Asymptotic Properties of the Number of Matching Coalescent Histories for Caterpillar-Like Families of Species Trees

Filippo Disanto and Noah A. Rosenberg

Abstract—Coalescent histories provide lists of species tree branches on which gene tree coalescences can take place, and their enumerative properties assist in understanding the computational complexity of calculations central in the study of gene trees and species trees. Here, we solve an enumerative problem left open by Rosenberg (*IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10: 1253-1262, 2013) concerning the number of coalescent histories for gene trees and species trees with a matching labeled topology that belongs to a generic caterpillar-like family. By bringing a generating function approach to the study of coalescent histories, we prove that for any caterpillar-like family with seed tree t, the sequence $(h_n)_{n \geq 0}$ describing the number of matching coalescent histories of the nth tree of the family grows asymptotically as a constant multiple of the Catalan numbers. Thus, $h_n \sim \beta_t c_n$, where the asymptotic constant $\beta_t > 0$ depends on the shape of the seed tree t. The result extends a claim demonstrated only for seed trees with at most eight taxa to arbitrary seed trees, expanding the set of cases for which detailed enumerative properties of coalescent histories can be determined. We introduce a procedure that computes from t the constant β_t as well as the algebraic expression for the generating function of the sequence $(h_n)_{n \geq 0}$.

Index Terms—Catalan numbers, caterpillar-like trees, coalescent, enumeration, generating functions, phylogenetics

✦

1 INTRODUCTION

COALESCENT histories, mathematical structures representing combinatorially distinct ways in which a given gene tree can coalesce along the branches of a given species tree, are important in a variety of phylogenetic problems [6], [14], [15]. They arise, for example, in proofs concerning theoretical properties of species tree inference algorithms [1], [18], in empirical analyses of gene tree probability distributions [16], and in studies of gene trees under hybridization [21]. Many of these applications trace to the appearance of coalescent histories in a sum performed in a fundamental calculation for inference of species trees from information on multiple genetic loci, the evaluation of gene tree probabilities conditional on species trees [5].

Owing to uses of coalescent histories in sets over which sums are computed, as well as in state spaces of certain phylogenetic Markov chains [7], [10], [11], solutions to enumerative problems involving coalescent histories contribute to an understanding of the computational complexity of phylogenetic calculations. A recursion for the number of coalescent histories for a given gene tree and species tree has been established [13], and several studies have reported exact numerical results and closed-form expressions for the number of coalescent histories for small trees and for specific types of trees of arbitrarily large size [4], [5], [6], [13], [14], [15], [19]. The latter computations have proceeded both by

solving or deploying the recursion in specific cases [13], [14], [15], [19], as well as by identifying correspondences between coalescent histories and other combinatorial structures for which enumerative results have already been established [4], [5], [6].

One class of gene trees and species trees of particular interest for enumeration of coalescent histories is the *caterpillar-like families*, trees that have a caterpillar shape, except that the caterpillar subtree with r taxa is replaced by a subtree of size r that is not necessarily a caterpillar subtree (Fig. 1). For the simplest caterpillar-like family, the caterpillar trees themselves, if the gene tree and species tree have the same caterpillar labeled topology with n taxa, then, as reported in [5], the number of coalescent histories is a Catalan number,

$$c_{n-1} = \frac{1}{n} \binom{2n-2}{n-1}. \tag{1}$$

For T_r-caterpillar-like families, in which the r-taxon subtree of an n-taxon caterpillar species tree is replaced by an r-taxon subtree T_r (Fig. 1), by employing the recursion, Rosenberg [14] obtained the exact number of coalescent histories for all n, for each T_r with $r \leq 8$, in the case that the gene tree and species tree have the same labeled topology. Rosenberg [14] argued that in each of these cases, as $n \to \infty$, the number of coalescent histories is asymptotic to a constant multiple of the Catalan numbers. A proof of this result has been presented in full for each case with $r \leq 5$ [4], [13], [14], and by computer algebra for cases with $r = 6, 7$, and 8 [14].

Each case considered by [14] involved cumbersome computations specific to the choice of T_r, limiting the generality of the approach. While no reason exists to suspect that the method of [14] would not extend to larger r, it is
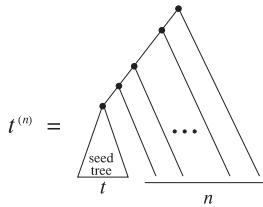
Fig. 1. A caterpillar-like family of species trees $(t^{(n)})_{n \geq 0}$. For a seed tree t, by adding $n \geq 0$ branches each with 1 leaf, we obtain the nth tree of the family, $t^{(n)}$. If t has two taxa, then $(t^{(n)})_{n \geq 0}$ is simply the caterpillar family.

desirable to find another method that is practical for a general T_r. Here, using a substantially different strategy that brings to studies of coalescent histories the methods of analytic combinatorics, we produce an enumeration result that covers caterpillar-like families in general. We show that the result of [14] applies to all caterpillar-like families, not only those for which T_r has $r \leq 8$. That is, we demonstrate that for any T_r, as $n \to \infty$, the number of coalescent histories in the T_r-caterpillar-like family is asymptotic to a constant multiple of the Catalan numbers—thus extending a result known only for $r \leq 8$ to arbitrarily large r. We describe a method and symbolic tool for computing the constant. Finally, we discuss the impact of the results in mathematical phylogenetics.

2 PRELIMINARIES

2.1 Species Trees and Coalescent Histories

We consider binary rooted leaf-labeled species trees, taking a single arbitrary labeling (without loss of generality) to represent a given unlabeled species tree topology. We consider an arbitrarily labeled species tree and its unlabeled tree interchangeably, treating the labeling as implicit.

We examine coalescent histories for the case in which gene trees and species trees have the same labeled topology t, terming a coalescent history in this case a *matching coalescent history*. To be a matching coalescent history, a mapping h from the internal nodes of t (viewed as the gene tree) to the branches of t (viewed as the species tree) must satisfy two conditions (Fig. 2): (a) for each leaf x in t, if x descends from node k in t, then x descends from branch $h(k)$ in t; (b) for each pair of internal nodes k_1 and k_2 in t, if k_2 descends from k_1 in t, then branch $h(k_2)$ descends from or coincides with branch $h(k_1)$ in t. We henceforth consider only matching coalescent histories, treating "matching" as implicit; we also refer simply to *histories* for short.

2.2 Caterpillar-Like Families of Species Trees

For a binary species tree t with at least two taxa, we denote by $(t^{(n)})_{n \geq 0}$ the caterpillar-like family generated by seed tree t. This family is recursively defined by taking $t^{(0)} = t$ and letting $t^{(n+1)}$ be the tree obtained by appending $t^{(n)}$ and a single leaf to a shared root (Fig. 1).

Our interest is in the number of matching coalescent histories of $t^{(n)}$ for $n \geq 0$, a quantity we denote by $h_n(t)$ or simply h_n. We note that whereas [14] indexed trees by their numbers of taxa, here n represents the number of taxa appended above the root of the seed tree, so that if seed tree t has $|t|$ taxa, then $|t| + n$ gives the number of taxa in $t^{(n)}$.
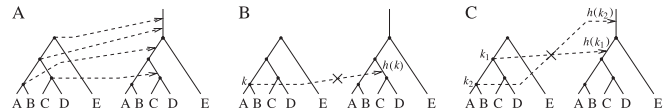

Fig. 2. Matching coalescent histories. (A) A matching coalescent history. (B) A mapping from the internal nodes of a tree to its branches that does not satisfy condition (a). Leaf B is descended from node k but does not descend from branch $h(k)$. (C) A mapping from the internal nodes of a tree to its internal branches that does not satisfy condition (b). Node k_2 is descended from node k_1, but branch $h(k_2)$ is strictly ancestral to branch $h(k_1)$.

2.3 Principles of Analytic Combinatorics

We rely on techniques of analytic combinatorics [8] to obtain our enumerative results, and recall several key points. In general, an integer sequence $(a_n)_{n \geq 0}$ can be associated with a formal power series $A(z) = \sum_{n=0}^{\infty} a_n z^n$, also termed the *generating function* of the integers a_n. Considering z as a complex variable, typically in a neighborhood of 0, features of the function $A(z)$ are related to the growth of the coefficients a_n.

More precisely, generating functions, considered as complex functions, enable analyses of the asymptotic growth of the associated integer sequences through the analysis of their singularities in the complex plane. In particular, under suitable conditions, there exists a general correspondence between the singular expansion of a generating function $A(z)$ near its dominant singularities—those nearest the origin—and the asymptotic behavior of the associated coefficients a_n (Chapter VI of [8]). We make use of theorems that describe this correspondence.

2.4 Catalan Numbers

The Catalan sequence appears often in combinatorics [8], [9], [17] and features prominently in our analysis. Rewriting Eq. (1) with index n rather than $n - 1$,

$$c_n = \frac{1}{n+1} \binom{2n}{n}. \tag{2}$$

The associated generating function is well known [17]:

$$C(z) = \sum_{n=0}^{\infty} c_n z^n = \frac{1 - \sqrt{1 - 4z}}{2z}. \tag{3}$$

By definition, if $[z^n]f(z)$ denotes the nth term in the power series expansion of $f(z)$ at $z = 0$, we have

$$\begin{aligned} c_n = [z^n]C(z) &= \frac{1}{2}[z^{n+1}](1 - \sqrt{1 - 4z}) \\ &= \frac{1}{2}[z^{n+1}](-\sqrt{1 - 4z}). \end{aligned} \tag{4}$$

Here, $1 - \sqrt{1 - 4z}$ is replaced by $-\sqrt{1 - 4z}$, as the constant 1 does not contribute to the power series expansion for terms of order $n + 1$, with $n \geq 0$. Asymptotically, applying Stirling's formula $n! \sim \sqrt{2\pi n}(n/e)^n$ to Eq. (2), the Catalan sequence satisfies

$$c_n \sim \frac{4^n}{n^{3/2}\sqrt{\pi}}. \tag{5}$$

3 THE NUMBER OF MATCHING COALESCENT HISTORIES FOR CATERPILLAR-LIKE FAMILIES

We aim to find a procedure that evaluates the number of coalescent histories $h_n(t)$ for matching gene trees and species trees in the caterpillar-like family that begins with seed tree t, and moreover, to show that

$$h_n(t) \sim \beta_t c_n, \tag{6}$$

where the multiplier $\beta_t > 0$ for the Catalan sequence is a constant depending on t. In other words, we wish to demonstrate that as $n \to \infty$, h_n/c_n converges to a constant $\beta_t > 0$ that depends on the seed tree t.

First, in Section 3.1, we determine a lower bound for the number of matching coalescent histories of the nth tree $t^{(n)}$ of the caterpillar-like family with seed tree t. Next, in Section 3.2, we introduce a concept of m-rooted histories of a species tree $t^{(n)}$. The section provides an iterative construction of the rooted histories of $t^{(n+1)}$ from those of $t^{(n)}$, describing the construction by means of a convenient labeling scheme. We follow a commonly used combinatorial enumeration strategy [2], [3] that determines a recursive succession rule for successive collections of objects in a sequence and then uses this rule to compute a generating function. In Section 3.3, we use the iterative construction to produce a bivariate generating function whose coefficients $h_{n,m}$ are the numbers of m-rooted histories for trees $t^{(n)}$. We next obtain the generating function for the integer sequence $(h_n)_{n \geq 0}$ describing the number of matching coalescent histories for the $t^{(n)}$. Finally, using the lower bound from Section 3.1, in Section 3.4, we apply methods of analytic combinatorics to study the asymptotic behavior of h_n.

3.1 Lower Bound for h_n

For our asymptotic analysis, we will need an initial lower bound for h_n. To produce this bound, we first define V as the tree with two taxa. Recalling that we index trees so that the number of taxa in a tree exceeds by n the number of taxa in the seed tree, we have [4], [13], [14]

$$h_n(\mathrm{V}) = c_{n+1}.$$

We can then use a constructive procedure, illustrated in detail in Fig. 3, to show that for any seed tree t with $|t| \geq 2$,

$$h_n(t) \geq h_n(\mathrm{V}) = c_{n+1}. \tag{7}$$

For a seed tree t, we can superimpose V on t so that the root r_{V} of V matches the root r_t of t (Fig. 3B). The two leaves of V are identified with two of the leaves of t, one on each side of the root of t. Generating caterpillar-like families by adding n single branches separately to V and to t, the superposition of V on t extends, so that $\mathrm{V}^{(n)}$ is superimposed on $t^{(n)}$ (Fig. 3C). The n caterpillar branches of $t^{(n)}$ and $\mathrm{V}^{(n)}$ then correspond.

Each matching coalescent history h of $t^{(n)}$ determines a corresponding matching coalescent history h' of $\mathrm{V}^{(n)}$ by considering the restriction of h to the set of internal nodes of $t^{(n)}$ that correspond to internal nodes of $\mathrm{V}^{(n)}$ (Fig. 3D). Thus, for any seed tree t, the number of matching coalescent histories of $t^{(n)}$ is greater than or equal to that of $\mathrm{V}^{(n)}$.
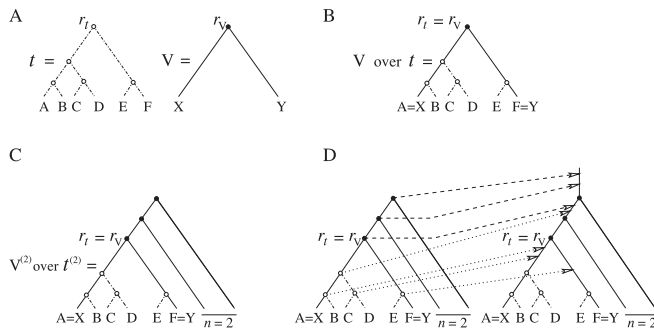

Fig. 3. Superposition of the caterpillar tree family on a caterpillar-like tree family with arbitrary seed tree of size $|t| \geq 2$. (A) A seed tree t and the seed tree V for the caterpillar family. (B) Superposition of V on t, so that the roots r_{V} and r_t overlap. (C) Superposition of $\mathrm{V}^{(2)}$ (shaded internal nodes) on $t^{(2)}$ (shaded and unshaded nodes). The $n = 2$ caterpillar branches in $\mathrm{V}^{(2)}$ and $t^{(2)}$ overlap, and r_{V} still matches r_t. (D) A matching coalescent history of $t^{(2)}$ (dashed and dotted arrows) determines a matching coalescent history of $\mathrm{V}^{(2)}$ (dashed arrows) by ignoring arrows from the unshaded nodes.

In symbols, we have Eq. (7). We will use this result in Section 3.4.

3.2 Iterative Generation of Rooted Histories

This section describes the iterative procedure that for a seed tree t eventually enables us to determine a formula for h_n. First, in Section 3.2.1, we discuss m-rooted histories, which extend the concept of matching coalescent histories, introducing an additional parameter m. Next, in Section 3.2.2, we examine the relationship between rooted histories and the *extended coalescent histories* of [13], importing results on extended coalescent histories into the more convenient framework of rooted histories. We expand our goal of enumerating matching coalescent histories for $t^{(n)}$, considering a more general problem of enumerating for $m \geq 1$ the m-rooted histories of $t^{(n)}$.

In Section 3.2.3, we define an operator Ω for constructing the rooted histories of $t^{(n+1)}$ from the rooted histories of $t^{(n)}$. Next, in Section 3.2.4, we introduce a labeling scheme that in Section 3.2.5 enables us to switch from counting rooted histories to counting multisets of labels. At the end of Section 3.2, we will have converted our enumeration problem into an enumeration that is more convenient for constructing a generating function.

3.2.1 m-Rooted Histories

Consider a tree t with $|t| \geq 2$, and suppose that the branch above the root of t (the *root-branch*) is divided into infinitely many components. A matching coalescent history mapping the internal nodes of t onto the branches of t is said to be m-*rooted* for $m \geq 1$ if the root of t is mapped *exactly* onto the mth component of the root (Fig. 4). It is said to be *rooted* if it is m-rooted for some m. Branches are numbered so that branch $m = 1$ is immediately above the root node, and m is greater for components that are farther from the root.

For a rooted history h of a tree t, $m = m(h)$ denotes the component of the root-branch of t that receives the image of the root of t. $H_{n,m}(t)$ denotes the set of m-rooted histories of $t^{(n)}$, and $H_n(t) = \bigcup_{m=1}^{\infty} H_{n,m}(t)$ the set of its rooted histories.
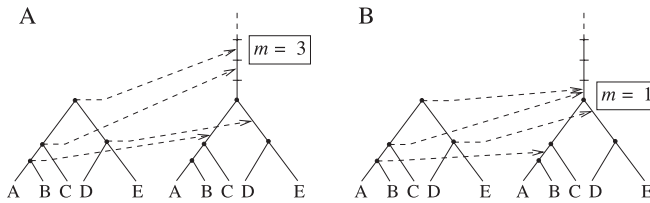
Fig. 4. Rooted histories of a tree. (A) A 3-rooted history. The root-branch is divided into infinitely many components, the third of which receives the image of the root. (B) A 1-rooted history. The number of 1-rooted histories corresponds to the number of matching coalescent histories of the tree.

The number of m-rooted histories of $t^{(n)}$ is $h_{n,m} = |H_{n,m}|$, and the number of 1-rooted histories $h_n = h_{n,1}$ is also the number of matching coalescent histories. Enumerating the matching coalescent histories of $t^{(n)}$ is equivalent to enumerating its 1-rooted histories.

3.2.2   Rooted Histories and Extended Histories

Rooted histories are closely related to *extended* coalescent histories, as defined by [13]. We use this relationship to study properties of rooted histories. Rosenberg [13] defined the set of *k-extended* coalescent histories of a tree t with $|t| \geq 1$ for integers $k \geq 1$; we also consider $k = 0$ by setting the number of 0-extended histories to 0.

A k-extended history is defined as a coalescent history for a species tree whose root-branch is divided into exactly $k \geq 0$ parts. In other words, the root-branch has exactly $k \geq 0$ possible components onto which a k-extended history can map the gene tree root. Here we consider matching k-extended histories, so that the internal nodes of a tree t are mapped to the branches of t and its k components above the root. For convenience, we refer to extended histories by the index k, reserving the index m for rooted histories.

By the definitions of k-extended and m-rooted histories, for each $k \geq 0$, the set of k-extended histories of a tree is exactly the set of all m-rooted histories with $1 \leq m \leq k$. Therefore, for a tree t with at least 2 leaves, if we label by $e_{t,k}$ its number of k-extended histories, then for each $m \geq 1$ the number of m-rooted histories of t is

$$h_{0,m} = e_{t,m} - e_{t,m-1}. \tag{8}$$

Note that for $m = 1$, we explicitly use in Eq. (8) the fact that $e_{t,0}$ is defined and equal to 0. In addition to setting $e_{t,0} = 0$ for any tree t, as in [13] we set $e_{t,k} = 1$ for all $k \geq 1$ in the case that t has exactly 1 leaf.

Suppose $|t| \geq 1$ and $k \geq 0$. Denote by t_L and t_R the left and right subtrees of the root of t. We can compute $e_{t,k}$ recursively as in Theorem 3.1 of [13]:

$$e_{t,k} = \begin{cases} 0 & \text{if } |t| \geq 1 \text{ and } k = 0 \\ 1 & \text{if } |t| = 1 \text{ and } k \geq 1 \\ \sum_{i=1}^{k} e_{t_L,i+1} e_{t_R,i+1} & \text{if } |t| \geq 2 \text{ and } k \geq 1. \end{cases} \tag{9}$$

As was already observed in the remarks following Corollary 3.2 of [13], by Eq. (9), for any tree t with $|t| \geq 1$, for positive integers $k \geq 1$, the function $f(k) = e_{t,k}$ is a polynomial in k. With our extension to permit $k = 0$, we can extend this fact to $k \geq 0$ for $|t| \geq 2$: for any tree t with $|t| \geq 2$, and for

$k \geq 0$, we claim that the function $f(k) = e_{t,k}$ is a polynomial in k. Note that in allowing $k = 0$, we claim $e_{t,k}$ is a polynomial in k only for $|t| \geq 2$; for $|t| = 1$, $e_{t,k}$ is not a polynomial in k because $e_{t,0} = 0$ and $e_{t,k} = 1$ for $k \geq 1$.

To prove the claim, fix t with $|t| \geq 2$ and consider the variable k over domain $[1, \infty)$. We demonstrate that $f(k)$ is a polynomial in k for domain $[0, \infty)$ by showing that the closed form for $f(k)$ has a factor of k, so that our choice $e_{t,0} = 0$ in Eq. (9) is compatible with the polynomial expression valid for $k \geq 1$.

Observe that for $i \geq 1$, $e_{t_L,i}$ and $e_{t_R,i}$ are polynomials in i, say $P_{t_L}(i)$ and $P_{t_R}(i)$. Replacing terms $e_{t_L,i+1}$ and $e_{t_R,i+1}$ in the recursion in Eq. (9) by polynomials $P_{t_L}(i+1)$ and $P_{t_R}(i+1)$, we obtain

$$\sum_{i=1}^{k} e_{t_L,i+1} e_{t_R,i+1} = \sum_{i=1}^{k} P_{t_L}(i+1) P_{t_R}(i+1) = \sum_{i=1}^{k} P'(i), \tag{10}$$

where $P'(i)$ denotes a polynomial in i that results from the product of $P_{t_L}(i+1)$ and $P_{t_R}(i+1)$. By Faulhaber's formula for sums of powers of integers, symbolic sums of the form $\sum_{i=1}^{k} i^p$ for a fixed integer $p \geq 0$ are polynomials containing a factor of k in their closed forms (Section 6.5 of [9])—for example, $\sum_{i=1}^{k} i^3 = k^2(k+1)^2/4$. Thus, because the polynomial $P'(i)$ is a linear combination of terms of the form i^p, the closed-form expression for the sum $\sum_{i=1}^{k} P'(i)$ appearing in Eq. (10) also has a factor of k. It therefore has a value of 0 at $k = 0$.

Functions $e_{t,k}$ for trees t with $1 \leq |t| \leq 9$ and $k \geq 1$ appear in Tables 1, 2, 3, and 4 of [13]. For $|t| \geq 2$, as we have shown, these example polynomials are divisible by the variable representing the number of components of the root-branch. By Eq. (8), we immediately obtain the following result.

Proposition 1. *For any tree t with $|t| \geq 2$ and for $m \geq 1$, the number $h_{0,m}$ of m-rooted histories of t is a polynomial in m that can be computed by the difference in Eq. (8) using $e_{t,k}$ as in Eq. (9).*

As an example of Proposition 1, consider the tree $t = ((A, B), (C, D))$, identifying this arbitrary labeling with the unlabeled tree $(()())$. By applying the recursive procedure in Eq. (9), we find that for $k \geq 0$, the number of k-extended coalescent histories for t is $e_{t,k} = \frac{1}{6} k(2k^2 + 9k + 13)$ [13]. The difference Eq. (8) yields that for $m \geq 1$, the number of m-rooted histories of t is $h_{0,m} = e_{t,m} - e_{t,m-1} = m^2 + 2m + 1$.

3.2.3   Rooted Histories of $t^{(n+1)}$ from Those of $t^{(n)}$

This section introduces an operator Ω that generates the rooted histories of $t^{(n+1)}$ from those of $t^{(n)}$. For each rooted history h' of $t^{(n+1)}$, there exists exactly one rooted history h of $t^{(n)}$ with $h' \in \Omega(h)$. Recalling the definitions of the sets $H_{n,m}(t)$ and $H_n(t)$ of m-rooted and rooted histories of $t^{(n)}$, we define Ω as follows.

Definition. *Let $\mathcal{P}(X) = \{x : x \subseteq X\}$ denote the power set of set X, and fix tree t. The operator Ω is a function*

$$\Omega : H_n(t) \to \mathcal{P}(H_{n+1}(t)),$$

where for a rooted history $h \in H_n(t)$, $\Omega(h)$ is the set of rooted histories $h' \in H_{n+1}(t)$ for which the restriction of h' to $t^{(n+1)}$ excluding its most basal caterpillar branch coincides with the rooted history h of $t^{(n)}$.

Denote by $b_1, b_2, \ldots, b_{n+1}$ the caterpillar branches in $t^{(n+1)}$, from the least basal b_1 to the most basal b_{n+1} (Fig. 5). Upon removal of the most basal caterpillar branch b_{n+1} from $t^{(n+1)}$, the root of $t^{(n+1)}$—to which branch b_{n+1} is attached—is replaced by a demarcation between the first and second components of the root-branch of $t^{(n)}$. For instance, in Fig. 5A, starting from tree $t = ((A, B), (C, D))$, we consider h''', a 3-rooted history of $t^{(3)}$. By removing the most basal caterpillar branch b_3 of $t^{(3)}$, we reduce to the 1-rooted history h'' of $t^{(2)}$ (Fig. 5B). Next, by removing the caterpillar branch b_2 of $t^{(2)}$, we reduce to the 2-rooted history h' of $t^{(1)}$ (Fig. 5C). By removing the remaining caterpillar branch b_1 from $t^{(1)}$, we reduce to the 2-rooted history h of $t = t^{(0)}$ (Fig. 5D). Therefore, by the definition of Ω, we have $h' \in \Omega(h), h'' \in \Omega(h')$, and $h''' \in \Omega(h'')$.

By definition, Ω has the property that for each rooted history $h' \in H_{n+1}(t)$, with $n \geq 0$, there exists exactly one rooted history $h \in H_n(t)$ such that $h' \in \Omega(h)$. In other words, for each $n \geq 0$, the set of rooted histories $H_{n+1}(t)$ can be partitioned as a disjoint union,

$$H_{n+1}(t) = \bigsqcup_{h \in H_n(t)} \Omega(h). \quad (11)$$

The set $H_{n+1}(t)$ is therefore generated without double occurrences of any rooted history by applying Ω to the rooted histories in $H_n(t)$. It follows immediately that in performing n iterations of Ω to obtain $\Omega[\ldots [\Omega[\Omega[H_0]]] \ldots]$ from the set H_0 of rooted histories of $t^{(0)}$, all the rooted histories of $t^{(n)}$ are generated exactly once.

3.2.4 Labels for Rooted Histories

The operator Ω, starting from the rooted histories of $t^{(n)}$, generates the rooted histories of $t^{(n+1)}$. In this section, we introduce a labeling scheme, giving each m-rooted history h of $t^{(n)}$ a label $L(h) = (n, m)$. We then describe how Ω acts on the labels of the rooted histories, characterizing the set of labels $L[\Omega(h)] = \{L(h') : h' \in \Omega(h)\}$. Our goal is to represent each set H_n of rooted histories of $t^{(n)}$ by the multiset of its labels, reducing the enumeration of $|H_{n,m}|$ to the problem of counting certain ordered pairs (n, m) iteratively generated by simple rules that reflect how the rooted histories in H_{n+1} are generated according to rule Ω from the rooted histories in H_n by Eq. (11).

In our labeling, each rooted history $h \in H_n(t)$ that maps the root of $t^{(n)}$ onto the mth component of the root-branch of $t^{(n)}$ receives label $L(h) = (n, m)$. Enumeration of $h_n = |H_{n,1}|$ then reduces to enumeration of those rooted histories labeled by $(n, 1)$.

Note that a label (n, m) does not uniquely specify an m-rooted history of $t^{(n)}$: a tree $t^{(n)}$ has in general many m-rooted histories, each receiving the label (n, m). In other words, if $h, \overline{h} \in H_n(t)$ and $L(h) = L(\overline{h})$, then h and $\overline{h}$ are not necessarily the same rooted history of $t^{(n)}$. We will, however,
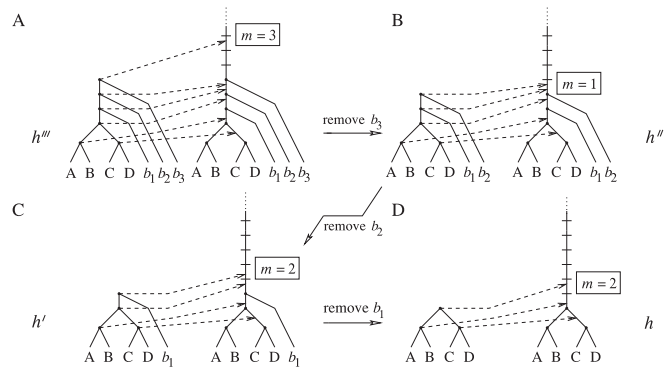

Fig. 5. The relationships among rooted histories for sequential members of caterpillar-like families. For a rooted history h''' of $t^{(3)}$, with $t = ((A, B), (C, D))$, the figure sequentially removes caterpillar branches. By definition, a rooted history h' of $t^{(n+1)}$ belongs to the set $\Omega(h)$ if, by removing the most basal caterpillar branch b_{n+1} in $t^{(n+1)}$, we recover the rooted history h of $t^{(n)}$. Note that when we remove the basal caterpillar branch b_{n+1} from $t^{(n+1)}$, the root of $t^{(n+1)}$—to which the branch b_{n+1} is attached—becomes the boundary between the first and second components of the root-branch of $t^{(n)}$, and is depicted as a horizontal segment. (A) $h''' \in \Omega(h'')$. (B) $h'' \in \Omega(h')$. (C) $h' \in \Omega(h)$. (D) h. For each rooted history, the value of the parameter m, representing the component of the root-branch that receives the image of the root, is shown.

consider for $n \geq 0$ *multisets* of labels in which we find a copy of the label (n, m) for each m-rooted history of $t^{(n)}$.

To characterize how the operator Ω acts on the labels for rooted histories, consider an m-rooted history $h \in H_n(t)$, so that h maps the root of $t^{(n)}$ onto the mth component of the root-branch of $t^{(n)}$. This history is labeled $L(h) = (n, m)$. For instance, taking the seed tree $t = ((A, B), (C, D))$, the history h of $t = t^{(0)}$ depicted in Fig. 6A is labeled $L(h) = (0, 3)$, whereas the history h of $t^{(1)}$ in Fig. 6C has $L(h) = (1, 1)$.

By applying Ω to a history h of $t^{(n)}$ with $L(h) = (n, m)$, we produce a set of rooted histories $\Omega(h) \subseteq H_{n+1}(t)$. The set of labels for $\Omega(h)$,

$$L[\Omega(h)] = \{L(h') : h' \in \Omega(h)\},$$

is determined according to the rule:

$$L[\Omega(h)] = \begin{cases} \{(n+1, m') : m' \geq m\} & \text{if } m = 1 \\ \{(n+1, m') : m' \geq m - 1\} & \text{if } m \geq 2, \end{cases} \quad (12)$$

where m' denotes the value of the parameter m—the component of the root-branch of $t^{(n+1)}$ to which the root is mapped—for the rooted histories $h' \in \Omega(h)$ of $t^{(n+1)}$.

The rule in Eq. (12) distinguishes between two cases depending on whether the value of the parameter $m = m(h)$ of the generating rooted history h is equal to or exceeds 1. In both cases, the set $L[\Omega(h)]$ contains infinitely many labels, each with its first component equal to $n + 1$, as the labels refer to rooted histories of $t^{(n+1)}$. The value of the second component m' ranges in $[m - 1, \infty)$ if $m \geq 2$, and in $[1, \infty)$ if $m = 1$.

Recall that according to the definition of Ω, from an m-rooted history h of $t^{(n)}$ (Figs. 6A and 6C), we generate an m'-rooted history $h' \in \Omega(h)$ of $t^{(n+1)}$ (Figs. 6B and 6D) by (i) choosing the component m' of the root-branch of $t^{(n+1)}$ onto which h' maps the root of $t^{(n+1)}$, and (ii) letting h'
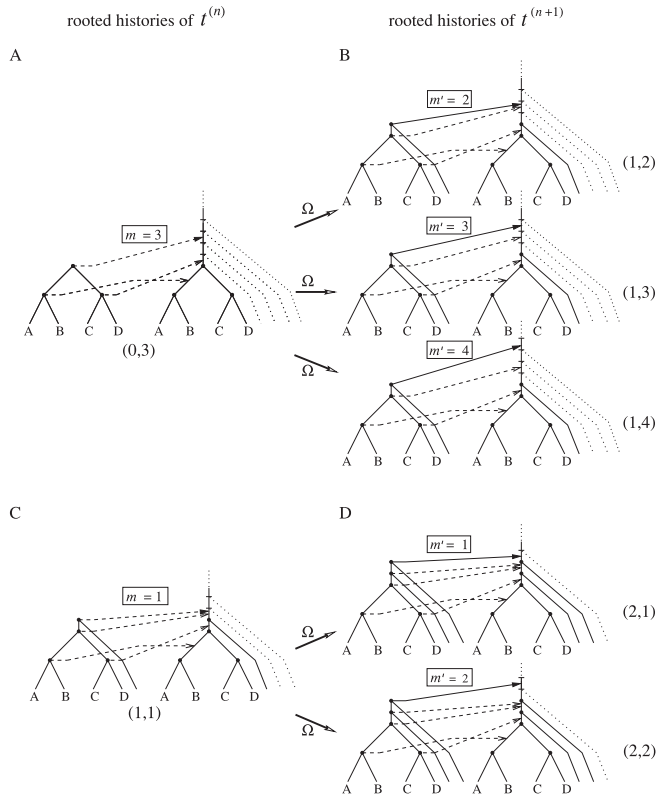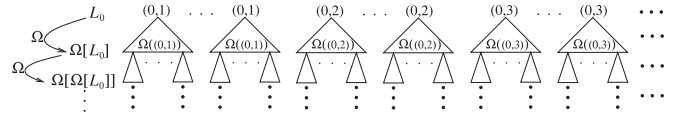
Fig. 7. Iterative application of a rule for generating the multiset of the labels of the rooted histories of a tree $t^{(n)}$. The iterative procedure starts with the multiset L_0 that contains those labels of the form $\{(0, m) : m \geq 1\}$ associated with the rooted histories of a seed tree $t = t^{(0)}$. In the first step of the iteration, we apply Ω (Eq. (13)) to each label of L_0. In the second step, we apply Ω to each label resulting from the first step, and so on. The number of m-rooted histories of $t^{(n)}$ corresponds to the number of labels (n, m), considered with their multiplicity, generated after the nth step of the iteration.

Fig. 6. Generation of rooted histories of $t^{(n+1)}$ from rooted histories of $t^{(n)}$, as given by rule Ω applied to seed tree $t = ((A, B), (C, D))$. To obtain rooted histories of $t^{(n+1)}$ (right) from rooted histories of $t^{(n)}$ (left), we choose the component m' of the root-branch of $t^{(n+1)}$ onto which the root of $t^{(n+1)}$ is mapped (solid arrows). The smallest among infinitely many possible choices are depicted. For all nodes of $t^{(n+1)}$ except the root, the rooted history generated for $t^{(n+1)}$ coincides with the generating rooted history of $t^{(n)}$ (dashed arrows). (A) A case with $m \geq 2$. A 3-rooted history h of $t^{(0)}$, labeled $(0, 3)$, is shown. (B) $\Omega(h)$ for h in (A). 2-, 3-, and 4-rooted histories of $t^{(1)}$ belonging to $\Omega(h)$ are shown and are labeled $(1, 2)$, $(1, 3)$, and $(1, 4)$, respectively. Because $m \geq 2$, $m' \geq m - 1$ as in eq. (12). (C) A case with $m = 1$. A 1-rooted history h of $t^{(1)}$, labeled $(1, 1)$, is shown. (D) $\Omega(h)$ for h in (C). 1- and 2-rooted histories of $t^{(2)}$ belonging to $\Omega(h)$ are shown and are labeled $(2, 1)$ and $(2, 2)$, respectively. Because $m = 1$, $m' \geq m$.

coincide with h on all nodes of $t^{(n+1)}$ except the root. The rooted history h' coincides with h once we remove the most basal caterpillar branch of $t^{(n+1)}$.

Fig. 6 illustrates both cases of Eq. (12). In step (i), infinitely many choices of m' are possible, because the root-branch of $t^{(n+1)}$ is divided into infinitely many parts. The most basal caterpillar branch in $t^{(n+1)}$ is attached at the border between the first and second components of the root-branch of $t^{(n)}$. Thus, the addition of the $(n + 1)$st caterpillar branch eliminates a component of the root-branch, so that if the starting rooted history h has $m \geq 2$ (Fig. 6A), then the root of $t^{(n)}$ maps to component $m - 1$ of the root-branch of $t^{(n+1)}$. The root of $t^{(n+1)}$ can map to this same branch, or to any branch m' with $m' \geq m - 1$. For instance, in Fig. 6B, one of the rooted histories h' generated by a rooted history h with $m = 3$ has $m' = m - 1 = 2$.

If h has $m = 1$, however, then production of h' is slightly different (Fig. 6C). By definition, the parameter m for a rooted history cannot be smaller than 1. The value $m' = m - 1$ is not permitted, and m' remains greater than or equal to $m = 1$ (Fig. 6D).

3.2.5   Counting the Labels of Rooted Histories

The labeling scheme in Section 3.2.4 encodes the application of the operator Ω to the rooted histories of $t^{(n)}$. Now that we have described the set of labels $L[\Omega(h)]$ arising from the label $L(h)$ according to the rule in Eq. (12), the problem of counting a set of rooted histories becomes a problem of counting the set of the associated labels along with their multiplicities—or the *multiset* of the labels.

For $n \geq 0$ and $m \geq 1$, we use $\Omega((n, m))$ to denote, with an abuse of notation, the set of labels $L[\Omega(h)]$ when $L(h) = (n, m)$. Recalling that iterative application of Ω to the rooted histories H_0 of tree $t^{(0)}$ generates the rooted histories H_n of $t^{(n)}$, the enumeration of $|H_{n,m}|$ for tree $t = t^{(0)}$ becomes a problem of counting those labels of the form (n, m) that are generated when we iteratively apply the operator Ω as $\Omega[\ldots [\Omega[\Omega[L_0]]] \ldots]$ starting from the multiset of labels $L_0 = \{L(h) : h \in H_0(t)\}$ (Fig. 7).

Eq. (12) characterizes the set of labels $L[\Omega(h)]$ of the rooted histories in $\Omega(h)$ in terms of the label $L(h)$ of rooted history h. If $L(h) = (n, m)$, then $\Omega((n, m))$ denotes the set of labels $L[\Omega(h)]$. Thus, converting the notation from histories to labels, Eq. (12) becomes

$$\Omega((n, m)) = \begin{cases} \{(n + 1, m') : m' \geq m\} & \text{if } m = 1 \\ \{(n + 1, m') : m' \geq m - 1\} & \text{if } m \geq 2. \end{cases} \quad (13)$$

For the seed tree t, we count $h_{n,m} = |H_{n,m}|$ by evaluating the number of occurrences of the ordered pair (n, m) in the multiset L_n defined as

$$L_n = L[H_n(t)] = \{L(h) : h \in H_n(t)\}. \quad (14)$$

In symbols, we have

$$h_{n,m} = |\{\ell \in L_n : \ell = (n, m)\}|. \quad (15)$$

By Eq. (11), each multiset L_n is generated iteratively (Fig. 7). We start with the multiset of labels

$$L_0 = \{L(h) : h \in H_0(t)\}. \quad (16)$$

For each $n \geq 0$, the multiset L_{n+1} is obtained as

$$L_{n+1} = \biguplus_{(n,m) \in L_n} \Omega((n, m)), \quad (17)$$

where the symbol $\uplus$ denotes the union operator for multi-sets. Thus, in $M = M_1 \uplus M_2$, if an element x appears n_1 times in M_1 and n_2 times in M_2, then it appears $n_1 + n_2$ times in M. Eq. (17) provides an iterative generation of the labels for the rooted histories of $H_{n+1}(t)$ from the labels of the rooted histories of $H_n(t)$, retaining information about the multiplicity of occurrences of each label.

3.3 Rooted Histories and Generating Functions

We have now obtained Eq. (15), which gives an equivalence between the number of m-rooted histories of $t^{(n)}$ and the number of labels (n, m) in the multiset L_n, and Eqs. (16) and (17), which give through Ω (Eq. (13)) an iterative procedure that generates the family of multisets $(L_n)_{n \geq 0}$. In this section, we translate the iterative procedure into algebraic terms, determining the generating function associated with the integer sequence $(h_n)_{n \geq 0}$.

First, in Section 3.3.1, we characterize a generating function $g(y)$ for the sequence $(h_{0,m})_{m \geq 1}$. Next, in Section 3.3.2, we deduce an equation satisfied by the bivariate generating function $F(y, z)$ for $(h_{n,m})_{n \geq 0, m \geq 1}$. In Section 3.3.3, we solve the equation, obtaining the desired generating function $f(z)$ for the sequence $(h_{n,1})_{n \geq 0}$. This generating function can be written in turn as a function of $g(y)$.

3.3.1 Generating Function for $(h_{0,m})_{m \geq 1}$

In this section, we characterize the generating function $g(y)$ that counts for a given seed tree t the labels in the multiset L_0 describing the labels of the rooted histories of t.

Fix the seed tree t. Recalling the equivalence in Eq. (15), define the generating function

$$g(y) = \sum_{(0,m) \in L_0} y^m = \sum_{m=1}^{\infty} h_{0,m} y^m, \qquad (18)$$

the mth coefficient of whose power series expansion provides the number $h_{0,m}$ of labels $(0, m)$ appearing in L_0. By Proposition 1, $h_{0,m}$ can be expressed as a polynomial in the variable m and can thus be decomposed as a finite linear combination of terms of the form m^k, where k is a non-negative integer. That is, for a certain finite set of non-negative integers with largest element K,

$$h_{0,m} = \sum_{k=0}^{K} w_k m^k, \qquad (19)$$

where the w_k are constants.

We introduce generating functions g_{m^k}, one for each k from 0 to K, in which the mth coefficient is m^k:

$$g_{m^k}(y) = \sum_{m=1}^{\infty} m^k y^m. \qquad (20)$$

Because K is finite, the desired generating function $g(y)$ can be written as a finite linear combination of this new collection of generating functions $g_{m^0}(y), g_{m^1}(y), \ldots, g_{m^K}(y)$. More precisely, by substituting in Eq. (18) the polynomial in Eq. (19) and switching the order of summation, we obtain

$$g(y) = \sum_{k=0}^{K} w_k g_{m^k}(y). \qquad (21)$$

We now state a lemma that characterizes the generating functions $g_{m^k}(y)$.

Lemma 1. *For each non-negative integer k from 0 to K, the generating function $g_{m^k}(y)$ in Eq. (20) is rational with denominator $(1 - y)^{k+1}$. That is, $g_{m^k}(y)$ has the form*

$$g_{m^k}(y) = \frac{P(y)}{(1 - y)^{k+1}},$$

where $P(y)$ is a polynomial in y.

Proof. We proceed by induction on k. If $k = 0$, then by Eq. (20), $g_{m^0}(y) = 1/(1 - y) - 1 = y/(1 - y)$. Assume the inductive hypothesis for $g_{m^k}(y)$. Applying Eq. (20) to $g_{m^{k+1}}(y)$, we can recover $g_{m^{k+1}}(y)$ as

$$g_{m^{k+1}}(y) = y \frac{\partial g_{m^k}(y)}{\partial y}, \qquad (22)$$

which by the quotient rule for derivatives is a rational function with denominator $(1 - y)^{k+2}$. $\qquad \square$

The proof of the lemma gives a recursive procedure in Eq. (22) to compute the functions $g_{m^k}(y)$. By Eq. (21), we immediately obtain from the lemma a result about the generating function $g(y)$.

Proposition 2. *The generating function $g(y)$ whose mth coefficient $[y^m]g(y)$ is the number of m-rooted histories $h_{0,m}$ of a seed tree t can be written as a finite linear combination*

$$g(y) = \sum_{j=1}^{J} q_j \frac{y^{a_j}}{(1 - y)^b}, \qquad (23)$$

where $b \geq 1$ and $J \geq 1$ are positive integers, each a_j is a non-negative integer, and the q_j are constants.

As an example, we show how the procedure in Proposition 2 can determine the generating function $g(y)$ for $t = ((A, B), (C, D))$, the same example seed tree for which we computed the polynomial $h_{0,m}$ via Proposition 1. Recall from Section 3.2.2 that $h_{0,m} = m^2 + 2m + 1$. To obtain the generating function $g(y)$ that has coefficients $[y^m]g(y) = m^2 + 2m + 1$, we sum generating functions for monomials m^2, $2m$, and 1. We already know $g_{m^0}(y)$, and by applying Eq. (22), we have

$$g_{m^0}(y) = \frac{y}{1 - y}$$

$$g_{m^1}(y) = y \frac{\partial g_{m^0}(y)}{\partial y} = \frac{y}{(1 - y)^2}$$

$$g_{m^2}(y) = y \frac{\partial g_{m^1}(y)}{\partial y} = \frac{y(y + 1)}{(1 - y)^3}.$$

Thus,

$$g(y) = g_{m^0}(y) + 2g_{m^1}(y) + g_{m^2}(y) = \frac{y^3 - 3y^2 + 4y}{(1 - y)^3}. \qquad (24)$$

In Eq. (24), $g(y)$ is written as in Eq. (23), taking $b = 3$, $J = 3$, $(a_1, a_2, a_3) = (1, 2, 3)$, and $(q_1, q_2, q_3) = (4, -3, 1)$.

3.3.2   Bivariate Generating Function for $(h_{n,m})_{n \geq 0, m \geq 1}$

Given t, the polynomial nature of $h_{0,m}$ in m enabled us to obtain a generating function for $h_{0,m}$. We now use the iterative procedure in Eq. (17) to determine an equation that characterizes the bivariate generating function with coefficients $h_{n,m}$. We represent each label of the form (n, m) by a symbolic algebraic expression in the variables y and z, so that (n, m) is replaced by $z^n y^m$. Let $L = \cup_{n=0}^{\infty} L_n$ be the multiset of all m-rooted histories for all trees $t^{(n)}$. Considering y and z as complex variables in two sufficiently small neighborhoods of 0, we aim to characterize the bivariate function $F(y, z)$ that admits the expansion

$$F(y, z) = \sum_{(n,m) \in L} z^n y^m,$$

where the sum is over all labels in the multiset L and thus has a term for each m-rooted history of each $t^{(n)}$. In particular, the function $F(y, z)$ is the bivariate generating function of the integers $h_{n,m}$, and its Taylor expansion can be written as

$$F(y, z) = \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} h_{n,m} z^n y^m, \tag{25}$$

where the coefficients $h_{n,m}$ appear explicitly.

By differentiating $F(y, z)$ with respect to y and then taking $y = 0$, we obtain

$$\frac{\partial F}{\partial y}(0, z) = \sum_{n=0}^{\infty} h_{n,1} z^n. \tag{26}$$

Thus, for each $n \geq 0$, we have

$$h_n = h_{n,1} = [z^n] \left(\frac{\partial F}{\partial y}(0, z) \right).$$

By representing each label of the form (n, m) by the symbolic expression $z^n y^m$ and assuming the complex variables y and z are sufficiently close to 0, the recursive generation in Eq. (17) of the multisets of labels $L_0, L_1, L_2, \ldots$ determines an equation for $F(y, z)$, demonstrated in Appendix 1:

$$F(y, z) \left[1 - \frac{z}{y(1 - y)} \right] = g(y) - z \frac{\partial F}{\partial y}(0, z). \tag{27}$$

Eq. (27) holds if the complex variables y and z are in two sufficiently small neighborhoods of 0, and it characterizes the generating function $F(y, z)$.

3.3.3   Generating Function for $(h_{n,1})_{n \geq 0}$

We now have an equation satisfied by the bivariate generating function $F(y, z)$. Further, we have Eq. (26), which demonstrates that the desired generating function for the sequence $(h_n)_{n \geq 0}$ is obtained from $\frac{\partial F}{\partial y}(0, z)$. By applying the *kernel method* [2], [12], we can determine the power series $\frac{\partial F}{\partial y}(0, z)$ from Eq. (27).

The idea of the method consists of coupling the two variables (z, y) as $(z, y(z))$ in such a way that two conditions

hold. First, (i) substituting $y = y(z)$ cancels the *kernel* of the equation, that is, the factor $1 - z/[y(1 - y)]$ on the left-hand side of Eq. (27). Second, (ii) for z near 0, the value of $y(z)$ remains in a sufficiently small neighborhood of $y = 0$, so that Eq. (27) still holds near $z = 0$ after substituting $y = y(z)$. This condition is required, as the power series expansion in Eq. (25) for $F(y, z)$ has been assumed to be valid in a neighborhood of $(y, z) = (0, 0)$, and the derivation of Eq. (27) relies on the fact that y and z are sufficiently close to 0. If the two conditions hold, then

$$z \frac{\partial F}{\partial y}(0, z) = g(y(z)),$$

so that $g(y(z))$ must be a power series for $z = 0$, because so must be $z \frac{\partial F}{\partial y}(0, z)$.

The required substitution couples y and z in such a way that $1 - z/[y(1 - y)] = 0$, so that $y(z) = (1 \pm \sqrt{1 - 4z})/2$. To determine whether to take the negative root $y_1(z)$ or the positive root $y_2(z)$, we note that if z is near 0, then $y_1(z)$ approaches 0, so that $y_1(z)$ lies in a neighborhood of $y = 0$ and $g(y_1(z))$ admits a power series expansion for z near 0. For $y_2(z)$, however, if z is near 0, then $y_2(z)$ approaches 1, and thus, $g(y_2(z))$ is not a power series for z near 0 due to the pole of the function $g(y)$ at $y = 1$ (Proposition 2). The only solution satisfying both (i) and (ii) is consequently

$$Y(z) = y_1(z) = \frac{1 - \sqrt{1 - 4z}}{2}, \tag{28}$$

which, with the generating function $C(z)$ of the Catalan numbers as in Eq. (3), satisfies $Y(z) = zC(z)$. Substituting $y = Y(z)$ in Eq. (27), we have $\frac{\partial F}{\partial y}(0, z) = g(Y(z))/z$, yielding the following result.

Proposition 3. *Fix tree t. Let $g(y)$ be the generating function associated with the polynomial $h_{0,m}$ (Eq. (18)). Let $Y(z)$ be as in Eq. (28). Then the generating function $f(z) = \sum_{n=0}^{\infty} h_n z^n$ is given by*

$$f(z) = \frac{\partial F}{\partial y}(0, z) = \frac{g(Y(z))}{z} = \frac{g\left(\frac{1 - \sqrt{1 - 4z}}{2} \right)}{z}. \tag{29}$$

The proposition thus determines the generating function $f(z) = g(Y(z))/z$ for the integer sequence describing the number of matching coalescent histories of species trees in the caterpillar-like family $(t^{(n)})_{n \geq 0}$. The function g depends on the seed tree t, whereas $Y(z)$ is fixed in Eq. (28) and does not depend on t.

As an example, recall that for $t = ((A, B), (C, D))$, in Eq. (24), we have computed the generating function g for the number $h_{0,m}$ of m-rooted histories of $t = t^{(0)}$. By Proposition 3, the generating function for the number h_n of matching coalescent histories of $t^{(n)}$ is

$$f(z) = \sum_{n=0}^{\infty} h_n z^n = \frac{g\left(\frac{1 - \sqrt{1 - 4z}}{2} \right)}{z}$$

$$= \frac{4(1 - \sqrt{1 - 4z})(3 - z + \sqrt{1 - 4z})}{z(1 + \sqrt{1 - 4z})^3}.$$

Taking the Taylor expansion of f, we obtain

$$f(z) = 4 + 13z + 42z^2 + 138z^3 + 462z^4 + 1{,}573z^5$$
$$+ 5{,}434z^6 + 19{,}006z^7 + 67{,}184z^8 + \dots \tag{30}$$

The coefficients h_n accord with the enumeration of matching coalescent histories reported in Corollary 3.9 of [13] and Table 3 of [14] for caterpillar-like families with seed tree $t = ((A, B), (C, D))$, except that those results tabulated numbers of coalescent histories by the number of taxa, whereas here, we use the index of the caterpillar-like family. Thus, in this example, the coefficient of z^n gives the number of matching coalescent histories for a tree with $n + 4$ taxa, as $|t| = 4$. Shifting the index in the formula from [13], [14] to agree with our indexing scheme, we obtain $[(5(n+4) - 12)/(4(n+4) - 6)]c_{(n+4)-1} = [(5n + 8)/(4n + 10)]c_{n+3}$ for the number of matching coalescent histories of $t^{(n)}$. This formula gives precisely the coefficients in the Taylor expansion in Eq. (30).

3.4 Asymptotic Behavior of h_n

From Proposition 3, we have the generating function f that counts matching histories of $t^{(n)}$ for a given fixed seed tree t. Applying techniques of analytic combinatorics as introduced in Section 2.3, we can determine the asymptotic behavior of the coefficients of the generating function

$$\tilde{f}(z) = \sum_{n=1}^{\infty} h_{n-1} z^n = z f(z) = g(Y(z)), \tag{31}$$

with $Y(z)$ as in Eq. (28). To simplify notation, we work with $\tilde{f}$ instead of f.

First, in Section 3.4.1, we obtain an asymptotic equivalence between h_n and $\beta_t c_n$, where β_t is a constant depending on the seed tree t, and the c_n are the Catalan numbers (Eq. (1)). Next, in Section 3.4.2, we produce a general procedure to determine the constants β_t, employing this procedure to obtain values of β_t for all seed trees t with $|t| \le 9$. We demonstrate that our values of β_t accord with constant multiples of the Catalan numbers previously obtained according to a different method [14] for seed trees with $|t| \le 8$.

3.4.1 A General Asymptotic Result

Recall that given t, Proposition 2 provides a procedure to determine the rational function g in Eq. (31). Writing g as the finite linear combination in Eq. (23), the values of b, J, and the $(a_j)_{1 \le j \le J}$ and $(q_j)_{1 \le j \le J}$ can all be computed.

As noted in Section 2.3, the expansion of $\tilde{f}$ at its dominant singularity characterizes the asymptotic behavior of the coefficients h_{n-1}. Appendix 2 obtains this expansion at the dominant singularity $z = \frac{1}{4}$,

$$\tilde{f}(z) = \alpha_t + \beta_t \left(-\frac{\sqrt{1 - 4z}}{2} \right) \pm \mathcal{O}(1 - 4z) \tag{32}$$

$$\sim \alpha_t + \beta_t \left(-\frac{\sqrt{1 - 4z}}{2} \right), \tag{33}$$

with

$$\alpha_t = \sum_{j=1}^{J} 2^{b-a_j} q_j \tag{34}$$

$$\beta_t = \sum_{j=1}^{J} 2^{b+1-a_j} (a_j + b) q_j. \tag{35}$$

Note that in Eq. (32), the seed tree affects only the constants α_t and β_t computed in Eqs. (34) and (35) from g, as written in the linear combination in Eq. (23). Excluding the constant α_t that does not influence the asymptotic behavior of the coefficients, the main term of the expansion of $\tilde{f}(z)$ (Eq. (33)) is the product of β_t and the generating function $-\sqrt{1 - 4z}/2$, whose nth coefficient is Catalan number c_{n-1} (Eq. (4)).

Theorem VI.4 of [8] indicates that under conditions satisfied by $\tilde{f}$, the asymptotic coefficients of a generating function as $n \to \infty$ are obtained from the expansion of the function at the dominant singularity; moreover, the error term in the asymptotic coefficients can be computed from the error term in the singular expansion. Applying the theorem to the expansion in Eq. (32), we obtain the asymptotic behavior of the coefficients $[z^n]\tilde{f}(z) = h_{n-1}$.

Proposition 4. *For any seed tree t, when $n \to \infty$, the number h_n of matching coalescent histories for $t^{(n)}$ satisfies*

$$h_{n-1} = [z^n]\tilde{f}(z) \sim \beta_t [z^n] \left(-\frac{\sqrt{1 - 4z}}{2} \right) \pm \mathcal{O}\left(\frac{4^n}{n^2} \right)$$

$$= \beta_t c_{n-1} \pm \mathcal{O}\left(\frac{4^n}{n^2} \right), \tag{36}$$

where β_t is a constant that depends on t. The constant β_t is computed in Eq. (35) once the function g, defined in Eq. (18), is written as the linear combination in Eq. (23).

We immediately obtain the following corollary, corresponding to our initial claim in Eq. (6).

Corollary 1. *For any seed tree t, there exists a constant $\beta_t > 0$ (Eq. (35)) such that when $n \to \infty$,*

$$h_n \sim \beta_t c_n. \tag{37}$$

Proof. The result follows from Proposition 4 by noting that if $\beta_t > 0$, then

$$\lim_{n \to \infty} \frac{h_{n-1}}{\beta_t c_{n-1}} = 1 \pm \lim_{n \to \infty} \frac{\mathcal{O}(4^n/n^2)}{\beta_t c_{n-1}} = 1.$$

Note that we are claiming $\beta_t > 0$. From the definition of β_t in Eq. (35), because the q_j are permitted to be negative, it is not immediately clear that $\beta_t > 0$. Proposition 4 eliminates the possibility that β_t is negative, as h_{n-1} is necessarily positive. To show that $\beta_t \ne 0$, note that by Eq. (36), $\beta_t = 0$ would give

$$h_{n-1} = \mathcal{O}\left(\frac{4^n}{n^2} \right), \tag{38}$$

so that $h_{n-1}/(4^n/n^2)$ would remain bounded by a constant as $n \to \infty$.

We now apply the lower bound $h_n \geq c_{n+1}$ from Eq. (7). By Eq. (7), we have

$$\frac{h_{n-1}}{4^n/n^2} \geq \frac{c_n}{4^n/n^2} = \frac{\sqrt{n}}{\sqrt{\pi}} \frac{c_n}{4^n/(n^{3/2}\sqrt{\pi})}.$$

As $n \to \infty$, $\sqrt{n}/\sqrt{\pi}$ diverges to ∞, while $c_n/[4^n/(n^{3/2}\sqrt{\pi})]$ converges to 1 by Eq. (5). Therefore, the sequence $h_{n-1}/(4^n/n^2)$ must diverge and Eq. (38) cannot hold. Thus, $\beta_t \neq 0$.   □

As an example of Corollary 1, consider $t = ((A, B), (C, D))$. By decomposing the function g of Eq. (24) as in Eq. (23), we have already obtained the parameters b, J, $(a_j)_{1 \leq j \leq J}$, and $(q_j)_{1 \leq j \leq J}$ in Section 3.3.1. Therefore, computing β_t as in Eq. (35), we obtain

$$\beta_t = 2^{1+3-1}(1+3)(4) + 2^{1+3-2}(2+3)(-3)$$
$$+ 2^{1+3-3}(3+3)(1) = 80.$$

Eq. (37) then produces $h_n \sim 80c_n$. Note that the limit $h_n \sim \frac{5}{4}c_{n+3}$ produced for this tree from $h_n = [(5n+8)/(4n+10)]c_{n+3}$ in Section 3.3.3 agrees with the limiting result $h_n \sim 80c_n$. Recalling Eq. (2),

$$\frac{h_n}{c_n} = \frac{5n+8}{4n+10} \frac{c_{n+3}}{c_n} \sim \frac{5}{4} \frac{\binom{2n+6}{n+3}/(n+3)}{\binom{2n}{n}/(n+1)} \sim \frac{5}{4} 4^3 = 80.$$

3.4.2   Determining β_t from the Seed Tree t

We have shown in Corollary 1 that the number of matching coalescent histories h_n for the caterpillar-like family $t^{(n)}$ is, for a constant β_t, asymptotic to $\beta_t c_n$. We can now assemble our results to describe a procedure that given a seed tree t with $|t| \geq 2$ determines both the generating function with coefficients h_n and the constant β_t.

(i)   Determine by Eq. (9) the polynomial $e_{t,k}$ in $k \geq 0$ that counts k-extended histories of t.

(ii)   Compute from Eq. (8) the polynomial in m that counts for $m \geq 1$ the number of m-rooted histories of t.

(iii)   Obtain the generating function $g(y) = \sum_{m=1}^{\infty} h_{0,m} y^m$ with coefficients $h_{0,m}$ by using Proposition 2.

(iv)   Determine the generating function $f(z) = \sum_{n=0}^{\infty} h_n z^n$ with coefficients h_n by applying Proposition 3.

(v)   Write $g(y)$ as a linear combination according to Eq. (23), determining the values of b, J, and the a_j and q_j.

(vi)   Compute the asymptotic constant β_t from Eq. (35).

We have programmed this procedure in Mathematica; starting from a given seed tree t, our program `CatFamily.nb` can automatically compute for the caterpillar-like family $t^{(n)}$ the generating function with coefficients h_n and the asymptotic constant β_t. Using this program, we have evaluated β_t for each seed tree with nine taxa, collecting the results in Table 1.

Recall that Rosenberg [14] reported the asymptotic constant multiples of the Catalan numbers, β_t^*, which represent asymptotic numbers of coalescent histories for seed trees with up to 8 taxa, indexing the results by the number of taxa m rather than by the index n of the caterpillar-like family. Also recall that for seed tree t, tree $t^{(n)}$ has $m = |t| + n$ taxa (Fig. 1). In the notation of [14], writing $A_{t_m,1}$ as the number of matching coalescent histories in the caterpillar-like tree with seed tree t and $m \geq |t|$ taxa, we have $h_n = A_{t_m,1}$.

By Eq. (5), we have the asymptotic equivalence $c_n \sim c_{n+k}/4^k$ for each positive integer k. Therefore,

$$A_{t_m,1} = h_n \sim \beta_t c_n \sim \frac{\beta_t}{4^{|t|-1}} c_{n+|t|-1} = \beta_t^* c_{m-1}, \qquad (39)$$

where the asymptotic constant β_t of Corollary 1 is normalized to obtain

TABLE 1
Asymptotic Constants β_t with $h_n \sim \beta_t c_n$, for Seed Trees t with Nine Taxa

Seed tree t	β_t	β_t^*	Seed tree t	β_t	β_t^*
(tree)	65,536	1	(tree)	128,864	$\frac{4{,}027}{2{,}048}$
(tree)	81,920	$\frac{5}{4}$	(tree)	166,624	$\frac{5{,}207}{2{,}048}$
(tree)	94,208	$\frac{23}{16}$	(tree)	197,296	$\frac{12{,}331}{4{,}096}$
(tree)	104,448	$\frac{51}{32}$	(tree)	224,704	$\frac{3{,}511}{1{,}024}$
(tree)	138,240	$\frac{135}{64}$	(tree)	308,576	$\frac{9{,}643}{2{,}048}$
(tree)	118,784	$\frac{29}{16}$	(tree)	262,000	$\frac{16{,}375}{4{,}096}$
(tree)	113,408	$\frac{443}{256}$	(tree)	250,272	$\frac{7{,}821}{2{,}048}$
(tree)	148,480	$\frac{145}{64}$	(tree)	339,504	$\frac{21{,}219}{4{,}096}$
(tree)	177,664	$\frac{347}{128}$	(tree)	417,632	$\frac{13{,}051}{2{,}048}$
(tree)	141,312	$\frac{69}{32}$	(tree)	326,240	$\frac{10{,}195}{2{,}048}$
(tree)	193,536	$\frac{189}{64}$	(tree)	464,128	$\frac{1{,}813}{256}$
(tree)	121,472	$\frac{949}{512}$	(tree)	182,912	$\frac{1{,}429}{512}$
(tree)	157,888	$\frac{2{,}467}{1{,}024}$	(tree)	243,904	$\frac{3{,}811}{1{,}024}$
(tree)	187,776	$\frac{1{,}467}{512}$	(tree)	296,064	$\frac{2{,}313}{512}$
(tree)	214,720	$\frac{3{,}355}{1{,}024}$	(tree)	344,512	$\frac{5{,}383}{1{,}024}$
(tree)	296,192	$\frac{1{,}157}{256}$	(tree)	487,808	$\frac{3{,}811}{512}$
(tree)	251,136	$\frac{981}{256}$	(tree)	410,112	$\frac{801}{128}$
(tree)	162,560	$\frac{635}{256}$	(tree)	214,016	$\frac{209}{64}$
(tree)	219,136	$\frac{107}{32}$	(tree)	306,112	$\frac{4{,}783}{1{,}024}$
(tree)	268,288	$\frac{131}{32}$	(tree)	294,784	$\frac{2{,}303}{512}$
(tree)	177,664	$\frac{347}{128}$	(tree)	425,216	$\frac{1{,}661}{256}$
(tree)	249,344	$\frac{487}{128}$	(tree)	366,720	$\frac{2{,}865}{512}$
(tree)	353,536	$\frac{1{,}381}{256}$	(tree)	532,224	$\frac{2{,}079}{256}$

Values of β_t appear for each of the 46 unlabeled species trees with nine taxa. For each species tree t, we also provide the constant $\beta_t^* = \beta_t/4^8$ (Eq. (40)). Trees are listed in increasing order by rank as defined in Section 2 of [14]. In the left column, each seed tree t belongs to a caterpillar-like family $(t^{(n)})_n$, with $|t| < 9$. In these cases, we recover the values of β_t^* as determined in Table 3 of [14].

$$\beta_t^* = \frac{\beta_t}{4^{|t|-1}}. \qquad (40)$$

This computation converts the asymptotic constant multiple β_t of c_n into a corresponding multiple β_t^* of c_{m-1}, as reported in [14] for small trees. Comparing Table 1 with Table 3 of [14], we see that for the cases examined by [14], the values of β_t^* we compute from the associated β_t agree with the values that were previously reported. This agreement is unsurprising; our method for calculating the constants β_t and β_t^* is simply a computational implementation based on our theorems, and the agreement confirms the validity of the implementation. Although [14] considered only $|t| \leq 8$, our method applies for arbitrary $|t|$.

Evaluation of β_t proceeds quadratically in $|t|$. The recursive step (i) requires at most $|t| - 1$ recursive calls, one for each internal node of t. Step (ii) is a polynomial subtraction at most linear in $|t|$, producing the polynomial $h_{0,m}$ with order at most equal to the order of $e_{t,m}$ minus 1—that is, at most $|t| - 2$. Step (iii) determines the generating function $g(y)$ (Eq. (18)) from $h_{0,m}$ and the generating functions $g_{m^k}(y)$ (Eq. (20)). For each k with $0 \leq k \leq |t| - 2$, $g_{m^k}(y)$ is computed in k recursive calls of Eq. (22). As the order of $h_{0,m}$ is at most $|t| - 2$, the total cost for calculating $g(y)$ is thus quadratic in $|t|$. Steps (iv), (v), and (vi) do not involve recursion and are at most linear in $|t|$. Thus, because step (iii) is the most expensive step, we see that the cost of the procedure that determines the asymptotic constant β_t increases as $\mathcal{O}(|t|^2)$.

4 CONCLUSIONS

In this paper, we have solved a problem left open by [14] on determining the number of coalescent histories for gene trees and species trees that have a matching labeled topology and that belong to a generic caterpillar-like family. We have proven that for any seed tree t, the integer sequence $(h_n)_{n \geq 0}$, whose nth element represents the number of matching coalescent histories of the caterpillar-like tree $t^{(n)}$, grows asymptotically as a constant multiple of the Catalan numbers, that is, $h_n \sim \beta_t c_n$, where the constant $\beta_t > 0$ depends on the shape of the seed tree t. Rosenberg [14] had previously obtained this result for seed trees with at most eight taxa; here, by using a succession rule for recursive enumeration and then applying techniques of analytic combinatorics, we have not only proven the existence of the constant β_t for seed trees of any size, we have also produced a procedure that computes β_t as well as the expression for the generating function of the integers $(h_n)_{n \geq 0}$.

The numerical results on the constants β_t extend the empirical observation of [14] that the caterpillar-like families that produce the largest numbers of matching coalescent histories are those whose seed tree has a high level of balance. By extending from seed trees with $|t| \leq 8$ taxa to those with $|t| = 9$, we observe that the constants β_t for the caterpillar-like families with the largest and smallest numbers of matching coalescent histories become further separated, so that for n large, many more coalescent histories exist by which a gene tree can match the species tree for some species trees than for others. For the 9-taxon seed tree

with the largest β_t^*, $\beta_t^* \approx 8.12$ compared to $\beta_t^* = 1$ for the seed tree with the smallest β_t^*. Our procedure for evaluating β_t and β_t^* as a function of the seed tree can now enable further systematic analyses of the correlates of the constants β_t and β_t^*, to facilitate additional explorations of determinants of the numbers of matching coalescent histories.

Nevertheless, although the constants β_t and β_t^* do depend on the seed tree, we have shown that otherwise, all caterpillar-like families are asymptotically equivalent in their numbers of matching coalescent histories. Computation time is often a challenge in phylogenetic problems, as the discrete structures of phylogenetics can grow rapidly in number with the number of taxa. Our results contribute to the study of computational complexity in phylogenetics, as the complexity of the evaluation of probabilities important in characterizing gene tree distributions [5] is proportional to the number of coalescent histories. That all caterpillar-like families have the same growth pattern up to a constant suggests that as the number of taxa increases, such evaluations will be comparably complex for all caterpillar-like trees. In large trees, the caterpillar branches contribute to the asymptotic growth of the number of matching coalescent histories—which follows a multiple of the Catalan numbers—and the seed tree only to the constant by which the Catalan numbers are multiplied.

The extent to which other tree families follow the Catalan sequence in their numbers of matching coalescent histories remains unknown, though we have recently found a family, the *lodgepole* family—defined iteratively by setting λ_0 to a tree with one taxon and sequentially forming λ_{n+1} by appending λ_n and a cherry to a shared root—for which the number of matching coalescent histories grows faster than with a constant multiple of the Catalan numbers [6]. Further analysis of this heterogeneous behavior of the increase in the number of coalescent histories will be useful in performing comparisons of coalescent history algorithms with algorithms that obtain similar phylogenetic probabilities but that do not rely on coalescent histories [20]. The use of our substantially different approach employing analytic combinatorics opens new methods for theoretical analysis of coalescent histories and can potentially assist in understanding when Catalan-like growth, the rapid growth of the lodgepole family, and intermediate or perhaps still faster growth patterns will apply.

We note, however, that our strategy for evaluating the asymptotic properties of the number of coalescent histories in caterpillar-like families has, like the work of [14], relied on the fact that the difficulty of the general problem of enumerating coalescent histories is partly evaded by restricting attention to caterpillar-like trees. In the recursion for the number of coalescent histories given a matching gene tree and species tree [13, Eq. (1)], a term arising from the subtree with fewer branches collapses to 1 for the caterpillar case, greatly simplifying the recursion. This reduction enabled the work of [14] for caterpillar-like families, and it also enables our approach of iteratively adding single-taxon branches to define the operator Ω and the generating function for $h_{n,m}$. Thus, in enumerating coalescent histories for matching lodgepole gene trees and species trees, we proceeded by a different method, establishing a bijection between coalescent

histories and established combinatorial structures [6]. We do expect, however, that a generating function approach will be fruitful in other scenarios, perhaps including cases with gene trees and species trees that are caterpillar-like, but non-matching.

APPENDIX 1
THE EQUATION FOR $F(y, z)$

In this appendix, we complete the derivation of Eq. (27) satisfied by $F(y, z)$. In the generating function $F(y, z)$ (Eq. (25)), each monomial $z^n y^m$ corresponds to a label $(n, m) \in L_n$ that in turn represents an m-rooted history of $t^{(n)}$. Recall that the multisets of labels $L_0, L_1, L_2, \dots$ (Eq. (14)) can be iteratively generated according to Eq. (17) through the operator Ω defined in Eq. (13), starting from the multiset L_0. Also recall that by considering the multiset of labels $L = \cup_{n=0}^{\infty} L_n$, we can write $F(y, z) = \sum_{(n,m) \in L} z^n y^m$. We use the iterative generation of the family of multisets $(L_n)_{n \geq 0}$ to obtain an equation for F.

By Eq. (13), for $n \geq 0$ and $m \geq 2$, for each occurrence in L_n of a label (n, m), a copy of each label in set

$$\Omega\big((n, m)\big) = \{(n + 1, m + j) : j \geq -1\}$$

belongs to the multiset L_{n+1}. Thus, in algebraic terms, each time that an expression $z^n y^m$ with $n \geq 0$ and $m \geq 2$ is counted in the generating function F—written $z^n y^m \in F$ in what follows—the terms $z^{n+1} \sum_{j=m-1}^{\infty} y^j$ appear in F as well. Summing over all $z^n y^m \in F$ with $n \geq 0$ and $m \geq 2$, we obtain

$$\sum_{z^n y^m \in F : n \geq 0, m \geq 2} \left(z^{n+1} \sum_{j=m-1}^{\infty} y^j \right) = \frac{z}{y} \sum_{z^n y^m \in F : n \geq 0, m \geq 2} \left(z^n y^m \sum_{j=0}^{\infty} y^j \right). \tag{41}$$

Similarly, for $n \geq 0$ and $m = 1$, for each occurrence in L_n of a label $(n, 1)$, a copy of each label in set $\Omega\big((n, 1)\big) = \{(n + 1, j) : j \geq 1\}$ appears in multiset L_{n+1}. Thus, for each term $z^n y \in F$, with $n \geq 0$, the terms $z^{n+1} \sum_{j=1}^{\infty} y^j$ are counted in F as well. Summing these terms for all $z^n y \in F$ with $n \geq 0$,

$$\sum_{z^n y \in F : n \geq 0} \left(z^{n+1} \sum_{j=1}^{\infty} y^j \right) = zy \sum_{z^n y \in F : n \geq 0} \left(z^n \sum_{j=0}^{\infty} y^j \right). \tag{42}$$

Notice that the sum of the expressions in Eqs. (41) and (42) is the algebraic representation of the multiset of labels $L \setminus L_0$. More precisely, each term $z^n y^m \in F$ associated with a label $(n, m) \in L_n$, with $n \geq 1$, is counted—and counted exactly once—in the sum of Eqs. (41) and (42). Therefore, to complete the description of F, we require only those terms $z^0 y^m$ associated with labels $(0, m) \in L_0$. These terms are represented

$$\sum_{(0,m) \in L_0} z^0 y^m = \sum_{m=1}^{\infty} h_{0,m} y^m = g(y), \tag{43}$$

considering that $h_{0,m} = |\{\ell \in L_0 : \ell = (0, m)\}|$ (Eq. (15)) and that by definition, $g(y) = \sum_{m=1}^{\infty} h_{0,m} y^m$ (Eq. (18)).

We can now equate the full generating function $F(y, z)$ to the sum of Eqs. (43), (41), and (42), obtaining

$$F(y, z) = g(y) + \frac{z}{y} \sum_{z^n y^m \in F : n \geq 0, m \geq 2} \left(z^n y^m \sum_{j=0}^{\infty} y^j \right)$$

$$+ zy \sum_{z^n y \in F : n \geq 0} \left(z^n \sum_{j=0}^{\infty} y^j \right).$$

Applying the fact that $\sum_{j=0}^{\infty} y^j = 1/(1 - y)$ for y near 0 in the complex plane, we then have

$$F(y, z) = g(y) + \frac{z}{y(1-y)} \left(\sum_{z^n y^m \in F : n \geq 0, m \geq 2} z^n y^m \right)$$

$$+ \frac{zy}{1-y} \left(\sum_{z^n y \in F : n \geq 0} z^n \right). \tag{44}$$

By Eq. (25) and the fact that the multisets L_n of labels (n, m) for m-rooted histories of $t^{(n)}$ have $h_{n,m}$ elements,

$$\sum_{z^n y \in F : n \geq 0} z^n = \frac{\partial F}{\partial y}(0, z)$$

$$\sum_{z^n y^m \in F : n \geq 0, m \geq 2} z^n y^m = \left(\sum_{z^n y^m \in F : n \geq 0, m \geq 1} z^n y^m \right) - \left(\sum_{z^n y \in F : n \geq 0} z^n y \right)$$

$$= F(y, z) - y \frac{\partial F}{\partial y}(0, z).$$

Substituting in Eq. (44), the last two expressions yield

$$F(y, z) = g(y) + \frac{z}{y(1-y)} \left(F(y, z) - y \frac{\partial F}{\partial y}(0, z) \right)$$

$$+ \frac{zy}{1-y} \frac{\partial F}{\partial y}(0, z), \tag{45}$$

which can be rewritten as in Eq. (27).

APPENDIX 2
THE DOMINANT SINGULARITY AND SINGULAR EXPANSION OF $\tilde{f}(z)$

This appendix obtains the singular expansion of $\tilde{f}(z)$ described in Eq. (32). In Eq. (31), we have defined $\tilde{f}(z)$ as a composition $\tilde{f}(z) = g(Y(z))$, with the internal function $Y(z)$ as in Eq. (28) and the external function $g(y)$ as in Eq. (23). Owing to the presence of the square root in the expression for $Y(z)$, the dominant singularity of the internal function $Y(z)$—the singularity nearest the origin of the complex plane—is at $z = \frac{1}{4}$. Computing the value of $Y(z)$ at its dominant singularity, we obtain $Y(\frac{1}{4}) = \frac{1}{2}$. In particular, we have $Y(\frac{1}{4}) < 1$, where 1 is the radius of convergence of the finite series corresponding to the external function g in $\tilde{f}$. Indeed, it immediately follows from Proposition 2 that $y = 1$ is the dominant singularity of $g(y)$.

As detailed in Section VI.9 of [8], on dominant singularities of compositions, we are in the setting of the subcritical case, in which the inequality $Y(\frac{1}{4}) < 1$ implies that the dominant singularity of $g(Y(z))$ coincides with the dominant singularity $z = \frac{1}{4}$ of the internal function $Y(z)$ rather than the

dominant singularity $y = 1$ of the external function $g(y)$. The desired singular expansion of $\tilde{f}(z) = g(Y(z))$ at the dominant singularity $z = \frac{1}{4}$ can be obtained by inserting $y = Y(z)$ in the regular (non-singular) expansion of $g(y)$ at $y = Y(\frac{1}{4}) = \frac{1}{2}$.

To recover the expansion of $g(y)$ at $y = \frac{1}{2}$, we expand and then sum each term $q_j[y^{a_j}/(1-y)^b]$ of the finite linear combination in Eq. (23). At $y = \frac{1}{2}$, each of these terms is an analytic function, and we can thus use Taylor's formula to produce the desired expansion. We obtain at $y = \frac{1}{2}$

$$q_j \frac{y^{a_j}}{(1-y)^b} = 2^{b-a_j} q_j + 2^{b+1-a_j}(a_j + b) q_j \left(y - \frac{1}{2} \right)$$
$$\pm \mathcal{O}\left(\left(y - \frac{1}{2} \right)^2 \right).$$

By summing over the indices $1 \le j \le J$ of Eq. (23), the expansion of $g(y)$ at $y = \frac{1}{2}$ is

$$g(y) = \alpha_t + \beta_t \left(y - \frac{1}{2} \right) \pm \mathcal{O}\left(\left(y - \frac{1}{2} \right)^2 \right), \qquad (46)$$

with the constants α_t and β_t defined as in Eqs. (34) and (35). Plugging $y = Y(z)$ from Eq. (28) into Eq. (46), we finally obtain the singular expansion of $\tilde{f}(z)$ at $z = \frac{1}{4}$ as in Eq. (32).

ACKNOWLEDGMENTS

REFERENCES

[1] E. S. Allman, J. H. Degnan, and J. A. Rhodes, "Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent," *J. Math. Biol.*, vol. 62, pp. 833–862, 2011.

[2] C. Banderier, M. Bousquet-Mélou, A. Denise, P. Flajolet, D. Gardy, and D. Gouyou-Beauchamps, "Generating functions for generating trees," *Discr. Math.*, vol. 246, pp. 29–55, 2002.

[3] E. Barcucci, A. Del Lungo, E. Pergola, and R. Pinzani, "ECO: A methodology for the enumeration of combinatorial objects," *J. Differ. Equ. Appl.*, vol. 5, pp. 435–490, 1999.

[4] J. H. Degnan, "Gene tree distributions under the coalescent process," Ph.D. dissertation, Univ. New Mexico, Albuquerque, NM, USA: 2005.

[5] J. H. Degnan and L. A. Salter, "Gene tree distributions under the coalescent process," *Evolution*, vol. 59, pp. 24–37, 2005.

[6] F. Disanto and N. A. Rosenberg, "Coalescent histories for lodgepole species trees," *J. Comput. Biol.*, vol. 22, pp. 918–929, 2015.

[7] J. Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup, "Ancestral population genomics: The coalescent hidden Markov model approach," *Genetics*, vol. 183, pp. 259–274, 2009.

[8] P. Flajolet and R. Sedgewick, "*Analytic Combinatorics*. Cambridge, U.K.: Cambridge Univ. 2009.

[9] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics*. 2nd ed. Boston, MA, USA: Addison-Wesley, 2008.

[10] A. Hobolth, O. F. Christensen, T. Mailund, and M. H. Schierup, "Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model," *PLoS Genet.*, vol. 3, pp. 294–304, 2007.

[11] A. Hobolth, J. Y. Dutheil, J. Hawks, M. H. Schierup, and T. Mailund, "Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widepsread selection," *Genome Res.*, vol. 21, pp. 349–356, 2011.

[12] H. Prodinger, "The kernel method: A collection of examples," *Sém. Lothar. Combin.*, vol. 50, article B50f, 2004.

[13] N. A. Rosenberg, "Counting coalescent histories," *J. Comput. Biol.*, vol. 14, pp. 360–377, 2007.

[14] N. A. Rosenberg, "Coalescent histories for caterpillar-like families," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, pp. 1253–1262, Sept./Oct. 2013.

[15] N. A. Rosenberg and J. H. Degnan, "Coalescent histories for discordant gene trees and species trees," *Theor. Pop. Biol.*, vol. 77, pp. 145–151, 2010.

[16] N. A. Rosenberg and R. Tao, "Discordance of species trees with their most likely gene trees: The case of five taxa," *Syst. Biol.*, vol. 57, pp. 131–140, 2008.

[17] R. P. Stanley, *Enumerative Combinatorics Volume 2*. New York, NY, USA: Cambridge Univ. Press, 1999.

[18] C. V. Than and N. A. Rosenberg, "Consistency properties of species tree inference by minimizing deep coalescences," *J. Comput. Biol.*, vol. 18, pp. 1–15, 2011.

[19] C. Than, D. Ruths, H. Innan, and L. Nakhleh, "Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions," *J. Comput. Biol.*, vol. 14, pp. 517–535, 2007.

[20] Y. Wu, "Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood," *Evolution*, vol. 66, pp. 763–775, 2012.

[21] Y. Yu, C. Than, J. H. Degnan, and L. Nakhleh, "Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting," *Syst. Biol.*, vol. 60, pp. 138–149, 2011.

Filippo Disanto received the PhD degree in theoretical computer science from both the University of Siena and the University of Paris VII in 2010. After receiving the PhD degree, he was a postdoc at CNRS in Montpellier and at the Institut für Genetik, University of Cologne. Since November 2013, he has been a postdoc in the Rosenberg Laboratory, Stanford University, CA. His main research interests include combinatorics and its applications.

Noah A. Rosenberg received the PhD degree in biological sciences from Stanford University in 2001 and completed postdoctoral training at the University of Southern California. He was on the faculty of the University of Michigan from 2005 to 2011, and he is currently a professor in the Department of Biology, Stanford University, CA. His research interests include human evolutionary genetics, population-genetic theory, and mathematical phylogenetics.