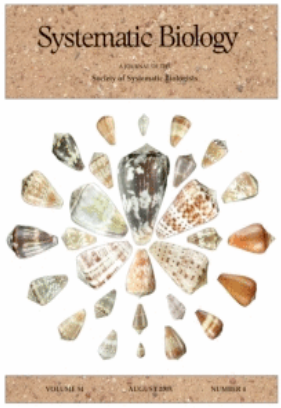

Systematic Biology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713658732

Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa

Noah A. Rosenberg [abc]; Randa Tao [b]
[a] Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA
[b] Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, Michigan, USA
[c] The Life Sciences Institute, University of Michigan, Ann Arbor, Michigan, USA

First Published on: 01 February 2008
To cite this Article: Rosenberg, Noah A. and Tao, Randa (2008) 'Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa', Systematic Biology, 57:1, 131 - 140
To link to this article: DOI: 10.1080/10635150801905535
URL: http://dx.doi.org/10.1080/10635150801905535

Syst. Biol. 57(1):131–140, 2008

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150801905535

Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa

NOAH A. ROSENBERG[1,2,3] AND RANDA TAO[2]

[1]*Department of Human Genetics,* [2]*Center for Computational Medicine and Biology, and* [3]*The Life Sciences Institute, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan 48109-2218, USA; E-mail: rnoah@umich.edu (N.A.R.)*

Abstract.— Under a coalescent model for within-species evolution, gene trees may differ from species trees to such an extent that the gene tree topology most likely to evolve along the branches of a species tree can disagree with the species tree topology. Gene tree topologies that are more likely to be produced than the topology that matches that of the species tree are termed *anomalous*, and the region of branch-length space that gives rise to anomalous gene trees (AGTs) is the *anomaly zone*. We examine the occurrence of anomalous gene trees for the case of five taxa, the smallest number of taxa for which every species tree topology has a nonempty anomaly zone. Considering all sets of branch lengths that give rise to anomalous gene trees, the largest value possible for the smallest branch length in the species tree is greater in the five-taxon case (0.1934 coalescent time units) than in the previously studied case of four taxa (0.1568). The five-taxon case demonstrates the existence of three phenomena that do not occur in the four-taxon case. First, anomalous gene trees can have the same unlabeled topology as the species tree. Second, the anomaly zone does not necessarily enclose a ball centered at the origin in branch-length space, in which all branches are short. Third, as a branch length increases, it is possible for the number of AGTs to *increase* rather than decrease or remain constant. These results, which help to describe how the properties of anomalous gene trees increase in complexity as the number of taxa increases, will be useful in formulating strategies for evading the problem of anomalous gene trees during species tree inference from multilocus data. [Coalescence; lineage sorting; probability.]

It is well known that the sorting process of genealogical lineages during speciation can cause gene tree topologies to differ from each other and from the topology of the species tree on which they have evolved (Pamilo and Nei, 1988; Takahata, 1989; Maddison, 1997; Nichols, 2001; Rosenberg, 2002; Degnan and Salter, 2005). Looking backwards in time, this discordance is produced largely by short internal branch lengths of the species tree, which increase the probability that genetic lineages persist far enough into the past that they have the opportunity to coalesce with lineages from distant species.

Employing a commonly used model for the stochastic evolution of gene trees along the branches of fixed species trees, Degnan and Rosenberg (2006) showed that discordance of gene trees and species trees can be sufficiently probable that the gene tree topology most likely to evolve along the branches of a species tree might differ from the species tree topology. Gene tree topologies with probability greater than that of the matching topology were termed *anomalous gene trees* (AGTs), and for a given species tree topology, the set of possible branch lengths giving rise to AGTs was termed the *anomaly zone*.

Degnan and Rosenberg (2006) proved that the asymmetric species tree topology with four taxa—and any species tree topology with $n \geq 5$ taxa—has a nonempty anomaly zone. However, the proof for $n \geq 5$ did not characterize the size and boundaries of the anomaly zone, and this explicit characterization was performed only for the four-taxon asymmetric species tree topology.

To gain further insight into the properties of anomalous gene trees, we investigate the anomaly zones for the three unlabeled species tree topologies with five taxa. The five-taxon case is the smallest instance of the general existence result for anomalous gene trees, in that it is the smallest number of taxa for which *each* species tree topology has AGTs. Additionally, because five-taxon trees have only three internal branches, the number of model parameters in the five-taxon case is small enough that the anomaly zones can be visualized relatively easily. Thus, we count the number of AGTs for all three unlabeled five-taxon species tree topologies, and we characterize the three anomaly zones. We examine how the anomaly zones depend on the three internal branch lengths, as well as how they differ across the three species tree topologies. Interestingly, the five-taxon case reveals several phenomena that are not observed with four taxa: AGTs can have the same unlabeled topology as the species tree, the anomaly zone need not enclose a ball in branch-length space centered at the origin (in which all species tree branches are short), and the number of AGTs can increase with increasing branch lengths rather than decreasing or remaining constant.

GENE TREE PROBABILITIES

General Model

To compute the probabilities of gene trees conditional on species trees, we use the same coalescent-based model for the evolution of gene trees on species trees that has been used in previous investigations (Pamilo and Nei, 1988; Degnan and Salter, 2005; Degnan and Rosenberg, 2006). In this model, the species tree is treated as fixed. Gene lineages travel backwards in time, eventually coalescing to a single lineage. Along each branch of the species tree, gene lineages entering the branch from a more recent time period may coalesce, with coalescence equiprobable for each pair of lineages, as specified by the Yule model (Harding, 1971; Brown, 1994; Aldous, 2001; Steel and McKenzie, 2001; Rosenberg, 2006), and with the coalescence rate following the coalescent process (Nordborg, 2003; Hein et al., 2005). We consider gene trees that are known exactly, so that mutations do not obscure the underlying relationships among genealogical lineages.

Consider a rooted binary species tree σ with topology ψ and with a vector of nonnegative branch lengths $\boldsymbol{T}$, where T_i denotes the length of branch i (technically if

TABLE 1. The probability $g_{ij}(T)$ that the number of gene lineages ancestral at time T in the past to i lineages in the present is j, for $i < 5$.

i	j	$g_{ij}(T)$
2	1	$1 - e^{-T}$
2	2	e^{-T}
3	1	$1 - \frac{3}{2}e^{-T} + \frac{1}{2}e^{-3T}$
3	2	$\frac{3}{2}e^{-T} - \frac{3}{2}e^{-3T}$
3	3	e^{-3T}
4	1	$1 - \frac{9}{5}e^{-T} + e^{-3T} - \frac{1}{5}e^{-6T}$
4	2	$\frac{9}{5}e^{-T} - 3e^{-3T} + \frac{6}{5}e^{-6T}$
4	3	$2e^{-3T} - 2e^{-6T}$
4	4	e^{-6T}

a branch length is zero, then the tree is not binary, but we will include this possibility). Branch lengths are measured in coalescent time units, which can be converted to units of generations under any of several choices for models of evolution within species (Nordborg, 2003; Hein et al., 2005; Sjödin et al., 2005). In the simplest model for diploids, each species has constant population size $N/2$ individuals, and λ_i coalescent units equals $\lambda_i N$ generations.

For the fixed species tree $\sigma = (\psi, \boldsymbol{T})$, the gene tree topology H (for a single gene lineage per species) is viewed as a random variable with distribution depending on σ. Under the model, this distribution is known for arbitrary rooted binary species trees (Degnan and Salter, 2005). It can be viewed as a sum over coalescent histories compatible with the topologies H and ψ of the probabilities of these histories, where each coalescent history refers to a list of branches of the species tree on which the coalescences in the gene tree can take place (Degnan and Salter, 2005; Rosenberg, 2007). The probabilities of individual coalescent histories are obtained using $g_{ij}(T)$, the known probability distribution under the coalescent model for the number of gene lineages j ancestral at time T in the past to a sample of i lineages in the present (Tavaré, 1984; Takahata and Nei, 1985). For fixed i and j, the function $g_{ij}(T)$ is a polynomial in e^{-T}. As a result, for n-taxon trees, the overall probability of a particular gene tree topology given a species tree topology together with branch lengths is a polynomial in $(e^{-T_2}, e^{-T_3}, ..., e^{-T_{n-1}})$, where $T_2, \ldots, T_{n-1}$ denote the lengths of the internal branches of the species tree (excluding the infinitely long branch above the root). For our five-taxon analysis, the $g_{ij}(T)$ polynomials with $i < 5$ are needed, and they are shown in Table 1.

Using $P_\sigma(H = h)$ to denote the probability under the model that a random gene tree has topology h when the species tree is $\sigma = (\psi, \boldsymbol{T})$, anomalous gene trees are defined as follows (Degnan and Rosenberg, 2006):

Definition 1. *(i) A gene tree topology h is* anomalous *for a species tree $\sigma = (\psi, \boldsymbol{T})$ if $\mathbb{P}_\sigma(H = h) > \mathbb{P}_\sigma(H = \psi)$.*
(ii) A topology ψ produces anomalies *if there exists a vector of branch lengths $\boldsymbol{T}$ such that the species tree $\sigma = (\psi, \boldsymbol{T})$ has at least one anomalous gene tree.*

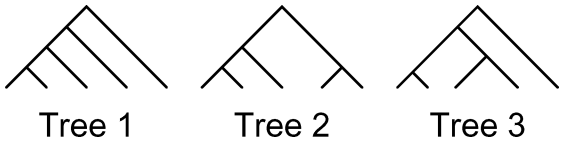

FIGURE 1. The three unlabeled topologies possible for five taxa. For a given set of five labels, trees 1, 2, and 3 have 60, 30, and 15 distinct labelings, respectively. Trees 1, 2, and 3 are also termed topologies 1, 2, and 3, respectively.

(iii) The anomaly zone *for a topology ψ is the set of vectors of branch lengths $\boldsymbol{T}$ for which $\sigma = (\psi, \boldsymbol{T})$ has at least one anomalous gene tree.*

In other words, a gene tree topology h is anomalous for a species tree σ if a gene tree evolving along the branches of σ is more likely to have the topology h than it is to have the same topology as the species tree.

Five-Taxon Case

Our aim is to describe the anomaly zones for all possible species tree topologies with five taxa. To perform this characterization, we compute the probabilities of all possible gene tree topologies for each possible species tree topology. For each species-tree topology, we then determine the region of branch-length space in which at least one nonmatching gene tree topology has a higher probability than the gene tree topology that matches the topology of the species tree.

Figure 1 depicts the three unlabeled topologies possible for five taxa. We refer to these topologies as trees 1, 2, and 3. In order to use these topologies both in gene tree and in species tree contexts, we sometimes refer to topologies ψ_1, ψ_2, and ψ_3 when considering species trees and to topologies γ_1, γ_2, and γ_3 when considering gene trees.

Without loss of generality, for the topology of the species tree, it suffices to consider one labeling of each of the three distinct unlabeled topologies (Fig. 2). Although there are 105 distinct labeled topologies for five taxa, symmetry considerations demonstrate that for each of the three species tree topologies, many subsets of the 105 gene tree topologies exist for which all topologies in a given subset have equal probability conditional on the species tree (Table 2). For example, for species tree ψ_1, the 105 gene tree topologies can be collapsed into 31 "history classes," each of which is distinguished by having

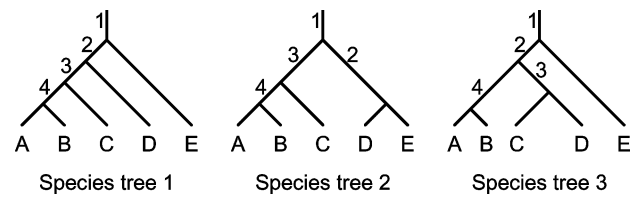

FIGURE 2. Labeled species tree topologies for five taxa. Letters denote species and numbers denote branches of species trees on which coalescences of gene lineages may occur. The length of branch i in coalescent time units ($i = 1, 2, 3, 4$) is denoted T_i, with $T_1 = \infty$.

TABLE 2. The number of distinct lists of coalescent histories (that is, the number of history classes) for labeled gene tree topologies with a given unlabeled topology, for each species tree labeled topology (see Fig. 2).

	Species tree 1	Species tree 2	Species tree 3	Total
Gene tree 1	14	6	10	30
Gene tree 2	12	9	9	30
Gene tree 3	5	5	5	15
Total	31	20	24	75

its own unique list of possible coalescent histories. For species tree ψ_2, the number of history classes is 20, and it is 24 for species tree ψ_3. Given the species tree topology, all gene tree topologies in the same history class have the same probability, for any branch lengths of the species tree. Tables 3–5 enumerate the history classes for the five-taxon species tree topologies.

Using the method of Degnan and Salter (2005), for each species tree topology, we can compute the probability of each gene tree topology as a function of the species tree branch lengths. This exhaustive computation proceeds by listing all history classes for the species tree topology, and for each history class, enumerating the coalescent histories associated with the history class, computing the probability of each coalescent history, and summing over coalescent histories. The full enumeration of gene tree probabilities for all history classes (and hence, for all gene tree topologies) is given in Supplementary Tables 1 to 15 (available online at http://www.systematicbiology.org). We have checked the accuracy of the gene tree probability formulas by (1) confirming that for each species tree topology, the sum of all gene tree probabilities is 1; (2) verifying that all formulas—which were derived using the reasoning of Degnan and Salter (2005) but without using their software for calculating gene tree probabilities (COAL)—produced identical values to those obtained by COAL for specific choices of branch lengths; (3) verifying that the five-taxon anomaly zone approached the four-taxon anomaly zone as appropriate branch lengths were sent to ∞. This last computation is described under Collapse of the Five-Taxon Anomaly Zone to the Four-Taxon Anomaly Zone.

TABLE 3. History classes for species tree topology ψ_1.

Gene tree topology	Number for history class	Description of class of labeled topologies	Definitions of V, W, X, Y, Z	Number of labeled topologies in class
γ_1	1	((((WE)X)Y)Z)	{W,X,Y,Z}={A,B,C,D}	24
	2	((((XD)E)Y)Z)	{X,Y,Z}={A,B,C}	6
	3	((((XD)Y)E)Z)	{X,Y,Z}={A,B,C}	6
	4	((((XD)Y)Z)E)	{X,Y,Z}={A,B,C}	6
	5	((((XC)E)Y)Z)	{X,Y,Z}={A,B,D}, X≠D	4
	6	((((XC)D)E)Y)	{X,Y}={A,B}	2
	7	((((XC)Y)E)D)	{X,Y}={A,B}	2
	8	((((XC)D)Y)E)	{X,Y}={A,B}	2
	9	((((XC)Y)D)E)	{X,Y}={A,B}	2
	10	((((AB)E)X)Y)	{X,Y}={C,D}	2
	11	((((AB)D)E)C)		1
	12	((((AB)C)E)D)		1
	13	((((AB)D)C)E)		1
	14	((((AB)C)D)E)		1
γ_2	1	(((XE)Y)(ZD))	{X,Y,Z}={A,B,C}	6
	2	(((XE)Y)(ZC))	{X,Y,Z}={A,B,D}, Z≠D	4
	3	(((XE)Y)(AB))	{X,Y}={C,D}	2
	4	(((XD)E)(YC))	{X,Y}={A,B}	2
	5	(((CD)E)(AB))		1
	6	(((XC)E)(YD))	{X,Y}={A,B}	2
	7	(((AB)E)(CD))		1
	8	(((XD)Y)(ZE))	{X,Y,Z}={A,B,C}	6
	9	(((XC)D)(YE))	{X,Y}={A,B}	2
	10	(((AB)D)(CE))		1
	11	(((XC)Y)(DE))	{X,Y}={A,B}	2
	12	(((AB)C)(DE))		1
γ_3	1	(((XD)(YE))Z)	{X,Y,Z}={A,B,C}	6
	2	(((XC)(YE))Z)	{X,Y,Z}={A,B,C}, X≠D	4
	3	(((AB)(XE))Y)	{X,Y}={C,D}	2
	4	(((XC)(YD))E)	{X,Y}={A,B}	2
	5	(((AB)(CD))E)		1

RESULTS

History Classes That Produce AGTs

For a given species tree labeled topology ψ and gene tree topology $h \neq \psi$, we can use the Degnan-Salter formula to determine if h can be anomalous for ψ. Letting $\sigma = (\psi, \boldsymbol{T})$ denote the species tree labeled topology together with branch lengths, by definition of anomalous gene trees, h is an AGT for ψ if and only if the set for which

$$\mathbb{P}_\sigma(H = h) > \mathbb{P}_\sigma(H = \psi) \tag{1}$$

contains at least one vector $\boldsymbol{T}$.

For each of the three species tree topologies, the polynomials (in e^{-T_2}, e^{-T_3}, and e^{-T_4}) on both sides of Equation 1 are linear in e^{-T_4}. This is a consequence of the fact that on branch 4 of the species tree, regardless of the eventual gene tree topology that is produced by the sequence of coalescences, two gene lineages "enter" this branch, and either they coalesce to one lineage—an event that has probability $g_{21}(T_4) = 1 - e^{-T_4}$—or they do not coalesce, an event with probability $g_{22}(T_4) = e^{-T_4}$. Each of the coalescent histories in the sum required for computing the two sides of Equation 1 therefore contains a linear term in e^{-T_4}—and no terms with higher powers of e^{-T_4}. Because of this linearity, for any given gene tree topology it is not hard to solve the inequality for T_4 in terms of T_2 and T_3 and to determine if some nonnegative combination of T_2, T_3, and T_4 satisfies the inequality. For each species tree topology and each gene tree history class, Supplementary Tables 16 to 24 (http://www.systematicbiology.org) give the polynomial inequality in e^{-T_2}, e^{-T_3}, and e^{-T_4} whose solution space is the region where AGTs in the history class occur.

Solving Equation 1 for each combination of a gene tree topology and species tree topology, we can count the number of history classes that contain AGTs (Table 6). For each of the three species tree topologies, we can then sum the numbers of gene tree topologies across history classes to obtain the total number of possible AGTs (Table 7).

TABLE 4. History classes for species tree topology ψ_2. Some elements of {V,W,X,Y,Z} appear under "Description of class of labeled topologies" but not under "Definitions of V, W, X, Y, Z." After the assignment to specific taxa of those elements of {V,W,X,Y,Z} that do appear under "Definitions of V, W, X, Y, Z," the remaining elements of {V,W,X,Y,Z} are chosen from the remaining taxa in {A,B,C,D,E}.

Gene tree topology	Number for history class	Description of class of labeled topologies	Definitions of V, W, X, Y, Z	Number of labeled topologies in class
γ_1	1	((((VW)X)Y)Z)	V ∈ {A,B,C}, W ∈ {D,E}	36
	2	((((DE)X)Y)Z)	{X,Y,Z}={A,B,C}	6
	3	((((WC)X)Y)Z)	W ∈ {A,B}, X ∈ {D,E}	8
	4	((((WC)X)Y)Z)	{W,X}={A,B}, {Y,Z}={D,E}	4
	5	((((AB)X)Y)Z)	{X,Y,Z}={C,D,E}, X≠C	4
	6	((((AB)C)X)Y)	{X,Y}={D,E}	2
γ_2	1	(((VW)X)(YZ))	{V,X,Y}={A,B,C}, {W,Z}={D,E}	12
	2	(((WX)Y)(ZC))	{W,Z}={A,B}, {X,Y}={D,E}	4
	3	(((CX)Y)(AB))	{X,Y}={D,E}	2
	4	(((DE)X)(YC))	{X,Y}={A,B}	2
	5	(((DE)C)(AB))		1
	6	(((WC)X)(YZ))	{W,Y}={A,B}, {X,Z}={D,E}	4
	7	(((AB)X)(CY))	{X,Y}={D,E}	2
	8	(((XC)Y)(DE))	{X,Y}={A,B}	2
	9	(((AB)C)(DE))		1
γ_3	1	(((XD)(YE))Z)	{X,Y,Z}={A,B,C}	6
	2	(((VW)(XY))Z)	{V,W,X}={A,B,C}, {V,W} ≠ {A,B}	4
	3	(((AB)(CX))Y)	{X,Y}={D,E}	2
	4	(((XC)(DE))Y)	{X,Y}={A,B}	2
	5	(((AB)(DE))C)		1

In the case of species tree ψ_1, all gene trees with topology γ_2 or γ_3 are AGTs. For species tree ψ_3, all gene trees with topology γ_2 are AGTs. Only in the case of species tree ψ_2 is it possible for an AGT to have the same unlabeled topology as the species tree; there are three such AGTs, falling into two different history classes, and these are the only AGTs for species tree topology ψ_2. In no case is a gene tree with topology γ_1 an AGT.

TABLE 5. History classes for species tree topology ψ_3. Some elements of {V,W,X,Y,Z} appear under "Description of class of labeled topologies" but not under "Definitions of V, W, X, Y, Z." After the assignment to specific taxa of those elements of {V,W,X,Y,Z} that do appear under "Definitions of V, W, X, Y, Z," the remaining elements of {V,W,X,Y,Z} are chosen from the remaining taxa in {A,B,C,D,E}.

Gene tree topology	Number for history class	Description of class of labeled topologies	Definitions of V, W, X, Y, Z	Number of labeled topologies in class
γ_1	1	((((WE)X)Y)Z)	{W,X,Y,Z}={A,B,C,D}	24
	2	((((WX)E)Y)Z)	W ∈ {A,B}, X ∈ {C,D}	8
	3	((((WX)Y)E)Z)	W ∈ {A,B}, X ∈ {C,D}	8
	4	((((WX)Y)Z)E)	W ∈ {A,B}, X ∈ {C,D}	8
	5	((((CD)E)X)Y)	{X,Y}={A,B}	2
	6	((((CD)X)E)Y)	{X,Y}={A,B}	2
	7	((((CD)X)Y)E)	{X,Y}={A,B}	2
	8	((((AB)E)X)Y)	{X,Y}={C,D}	2
	9	((((AB)X)E)Y)	{X,Y}={C,D}	2
	10	((((AB)X)Y)E)	{X,Y}={C,D}	2
γ_2	1	(((WE)X)(YZ))	Y ∈ {A,B}, Z ∈ {C,D}	8
	2	(((XE)Y)(CD))	{X,Y}={A,B}	2
	3	(((XE)Y)(AB))	{X,Y}={C,D}	2
	4	(((WX)E)(YZ))	W ∈ {A,B}, X ∈ {C,D}	4
	5	(((CD)E)(AB))		1
	6	(((AB)E)(CD))		1
	7	(((WX)Y)(ZE))	W ∈ {A,B}, X ∈ {C,D}	8
	8	(((AB)X)(YE))	{X,Y}={C,D}	2
	9	(((CD)X)(YE))	{X,Y}={A,B}	2
γ_3	1	(((WX)(YE))Z)	W ∈ {A,B}, X ∈ {C,D}	8
	2	(((CD)(XE))Y)	{X,Y}={A,B}	2
	3	(((AB)(XE))Y)	{X,Y}={C,D}	2
	4	(((XC)(YD))E)	{X,Y}={A,B}	2
	5	(((AB)(CD))E)		1

The Anomaly Zone

For each history class that gives rise to AGTs, we can plot the portion of the space of branch lengths for which the AGTs occur (Figs. 3 to 5). For each species tree, for each pair of values of two of the branch lengths, the maximal value of the third branch length that produces AGTs in a given history class is identified. The full anomaly zone is then obtained by taking the maximum across all history classes. Each species tree requires an assignment of branches (T_2, T_3, T_4) to the x, y, and z axes, with the z-axis depicted by contours. In each case, this assignment is made so that if (x, y, z) is in the anomaly zone and $Z < z$, then all tested points (x, y, Z)—with four exceptions in the case of species tree ψ_3—are also in the anomaly zone.

Species tree ψ_1.—Figure 3a shows the regions in which AGTs occur for species tree ψ_1, for the 12 history classes in which the AGT has gene tree γ_2. The zones with AGTs from history classes 1 to 3 are quite small, as these cases (see Table 3) require that at least three of the four gene

TABLE 6. The number of distinct anomaly-producing lists of coalescent histories (history classes) among labeled gene tree topologies with a given unlabeled topology, for each species tree labeled topology (see Fig. 2).

	Species tree 1	Species tree 2	Species tree 3	Total
Gene tree 1	0	0	0	0
Gene tree 2	12	2	9	23
Gene tree 3	5	0	0	5
Total	17	2	9	28

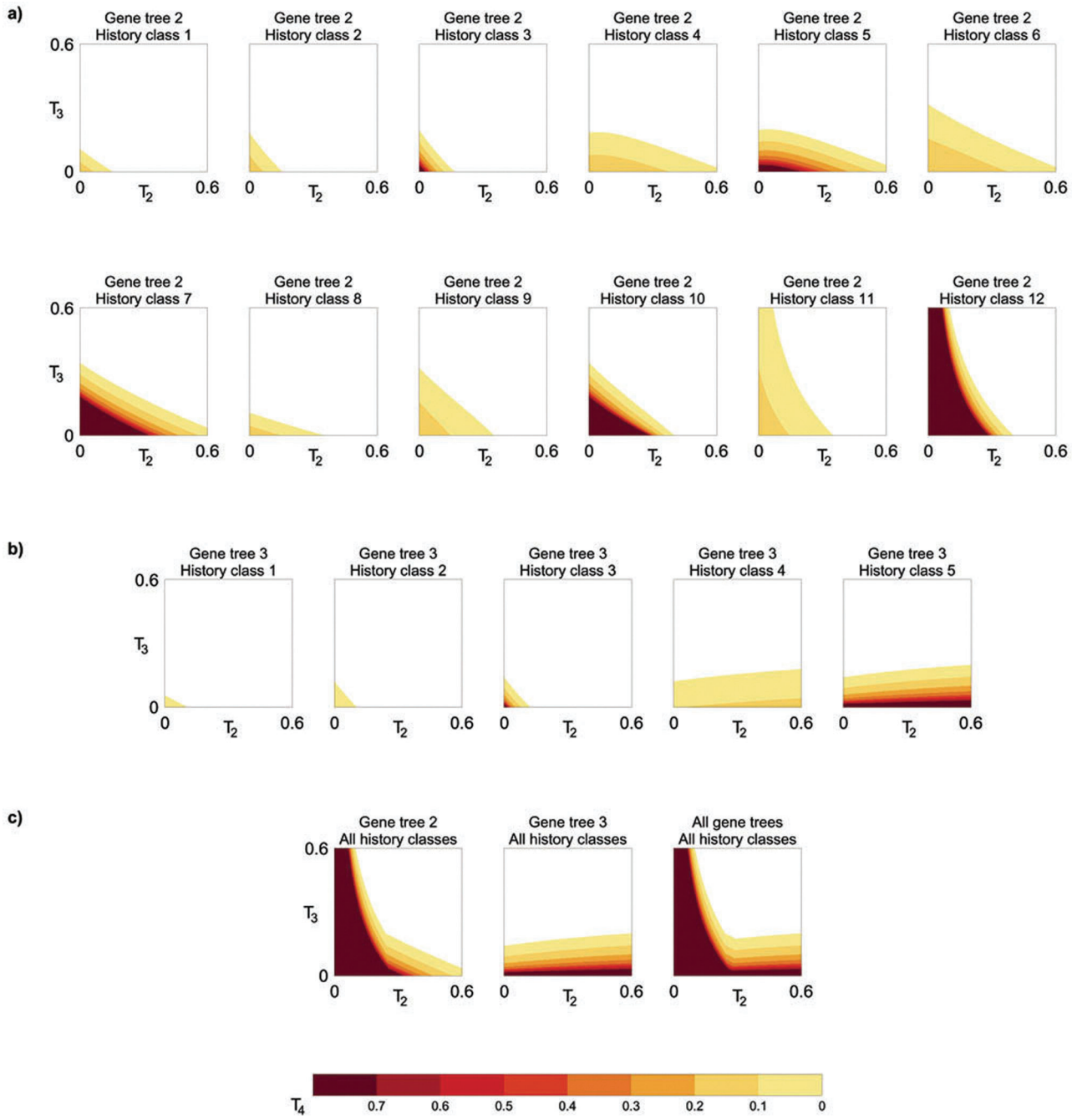

FIGURE 3. The anomaly zone for species tree 1. For each of the 17 history classes that produce anomalous gene trees (AGTs)—12 with gene tree 2 (a) and 5 with gene tree 3 (b)—each point (T_2, T_3) is shaded according to the largest value of T_4 for which AGTs in the history class occur. Each panel corresponds to a different history class. The darkest shade corresponds to $T_4 > 0.7$, and at some points, AGTs can occur when T_4 is arbitrarily large. The largest T_4 producing AGTs for any history class associated with gene tree 2, the largest T_4 producing AGTs for any history class associated with gene tree 3, and the largest T_4 producing AGTs for any history class and any gene tree are shown in (c). The figure was constructed by evaluating the equations in Supplementary Tables 16 to 18 at a grid of points with $T_2 \in [0, 0.6]$ at intervals of 0.002, $T_3 \in [0, 0.6]$ at intervals of 0.002, and $T_4 \in [0, 0.8]$ at intervals of 0.02. It was always observed that if a point (T_2, T_3, t) gave rise to an AGT, then (T_2, T_3, T_4) also produced AGTs for each $T_4 \in [0, t]$ that was investigated. If the largest T_4 producing AGTs occurred on a boundary between two colors, then a point was associated with the color shown to the left in the legend. Note that the meaning of colors differs between Figure 3 and Figures 4 and 5.

tree coalescences occur more anciently than the root of the species tree. Branch lengths must be very small in order to prevent coalescences from being likely along the internal branches.

History classes 4 to 7 have the form (((WX)E)(YZ)) rather than (((WE)X)(YZ)) as in history classes 1 to 3, where {W,X,Y,Z} = {A,B,C,D}. Thus, because these classes can have two gene coalescences more recent than

TABLE 7. The number of anomalous gene trees for each species tree labeled topology.

	Species tree 1	Species tree 2	Species tree 3	Total
Gene tree 1	0	0	0	0
Gene tree 2	30	3	30	63
Gene tree 3	15	0	0	15
Total	45	3	30	78

the root rather than one larger values of T_2 produce AGTs in these cases than for history classes 1 to 3.

As $T_4 \to \infty$, the gene lineages from species A and B become increasingly likely to coalesce. The large value of T_4 causes the case of species tree ψ_1 to approach the case of the four-taxon asymmetric species tree. The three AGTs for the limiting four-taxon case are represented by history classes 7, 10, and 12, each of which—especially history class 12, which corresponds to the AGT that occurs over the largest range of branch lengths in the four-taxon case—is anomalous over a comparatively large region of the parameter space. For these history classes, the regions of (T_2, T_3)-space in which the anomaly zone contains points with $T_4 > 0.7$ are noticeably similar to the regions in which corresponding AGTs occur in the four-taxon case (Degnan and Rosenberg, 2006, fig. 2).

For the five history classes in which the AGT has gene tree γ_3, Figure 3b shows the regions of branch-length space in which the AGTs occur. When all gene coalescences occur more anciently than the species tree root, each labeled topology for gene tree γ_3 has probability 1/90, compared to a larger probability of 1/60 for each labeled topology with gene tree γ_2. Thus, AGTs for most history classes with gene tree γ_3 occur over a smaller range than for those with gene tree γ_2. Interestingly, however, for larger T_2, the set of values of (T_3, T_4) for which AGTs occur in history classes 4 and 5 *increases* in size rather than decreases: an increase in the branch length leads to AGTs that cannot occur when the branch is shorter. This phenomenon is a consequence of the fact that larger values of T_2 increase the probability that gene lineages from species A, B, C, and D coalesce among themselves, without involving lineages from species E. Similarly to an increase in T_4, an increase in T_2 causes the scenario of species tree 1 to approach the case of the four-taxon asymmetric species tree. AGTs in the four-taxon case then correspond to history classes 4 and 5, and the larger parameter space producing AGTs in history class 5 reflects the fact that this history class corresponds to the four-taxon AGT that occurs over the broadest range of parameter values.

Figure 3c shows the regions of branch-length space that produce any AGT with gene tree γ_2, any AGT with gene tree γ_3, and any AGT of either type. These graphs are obtained by identifying at each point the history class with the largest value of T_4 that gives rise to AGTs. If a species tree has AGTs with gene tree γ_2, then this history class is always observed to be either class 7 or class 12. For AGTs with gene tree γ_3, it is always history class 5. Thus, considering both gene tree topologies, the overall

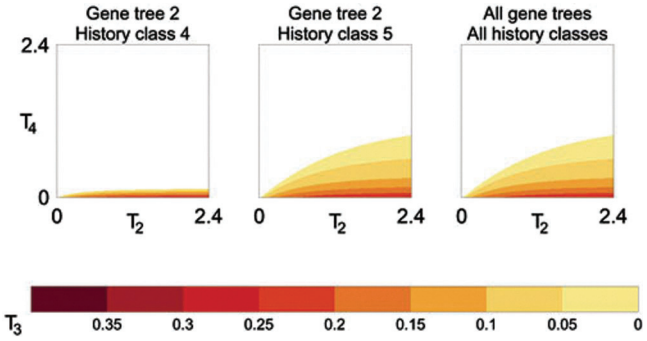

FIGURE 4. The anomaly zone for species tree 2. For both of the history classes that produce anomalous gene trees (AGTs), each point (T_2, T_4) is shaded according to the largest value of T_3 for which AGTs in the history class occur. The two panels on the left correspond to the two different history classes that produce AGTs. The largest T_3 producing AGTs for any history class and any gene tree is shown in the third panel. The figure was constructed by evaluating the equations in Supplementary Tables 19 to 21 at a grid of points with $T_2 \in [0, 2.4]$ at intervals of 0.008, $T_3 \in [0, 0.3]$ at intervals of 0.025, and $T_4 \in [0, 2.4]$ at intervals of 0.008. It was always observed that if a point (T_2, t, T_4) gave rise to an AGT, then (T_2, T_3, T_4) also produced AGTs for each $T_3 \in [0, t]$ that was investigated. If the largest T_3 producing AGTs occurred on a boundary between two colors, then a point was associated with the color shown to the left in the legend.

anomaly zone for species tree ψ_1 is constructed from the regions that produce AGTs in history classes 7 and 12 with gene tree γ_2 and history class 5 with gene tree γ_3.

Species tree ψ_2.—For species tree ψ_2, the anomaly zone is less extensive than for species tree ψ_1 (Fig. 4). AGTs occur in only two history classes, both of which are among the history classes of gene tree γ_2. The region in which AGTs occur for history class 5 entirely subsumes the corresponding region for history class 4, so that the full anomaly zone for species tree ψ_2 is identical to the zone that produces AGTs with history class 5. In the same way that the anomaly zone for species tree ψ_1 can increase in size as T_2 increases, the size of the anomaly zone for species tree ψ_2 can also increase as T_2 increases. Because topology 2 is the topology most probable when all coalescences occur more anciently than the species tree root, it is difficult to produce AGTs for species tree ψ_2 when all branch lengths are small. However, an increase in T_2 causes the case of species tree ψ_2 to approach the case of the four-taxon asymmetric species tree. Thus, as T_2 increases, the set of values of T_3 and T_4 that enable the production of AGTs becomes more extensive.

Species Tree ψ_3.—Figure 5 shows the regions of branch-length space that produce AGTs for species tree ψ_3. For this species tree, five of the nine history classes that produce AGTs do so in a nearly negligible region of the parameter space. Due to a symmetry between the roles of T_3 and T_4, AGTs occur most extensively both for history classes 5 and 6 and for history classes 8 and 9. For these history classes, if T_2 is small, then enough of the coalescences can occur more anciently than the root of the species tree so that topologies with gene tree γ_2 can be more probable than the matching gene tree topology. This occurs most easily for history classes 5 and 6, which

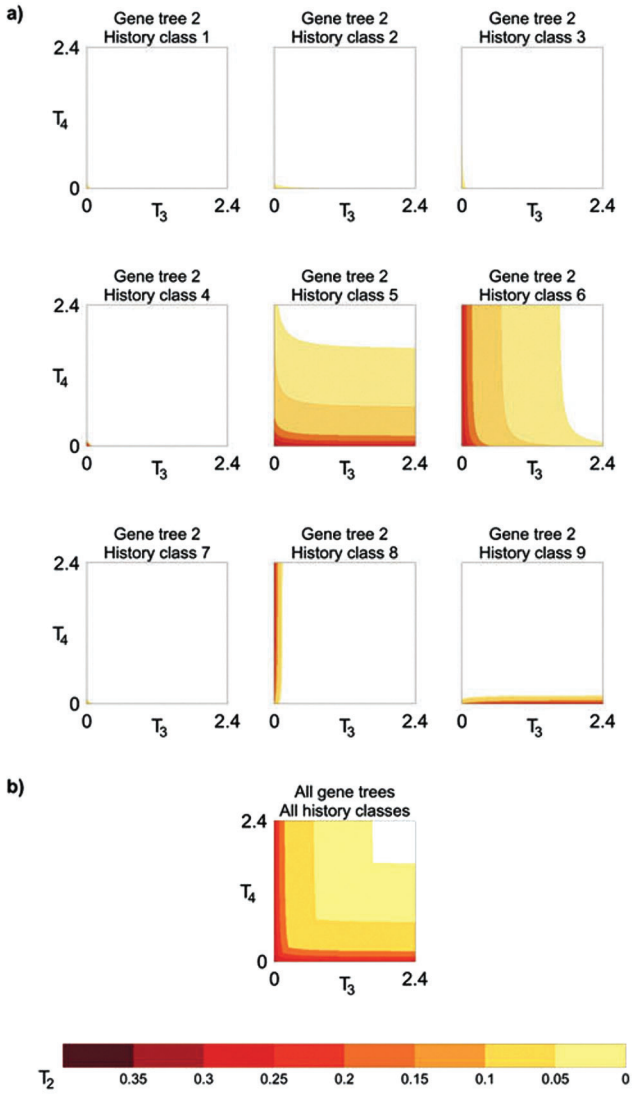

FIGURE 5. The anomaly zone for species tree 3. For each of the history classes that produce anomalous gene trees (AGTs), each point (T_3, T_4) is shaded according to the largest value of T_2 for which AGTs in the history class occur. The panels in (a) correspond to the nine different history classes that produce AGTs. The largest T_2 producing AGTs for any history class and any gene tree is shown in (b). The figure was constructed by evaluating the equations in Supplementary Tables 22 to 24 at a grid of points with $T_2 \in [0, 0.4]$ at intervals of 0.0125, $T_3 \in [0, 2.4]$ at intervals of 0.008, and $T_4 \in [0, 2.4]$ at intervals of 0.008. Except at four points, it was always observed that if a point (t, T_3, T_4) gave rise to an AGT, then (T_2, T_3, T_4) also produced AGTs for each $T_2 \in [0, t]$ that was investigated. If the largest T_2 producing AGTs occurred on a boundary between two colors, then a point was associated with the color shown to the left in the legend.

contain the two pairs of sister taxa (A and B, and C and D) present in the species tree, but which do not join these two pairs together. The full anomaly zone for species tree ψ_3 is obtained from the regions in which AGTs occur in history classes 5 and 6. In contrast to the anomaly zone for species tree ψ_1, but similarly to the anomaly zone for species tree ψ_2, the history classes that most easily produce AGTs tend to do so over a relatively wide range for the x and y variables but over a narrow range for the z variable.

Collapse of the Five-Taxon Anomaly Zone to the Four-Taxon Anomaly Zone

To verify that our results on the five-taxon anomaly zone are sensible, it is possible to show that the five-taxon anomaly zone collapses to the four-taxon anomaly zone when appropriate branch lengths are sent to ∞. For example, with species tree ψ_1, if $T_4 \to \infty$, then the species tree behaves as if it is a four-taxon species tree with taxa (AB), C, D, and E.

Using x and y for the more ancient and more recent internal branches of an asymmetric four-taxon species tree topology, there are five limiting scenarios involving five taxa that produce the asymmetric four-taxon species tree topology with internal branch lengths x and y (Table 8). For each of these scenarios, Fig. 6 plots four cross sections of the anomaly zone, with the variable approaching ∞ equal to 0, 0.1, 1, and 10. In each case, it can be observed that the five-taxon anomaly zone approaches the same form as the four-taxon anomaly zone in figure 2 of Degnan and Rosenberg (2006).

Several properties of AGTs are apparent from Fig. 6. For species tree ψ_1 (Figs. 6a and 6b), the anomaly zone is quite complex when all three branches are short, and it is subdivided into many compartments. For species tree ψ_2 (Fig. 6c), unlike for species trees ψ_1 and ψ_3, for which it declines, the anomaly zone grows across the four plots shown. Finally, for species tree ψ_3, as a result of symmetry in the species tree, the two sets of graphs (Figs. 6d and 6e) are identical. For species trees ψ_1 and ψ_3, it is possible to see that there are distinct regions with the same number of AGTs, but with nonidentical *sets* of AGTs. All three species trees illustrate the phenomenon that an increase in a branch can lead to more AGTs. For example, for ψ_1, consider an increase of T_2 into the dark red area towards the right side of the plot (5 AGTs) when $T_4 = 0.1$ (Fig. 6a); for ψ_2, this phenomenon is seen in the formation of the anomaly zone when T_2 is increased (Fig. 6c); finally, for ψ_3, consider an increase of T_3 into the higher of the two dark red areas (4 AGTs) when $T_4 = 0.1$ (Fig. 6d).

Maximal Length of the Shortest Internal Branch of a Species Tree with AGTs

A useful indicator of the severity of AGTs for a given species tree topology is the minimum value $T_{\min}$ so that

TABLE 8. Limiting scenarios for five-taxon species trees that produce the asymmetric four-taxon species tree with deeper and shallower internal branch lengths x and y, respectively.

Species tree	Scenario for limits of branch lengths		
1	$T_2 \to x$	$T_3 \to y$	$T_4 \to \infty$
1	$T_3 \to x$	$T_4 \to y$	$T_2 \to \infty$
2	$T_3 \to x$	$T_4 \to y$	$T_2 \to \infty$
3	$T_2 \to x$	$T_3 \to y$	$T_4 \to \infty$
3	$T_2 \to x$	$T_4 \to y$	$T_3 \to \infty$

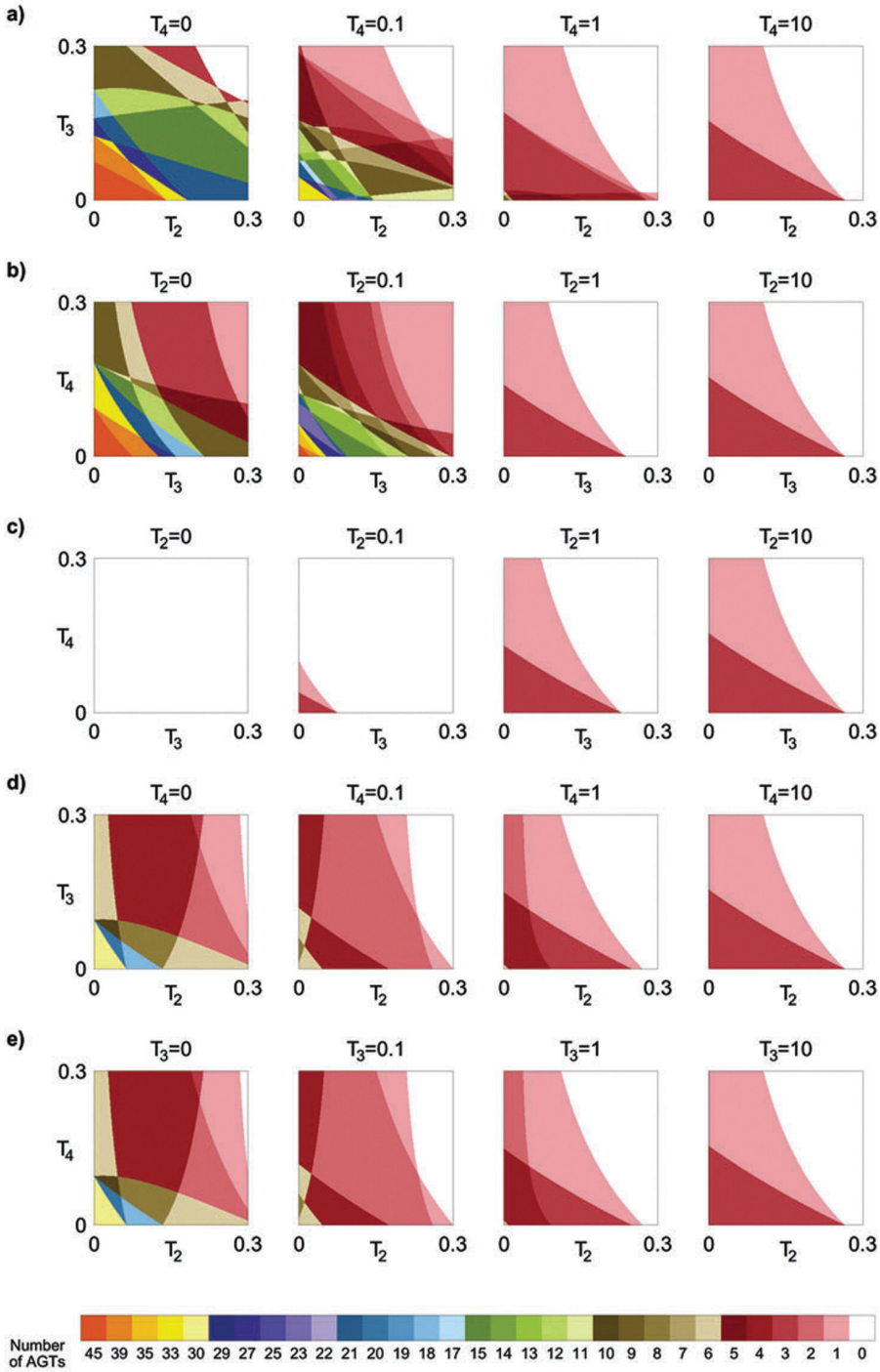

FIGURE 6. The collapse of the five-taxon anomaly zone to the four-taxon anomaly zone. Each row depicts four cross sections of the anomaly zone that show the increase of one branch length from a small value to a large value. In each graph, one of the three branches is held constant, as shown above the graph, and the other two are allowed to vary, as shown on the axes. The color of a point in a given plot represents the number of anomalous gene trees (AGTs) produced at the point. The right-most plots, where the branch length equals 10, illustrate in each case that when the appropriate branch is long, the anomaly zone for five taxa matches the anomaly zone for four taxa. (a, b) Species tree 1. (c) Species tree 2. (d, e) Species tree 3.

if all internal branches of the species tree have length at least $T_{\min}$, the species tree cannot have any AGTs. For four taxa, Degnan and Rosenberg (2006) found that $T_{\min} \approx 0.1569$.

To determine the values of $T_{\min}$ for the three species tree topologies with five taxa, for each inequality describing the anomaly zone in Supplementary Tables 16 to 24, we can set $T_4 = T_3 = T_2$ and solve for the smallest value

TABLE 9. The minimum value $T_{\min}$ so that if all internal branches of a species tree have length at least $T_{\min}$, the species tree has no anomalous gene trees in the specified history class. Each value shown is the smallest four-digit decimal number $T_{\min}$ so that the given species tree topology with branch lengths $(T_2, T_3, T_4) = (T_{\min}, T_{\min}, T_{\min})$ has no AGTs in the specified gene tree history class.

Species tree	Gene tree	History class	$T_{\min}$
ψ_1	γ_2	1	0.0477
		2	0.0568
		3	0.0690
		4	0.0890
		5	0.1226
		6	0.1044
		7	0.1604
		8	0.0570
		9	0.0884
		10	0.1432
		11	0.0939
		12	0.1935
ψ_1	γ_3	1	0.0307
		2	0.0394
		3	0.0523
		4	0.0601
		5	0.1014
ψ_3	γ_2	1	0.0254
		2	0.0322
		3	0.0322
		4	0.0431
		5	0.1803
		6	0.1803
		7	0.0325
		8	0.0643
		9	0.0643

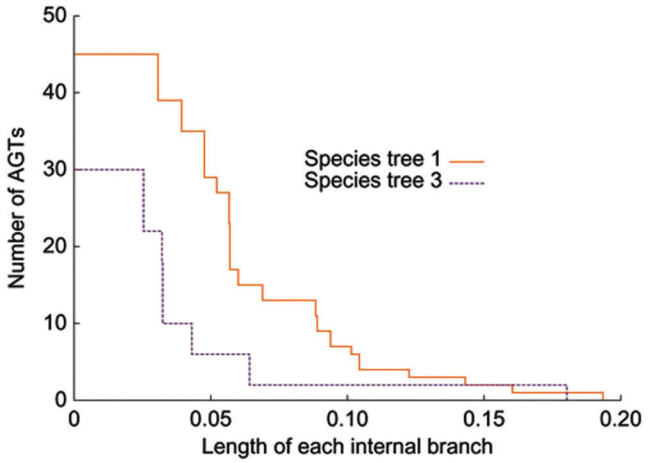

FIGURE 7. The number of anomalous gene trees (AGTs) when all branch lengths are equal, as a function of branch length (for each branch). For species tree 2, there are no AGTs when all branch lengths are equal. As the branch length increases, the specific history classes that drop out of the count of AGTs can be identified from the cutoff values in Table 9.

$T_{\min}$ that violates the inequality. For species tree ψ_2, we observe that $T_{\min} = 0$; in other words, if all branches of the species tree are equal, then the species tree has no AGTs.

For species trees ψ_1 and ψ_3, however, a portion of the anomaly zone lies on the line described by $T_4 = T_3 = T_2$, and the values of $T_{\min}$—restricting attention to AGTs of individual history classes—are shown in Table 9. From Table 9, we can observe that for species tree ψ_3, $T_{\min} \approx 0.1803$, and for species tree ψ_1, $T_{\min} \approx 0.1935$. These quantities indicate that internal branches must be slightly longer in the five-taxon case than in the four-taxon case to guarantee that a species tree topology has no AGTs.

Using the numbers of labeled topologies in each history class, we can plot the number of AGTs for a species tree, traveling away from the origin on the $T_4 = T_3 = T_2$ line (Fig. 7). For species tree ψ_1, the number of AGTs begins at 45, and for species tree ψ_3, it begins at 30. In both cases, the number of AGTs declines quickly, reaching 0 at 0.1935 and 0.1803 for species trees ψ_1 and ψ_3, respectively. For a small region between 0.1604 and 0.1803, the number of AGTs for species tree ψ_3, or 2, exceeds the number for species tree ψ_1, or 1.

Wicked Forests

Degnan and Rosenberg (2006) defined a "wicked forest" as a collection of mutually anomalous species trees—that is, a set W of species trees (with branch lengths) in which for each $\sigma_i, \sigma_j \in W$ the labeled topology ψ_i of species tree σ_i is an AGT for σ_j (see also Degnan, 2005). A gene tree evolving on a species tree σ_j in a wicked forest W is less likely to match the topology of σ_j than it is to match the topology of any other species tree in the forest. Wicked forests do not exist for four taxa: each AGT must have the symmetric unlabeled topology, and each species tree with a symmetric unlabeled topology has no AGTs.

Our results indicate that five is the smallest number of taxa for which wicked forests exist, and that all wicked forests with five taxa contain exactly two trees. To observe why this is true, note that a wicked forest cannot contain any tree with topology 1, as no AGTs have this unlabeled topology. It also cannot contain any tree with topology 3, as this topology is only anomalous if the species tree has topology 1, and no species trees with topology 1 can be in a wicked forest. Consequently, every wicked forest must contain only trees with topology 2.

Considering only trees with topology 2, suppose without loss of generality that a wicked forest W contains a tree with topology (((AB)C)(DE)). A species tree with this topology has AGTs with topologies (((DE)A)(BC)), (((DE)B)(AC)), and (((DE)C)(AB)). Because the AGTs for species tree topologies (((DE)A)(BC)) and (((DE)B)(AC)) do not include (((AB)C)(DE)), no trees with these topologies can be in W. However, species trees with topology ((DE)C)(AB)) can have AGTs with topology (((AB)C)(DE)). Thus, we can construct wicked forests that contain one tree with topology ((DE)C)(AB)) and branch lengths in the part of the anomaly zone that gives rise to AGTs with topology (((AB)C)(DE)), and one tree with topology ((AB)C)(DE)) and branch lengths in the part of the anomaly zone that gives rise to AGTs with topology (((DE)C)(AB)). Because a wicked forest contains at most one tree with a given labeled topology,

and because we have shown that no additional labeled topologies can be in a wicked forest with five taxa, every wicked forest with five taxa must contain exactly two trees.

DISCUSSION

In this article, we have analyzed the relationship of gene trees and species trees with five taxa, paying particular attention to the occurrence of anomalous gene trees. We have identified several qualitative aspects of gene tree discordance that first become apparent with five taxa. For example, AGTs can have the same unlabeled topology as the species tree topology, and the anomaly zone need not enclose ball-shaped regions around the origin in which all branches are short (species tree ψ_2). AGTs can occur in situations in which some branches are short and some are long (all three species trees). We have additionally shown that the smallest wicked forests occur for five taxa, and that wicked forests with five taxa must contain only two trees.

By counting AGTs for each species tree topology with five taxa, we have found that the number of AGTs in the anomaly zone can be considerable, especially for species trees ψ_1 and ψ_3. A fairly large fraction of species tree/gene tree combinations give rise to AGTs: if a random species tree labeled topology and a random gene tree labeled topology (different from the species tree topology) are chosen, then the probability that the species tree topology can have AGTs with the gene tree topology is $(60/105)(45/104) + (30/105)(3/104) + (15/105)(30/104) = 27/91 \approx 0.297$. The corresponding probability in the case of four taxa is $(12/15)(3/14) + (3/15)(0/14) = 6/35 \approx 0.171$.

Although some of the AGTs possible with five taxa occur over relatively large regions of branch-length space—especially for species tree ψ_1—others occur only in very small regions. Particularly for species tree ψ_3, for which 30 AGTs are possible, only two of these AGTs occur over branch-length regions that are reasonably large (history classes 5 and 6 of gene tree γ_2).

This collection of results points to an increasing complexity of the anomalous gene tree problem as the number of taxa increases from four to five. The increase in complexity is coupled with somewhat of an increase in the size of the anomaly zone. If the smallest species tree branch length in a four-taxon tree is at least 0.1569, then the tree is outside the anomaly zone (Degnan and Rosenberg, 2006); the corresponding value for five taxa is 0.1935. Although it is encouraging for phylogenetic inference that the anomaly zone increases in size by a relatively small amount as the number of taxa increases from four to five, the rapid increase in the number of AGTs and in the diversity of AGT-related phenomena highlights the importance of taking AGTs into consideration in procedures that aim to estimate species phylogenies from data on multiple loci.

ACKNOWLEDGMENTS

We thank J. Degnan for assistance with COAL, M. Blum and M. Jakobsson for helpful discussions, and D. Bryant and L. Kubatko for comments on a draft of the manuscript. This work was supported by grants from the National Science Foundation (DEB-0609760 and DEB-0716904), the Burroughs Wellcome Fund, and the Alfred P. Sloan Foundation.

REFERENCES

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16:23–34.

Brown, J. K. M. 1994. Probabilities of evolutionary trees. Syst. Biol. 43:78–91.

Degnan, J. H. 2005. *Gene tree distributions under the coalescent process*. PhD thesis, University of New Mexico, Albuquerque.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762–768.

Degnan, J. H. and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3:44–77.

Hein, J., M. H. Schierup, and C. Wiuf. 2005. Gene genealogies, variation and evolution. Oxford University Press, Oxford, UK.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Nichols, R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.

Nordborg, M. 2003. Coalescent theory. Pages 602–635. in *Handbook of statistical genetics*, 2nd ed. (D. J. Balding, M. Bishop, and C. Cannings, editors), Wiley, Chichester, UK.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.

Rosenberg, N. A. 2006. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann. Comb. 10:129–146.

Rosenberg, N. A. 2007. Counting coalescent histories. J. Comput. Biol. 14:360–377.

Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg. 2005. On the meaning and existence of an effective population size. Genetics 169:1061–1070.

Steel, M. and A. McKenzie. 2001. Properties of phylogenetic trees generated by Yule-type speciation models. Math. Biosci. 170:91–112.

Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957–966.

Takahata, N. and M. Nei. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110:325–344.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

First submitted 5 May 2007; reviews returned 9 August 2007; final acceptance 15 November 2007
Associate Editor: Laura Kubatko