. We use our formula to compute the probabilities of monophyly of S and reciprocal monophyly with the settings from ref. 4, showing that we obtain the same results.

In our notation, the root node of a two-taxon tree is x, and the leaves are x_L and x_R. The species tree initial conditions $\mathscr{T}_{SC}$ from ref. 4 are $\mathbf{n}^I_{x_L} = (r_A, 0{,}0)$ and $\mathbf{n}^I_{x_R} = (0, r_B, 0)$, with $T_{x_L} = T_A$, $T_{x_R} = T_B$, and $r_A, r_B \geq 1$. The outputs of the leaves are $s^L_x = s^O_{x_L} = q_A$ and $c^R_x = c^O_{x_R} = q_B$.

For the probability of monophyly of S, applying the initial conditions in Eq. **2** yields:

$$\begin{aligned}\mathbb{P}\big(\mathbf{Z}_x = (0{,}0{,}1), E^x_S | \mathscr{T}^x_{SC}\big) &= \sum_{q_A=0}^{r_A} \sum_{q_B=0}^{r_B} \mathbb{P}\big(\mathbf{Z}_{x_L} = (q_A, 0{,}0), E^{x_L}_S | \mathscr{T}^{x_L}_{SC}\big) \\ &\times \mathbb{P}\big(\mathbf{Z}_{x_R} = (0, q_B, 0), E^{x_R}_S | \mathscr{T}^{x_R}_{SC}\big) g_{q_A+q_B,1}(T_x) K_S. \end{aligned} \quad \textbf{[S1]}$$

We verify that this equation accords with the sum of equations 14 and 15 of ref. 4, representing $\mathbb{P}(C1 \cup C2)$. Neither $q_A = 0$ nor $q_B = 0$ is possible because neither r_A nor r_B is 0; we can ignore the 0 summation indices in Eq. **S1**, and the limits of summation therefore agree with ref. 4.

For the combinatorial term K_S in Eq. **S1**, the only possibility is case 2 in Eq. **4**:

$$K_S = \frac{\sum_{k=1}^{q_B} I_{q_A,1} I_{q_B,k} W_2(q_A - 1, q_B - k) k I_{k,1}}{I_{q_A+q_B,1}}. \quad \textbf{[S2]}$$

The summand is equivalent to the quantity in the line above equation 10 of ref. 4, which provides an intermediate step in computing the probability of monophyly of S. Expression S2 therefore accords with the corresponding equation 11 of ref. 4. Note that the line above equation 10 of ref. 4 contains a known typographical error, with H_{k-1} written in place of the correct H_k; this error, which is corrected by our Eq. **S2**, did not produce an error in the numbered equation 10 of ref. 4.

The next step is to verify that the probability terms in Eq. **S1** for the left and right species tree leaves agree with ref. 4. For the left leaf, our definitions of probabilities for leaves force the summation to have only one nonzero term, with input probability 1. Because only input S lineages are present, case 1e for K_S (Eq. **4**) applies, so that $\mathbb{P}(\mathbf{Z}_{x_L} = (q_A, 0{,}0), E^{x_L}_S | \mathscr{T}^{x_L}_{SC}) = F((r_A, 0{,}0), (q_A, 0{,}0), E^{x_L}_S | \mathscr{T}^{x_L}_{SC}) = g_{r_A,q_A}(T_A) K_S = g_{r_A,q_A}(T_A)$. Analogously, for the right leaf, $\mathbb{P}(\mathbf{Z}_{x_R} = (0, q_B, 0), E^{x_R}_S | \mathscr{T}^{x_R}_{SC}) = g_{r_B,q_B}(T_B)$. These two results accord with equations 14 and 15 of ref. 4.

The recursion terminates at the leaves. Because node x is the root, $T_x = \infty$, and $g_{q_A+q_B,1}(T_x) = 1$ in Eq. **S1**. Thus, the probability of E_S, or in the notation of ref. 4, $\mathbb{P}(C1 \cup C2)$, is

$$\begin{aligned}\mathbb{P}\big(\mathbf{Z}_x = (0{,}0{,}1), E^x_S | \mathscr{T}^x_{SC}\big) &= \sum_{q_A=1}^{r_A} \sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \\ &\times \frac{\sum_{k=1}^{q_B} I_{q_A,1} I_{q_B,k} W_2(q_A - 1, q_B - k) k I_{k,1}}{I_{q_A+q_B,1}}. \end{aligned} \quad \textbf{[S3]}$$

For the probability of E_{SC}, or in the notation of ref. 4, $\mathbb{P}(C1)$, we use the same process, but with K_{SC} instead of K_S, obtaining

$$\begin{aligned}\mathbb{P}\big(\mathbf{Z}_x = (0{,}0{,}1), E^x_{SC} | \mathscr{T}^x_{SC}\big) &= \sum_{q_A=1}^{r_A} \sum_{q_B=1}^{r_B} g_{r_A,q_A}(T_A) g_{r_B,q_B}(T_B) \\ &\times \frac{I_{q_A,1} I_{q_B,1} W_2(q_A - 1, q_B - 1)}{I_{q_A+q_B,1}}. \end{aligned} \quad \textbf{[S4]}$$

This equation simplifies to equation 14 ref. 4. Taking the difference of our results in Eqs. **S3** and **S4** and simplifying produces equation 15 of ref. 4, confirming agreement of our formulas with those of ref. 4.

Probabilities in the Relative-Branch-Length Scenario

The four cases of Fig. 3, representing different distributions across species of lineages in lineage classes S and C, illustrate different effects of the tree height T and relative-branch-length parameter r. These differing effects can be explained by considering the way in which likely coalescence patterns for the sampled lineages differ as a function of the locations of those lineages.

In Fig. 3*B*, with $(S_1, S_2, S_3) = (2{,}0{,}2)$ and $(C_1, C_2, C_3) = (2{,}2{,}2)$, we observe similar monotonic decreases in monophyly probability as a function of tree height for all values of r. With T fixed, r modulates the time during which the $S_1 = 2$ lineages might coalesce with the $C_2 = 2$ lineages before reaching the root: for larger r, a monophyly-violating coalescence is more likely. Because the S_1 lineages already have the possibility of coalescing with the $C_1 = 2$ lineages before reaching the root, however, r is less important than T in predicting the monophyly probability. As T increases, because the minimal subtree with respect to S is the full species tree—containing C lineages but not occurring at a leaf—the probability approaches 0.

For Fig. 3*C*, $(S_1, S_2, S_3) = (2{,}0{,}2)$ and $(C_1, C_2, C_3) = (0{,}2{,}0)$. Here, r controls whether a coalescence violating E_S can happen before the root at all. For $r = 0$, E_S is determined above the root. As r increases at fixed T, the $S_1 = 2$ and $C_2 = 2$ lineages coexist longer before reaching the root, and the monophyly probability decreases. Again, because the minimal subtree with respect to S is the full species tree, for $r > 0$, the probability nears 0 as $T \to \infty$. For $r = 0$, it approaches 1/3, the monophyly probability for two lineages in a set of three. The shape of the probability function in terms of T differs with r: it increases monotonically at $r = 0$, decreases monotonically at $r = 1$, and is not monotonic for intermediate r.

In Fig. 3*D*, with $(S_1, S_2, S_3) = (2{,}2{,}0)$ and $(C_1, C_2, C_3) = (2{,}0{,}2)$, as r increases with T fixed, the time before the $S_1 = 2$ and $C_1 = 2$ lineages can coalesce with the $S_2 = 2$ lineages decreases. Thus, the monophyly probability increases with r. For $r < 1$, because the minimal subtree with respect to S lies at the MRCA for the sister species pair, the probability approaches 0 as $T \to \infty$; for $r = 1$, it approaches 1/25, the monophyly probability for four lineages among six (equation 11 in ref. 4). It is monotonic in T only for $r = 1$, for which it increases.

Finally, in Fig. 3*E*, we set $(S_1, S_2, S_3) = (2{,}2{,}0)$ and $(C_1, C_2, C_3) = (2{,}2{,}2)$. For fixed T, increasing r increases the time during which the $S_1 = 2$ and $S_2 = 2$ lineages might coalesce with each other, so that the monophyly probability increases. For $r < 1$, the probability again nears 0 as $T \to \infty$; for $r = 1$, it approaches 2/175, the monophyly probability for four lineages among eight. A monotonic decrease is observed for low r and a monotonic increase is observed for $r = 1$; a change in monotonicity occurs as r increases from 0 to 1.

Pairs of Maize and Teosinte Lineages

The eight-lineage subsamples of maize and teosinte lineages all contained the only two mexicana individuals in the ref. 47 dataset; one of four pairs of parviglumis individuals: {TIL07, TIL09}, {TIL10, TIL17}, {TIL01, TIL11}, and {TIL03, TIL14}; either {MR12, MR20}, any two individuals from {MR03, MR23, MR21, MR18, MR06}, or any two from {MR05, MR09, MR24, MR26, MR01}, for a total of 21 possible landrace pairs; and two individuals chosen from the pairs {IL14H, P39}, {KY21, M162W}, {CML103, TX303}, {CML247, CML322}, any two from {CAU178, OH78, MS71, B97, W22, W64A, CAUMO17, MO17, OH43, B73, CAUZHENG58, CAU478, CAU5003, CML333, CML52}, or any two from {NC350, NC358, CML69, KI11, CML228, KI3}, for a total of 124 possible improved pairs.

The outlier samples all contain the pair {CAUMO17, MO17} (improved) or the pair {TIL03, TIL14} (parviglumis), two recently coalescing pairs for which the model species tree least adequately reflects the original species tree of ref. 47. In the case of parviglumis, one of the four pairs produces substantially different results from the others, and for convenience, we regard it as an outlier.

Numerical Implementation

In our numerical implementation, although a leaf has no input nodes, without loss of generality, we let all its inputs "enter" from the left. To reduce numerical challenges, we use a binomial coefficient representation that avoids large numerators and denominators:

$$\binom{a}{b} = \prod_{i=0}^{b-1}\left(1 + \frac{a-b}{b-i}\right).$$

The function $g_{n,j}(T)$ (equation 6.1 in ref. 45) is implemented using binomial coefficients as:

$$g_{n,j}(T) = \sum_{k=j}^{n} e^{\frac{-k(k-1)T}{2}}(-1)^{k-j}\binom{2k-1}{n+k-1} \times \frac{\binom{j+k-2}{j}}{\binom{n+k-2}{n}}\binom{n-1}{k-1}\binom{k-1}{j-1}.$$

We express case 2 from Eq. **4** using binomial coefficients as

$$K_S = 2\binom{s_1+c_1}{s_1}^{-1}\sum_{k=c_2+1}^{c_1}\binom{c_1-1}{k-1}\binom{s_1+c_1-1}{k}^{-1}.$$

For case 3 from Eq. **4**, we have

$$K_S = \binom{s_2+c_2}{s_2}\binom{c_1-1}{c_2-1}\binom{s_1-1}{s_2-1}\binom{s_1+c_1}{s_1}^{-1}\binom{s_1+c_1-1}{s_2+c_2-1}^{-1}.$$

Finally, for case 2 from Eq. **5**,

$$K_{SC} = 2\binom{s_1+c_1}{s_1}^{-1}(s_1 + c_1 - 1)^{-1}.$$

Note that this representation accords with equation 9 in ref. 4.

Table S1. Summary of notation

Notation	Meaning
$\mathscr{T}$	Species tree: topology and branch lengths
ℓ	Number of leaves of the species tree
S	Subsampled lineage class
C	Complement (non *S*) lineage class
M	Mixed lineage class
S_i, C_i	Number of *S* or *C* lineages for the *i*th leaf of the species tree
$\mathscr{T}_{SC}$	Initialized species tree
$\mathscr{T}^*_{SC}$	Minimal subtree with respect to *S*
E_i	Monophyly event (Table 2)
x	Species tree node or its corresponding branch
x_L, x_R	Node or corresponding branch directly below node or branch *x* on the left (*L*) or right (*R*)
$\mathbf{Z}(T_x)$	Vector of random variables: the output lineages of classes *S*, *C*, and *M* given time T_x
s^I_x, s^O_x, c^I_x, c^O_x	Number of lineages of class *S* or *C* entering (*I*) or exiting (*O*) node *x*
$\mathbf{n}^I_x$, $\mathbf{n}^O_x$	Input and output states of branch *x*
$\mathbf{n}^L_x$, $\mathbf{n}^R_x$	Portion of input states of branch *x* from branches x_L or x_R
$g_{n,j}(T)$	Probability that *n* lineages coalesce to *j* lineages in time *T* (49)
$I_{n,k}$	Number of coalescence sequences in which *n* lineages can coalesce to *k* lineages (4)
$W_2(r_1, r_2)$	Number of ways the coalescences of two disjoint groups of lineages can be ordered, with r_i coalescences occurring for group *i* (50)
$s^{\mathscr{T}}_x$, $c^{\mathscr{T}}_x$	Total number of *S* or *C* lineages extant at node *x*

Table S2. Comparison of theoretical and observed monophyly frequencies in four maize and teosinte groups

Clade	Expected	Observed Mean	Observed SD	95% CI
Improved	0.129	0.140	0.0380	[0.133, 0.147]
Improved (99 samples)	0.129	0.137	0.0256	[0.132, 0.142]
Landraces	0.129	0.138	0.0237	[0.133, 0.142]
Parviglumis	0.131	0.128	0.0327	[0.122, 0.134]
Parviglumis (74 samples)	0.131	0.109	0.0045	[0.108, 0.110]
Mexicana	0.133	0.121	0.0079	[0.119, 0.122]

Observed means and SDs are computed over 100 samples except where outliers are removed as noted. For the improved lines with outliers excluded, frequencies exceeding 0.4 are excluded (1 sample), and for parviglumis, frequencies exceeding 0.17 are excluded (26 samples). CI, confidence interval.