The probability of monophyly of a sample of gene lineages on a species tree

Rohan S. Mehta[a,1], David Bryant[b], and Noah A. Rosenberg[a]

[a]Department of Biology, Stanford University, Stanford, CA 94305; and [b]Department of Mathematics and Statistics, University of Otago, Dunedin 9054, New Zealand

Edited by John C. Avise, University of California, Irvine, CA, and approved April 18, 2016 (received for review February 5, 2016)

Monophyletic groups—groups that consist of all of the descendants of a most recent common ancestor—arise naturally as a consequence of descent processes that result in meaningful distinctions between organisms. Aspects of monophyly are therefore central to fields that examine and use genealogical descent. In particular, studies in conservation genetics, phylogeography, population genetics, species delimitation, and systematics can all make use of mathematical predictions under evolutionary models about features of monophyly. One important calculation, the probability that a set of gene lineages is monophyletic under a two-species neutral coalescent model, has been used in many studies. Here, we extend this calculation for a species tree model that contains arbitrarily many species. We study the effects of species tree topology and branch lengths on the monophyly probability. These analyses reveal new behavior, including the maintenance of nontrivial monophyly probabilities for gene lineage samples that span multiple species and even for lineages that do not derive from a monophyletic species group. We illustrate the mathematical results using an example application to data from maize and teosinte.

coalescent theory | monophyly | phylogeography

Mathematical computations under coalescent models have been central in developing a modern view of the descent of gene lineages along the branches of species phylogenies. Since early in the development of coalescent theory and phylogeography, coalescent formulas and related simulations have contributed to a probabilistic understanding of the shapes of multispecies gene trees (1–3), enabling novel predictions about gene tree shapes under evolutionary hypotheses (4, 5), new ways of testing hypotheses about gene tree discordances (6, 7), and new algorithms for problems of species tree inference (8, 9) and species delimitation (10, 11). A "multispecies coalescent" model, in which coalescent processes on separate species tree branches merge back in time as species reach a common ancestor (12), has become a key tool for theoretical predictions, simulation design, and evaluation of inference methods, and as a null model for data analysis.

A fundamental concept in genealogical studies is that of monophyly. In a genealogy, a group that is monophyletic consists of all of the descendants of its most recent common ancestor (MRCA): every lineage in the group—and no lineage outside it—descends from this ancestor. Backward in time, a monophyletic group has all of its lineages coalesce with each other before any coalesces with a lineage from outside the group.

The phylogenetic and phylogeographic importance of monophyly traces to the fact that monophyly enables a natural definition of a genealogical unit. Such a unit can describe a distinctive set of organisms that differs from other groups of organisms in ways that are evolutionarily meaningful. Species can be delimited by characters present in every member of a species and absent outside the species, and that therefore can reflect monophyly (13, 14). In conservation biology, monophyly can be used as a prioritization criterion because groups with many monophyletic loci are likely to possess unique evolutionary features (15). Reciprocal monophyly, in which a set of lineages is divided into two groups that are simultaneously monophyletic, is often used in a genealogical approach to species divergence (16, 17). The proportion of loci that are reciprocally monophyletic is informative about the time since species divergence and can assist in representing the level of differentiation between groups (4, 18).

Many empirical investigations of genealogical phenomena have made use of conceptual and statistical properties of monophyly (19). Comparisons of observed monophyly levels to model predictions have been used to provide information about species divergence times (20, 21). Model-based monophyly computations have been used alongside DNA sequence differences between and within proposed clades to argue for the existence of the clades (22), and tests involving reciprocal monophyly have been used to explain differing phylogeographic patterns across species (23). Comparisons of observed levels of monophyly with the level expected by chance alone (24) have assisted in establishing the distinctiveness of taxonomic groups (25, 26). Loci that conflict with expected monophyly levels have provided signatures of genic roles in species divergences (27–29).

For lineages from two species under a model of population divergence, Rosenberg (4) computed probabilities of four different genealogical shapes: reciprocal monophyly of both species, monophyly of only one of the species, monophyly of only the other species, and monophyly of neither species. The computation permitted arbitrary species divergence times and sample sizes—generalizing earlier small-sample computations (1–3, 30, 31)—and illustrated the transition from the species divergence, when monophyly is unlikely for both species, to long after divergence, when reciprocal monophyly becomes extremely likely. Between these extremes, the species can pass through a period during which monophyly of one species but not the other is the most probable state.

Although this two-species computation has contributed to various insights about empirical monophyly patterns (21–23, 32–34), many scenarios deal with more than two species. Because multispecies monophyly probability computations have been unavailable—except in limited cases with up to four species (4, 35–38)—multispecies studies have been forced to rely on two-species models, restricting attention to species pairs (25, 34, 39) or pooling disparate lineages and disregarding their taxonomic distinctiveness (23, 26).

Here, we derive an extension to the two-species monophyly probability computation, examining arbitrarily many species related by an evolutionary tree. Furthermore, we eliminate the past restriction (4) that the lineages whose monophyly is examined all derive from the same population. This generalization is analogous to the assumption that in computing the probability of a binary evolutionary character (40–42), one or both character states can appear in multiple species. Our approach uses a

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "In the Light of Evolution X: Comparative Phylogeography," held January 8–9, 2016, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/ILE_X_Comparative_Phylogeography.

Author contributions: R.S.M., D.B., and N.A.R. designed research; R.S.M. and N.A.R. performed research; R.S.M. analyzed data; and R.S.M., D.B., and N.A.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: rsmehta@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1601074113/-/DCSupplemental.

pruning algorithm, generalizing the two-species formula in a conceptually similar manner to other recursive coalescent computations on arbitrary trees (9, 40–44).

Like the work of Degnan and Salter (5), which considered probability distributions for gene tree topologies under the multispecies coalescent model, our work generalizes a coalescent computation known only for small trees (4, 35) to arbitrary species trees. We study the dependence of the monophyly probability on the model parameters, providing an understanding of factors that contribute to monophyly in species trees of arbitrary size. Finally, we explore the utility of monophyly probabilities in an application to genomewide data from maize and teosinte.

Results

Model and Notation.

Overview. Consider a rooted binary species tree $\mathscr{T}$ with ℓ leaves and specified topology and branch lengths. For each of the ℓ species represented by leaves of $\mathscr{T}$, a number of sampled lineages is specified. Given a specified partition of the lineages into two subsets, we consider a condition describing whether one, the other, both, or neither of the two subsets of lineages is monophyletic. Our goal is to provide a recursive computation of the probability that the condition is obtained under the multispecies coalescent model. Notation appears in Table S1.

Lineage classes. The initial sampled lineages are partitioned into class S (subset) for lineages within a chosen subset, and class C (complement) for all lineages not included in S. Coalescence between an S lineage and a C lineage produces an M (mixed) lineage. Any coalescence involving an M lineage also produces an M lineage. Coalescences between two S or two C lineages produce S and C lineages, respectively (Table 1).

Letting the number of S and C lineages present initially in the ith leaf be S_i and C_i, respectively, the model parameters are S_i and C_i for $1 \le i \le \ell$, and the species tree $\mathscr{T}$. For convenience, we aggregate the S_i and C_i with $\mathscr{T}$ into a parameter collection $\mathscr{T}_{SC}$ that we call the initialized species tree.

Monophyly events. A monophyly event E_i is an assignment of labels to lineage classes S and C. We can choose to label a class "monophyletic" or "not monophyletic," or assign no label at all, so that nine monophyly events are possible, six of which are relevant for our purposes (Table 2). All lineages in a monophyletic class must coalesce within the class to a single lineage before any coalesces outside the class. If multiple classes are labeled monophyletic, then each class must be separately monophyletic.

Species-merging events. We orient the species tree vertically, "up" toward the root and "down" toward the leaves. From a coalescent backward-in-time perspective, at every internal node of the species tree—representing a species-merging event—lineages enter from two branches directly below the node. We label one of these branches "left" and the other "right," based on an arbitrarily labeled diagram of species tree $\mathscr{T}$. These labels are used only for bookkeeping; the labeling does not affect subsequent calculations. Lineages entering from the left and right branches are called "left inputs" and "right inputs," respectively. Each node x of $\mathscr{T}$ is associated with exactly one branch, leading from node x to its immediate predecessor on $\mathscr{T}$. We refer to this branch with the shared label x.

Table 1. Lineage classes produced by coalescence events

		Class 2		
		S	*C*	*M*
Class 1	*S*	S → S S	M → S C	M → S M
	C	M → C S	C → C C	M → C M
	M	M → M S	M → M C	M → M M

Intraclass coalescences between pairs of lineages preserve the class; interclass coalescences result in *M* lineages.

Table 2. Possible monophyly events for two disjoint lineage classes, *S* and *C*

Monophyletic groups	Description	Notation
S	Monophyly of *S*	E_S
C	Monophyly of *C*	E_C
Only *S*	Paraphyly of *C*	$E_{SC'}$
Only *C*	Paraphyly of *S*	$E_{S'C}$
Both *S* and *C*	Reciprocal monophyly	E_{SC}
Neither *S* nor *C*	Polyphyly	$E_{S'C'}$

For an internal branch x in $\mathscr{T}$, the number of class-S left inputs is s_x^L (c_x^L for class C, m_x^L for class M); the number of class-S right inputs is s_x^R (c_x^R for class C, m_x^R for class M). The total number of class-S inputs of x is $s_x^I = s_x^L + s_x^R$ ($c_x^I = c_x^L + c_x^R$ for class C, $m_x^I = m_x^L + m_x^R$ for class M). The number of lineages that exit branch x, entering a branch farther up the species tree, is the set of outputs of branch x: s_x^O, c_x^O, or m_x^O.

We combine the input and output values into two three-entry vectors: the "input states" $\mathbf{n}_x^I = (s_x^I, c_x^I, m_x^I)$ and the "output states" $\mathbf{n}_x^O = (s_x^O, c_x^O, m_x^O)$. Note that $\mathbf{n}_x^I = \mathbf{n}_x^L + \mathbf{n}_x^R$. We refer to the nodes directly below node x corresponding to its left and right incoming branches by x_L and x_R, respectively, and to nodes farther down the tree by sequences of Ls and Rs, which, read from left to right, give the steps needed to reach them from x. For example, x_{RL} follows down from x to the right (x_R), then from x_R to the left (x_{RL}).

The time interval associated with node x is T_x, the length of branch x. Branch lengths are measured in coalescent time units of N generations, where N represents the haploid population size along the branch and is assumed to be constant. Thus, larger population sizes correspond to shorter lengths of time in coalescent units. Coalescences between inputs during time T_x yield the outputs of x. The root branch of $\mathscr{T}$ has infinite length.

The outputs of any nonroot branch are exactly the left or the right inputs of another branch farther up the tree; the outputs of the root are the outputs of the species tree. The root has only one output lineage: $\mathbf{n}_{\text{root}}^O = (0,0,1)$. Inputs of a node x are the outputs of x_L and x_R, so that $\mathbf{n}_x^L = \mathbf{n}_{x_L}^O$ and $\mathbf{n}_x^R = \mathbf{n}_{x_R}^O$. For convenience, when node x corresponds to leaf i, we let $s_x^I = s_x^L = S_i$ and $c_x^I = c_x^L = C_i$ (Fig. 1).

We define $\mathscr{T}_{SC}^x$ to be the initialized species subtree with root x and E_i^x to be the monophyly event E_i for the subtree with root x, ignoring the rest of the species tree.

Coalescence sequences. A coalescence sequence is a sequence of coalescences that reduces a set of lineages to another set of lineages. As an example, consider four lineages—labeled A, B, C, and D—that coalesce to a single lineage. One sequence has A and C coalesce first, followed by B and D, then the lineages resulting from the AC and BD coalescences. This sequence could be described as (A, C), (B, D), (AC, BD). If the first two coalescences happened in opposite order, the sequence would be (B, D), (A, C), (AC, BD).

Combinatorial functions. The probability $g_{nj}(T)$ that n lineages coalesce to j lineages in time T is given by equation 6.1 of ref. 45. It is nonzero only when $n \ge j \ge 1$ and $T \ge 0$, except that we set $g_{0,0}(T) = 1$.

Following equation 4 of ref. 4, the number of coalescence sequences that reduce n lineages to k lineages is $I_{n,k} = [n!(n-1)!] / [2^{n-k} k!(k-1)!]$. This function is nonzero only when $n \ge k \ge 1$, with the convention $I_{0,0} = 1$.

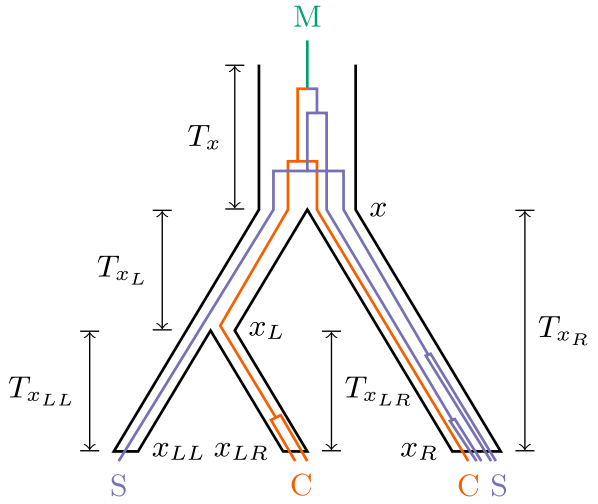

Fig. 1. Notation for computing monophyly probabilities above a species tree node x. Nodes x_{LL}, x_{LR}, and x_R are leaves. *S* lineages appear in blue, *C* lineages in orange, and *M* lineages in green. The figure illustrates reciprocal monophyly. Sequentially listing the numbers of *S*, *C*, and *M* lineages as a vector, the outputs of branch x are $\mathbf{n}_x^O = (0,0,1)$. Inputs are $\mathbf{n}_x^L = (1,1,0)$ and $\mathbf{n}_x^R = (2,1,0)$. Farther down the tree, branch x_L has inputs $\mathbf{n}_{x_L}^L = (1,0,0)$ and $\mathbf{n}_{x_L}^R = (0,1,0)$. Adopting the convention that leaf inputs enter from the left, branch x_R has inputs $\mathbf{n}_{x_R}^L = (4,1,0)$ and $\mathbf{n}_{x_R}^R = (0,0,0)$. Descending one more level—which is only possible for x_L—the inputs for branch x_{LL} are $\mathbf{n}_{x_{LL}}^L = (1,0,0)$ and $\mathbf{n}_{x_{LL}}^R = (0,0,0)$, and for branch x_{LR}, they are $\mathbf{n}_{x_{LR}}^L = (0,2,0)$ and $\mathbf{n}_{x_{LR}}^R = (0,0,0)$. Branch widths represent constant population sizes but do not indicate relative magnitudes of these sizes.

Finally, the binomial coefficient $W_2(r_1,r_2) = \binom{r_1 + r_2}{r_1}$, by equation 5 from ref. 4, gives the number of ways that separate coalescence sequences consisting of r_1 and r_2 coalescences can be ordered in a larger sequence containing them both as subsequences. $W_2(r_1,r_2)$ is defined when $r_1, r_2 \geq 0$.

The Central Recursion.

Overview. We develop a recursion for the probability of a particular output state $\mathbf{n}_x^O$ and monophyly event E_i^x for a branch x given the initialized species subtree $\mathscr{T}_{SC}^x$. We use the law of total probability to write the desired probability as a sum over all possible input states $\mathbf{n}_x^I$ of the probability of the input state multiplied by the conditional probability of the output given the input. Keeping in mind that $\{\text{inputs of } x\} = \{\text{outputs of } x_L\} \cup \{\text{outputs of } x_R\}$, we then use the independence of the outputs for branches x_L and x_R to decompose the probability of the input state of x into a product of the probabilities of the output states of x_L and x_R. Schematically,

$$\mathbb{P}\left(\text{outputs of } x, E_i^x \middle| \mathscr{T}_{SC}^x\right) = \sum_{\substack{\text{possible} \\ \text{inputs of } x}} \mathbb{P}\left(\text{outputs of } x_L, E_i^{x_L} \middle| \mathscr{T}_{SC}^{x_L}\right) \times \mathbb{P}\left(\text{outputs of } x_R, E_i^{x_R} \middle| \mathscr{T}_{SC}^{x_R}\right) \times \mathbb{P}\left(\text{outputs of } x, E_i^x \middle| \text{inputs of } x, \mathscr{T}_{SC}^x\right). \quad [1]$$

The third term on the right-hand side of Eq. **1**, which we represent by F, is the probability that the inputs coalesce to the specified outputs during time T_x in accord with the monophyly event. We write the random variable for the output state of branch x as $\mathbf{Z}_x$, labeling the particular values attained by the random variable by $\mathbf{n}_x^O$. By formalizing Eq. **1**, we can write the central recursion of our analysis:

$$\mathbb{P}\left(\mathbf{Z}_x = \mathbf{n}_x^O, E_i^x \middle| \mathscr{T}_{SC}^x\right) = \sum_{\mathbf{n}_x^L=0}^{\left(S_{x_L}^{\text{subt}}, C_{x_L}^{\text{subt}}, 1\right)} \sum_{\mathbf{n}_x^R=0}^{\left(S_{x_R}^{\text{subt}}, C_{x_R}^{\text{subt}}, 1\right)} \mathbb{P}\left(\mathbf{Z}_{x_L} = \mathbf{n}_{x_L}^O, E_i^{x_L} \middle| \mathscr{T}_{SC}^{x_L}\right) \times \mathbb{P}\left(\mathbf{Z}_{x_R} = \mathbf{n}_{x_R}^O, E_i^{x_R} \middle| \mathscr{T}_{SC}^{x_R}\right) F\left(\mathbf{n}_x^O, E_i^x \middle| \mathbf{n}_x^I, \mathscr{T}_{SC}^x\right). \quad [2]$$

In this equation, we denote the total number of inputs of class S across all of the leaves subtended by x_L or x_R by $S_{x_L}^{\text{subt}}$ or $S_{x_R}^{\text{subt}}$ ($C_{x_L}^{\text{subt}}$ or $C_{x_R}^{\text{subt}}$ for class C). Each of the two summations is a nested triple sum, proceeding componentwise over the three entries in the vectors $\mathbf{n}_x^L$ and $\mathbf{n}_x^R$—e.g., for $\mathbf{n}_x^L$, we sum from 0 to $S_{x_L}^{\text{subt}}$, from 0 to $C_{x_L}^{\text{subt}}$, and from 0 to 1. We now explain the basis for this recursion.

Bounds of summation. The sums in Eq. **2** traverse all possible inputs of branch x. We use summation bounds that only require information contained in the initialized species subtree $\mathscr{T}_{SC}^x$. Numbers of inputs are nonnegative, and for each lineage class, some branches have the possibility of having no inputs in the class. Thus, all lower bounds are 0.

For the upper bounds, because coalescence does not create new *S* and *C* lineages (Table 1), the numbers of *S* and *C* lineages never exceed the numbers of *S* and *C* leaves in the gene tree, respectively. Thus, for branch x, an upper bound for the possible number of inputs of class *S* or *C* from one side (*L* or *R*) is $S_{x_L}^{\text{subt}}$ or $S_{x_R}^{\text{subt}}$ for class *S* and $C_{x_L}^{\text{subt}}$ or $C_{x_R}^{\text{subt}}$ for class *C*.

We use Eq. **2** to calculate probabilities only for E_S, E_C, and E_{SC} (Table 2), using them to obtain probabilities for the remaining events. These three events require complete intraclass coalescence separately in the appropriate classes before interclass coalescences are possible. As a result, they permit exactly one coalescence between an *S* lineage and a *C* lineage. Because the leaves possess no *M* lineages and because only the unique coalescence between an *S* and a *C* lineage creates an *M* lineage (Table 1), the number of *M* lineages never exceeds 1.

Probability of the outputs of a node given the inputs. Separating the function F from Eq. **2** into a term for the probability that the correct number of outputs is produced from the inputs and a combinatorial term K_i for the probability that the coalescence

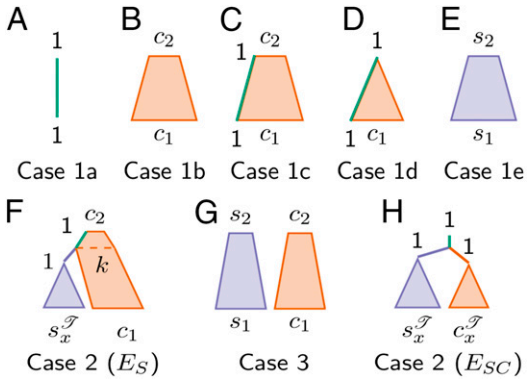

Fig. 2. All cases required for computing combinatorial terms K_S and K_{SC} in monophyly probabilities. (*A*–*G*) Cases for monophyly of *S* (Eq. **4**). (*H*) A case for reciprocal monophyly (Eq. **5**). In each panel, lineages coalesce from bottom to top, with the width of a shape corresponding to the number of lineages present. A single lineage is represented by a line, and multiple freely coalescing lineages are represented by shaded polygons with horizontal cross-section proportional to the number of extant lineages. Lineages represented in the same shape or in touching shapes can coalesce with each other. Lineage colors follow Fig. 1.

sequence generating those outputs occurs in accord with the monophyly event E_i, F takes the form

$$F\left(\mathbf{n}_x^O, E_i^x \middle| \mathbf{n}_x^I, \mathscr{T}_{SC}^x\right) = g_{\left|\mathbf{n}_x^I\right|,\left|\mathbf{n}_x^O\right|}(T_x)\, K_i\left(\mathbf{n}_x^O, E_i^x \middle| \mathbf{n}_x^I, \mathscr{T}_{SC}^x\right), \qquad [3]$$

where $\left|\mathbf{n}_x^I\right| = s_x^I + c_x^I + m_x^I$ and $\left|\mathbf{n}_x^O\right| = s_x^O + c_x^O + m_x^O$. For the case of $i = S$, in which monophyly of S is of interest, we have:

$$K_S\left(\mathbf{n}_x^O, E_S^x \middle| \mathbf{n}_x^I, \mathscr{T}_{SC}^x\right) = \begin{cases} 1 & \text{Case 1a: } \mathbf{n}_x^I = (0,0,1), \mathbf{n}_x^O = (0,0,1) \\ & \text{Case 1b: } \mathbf{n}_x^I = (0,c_1,0), \mathbf{n}_x^O = (0,c_2,0) \\ & \text{Case 1c: } \mathbf{n}_x^I = (0,c_1,1), \mathbf{n}_x^O = (0,c_2,1) \\ & \text{Case 1d: } \mathbf{n}_x^I = (0,c_1,1), \mathbf{n}_x^O = (0,0,1) \\ & \text{Case 1e: } \mathbf{n}_x^I = (s_1,0,0), \mathbf{n}_x^O = (s_2,0,0) \\ \dfrac{\sum_{k=c_2+1}^{c_1} I_{s_x^{\mathscr{T}},1} I_{c_1,k} W_2\left(s_x^{\mathscr{T}}-1, c_1-k\right) k I_{k,c_2+1}}{I_{s_x^{\mathscr{T}}+c_1,c_2+1}} & \text{Case 2: } \mathbf{n}_x^I = \left(s_x^{\mathscr{T}}, c_1, 0\right), \mathbf{n}_x^O = (0,c_2,1) \\ \dfrac{I_{s_1,s_2} I_{c_1,c_2} W_2(s_1-s_2, c_1-c_2)}{I_{s_1+c_1,s_2+c_2}} & \text{Case 3: } \mathbf{n}_x^I = (s_1,c_1,0), \mathbf{n}_x^O = (s_2,c_2,0) \\ 0 & \text{otherwise} \end{cases} . \qquad [4]$$

Here, $s_x^{\mathscr{T}}$ records the total number of class-S lineages in the species tree $\mathscr{T}$ at the species merging event corresponding to node x. For cases 1 and 3, $0 < c_2 \le c_1$ and $0 < s_2 \le s_1$. For case 2, $0 \le c_2 < c_1$, $0 < c_1$, and $0 < s_x^{\mathscr{T}}$. Note that it is not strictly necessary for $s_x^I = s_x^{\mathscr{T}}$ in case 2 (violation of E_S would be accommodated elsewhere in the calculation, on another species tree branch), but we retain this condition for clarity.

Function F (Eq. **3**) describes the probability of an output state and monophyly event given an input state and the initialized species tree. Its g term records the probability that the correct number of coalescences occur during the time T_x, defining a space of coalescence sequences from the input state to any output state with the same number of lineages as the desired output. K_i (Eq. **4**) records the fraction of those sequences that produce the correct output and preserve the monophyly event E_i (in this case, E_S).

The cases in Eq. **4** represent distinct scenarios for the types of input and output lineages present (Fig. 2 *A*–*G*). In case 1 (Fig. 2 *A*–*E*), no coalescence violates E_S, as all coalescences have types (S,S) (case 1e), (C,C) (cases 1b, 1c, 1d), or (C,M) (cases 1c, 1d). No coalescences occur in case 1a. The correct output state is guaranteed ($K_S = 1$), as each coalescence decrements the number of S (case 1e) or C lineages (cases 1b, 1c, 1d), and the only change from input to output is a reduction in S or C lineages.

In cases 2 and 3, both S and C lineages are present, and we enumerate the ways to obtain the desired output state from the input state in accord with the monophyly event. To obtain K_S, we divide by the total number of coalescence sequences of correct length.

Case 2 describes the only possible way an S lineage and a C lineage can coalesce with each other under E_S (Fig. 2*F*). All extant S lineages at the time of node x ($s_x^I = s_x^{\mathscr{T}}$) must coalesce to a single lineage, and that lineage must coalesce with a C lineage when k class-C lineages remain from the $c_x^I = c_1$ extant C lineages present in both species at node x. This coalescence results in a single M lineage and $k-1$ lineages of class C, which can coalesce in any order to a single class-M lineage and $c_x^O = c_2$ class-C lineages.

The number of ways that $s_x^{\mathscr{T}}$ lineages can coalesce to one lineage is $I_{s_x^{\mathscr{T}},1}$. The number of ways that c_1 lineages can coalesce to k lineages is $I_{c_1,k}$. These separate sequences of $s_x^{\mathscr{T}} - 1$ and $c_1 - k$ coalescences can be ordered in $W_2(s_x^{\mathscr{T}} - 1, c_1 - k)$ ways. The number of ways that a single S lineage can coalesce with one of k lineages of class C is k. Finally, k lineages—one M lineage and $k-1$ class-C lineages—can coalesce to $c_2 + 1$ lineages in I_{k,c_2+1} ways. The desired number of coalescence sequences of correct length that result in the correct output state without violating E_S is obtained by summing the product of these terms over possible values of k, which ranges from just enough C lineages $(c_2 + 1)$ to allow the correct number of output lineages (c_2)—the resultant single S lineage coalesces with one C lineage and then no other coalescence occurs—to the total number c_1 of incoming C lineages, when all of the S lineages coalesce before any of the C lineages coalesce. The denominator of ratio K_S is the total number of ways of coalescing $s_x^{\mathscr{T}} + c_1$ input lineages to $c_2 + 1$ output lineages: $I_{s_x^{\mathscr{T}}+c_1,c_2+1}$. Note that setting $c_2 = 0$ in the ratio, reflecting a scenario with only one output lineage, of class M, reduces the formula to the two-species equation 11 from ref. 4 (*Supporting Information*).

Case 3 describes any situation with S and C lineages present and no interclass coalescence (Fig. 2*G*). At node x, the $s_x^I = s_1$ class-S lineages coalesce to $s_x^O = s_2$ class-S lineages, and the $c_x^I = c_1$ class-C lineages to $c_x^O = c_2$ class-C lineages. Group S has not yet coalesced with the other sampled lineages and does not do so within this species tree branch; its monophyly is not necessarily determined on the branch. The number of ways s_1 lineages can coalesce to s_2 lineages is I_{s_1,s_2}; c_1 lineages can coalesce to c_2 lineages in I_{c_1,c_2} ways. These sequences can be ordered in $W_2(s_1 - s_2, c_1 - c_2)$ ways. The numerator in the fraction of coalescence sequences of the correct length that result in the correct output state without violating E_S is the product of these three terms. The denominator is the total number of ways of coalescing $s_1 + c_1$ input lineages to $s_2 + c_2$ outputs: $I_{s_1+c_1,s_2+c_2}$.

Any pairing of an input state and an output state that does not belong in cases 1–3 of Eq. **4** must violate E_S. This violation yields an output probability of $K_S = 0$.

Reciprocal monophyly. Monophyly events E_{SC} and E_S differ in that for E_{SC}, unlike for E_S, C and M lineages cannot coexist. Thus, cases 1c and 1d of Eq. **4** move to "otherwise" for K_{SC}, producing $K_{SC} = 0$ for the input states of those cases. Additionally, for E_{SC}, an interclass coalescence can occur only after all S lineages have coalesced to a single S lineage and all C lineages have coalesced to a single C lineage, whereas E_S required only that all S lineages coalesce. For E_S, interclass coalescences occur only in case 2 of Eq. **4**; for E_{SC}, we modify this case by requiring first that before the interclass coalescence, the C lineages must be all C lineages in the tree at the time of node x (as we did for S lineages for case 2 of Eq. **4**; $c_x^I = c_x^{\mathscr{T}}$). Second we require $k = 1$ and $c_2 = 0$, so all C lineages coalesce to a single lineage before the interclass coalescence. Setting $k = 1$, $c_2 = 0$, substituting $c_x^{\mathscr{T}}$

for c_1 in case 2 of Eq. **4**, and noting that $I_{1,1}=1$, we obtain case 2 for K_{SC} (Fig. 2*H*), applicable when $\mathbf{n}_x^I=(s_x^{\mathscr{T}},c_x^{\mathscr{T}},0)$ and $\mathbf{n}_x^O=(0,0,1)$:

$$K_{SC}\left(\mathbf{n}_x^O,E_{SC}^x\middle|\mathbf{n}_x^I,\mathscr{T}_{SC}^x\right)=\frac{I_{s_x^{\mathscr{T}},1}I_{c_x^{\mathscr{T}},1}W_2\left(s_x^{\mathscr{T}}-1,c_x^{\mathscr{T}}-1\right)}{I_{s_x^{\mathscr{T}}+c_x^{\mathscr{T}},1}}. \quad [5]$$

For E_{SC}, the input condition for case 2 can be satisfied only at the root of $\mathscr{T}$. For all input states other than those of Eq. **5** or cases 1c and 1d of Eq. **4**, $K_{SC}=K_S$.

Completing the calculation. Having obtained a recursion that propagates monophyly probabilities through a species tree, we apply Eq. **2** at the root to complete the calculation of the probability of a monophyly event on $\mathscr{T}_{SC}$:

$$\mathbb{P}(E_i|\mathscr{T}_{SC})=\mathbb{P}\left(\mathbf{Z}_{\text{root}}=(0,0,1),E_i^{\text{root}}\middle|\mathscr{T}_{SC}^{\text{root}}\right). \quad [6]$$

Specifying each possible monophyly event E_i^{root} in Eq. **6**,

$$\mathbb{P}(E_S|\mathscr{T}_{SC})=\mathbb{P}\left(\mathbf{Z}_{\text{root}}=(0,0,1),E_S^{\text{root}}\middle|\mathscr{T}_{SC}^{\text{root}}\right) \quad [7]$$

$$\mathbb{P}(E_C|\mathscr{T}_{SC})=\mathbb{P}(E_S|\mathscr{T}_{CS}) \quad [8]$$

$$\mathbb{P}(E_{SC}|\mathscr{T}_{SC})=\mathbb{P}\left(\mathbf{Z}_{\text{root}}=(0,0,1),E_{SC}^{\text{root}}\middle|\mathscr{T}_{SC}^{\text{root}}\right) \quad [9]$$

$$\mathbb{P}(E_{SC'}|\mathscr{T}_{SC})=\mathbb{P}(E_S|\mathscr{T}_{SC})-\mathbb{P}(E_{SC}|\mathscr{T}_{SC}) \quad [10]$$

$$\mathbb{P}(E_{S'C}|\mathscr{T}_{SC})=\mathbb{P}(E_C|\mathscr{T}_{SC})-\mathbb{P}(E_{SC}|\mathscr{T}_{SC}) \quad [11]$$

$$\mathbb{P}(E_{S'C'}|\mathscr{T}_{SC})=1-\mathbb{P}(E_{SC'}|\mathscr{T}_{SC})-\mathbb{P}(E_{S'C}|\mathscr{T})-\mathbb{P}(E_{SC}|\mathscr{T}_{SC}), \quad [12]$$

where $\mathscr{T}_{CS}$ is $\mathscr{T}_{SC}$ with the labels S and C switched. These recursive computations reduce to the known values for the two-species case (*Supporting Information*).

Effect of Species Tree Height *T*. To illustrate the features of monophyly probabilities, we now examine the effects on the probabilities of model parameters. First, we vary the tree height T and preserve relative branch length proportions, studying the limiting cases of $T=0$ and $T\to\infty$.

***T* = 0.** At $T=0$, nonroot species tree branches have length 0, so the species tree is a single infinitely long branch—the root—with initial sample sizes equal to the sums of the values at the leaves. Formally, because $g_{ij}(0)=1$ if $i=j$, every nonroot branch outputs exactly its inputs. All $s=\sum_{i=1}^{\ell}S_i$ class-S lineages and all $c=\sum_{i=1}^{\ell}C_i$ class-C lineages enter the root. Using Eq. **7**, and noting that $g_{i,1}(\infty)=1$, we find that $\mathbb{P}(E_S|\mathscr{T}_{SC})$ is a simple function of the total numbers of S and C lineages:

$$f(s,c)=\frac{\sum_{k=1}^{c}I_{s,1}I_{c,k}W_2(s-1,c-k)kI_{k,1}}{I_{s+c,1}}=\frac{2(s+c)}{s(s+1)\binom{s+c}{s}}, \quad [13]$$

with the last equality from equation 11 in ref. 4. Function f decreases with increasing s or c, as adding any lineage increases the chance of a monophyly-violating interclass coalescence.

T → ∞. As $T\to\infty$, because $\lim_{T\to\infty}g_{ij}(T)=1$ when $j=1$, every branch exhibits complete coalescence. We define the minimal subtree with respect to S, $\mathscr{T}_{SC}^*$, as the smallest subtree of the species tree whose leaves contain all of the initial S lineages in the tree.

For large T, the monophyly probability depends on properties of $\mathscr{T}_{SC}^*$. To be monophyletic, the S lineages must encounter C lineages only above its root. If $\mathscr{T}_{SC}^*$ contains no C lineages, then complete coalescence in each branch implies monophyly of S lineages, and the monophyly probability is 1. If $\mathscr{T}_{SC}^*$ contains C lineages and is at a leaf, k, then the limiting probability is $f(S_k,C_k)$. Complete coalescence in every branch makes this leaf analogous to the root in the $T=0$ case. Note that if $S_k>1$ then the limit $f(S_k,C_k)$ lies in the interior of the unit interval. This result contrasts with ref. 4, where lineage classes correspond to species tree leaves and the $T\to\infty$ probability of E_S is 1. In our scenario, because multiple lineage classes are permitted at a leaf, a nonzero limit can be below 1.

If $\mathscr{T}_{SC}^*$ contains C lineages but is not a leaf, however, then complete coalescence in every branch implies that some proper subset of S lineages must coalesce with C lineages before all of the S lineages can coalesce with each other. In this case, the limiting monophyly probability is 0.

***Finite, nonzero* T.** The extreme cases assist in understanding the behavior of the probability of E_S for intermediate T. We enumerate the possible situations based on $\mathscr{T}_{SC}^*$, continuing to assume that relative branch lengths are fixed and that a changing tree height changes all branch lengths proportionally.

If $\mathscr{T}_{SC}^*$ contains no C lineages, then decreasing the tree height decreases the probability of monophyly by decreasing the time during which S lineages are able to coalesce with only themselves, eventually approaching a minimum $f(s,c)$ achieved at $T=0$. Similarly, increasing T increases the monophyly probability toward 1 as $T\to\infty$.

If $\mathscr{T}_{SC}^*$ contains C lineages and is a leaf, then decreasing the tree height decreases the monophyly probability by decreasing the time before more C lineages are added to the population that contains the S lineages. Shrinking the tree also increases the expected number of additional C lineages introduced at species merging events, further decreasing the monophyly probability. The minimal probability of monophyly therefore occurs at $T=0$. Similarly, increasing the tree height increases the probability of monophyly, approaching a maximal value as $T\to\infty$. Consequently, in this case, like in the previous case, the probability also increases monotonically in T.

If $\mathscr{T}_{SC}^*$ contains C lineages and is not a leaf, then the minimal probability of monophyly, approached as $T\to\infty$, is 0. As we will see in numerical examples, however, monotonicity of the monophyly probability with T is not guaranteed, and different initial sample sizes on the same species tree can generate different behavior.

Effect of Relative Branch Lengths. Next, to investigate the behavior of the monophyly probability as T increases, we devise a simple three-species, two-parameter scenario, subdividing the tree height T by a parameter r. We calculate the probability of E_S for different sample-size conditions, varying r and T.

Fig. 3 shows the species tree and its resulting monophyly probabilities for four representative initial conditions. For each lineage class, S and C, the four cases place one or more lineage pairs into the three species, using different placements across the four cases. The cases include scenarios in which at least one species contains both S and C lineages (B, D, E), in which one (C) or both lineage classes spans multiple species (B, D, E), and in which the species containing S lineages are not monophyletic in the species tree (B, C).

The four cases (Fig. 3 *B*–*E*) illustrate differences in the pattern of increase or decrease in the monophyly probability with changes in r at fixed tree height T (*Supporting Information*). In most cases with fixed r, the probability decreases to 0 with increasing T, although in some boundary cases with $r=0$ and $r=1$ that change the case for the limiting behavior with T (see above on $T\to\infty$), it approaches a positive value strictly within the unit interval. These scenarios highlight the fact that depending on the relative branch lengths and distribution of lineage classes across species, the monophyly probability can be monotonically increasing in T, monotonically decreasing, or not monotonic at all.

Effect of Pooling. Our next scenario simulates the difference between separating and pooling distinct species when computing monophyly probabilities, recalling that tests with more than two species have until now required the pooling of multiple clades (23, 26).

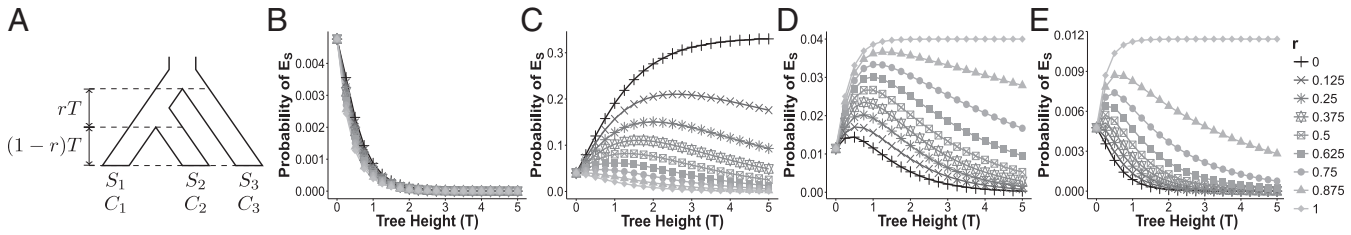

Fig. 3. The effect on monophyly probabilities of changing two branch lengths in relation to each other. (*A*) Model species tree. If the branch length coefficient *r* is 0, then the tree has a polytomy, and if $r = 1$, then the tree reduces to a two-species tree. (*B–E*) The probabilities of E_S (Eq. **7**) for monophyly of *S* for the tree in *A* under different scenarios: (*B*) $(S_1, S_2, S_3) = (2,0,2)$, $(C_1, C_2, C_3) = (2,2,2)$. (*C*) $(S_1, S_2, S_3) = (2,0,2)$, $(C_1, C_2, C_3) = (0,2,0)$. (*D*) $(S_1, S_2, S_3) = (2,2,0)$, $(C_1, C_2, C_3) = (2,0,2)$. (*E*) $(S_1, S_2, S_3) = (2,2,0)$, $(C_1, C_2, C_3) = (2,2,2)$.

We consider four species trees with equal height and 12 lineages (Fig. 4). Six class-*C* lineages appear in one species descended from the root. The other six—the *S* lineage class—are evenly divided between one, two, three, or six other leaves. If we interpret the seven-leaf tree in Fig. 4*D* to be the "true" species tree, then the other trees represent pooling schemes, the two-leaf tree (Fig. 4*A*) being the only one possible to analyze using previous results.

Fig. 4 *E–J* displays the probabilities of all possible monophyly events for each tree. For each event, pooling does not affect the extreme cases $T = 0$ and $T \to \infty$. For intermediate *T*, the monophyly probability for the *S* lineages decreases as pooling is reduced from the case in which the six class-*S* lineages are treated as belonging to a single species to the case in which each lineage is in its own species (Fig. 4*E*); the monophyly probability for *C* remains largely unchanged (Fig. 4*F*). As pooling is reduced, the probability of monophyly of only *S* and not *C* decreases (Fig. 4*G*), and that of only *C* and not *S* increases (Fig. 4*H*). The reciprocal monophyly probability decreases (Fig. 4*I*) and the probability of no monophyly increases (Fig. 4*J*).

In this scenario, the *S* and *C* lineages meet only at the species tree root, and the monophyly probabilities are determined by the numbers of lineages that reach the root. Coalescence is faster with more nonisolated lineages; pooling species together results in more coalescence events and fewer *S* lineages entering the root, increasing the probability of monophyly of both *S* and *C* lineages as well as the reciprocal monophyly probability (Fig. 4 *E, F,* and *I*). Decreasing the number of *S* lineages at the root decreases the number of coalescences needed to produce E_S above the root, decreasing the chance of an interclass coalescence, whereas decreasing the number of *S* lineages does not change the number of coalescences necessary to produce E_C and has a smaller effect on its probability (cf. Fig. 4 *E* and *F*). The probability for E_{SC} closely follows that of E_S, as production of reciprocal monophyly is limited by the monophyly of the individual classes.

As can be seen from the increase in probability for E_S as pooling is increased (Fig. 4*E*), the correct monophyly probability for clades that have been pooled tends to be lower than that obtained under a model where the pooled clades are treated as a single clade. The monophyly probability will likely be overestimated if populations are pooled.

Application to Data. To illustrate the empirical use of Eq. **7** and to test if our theoretical results reasonably replicate patterns in real data, we perform an analysis of monophyly frequencies using *Zea mays* maize and teosinte genomic data (46).

Hufford et al. (47) analyzed 75 individuals from the data of Chia et al. (46), considering four groups: teosinte varieties var. *parviglumis* ("parviglumis") and var. *mexicana* ("mexicana") and domesticated maize landraces ("landraces") and improved lines ("improved"). Modifying the estimated tree of individuals from figure 1 in Hufford et al. (47) to make a model "species" tree the leaves of which are the four groups (Fig. 5*A*), we compute theoretical monophyly probabilities for each of the groups via Eq. **7**. We also estimate the empirical frequency of monophyly for each group by randomly sampling individuals from each group, constructing multiple gene trees per sample from SNP blocks, and averaging frequencies of monophyly in the gene trees over the random samples. This procedure employs 100 unique random samples of eight individuals from the Hufford et al. subset, each containing two individuals from each of the four groups. Finally, we compare the observed and theoretical monophyly frequencies.

The monophyly frequencies appear in Fig. 5*B* and are summarized in Table S2. The theoretical frequencies predict the observations reasonably well. For each clade, especially parviglumis and mexicana, the mean observed monophyly frequency

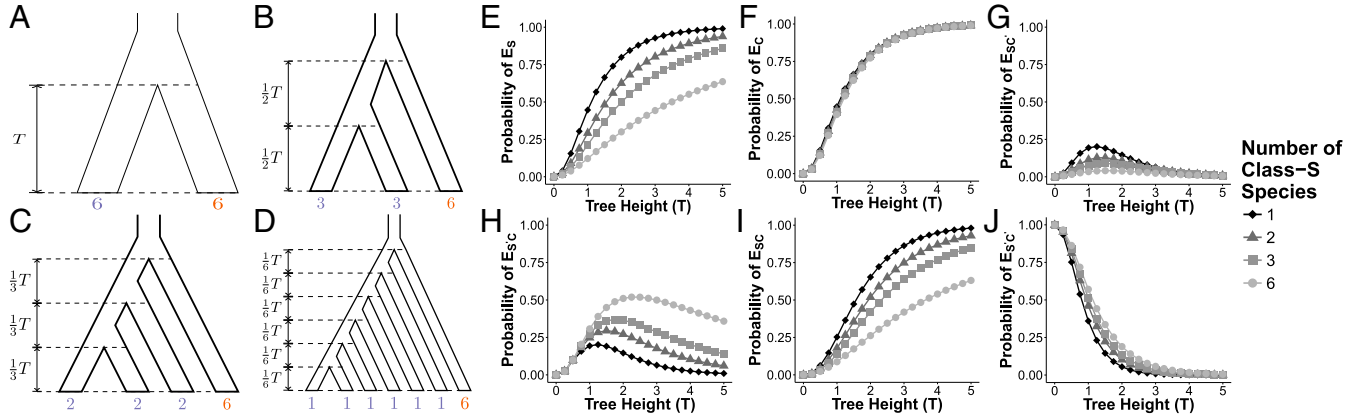

Fig. 4. The effect on monophyly probabilities of pooling lineages from separate species. (*A–D*) Model species trees. Labels record numbers of input lineages (*S* in blue, *C* in orange). (*E–J*) Probabilities of monophyly events. The trajectories represent species trees with six class-*S* lineages evenly distributed over one (*A*), two (*B*), three (*C*), and six (*D*) species. (*E*) E_S (Eq. **7**). (*F*) E_C (Eq. **8**). (*G*) $E_{SC'}$ (Eq. **10**). (*H*) $E_{S'C}$ (Eq. **11**). (*I*) E_{SC} (Eq. **9**). (*J*) $E_{S'C'}$ (Eq. **12**).

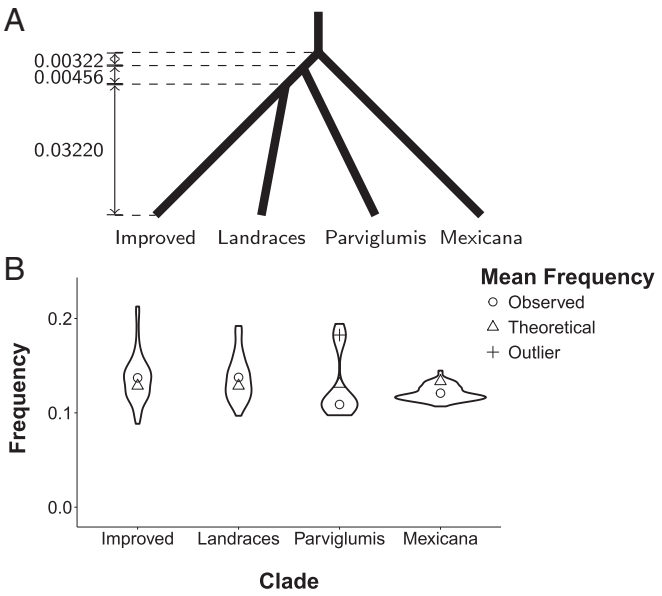

Fig. 5. Monophyly frequencies in maize and teosinte. (*A*) Model species tree. (*B*) Violin-plot distributions across lineage subsamples of monophyly frequencies for four clades. Means of the observed distributions (excluding outliers for the improved and parviglumis clades) appear as circles and theoretical values appear as triangles. Outliers appear for a single point at frequency ~0.43 in the improved clade and for several points at frequency >0.17 in the parviglumis clade, with the cross indicating the mean of the parviglumis outliers (*Supporting Information*).

over 100 samples closely coincides with the theoretical monophyly probability (Fig. 5*B*). Although the theoretical probability is noticeably below the mean for the improved and landrace clades and above the mean for parviglumis and mexicana, it lies well inside the observed distributions.

Eq. **7** relies on a model with selectively neutral loci and constant population size; a deviation from theoretical probabilities could suggest a violation of one of the model assumptions. Domestication imposes strong selection and population bottlenecks (27, 48, 49), factors that violate our model in a manner that would increase monophyly frequencies. Excess empirical monophyly in the improved and landrace clades (Fig. 5*B*, Table S2) is thus compatible with domestication in the history of these domesticated groups.

Discussion

Extending a past computation (4) from 2 to n species, we have obtained a general algorithm for the probability of any monophyly event of two lineage classes in a species tree of any size. In our generalization, unlike in previous calculations, no restriction exists on the class labeling of lineages, so that monophyly probabilities can be computed on samples aggregated across multiple species. We have uncovered behaviors absent in the two-species case, including nonmonotonicity of the monophyly probability in the tree height and positive limiting probabilities below 1. Both phenomena occur in scenarios newly possible to include in monophyly calculations, in which the lineage set whose monophyly is of interest spans multiple species, or in which lineages of at least one species span both classes.

We have used a pruning algorithm similar to other species tree computations (9, 40–44) that evaluate a quantity at a parent node in terms of corresponding values for daughter nodes. In previous applications of this idea, the states recorded at a node are generally simpler than our input and output states. For example, in evaluating the time to the MRCA (43), they are one-dimensional; our approach instead tracks lineage classes as three variables, accommodating complex transitions that occur at interclass coalescences.

Previous work on monophyly probabilities has been limited to small numbers of species (4, 35–38). This limitation has forced investigators to either group multiple species together into a single clade (23, 26)—a choice that our tree-pooling experiment shows can overestimate monophyly probabilities—or to consider pairwise comparisons when multispecies analyses would be preferable (25, 34, 39). By identifying a bias that occurs when pooling distinct species in monophyly probability computations, our experiment suggests that pooling should be avoided when possible. Our results allow researchers to move beyond such simplifications by performing monophyly calculations in larger species groups.

One application of our results is to extend a test of a null hypothesis that an observed monophyletic pattern is due to chance alone (24). This test has been available only in situations with species-specific lineages and two-species trees; it can now be extended to arbitrary trees and non-species-specific lineages. The results also provide a step toward computations for monophyly events on three or more lineage groups considered jointly.

As an empirical demonstration, we analyzed data from maize and teosinte, calculating theoretical and observed monophyly frequencies in four groups. The empirical frequencies generally match the predictions; frequencies exceeding predicted values in the domesticated species may reflect the fact that domestication bottlenecks and strong selection can violate our model in a manner that increases the likelihood of monophyly.

We note that our *Z. mays* results should be viewed with caution. We assumed a model of instantaneous divergence events without incorporating the subsequent gene flow that likely occurred in this system (47). Furthermore, our model species tree contains uncertainty; however, we do not expect a bias in any specific direction to have resulted from its construction. Perhaps more seriously, we generated the model tree from the same study whose data we used for constructing gene trees. However, considerations of monophyly were irrelevant in producing the model tree, so that construction of the model did not guarantee the agreement we obtained between theoretical and observed monophyly.

The maize analysis illustrates how our framework can be used to study monophyly in multispecies genomic data. The formulas derived here allow for greater flexibility in studies of monophyly and its relationship to species trees, contributing to a more comprehensive toolkit for phylogeographic, systematic, and evolutionary studies.

Materials and Methods

Maize Species Tree. We used maize HapMap V2 SNP data from www.panzea.org/#!genotypes/cctl (46) consisting of 55 million SNPs and small indels from 103 *Z. mays* inbred lines. To construct Fig. 5*A*, we determined relative branch lengths from figure 1 in Hufford et al. (47). We chose a tree height of 0.04, measured in units of N generations, where N is the haploid population size, noting that a ~10,000-y domestication time (47) translates via conversion factors calculated from figure 7 in ref. 50 (top panel, T_D column) to 0.036 units of N generations. We chose our root as the root of the Hufford et al. ingroup tree (second node from left in figure 1 of ref. 47, call it x), our Parviglumis/Domesticated node as the MRCA of all domesticated lineages and parviglumis lineages TIL01, TIL03, TIL11, and TIL14 ($y = x_{LLLL}$ in figure 1 of ref. 47, oriented so that L is "down" rather than "left"), and our Landrace/Improved node as the MRCA of all domesticated lineages (y_L in figure 1 of ref. 47).

Maize Samples. We chose 100 samples of four lineage pairs, selecting randomly among 29 improved, 12 landrace, 8 parviglumis, and 2 mexicana individuals. We chose pairs within groups so that the Hufford et al. tree, a genome-wide tree of individuals, restricted to each eight-lineage sample would display the model species tree in Fig. 5*A*, irrespective of which lineage in a pair was chosen to represent its group (*Supporting Information*).

Maize Gene Trees. The maize genome has ~2.3×10^9 bp (51), with linkage disequilibrium (LD) decay at ~1,500 bp (52). For simplicity and to

accommodate large quantities of missing data, despite genome-wide variation in recombination rate and SNP density, we fixed a single block size for analyses throughout the genome. With $\sim 5 \times 10^7$ SNPs in the dataset, SNP density per "LD block" is 32.6, which we round to 30. We divided the SNPs into nonoverlapping 30-SNP blocks and used every hundredth block in a concatenated genome starting from chromosome 1, resulting in ~6,000–7,000 gene trees per sample after removing blocks monomorphic in the sample and gene trees polytomic for the sample. We concatenated SNPs within blocks, computed blockwise Hamming distance matrices, and obtained gene trees using the *hclust* UPGMA (unweighted pair group method with arithmetic mean) clustering function in the R *stats* package. SNPs with missing data for a lineage pair were excluded in distance calculations.

Software Implementation. The *Monophyler* software package implementing Eqs. **7**, **8**, and **9** can be found at rosenberglab.stanford.edu/monophyler.html.

ACKNOWLEDGMENTS. We thank Jeff Ross-Ibarra for assistance with the maize data and John Rhodes and two reviewers for comments on a draft of the manuscript. We acknowledge support from NIH Grant R01 GM117590, NSF Grant DBI-1458059, a New Zealand Marsden grant, and a Stanford Graduate Fellowship.

1. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.
2. Takahata N, Nei M (1985) Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110(2):325–344.
3. Neigel J, Avise J (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. *Evolutionary Processes and Theory*, eds Karlin S, Nevo E (Academic Press, San Diego), pp 515–534.
4. Rosenberg NA (2003) The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57(7):1465–1477.
5. Degnan JH, Salter LA (2005) Gene tree distributions under the coalescent process. *Evolution* 59(1):24–37.
6. Wu CI (1991) Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127(2):429–435.
7. Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet* 8(4): e1002660.
8. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol* 53(1):320–328.
9. Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66(3):763–775.
10. Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Syst Biol* 56(6):887–895.
11. Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci USA* 107(20):9264–9269.
12. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* 24(6):332–340.
13. Sites JW, Jr, Marshall JC (2004) Operational criteria for delimiting species. *Annu Rev Ecol Evol Syst* 35:199–227.
14. De Queiroz K (2007) Species concepts and species delimitation. *Syst Biol* 56(6): 879–886.
15. Moritz C (1994) Defining "evolutionarily significant units" for conservation. *Trends Ecol Evol* 9(10):373–375.
16. Baum DA, Shaw KL (1995) Genealogical perspectives on the species problem. *Experimental and Molecular Approaches to Plant Biosystematics*, eds Hoch PC, Stephenson AC (Missouri Botanical Garden, St. Louis), pp 289–303.
17. Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution* 56(8):1557–1565.
18. Edwards SV, Beerli P (2000) Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54(6): 1839–1854.
19. Funk D, Omland K (2003) Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu Rev Ecol Evol Syst* 34:397–423.
20. Hare MP, Weinberg JR (2005) Phylogeography of surfclams, *Spisula solidissima*, in the western North Atlantic based on mitochondrial and nuclear DNA sequences. *Mar Biol* 146(4):707–716.
21. Syring J, Farrell K, Businský R, Cronn R, Liston A (2007) Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Syst Biol* 56(2):163–181.
22. Birky CW, Jr, Wolf C, Maughan H, Herbertson L, Henry E (2005) Speciation and selection without sex. *Hydrobiologia* 546(1):29–45.
23. Carstens BC, Richards CL (2007) Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* 61(6):1439–1454.
24. Rosenberg NA (2007) Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution* 61(2):317–323.
25. Neilson ME, Stepien CA (2009) Evolution and phylogeography of the tubenose goby genus *Proterorhinus* (Gobiidae: Teleostei): Evidence for new cryptic species. *Biol J Linn Soc Lond* 96(3):664–684.
26. Kubatko LS, Gibbs HL, Bloomquist EW (2011) Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. *Syst Biol* 60(4):393–409.
27. Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. *Nature* 398(6724):236–239.
28. Ting CT, Tsaur SC, Wu CI (2000) The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. *Proc Natl Acad Sci USA* 97(10): 5313–5316.
29. Dopman EB, Pérez L, Bogdanowicz SM, Harrison RG (2005) Consequences of reproductive barriers for genealogical discordance in the European corn borer. *Proc Natl Acad Sci USA* 102(41):14706–14711.
30. Takahata N, Slatkin M (1990) Genealogy of neutral genes in two partially isolated populations. *Theor Popul Biol* 38(3):331–350.
31. Wakeley J (2000) The effects of subdivision on the genetic divergence of populations and species. *Evolution* 54(4):1092–1101.
32. Hickerson MJ, Meyer CP, Moritz C (2006) DNA barcoding will often fail to discover new animal species over broad parameter space. *Syst Biol* 55(5):729–739.
33. Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. *Syst Biol* 56(3):400–411.
34. Bergsten J, et al. (2012) The effect of geographical scale of sampling on DNA barcoding. *Syst Biol* 61(5):851–869.
35. Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. *Theor Popul Biol* 61(2):225–247.
36. Degnan JH (2010) Probabilities of gene trees with intraspecific sampling given a species tree. *Estimating Species Trees: Practical and Theoretical Aspects*, eds Knowles LL, Kubatko LS (Wiley-Blackwell, Hoboken, NJ), pp 53–78.
37. Zhu S, Degnan JH, Steel M (2011) Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theor Popul Biol* 79(4):220–227.
38. Eldon B, Degnan JH (2012) Multiple merger gene genealogies in two species: Monophyly, paraphyly, and polyphyly for two examples of Lambda coalescents. *Theor Popul Biol* 82(2):117–130.
39. Baker AJ, Tavares ES, Elbourne RF (2009) Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Mol Ecol Resour* 9(Suppl s1):257–268.
40. RoyChoudhury A, Felsenstein J, Thompson EA (2008) A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180(2): 1095–1105.
41. Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* 29(8):1917–1932.
42. RoyChoudhury A, Thompson EA (2012) Ascertainment correction for a population tree via a pruning algorithm for likelihood computation. *Theor Popul Biol* 82(1):59–65.
43. Efromovich S, Kubatko LS (2008) Coalescent time distributions in trees of arbitrary size. *Stat Appl Genet Mol Biol* 7(1):2.
44. Stadler T, Degnan JH (2012) A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Algorithms Mol Biol* 7(1):7.
45. Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol* 26(2):119–164.
46. Chia JM, et al. (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* 44(7):803–807.
47. Hufford MB, et al. (2012) Comparative population genomics of maize domestication and improvement. *Nat Genet* 44(7):808–811.
48. Wright SI, et al. (2005) The effects of artificial selection on the maize genome. *Science* 308(5726):1310–1314.
49. Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci USA* 101(29):10667–10672.
50. Ross-Ibarra J, Tenaillon M, Gaut BS (2009) Historical divergence and gene flow in the genus *Zea*. *Genetics* 181(4):1399–1413.
51. Schnable PS, et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.
52. Remington DL, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98(20):11479–11484.