Version dated: September 29, 2020

ILS-Aware Analyses of Retroelement Insertions in the Anomaly Zone

Erin K. Molloy[1], John Gatesy[2], Mark S. Springer[3]

[1]*Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA*

[2]*Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY , USA*

[3]*Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, Riverside, CA, USA*

Corresponding authors:

E. K. Molloy, E-mail: ekmolloy@cs.ucla.edu

M. S. Springer, Email: springer@ucr.edu

Abstract.— A major shortcoming of concatenation methods for species tree estimation is their failure to account for incomplete lineage sorting (ILS). Coalescence methods explicitly address this problem, but make various assumptions that, if violated, can result in worse performance than concatenation. Given the challenges of analyzing DNA sequences with both concatenation and coalescence methods, retroelement insertions have emerged as powerful phylogenomic markers for species tree estimation. We show that two recently

proposed methods, SDPquartets and ASTRAL_BP, are statistically consistent estimators of the species tree under the multispecies coalescent model, with retroelement insertions following a neutral infinite sites model of mutation. The accuracy of these and other methods for inferring species trees with retroelements has not been assessed in simulation studies. We simulate retroelements for four different species trees, including three with short branch lengths in the anomaly zone, and assess the performance of eight different methods for recovering the correct species tree. We also examine whether ASTRAL_BP recovers accurate internal branch lengths for internodes of various lengths (in coalescent units). Our results indicate that two recently proposed ILS-aware methods, ASTRAL_BP and SDPquartets, as well as the newly proposed ASTRID_BP, always recover the correct species tree on data sets with large numbers of retroelements even when there are extremely short species-tree branches in the anomaly zone. Dollo parsimony performed almost as well as these ILS-aware methods. By contrast, unordered parsimony, polymorphism parsimony, and MDC recovered the correct species tree in the case of a pectinate tree with four ingroup taxa in the anomaly zone, but failed to recover the correct tree in more complex anomaly-zone situations with additional lineages impacted by extensive incomplete lineage sorting. Camin-Sokal parsimony always reconstructed an incorrect tree in the anomaly zone. ASTRAL_BP accurately estimated branch lengths when internal branches were very short as in anomaly zone situations, but branch lengths were upwardly biased by more than 35% when species tree branches were longer. We derive a mathematical correction for these distortions, assuming the expected number of new retroelement insertions per generation is constant across the species tree. We also show that short branches do not need to be corrected even when this assumption does not hold; therefore, the branch lengths estimates produced by ASTRAL_BP may provide insight into whether an estimated species tree is in the anomaly zone.

(Keywords: coalescence; incomplete lineage sorting; Laurasiatheria; Palaeognathae;

polymorphism parsimony; transposon)

Concatenation methods for species tree construction have been and continue to be widely used in analyses of phylogenomic data sets. However, a pitfall of these methods is that they fail to account for incomplete lineage sorting (ILS). The consequences of this problem are most pronounced when the species tree has consecutive short branches in the anomaly zone. In these instances, concatenation may fail because the most probable gene tree(s) is different from the species tree (Degnan and Rosenberg 2006, 2009). Given this situation, numerous authors have proposed coalescence-based methods for species tree reconstruction that explicitly account for ILS. The three main approaches for estimating species trees in the framework of the multispecies coalescent (MSC) are (1) methods such as *BEAST (Heled and Drummond 2010) that co-estimate gene trees and species trees, (2) summary coalescence methods such as ASTRAL (Mirarab and Warnow 2015) that estimate species trees from gene trees, and (3) SNP methods such as SVDquartets (Chifman and Kubatko 2015) that infer species trees from nucleotide site patterns. Many of these methods are known to be statistically consistent under the multispecies coalescent given their assumptions (Nute et al. 2018; Roch et al. 2019; Islam et al. 2020). In the case of summary coalescence methods, where species trees are inferred from sequence-based gene trees, important assumptions of the MSC include neutral evolution, gene tree heterogeneity that results exclusively from ILS, and free recombination between loci but no intralocus recombination, where each locus is a coalescence gene (c-gene) (Liu et al. 2009). Intralocus recombination and violations of neutral evolution are also problematic for *BEAST. If these assumptions are violated there is no guarantee that coalescence methods for species tree estimation will perform any better than concatenation, and in many empirical analyses seem to perform worse (e.g., Xi et al. 2014; Hosner et al. 2016; Oliveros

et al. 2019). Indeed, theoretical arguments and empirical evidence suggest that violations of these assumptions may be problematic for the application of summary coalescence methods with sequence-based gene trees (Huang et al. 2010; Meredith et al. 2011; Patel et al. 2013; Gatesy et al. 2013; Gatesy and Springer 2014; Springer and Gatesy 2016, 2018a; Scornavacca and Galtier 2017; He et al. 2020). The problem of gene tree reconstruction error is especially troublesome and has been documented for numerous phylogenomic data sets (Mirarab and Warnow 2015; Simmons and Gatesy 2015; Springer and Gatesy 2016, 2017, 2018b; Gatesy et al. 2017, 2019; Shen et al. 2017). SNP methods avoid gene tree reconstruction error and problems that stem from intralocus recombination, but they can still be negatively impacted by non-neutral evolution, violations of the site substitution model, and deviations from ultrametricity.

Given these problems with existing coalescence methods, Springer et al. (2020) suggested that retroelement insertions are ideal markers for coalescence-based analyses because they satisfy the assumptions of the MSC much better than sequence-based gene trees or SNPs. Unlike DNA sequences, homoplasy is almost unknown for retroelement insertions (Shedlock et al. 2000, 2004; Ray et al. 2006; Kuritzin et al. 2016; Doronina et al. 2017, 2019), so conflicting patterns that look like homoplasy may be attributed to ILS, i.e. hemiplasy (Avise and Robinson 2008). In addition, retroelements likely come closer to satisfying the neutral evolution assumption of the MSC because they generally occur in regions of the genome that are safe havens from selection (e.g., introns, intergenic regions) (Chuong et al. 2017). Finally, retroelements are singular events, and the presence/absence of a retroelement insertion is not subject to intralocus recombination (Springer et al. 2020).

Given these desirable properties of retroelement insertions that match the MSC, Springer et al. (2020) proposed two quartet-based ILS-aware methods (ASTRAL_BP, SDPquartets) that can be applied to these markers. We show that both of these methods are statistically consistent under the MSC model, with retroelement insertions following a

neutral infinite sites model of mutation (see Mendes and Hahn 2017 for related theoretical results).

Whereas Springer et al. (2020) applied ASTRAL_BP and SDPquartets to published retroelement data sets for Placentalia (Nishihara et al. 2009), Laurasiatheria (Doronina et al. 2017), Balaenopteroidea (Lammers et al. 2019), and Palaeognathae (Cloutier et al. 2019; Sackton et al. 2019), these methods have not yet been tested on simulated data sets, where the true species trees are known. Variants of parsimony that commonly have been applied to retroelement data sets (Nikaido et al. 1999; Suh et al. 2015a; Lammers et al. 2019) also have not been assessed in simulation studies. Here, we show how the *ms* program (Hudson 2002) can be used to simulate retroelement insertions and use this approach for four model species trees, three of which include consecutive short branch lengths that are in the anomaly zone. We then analyze these data sets with eight different methods for species tree reconstruction that take retroelement insertions or other low-homoplasy binary (01) genomic characters such as nuclear copies of mitochrondrial genes (NUMTs) or large indels as input.

Of the tested methods, unordered parsimony, Camin-Sokal parsimony, and Dollo parsimony apply equal (unordered) or differential (Camin-Sokal, Dollo) weights to forward changes and reversals. Two additional parsimony methods (MDC, polymorphism) infer species trees by minimizing deep coalescences or the extent of polymorphism on the tree, respectively. Both SDPquartets and ASTRAL_BP effectively estimate species trees by analyzing retroelement insertions on subsets of four taxa; their utilization of quartets enables proofs of statistical consistency. ASTRAL_BP gets its name, because it is implemented by encoding each retroelement as a "gene tree" with a single bipartition and then applying ASTRAL (Zhang et al. 2018) to the resulting set of incompletely resolved "gene trees." We also explore using the gene tree summary method ASTRID (Vachaspati and Warnow 2015) in a similar fashion, calling this approach ASTRID_BP.

Among the eight methods that we assessed, ASTRAL_BP is the only method that returns internal branch lengths in coalescent units (CUs) (Sayyari and Mirarab 2016). Therefore, we examined whether ASTRAL_BP branch lengths are biased relative to the model species tree. This issue is critical because there are empirical examples of species trees that are in the anomaly zone based on analyses of sequence-based gene trees (Cloutier et al. 2019; Sackton et al. 2019) but are outside of the anomaly zone based on analyses of retroelement insertions (Springer et al. 2020). We show that ASTRAL's technique for branch length estimation is accurate when internal branch lengths are sufficiently short so that the small angle approximation applies. For longer branch lengths, a correction is required; we derive how to estimate branch lengths for retroelement insertion data sets when the expected number of new retroelement insertions is constant across the species tree. We evaluate our technique compared to the branch length estimates produced by ASTRAL on the retroelement insertion data sets simulated from four model species trees. Three of the species trees are in the anomaly zone, including one based on an empirical species tree for palaeognath birds (Cloutier et al. 2019; Sackton et al. 2019); this allowed us to test whether branch lengths inferred from ASTRAL_BP analysis are accurately estimated in anomaly zone conditions. The fourth species tree (26 taxa) includes internal branch lengths ranging from 0.1 to 7 CUs.

Methods

Model Species Trees

We simulated retroelements from four model species trees, three of which have very short branches, in coalescent units (CUs), and are in the anomaly zone (Fig. 1). The first three trees have a long ingroup stem branch (20 CUs) to preclude deep coalescences between any

ingroup taxa and the outgroup. Because of this long branch, the anomaly zone for unrooted gene trees converges to the anomaly zone for rooted gene trees on the ingroup taxa (Degnan 2013).

The first model species tree (4-ingroup taxa anomaly zone tree; Fig. 1A) is a pectinate tree with four ingroup taxa (A,B,C,D) and an outgroup (Out) where the two shallowest internal branches each have length 0.01 CUs. We let x denote the deeper branch and y the shallower branch. This tree is based on the 4-taxa anomaly zone tree employed by Mendes and Hahn (2017). A 4-taxa pectinate tree is the simplest case for the anomaly zone for rooted gene trees (Degnan and Rosenberg 2006), and thus a 5-taxa tree is the simplest case for unrooted anomalous genes trees (Degnan 2013). The two very short branches (0.01 CUs) within the ingroup ensure that this pectinate tree is deep in the anomaly zone because the minimum requirement for equal branch lengths x and y in the anomaly zone is 0.1542 CUs (Degnan and Rosenberg 2006); this is more than an order of magnitude longer than the branch lengths in our species tree.

The second model species tree (5-ingroup taxa anomaly zone; Fig. 1B) is a pectinate tree for five ingroup taxa (A,B,C,D,E) and an outgroup (Out) and includes three consecutive short internal branches of 0.01, 0.01, and 0.1 CUs (Fig. 1B) where x is the deepest branch, branch y is intermediate in depth, and z is the shallowest branch. This tree is also in the anomaly zone based on these internal branch lengths and represents one example of a 5-taxa anomaly zone tree (Rosenberg and Tao 2008; Degnan and Rosenberg 2009). The third model species tree (Palaeognathae anomaly zone; Fig. 1C) is based on Cloutier et al.'s (2019) ASTRAL analysis of 20,850 loci (12,676 CNEEs, 5,016 introns, 3,158 UCEs) for palaeognath birds (ratites, tinamous) and a chicken outgroup. We shortened the lengths of some of the terminal branches so that the final tree was ultrametric. Cloutier et al.'s (2019) species trees based on ASTRAL analysis contains three successive short branches that are within the anomaly zone. The fourth model species tree

(26-taxa species tree; Fig. 1D) does not have consecutive short branches in the anomaly zone but instead includes a wider range of internal branch lengths for examining potential branch length distortions in ASTRAL_BP analysis.

Simulations

Retroelement insertions were simulated with the *ms* program (Hudson 2002), which enables coalescent simulations with 0/1 mutations occurring under a neutral infinite sites model of mutation. Kuritzin et al. (2016) and Doronina et al. (2019) previously utilized such a model in developing methods for detecting introgression using retroelement data sets. The infinite sites model with 0/1 mutations is appropriate for simulating retroelement insertion data because (1) retroelements are presence/absence (1/0) characters and (2) retroelement insertions at specific genomic sites are rare events, as are back mutations (i.e., precise excision of an inserted sequence) (Shedlock and Okada 2000; Doronina et al. 2019). Our simulations further assume free recombination among loci, no intralocus recombination, neutrality, no missing data, constant effective population size, and a uniform rate of retroelement insertions per unit length of the species tree. We simulated 25 replicate data sets from each of the four model species trees (Fig. 1) with one segregating site for each gene tree locus, where the probability of selecting a site on a given branch is proportional to its branch length divided by the total length of the gene tree (Hudson 2002). Given that we were primarily interested in whether different analytical methods converge on the correct species tree when there are short consecutive branches in the anomaly zone, we simulated data sets that were sufficiently large to contain more than 100,000 informative retroelements (i.e. the retroelement insertion induces at least one quartet) and then pruned these data sets to exactly 100,000 informative retroelements (note that species tree branch lengths were halved prior to simulating data, because the *ms* program uses a currency of $4N$ generations per unit for species tree and gene tree branch lengths, whereas a coalescent

unit is $2N$ generations for a population of diploid individuals). We used a custom script for each species tree to simulate 25 data sets with *ms* and convert the output of each of these data sets into a nexus file (available on Dryad). Next, we used a batchfile command (available on Dryad) in PAUP* to perform the following operations: (1) execute each of 25 data sets, and for each data set, (2) exclude uninformative characters, (3) export a nexus file with informative characters only, (4) execute the new data set with informative characters only, (5) exclude all characters after the first 100,000 characters, (6) export a nexus file with the 100,000 informative characters, and (7) export a phylip file with the 100,000 informative characters. In addition, each phylip file with 100,000 binary characters was converted into a Newick tree file with 100,000 bipartitions (each represented by a Newick string) using a script from Springer et al. (2020).

Species Tree Estimation

We estimated species trees using eight different phylogenetic methods: unordered parsimony, Camin-Sokal parsimony, Dollo parsimony, polymorphism parsimony, minimize deep coalescences (MDC), ASTRAL_BP, ASTRID_BP, and SDPquartets. All eight methods were applied to data sets that were simulated with the four species trees shown in Figure 1. Unordered, Camin-Sokal, and Dollo parsimony analyses were executed with PAUP* 4.0a168 (Swofford 2002). Unordered parsimony applies equal weights to forward (0 to 1) and reverse (1 to 0) changes; Camin-Sokal parsimony only allows forward changes; and Dollo parsimony allows for one forward change and as many reversals as are necessary to explain the character data (Felsenstein 2004). We used branch-and-bound searches for all analyses with the exception of the 26-taxa data set where we employed heuristic searches for Camin-Sokal and Dollo parsimony. In these cases, heuristic searches employed tree-bisection and reconnection branch swapping and stepwise addition with 100 randomized input orders of taxa. Polymorphism parsimony analyses were performed with

the dollop program in PHYLIP version 3.695 (Felsenstein 1989) with the jumble option set to 50. For presence/absence (01) characters, polymorphism parsimony assumes that after a state of polymorphism for the two alleles is established in an ancestral population, all subsequent occurrences of state 0 or state 1 in terminal taxa result from losses of one or the other allele (Felsenstein 2004). We used PhyloNet (Than et al. 2008; Than and Nakhleh 2009) to implement the MDC approach of Maddison (1997). MDC is a parsimony-based approach that infers a species tree from a set of gene trees, which in our case are incompletely resolved and include only a single bipartition, by minimizing the number of extra allelic lineages. Sanderson et al. (2020) suggested that polymorphism parsimony and MDC are equivalent approaches for inferring species trees. ASTRAL-III (Zhang et al. 2018) and ASTRID (Vachaspati and Warnow 2015) are summary coalescence methods that allow for polytomies, but only the former returns branch lengths in CUs. As previously mentioned, ASTRAL_BP (Springer et al. 2020) and ASTRID_BP construct a species trees by representing each retrolement insertion as a newick string with a single bipartition and then running ASTRAL-III (version 5.7.3) or ASTRID, respectively. SDPquartets (Springer et al. 2020) is a quartet-based method that was developed for low-homoplasy 01 (absence/presence) data such as retroelements. The first step with SDPquartets is to perform parsimony analyses with all possible subsets of four species. In the second step, optimal species trees on four taxa are assembled into a species tree on the full set of taxa using Matrix Representation with Parsimony (MRP) (Ragan 1992). We performed SDPquartets analyses with a custom Perl script (`https://github.com/dbsloan/SDPquartets`) that directs PAUP* (Swofford 2002) to perform both steps of the analysis. We used branch-and-bound searches for the parsimony analyses of the MRP matrices to ensure recovery of all most parsimonious trees.

Results

Theory

Statistical consistency. Given a retroelement insertion on four species $\{A, B, C, D\}$, patterns 1100 and 0011 correspond to quartet $AB|CD$, patterns 1010 and 0101 correspond to quartet $AC|BD$, and patterns 1001 and 0110 correspond to quartet $AD|BC$. We assume that retroelement insertions are generated under the MSC + infinite sites neutral mutation model, parameterized by a rooted species tree topology on $\{A, B, C, D\}$, where each branch is annotated by the amount of time in generations, the effective population size, and the probability of new insertions for each individual allele in the population (note that the latter two parameters must also be specified for the population above the root).

Doronina et al. (2017) provided an approximation for the expected number of retroelement insertions displaying each of the six patterns when retroelement insertions are generated from four-taxon species networks. Under their approximation, which is based on the diffusion approximation of the Wright-Fisher coalescent model (Fisher 1922; Wright 1931) and the neutral mutation model (Kimura 1955a,b), we show that for the pectinate rooted species tree $(((A, B), C), D)$ and for the balanced rooted species tree $((A, B), (C, D))$,

$$P(1100) + P(0011) > P(1010) + P(0101) = P(1001) + P(0110) \tag{1}$$

where $P(1100)$ is the probability that a retroelement insertion displaying one of the six informative patterns displays pattern 1100 (Theorem 6 in the Appendix). Theorem 6 does not require the expected number of new insertions per generation to be constant across the tree, and we use this result to show that SDPquartets and ASTRAL_BP are statistically consistent.

Theorem 1. Suppose that retroelement insertions are generated under the MSC with

insertions following an infinite sites neutral model (as approximated by Doronina et al. 2017), with a constant rate of insertions per generation across the four-taxon species tree. Then, SDPquartets using a branch-and-bound algorithm is statistically consistent.

Proof. SDPquartets uses parsimony to identify the species tree from the retroelement insertions restricted to every possible subset of four taxa. Specifically, for each of the three possible quartet topologies, denoted t_1, t_2, t_3, on four taxa, the parsimony score is computed as

$$score(t_i) = N(t_i) + \left(2 \times \left(N(t_{j \neq i,j}) + N(t_{k \neq i,j})\right)\right)$$

where $N(t_i)$ is the number of retroelement insertions that display topology t_i, and the tree with the lowest parsimony score is added to the set $\mathcal{T}$ of source trees. By Theorem 6 in the Appendix, the most probable quartet agrees with the species tree and the two alternative quartets have equal probability. Therefore, as the number of retroelement insertions goes to infinity, SDPquartets identifies the true species tree on subsets of four taxa with probability going to one, so the true species tree T^* will be the unique compatibility supertree for $\mathcal{T}$ with high probability.

SDPquartets runs the supertree method Matrix Representation with Parsimony (Ragan 1992) given $\mathcal{T}$. By Theorem 7.8 in Warnow (2017), when $\mathcal{T}$ are compatible, any optimal solution to MRP is a refined compatibility supertree for $\mathcal{T}$ (see Sections 3.2.1, 7.2, and 7.5 in Warnow 2017 for details). MRP is an NP-hard problem (Theorem 7.8 in Warnow 2017); however, branch-and-bound algorithms (Hendy and Penny 1982) guaranteed to find the optimal solution can be utilized whenever the number of taxa is sufficiently small. In this case, as the number of retroelement insertions goes to infinity, the optimal solution to MRP given $\mathcal{T}$ equals T^* with probability going to one, so SDPquartets returns the true species tree with high probability. □

The proof of statistical consistency for ASTRAL_BP is closely related to the proof

of statistical consistency for ASTRAL (Theorem 2 in Mirarab et al. (2014)), so we provide the proof in the Appendix (Theorem 3).

Branch length estimation. ASTRAL not only estimates the species tree topology but also the internal branch lengths (in CUs). Branch length estimation is based on quartet frequencies (i.e. the number z_1 of gene trees that display the quartet induced by the branch divided by the total number n of gene trees); see Sayyari and Mirarab 2016 for details. Assuming the branch in question is correct, the maximum likelihood (ML) estimate of its length is $\hat{\tau} = -\log(\frac{3}{2}(1 - \frac{z_1}{n}))$ (Theorem 2 in Sayyari and Mirarab 2016). This follows from their statistical framework (Lemma 1 in Sayyari and Mirarab 2016) and from the probability of gene trees under the MSC:

$$p^G_{A,B|C,D} = 1 - \frac{2}{3}e^{-\tau} \quad \text{and} \quad p^G_{A,C|B,D} = p^G_{A,D|B,C} = \frac{1 - p^G_{A,B|C,D}}{2} \tag{2}$$

where τ is the length (in CUs) of the internal branch inducing $A, B|C, D$ in the model species tree (Section 4.1 in Allman et al. 2011). As $A, B|C, D$ agrees with the species tree, we refer to it as the "dominant quartet"; we refer to $A, C|B, D$ and $A, D|B, C$ as the "alternative quartets."

The statistical framework proposed by Sayyari and Mirarab (2016) can be applied to retroelement insertions (Appendix); however, the formula for the probability of the dominant quartet is more complicated and depends on whether the model species tree is pectinate or balanced (Appendix). When internal branches of the model species tree are short enough so that the small angle approximation $e^{-\tau} = 1 - \tau$ can be applied, the probability of the dominant quartet for the pectinate and balanced species tree simplifies to

$$p_{A,B|C,D} \approx \frac{1}{3} + \frac{2}{3}\tau \quad \text{and} \quad p_{A,C|B,D} = p_{A,D|B,C} \approx \frac{1 - p_{A,B|C,D}}{2}. \tag{3}$$

Applying the small angle approximation to Equation 2 also yields Equation 3; therefore, the ML branch lengths estimated using ASTRAL are applicable to retroelement insertions whenever the internal branches are sufficiently short (Figure 2).

To estimate longer branch lengths from retroelement insertions, a correction is required. If the expected number of new retroelement insertions per generation is constant across the species tree, the probability of the dominant quartet simplifies to

$$p^R_{A,B|C,D} = \frac{\frac{1}{3}e^{-\tau} + \tau}{e^{-\tau} + \tau} \quad \text{and} \quad p^R_{A,C|B,D} = p^R_{A,D|B,C} = \frac{1 - p^R_{A,B|C,D}}{2} \tag{4}$$

for both the pectinate and balanced species tree (Appendix). Then, using the statistical framework proposed by Sayyari and Mirarab (2016), we show that the ML estimate of the branch length is

$$\hat{\tau} = W[\frac{2}{3}(\frac{z_1}{n} - 1)^{-1} - 1] \tag{5}$$

where W is Lambert's W (Theorem 4 in the Appendix). This correction can be applied by running ASTRAL with the "-t 2" option to get the average quartet frequency for each branch (referred to as the normalized quartet support) and then substituting this value into Equation 5 for $\frac{z_1}{n}$. The ML estimate of the branch length does not exist when $\frac{z_1}{n} = 1$ (i.e. there is no conflict); in this case, we set the branch length to ∞. We set the $\hat{\tau}$ to 0 when $\frac{z_1}{n} < \frac{1}{3}$ (as the branch is not in the species tree). A simple Python script for correcting branch lengths is available on Dryad; our hope is to integrate this as an option of ASTRAL in the near future.

Branch support. Lastly, ASTRAL provides a measure of branch support: the local posterior probability (local PP). This measure of support is appropriate for retroelement insertions with two caveats. First, the calculation of local PP is based on the effective number (EN) of gene trees (in this case retroelement insertions) for the branch. Because

retroelement insertion does not induce quartets on all subsets of four taxa, some insertions will not have any information about the resolution of the branch in question. For retroelement insertion data sets, the EN can be quite low on some branches, and local PP should be interpreted cautiously in this case. We recommend reporting EN when analyzing retroelement insertion data sets with ASTRAL_BP. Second, Lemma 2 in Sayyari and Mirarab (2016) states that local PP corresponds to the species tree being generated under a Yule process with birth rate λ; furthermore, when $\lambda = \frac{1}{2}$ (the default in ASTRAL), this corresponds to the prior on the probability of the dominant quartet being uniform. This interpretation (regarding the generation of the species tree under the Yule process) does not hold for retroelement insertions (Appendix); nevertheless, it seems reasonable to put a uniform prior for the probability of the dominant quartet. Lastly, ASTRAL returns the maximum a posteriori (MAP) estimate of branch lengths by default, which is based on the branch lengths being exponentially distributed. When the number of gene trees is large, this converges to the ML estimate, so we report the MAP estimate in the main text and the ML estimate in the Supplementary Text.

Simulation Study

4-ingroup taxa anomaly zone species tree. For the simulated tree with four ingroup taxa in the anomaly zone and a long outgroup branch, seven of eight methods (ASTRAL_BP, ASTRID_BP, SDPquartets, unordered parsimony, Dollo parsimony, polymorphism parsimony, MDC) returned the correct species-tree topology for all 25 simulated data sets (Fig. 3A). Camin-Sokal parsimony always recovered an incorrect position for Taxon C as the sister to Taxon D instead of sister to Taxon A + Taxon B (Fig. 3A).

Our results for unordered parsimony are consistent with those of Mendes and Hahn (2017) who also recovered the correct species tree with parsimony. A minor difference is

that these authors simulated mutations down each gene tree under a Jukes-Cantor model. Mendes and Hahn (2017) hypothesized that parsimony should return the correct species tree inside the 4-taxa (ingroup) anomaly zone even though the most probable gene tree(s) differs from the species tree. This is because the anomalous gene trees have very short internal branches, on average, relative to internal branches on gene trees that agree with the species tree. The net effect of these branch length differences is that the most common (democratic) site patterns will still support the correct species tree.

Than and Rosenberg (2011) showed that for a pectinate species tree with four ingroup taxa, the MDC criterion is statistically inconsistent if branch $x = y < 0.2215$ CUs. By contrast, the corresponding length for the democratic vote criterion (i.e., favor species tree that matches the most common gene tree) is $x = y < 0.1542$ CUs (Degnan and Rosenberg 2006). Thus, the anomaly zone is larger with MDC than with a simple democratic vote even though the MDC criterion specifically considers the mechanism of deep coalescence (Than and Rosenberg 2011). However, MDC is based on a parsimony criterion and fails to consider all elements of the multispecies coalescent such as the probability of a gene tree given a species tree. MDC also ignores branch lengths in gene trees. Than and Rosenberg (2011) suggested that these deficiencies may explain the statistical inconsistency of the MDC criterion. Given the above points, it is notable that MDC recovered the correct species tree in 25 of 25 simulations with four ingroup taxa even though the x and y branch lengths on the species tree are both 0.01, which is well below the threshold of 0.2215 CUs that results in an incorrect species tree when MDC is applied to full gene trees that are simulated from a species tree. We suggest that MDC infers the correct four-ingroup species tree with simulated retroelements because these are presence/absence characters, each of which corresponds to a single bipartition on a gene tree, and are more likely to occur on the generally longer internal branches of gene trees that agree with the species tree (Mendes and Hahn 2017).

Mean branch lengths on the ASTRAL_BP tree for branches x and y have lengths 0.0096 and 0.0102 CUs, respectively (Fig. 3A). Differences between the estimated and true branch lengths of 0.01 are minor, with mean error of 0.0021 and 0.0025, respectively. This is consistent with our theoretical results showing that small branch lengths do not need to be corrected, making ASTRAL_BP a useful tool for determining whether the estimated species tree is in the anomaly zone.

5-ingroup taxa anomaly zone species tree. By contrast with the 4-ingroup taxa anomaly zone species tree, only four of eight methods (ASTRAL_BP, ASTRID_BP, SDPquartets, Dollo parsimony) recovered the correct 5-ingroup taxa anomaly zone tree (Fig. 3B). Unlike the three ILS-aware methods and Dollo parsimony that recovered the correct species tree for all 25 simulated datasets, unordered parsimony, Camin-Sokal parsimony, polymorphism parsimony, and MDC always recovered incorrect species trees that were not fully pectinate. These results demonstrate that many methods that have been previously applied to retroelement data sets are not immune to anomaly zone problems when there are more than four ingroup taxa. Indeed, Roch and Steel (2015) showed that concatenation (parsimony or maximum likelihood) can be positively misleading under the coalescent + infinite sites neutral mutation model for a 6-taxa species tree in the anomaly zone. Among methods that estimated the incorrect species tree, MDC always recovered the ingroup topology $((E,(A,B),(C,D))$, but polymorphism parsimony only recovered this topology for 15 of 25 data sets and in ten other cases recovered different incorrect topologies. These results suggest that MDC and polymorphism parsimony do not always generate the same results (contra Sanderson et al. (2020)), at least as we have executed analyses using the programs for MDC and polymorphism parsimony.

ASTRAL_BP recovered average branch lengths of 0.0103, 0.0103, and 0.1059 for branch x (0.01 CUs), branch y (0.01 CUs), and branch z (0.1 CUs), respectively (Fig. 3B).

The mean error was again small: 0.0018, 0.0018, and 0.0060, respectively.

Palaeognathae anomaly zone species tree. ASTRAL_BP, ASTRID_BP, and SDPquartets recovered the correct species tree for all 25 simulated data sets (Fig. 3C). On these trees, rheas are the sister-taxon to kiwis + emu + cassowary. Dollo parsimony recovered the correct tree for 22 of 25 simulated data sets and in the other three instances reconstructed rheas as the sister-taxon to kiwis + emu + cassowary + tinamous. The other four methods (unordered parsimony, Camin-Sokal parsimony, polymorphism parsimony, MDC) recovered the correct species tree except for the placement of rheas, which were always estimated as the sister taxon to tinamous (Fig. 3C). This misplacement of rheas occurs in a region of the species tree where there are consecutive short branches in the anomaly zone. Together with our results for the 5-ingroup taxa anomaly zone tree, the palaeognath results suggest that unordered parsimony, Camin-Sokal parsimony, polymorphism parsimony, and MDC are inappropriate methods for estimating species trees from retroelements when the anomaly zone is more complicated than a pectinate tree with four ingroup taxa. Dollo parsimony performs much better than the other parsimony methods in the anomaly zone situations examined here, although it was not as efficient or accurate as ASTRAL_BP, ASTRID_BP, and SDPquartets. These results are significant because retroelement data sets are commonly analyzed using variants of parsimony, including Camin-Sokal (e.g. Nikaido et al. 1999; Nilsson et al. 2010; Suh et al. 2011), unordered (e.g. Gatesy et al. 2013, 2019), polymorphism (e.g. Suh et al. 2015b; Doronina et al. 2015), and Dollo (e.g. Lammers et al. 2019).

The palaeognath species tree (Cloutier et al. 2019) based on ASTRAL analysis of sequence-based gene trees (Fig. 1C) includes three consecutive short branches with lengths of $x = 0.3874$, $y = 0.0194$, and $z = 0.0532$ CUs. These three consecutive branches are consistent with an anomaly zone situation for five taxa (Rosenberg 2013) and are the basis

for the claim that the palaeognath tree provides an empirical example of the anomaly zone (Cloutier et al. 2019; Sackton et al. 2019). By contrast, Springer et al. (2020) reconstructed a palaeognath species tree based on ASTRAL_BP analysis of 4301 retroelement insertions from Cloutier et al. (2019) and recovered much longer branch lengths: x has length 2.5390 (∞ corrected—because the ML estimate does not exist), y has length 0.8939 (0.8657 corrected), and z has length 0.2528 (0.2587 corrected) CUs (Table 1). This suggests that the palaeognath species tree based on retroelements is well outside of the anomaly zone (although this result should be interpreted cautiously, as the effective numbers of retroelement insertions that induce quartets around branches x, y, and z are 18, 26.23, and 13.3, respectively). This result is in contrast with the results of our simulation study, where we simulated 25 retroelement insertion data sets from Cloutier et al.'s (2019) ASTRAL species tree and found that ASTRAL_BP given these data produced species trees in the anomaly zone. Specifically, ASTRAL_BP produced trees with the following mean branch lengths: x has length 0.452 (0.388 corrected), y has length 0.0189 (0.0188 corrected), and z has length 0.0563 (0.0549 corrected). These branch lengths are also consistent with an anomaly zone situation (Table 1 in Supplemental Text). Lastly, while our correction tool does not have a large impact on these short branches, for branches greater than 1 CU, the mean percent error dropped from above 30% to 1% following correction (Figure 4A–C).

26-taxa species trees. Given the biased increase in branch lengths for longer branches on the Palaeognathae anomaly zone tree, we simulated retroelement data sets for a 26-taxa tree with internal branch lengths that range from 0.1 to 6.0 CUs. This range of branch lengths does not include the stem branch for the clade comprised of Taxa A–T that has a length of 7.0 CUs because this internal branch is merged with the stem branch leading to Taxa U–Z on the inferred ASTRAL_BP species trees (Fig. 1D). There is no anomaly zone for the 26-taxa tree, and seven of eight phylogenetic methods estimated the correct species

tree for all 25 simulated data sets (Fig. 3D). Camin-Sokal parsimony recovered an incorrect phylogeny in 24 of 25 replicates with Taxon E + Taxon F misplaced as the sister to Taxon G + Taxon H. Applying our correction tool to branch lengths larger than 1 CU and shorter than 4 CUs reduced the mean percent error from over 30% to 1–3% (Fig. 4D–F). The percent error increases for uncorrected branch lengths larger than 4 CUs, but this increase in error is expected as the conditioning of the ML branch length estimation problem worsens with increasing branch lengths (Table 2 in the Supplementary Text).

Discussion and Conclusions

Comparison of different methods for estimating species trees with retroelements. We developed a pipeline for simulating retroelements based on the *ms* program (Hudson 2002) and used this simulation approach to compare the accuracy of eight phylogenetic methods for inferring the correct species tree from retroelement data. These methods include two summary coalescence approaches (quartet-based ASTRAL_BP, distance-based ASTRID_BP), one quartet-based coalescent method for 01 characters (SDPquartets), one method that minimizes deep coalescence (MDC), one method that minimizes the extent of polymorphism on the tree (polymorphism parsimony option of dollop), and three character-based parsimony methods (unordered, Camin-Sokal, Dollo) that give different weights to forward and reverse changes. All of these methods were tested on model species trees in the anomaly zone including the 4-ingroup taxa anomaly zone and 5-ingroup taxa anomaly zone trees that have consecutive short branch lengths (0.01 CUs). These branch lengths are much shorter than the minimum length of 0.1542 CUs that is required for two consecutive and equal short branches to remain in the anomaly zone for gene-tree based analyses (Degnan and Rosenberg 2006). Moreover, Patel et al. (2013) suggested 400,000 years of evolution along a branch is a reasonable

approximation for a coalescent unit in vertebrates. If we use this approximation, then branch lengths of 0.01 CUs are equivalent to just 4,000 years and highlight the challenging conditions that we modeled.

Even in these extreme anomaly zone situations, ASTRAL_BP, ASTRID_BP, and SDPquartets consistently recovered the correct species tree. While ASTRAL_BP and SDPquartets are statistically consistent for the assumptions that we have made here (see Methods and Appendix), the good performance of ASTRID_BP suggests that it too may be statistically consistent under these conditions; future research should investigate this further. Dollo parsimony also performed well and only failed to recover the correct species tree in 3 of 25 simulations for the Palaeognathae anomaly zone tree (Fig. 3C). MDC, polymorphism parsimony, and unordered parsimony generally performed well in the simplest anomaly zone situation with a pectinate tree and four ingroup taxa (Fig. 1A). However, these three methods all failed in more complex anomaly zone situations with greater than four ingroup taxa (Fig. 3B-C). First, these methods failed to recover the pectinate species tree for five ingroup taxa in the anomaly zone, and as expected, more symmetrical species trees were recovered that are consistent with the occurrence of anomalous gene trees that are also more symmetrical (Table 7 in Rosenberg and Tao 2008). Second, these methods recovered an incorrect position for rheas in the palaeognath simulations. Finally, Camin-Sokal failed to recover the correct topology in all cases for species trees with anomaly zone situations, and only recovered the correct topology in 1 of 25 simulations for the 26-taxa data set (Fig. 3).

Based on these experimental results, and on theoretical considerations pertaining to statistical consistency (for ASTRAL_BP and SDPquartets), we suggest that ASTRAL_BP, ASTRID_BP, and SDPquartets are the most appropriate of the tested methods for inferring species trees with retroelements. We expect that these methods should also perform well with other low-homoplasy absence/presence characters such as NUMTs and

large indels that, along with retroelements, are becoming increasingly easy to mine from genomic sequences (Schull et al. 2019; Churakov et al. 2020).

Branch length bias and the anomaly zone. In our simulation study, the mean branch length distortion on ASTRAL_BP trees based on retroelements was minimal for very short branches (<0.1 CUs) in the anomaly zone (Figure 4B,E), but became progressively larger as species tree branches lengths increased from 0.2 CUs (+12%) to 1.5 CUs (+38%). Distortion levels off at ∼37–39% for species tree branches in the range of ∼1.5 to ∼4.0 CUs (Fig. 4C,F).

The recovery of accurate branch lengths for short branches is predicted based on our theoretical results and suggests that ASTRAL_BP branch lengths without correction can be used to assess claims of empirical anomaly zones that are inferred from sequence-based gene trees (provided there are a sufficiently large number of retroelement insertions available to estimate the probabilities of quartets around the branch in question with high accuracy). By contrast, simulations show that gene tree reconstruction error in sequence-based analyses can result in branch length estimates on ASTRAL species trees that are too short by almost an order of magnitude when gene tree reconstruction error is high (Sayyari and Mirarab 2016).

A case in point is the species tree for palaeognath birds that Cloutier et al. (2019) claimed is in the anomaly zone based on both MP-EST and ASTRAL analyses, although many of gene trees were arbitrarily resolved and therefore inaccurately reconstructed (Springer et al. 2020). Gene tree reconstruction error is prevalent among phylogenomic studies and can occur because of long-branch misplacement, missing data, model misspecification, homology errors, arbitrary resolution of polytomies by programs such as RAxML and PhyML, and other causes (Gatesy and Springer 2014; Springer M. S. 2014; Springer and Gatesy 2016). Notably, Cloutier et al.'s (2019) ASTRAL species tree based

on sequence-based gene trees had much shorter branch lengths than Springer et al.'s (2020) ASTRAL_BP species tree based on low-homoplasy retroelement insertions. The simulation results presented here provides additional support for the conclusion that the palaeognath species tree is not in the anomaly zone.

While short branches are typically of the most interest, longer branch lengths can be corrected using the technique proposed here, although recall that this technique assumes that the rate of retroelement insertions per generation is constant across the tree. By contrast, the result for short branches not needing the correction does not make this assumption, as it is derived using the small angle approximation. Overall, our results suggest that ASTRAL_BP analysis of retroelement insertions is an effective approach for evaluating whether a species tree is in the anomaly zone.

Future directions. All of the analyses that we performed are based on large simulated data sets with 100,000 informative retroelements. These data sets are much larger than most published data sets for empirical retroelements (e.g., Doronina et al. 2017; Cloutier et al. 2019). We chose to simulate large data sets because the major motivation of our study was to determine if different species tree methods that have been applied to retroelement data sets converge on the correct or incorrect species tree, and in the case of ASTRAL_BP if branch lengths in the anomaly zone are upwardly or downwardly biased. It remains for future studies to determine how many retroelements are required to estimate a correct species tree with high confidence for species trees with different numbers of taxa and varying branch lengths. Our computational pipeline based on the *ms* program should be useful for exploring this question experimentally.

It will also be important to use simulations to compare species trees that are inferred from sequence-based gene trees versus retroelement insertions. These simulations should be performed at various phylogenetic depths and with difficult anomaly zone branch

lengths. We expect that retroelements will fare well in such simulations, especially at deep divergences where the estimation of gene trees can be challenging, because these characters better satisfy assumptions of the MSC. Specifically, retroelements avoid or reduce problems with small c-gene size, recombination, and selection that impact the accurate reconstruction of sequence-based gene trees (Springer et al. 2020). Unlike DNA sequences, which show increased homoplasy with depth, retroelements are low-homoplasy markers in both shallow and deep phylogenetic settings when accurately coded. Retroelement insertions become more difficult to characterize at deep divergences because indels and other mutations can erase or obscure their history, but remain useful for phylogenetic problems that are at least as old as the radiations of placental mammals, crocodylians, and birds that each extend to the Cretaceous (Nishihara et al. 2009; Haddrath and Baker 2012; Suh et al. 2015a,b; Doronina et al. 2017). We emphasize that these methods should only be applied to empirical data sets with well-vetted coding of retroelements (Doronina et al. 2019).

A critical issue concerns the number of retroelements that are available for estimating species trees. Published data sets for mammalian retroelement insertions range from those with <100 retroelements (e.g., placental root, [Nishihara et al. 2009]) to 91,859 for eight species of baleen whales (Lammers et al. 2019). In the latter case, 24,598 of these insertions are phylogenetically informative and occur in two to six of the balaenopteroid species. For protein-coding genes, the number of available loci is relatively fixed whether a data set includes genomes from five mammalian species or 500, because the majority of protein-coding genes are shared among these taxa. In humans, a recent estimate for the total number of protein-coding genes is 19,116 (Piovesan et al. 2019). By contrast, retroelement insertions are segregating sites as are single nucleotide mutations, albeit without the attendant homoplasy in the latter, and retroelement data sets are expected to increase in size as more taxa are added to a data set. For a taxonomically diverse genomic data set with more than 200 mammal species (e.g., Genereux et al. 2020), we are optimistic

that it will soon be possible to extract hundreds of thousands or even millions of informative retroelements as improved methods become available for efficiently extracting and applying quality-control filtering steps to assemble these data sets (Churakov et al. 2020). Indeed, the number of informative markers will grow even larger if such data sets also include NUMTs and large indels (Schull et al. 2019). Combining these low-homoplasy markers with sequence-based gene trees is a valuable direction of future research for mitigating the impact of gene tree estimation error and maximizing the amount of high quality data available for species tree estimation (Houde et al. 2019).

Because accurately-coded retroelement characters are unlikely to be impacted by homoplasy, such data can be modeled under the MSC, with insertions following an infinite sites model. ASTRAL_BP and SDPquartets are provably statistically consistent under this model and perform well in simulations in the anomaly-zone, at least when the number of retroelement insertions is quite large. Future phylogenomic studies should leverage the power of retroelements and other low-homoplasy presence/absence characters that can now be analyzed with ILS-aware methods to resolve some of the most challenging phylogenetic problems that remain.

Acknowledgements

We thank Rick Baker for technical advice and Siavash Mirarab and Arun Durvasula for discussions.

Funding

This research was funded by the U.S. National Science Foundation (Grant No. DEB-1457735), the NSF Graduate Research Fellowship (Grant No. DGE-1144245), and the Ira and Debra Cohen Graduate Fellowship in Computer Science.

Supplementary Material

Data available from the Dryad Digital Repository:

http://dx.doi.org/10.5061/dryad.[NNNN]

Appendix

Quartet Probabilities

Kuritzin et al. (2016) and Doronina et al. (2017) model retroelement insertions under the MSC model with insertions following an infinite sites neutral mutation model. They use $\omega_{i,j}$ to represent the scenario where a retrolement insertion in the orthologous locus is absent (0) from lineages A_i and A_j and present (1) in lineages A_k and A_l, so

- $\omega_{1,2} = 0011$ and $\omega_{3,4} = 1100$ both display quartet $A_1A_2|A_3A_4$,
- $\omega_{1,3} = 0101$ and $\omega_{2,4} = 1010$ both display quartet: $A_1A_3|X_2A_4$, and
- $\omega_{1,4} = 0110$ and $\omega_{2,3} = 1001$ both display quartet: $A_1A_4|A_2A_3$.

Doronina et al. (2017) derive an approximation for the expected number $a_{i,j}$ of retroelement insertions with property $\omega_{i,j}$ for three different phylogenetic networks on four species based on the diffusion approximation of the Wright-Fisher coalescent model (Fisher 1922; Wright 1931) and the neutral mutation model (Kimura 1955a,b). Their "Hybridization model 1" is equivalent to

- a pectinate species tree $(((A_4, A_3), A_2), A_1)$ when $\gamma_1 = 0$ and $\gamma_2 = 1$ and
- a balanced model species tree $((A_4,A_3),(A_2, A_1))$ when $\gamma_1 = 1$ and $\gamma_2 = 0$

(see Figure 6 in Doronina et al. (2017) Supplemental Materials S1). We simplify the equations that they derived for "Hybridization model 1" in order to compute the probability of observing retroelement insertions corresponding to each quartet $A_iA_j|A_kA_l$:

$$p^R_{i,j|k,l} = \frac{a_{i,j} + a_{k,l}}{a_{i,j} + a_{i,k} + a_{i,l} + a_{j,k} + a_{j,l} + a_{l,k}} \tag{6}$$

We then verify that $p^R_{1,2|3,4} > p^R_{1,3|2,4} = p^R_{1,4|2,3}$ for the pectinate and balanced model species trees with unrooted topology: $A_1A_2|A_3A_4$. This is summarized in the following theorem.

Theorem 2. Suppose that retroelement insertions are generated under the MSC with insertions following an infinite sites neutral model (as approximated by Doronina et al. 2017), with a constant rate of insertions per generation across the four taxa species tree. Then, the most probable quartet agrees with the unrooted species tree, and the two alternative quartets have equal probability.

Proof. For the **pectinate model species tree**, let τ_3 be the length (in CUs) of the internal branch separating A_2, A_3, A_4 from A_1, and let τ_2 be the length (in CUs) of the internal branch separating A_3, A_4 from A_1, A_2. Let n_i be the expected number of new retroelement insertions per generation on the branch with length τ_i or on the above the root population when $i = 0$ (note that n_i is the probability of a new retroelement insertion occurring in an individual times the effective population size). Simplifying the equations from Doronina et al. (2017), the expected number of retroelement insertions that display the quartet topology that agrees with the unrooted model species tree (i.e. $A_1, A_2|A_3, A_4$) is

$$\begin{aligned} a_{1,2} + a_{3,4} = {} & n_0\left(e^{-\tau_2} - \frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\ & + n_2\left(1 - e^{-\tau_2} - \frac{2}{3}e^{-\tau_3} + \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\ & + n_3\left(\tau_3 - 1 + e^{-\tau_3}\right) \end{aligned} \tag{7}$$

and the expected number of retroelement insertions that display one of the two alternative quartets (i.e. $A_1, A_3|A_2, A_4$ and $A_1, A_4|A_2, A_3$) is

$$a_{1,3} + a_{2,4} = a_{1,4} + a_{2,3} = n_0\left(\frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) + n_2\left(\frac{1}{3}e^{-\tau_3} - \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \tag{8}$$

(see Section 1 of the Supplementary Text for details). Now we verify that

$$\begin{aligned}
(p^R_{1,2} + p^R_{3,4}) - (p^R_{1,3} + p^R_{2,4}) &> 0 \\
(a_{1,2} + a_{3,4}) - (a_{1,3} + a_{2,4}) &> 0 \\
n_0\left(e^{-\tau_2} - \frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) - n_0\left(\frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) & \\
+n_2\left(1 - e^{-\tau_2} - \frac{2}{3}e^{-\tau_3} + \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) - n_2\left(\frac{1}{3}e^{-\tau_3} - \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) & \\
+n_3\left(\tau_3 - 1 + e^{-\tau_3}\right) &> 0 \\
n_0\left(e^{-\tau_2} - e^{-\tau_2-\tau_3} + \frac{1}{3}e^{-3\tau_2-\tau_3}\right) + n_2\left(1 - e^{-\tau_2} - e^{-\tau_3} + e^{-\tau_2-\tau_3}\right) + n_3\left(\tau_3 - 1 + e^{-\tau_3}\right) &> 0.
\end{aligned}$$

This inequality holds for $n_0, n_1, n_2, \tau_1, \tau_2 > 0$, because the first term is positive by

$$\begin{aligned}
1 &> e^{-\tau_2} \\
1 - e^{-\tau_2} &> 0 \\
(1 - e^{-\tau_2}) \times e^{-\tau_3} &> 0 \times e^{-\tau_3} \\
e^{-\tau_2} - e^{-\tau_2-\tau_3} &> 0,
\end{aligned}$$

the second term is positive by $1 - e^{-\tau_2} - e^{-\tau_3} + e^{-\tau_2-\tau_3} = (1 - e^{-\tau_2})(1 - e^{-\tau_3})$ and the third term is positive by $1 - \tau_2 < e^{-\tau_2}$ (Bernoulli's inequality). For the pectinate model species tree, the most probable quartet on species A_1, A_2, A_3, and A_4 agrees with the species tree

on species A_1, A_2, A_3, and A_4, and the two alternative quartet trees have equal probability.

For the **balanced model species tree**, let τ_1 be the length (in CUs) of the internal branch above A_1, A_2, and let τ_3 be the length (in CUs) of the internal branch above A_3, A_4. Let n_i be the expected number of new retroelement insertions per generation corresponding to the same branch as length τ_i or the above the root population when $i = 0$. Simplifying the equations from Doronina et al. (2017), the expected number of retroelement insertions that display the quartet topology that agrees with the species tree (i.e. $A_1, A_2|A_3, A_4$) is

$$a_{1,2} + a_{3,4} = n_0\left(2 - e^{-\tau_1} - e^{-\tau_3} + \frac{1}{3}e^{-\tau_1-\tau_3}\right) + n_1\left(\tau_1 - 1 + e^{-\tau_1}\right) + n_3\left(\tau_3 - 1 + e^{-\tau_3}\right), \quad (9)$$

and the expected number of retroelement insertions that display one of the two alternative quartets (i.e. $A_1, A_3|A_2, A_3$ and $A_1, A_4|A_2, A_3$) is

$$a_{1,3} + a_{2,4} = a_{1,4} + a_{2,3} = n_0\frac{1}{3}e^{-\tau_1-\tau_3}. \quad (10)$$

(see Section 2 of the Supplementary Text for details). Now we verify that

$$\begin{aligned}
(p^R_{1,2} + p^R_{3,4}) - (p^R_{1,3} + p^R_{2,4}) &> 0 \\
(a_{1,2} + a_{3,4}) - (a_{1,3} + a_{2,4}) &> 0 \\
n_0\left(2 - e^{-\tau_1} - e^{-\tau_3} + \frac{1}{3}e^{-\tau_1-\tau_3}\right) - n_0\frac{1}{3}e^{-\tau_1-\tau_3} + n_1\left(\tau_1 - 1 + e^{-\tau_1}\right) + n_2\left(\tau_3 - 1 + e^{-\tau_3}\right) &> 0 \\
n_0\left(2 - e^{-\tau_1} - e^{-\tau_3}\right) + n_1\left(\tau_1 - 1 + e^{-\tau_1}\right) + n_2\left(\tau_3 - 1 + e^{-\tau_3}\right) &> 0
\end{aligned}$$

This inequality holds for $n_0, n_1, n_2, \tau_1, \tau_3 > 0$, because the first term is positive as $1 > e^{-\tau_i}$ and the second and third terms are positive as $e^{-\tau_i} > 1 - \tau_i$ (Bernoulli's inequality). For the balanced model species tree, the most probable quartet on species A_1, A_2, A_3, and A_4 agrees with the species tree on species A_1, A_2, A_3, and A_4, and the alternative two quartet

trees have equal probability. □

The theorem above enables proofs of statistical consistency for two different quartet-based methods: SDPquartets (Theorem 1) and ASTRAL_BP (below).

Theorem 3. Under the conditions of Theorem 2, ASTRAL_BP is statistical consistent.

Proof. Let T^* be the true species tree on taxon set S, and let $w_{\mathcal{R}}(T|_{S_i})$ denote the number of retroelement insertions in $\mathcal{R}$ that displays the same quartet topology as tree T restricted to a subset S_i of four taxa, with $1 \leq i \leq m = \binom{|S|}{4}$. Note that $w_{\mathcal{R}}(T|_{S_i})$ can be 0 either because the retroelement insertion displays a different topology than T or because the retroelement insertion does not display any of the three possible quartet topologies on S_i (e.g. the insertion 11000 for taxon set $\{A, B, C, D, E\}$, represented as $((A, B), C, D, E)$, does not display a quartet for taxon subset $\{A, C, D, E\}$).

Let n_i be the number of retroelement insertions in $\mathcal{R}$ that displays any of the three possible quartet topologies on taxon subset S_i. For all $i \in \{1, 2, \ldots, m\}$, as the total number of retroelement insertions goes to infinity, n_i also goes to infinity (i.e. n_i is not bounded). Then, because the most probable quartet agrees with the true species tree T^* and the two alternative quartets have lesser probability (Theorem 2), for any possible tree topology T on taxon set S and for all $i \in \{1, 2, \ldots, m\}$, $w_{\mathcal{R}}(T|_{S_i}) \leq w_{\mathcal{R}}(T^*|_{S_i})$ with probability going to one, as the number of retroelement insertions goes to infinity. It follows that the true species tree T^* is the unique optimal solution to maximum quartet support supertree (MQSS) problem with high probability (recall that the MQSS problem is to find T such that $\sum_{i=1}^{m} w_{\mathcal{R}}(T_{s_i})$ is maximized).

The MQSS problem is NP-hard (Jiang et al. 2001; Lafond and Scornavacca 2019); however, when the solution space is constrained by a set Σ of bipartitions, it can be solved in polynomial time (Bryant and Steel 2001; Mirarab et al. 2014). ASTRAL implements an exact algorithm for solving the bipartition-constrained version of the MQSS problem, and

by default, every bipartition in the input (in this case $\mathcal{R}$) is added to the constraint set Σ.

Because every retroelement insertion (0/1) pattern occurs under the MSC + neutral infinite sites model with non-zero probability, the probability that every bipartition in T^* is represented by a retroelement insertion in $\mathcal{R}$ goes to 1, as the number of retroelement insertions goes to infinity; therefore, ASTRAL given $\mathcal{R}$ returns the true species tree with high probability. □

Quartet probabilities for short internal branches

We now consider what happens **both** internal branches are short enough so that the small angle approximation can be applied.

For the **pectinate model species tree**, suppose both τ_2 and τ_3 are sufficiently short, then we can apply the small angle approximation $e^{-\tau_i} \approx 1 - \tau_i$ and drop the higher order terms (e.g. $\tau_2\tau_3$ and τ_2^2). From Equation 7, the expected number of retroelement

insertions displaying the quartet that agrees with the species tree is

$$\begin{aligned}
a_{1,2}+a_{3,4} &= n_0\left(e^{-\tau_2}-\frac{1}{2}e^{-\tau_2-\tau_3}-\frac{1}{6}e^{-3\tau_2-\tau_3}\right)\\
&\quad+n_2\left(1-e^{-\tau_2}-\frac{2}{3}e^{-\tau_3}+\frac{1}{2}e^{-\tau_2-\tau_3}+\frac{1}{6}e^{-3\tau_2-\tau_3}\right)\\
&\quad+n_3\left(\tau_3-1+e^{-\tau_3}\right)\\
&\approx n_0\left((1-\tau_2)-\frac{1}{2}(1-\tau_2)(1-\tau_3)-\frac{1}{6}(1-\tau_2)^3(1-\tau_3)\right)\\
&\quad+n_2\left(1-(1-\tau_2)-\frac{2}{3}(1-\tau_3)+\frac{1}{2}(1-\tau_2)(1-\tau_3)+\frac{1}{6}(1-\tau_2)^3(1-\tau_3)\right)\\
&\quad+n_3\left(\tau_3-1+(1-\tau_3)\right)\\
&\approx n_0\left((1-\tau_2)-\frac{1}{2}(1-\tau_2-\tau_3)-\frac{1}{6}(1-3\tau_2-\tau_3)\right)\\
&\quad+n_2\left(1-(1-\tau_2)-\frac{2}{3}(1-\tau_3)+\frac{1}{2}(1-\tau_2-\tau_3)+\frac{1}{6}(1-3\tau_2-\tau_3)\right)\\
&= n_0\left(\frac{1}{3}+\frac{2}{3}\tau_3\right)
\end{aligned}$$

and from Equation 8, the expected number of retroelement insertions displaying the alternative quartets is

$$a_{1,3}+a_{2,4}=a_{1,4}+a_{2,3}\approx n_0\left(\frac{1}{3}-\frac{1}{3}\tau_3\right)$$

(see Section 1 of the Supplementary Text for details). Repeating this approximation for the **balanced model species tree** using Equations 9 and 10 gives

$$a_{1,2}+a_{3,4}\approx n_0\left(\frac{1}{3}+\frac{2}{3}(\tau_1+\tau_3)\right)\text{ and }a_{1,3}+a_{2,4}=a_{1,4}+a_{2,3}\approx n_0\left(\frac{1}{3}-\frac{1}{3}(\tau_1+\tau_3)\right)$$

(see Section 2 of the Supplementary Text for details). Plugging these formulas into

Equation 6 gives

$$p^R_{1,2|3,4} \approx \frac{1}{3} + \frac{2}{3}\tau \quad \text{and} \quad p^R_{1,3|2,4} = p^R_{1,4|2,3} \approx \frac{1}{3} - \frac{1}{3}\tau \tag{11}$$

where τ is the length of the internal branch that induces quartet $A_1, A_2|A_3, A_4$. For the pectinate model species tree, $\tau = \tau_3$ (but recall that τ_2 must also be short for the approximation to apply) and $\tau = \tau_1 + \tau_3$ for the balanced model species tree.

Quartet probabilities when the expected number of new retroelement insertions per generation is constant

We now consider what happens when the expected number of retroelement insertions per generation is constant across the species tree.

For the **pectinate model species tree**, we set $n_0 = n_2 = n_3$. From Equation 7, the expected number of retroelement insertions displaying the quartet that agrees with the species tree is

$$\begin{aligned}
a_{1,2} + a_{3,4} &= n_0\left(e^{-\tau_2} - \frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\
&\quad + n_2\left(1 - e^{-\tau_2} - \frac{2}{3}e^{-\tau_3} + \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\
&\quad + n_3\left(\tau_3 - 1 + e^{-\tau_3}\right) \\
&= n_3\tau_3 + n_3e^{-\tau_3} - n_2\frac{2}{3}e^{-\tau_3} + (n_2 - n_3) + (n_0 - n_2)\left(e^{-\tau_2} - \frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\
&= n_0\left(\tau_3 + \frac{1}{3}e^{-\tau_3}\right)
\end{aligned}$$

and from Equation 8, the expected number of retroelement insertions displaying the

alternative quartets is

$$\begin{aligned} a_{1,3} + a_{2,4} = a_{1,4} + a_{2,3} &= n_0\left(\frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) + n_2\left(\frac{1}{3}e^{-\tau_3} - \frac{1}{2}e^{-\tau_2-\tau_3} + \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\ &= n_2\frac{1}{3}e^{-\tau_3} + (n_0 - n_2)\left(\frac{1}{2}e^{-\tau_2-\tau_3} - \frac{1}{6}e^{-3\tau_2-\tau_3}\right) \\ &= n_0\frac{1}{3}e^{-\tau_3} \end{aligned}$$

(see Section 1 of the Supplementary Text for details). Repeating this simplification (i.e. $n_0 = n_1 = n_3$) for the **balanced model species tree** using Equations 9 and 10 gives

$$a_{1,2} + a_{3,4} = n_0\left((\tau_1 + \tau_3) + \frac{1}{3}e^{-(\tau_1+\tau_3)}\right) \text{ and } a_{1,3} + a_{2,4} = a_{1,4} + a_{2,3} = n_0\frac{1}{3}e^{-(\tau_1+\tau_3)}$$

(see Section 1 in the Supplementary Text for details). Plugging these formulas into Equation 6 gives

$$p^R_{1,2|3,4} = \frac{\frac{1}{3}e^{-\tau} + \tau}{e^{-\tau} + \tau} \quad \text{and} \quad p^R_{1,3|2,4} = p^R_{1,4|2,3} = \frac{\frac{1}{3}e^{-\tau}}{e^{-\tau} + \tau} \tag{12}$$

where τ is the length of the internal branch that induces quartet $A_1, A_2|A_3, A_4$. For the pectinate model species tree, $\tau = \tau_3$ and $\tau = \tau_1 + \tau_3$ for the balanced model species tree.

Maximum Likelihood Branch Length Estimation

We now show how to compute the maximum likelihood (ML) estimate of the lengths of the internal branches for retroelement data sets using the framework (and notation) proposed by Sayyari and Mirarab (2016).

Consider a species tree on four taxa with unrooted topology: $A_1, A_2|A_3, A_4$. Let θ_1 denote the probability of the dominant topology (i.e. $A_1, A_2|A_3, A_4$); similarly, let θ_2 and θ_3 denote the probabilities of the two alternative topologies (i.e. $A_1, A_3|A_2, A_4$ and

$A_1, A_4|A_2, A_3$). When retroelement insertions are generated under the MSC + infinite sites neutral mutation model with a constant rate of insertions per generation across the species tree, θ_1,θ_2, and θ_3 are functions of the true internal branch length τ^* (Equation 12). For n retroelement insertions generated under the model described above, we can count the number of retroelement insertions that display one of the three quartet topologies. We denote these counts as the vector $\bar{Z} = (Z_1, Z_2, Z_3)$ and assume that $\bar{Z}$ is generated from a multinomial distribution parameterized by $(\theta_1, \theta_2, \theta_3)$. In practice, $\bar{Z}$ is estimated from the retroelement insertion data, and these estimates are denoted $\bar{z} = (z_1, z_2, z_3)$.

Theorem 4. Suppose that n retroelement insertions are generated under the MSC with insertions following an infinite sites neutral model (as approximated by Doronina et al. (2017)), with a constant rate of insertions per generation across the four-taxon species tree. Let z_1 be the number of quartets associated with branch Q. Given that Q corresponds to the internal branch in the true four-taxon species tree and given the modeling of z_1 described above, the ML estimate of its length is $W\left[\frac{2}{3}(1-\frac{z_1}{n})^{-1}-1\right]$, where W is Lambert's function. This holds for $\frac{1}{3} \leq \frac{z_1}{n} < 1$. When $\frac{z_1}{n} = 1$, the ML estimate does not exist (note that we set the length equal to ∞ in this case), and when $\frac{z_1}{n} < \frac{1}{3}$, the branch Q cannot be in the true species tree (note that we set the branch length equal to 0 in this case).

Proof. Let $D \in [0, \infty)$ be a branch length. We model D as a random variable, so the ML estimate of the branch length is $\arg\max_{\tau \geq 0} P_{Z_1|D}(z_1|\tau; n)$, where $P_{Z_1|D}(z_1|\tau; n)$ is the likelihood of D. Given that Q induces the true quartet topology, by Lemma 1 in Sayyari and Mirarab (2016) and Equation 12,

$$P_{Z_1|D}(z_1|\tau; n) \propto \left(\frac{\frac{1}{3}e^{-\tau}+\tau}{e^{-\tau}+\tau}\right)^{z_1}\left(\frac{\frac{1}{3}e^{-\tau}}{e^{-\tau}+\tau}\right)^{n-z_1} \tag{13}$$

We now compute the log-likelihood function

$$\begin{aligned} L(\tau; z, n) &= z_1 \ln\left(\frac{\frac{1}{3}e^{-\tau}+\tau}{e^{-\tau}+\tau}\right) + (n-z_1)\ln\left(\frac{\frac{1}{3}e^{-\tau}}{e^{-\tau}+\tau}\right) \\ &= z_1 \ln\left(\frac{1}{3}e^{-\tau}+\tau\right) + (z_1-n)\tau - n\ln(e^{-\tau}+\tau) \end{aligned}$$

dropping the constant terms. To find the critical point, we take the first derivative of the log likelihood function

$$\frac{dL(\tau; z_1, n)}{d\tau} = z_1 \frac{1-\frac{1}{3}e^{-\tau}}{\frac{1}{3}e^{-\tau}+\tau} - n + z_1 - n\frac{1-e^{-\tau}}{e^{-\tau}+\tau}$$

and set it equal to 0. Therefore, the critical point is given by

$$\begin{aligned} z_1 \frac{1-\frac{1}{3}e^{-\tau}}{\frac{1}{3}e^{-\tau}+\tau} - n + z_1 - n\frac{1-e^{-\tau}}{e^{-\tau}+\tau} &= 0 \\ \frac{z_1}{n}\frac{1-\frac{1}{3}e^{-\tau}}{\frac{1}{3}e^{-\tau}+\tau} - 1 + \frac{z_1}{n} - \frac{1-e^{-\tau}}{e^{-\tau}+\tau} &= 0 \\ \frac{z_1}{n}\left(\frac{1-\frac{1}{3}e^{-\tau}}{\frac{1}{3}e^{-\tau}+\tau}+1\right) - \left(1+\frac{1-e^{-\tau}}{e^{-\tau}+\tau}\right) &= 0 \\ \frac{z_1}{n}\left(\frac{\frac{1}{3}e^{-\tau}+\tau+1-\frac{1}{3}e^{-\tau}}{\frac{1}{3}e^{-\tau}+\tau}\right) - \frac{e^{-\tau}+\tau+1-e^{-\tau}}{e^{-\tau}+\tau} &= 0 \\ \frac{\frac{1}{3}e^{-\tau}+\tau}{e^{-\tau}+\tau} &= \frac{z_1}{n} \qquad (14) \\ \tau e^{\tau} = \frac{(\frac{z_1}{n}-\frac{1}{3})}{(1-\frac{z_1}{n})} &= \frac{2}{3}(1-\frac{z_1}{n})^{-1} - 1 \\ \tau &= W\left[\frac{2}{3}(1-\frac{z_1}{n})^{-1} - 1\right] \qquad (15) \end{aligned}$$

where W is the Lambert's function. Note that $y = \frac{2}{3}(1-\frac{z_1}{n})^{-1} - 1$ is positive when $\frac{1}{3} < \frac{z_1}{n} < 1$ and is undefined at 1. By Proposition 5 in Borwein and Lindstrom (2016), $W[y]$ is positive for $y \in [0, \infty)$ and concave on $(-1/e, \infty)$ (also see `https://www.carma.newcastle.edu.au/resources/jon/WinOpt.pdf`). Therefore, the

critical point is a local maximum of the likelihood function of D when $\frac{1}{3} \leq \frac{z_1}{n} < 1$. When $\frac{z_1}{n} = 1$, the maximum likelihood estimate does not exist. □

When there are more than four taxa in the species tree, the situation is more complicated. Consider a branch Q that splits the leaf set into four disjoint sets A, B, C, D by the four branches that were incident to Q (but not their endpoints) and deleting branch Q including its endpoints. This implies that Q induces $m' = |A| \times |B| \times |C| \times |D|$ quartets; in this case we say that there are m' quartets "around" Q. If there are n retroelement insertions, then each of the m' quartets has its own values for n and z_1. For a quartet k ($1 \leq k \leq m$), let n^k denote the number of insertions (trials) that display any of the three possible quartet topologies, and let z_1^k denote the number of insertions (trials) that display the quartet that agrees with branch Q.

One possibility is to take just one of the m' quartets around branch Q and compute the maximum likelihood estimate of the branch length from z_1^k and n^k (where k is the index of the selected quartet). For example, we could choose k to maximize n^k for $1 \leq k \leq m'$. Alternatively, because Equation 14 above is equal to Equation 12, where $\frac{z_1}{n}$ is the frequentist estimate of the probability of the dominant quartet, we could utilize the m' quartets to get a better estimate by taking the average value

$$\frac{1}{m'} \sum_{k=1}^{m'} \frac{z_1^k}{n^k} \tag{16}$$

Sayyari and Mirarab (2016) provide an efficient algorithm for approximating this quantity ("q1" when running ASTRAL with option "-t 2"). We corrected the branch lengths by plugging "q1" into Equation 15 for $\frac{z_1}{n}$, and in our simulations, branch length estimation was accurate (Figure 4).

Local Posterior Probability

Sayyari and Mirarab (2016) also show how to compute the local posterior probability (local PP) for branch Q (Therem 1 in Sayyari and Mirarab 2016). In particular, the local PP calculation assumes that the gene trees are generated under the MSC from a model species tree generated under the Yule process with birth rate λ. Under this assumption, branch lengths in the species tree are exponentially distributed, and we take $f_D(\tau) = 2\lambda e^{2\lambda\tau}$ as the prior for branch lengths (Stadler and Steel 2012). By default, ASTRAL sets $\lambda = \frac{1}{2}$, which corresponds to a uniform prior on θ_j (Lemma 1 in Sayyari and Mirarab 2016).

We show that for retroelement insertions the prior on θ_j, denoted f_{θ_j}, is not uniform when the species tree is generated under a Yule process with birth rate $\lambda = \frac{1}{2}$. For $t = \frac{z_1}{n} \geq \frac{1}{3}$,

$$\begin{aligned} f_{\theta_j}(t) &= \frac{1}{3}\frac{1}{|\frac{d\theta_j}{dx}|} f_D(x)\Big|_{x=W[y]} \\ &= \frac{1}{3}\left(\frac{(e^{-x}+x)^2}{\frac{2}{3}e^{-x}(x+1)}\right) \times 2\lambda e^{-2\lambda x}\Big|_{x=W[y]} = \lambda\left(\frac{(e^{-x}+x)^2}{(x+1)}\right)e^{-\lambda x}\Big|_{x=W[y]} \\ &= \lambda\left(\frac{(\frac{W[y]}{y}+W[y])^2}{(W[y]+1)}\right)\left(\frac{W[y]}{y}\right)^{\lambda} = \lambda\left(\frac{(W[y](\frac{1+y}{y}))^2}{(W[y]+1)}\right)\left(\frac{W[y]^\lambda}{y^\lambda}\right) \\ &= \lambda\left(\frac{(1+y)^2}{(W[y]+1)}\right)\left(\frac{W[y]}{y}\right)^{2+\lambda} \end{aligned} \tag{17}$$

where $y = \frac{2}{3}(1-\frac{z_1}{n})^{-1} - 1$. Recall that $e^{-kW[y]} = (\frac{W[y]}{y})^k$ (shown as part of Proposition 6 in Borwein and Lindstrom 2016) and Lambert's W is positive on $[0, \infty)$ (Proposition 5 in Borwein and Lindstrom 2016). The prior on θ_j is not uniform on $[\frac{1}{3}, 1]$ when $\lambda = \frac{1}{2}$ (Supplementary Figure 1); furthermore, this prior makes it difficult to integrate

$$\int_{\frac{1}{3}}^{1} t^{z_j}\left(\frac{1-t}{2}\right)^{n-z_j} f_{\theta_j}(t)\, dt$$

when computing local PP.

It is reasonable to have a uniform prior on θ_j in the absence of other information,

and therefore, one could justify utilizing the branch support returned by ASTRAL for retroelement data sets for four species. However, we cannot keep the interpretation that the prior on branch lengths comes from species trees generated under the Yule process with birth rate λ.

Another issue arises when the number of species is greater than four, because as discussed previously, z_1, z_2, z_3, and n can take on different values for each of the m' quartets induced by a branch. ASTRAL uses the m' quartets to estimate each z_i, taking n to be the effective number (EN) of gene trees (in this case retroelement insertions) for the branch (this is "EN" when running ASTRAL with option "-t 2"). Computing EN for each branch is important when there are missing data (i.e. gene trees can be missing taxa) or when gene trees are unresolved. The latter is always the case for retroelement data sets, as insertions are represented as unresolved "gene trees" with a single bipartition.

Estimates of the local PP should be interpreted cautiously when EN is low, and this can be an issue when analyzing retroelement insertion data sets. For example, when analyzing the 4,301 retroelement insertions from Cloutier et al. (2019), the branch that separates Chilean tinamou (CT) and Greater Rhea (GR) from Emu (E) and Great spotted kiwi (GK) (i.e. the branch with length 0.0532 in Fig. 1C) had $EN = 26.23$, $z_1 = 0.50 \times 26.23 = 13.12$, $z_2 = 0.24 \times 26.23 = 6.30$ and $z_3 = 0.26 \times 26.23 = 6.81$ (Table 1). These estimates by ASTRAL are consistent with the analysis by Springer et al. (2020) that showed only 28 retroelement insertions displayed quartets on these four taxa: 15 insertions supported $CT, GR|GK, E$, 6 insertions supported $CT, E|GK, GR$, and 7 insertions supported $CT, GK|GR, E$ (Table 3 in Springer et al. 2020). Therefore, we recommend running ASTRAL with the "-t 2" option and then explicitly checking EN of each branch when analyzing retroelement data sets. Investigation of whether the quantities (e.g. z_1, z_2, z_3) computed by ASTRAL are good approximations for retroelement insertion data sets is a valuable direction for future research.

752 *

753 References

754 Allman, E. S., J. H. Degnan, and J. A. Rhodes. 2011. Identifying the rooted species tree
755 from the distribution of unrooted gene trees under the coalescent. J. Math. Biol.
756 62:833–862.

757 Avise, J. C. and T. J. Robinson. 2008. Hemiplasy: a new term in the lexicon of
758 phylogenetics. Syst. Biol. 57:503–507.

759 Borwein, J. M. and S. B. Lindstrom. 2016. Meetings with Lambert W and other special
760 functions in optimization and analysis. Pure and Applied Functional Analysis 1:361–396.

761 Bryant, D. and M. Steel. 2001. Constructing optimal trees from quartets. Journal of
762 Algorithms 38:237–259.

763 Chifman, J. and L. Kubatko. 2015. Identifiability of the unrooted species tree topology
764 under the coalescent model with time-reversible substitution processes, site-specific rate
765 variation, and invariable sites. J. of Theoret. Biol. 374:35–47.

766 Chuong, E. B., N. C. Elde, and C. Feschotte. 2017. Regulatory activities of transposable
767 elements: from conflicts to benefits. Nat. Rev. Genet. 18:71–86.

768 Churakov, G., F. Zhang, N. Grundmann, W. Makalowski, A. Noll, L. Doronina, and
769 J. Schmitz. 2020. The multi-comparative 2-n-way genome suite. Genome Res. .

770 Cloutier, A., T. B. Sackton, P. Grayson, M. Clamp, A. J. Baker, and S. V. Edwards. 2019.
771 Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the
772 presence of an empirical anomaly zone. Syst. Biol. 68:937–955.

773 Degnan, J. H. 2013. Anomalous unrooted gene trees. Systematic Biology 62:574–590.

774 Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most
775 likely gene trees. PLOS Genetics 2:1–7.

776 Degnan, J. H. and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference
777 and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

778 Doronina, L., G. Churakov, A. Kuritzin, J. Shi, R. Baertsch, H. Clawson, and J. Schmitz.
779 2017. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome Res.
780 27:997–1003.

781 Doronina, L., G. Churakov, J. Shi, J. Brosius, R. Baertsch, H. Clawson, and J. Schmitz.
782 2015. Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria,
783 Carnivora). Mol. Biol. Evol. 32:3194–3204.

784 Doronina, L., O. Reising, H. Clawson, D. A. Ray, and J. Schmitz. 2019. True homoplasy of
785 retrotransposon insertions in Primates. Syst. Biol. 68:482–493.

786 Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics
787 5:164–166.

788 Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Inc.

789 Fisher, R. A. 1922. On the dominance ratio. Proc. Roy. Soc. B 42:321–341.

790 Gatesy, J., J. H. Geisler, J. Chang, C. Buell, A. Berta, R. W. Meredith, M. S. Springer,
791 and M. R. McGowen. 2013. A phylogenetic blueprint for a modern whale. Mol.
792 Phylogenet. Evol. 66:479–506.

793 Gatesy, J., R. W. Meredith, J. E. Janecka, M. P. Simmons, W. J. Murphy, and M. S.
794 Springer. 2017. Resolution of a concatenation/coalescence kerfuffle: partitioned
795 coalescence support and a robust family-level tree for Mammalia. Cladistics 33:295–332.

796 Gatesy, J., D. B. Sloan, J. M. Warren, R. H. Baker, M. P. Simmons, and M. S. Springer.
797 2019. Partitioned coalescence support reveals biases in species-tree methods and detects
798 gene trees that determine phylogenomic conflicts. Mol. Phylogenet. Evol. 139:106539.

799 Gatesy, J. and M. S. Springer. 2014. Phylogenetic analysis at deep timescales: Unreliable
800 gene trees, bypassed hidden support, and the coalescence/ concatalescence conundrum.
801 Mol. Phylogenet. Evol. 80:231–266.

802 Genereux, D. P., A. Serres, J. Armstrong, J. Johnson, V. D. Marinescu, E. Murén,
803 D. Juan, G. Bejerano, N. R. Casewell, L. G. Chemnick, J. Damas, F. Di Palma,
804 M. Diekhans, I. Fiddes, M. Garber, V. N. Gladyshev, L. Goodman, W. Haerty, M. L.
805 Houck, R. Hubley, T. Kivioja, K.-P. Koepfli, L. F. K. Kuderna, E. S. Lander, J. R. W.
806 Meadows, W. J. Murphy, W. Nash, H. J. Noh, M. Nweeia, A. R. Pfenning, K. S. Pollard,
807 D. Ray, B. Shapiro, A. Smit, M. S. Springer, C. C. Steiner, R. Swofford, J. Taipale, E. C.
808 Teeling, J. Turner-Maier, J. Alfoldi, B. Birren, O. A. Ryder, H. Lewin, B. Paten,
809 T. Marques-Bonet, K. Lindblad-Toh, and K. E. K. 2020. A comparative genomics
810 multitool for scientific discovery and conservation. Nature In press.

811 Haddrath, O. and A. J. Baker. 2012. Multiple nuclear genes and retroposons support
812 vicariance and dispersal of the palaeognaths, and an Early Cretaceous origin of modern
813 birds. Proc. R. Soc. B 279:4617–4625.

814 He, C., D. Liang, and P. Zhang. 2020. Asymmetric distribution of gene trees can arise under
815 purifying selection if differences in population size exist. Mol. Biol. Evol. 37:881–892.

816 Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus
817 data. Mol. Biol. Evol. 27:570–580.

818 Hendy, M. D. and D. Penny. 1982. Branch and bound algorithms to determine minimal
819 evolutionary trees. Mathematical Biosciences 60:133–142.

820 Hosner, P. A., B. C. Faircloth, T. C. Glenn, E. L. Braun, and R. T. Kimball. 2016.
821 Avoiding missing data biases in phylogenomic inference: An empirical study in the
822 landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.

823 Houde, P., E. L. Braun, N. Narula, U. Minjares, and S. Mirarab. 2019. Phylogenetic signal
824 of indels and the neoavian radiation. Diversity 11:108.

825 Huang, H., Q. He, L. S. Kubatko, and L. L. Knowles. 2010. Sources of error inherent in
826 species-tree estimation: impact of mutational and coalescent effects on accuracy and
827 implications for choosing among different methods. Syst. Biol. 59:573–583.

828 Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic
829 variation. Bioinformatics 18:337–338.

830 Islam, M., K. Sarker, T. Das, R. Reaz, and M. S. Bayzid. 2020. STELAR: A statistically
831 consistent coalescent-based species tree estimation method by maximizing triplet
832 consistency. BMC Genomics 21:136.

833 Jiang, T., P. Kearney, and M. Li. 2001. A polynomial time approximation scheme for
834 inferring evolutionary trees from quartet topologies and its application. SIAM Journal on
835 Computing 30:1942–1961.

836 Kimura, M. 1955a. Solution of a process of random genetic drift with a continuous model.
837 Proc. Natl. Acad. Sci. USA 41:144–150.

838 Kimura, M. 1955b. Stochastic processes and distribution of gene frequencies under natural
839 selection. Cold Spring Harb Symp Quant Biol 20:33–53.

840 Kuritzin, A., T. Kischka, J. Schmitz, and G. Churakov. 2016. Incomplete lineage sorting
841 and hybridization statistics for large-scale retroposon insertion data. PLOS
842 Computational Biology 12:1–20.

Lafond, M. and C. Scornavacca. 2019. On the weighted quartet consensus problem. Theoretical Computer Science 769:1–17.

Lammers, F., M. Blumer, C. Ruckle, and M. A. Nilsson. 2019. Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation. Mobile DNA 10.

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

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

Mendes, F. K. and M. W. Hahn. 2017. Why concatenation fails near the anomaly zone. Systematic Biology 67:158–169.

Meredith, R. W., J. E. Janecka, J. Gatesy, O. A. Ryder, C. A. Fisher, E. C. Teeling, A. Goodbla, E. Eizirik, T. L. L. Simão, T. Stadler, D. L. Rabosky, R. L. Honeycutt, J. J. Flynn, C. M. Ingram, C. Steiner, T. L. Williams, T. J. Robinson, A. Burk-Herrick, M. Westerman, N. A. Ayoub, M. S. Springer, and W. J. Murphy. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Mirarab, S. and T. Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Nikaido, M., A. P. Rooney, and N. Okada. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements:

hippopotamuses are the closest extant relatives of whales. Proc. Natl. Acad. Sci. USA 96:10261–10266.

Nilsson, M. A., G. Churakov, M. Sommer, N. Van Tran, A. Zemann, J. Brosius, and J. Schmitz. 2010. Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol. 8:e1000436.

Nishihara, H., S. Maruyama, and N. Okada. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc. Natl. Acad. Sci. USA 106:5235–5240.

Nute, M., J. Chou, E. K. Molloy, and T. Warnow. 2018. The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics 19:286.

Oliveros, C. H., D. J. Field, D. T. Ksepka, F. K. Barker, A. Aleixo, M. J. Andersen, P. Alström, B. W. Benz, E. L. Braun, M. J. Braun, G. A. Bravo, R. T. Brumfield, R. T. Chesser, S. Claramunt, J. Cracraft, A. M. Cuervo, E. P. Derryberry, T. C. Glenn, M. G. Harvey, P. A. Hosner, L. Joseph, R. T. Kimball, A. L. Mack, C. M. Miskelly, A. T. Peterson, M. B. Robbins, F. H. Sheldon, L. F. Silveira, B. T. Smith, N. D. White, R. G. Moyle, and B. C. Faircloth. 2019. Earth history and the passerine superradiation. Proc. Natl. Acad. Sci. USA 116:7916–7925.

Patel, S., R. T. Kimball, and E. L. Braun. 2013. Error in phylogenetic estimation for bushes in the tree of life. J. Phylogenet. Evol. Biol. 1:110.

Piovesan, A., F. Antonaros, L. Vitale, P. Strippoli, M. C. Pelleri, and M. Caracausi. 2019. Human protein-coding genes and gene feature statistics in 2019. BMC Res. Notes 12:315.

887 Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees.
888 Molecular Phylogenetics and Evolution 1:53–58.

889 Ray, D. A., J. Xing, A. H. Salem, and M. A. Batzer. 2006. SINEs of a nearly perfect
890 character. Syst Biol. 55:928–935.

891 Roch, S., M. Nute, and T. Warnow. 2019. Long-branch attraction in species tree
892 estimation: inconsistency of partitioned likelihood and topology-based summary
893 methods. Syst. Biol. 68:281–297.

894 Roch, S. and M. Steel. 2015. Likelihood-based tree reconstruction on a concatenation of
895 aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol.
896 100:56–62.

897 Rosenberg, N. A. 2013. Discordance of species trees with their most likely gene trees: a
898 unifying principle. Mol. Biol. Evol. 30:2709–2713.

899 Rosenberg, N. A. and R. Tao. 2008. Discordance of species trees with their most likely gene
900 trees: the case of five taxa. Syst. Biol. 57:131–140.

901 Sackton, T. B., P. Grayson, A. Cloutier, Z. Hu, J. S. Liu, N. E. Wheeler, P. P. Gardner,
902 J. A. Clarke, A. J. Baker, M. Clamp, and S. V. Edwards. 2019. Convergent regulatory
903 evolution and loss of flight in paleognathous birds. Science 364:74–78.

904 Sanderson, M. J., A. Búrquez, D. Copetti, M. M. McMahon, Y. Zeng, and M. F.
905 Wojciechowski. 2020. A new (old) approach to genotype-based phylogenomic inference
906 within species, with an example from the saguaro cactus (*Carnegiea gigantea*). bioRxiv
907 Page 2020.06.17.157768.

908 Sayyari, E. and S. Mirarab. 2016. Fast coalescent-based computation of local branch
909 support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

910 Schull, J. K., Y. Turakhia, W. J. Dally, and G. Bejerano. 2019. Champagne:
911 Whole-genome phylogenomic character matrix method places Myomorpha basal in
912 Rodentia. bioRxiv Page 803957.

913 Scornavacca, C. and N. Galtier. 2017. Incomplete lineage sorting in mammalian
914 phylogenomics. Syst. Biol. 66:112–120.

915 Shedlock, A., K. Takahashi, and N. Okada. 2004. SINEs of speciation: tracking lineages
916 with retroposons. Trends Ecol. Evol. 19:545–553.

917 Shedlock, A. M., M. C. Milinkovitch, and N. Okada. 2000. SINE evolution, missing data,
918 and the origin of whales. Syst Biol. 49:808–817.

919 Shedlock, A. M. and N. Okada. 2000. SINE insertions: powerful tools for molecular
920 systematics. BioEssays 22:148–160.

921 Shen, X.-X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in
922 phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:1–10.

923 Simmons, M. P. and J. Gatesy. 2015. Coalescence vs. concatenation: Sophisticated analyses
924 vs. first principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91:98–122.

925 Springer, M. S. and J. Gatesy. 2016. The gene tree delusion. Mol. Phylogenet. Evol.
926 94:1–33.

927 Springer, M. S. and J. Gatesy. 2017. Pinniped diphyly and bat triphyly: more homology
928 errors drive conflicts in the mammalian tree. J. Hered. 109:297–307.

929 Springer, M. S. and J. Gatesy. 2018a. Delimiting coalescence genes (c-genes) in
930 phylogenomic data sets. Genes 9:123.

931 Springer, M. S. and J. Gatesy. 2018b. On the importance of homology in the age of
932 phylogenomics. Syst. Biodivers. 16:210–228.

933 Springer, M. S., E. K. Molloy, D. B. Sloan, M. P. Simmons, and J. Gatesy. 2020.
934 ILS-Aware analysis of low-homoplasy retroelement insertions: Inference of species trees
935 and introgression Using quartets. Journal of Heredity 111:147–168.

936 Springer M. S., G. J. 2014. Land plant origins and coalescence confusion. Trends Plant Sci.
937 19:267–269.

938 Stadler, T. and M. Steel. 2012. Distribution of branch lengths and phylogenetic diversity
939 under homogeneous speciation models. Journal of Theoretical Biology 297:33 – 40.

940 Suh, A., G. Churakov, M. P. Ramakodi, R. N. Platt II, J. Jurka, K. K. Kojima,
941 J. Caballero, A. F. Smit, K. A. Vliet, F. G. Hoffmann, J. Brosius, R. E. Green, E. L.
942 Braun, D. A. Ray, and J. Schmitz. 2015a. Multiple lineages of ancient CR1 retroposons
943 shaped the early genome evolution of amniotes. Genome Biol. Evol. 7:205–217.

944 Suh, A., M. Paus, M. Kiefmann, G. Churakov, F. A. Franke, J. Brosius, J. O. Kriegs, and
945 J. Schmitz. 2011. Mesozoic retroposons reveal parrots as the closest living relatives of
946 passerine birds. Nat. Comm. 2:443.

947 Suh, A., L. Smeds, and H. Ellegren. 2015b. The dynamics of incomplete lineage sorting
948 across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:e1002224.

949 Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other
950 methods). 4.0b10 ed.

951 Than, C. and L. Nakhleh. 2009. Species tree inference by minimizing deep coalescences.
952 PLoS Comput. Biol. 5:e1000501.

953 Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: a software package for analyzing and
954 reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.

955 Than, C. V. and N. A. Rosenberg. 2011. Consistency properties of species tree inference by
956 minimizing deep coalescences. J. Comput. Biol. 18:1–15.

957 Vachaspati, P. and T. Warnow. 2015. ASTRID: Accurate species TRees from internode
958 distances. BMC Genomics 16:S3.

959 Warnow, T. 2017. Computational Phylogenetics: An Introduction to Designing Methods
960 for Phylogeny Estimation. Cambridge University Press, Cambridge, United Kingdom.

961 Wright, S. 1931. Evolution in mendelian populations. Genetics 16:97–159.

962 Xi, Z., L. Liu, J. S. Rest, and C. C. Davis. 2014. Coalescent versus concatenation methods
963 and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63:919–932.

964 Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time
965 species tree reconstruction from partially resolved gene trees. BMC Bioinformatics
966 19:153.

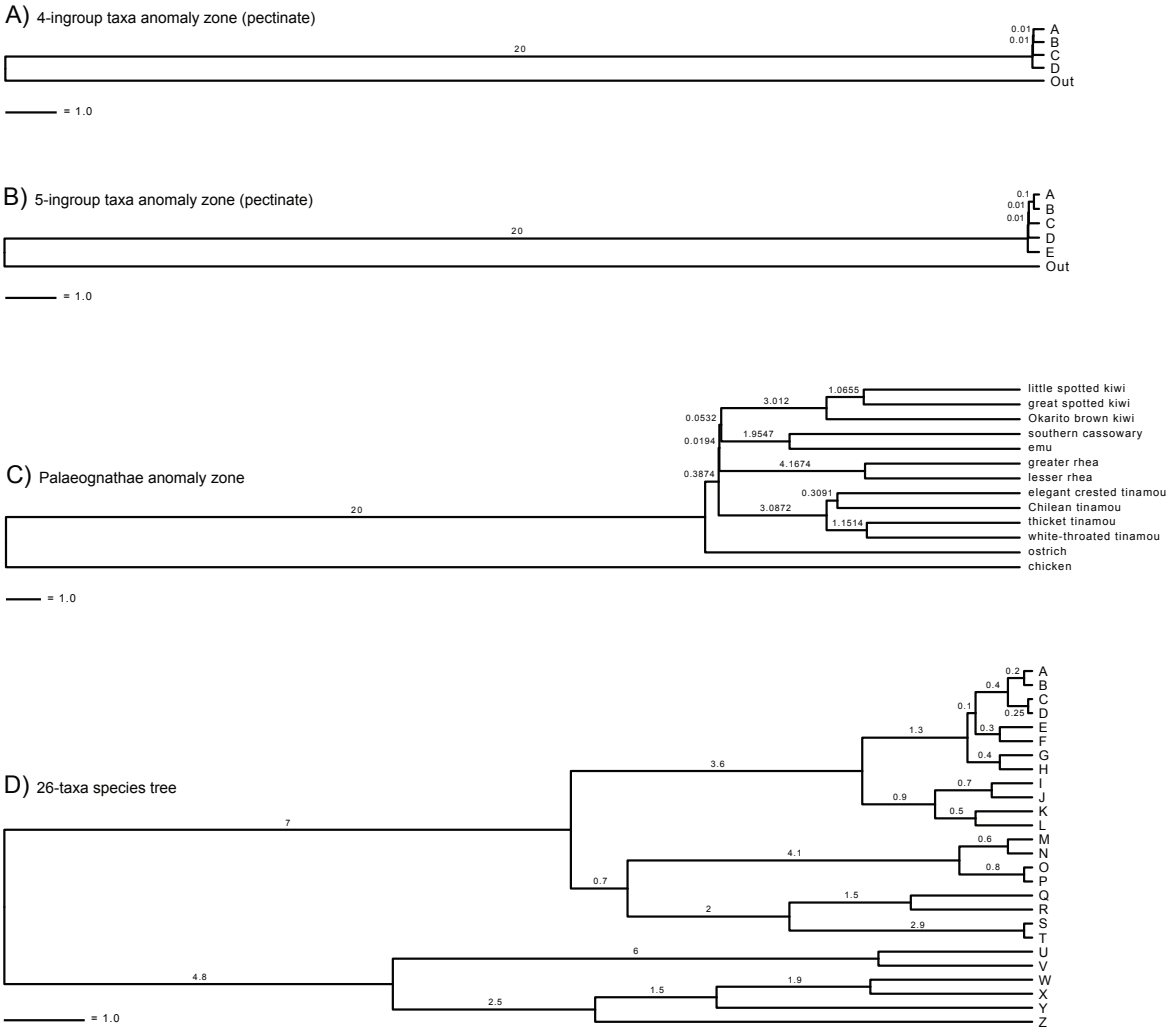

Figure 1: Four species trees that were employed in simulations. (A) 4-ingroup taxa anomaly zone tree, (B) 5-ingroup taxa anomaly zone tree, (C) ASTRAL tree for Palaeognathae from Cloutier et al. (2019), and (D) 26-taxa tree. Branch lengths are in coalescent units (CUs). Newick versions of all trees with branch lengths are available in Supplementary Material.

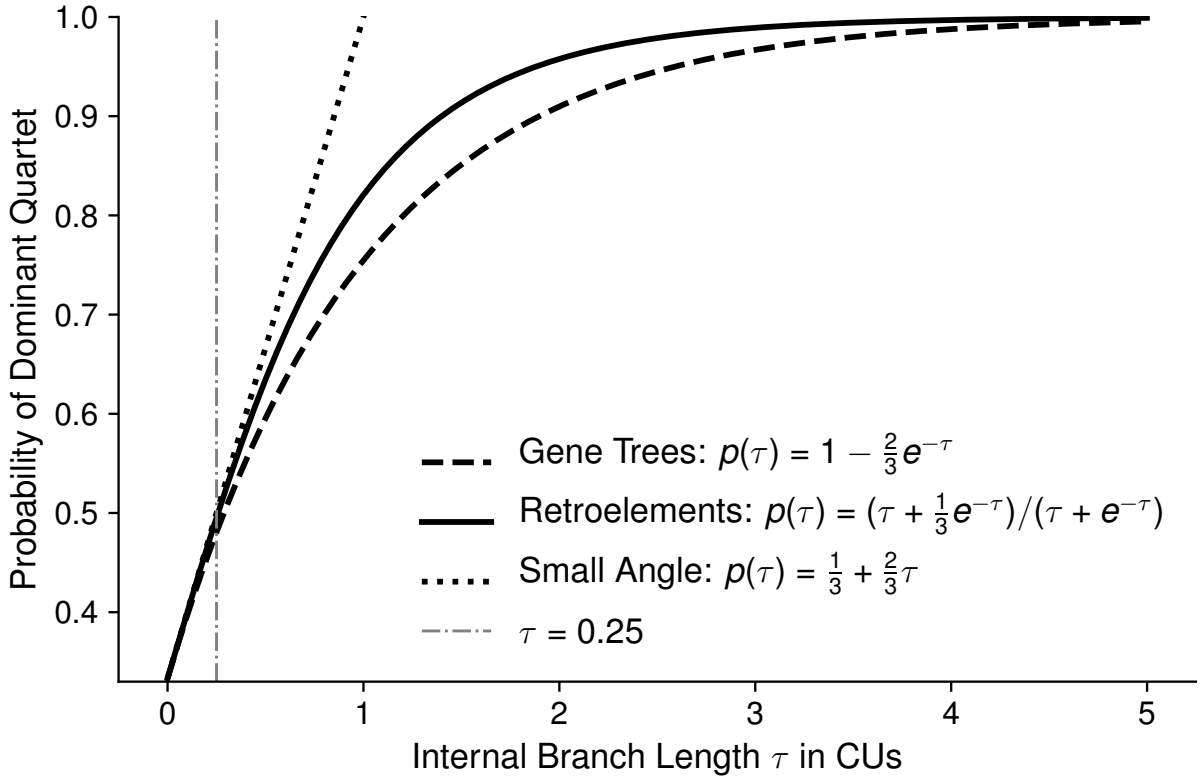

Figure 2: Relationship between the internal branch lengths and the probability of the dominant quartet (i.e. the quartet that agrees with the species tree). The formula for gene trees is under the MSC model (Allman et al. 2011). The formula for retroelements is under the MSC + infinite sites neutral mutation models, assuming that the expected number of new retroelement insertions per generation is constant across the species tree. When the internal branch lengths are sufficiently short so that we can use the small angle approximation $e^{-\tau} = 1–\tau$, the formula for gene trees and the formula for retroelements reduces to the equation shown above; this is the case even when the expected number of new retroelement insertions per generation is not constant across the tree.

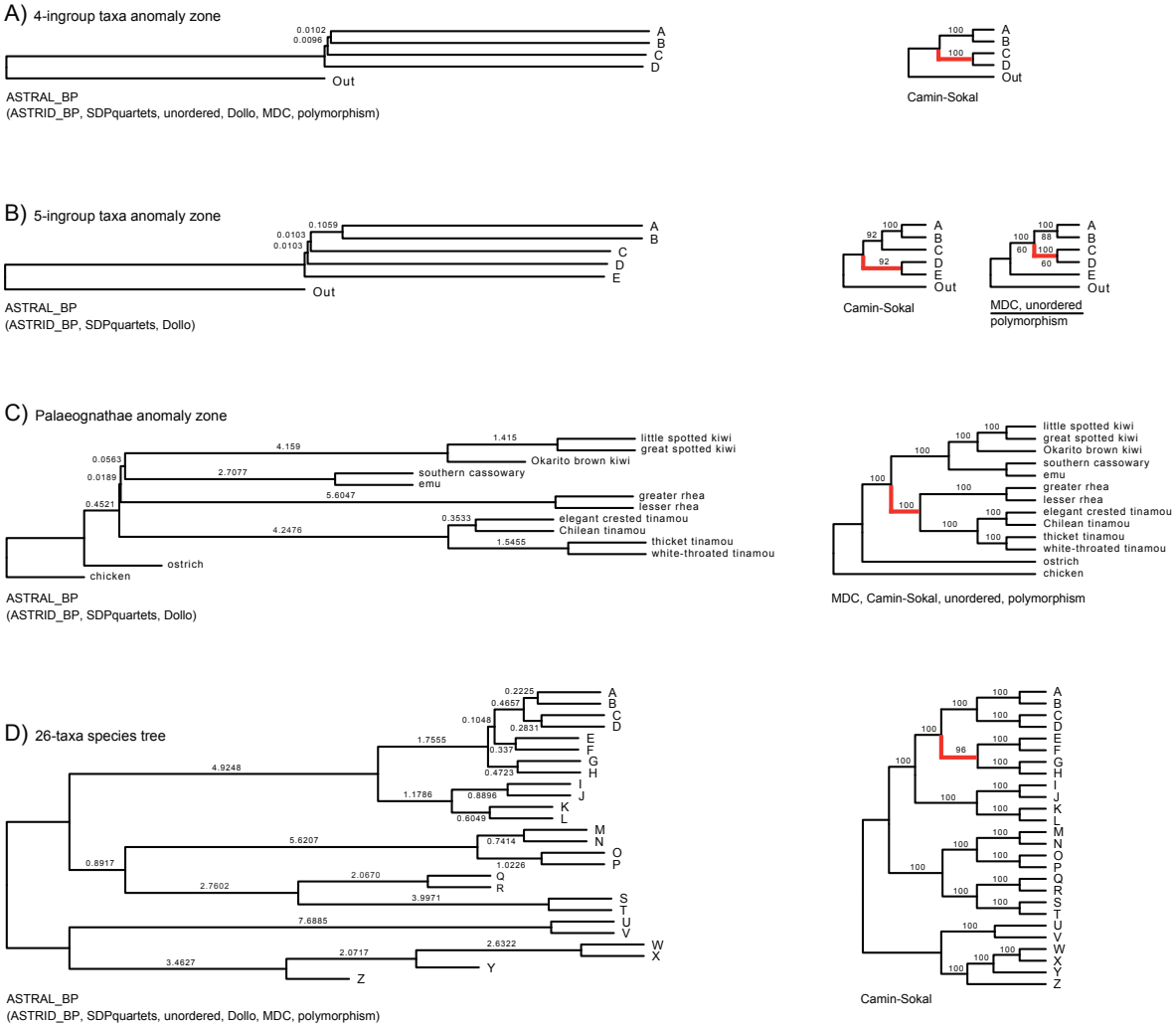

Figure 3: Summary of results for eight different phylogeny reconstruction methods (ASTRAL_BP, ASTRID_BP, SDPquartets, Dollo parsimony, Camin-Sokal parsimony, unordered parsimony, polymorphism parsimony, MDC) that were employed to estimate species trees for 25 simulated data sets for each of four different species trees: (A) 4-taxa anomaly zone tree, (B) 5-taxa anomaly zone tree, (C) ASTRAL TENT tree for Palaeognathae from Cloutier et al. (2019), and (D) 26-taxa tree. ASTRAL_BP species trees with mean MAP branch lengths (in coalescent units) based on analyses of 25 data sets per species tree are shown on the left (see Table 1 in the Supplementary Text for ML branch lengths and corrected branch lengths). ASTRAL_BP always recovered the correct species tree. Species trees for other methods that recovered the correct species tree for all 25 simulated data sets are shown in parentheses. Dollo parsimony recovered the correct tree for Palaeognathae in 22 of 25 simulations (also in parentheses on the left). Majority-rule consensus species trees for methods that never recovered the correct topology are shown on the right. Numbers above and below branches on these incorrect species trees indicate the percentage of analyses (out of 25) for which each clade was reconstructed. Red branches are those that conflict with the model species tree used to simulate retroelement insertions.

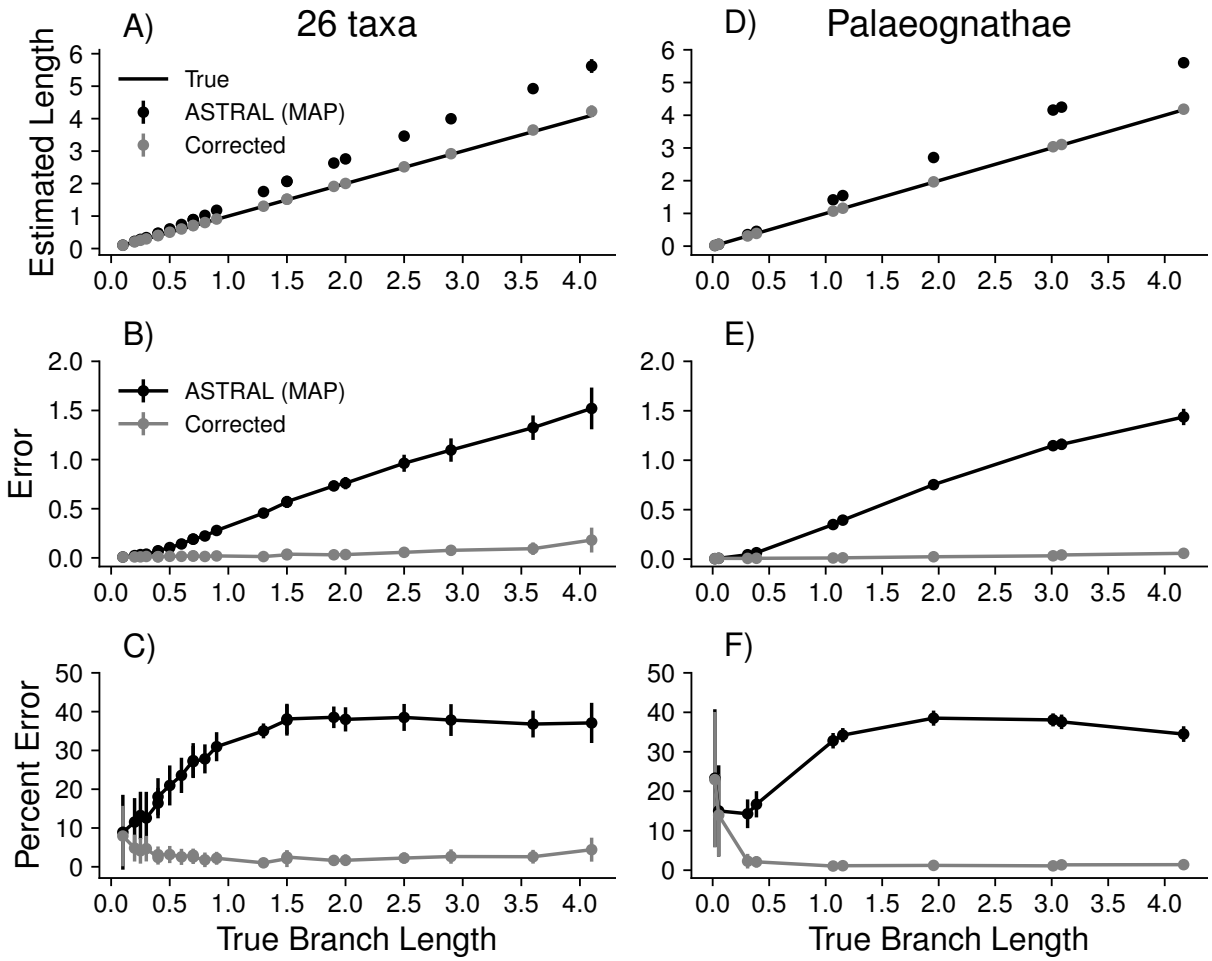

Figure 4: Branch length estimation for two of the four simulated data sets. Only true branch lengths of less than 5 coalescent units are shown, as computing large branch lengths is an ill-conditioned problem (Table 2 in Supplementary Text); results for longer branch lengths are in Table 1 in the Supplementary Text. Subfigures (A) and (D) show the true species tree branch length (x-axis) plotted against either the true branch lengths, the default (MAP) branch lengths estimated by ASTRAL, or the estimated by ASTRAL and then corrected with Equation 5. Subfigures (B) and (E) show the absolute value of the error: $abs(\tau^* - \hat{\tau})$, where $\hat{\tau}$ is the estimated branch length and τ^* is the true branch length. Note that when the true branch length τ^* is greater than 0.25 CUs, the both ASTRAL MAP and ML branch length estimates are greater than the true branch length for all 25 replicates (Table 1 in the Supplementary Text). Subfigures (C) and (F) show percent error: $(abs(\tau^* - \hat{\tau})/\tau^*) \times 100$. All values are averaged over 25 replicate data sets; dots are means, and bars are standard deviations.

Table 1: Branch lengths for species trees estimated using ASTRAL. ASTRAL TENT is the tree estimated by Cloutier et al. (2019) from 20,850 DNA-sequence-based gene trees. ASTRAL_BP is the tree produced by running ASTRAL_BP given the set of 4,301 retroelement insertions from Cloutier et al. (2019). The MAP branch length estimates are the default, the ML branch length estimates are returned by running ASTRAL with the option "-c 0.0," the corrected branch lengths are computed using Equation 5. When the normalized quartet support (defined below) is 1, Equation 5 is undefined, so we set the branch length to ∞ by default. EN, the effective number of retroelement insertions for that branch, can be smaller than the total number of insertions, because the insertions represent a single bipartition rather than a fully resolved gene tree. Note that all branches in the ASTRAL_BP tree had local PP of 1.0, except Kiwi, Casowary, & Emu, which had a local PP of 0.89. Regardless, when the EN is low, local PP should be interpreted cautiously.

Clade	ASTRAL TENT Analysis	ASTRAL_BP Analysis		
	Branch Length	Branch Length	Quartet Support	EN
	MAP	MAP / ML / Corrected		
Kiwi, Casowary, Emu, Rhea	0.0194	0.8938 / 1.1176 / 0.8657	0.7820 / 0.1429 / 0.0752	13.30
Kiwi, Casowary, Emu	0.0532	0.2528 / 0.2890 / 0.2587	0.5006 / 0.2402 / 0.2592	26.23
Chilean, elegant created tinamou	0.3091	1.7081 / 1.8124 / 1.3408	0.8912 / 0.0272 / 0.0816	73.50
All but chicken & ostrich	0.3874	2.5390 / ∞ / ∞	1.0000 / 0.0000 / 0.0000	18.00
Spotted kiwi	1.0655	3.4999 / 3.8986 / 2.8358	0.9865 / 0.0000 / 0.0135	148.00
White-throated, thicket tinamou	1.1514	4.7155 / 5.4057 / 4.0119	0.9970 / 0.0030 / 0.0000	334.00
Cassowary & emu	1.9547	3.6966 / ∞ / ∞	1.0000 / 0.0000 / 0.0000	59.46
All kiwi	3.0120	6.3045 / ∞ / ∞	1.0000 / 0.0000 / 0.0000	819.54
All tinamou	3.0872	5.6043 / 7.1065 / 5.4162	0.9994 / 0.0000 / 0.0006	522.79
All rhea	4.1674	7.3466 / ∞ / ∞	1.0000 / 0.0000 / 0.0000	2325.47