Journal of Phylogenetics & Evolutionary Biology

Patel et al., J Phylogen Evolution Biol 2013, 1:2
DOI: 10.4172/2329-9002.1000110

Research Article Open Access

Error in Phylogenetic Estimation for Bushes in the Tree of Life

Swati Patel[1,2]*, Rebecca T. Kimball[1] and Edward L Braun[1]

[1]*Department of Biology, 223 Bartram Hall, P.O. Box 118525, University of Florida, Gainesville, Florida 32611-8525, USA*
[2]*Department of Mathematics, One Shields Avenue, University of California, Davis, USA*

Abstract

Many rapid radiations, or bushes, throughout the Tree of Life remain unresolved. Here, we investigated how the shape of a bush interacts with two key processes - coalescence and mutation - that can lead to errors in phylogenetic inference under specific conditions. For this study, we focused on the tradeoff between sampling more individuals per species and sampling more loci as well as the utility of a species tree method based upon gene tree reconciliation and the concatenation of multiple loci for resolving bushes. We examined different bush shapes, varying both the speciation rate during the radiation and the depth of the radiation, to encompass a broad range of situations. Using simulations based upon parameters derived from empirical studies, we investigated the performance of phylogenetic analyses under different conditions to identify approaches with the greatest potential to resolve difficult phylogenies. Sampling a single individual for more loci outperformed sampling multiple individuals for one locus in all cases except the most recent radiations. We found that error due to homoplastic mutations increased with depth, while error due to the coalescent process remained unchanged. These simulations also revealed that, for certain ancient bushes, analyses of concatenated data matrices surprisingly resulted in more accurate phylogenies than gene tree reconciliation. The poor performance of gene tree reconciliation in this study appeared to reflect the poor estimation of gene trees, not the superiority of concatenation per se. Our results suggest concatenation remains a useful approximate method for species tree estimation, even for rapid evolutionary radiations. However, improved estimation of gene trees combined with use of gene tree reconciliation has the greatest potential for resolving the remaining bushes of the Tree of Life.

Keywords: Bushes; Coalescence; Lineage sorting; Mutation; Phylogenetics; Simulation; Tree of life

Introduction

Despite the progress made on assembling the Tree of Life, many clades remain unresolved even after substantial effort. These difficult clades have been called "bushes in the Tree of Life" [1] and they are thought to reflect rapid evolutionary radiations. Empirical examples of bushes are ubiquitous throughout the Tree of Life [1-6] and they are especially difficult to resolve when they are ancient [6], leaving large gaps in our knowledge about the Tree of Life.

Bushes in the Tree of Life can be characterized based on the rate of speciation during the evolutionary radiation and the overall depth of that radiation (Figure 1). Rate captures the times between speciation events whereas depth captures the time since the radiation, and these two parameters can be viewed as defining a "bush shape". Depth shows especially striking variation; even if we restrict consideration to animals, there are bushes as deep as 550 million years ago (Ma) when the Cambrian explosion occurred [7], to as recent as 100,000 years ago when the Lake Victoria cichlids radiated [8]. Although it has long been recognized that trees with long terminal branches (large depth) and short internal branches (high rate) are often difficult to resolve [6,9] the relationship between rate, depth, and phylogenetic error remains unclear. Understanding the ways that these characteristics (rate and depth) influence phylogenetic estimation will be beneficial to making further progress in resolving bushes in the Tree of Life.

To explore the best approaches for resolving bushes in the Tree of Life, it is important to consider the processes that lead to genetic differences among taxa. Ultimately, the differences among taxa reflect a complex set of stochastic processes that include lineage sorting due to the coalescent, patterns of mutation, recombination, horizontal gene transfer and introgression due to hybridization, and the duplication and loss of genes [10-12]. If we restrict our attention to vertebrates the first two processes are likely to have the greatest impact upon the resolution of bushes, although all of those processes make important contributions to the differences among genomes. The coalescent describes the history of alleles in populations [13,14] and the random sorting of alleles into different lineages due to the coalescent can result in discordance between gene trees and species trees [10,12]. Mutational processes have an impact both upon the probability of finding synapomorphic mutations that unite taxa [15] and the likelihood that homoplasy will obscure phylogenetic signal [6,16]. The bush shape (i.e., the speciation rate and depth of the radiation) is likely to influence whether one or both of these factors, called "coalescent error" and "mutational error" hereafter, will obscure the true phylogeny.

The coalescent process, which can lead to discordance between gene trees and species trees, has been extensively studied from both a theoretical [10,17-19] and empirical [20-22] standpoint. It is known that there are extreme situations (the "anomaly zone") where the most common gene tree is discordant with the species tree [23]. However, these studies have typically focused on relatively recent, or shallow, radiations. There has been limited study of the problem for ancient rapid radiations, but the coalescent process should have just as much potential to result in gene tree discordance for ancient radiations [6,24]. Indeed, recent analyses of deep mammalian phylogeny using methods that accommodate discordance among gene trees due to the coalescent do yield different results than concatenation [25,26]. However, other empirical studies have not revealed clear evidence that gene tree discordance due to the coalescent has had an impact upon species tree estimation for ancient radiations [27-29]. Moreover, methods that accommodate gene tree discordance do appear to improve the efficiency and accuracy of phylogenetic estimation [30], at least for some problems. Regardless, the fact that gene tree-species tree

***Corresponding author:** Swati Patel, Department of Biology, 223 Bartram Hall, P.O. Box 118525, University of Florida, Gainesville, Florida 32611-8525, USA, Tel: 224-558-9786; E-mail: swpatel@ucdavis.edu

Received March 26, 2013; **Accepted** May 31, 2013; **Published** June 10, 2013

Citation: Patel S, Kimball RT, Braun EL (2013) Error in Phylogenetic Estimation for Bushes in the Tree of Life. J Phylogen Evolution Biol 1: 110. doi:10.4172/2329-9002.1000110

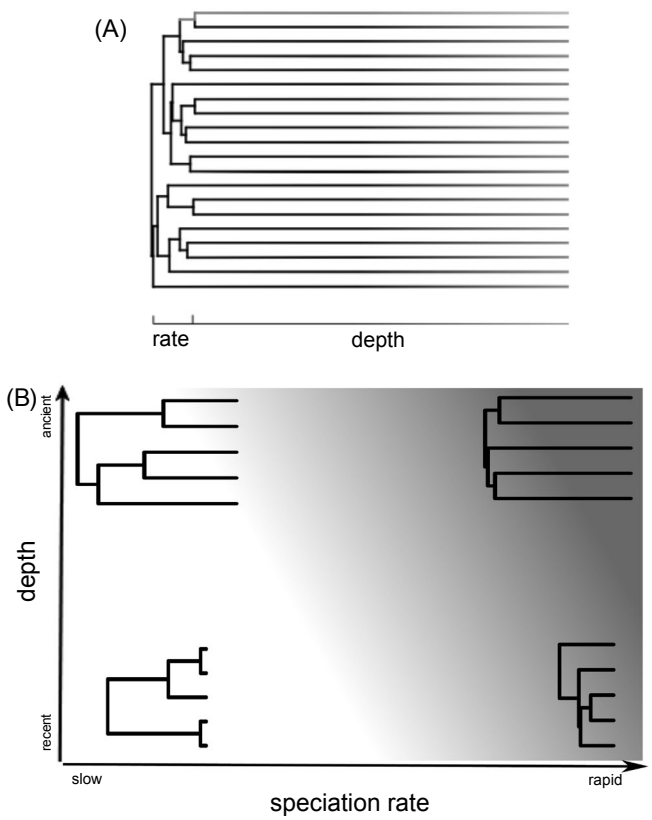

Figure 1: Parameter space describing bushes: (A) An example of a simulated phylogenetic tree defining two important characteristics, rate of radiation and depth. (B) The range of tree shapes we consider in this study. The shade of gray indicates how "bushlike" the phylogeny is, with lighter areas corresponding to "treelike" phylogenies and darker areas to more "bushlike" phylogenies.

discordance is as likely for ancient radiations as it is in recent radiations suggests that analyses of ancient rapid radiations should consider the impact of coalescent error.

The stochastic nature of the mutational process can also result in difficulties for the resolution of bushes in the Tree of Life. Even if gene tree-species tree discordance were ignored it would be difficult to obtain accurate estimates of gene trees for ancient rapid radiations [6,16]. This reflects a combination of three factors that we view as aspects of mutational error (we use the term "mutational error" to conform to the terminology used by Huang et al. [31] but note that this source of error includes both the origin of novel alleles by mutation and all other factors that influence the fixation of these alleles in lineages). First, present-day sequences are likely to exhibit substantial homoplasy that can obscure the branching pattern during an ancient radiation. In the extreme this homoplasy could lead to substitutional saturation [32,33]. Second, a limited number of informative characters are expected to be present in finite sequences that were generated by evolution on trees with short internodes [15]. This can lead to a requirement for very long sequences to accurately estimate gene trees [34]. Finally, there are cases where the expected site pattern spectra support a topology that conflicts with the true tree due to factors like long-branch attraction and base compositional convergence [32,35-37]. There has been substantial effort to develop phylogenetic methods that can detect and overcome bias, but it is important to recognize that the first two phenomena can be problematic for the estimation of gene trees even if long-branch attraction, base compositional convergence, and other sources of bias are absent. Specifically, the impact of homoplastic changes after the radiation is related to depth whereas the expected number of informative characters that define clades in gene trees is related to the rate of speciation during the radiation. Thus, both of the parameters we use to describe bushes are expected to affect the impact of mutational error upon phylogenetic estimation.

For each bush in the Tree of Life, it can be difficult to identify whether the lack of resolution reflects coalescent or mutational error (or both). Indeed, identifying the likely source of error may allow researchers to modify their sampling strategies and analytical methods to better address the specific source of phylogenetic error for a given problem. The relative contribution of coalescent and mutational error may depend on the shape of the bush. Several studies have explicitly examined how specific characteristics of a tree can affect phylogenetic inference [31,38-41]. While providing important insights into phylogenetic inference, these studies did not examine the impact of both coalescent and mutational variance upon radiations that are both ancient and rapid. However, it will be necessary to understand the impact of both processes and the ways that they interact to then identify methods with the greatest potential to resolve difficult bushes.

Here we use simulations to explore the impact of both coalescent and mutational processes upon phylogenetic estimation given different bush shapes (i.e., different rates of speciation and depths). We base our simulation parameters on empirical observations obtained from studies of tetrapods, specifically birds and mammals (see Supplementary Material), to better link theory with practice and provide useful recommendations for empirical studies. We use these simulations to examine the impact of depth and rate upon two methodological questions likely to be important. First, we examine the tradeoff between sampling more individuals per species and sampling more loci. Second, we compare the performance of a species tree method based upon gene tree reconciliation in a coalescent framework and simple phylogenetic analyses that use a concatenated alignment of multiple loci for resolving bushes. The goal of these analyses was to address the larger question of how to best resolve bushes in the Tree of Life and to gain insights into whether full resolution of the Tree of Life from sequence data is possible.

Methods

In this study, we generated random species trees, then we simulated gene trees based on these species trees, and finally we simulated nucleotide sequences using the gene trees. Phylogenetic trees obtained from analyses of the simulated data were compared with the true species trees that were used for the simulation (Figure 2). Details of these analyses are provided below.

Simulations

To examine the contribution of the coalescent and mutational processes to error in phylogenetic estimation, we simulated 20-taxon species trees assuming a Yule process [42] for various speciation rate and depth combinations. The simulations used code modified from the LASER package [43] in R. Speciation rate refers to the expected number of speciation events during each unit of time, while depth refers to the time since the last speciation event (Figure 1A). All times were measured in coalescent time units ($2N_e$ generations for diploids, where N_e is the effective population size). We used a broad range of speciation rates, with the full set of speciation rates corresponding to 0.01, 0.1, 0.5, 1, 2.5, 5, 7.5, and 10, which resulted in a range of tree shapes (Figure 1B). After simulating the species tree, the terminal branches were then

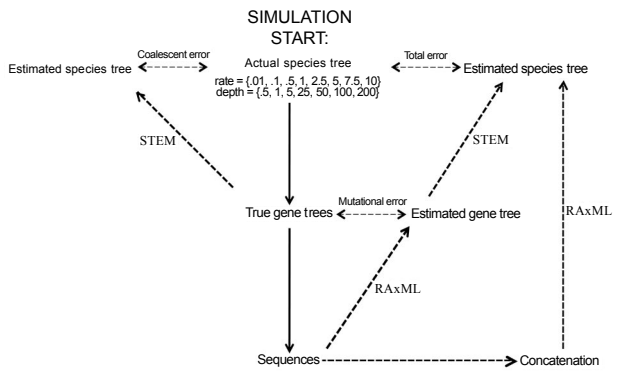

Figure 2: Schematic of our simulation procedure: Coalescent error was measured using the Robinson-Foulds (RF) distance between true species trees and species trees estimated using the true gene trees. Mutational error was measured using the RF distance between true gene trees and gene trees estimated using sequence data. Total discordance between the true species trees and species trees was estimated using gene trees that were estimated using sequence data. Solid lines represent simulated data, while dashed lines indicate estimations from the simulated data.

extended to place the radiation at various depths. The depth values ranged from 0.5 coalescent units to 200 coalescent units (the full set was 0.5, 1, 5, 25, 50, 100, and 200). We then simulated gene trees under the coalescent process using the Phybase package [44] in R using these simulated species trees.

We simulated 1000 base pair (bp) sequences (a typical length for many markers; [45-47] using each gene tree in SeqGen [48], assuming the HKY model of evolution with a range of parameters that are typical of vertebrate introns. We established ranges of parameter values from empirical studies (Supplementary Material) and then chose values for our simulated regions randomly from these distributions. We used these parameters since introns have been used in many empirical studies e.g., [49-52], and their faster rates of evolution appear to make them more suitable for resolving rapid radiations than nuclear coding regions [34]. Although introns may be unsuitable for very ancient radiations, they do appear useful for the resolution of clades at a range of depths within some vertebrate classes [34,52,53]. Rates of molecular evolution obtained from empirical studies are typically expressed as substitutions per site per year instead of substitutions per site per coalescent unit. To convert coalescent units to years we assumed a constant N_e of 200,000 diploid individuals and generations of one year, so coalescent time units can be converted to years by multiplying by 400,000. These values are likely to be reasonable for a number of vertebrates based upon empirical studies (e.g, [20,54]). Although the most appropriate mutational and population genetic parameters are likely to differ among taxonomic groups, our approach could be applied to other types of loci or organisms by adjusting these parameters.

Sampling strategies

To examine the impact of sampling upon error due to the coalescent process we also simulated different numbers of individuals, or alleles in the diploid case, per species in each gene tree. We use the word "individuals" to represent individual intraspecific lineages [18]. For this part of the study, we added smaller depths and omitted depths greater than 5 because theory suggests that most genes will coalesce within 5 coalescent units [55]. To examine coalescent error specifically, we compared the true species tree with the estimated species tree from true gene trees. The set of depths we used was 0.01, 0.25, 0.5, 0.75, 1, 1.5, 2, and 5. We tested the following sampling strategies:

1.) 1 individual per species and 1 locus.

2.) 2 individuals per species and 1 locus.

3.) 5 individuals per species and 1 locus.

4.) 1 individual per species and 5 loci.

5.) 2 individuals per species and 5 loci.

6.) 5 individuals per species and 5 loci.

Finally, we added simulations with 50 loci and one individual per species for rates of 1 and 5 to address whether this increased amount of data would provide accurate estimates of phylogeny in the most problematic part of parameter space.

Phylogenetic analyses

We used STEM v1.1a [56] to find maximum likelihood (ML) estimates of species trees by reconciling gene trees. The approach implemented in STEM represents a practical and commonly used ML method for species tree estimation, making it suitable for our simulation study. The gene trees were either the true gene trees obtained directly from simulations or estimates of gene trees obtained by analyzing simulated sequence data. ML estimates of gene trees were obtained from the simulated sequence data using RAxML version 7.2.8a [57] and the GTR+Γ (-m GTRGAMMA) model of evolution and converted into ultrametric trees using penalized likelihood [58] as implemented in the ape package [59] in R. RAxML was also used to obtain the ML trees for concatenated data matrices generated when a single individual per species was sampled.

To measure accuracy of phylogenetic inference, we used the Robinson-Foulds (RF) distance, a metric based upon the number of clades that differ between two given trees [60]. Depending upon the specific analysis conducted, as many as four pairwise comparisons were conducted: 1) the true gene tree and the ML estimate (from RAxML) of the tree from the sequence data; 2) the true species tree with the ML estimate (from STEM) of the species tree based upon the true gene trees; 3) the true species tree and the ML estimate of the species tree obtained by analysis of the sequence data (using RAxML) followed by analysis of the gene trees (using STEM); and 4) the true species tree and the ML estimate of the species tree obtained by analysis of concatenated data (using RAxML). These comparisons provided information about error due to the mutational process, coalescent process, both together, and concatenation, respectively.

Results and Discussion

Relative contributions of coalescent and mutational error

Error due to coalescence increased only with the speciation rate and was not affected by depth whereas error due to the mutational process increased with both rate and depth (Figure 3). The increase in coalescent error with rate was expected, given that high speciation rates result in less time between speciation events for alleles to sort into lineages. The independence of coalescent error and depth demonstrates that the problem of coalescent error can impact the phylogeny estimation for ancient radiations [25,26] just as it can for recent radiations. The increase in mutational error with rate likely reflects the low probability that a sufficient number of mutations will accumulate along short internal branches to provide an accurate estimate of the phylogeny. On the other hand, the increase in mutational error with depth probably reflects the tendency of homoplasy to confound phylogenetic estimation.

This differential dependence on rate and depth of the mutational

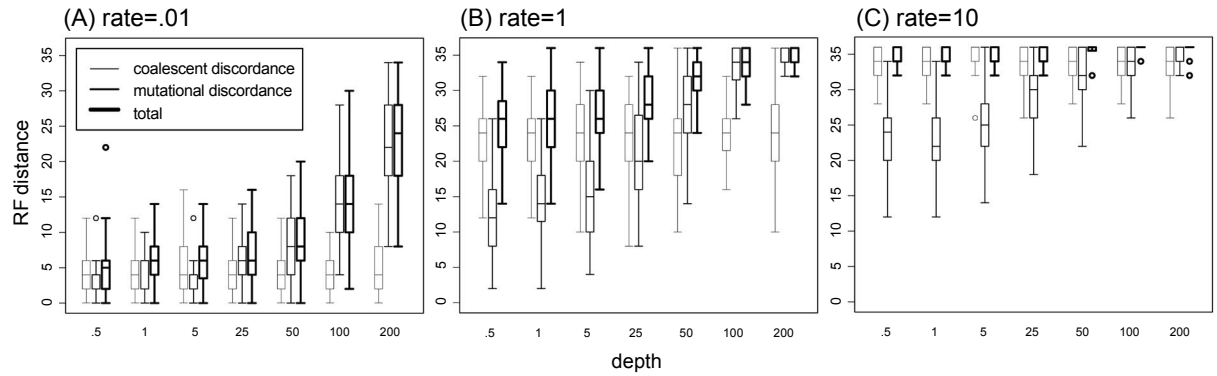

Figure 3: The impact of rate and depth upon coalescent and mutational variance: Box plots that show the effects of speciation rate and depth on the discordance of phylogenetic estimation due to coalescent and mutational variance and the total error (Figure 2). Rate and depth are given in coalescent time units. Discordance was measured using the RF distance and data reflects 100 simulated trees. The maximum discordance possible was 36.

and coalescent error led to a difference in how they contribute to the overall accuracy of phylogenetic estimation. For bushes shaped by slow speciation rates (e.g., rate=0.1) and shallow depths (e.g., depth=0.5), neither mutational nor coalescent error was large and relatively accurate estimates of phylogeny were obtained (Figure 3A). For the same rate but greater depths (e.g., depth=200), coalescent error remained negligible but mutational error increased (Figure 3A), suggesting that focusing on the latter problem could help resolution.

Regardless of depth, high speciation rates resulted in substantial error (e.g., Figure 3C; rate=10). Thus, the worst situation was a combination of a fast speciation rate with high depth (Figure 3C; e.g., rate=10, depth=200), where both coalescent and mutational error caused substantial incongruence between estimates of the species tree and the true species tree. Thus, it will be necessary to address both problems to correctly resolve relationships, and this may prove to be very difficult at the highest rates. Under these conditions, accurate gene trees were not reconstructed from sequences. Even if the gene trees had been accurate, estimating the true species tree would have remained problematic (Figure 3C). This result may reflect the limited number of genes analyzed here (see below for simulations using a larger number of genes). Overall, our results corroborated the idea that it is important to consider the depth and speciation rate when determining whether coalescent error, mutational error, or both are likely to affect phylogenetic reconstruction.

Some studies have considered the coalescent and mutational errors (as measured by RF distances) to be additive [31], but our results suggest that this is not always true. The total error was not equal to the sum of the error due to the coalescent and mutational processes in some cases (Figure 3). This means that reducing either mutational or coalescent error by a certain amount does not necessarily mean that the total error will also decrease. Thus, it is critical to consider both processes and the errors they can cause to accurately resolve phylogenies in difficult parts of parameter space.

Sampling strategy to alleviate coalescent variance

Data collection strategies are important for solving difficult phylogenetic problems. The tradeoff between sampling many loci and sampling multiple individuals has been debated for many years [38,39,61] and the best strategy appears to depend upon the bush shape [31,38]. To explore this, we focused on coalescent error for the region of bush shape space where this trade off is likely to be particularly important.

The optimal sampling strategy to overcome coalescent error depended on the rate and depth of the bush being considered. For shallow depths and fast rates, a sampling strategy that included multiple individuals rather than multiple loci was more beneficial to resolving relationships (Figure 4). In contrast, a transition occurred in the optimal sampling strategy as simulated radiations became slower and more ancient (Figure 4B; e.g., depth=0.25, rate between 2.5 and 5.0). For bushes characterized by either high depths or low speciation rates (or both), sampling additional loci rather than additional individuals per species resulted in greater accuracy (Figure 4). Thus, it appears that when resources for data collection are limited, empirical studies are likely to benefit from sampling more individuals rather than sampling multiple loci only when they are focused on a bush with a high rate and shallow depth.

We found that sampling more than one individual did not improve phylogenetic estimation at greater depths. Even by a depth of two coalescent units, increasing the number of sampled individuals resulted in limited improvement to our accuracy at great depths. This was even more pronounced at five coalescent units (Supplementary figure S2). Indeed, the RF distances for simulations that sampled either two or five individuals approach that of a single individual at these depths. This is in agreement with theory, which indicates that the expected time for two individuals within a population to coalesce is two coalescent units and that one can be 95% confident that any number of individuals within a population will have coalesced by five coalescent units [55,62,63].

The absolute time frame for the depths we are considering is surprisingly recent. Given the parameter space we examined, which we believe to be reasonable for many bushes in the vertebrate Tree of Life, two coalescent time units may reflect 800,000 years or less. Thus, sampling multiple individuals may only be beneficial with respect to overcoming coalescent error for Plio-Pleistocene radiations even if one considers the time from the beginning of the radiation. Although there may be additional reasons to sample multiple individuals (e.g., to limit the potential impact of errors in species identification; cf. [64], or to improve estimates of demographic parameters for the extant species), sampling multiple individuals does not appear to have a direct benefit for phylogenetic estimation with more ancient rapid radiations.

The utility of concatenation for rapid radiations

Two distinct approaches have been used for phylogenetic analyses of data from multiple independent loci. A common practice is to concatenate the data into a large supermatrix for a combined analysis

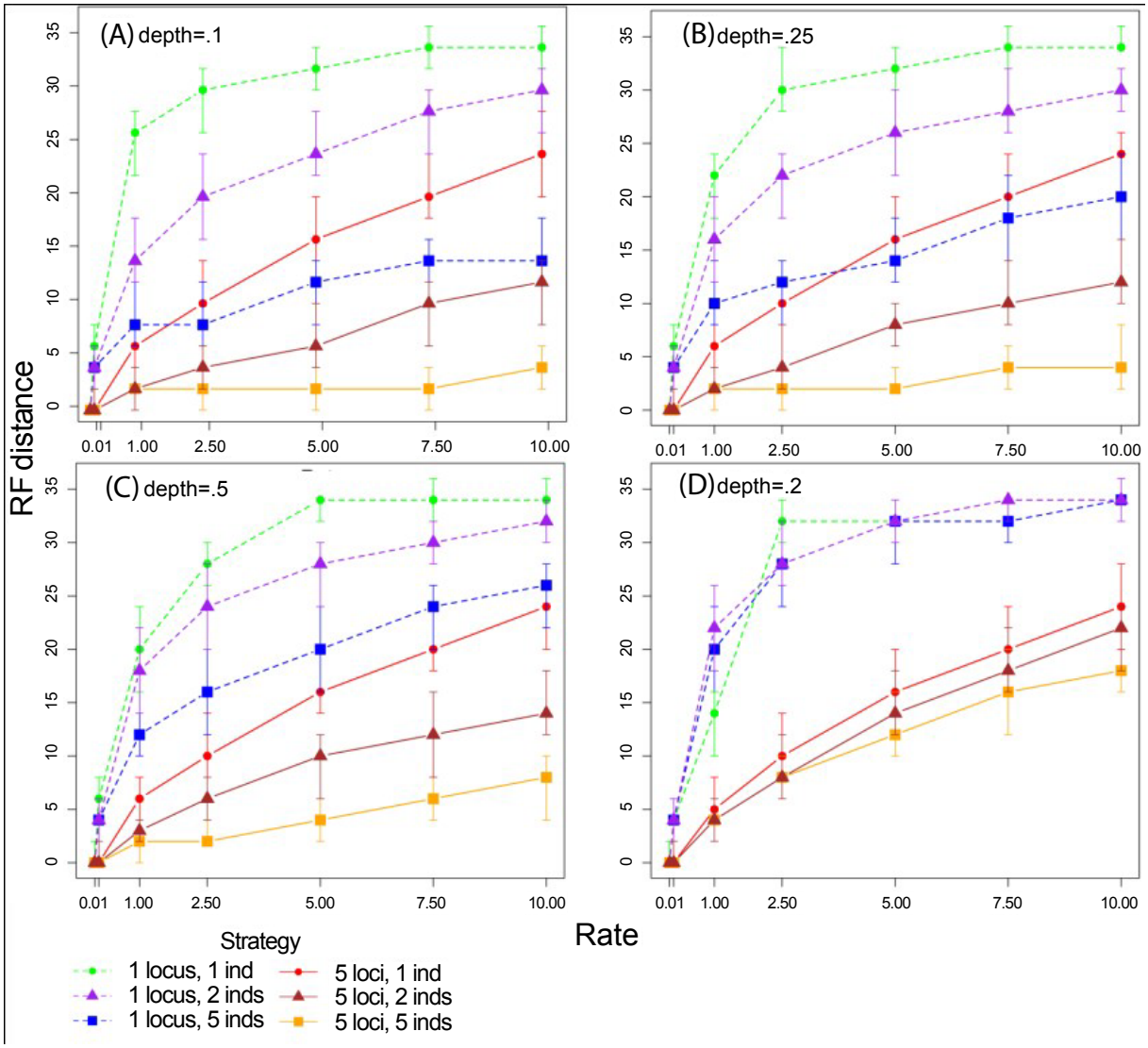

Figure 4: The impact of sampling strategy upon species tree estimation: Comparisons of sampling strategy on coalescent discordance are shown for various tree shapes. We varied the number of loci used in the analyses as well as the number of individual lineages per species to represent different sampling strategies. We then calculated coalescent discordance from our simulations.

([52,65-67]). This approach implicitly assumes that all loci have a single underlying tree topology. However, recent studies argue that concatenation can result in inaccurate estimates of relationships, sometimes with deceptively high support, when there is incongruence among gene trees [30,40,68,69]. Thus, methods of species tree estimation that allow topological differences among gene trees, such as gene tree reconciliation in a coalescent framework, are becoming more common.

The method used to analyze multiple loci had the greatest impact on the accuracy of phylogenetic estimation at slow (e.g., rate=0.01) to intermediate rates (e.g., rate=1). At shallow depths, both gene tree reconciliation (STEM with estimated gene trees) and concatenation resulted in comparable amounts of error. However, as depth increased, concatenation performed better than the gene tree reconciliation approach with gene trees estimated from sequence data (Figure 5A and 5B). Since coalescent error remained constant across depths (Figure 3), the relative advantage of concatenation at high depths likely reflects the impact of mutational error. In an empirical study of a deep radiation in iguanian lizards, Townsend et al. [70] also assert that concatenation will probably outperform estimation of a species tree using gene trees when a large number of the gene trees are poorly resolved. To address the hypothesis that the inferior performance of gene tree reconciliation in our simulations was due to mutational error at greater depths, we examined the performance of gene tree reconciliation using the true gene trees since these should exhibit no mutational error. The use of true gene trees resulted in substantially more accurate phylogenetic estimates than either the STEM analyses with estimated gene trees and concatenation (Figure 5). Moreover, the accuracy of the species tree using true gene trees remained constant across depths as expected if the differences observed above reflect mutational error.

On the other hand, at high speciation rates (e.g., rate=5), the method for analyzing multiple loci was inconsequential. Both approaches resulted in substantial error, often estimating trees that were maximally different from the true species tree (Figure 5C). At this speciation rate the evolutionary radiation could only be described as explosive, although the speciation rate was not be outside the likely

range of speciation rates for some known adaptive radiations (e.g., African cichlids; [71]). It is not surprising that resolving such a rapid radiation would prove extremely difficult. However, even for such a rapid radiation gene tree reconciliation using the true gene trees provided more accurate estimates of the species tree than concatenation or use of estimated gene trees (Figure 5C).

As depth increases, mutational error overwhelmed the coalescent error (Figure 3). In gene tree reconciliation, high mutational error led to inaccurate estimates of individual gene trees and thus of the species tree. The increased power that came from including a large number of sites in concatenated analyses appeared to reduce the impact of mutational error and compensated (at least to some degree) for the incorrect assumption that all loci have the same topology. Thus, for certain bush shapes and with the type of data simulated here, concatenation may perform better than gene tree reconciliation using estimated gene trees.

To further test the hypothesis that concatenation can be beneficial due to the increased power of sampling more sites, we analyzed 50 loci using concatenation and gene tree reconciliation. With 50 loci, concatenation performed better than gene tree reconciliation (using estimated gene trees) at all depths (Figure 6). Expectedly, the best performance came from use of the true gene trees, where even at very high speciation rates (Figure 6B) some estimated species trees matched the true species tree (RF distance = 0).

Comparing the RF distances from the analyses of 5 loci (Figure 5) and those with 50 loci (Figure 6) revealed that there is improved species tree estimation using concatenation when the number of loci is increased. In contrast, results using gene tree reconciliation were very similar when comparing 5 loci (Figure 5) and 50 loci (Figure 6), even though the amount of data analyzed is 10X greater. These results suggest that gene tree reconciliation depends heavily upon the quality of gene trees used to estimate the species tree. When there is a lot of error in the gene tree estimates, the number of gene trees becomes inconsequential. The substantial mutational error likely hindered the gene tree reconciliation approach. Thus, in the absence of accurate estimates of gene trees, our analyses indicate that concatenated analyses of increased number of loci have excellent potential to improve the resolution of difficult clades because of the increase in the overall power of these analyses.

Despite the evidence that concatenation can be inconsistent [40,68] the ML tree from a concatenated data matrix may still represent a good estimate of the phylogeny under some circumstances. Concatenation can be viewed as a type of model violation [12] and it is known that many phylogenetic methods are relatively robust to model violations

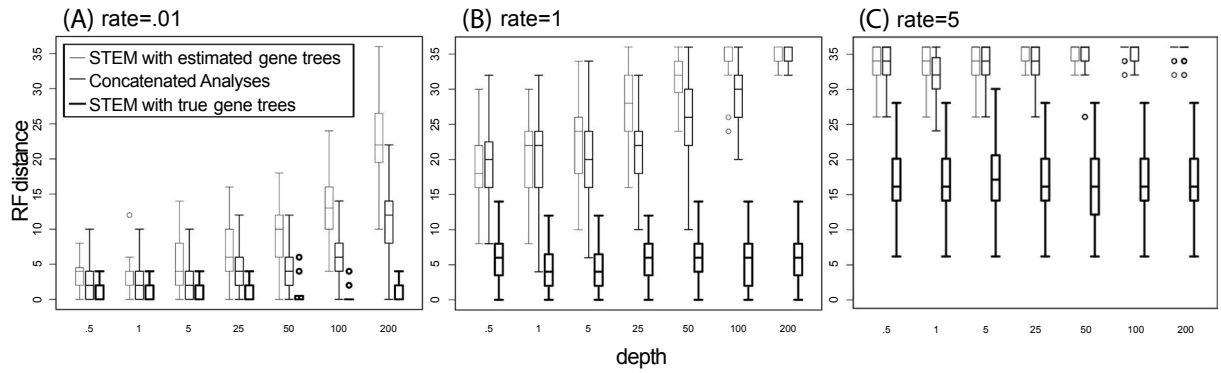

Figure 5: Comparison of concatenation and species-tree estimation: Box plots showing the effects of concatenation on the accuracy of phylogenetic estimation with 5 loci at various rates and depths. The "STEM with estimated gene trees" boxes represent the discordance between true species trees and a tree estimated using 5 gene trees that were estimated each from 1000 base pair (bp) simulated sequence data matrices. Concatenation analyses involved joining all 5 sequences into a single 5000 bp data matrix from which a tree was estimated and compared to the true gene tree. The "STEM with true gene trees" box represents the discordance between the true species tree and the estimated species tree from the true gene trees (i.e. the coalescent variance).

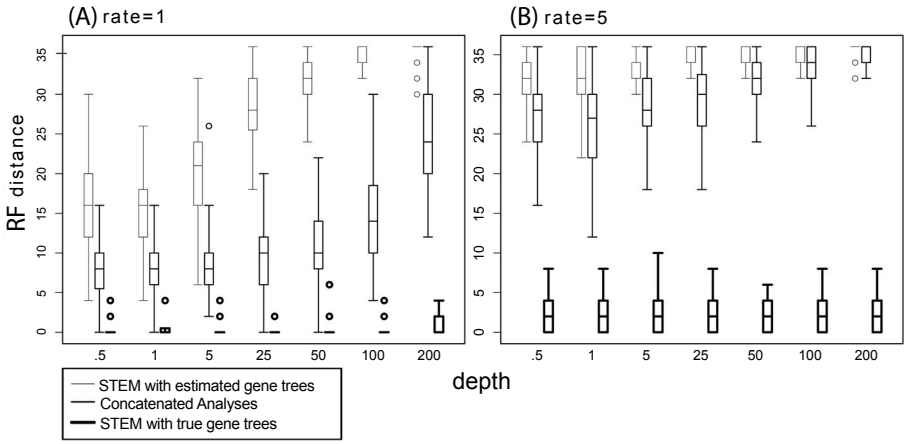

Figure 6: Comparison of concatenation and species-tree for large numbers of loci: Box plot showing the effects of increasing data to 50 loci on the accuracy of phylogenetic estimation. See Figure 5 for detailed explanation of the symbols and the results of analyses using 5 loci.

(cf. [36,72,73]). Thus, it is reasonable to postulate that analyses of concatenated data provide a useful, albeit biased, estimator of the species tree in many parts of parameter space. Although there is, as yet, no formal proof that ML analyses of concatenated data are inconsistent, simulations [68] are suggestive that analyses of concatenated data are inconsistent in the anomaly zone and further suggest that there may even be parts of parameter space where the tree favored by analyses of concatenated data is very different from the species tree. However, even in those parts of parameter space where concatenation is inconsistent, we expect the tree recovered by these analyses to be fairly close to the species tree under many circumstances. Thus, it may still be desirable to use concatenation to obtain an initial tree that can then be further rearranged to identify the optimal tree using a method that considers both the coalescent and mutational processes. This would allow the use of a computationally efficient approach (i.e., ML analysis of concatenated data) to obtain a tree topology that is fairly close to the species tree before refining that topology using a computationally difficult but consistent approach (e.g., the ML approach proposed by Maddison [10]).

The excellent performance of gene tree reconciliation when true gene trees are used suggests that it will be important to focus on ways to improve gene tree estimation. The simplest way to obtain better gene tree estimates may be to increase the sequence length of the regions analyzed [34]. Empirical studies are consistent with this hypothesis; STAR (a species tree methods) did not appear to perform as well as concatenation in an analysis of avian phylogeny [74] based upon a large number of short (<600 bp) loci but it did perform well in an analysis of mammalian phylogeny [25] based upon longer (>1000 bp) loci. However, the maximum length of regions that can be used to estimate gene trees is unclear since different parts of very long sequences may actually have distinct gene trees due to recombination or gene conversion within the individual regions. At this time, it is not clear how problematic this will be for vertebrates in practice.

An alternative to sequencing longer regions might be to focus on rare genomic changes (RGCs) to identify gene trees. Transposable element insertions are the most commonly used RGC in phylogenetics [5,75-77], although some studies have focused on other classes of RGCs such as microinversions [78,79], the presence/absence of microRNAs [80], and a subset of amino acid changes ("RGC_CAMs" [81]). The slow rate of accumulation for RGCs means they will not provide enough information to resolve gene trees completely. Instead, they are used to define specific bipartitions within gene trees. Indeed, conflict among transposable element insertions has been interpreted as *prima facie* evidence of conflict among gene trees due to lineage sorting [76,82]. However, there is also evidence that some RGCs exhibit homoplasy [79,83,84]. Indeed, several analyses [82,85] of these RGCs, in isolation from nucleotide or amino acid sequence data, have lead to conclusions that conflict with careful analyses of very large sequence datasets [52,86-88]. Nonetheless, the limited homoplasy associated with RGCs suggests that they will be useful, especially if they are combined with analyses of sequence data. Since it is clear that very accurate gene tree estimates will be very useful, even if they are challenging to obtain, identifying the best ways to obtain accurate estimates of gene trees seems critical.

Confronting theory with data

Our simulations examined an especially difficult phylogenetic problem: a rapid radiation followed by a period of time with no speciation (Figure 1A). This situation may seem extreme, but it is relevant to many known biological radiations and is of general interest for assembling the Tree of Life. Even when there is post-radiation speciation, the situation is expected to be similar to our model tree if there are long branches between the initial radiation and later speciation events. Thus, excellent examples of bushes might be found in the divergence among the three major supergroups of eutherian mammals (Boreoeutheria, Afrotheria, and Xenarthra), where analyses of transposable element insertions [5] suggest a polytomy but species tree analyses suggest an Afrotheria-Xenarthra clade [25,26]. Indeed, the Afrotheria-Xenarthra clade has been suggested to reflect an empirical example of a case where estimation of phylogeny using concatenated data is inconsistent [25,26]. However, we believe that conclusion should be approached with caution since the Afrotheria-Xenarthra clade was recovered both in some concatenated analyses [89] and in analyses using a model that accommodates gene duplication and loss but not lineage sorting [90]. Regardless, it seems reasonable to view these early divergences among eutherian supergroups as a radiation that is both relatively ancient and similar in rate to the bushes we simulated, albeit with fewer taxa.

Additional bushes that are similar to our model tree can be found in the birds. Both Notopalaeognathae (the clade comprising all extant paleognathous birds except the ostrich; Yuri et al. [91]) and Neoaves (the clade comprising the majority of extant bird; reviewed by Cracraft et al. [92]) Both of these groups include a number of highly divergent taxa characterized by long periods with no net speciation after the initial radiation, especially if we consider the subset of Neoaves designated "Metaves" by Fain and Houde [93]. Although these examples include some subsequent speciation, they are unified by the origin of a relatively large number of lineages during a short period of time followed by limited cladogenesis (and/or substantial extinction) afterward, at least in some lineages. The existence of these examples, along with examples in other lineages (e.g., iguanian lizards [70]), emphasizes the fact that the parts of parameter space we explored are relevant to important biological problems.

Future directions

Our study illuminates the ways that specific characteristics of bush shape can influence the phylogenetic error due to the coalescent and mutational process. We used empirical data to guide our simulations so the assumptions we made have both theoretical and empirical justification. However, as in all simulation studies, there are limitations in our choices and our results reflect the parts of parameter space that we chose to explore. We chose our bush shape (Figure 1) to reduce the tree characteristics to a set of two parameters. Although we have highlighted examples of empirical situations that are similar, all of our examples included some subsequent speciation. Given that subsequent speciation (and extinction) events do occur post-radiation, it would be interesting to pursue the effect of these events in future studies. Methods to characterize shifts in the rate of speciation have been developed [94] and it might be possible to use these methods to parameterize realistic model trees with shifting rates of speciation and extinction rather than our simple two-parameter bush model. Nonetheless, our bush models are likely to be informative since the most problematic parts of trees are likely to be during the rapid radiations.

Here we focused on simulations of up to 50 loci that evolve under patterns similar to the nuclear introns of birds and mammals, in part because studies using these types of markers in this number are becoming more common [29,95,96]. We recognize that the patterns of molecular evolution may differ among groups of organisms (e.g., turtles have evolved more slowly [87] and exhibit less GC-content heterogeneity [97] than birds or mammals) and types of markers (e.g., ultraconserved elements [25,74] exhibit different patterns of sequence evolution from the introns simulated herein). Furthermore, we avoided

including rate variation among lineages in order to limit the impact of bias upon our simulations. This allowed us to use a simple tree model (Figure 1) and focus on the other aspects of mutational variance. However, simulations that include rate variation among lineages (in addition to the variation among loci and among sites that we simulated) could be very interesting.

We also restricted our analyses to computationally tractable approaches that are commonly used in empirical studies. There are a number of these methods that rely upon a two step process, first estimating gene trees and later combining them to generate an estimate of the species tree (Liu et al. [98]). However, other methods exist and those methods may provide better estimates of species trees under certain conditions [31,40]. Specifically, the Bayesian MCMC approaches BEST [99] and *BEAST [100] simultaneously estimate gene trees and species trees from sequence data and avoid the two-step approach. Bayesian methods provide a straightforward means to assess parameter identifiability [101], although these methods are computationally demanding for studies with large numbers of species. A third Bayesian approach, BUCKy [102,103], does use a two-step procedure that has the potential to better accommodate uncertainty in the estimates of gene trees. Although it would be interesting to test the effectiveness of these and other methods for phylogenetic inference using the types of evolutionary radiations we explored here, we note that some simulation studies [104] have found that these much more computationally intensive methods have relatively limited increases in accuracy, at least in some parts of parameter space. Moreover, we note that STEM is a consistent estimator of the species tree when gene trees and their branch lengths are known [56]. Our focal question for this study was whether mutational error, given patterns of molecular evolution based upon empirical studies, was sufficient to degrade the performance of a representative coalescent-based gene tree reconciliation method and this does appear to be the case in specific parts of our "bush parameter space". This result leads us to suggest that the greatest benefit to improving these methods of phylogenetic estimation may come from identifying methods to improve gene tree estimates, whether those improvements reflect the joint estimation of gene trees and the species tree or two-step methods combined with other approaches to improve gene tree estimates.

Finally, it will be interesting to examine how processes such as recombination, selection, and hybridization can influence phylogenetic estimation for the bushes we considered. It seems reasonable to speculate that these processes will also exacerbate the difficulties associated with the phylogenetic inference problem. Few methods exist for making inferences when these problems are present. We note, however, that BUCKy [103] does not make assumptions regarding the source of discordance among the estimates of gene trees and that Kubatko [105] recently proposed a species tree approach that incorporates hybridization. Adding these types of complexities to simulations would provide further information about the best approaches for phylogenetic reconstruction and allow simulations of this type to be expanded beyond the focus of this study to many other parts of the Tree of Life.

The performance of phylogenetic methods has been evaluated in many ways, including mathematical analyses, simulations, and studies of "known" phylogenies (see Yuri et al. [91] for a discussion regarding the limitations of the last approach). These approaches are complementary; for example, the development of modern species tree methods was motivated in part by the proof that the anomaly zone exists [23]. Although proofs of consistency are important, phylogenetic methods should also be evaluated based upon their performance in the parts of parameter space that are most relevant to practicing systematists. These evaluations will help systematists determine appropriate approaches for phylogenetic estimation for their specific problem. Overall, we feel that it is possible to explore many of the parts of parameter space that are most relevant to practicing systematists to assess methods as they are being developed.

Conclusion

As we expand data collection to assemble the Tree of Life it is important to examine the performance of phylogenetic methods given realistic models (and parameter values) that describe the process of evolution. We demonstrated the ways coalescent and mutational error impact phylogenetic inference given bushes of different shapes and highlighted approaches that may reduce these errors and improve accuracy of phylogenetic inference. Surprisingly, we found that concatenation performed better than gene tree reconciliation for deep bushes when mutational error overwhelmed coalescent error. However, the poor performance of gene tree reconciliation appeared to be due to the use of poor gene tree estimates; using true gene trees with gene tree reconciliation always resulted in the best estimates of the species tree. Unless it is possible to obtain accurate gene trees, concatenation of many loci may provide a tractable approach to resolve difficult phylogenetic problems (albeit one that may exhibit biases for a subset of nodes under specific conditions). Regardless, the relatively good performance of concatenation in this study suggests that concatenation should continue to be compared to species tree methods in both empirical and simulation studies, at least for the time being. In the long term, however, identifying the best ways to improve gene tree estimation along with the continued development of improved approaches for species tree estimation will improve the resolution of the bushes in the Tree of Life.

Acknowledgements

We would like to thank José Miguel Ponciano, Gordon Burleigh, Tandy Warnow, members of the Braun-Kimball lab, and two anonymous reviewers for helpful discussions and comments on earlier versions of the manuscript. This research was supported by a National Science Foundation Graduate Research Fellowship to SP.

References

1. Rokas A, Carroll SB (2006) Bushes in the Tree of Life. PLoS Biol 4: 352-358.
2. Teichmann SA, Mitchison G (1999) Is there a phylogenetic signal in prokaryote proteins? J Mol Evol 49: 98-107.
3. Philippe H, Germot A, Moreira D (2000) The new phylogeny of eukaryotes. Curr Opin Genet Dev 10: 596-601.
4. Poe S, Chubb AL (2004) Birds in a bush: Five genes indicate explosive evolution of avian orders. Evolution 58: 404-415.
5. 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.
6. Whitfield JB, Lockhart PJ (2007) Deciphering ancient rapid radiations. Trends Ecol Evol 22: 258-265.
7. Rokas A, Krueger D, Carroll SB (2005) Animal evolution and the molecular signature of radiations compressed in time. Science 310: 1933-1938.
8. Verheyen E, Salzburger W, Snoeks J, Meyer A (2003) Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science 300: 325-329.
9. Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8: 616-623.
10. Maddison WP (1997) Gene trees in species trees. Syst Biol 46: 523-536.
11. Slowinski JB, Page RDM (1999) How should species phylogenies be inferred from sequence data? Syst Biol 48: 814-825.
12. Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63: 1-19.

13. Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13: 235-248.

14. Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7: 1-44.

15. Braun EL, Kimball RT (2001) Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: A comment on Walsh et al. (1999). Evolution 55: 1261-1263.

16. Mossel E (2003) On the impossibility of reconstructing ancestral data and phylogenies. J Comput Biol 10: 669-676.

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

18. Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122: 957-966.

19. Rosenberg NA (2002) The probability of topological concordance of gene trees and species trees. Theor Popul Biol 61: 225-47.

20. Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. Evolution 59: 2033-2047.

21. Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst Biol 56: 400-411.

22. Heckman KL, Mariani CL, Rasoloarison R, Yoder AD (2007) Multiple nuclear loci reveal patterns of incomplete lineage sorting and complex species history within western mouse lemurs (Microcebus). Mol Phylogenet Evol 43: 353-367.

23. Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2: e68.

24. Oliver JC (2013) Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution 67:1823-1830.

25. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, et al. (2012) Ultraconserved elements are novel phylogenetic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res 22: 746-754.

26. 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.

27. Ewing GB, Ebersberger I, Schmidt HA, von Haeseler A (2008) Rooted triple consensus and anomalous gene trees. BMC Evol Biol 8: 118-127.

28. Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RCK, et al. (2008) Phylogenomic evidence for multiple losses of flight in ratite birds. Proc Natl Acad Sci USA 105: 13462-13467.

29. Smith JV, Braun EL, Kimball RT (2013) Ratite non-monophyly: Independent evidence from 40 novel loci. Syst Biol 62: 35-49.

30. Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci USA 104: 5936-41.

31. Huang H, He Q, Kubatko LS, Knowles LL (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst Biol 59: 573-83.

32. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: The beginning of incongruence? Trends Genet 4: 225-231.

33. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol 9: e1000602.

34. Chojnowski JL, Kimball RT, Braun EL (2008) Introns outperform exons in analyses of basal avian phylogeny using clathrin heavy chain genes. Gene 410: 89-96.

35. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27: 401-410.

36. Kim J (2000) Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. Mol Phylogenet Evol 17: 58-75.

37. Katsu Y, Braun EL, Guillette LJ Jr., Iguchi T (2009) From reptilian phylogenomics to reptilian genomes: Analyses of c-Jun and DJ-1 proto-oncogenes. Cytogenet Genome Res 127: 79-93.

38. Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30.

39. McCormack JE, Huang H, Knowles LL (2009) Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. Syst Biol 58: 501-508.

40. Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: A comparison of methods. Syst Biol 60: 126-137.

41. Chung Y, Ané C (2011) Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst Biol 60: 261-275.

42. Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B Biol Sci 213: 21-87.

43. Rabosky DL (2006) LASER: A maximum likelihood toolkit for detecting temporal shifts in diversification rates. Evol Bioinform Online 2: 247-250.

44. Liu L, Yu L (2010) Phybase: an R package for species tree analysis. Bioinformatics 26: 962-963.

45. Backstrom N, Fagerberg S, Ellegren H (2008) Genomics of natural bird populations: a gene-based set of reference markers evenly spread across the avian genome. Mol Ecol 17: 964-980.

46. Kimball RT, Braun EL, Barker FK, Bowie RCK, Braun MJ, et al. (2009) A well-tested set of primers to amplify regions spread across the avian genome. Mol Phylogenet Evol 50: 654-660.

47. Spinks PQ, Thomson RC, Barkley AJ, Newman CE, Shaffer HB (2010) Testing avian, squamate, and mammalian nuclear markers for cross amplification in turtles. Conserv Genet Resour 2: 127-129.

48. Rambaut A, Grassly NC (1997) SEQ-GEN: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13: 235-238.

49. DeBry RW, Seshadri S (2001) Nuclear intron sequences for phylogenetics of closely related mammals: an example using the phylogeny of Mus. J Mammal 82: 280-288.

50. Fujita MK, Engstrom TN, Starkey DE, Shaffer HB (2004) Turtle phylogeny: Insights from a novel nuclear intron. Mol Phylogenet Evol 31: 1031-1040.

51. Kimball RT, Braun EL (2008) A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. J Avian Biol 39: 438-445.

52. Hackett SJ, Kimball RT, Reddy S, Bowie RCK, Braun EL, et al. (2008) A phylogenomic study of birds reveals their evolutionary history. Science 320: 1763-1768.

53. Matthee CA, Eick G, Willows-Munro S, Montgelard C, Pardini AT, et al. (2007) Indel evolution of mammalian introns and the utility of non-coding nuclear markers in the eutherian phylogenetics. Mol Phylogenet Evol 42: 827-837.

54. Peters JL, Zhuravlev Y, Fefelov I, Logie A, Omland KE (2007) Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (Anas spp.). Evolution 61: 1992-2006.

55. Wakeley J, Coalescent theory (2008) Roberts & Company Publishers.

56. Kubatko L, Carstens BC, Knowles LL (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25: 971-973.

57. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood–based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

58. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol Biol Evol 19: 101-109.

59. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20: 289-290.

60. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53: 131-147.

61. Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu Rev Ecol Evol Syst 34: 397-423.

62. Nichols R (2001) Gene trees and species trees are not the same. Trends Ecol Evol 7: 358-364.

63. Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24: 332-40.

64. Moore WS, DeFilippis VR (1997) The window of taxonomic resolution for phylogenies based on mitochondrial cytochrome b, in Avian Molecular Evolution and Systematics, DP Mindell, Editor, Academic Press: San Diego 84-120.

65. Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, et al. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348-51.

66. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approach to resolving incongruence in molecular phylogenies. Nature 425: 798-804.

67. Nylander J, Ronquist F, Huelsenbeck J, Nieves-Aldrey J (2004) Bayesian phylogenetic analysis of combined data. Syst Biol 53: 47-67.

68. Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol 56: 17-24.

69. Liu L, Pearl DK, Brumfield RT, Edwards SV (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62: 2080-91.

70. Townsend TM, Mulcahy DG, Noonan BP, Sites JW Jr, Kuczynski CA, et al. (2011) Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concatenated and species-tree approaches for an ancient, rapid radiation. Mol Phylogenet Evol 61: 363-380.

71. Seehausen O (2006) African cichlid fish: a model system in adaptive radiation research. Proc Biol Sci 237: 1987-1998.

72. Sullivan J, Swofford DL (2001) Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 50: 723-729.

73. Holder MT, Zwicki DJ, Dessimoz C (2008) Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Philos Trans R Soc Lond B Biol Sci 363: 4013-4021.

74. McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, et al. (2013) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS ONE 8: e54848.

75. Ray DA, Xing J, Salem AH, Batzer MA (2006) SINEs of a nearly perfect character. Syst Biol 55: 928-935.

76. Suh A, Paus M, Kiefmann M, Churakov G, Franke FA, et al. (2011) Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nat Commun 2: 443.

77. Suh A, Kriegs JO, Donnellan S, Brosius J, Schmitz J (2012) A universal method for the study of CR1 retroposons in nonmodel bird genomes. Mol Biol Evol 29: 2899-2903.

78. Chaisson MJ, Raphael BJ, Pevzner PA (2006) Microinversions in mammalian evolution. Proc Natl Acad Sci USA 103: 19824-19829.

79. Braun EL, Kimball RT, Han KL, Iuhasz-Velez NR, Bonilla AJ, et al. (2011) Homoplastic microinversions and the avian tree of life. BMC Evol Biol 11: 141-151.

80. Sperling E, Peterson K, microRNAs and metazoan phylogeny: big trees from little genes, in Animal evolution - genomes, trees, and fossils, MJ Telford and DTJ Littlewood, Editors. 2009, Oxford University Press: Oxford: 157-170.

81. Rogozin I, Wolf Y, Carmel L, Koonin EV (2007) Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol Biol Evol 24: 1080-1090.

82. Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, et al. (2012) Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. Mol Biol Evol 29: 1497-1501.

83. Han KL, Braun EL, Kimball RT, Reddy S, Bowie RCK, et al. (2011) Are transposons homoplasy free? An examination using the avian tree of life. Syst Biol 60: 375-386.

84. Rogozin I, Thomson K, Csuros M, Carmel L, Koonin EV (2008) Homoplasy in genome-wide analysis of rare amino acid replacements: The molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct 3: 7.

85. Lyson TR, Sperling E, Heimberg MA, Gauthier JA, King BL, et al. (2012) MicroRNAs support a turtle + lizard clade. Biol Lett 8: 104-107.

86. Philippe H, Brinkmann H, Copley RR, Moroz LL, Nakano H, et al. (2011) Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature 470: 255-258.

87. Shaffer HB, Minx P, Warren DE, Shedlock AM, Thomson RC, et al. (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol 14: R28.

88. Penny D (2013) Rewriting evolution--"been there, done that". Genome Biol Evol 5: 819-821.

89. Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, et al. (2007) Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci USA 104: 14395-14400.

90. Boussau B, Szollozi GJ, Duret L, Gouy M, Tannier E, et al. (2013) Genome-scale coestimation of species and gene trees. Genome Res 23: 323-330.

91. Yuri T, Kimball RT, Harshman J, Bowie RCK, Braun MJ, et al. (2013) Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. Biology 2: 419-444.

92. Cracraft J, Barker FK, Braun MJ, Harshman J, Dyke GJ, et al. (2004) Phylogenetic relationships among modern birds (Neornithes): towards an avian tree of life, in Assembling the Tree of Life, J Cracraft and M Donoghue, Editors. Oxford University Press: Oxford.

93. Fain MG, Houde P (2004) Parallel radiations in the primary clades of birds. Evolution 58: 2558-2573.

94. Alfaro ME, Santini F, Brock C, Alamillo H, Dornburg A, et al. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc Natl Acad Sci USA 106: 13410-13414.

95. Yu L, Luan PT, Jin W, Ryder OA, Chemnick LG, et al. (2011) Phylogenetic utility of nuclear introns in interfamilial relationships of Caniformia (order Carnivora). Syst Biol 60: 175-87.

96. Wang N, Braun EL, Kimball RT (2012) Testing hypotheses about the sister group of the passeriformes using an independent 30-locus data set. Mol Biol Evol 29: 737-750.

97. Chojnowski JL, Braun EL (2008) Turtle isochore structure is intermediate between amphibians and other amniotes. Integr Comp Biol 48: 454-462.

98. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV (2009) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53: 320-328.

99. Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24: 2542-2543.

100. Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27: 570-580.

101. Ponciano JM, Burleigh JG, Braun EL, Taper ML (2012) Assessing parameter identifiability in phylogenetic models using data cloning. Syst Biol 61: 955-972.

102. Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. Mol Biol Evol 24: 412-426.

103. Larget B, Kotha SK, Dewey CN, Ané C (2010) BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26: 2910-2911.

104. Yang J, Warnow T (2011) Fast and accurate methods for phylogenomic analyses. BMC Bioinformatics 9: S4.

105. Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. Syst Biol 58: 478-488.