Journal of Phylogenetics & Evolutionary Biology

Springer and Gatesy, J Phylogenetics Evol Biol
2018, 6:3
DOI: 10.4172/2329-9002.1000205

Short Communication Open Access

On the Illogic of Coalescence Simulations for Distinguishing the Causes of Conflict among Gene Trees

Mark S. Springer[1*] **and John Gatesy**[2*]

[1]*Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, California 92521, USA*

[2]*Division of Vertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA*

[*]**Corresponding authors:** Mark S. Springer, Department of Evolution, Ecology and Organismal Biology, University of California, Riverside, California 92521, USA, Tel: 951-827-6458; E-mail: springer@ucr.edu; John Gatesy, John Gatesy, Division of Vertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA, Tel: 212-313-7529; E-mail: jgatesy@amnh.org

Received date: August 07, 2018; **Accepted date:** August 24, 2018; **Published date:** August 31, 2018

Introduction

Coalescent methods for species tree reconstruction have emerged as important alternatives to concatenation (supermatrix) approaches for species tree inference because they explicitly address the problem of incomplete lineage sorting (ILS). For genome scale datasets that include thousands of loci, fully parametric coalescence methods such as *BEAST [1] are not computationally pragmatic, and summary coalescence methods (e.g., STAR, MP-EST, ASTRAL [2-4]) commonly are applied instead [5,6]. However, summary coalescence methods make their own assumptions including recombination between but not within loci, neutral evolution, and the supposition that all gene tree heterogeneity is the result of ILS. Violations of these assumptions may cause summary coalescence methods to perform poorly relative to supermatrix approaches for species tree inference, and it should not be blithely assumed that summary coalescence methods will always perform better than concatenation as some authors have suggested [7]. In this note, we focus on gene tree heterogeneity and the improper use of simulation procedures that have been employed to assess the relative proportion of gene tree heterogeneity that can be explained by ILS.

ILS is undoubtedly an important source of gene tree heterogeneity. However, there are other important contributors to gene tree incongruence including gene tree reconstruction errors, undetected paralogy, and interspecific hybridization. Gene tree reconstruction error can be extensive when there are long branches, rapid radiations with short internodes, deep divergences, inadequate models of sequence evolution, improper alignments of sequences, and/or short loci with few informative characters. Given the many sources of gene tree dissonance, the relative importance of ILS is an open question for most genomic data sets. Song et al. [8] and Zhong et al. [9] concluded that ILS could account for the majority of gene tree heterogeneity for phylogenomic data sets for mammals (77%) and land plants (68%), respectively. However, Springer and Gatesy [10,11] and Gatesy and Springer [12] disputed the relative contribution of ILS to gene tree heterogeneity in both of these studies. At issue is the simulation procedure that was used by both Song et al. [8] and Zhong et al. [9] to assess the relative contribution of ILS to gene tree heterogeneity. Springer and Gatesy [10,11] called attention to the circular logic that underpins the simulation procedure employed by these authors [8,9]. Edwards et al. [13] claimed otherwise and suggest "recent computer simulations used to test the robustness of MSC models are not circular and do not unfairly favor MSC models over concatenation." This simulation procedure has also been employed in a recent study of flightless birds (Palaeognathae) to assess the proportion of gene tree heterogeneity that is attributable to the coalescent process [14]. To borrow a phrase from Edwards et al. [13], the "smoke-and-mirrors" procedure employed by these authors [8,9,14] is flawed at each step of the procedure (Table 1) and should be diligently avoided in future simulation work because it wrongly converts gene tree reconstruction errors that have nothing to do with ILS into ILS-based gene tree heterogeneity in the simulated gene trees.

Step	Problem
1. Reconstruct gene trees for each locus.	1. Gene trees are highly inaccurate because of long branch misplacement, model misspecification, missing data, lack of phylogenetic signal, poor searches, and/or non-homologous sequences.
2. Construct species tree with MP-EST, which assumes that all gene tree heterogeneity results from ILS.	2. Internal branches on species tree are much too short because of numerous gene tree reconstruction errors that are unrelated to ILS.
3. Simulate gene trees based on the estimated MP-EST species tree.	3. Gene trees have too much topological variation because of stunted branches on MP-EST species tree.
4. Compare reconstructed gene trees (Step 1) and simulated gene trees (Step 3) and concludes that ILS explains a high percentage of gene tree variation.	4. Procedure is circular and is rigged to conclude that ILS explains a high percentage of gene tree heterogeneity.

Table 1: Outline of the steps and problems in the circular simulation procedure employed by Song et al. [8] and Zhong et al. [9] to assess the percentage of gene tree heterogeneity that is explained by incomplete lineage sorting (ILS) for mammalian and land plant data sets, respectively.

In Step 1 of their simulation procedure (Table 1), both Song et al. [8] and Zhong et al. [9] reconstructed gene trees for each locus in the dataset. In the case of Song et al. [8], these gene trees are highly inaccurate because of few informative sites in some alignments, switched taxon names that interchanged a marsupial (wallaby) and a primate (macaque), alignment of non-homologous sequences,

extensive missing data for some taxa, poor branch swapping, and an inadequate model of sequence evolution (GTR instead of GTR+Γ) [11]. We corrected the switched taxon names, discarded alignments with non-homologous sequences that are shared by multiple taxa, and performed more thorough searches with SPR + NNI branching and a GTR+Γ model of sequence evolution [11] for 413 alignments. These searches resulted in improved gene trees with much less gene tree reconstruction error. For example, we [11] recovered Boreoeutheria as monophyletic on 324 of 413 gene trees (78.5% congruence) whereas Song et al.'s [8] original analyses yielded a monophyletic Boreoeutheria on just 117 of 447 gene trees (26.2% congruence). Even with these improvements, the 413 fixed gene trees contain numerous reconstruction errors because of long-branch misplacement, model misspecification, long tracts of missing data, and or/ lack of phylogenetic signal [11]. For example, 154 of 413 gene trees conflict with Haplorhini (anthropoids+tarsiers) even though relationships at the base of this clade are not candidates for ILS or interspecific hybridization (see below). The plant gene trees from Zhong et al. [9] show even more conflicts than the mammal gene trees with an average of 73% incongruence in pairwise comparisons among gene trees [15,16]. After reconstructing gene trees, Song et al. [8] and Zhong et al. [9] reconstructed species trees with the summary coalescence method MP-EST in Step 2 of their simulation procedure (Table 1). However, estimates of internal branch lengths in the MP-EST species trees are much too short (in coalescent units) because of reconstruction errors in the input gene trees that are unrelated to ILS. In Step 3, the MP-EST species tree with its stunted internal branches was used to simulate gene trees (Table 1). These simulated gene trees exhibit excess topological variation because the gene trees were simulated from an MP-EST species tree with much shorter internal branches than those in the true species tree. In the final step (Step 4; Table 1), Robinson Foulds (RF) distances [17] for the simulated gene trees (Step 3) were compared to RF distances for the original gene trees (Step 1) to determine the percentage of gene tree variation that is accounted for by ILS. However, this procedure necessarily converts all gene tree variation, including variation that has nothing to do with ILS (Step 1 trees), into ILS variation (Step 3 trees) because MP-EST assumes that the only source of gene tree heterogeneity is ILS. This procedure therefore leads to the spurious conclusion that ILS accounts for the majority of gene tree variation, both for Song et al.'s [8] mammal data and Zhong et al.'s [9] land plant data.

A simple example with three primates (*Microcebus* [mouse lemur], *Tarsius* [tarsier], *Homo* [human]) and one outgroup (*Mus* [mouse]) illustrates the fundamental problems with this simulation procedure (Figure 2 and Table 1). The three primate species in this example are representatives from the three main groups of the order Primates: Strepsirrhini (mouse lemur), Tarsiiformes (tarsier), and Anthropoidea (human). Relationships among these three groups have a long history of debate, with all three alternatives (1=tarsier-first hypothesis [Strepsirrhini+Anthropoidea], 2=Prosimii [Strepsirrhini +Tarsiiformes], 3=Haplorhini [Anthropoidea+Tarsiiformes]) receiving support from molecular phylogenetic analyses in the last two decades [18-23]. More recent sequence-based studies based on multiple loci and genomic data consistently support Haplorhini [8,24-31], but with mixed support from the individual gene trees (Figures 1A-1C). By contrast, a recent study of rare genomic changes reported 104 conflict-free transposon insertions that are shared by Haplorhini (Anthropoidea+Tarsiiformes) to the exclusion of Strepsirrhini (Figure 1D) [32]. Transposon insertions are especially useful for discriminating between ILS and gene tree reconstruction errors because they have extremely low levels of standard homoplasy (reversal and convergence) that afflicts nucleotide substitutions [12,33-35]. At the same time, conflict among transposon characters is expected when successive speciation events occur in rapid succession and there is ILS [35,36].

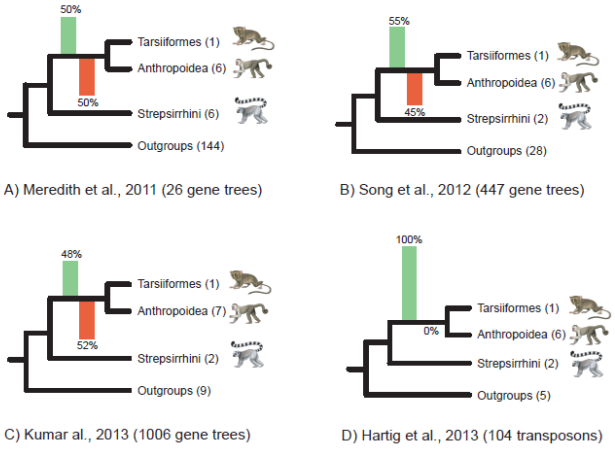

Figure 1: Conflict and congruence for the phylogenetic placement of Tarsiiformes (tarsiers) within Primates in recent large-scale systematic studies (A-D). All phylogenetic analyses of DNA sequences (A-C) robustly position Tarsiiformes as sister to anthropoid primates (e.g., Jameson et al. [24]), and 104 insertions of transposons that were mined from genomic comparisons uniformly support this clade (Haplorhini; green bar at internode) (D). One hundred percent congruence of 104 transposon inserts with no conflicts suggests an extremely low rate of incomplete lineage sorting (ILS) at this node. By contrast, DNA-based reconstructions of gene trees in multiple studies (A-C) document poor congruence (green) at this node (48%-55%), which therefore implies extensive gene tree reconstruction errors (red; 45%-52%) at a node where no conflict is expected. For each study, the number of gene trees reconstructed is shown, and the number of species sampled for each group is in parentheses to the right of each taxonomic name. 'Outgroups' refers to all species in an analysis that are not primates.

Hartig et al.'s [32] characterization of 104 uncontradicted transposon insertions for Haplorhini (Figure 2A) suggests that ILS (or gene flow between divergent lineages) does not contribute to gene tree heterogeneity for the phylogenetic placement of tarsiers relative to anthropoids and strepsirrhines. Indeed, the MP-EST tree based on these 104 transposon trees has a branch length of 7.0 coalescent units for the stem Haplorhini branch (i.e., ancestral branch leading to human and tarsier) (Figure 2A). This value of 7.0 coalescent units is the maximum branch length for an MP-EST species tree. We simulated 100000 gene trees based on this MP-EST species tree, and 99.9% of the trees support human+tarsier to the exclusion of lemur (Figure 2A). This result precludes any significant role for ILS in generating topological variation at this node. We also extracted 4-taxon subtrees (mouse, mouse lemur, tarsier, human) from the 413 RAxML gene trees that were estimated using a GTR+Γ model of sequence evolution for loci from Song et al. These subtrees were obtained by pruning all other taxa and then rooting on mouse. The consensus of these 413 gene trees (four taxa) shows ~65.4% support for human+tarsier to the exclusion

of mouse lemur (Figure 2B). Similar levels of mixed support for Haplorhini are evident in gene trees from other data sets with 45% to 52% conflict at this node (Figures 1A-1C). Next, we used MP-EST to construct a species tree based on these 413 gene trees for mouse, mouse lemur, human, and tarsier (Figure 2B). The inferred MP-EST species tree, with a short branch length of 0.656 coalescent units for stem Haplorhini, was then used to simulate 100000 gene trees. The consensus of these 100000 simulated gene trees shows ~65.5% support for human+tarsier. Song et al.'s [8] simulation procedure would then conclude that ILS explains nearly 100% of the gene tree heterogeneity in this example! This is because 35.5% of the simulated gene trees do not support human+tarsier just as 35.6% of the original gene trees do not support this clade (Figure 2B). However, the failure of ~35.5% of these gene trees to support Haplorhini (human+tarsier) cannot be attributed to ILS in view of unanimous, overwhelming support from 104 transposons [32]. Instead, the failure of gene trees to support Haplorhini must be attributed to gene tree reconstruction errors or other factors such as unrecognized paralogy in the DNA sequence alignments.

It is perhaps surprising that Haplorhini is frequently absent on individual gene trees given unanimous and compelling transposon support for this clade. Possible causes of gene tree reconstruction error include a relatively long Tarsiiformes branch on many gene trees and a divergence time for Haplorhini that extends as far back as the early Paleocene e.g., 65 million years ago [28]. If Haplorhini is so difficult to recover even though transposon support is unambiguous, then we should be especially wary of gene tree reconstruction error that affects difficult to resolve nodes that are bracketed by short internal branches and have equivocal transposon support, e.g., the root of Placentalia or relationships among laurasiatherian orders [35,37,38]. Future efforts to infer species trees from gene trees must confront real biological phenomena such as ILS and interspecific hybridization, but should not ignore the impact of gene tree reconstruction error, which in many cases can dwarf other causes of gene tree heterogeneity.

Song et al. [8], Zhong et al. [9], and Cloutier et al. [14] have all suggested that ILS can account for the majority of gene tree heterogeneity in diverse datasets for mammals (77%), land plants (78%), and flightless birds (>70% for most comparisons with CNEEs and UCEs, >90% for some comparisons with introns), respectively. These values are all based on the same circular simulation procedure and conflict with simulations where authors have compared true gene trees that were simulated under the multispecies coalescent with gene trees that were estimated from simulated DNA sequences (for the same true gene trees). Edwards et al. [13] performed simulations based on a species tree from Springer and Gatesy [11], and gene tree reconstruction error accounted for 50% of the total gene tree heterogeneity relative to the species tree. This estimate is higher than for any of the above empirical estimates that employed circular logic. Mirarab et al.'s [39] simulation results show that gene tree reconstruction errors impact 27%-79% of the nodes on the simulated gene trees for species trees modeled form bird and mammal datasets. Assuming that there is absolutely no gene tree reconstruction error and then blithely carrying on with simulations is therefore not a sound procedure when the output is so dependent on the input (Table 1 and Figure 2).

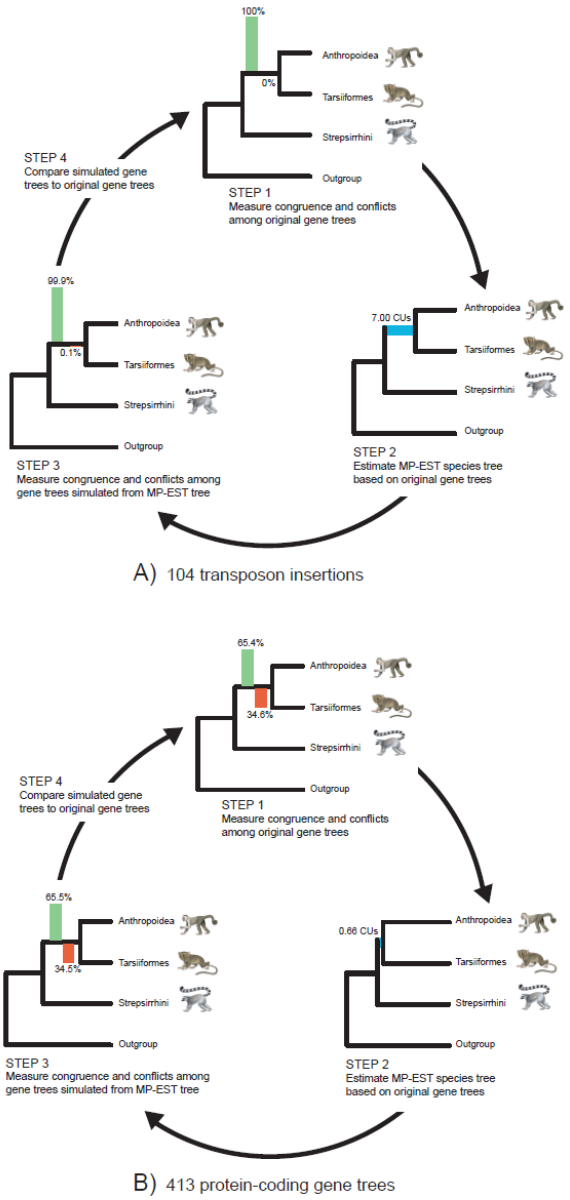

Figure 2: A schematic of the four main steps in the circular simulation procedure used by Song et al. [8] and Zhong et al. [9] to estimate how much conflict among gene trees is due to ILS for (A) Hartig et al.'s [32] 104 transposons and (B) Springer and Gatesy's [11] 413 protein-coding genes from Song et al. [8]. For Step 1, the percentages of congruence (green) and conflict (red) among loci are indicated at the Haplorhini node (Anthropoidea+Tarsiiformes) for gene trees inferred from the original data. For Step 2, the MP-EST species tree estimated from the original gene trees is shown with the branch length (in coalescent units=CUs) indicated for the common ancestor for Haplorhini (blue). In Step 3, the percentages of congruence (green) and conflict (red) among 100000 gene trees simulated from the MP-EST species trees are shown. In Step 4, gene trees estimated from the original data are compared to gene trees from the simulations. Note that that the amount of conflict among loci in the output of the simulations closely matches the amount of conflict in the output of the circular simulation procedure. For both datasets, just one genus was sampled from each group to simplify calculations (Anthropoidea=*Homo*, Tarsiiformes=*Tarsius*, Strepsirrhini=*Microcebus*, outgroup=*Mus*).

It is naive to assume, at the start, that every node in every gene tree is accurately reconstructed. We suspect that this is very rarely the case for DNA sequence based phylogenomic analyses, especially in view of problems with real sequence alignments (e.g., long tracts of missing data, non-homologous sequences, non-stationarity) that negatively impact gene tree accuracy but are rarely modeled in simulated sequences.

In summary, the multispecies coalescent is a fundamentally important concept in population genetics, and provides a conceptual model for addressing problems with ILS in systematic studies. However, the application of the multispecies coalescent to deep level phylogenomics is a challenge, in contrast to the opinions of many authors [6-9,13]. Concatenation may fail in the anomaly zone [40], but summary coalescence methods have their own problems and are not guaranteed to fare any better due to small coalescence genes that are the input for these methods [11,41], the recombination ratchet [11,12,42], and gene tree reconstruction errors [11,12,35,43]. Indeed, Scornavacca and Galtier [41] reported that ILS only makes a minor contribution to conflict among gene trees for a mammalian phylogenomic data set comprised of protein-coding sequences. It remains critical to assess the relative contribution of ILS to gene tree heterogeneity for other data sets, but the circular simulation procedure of Song et al. [8] and Zhong et al. [9], which has also been applied to assess gene tree heterogeneity in flightless birds [14], should not be used for this purpose because of its flawed consecution.

Acknowledgments

This research was supported by NSF grant DEB-1457735 to JG and MSS.

References

1. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27: 570-580.
2. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53: 320-328.
3. Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10: 302.
4. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, et al. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541-i548.
5. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al. (2014) Whole-genome analyses resolve early branches in the tree of life. Science 346: 1320-1321.
6. Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, et al. (2017) Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci USA. 114: E7282-E7290.
7. Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annu Rev Ecol Evol Syst 44: 99-121.
8. Song S, Liu L, Edwards SV, Wu S (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA 109: 14942-14947.
9. Zhong B, Liu L, Yan Z, Penny D (2013) Origin of land plants using the multispecies coalescent model. Trends Plant Sci 18: 492-495.
10. Springer MS, Gatesy J (2014) Land plant origins and coalescence confusion. Trends Plant Sci 19: 267-269.
11. Springer MS, Gatesy J (2016) The gene tree delusion. Mol Phylogenet Evol 94: 1-33.
12. Gatesy J, Springer MS (2014) Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol Phylogenet Evol 80: 231-266.
13. Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, et al. (2016) Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol Phylogenet Evol 94: 447-462.
14. Cloutier A, Sackton TB, Grayson P, Clamp M, Baker AJ, et al. (2018) Whole- genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. bioRxiv: 262949.
15. Simmons MP, Sloan DB, Gatesy J (2016) The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol Phylogenet Evol 97: 76-89.
16. Simmons MP, Sloan D, Springer MS, Gatesy J (in review) Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. Mol Phylogenet Evol.
17. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53: 131-147.
18. Andrews TD, Jermiin LS, Easteal S (1998) Accelerated evolution of cytochrome b in simian primates: adaptive evolution in concert with other mitochondrial proteins? J Mol Evol 47: 249-257.
19. Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, et al. (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol 9: 585-598.
20. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, et al. (2001a) Molecular phylogenetics and the origins of placental mammals. Nature 409: 614-618.
21. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, et al. (2001b) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348-2351.
22. Schmitz J, Ohme M, Zischler H (2002) The complete mitochondrial sequence of *Tarsius bancanus*: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol Biol Evol 19: 544-553.
23. Chatterjee HJ, Ho SY, Barnes I, Groves C (2009) Estimating the phylogeny and divergence times of primates using a supermatrix approach. BMC Evol Biol 9: 259.
24. Jameson NM, Hou Z-C, Sterner KN, Weckle A, Goodman M, et al. (2011) Genomic data reject the hypothesis of a prosimian primate clade. J Hum Evol 61: 295-305.
25. Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, et al. (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334: 521-524.
26. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342.
27. Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, et al. (2012) Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7: e49521.
28. dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, et al. (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc R Soc B 279: 3491-3500.
29. Kumar V, Hallström BM, Janke A (2013) Coalescent-based genome analyses resolve the early branches of the Euarchontoglires. PLoS ONE 8, e60019.
30. Foley NM, Springer MS, Teeling EC (2016) Mammal madness: is the mammal tree of life not yet resolved? Phil Trans R Soc B 371: 20150140.
31. Tarver JE, dos Reis M, Mirarab S, Moran RJ, Parker S, et al. (2016) The interrelationships of placental mammals and the limits of phylogenetic inference. Genome Biol Evol 8: 330-344.
32. Hartig G, Churakov G, Warren WC, Brosius J, Makałowski W, et al. (2013) Retrophylogenomics place tarsiers on the evolutionary branch of anthropoids. Sci Reports 3: 1756.
33. Shedlock AM, Okada N (2000) SINE insertions: powerful tools for molecular systematics. BioEssays 22: 148-160.

34. Doronina L, Churakov G, Shi J, Brosius J, Baertsch R, et al. (2015) Exploring massive incomplete lineage sorting in arctoids (Laurasiatheria, Carnivora). Mol Biol Evol 32: 3194-3204.
35. Gatesy J, Meredith RW, Janecka JE, Simmons MP, Murphy WJ, et al. (2017) Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. Cladistics 33: 295-332.
36. Suh A, Smeds L, Ellegren H (2015) The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol 13: e1002224.
37. Nishihara H, Hasegawa M, Okada N (2006) Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc Natl Acad Sci USA 103: 9929–9934.
38. Nishihara H, Maruyama S, Okada N (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.
39. Mirarab S, Bayzid MS, Boussau B, Warnow T (2014) Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346: 1250463.
40. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2: e68.
41. Scornavacca C, Galtier N (2017) Incomplete lineage sorting in mammalian phylogenomics. Syst Biol 66: 112-120.
42. Springer MS, Gatesy J (2018) Delimiting coalescence genes (c-genes) in phylogenomic data sets. Genes 9: 123.
43. Gatesy J, Springer MS (2017) Phylogenomic red flags: Homology errors and zombie lineages in the evolutionary diversification of placental mammals. Proc Natl Acad Sci USA 114: E9431-E9432.