Molecular Phylogenetics and Evolution 131 (2019) 80–92

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses

Mark P. Simmons[a,*], Daniel B. Sloan[a], Mark S. Springer[b], John Gatesy[c]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
[b] Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, CA 92521, USA
[c] Division of Vertebrate Zoology and Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY 10024, USA

ARTICLE INFO

Keywords:
Bootstrap
Branch support
Gene tree
Jackknife
MSC Tree Resampling
Summary coalescent methods

ABSTRACT

In summary ("two-step") coalescent analyses of empirical data, researchers typically apply the bootstrap to quantify branch support for clades inferred on the optimal species tree. We tested whether site-wise bootstrap analyses provide consistently more conservative support than gene-wise bootstrap analyses. We did so using data from three empirical phylogenomic studies and employed four coalescent methods (ASTRAL, MP-EST, NJst, and STAR). We demonstrate that application of site-wise bootstrapping generally resulted in gene-trees with substantial additional conflicts relative to the original data and this approach therefore cannot be relied upon to provide conservative support. Instead the site-wise bootstrap can provide high support for apparently incorrect clades. We provide a script (https://github.com/dbsloan/msc_tree_resampling) that implements gene-wise resampling, using either the bootstrap or the jackknife, for use with ASTRAL, MP-EST, NJst, and STAR. We demonstrate that the gene-wise bootstrap outperformed the site-wise bootstrap for the primary focal clades for all four coalescent methods that were applied to all three empirical studies. For summary coalescent analyses we suggest that gene-wise resampling support should be favored over gene + site or site-wise resampling when numerous genes are sampled because site-wise resampling causes substantially greater gene-tree-estimation error.

1. Introduction

Summary ("two-step") coalescent analyses infer the species (phylogenetic) tree from independently estimated gene trees rather than simultaneously inferring gene trees with the species tree. Researchers who use these methods typically apply the bootstrap (Felsenstein, 1985) to quantify branch support for clades inferred on the optimal species tree. There is heterogeneity in how bootstrapping has been implemented in these analyses (Fig. 1). Some authors apply the bootstrap within loci (i.e., site-wise bootstrap) but not among loci (e.g., Mirarab and Warnow, 2015; Linkem et al., 2016), whereas others implement the multilocus bootstrap method (Seo et al., 2005; Seo, 2008), in which bootstrap resampling is conducted at the level of both sites within genes as well as among genes (e.g., Song et al., 2012; Xi et al., 2014).

Incorporating site-wise resampling in summary coalescent analyses is fundamentally different from resampling sites in a traditional concatenation context. In concatenation, sites from all genes in the supermatrix are resampled, trees are inferred from the resampled datasets, and then the topological results from the different pseudoreplicates are summarized in a majority-rule consensus tree. By contrast, in a summary-coalescent context, sites within a particular gene are resampled, gene trees are inferred from the resampled sites, sets of bootstrapped gene trees are then used to infer species trees, and the species trees from the different pseudoreplicates are summarized in a majority-rule consensus tree. Hence there is a novel opportunity for topological error to be incorporated into coalescent resampling that is not applicable to traditional concatenation resampling. Even if a given summary coalescent method is statistically consistent when supplied with topologically accurate gene trees, the method may still be statistically inconsistent when it is applied to gene trees that have topological errors (Roch and Warnow, 2015), as may be caused by bootstrap resampling of sites within each gene.

Abbreviations: Chiari, Chiari et al. (2012); Linkem, Linkem et al. (2016); RF distance, Robinson-Foulds distance; Xi, Xi et al. (2014)

* Corresponding author at: Department of Biology, 200 West Lake Street, Colorado State University, Fort Collins, CO 80523-1878, USA.
E-mail address: psimmons@rams.colostate.edu (M.P. Simmons).

https://doi.org/10.1016/j.ympev.2018.10.001
Received 8 March 2018; Accepted 1 October 2018
Available online 02 November 2018
1055-7903/

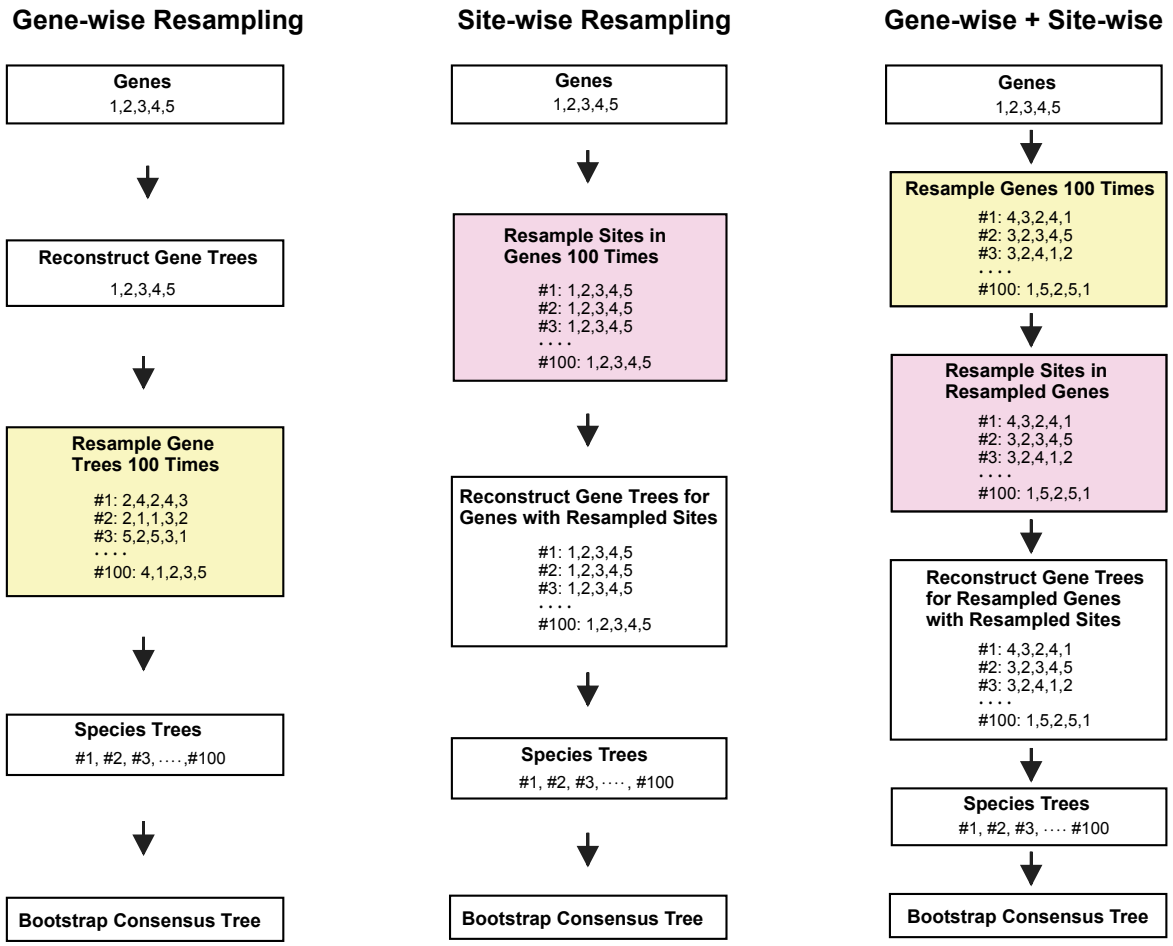

Fig. 1. Schematic illustration of the different steps that are associated with gene-wise resampling, site-wise resampling, and gene-wise + site-wise resampling (Seo, 2008) that are employed with summary coalescence methods such as ASTRAL (Mirarab et al., 2014), MP-EST, (Liu et al., 2010) and STAR (Liu et al., 2009). The examples in this figure are bootstrap analyses of a hypothetical dataset with 5 genes (1–5) and 100 pseudoreplicates. Yellow boxes indicate gene-wise resampling (genes or gene trees) and pink boxes indicate site-wise resampling. Site-wise resampling is not shown in the pink boxes; rather, site-wise resampling occurs within each of the genes that are listed in the pink boxes. Pound signs (#) correspond to individual pseudoreplicates and species trees based on these pseudoreplicates.

1.1. Seo's multilocus bootstrap

Seo et al. (2005) noted that the independently-and-identically-distributed (IID) bootstrapping assumption is violated when applied to loci ("genes") that have different histories, as in the case of incomplete lineage sorting. In such cases traditional one-step bootstrapping applied to concatenated loci can indicate high support for incorrect clades (Seo, 2008). To address this limitation of the one-step bootstrap, Seo et al. (2005) and Seo (2008) described how the two-step bootstrap (Rao and Wu, 1988) can be applied to multilocus sequence datasets when phylogenetic inference is performed using multiple genes. In their two-step bootstrap the first step is to resample genes and the second step is to resample characters within each resampled gene.

Seo et al. (2005) and Seo (2008) demonstrated the need to resample genes based on logic, simulations, and two empirical examples. Yet to our reading, they never demonstrated the importance of the second step of their procedure—resampling characters within each gene. Indeed, their primary comparison of their two-step (gene + site bootstrap) procedure was with the traditional one-step bootstrap applied to concatenated genes. As Seo (2008: 970) concluded, "By theoretical investigation, simulation studies, and empirical data analysis, we demonstrated that there is a potential problem associated with the sequence concatenation and that intergene variations should be considered during the measurement of phylogenetic uncertainty." However, they did not compare their gene + site bootstrap with a gene-wise bootstrap (i.e., sampling genes with replacement but not bootstrapping sites within each of the sampled genes; Fig. 1), which also takes into account intergene variations.

Seo's (2008) four-taxon simulations sampled genes with 501 – 3500 characters that evolved based on Jukes and Cantor's (1969) simple model with branch lengths obtained from a uniform distribution between 0 and 0.2 or 0 and 0.02. As such, these simulations represent nearly optimal conditions wherein gene-tree-estimation error would be minimal. Seo (2008) effectively used these simulations to demonstrate that his two-step bootstrap is more conservative (i.e., generally providing lower support values rather than very high support for alternative resolutions) than the one-step bootstrap. But these simulations did not represent a good test of how the two-step bootstrap would perform in the context of summary coalescence analyses, wherein there may be substantial gene-tree-estimation error (Townsend et al., 2011; Bayzid and Warnow, 2013; Betancur-R et al., 2014; Mirarab et al., 2014; Gatesy and Springer, 2014; Simmons and Gatesy, 2015; Rivers et al., 2016; Simmons et al., 2016; Springer and Gatesy, 2016; Gatesy et al., 2017; Richards et al., 2018). This gene-tree-estimation-error problem is exacerbated by employing the gene + site bootstrap.

Gene + site bootstrapping has been considered to be a more conservative method for assessing branch support in coalescent analyses given that it incorporates uncertainty at the level of both coalescent genes and sites within coalescent genes (Liu et al., 2010; Edwards, 2016). But our focus is on how the method performs when applied to

empirical datasets, for which the following two concerns apply. First, individual coalescent genes (i.e., contiguous DNA segments bounded by recombination events within the study lineage) can be as short as a single nucleotide and provide insufficient evidence with which to accurately infer gene trees—even when all characters from each coalescent gene are sampled (Hudson, 1990; Doyle, 1995). Bootstrap resampling can provide an accurate estimate of the underlying distribution from which the characters were sampled when it is applied to datasets composed of numerous characters (Efron, 1979; Felsenstein, 2004). But this condition is typically not met when sampling nuclear autosomal loci for phylogenetic analyses wherein a single functional gene may consist of many separate coalescent genes (Hobolth et al., 2011; Gatesy and Springer, 2014; Springer and Gatesy, 2016, 2018). Second, few, if any, synapomorphies can be expected from each coalescent gene when there are rapid cladogenic events or short ultraconserved elements are sampled (Gatesy and Springer, 2014; Springer and Gatesy, 2016). Even if one is fortunate enough to have a single uncontradicted synapomorphy that is shared by all descendent lineages on the intervening branch for each coalescent gene, that synapomorphy will not be sampled in ~37% of the bootstrap or jackknife (with e^{-1} deletion probability) pseudoreplicates (Farris et al., 1996).

Given the circumstances described above, it is inevitable that bootstrap pseudoreplicates with site resampling will have additional gene-tree-estimation error relative to gene trees inferred from the original sequence data. This error is exacerbated still further when the gene trees are inferred using maximum-likelihood programs that output only a single fully resolved gene tree, even when there are no informative characters, as is the case for both PhyML (Guindon et al., 2010), and RAxML (Stamatakis, 2014). The question is then: does this additional gene-tree-estimation error result in consistently conservative bootstrap support or can it instead provide unjustifiably high support, particularly for incorrect clades?

1.2. Alternatives to Seo's multilocus bootstrap

Problems caused by gene-tree-estimation error when sites within genes are resampled in bootstrap coalescent analyses include low support for true positives and high support for false positives in simulations (Bayzid et al., 2015), less accurate phylogenetic inference when bootstrap-pseudoreplicate rather than optimal gene trees are used for phylogenetic inference by MP-EST, and greater topological incongruence relative to the species tree on which the gene trees were simulated, which leads to bias (Sayyari and Mirarab, 2016). To address these problems, Bayzid et al. (2015) proposed weighted statistical binning and Sayyari and Mirarab (2016) proposed local posterior probabilities (based on gene-tree quartet frequencies for each internal branch in the inferred species tree).

Sayyari and Mirarab (2016) reported that local posterior probabilities outperformed, and were less conservative than, site-wise bootstrapping based on both simulated and empirical examples. They did note a limitation of their approach, which also applies to gene-wise resampling: it does not take into account gene-tree-estimation uncertainty, which some authors regard as essential (e.g., Edwards, 2016; Liu et al., 2017). But given that the additional gene-tree-estimation error entailed in site and gene + site bootstrapping causes a bias in coalescent analyses, we suggest that these are poor approaches to incorporate uncertainty. Instead we argue that gene-wise resampling and local posterior probabilities can more effectively account for phylogenetic uncertainty when the bias caused by resampling sites is eliminated—as long as these approaches are applied to numerous coalescent genes.

Zhang et al. (2017) proposed and tested another method to reduce the effect of gene-tree-estimation error in summary coalescent analyses: collapsing internal branches on gene trees with very low bootstrap support. They implemented their method in ASTRAL-III. Their method can be used together with both site-wise resampling as well as local posterior probabilities. But a qualifier for their method is that the bootstrap support of the optimal gene trees must be accurately estimated and some programs can indicate high support despite only having ambiguous data (e.g., Goloboff and Pol, 2005; Lemmon et al., 2009; Simmons, 2012, 2014). For likelihood implementations that only save a single optimal tree, such as GARLI (Zwickl, 2006), PhyML, and RAxML, a better choice may be to use the Shimodaira-Hasegawa-like approximate likelihood ratio test (Anisimova and Gascuel, 2006; Guindon et al., 2010; Simmons and Norton, 2013) or simply collapse internal branches that are exceptionally short (Gatesy and Springer, 2014; Giarla and Esselstyn, 2015).

1.3. Bootstrap versus jackknife resampling

Both gene-wise and gene + site resampling (Fig. 1) can be implemented using the bootstrap or the jackknife. The bootstrap involves resampling characters with replacement, whereas the jackknife involves resampling characters with a fixed deletion probability (p) for each character (Felsenstein, 1985; Farris et al., 1996). In the context of parsimony analyses, an advantage of the jackknife "… is that it simplifies the relationship between group frequency and support. Provided the data have no missing entries, the expected jackknife frequency of a group *G* set off by *r* uncontradicted characters is just $1 - p^r$" (Farris et al., 1996:114). As a result, jackknife support for a given clade is unaffected by characters that neither support nor contradict the clade in question, in contrast to the bootstrap (Carpenter, 1996; Farris et al., 1996). Furthermore, bootstrapping… "creates pseudoreplicates that comprise unobserved (non-real) character combinations, whereas [jackknifing] simply uses actual character subsets as its pseudoreplicates. The unobserved character combinations that arise in a bootstrap analysis introduce an element of arbitrary character weighting to the analysis…" (Freudenstein and Davis, 2010: 653-654). At the extreme, the arbitrary character weighting in a bootstrap analysis can produce 99% support values for properly unsupported clades that receive < 50% jackknife support. In other cases, the bootstrap can underestimate support for properly supported clades (Simmons and Freudenstein, 2011). Hence, there is reason to expect that the jackknife may outperform the bootstrap by avoiding the effects of characters that neither support nor contradict the clade in question as well as arbitrary character weighting.

Felsenstein (1985, 2004) proposed that the jackknife be applied with a 0.5 deletion probability, whereas Farris et al. (1996) proposed that the jackknife be used with the e^{-1} (~0.3679) deletion probability. Farris et al. (1996) objected to the delete-half jackknife because clades with 50% support by this resampling scheme may alternatively be unsupported by the data (i.e., absent in the strict consensus of all optimal trees; Nixon and Carpenter, 1996; Goloboff and Farris, 2001; Goloboff et al., 2003; Kopuchian and Ramírez, 2010) or unambiguously supported by a single uncontradicted synapomorphy. Subsequent studies have demonstrated that unsupported clades can be assigned up to 99% resampling values because of artifacts caused by problematic implementation of the resampling methods (Goloboff and Pol, 2005; Simmons and Freudenstein, 2011; Simmons and Goloboff, 2013). Farris et al.'s (1996) criticism of the delete-half jackknife can be obviated by simply mapping jackknife support onto the strict consensus of all equally optimal trees. Hence, beyond precedent and the connection between bootstrap and jackknife support at the limit with Hennigian (i.e., congruent, non-homoplasious) characters (Farris et al., 1996), there is no need to apply their favored e^{-1} deletion probability (Goloboff et al., 2003; Freudenstein and Davis, 2010; Kopuchian and Ramírez, 2010).

1.4. Foci of this study

In this study, we tested whether bootstrap analyses that incorporate site-wise resampling consistently provide conservative support or

instead can provide unjustifiably high support, particularly for apparently incorrect clades, in summary coalescent analyses. We did so using data from three empirical phylogenomic studies and four coalescent methods (ASTRAL, MP-EST [Liu et al., 2010], NJst [Liu and Yu, 2011], and STAR [Liu et al., 2009]). We demonstrate that application of site-wise bootstrapping generally resulted in gene-trees with substantial additional conflicts relative to gene trees inferred from the original sequence data. Due to this bias, site-wise bootstrapping cannot be relied upon to provide conservative support. Instead this approach can yield high support for apparently incorrect clades. We provide a script that implements gene-wise resampling, using either the bootstrap or the jackknife, for use with ASTRAL, MP-EST, NJst, and STAR. We demonstrate that the gene-wise bootstrap outperformed (i.e., gave more credible results) the site-wise bootstrap for the primary focal clades for all four of these coalescent methods applied to all three empirical studies. For summary-coalescent-based phylogenetic analyses, we therefore suggest that gene-wise resampling is preferable to site-wise or gene + site resampling (Fig. 1).

2. Materials and methods

2.1. Empirical studies sampled

We sampled the following three empirical phylogenomic studies wherein the original authors applied at least one summary coalescent method to analyze their data: Chiari et al. (2012; hereafter "Chiari"), Linkem et al. (2016; hereafter "Linkem"), and Xi et al. (2014; hereafter "Xi"). Chiari compiled 248 nuclear genes for 16 amniote species, with a focus on resolving the relationship of turtles (Testudines) to birds (Aves), crocodiles (Crocodylia), lizards and snakes (Squamata), and mammals (Mammalia). They concluded that their data best support the relationship (Testudines (Aves, Crocodylia)). Their MP-EST-based inference using DNA-based gene trees grouped Crocodylia with Testudines (87% bootstrap), but Chiari concluded that this is an artifact caused by saturation of some nucleotide substitutions. Despite obtaining relatively high MP-EST bootstrap support for the (Crocodylia, Testudines) clade, the difference between this topology and the (Aves, Crocodylia) topology is just 7.95 log pseudolikelihood units (−28231.93 vs. −28239.88), as reported by MP-EST ver. 1.5. MP-EST is the only coalescence method that they applied, and they calculated bootstrap support by resampling nucleotide sites while retaining all 248 of the originally sampled genes in each bootstrap pseudoreplicate. We used Chiari's bootstrapped DNA-based gene trees (100 pseudoreplicates) for our analyses.

Linkem sampled 429 loci for 16 Scincidae lizard species, with a focus on determining whether Scincinae is a monophyletic subfamily (in particular, whether *Brachymeles* is sister to the six other Scincinae species sampled or sister to the Lygosominae). They concluded that their data best support Scincinae as a clade based on their MP-EST coalescent analysis, which assigned 77% bootstrap support to the Scincinae. MP-EST is the only coalescence method that they applied, and they calculated bootstrap support by resampling sites while retaining all 429 of the originally sampled loci in each bootstrap pseudoreplicate. The difference between the two alternative resolutions of *Brachymeles* differ by 14.22 log pseudolikelihood units (−182286.71 vs. −182300.93) based on MP-EST ver. 1.5 analyses. We used Linkem's original bootstrapped gene trees (1000 pseudoreplicates) for our analyses. Chiari, Linkem, and Xi (see below) all performed a single MP-EST tree search for each of their bootstrap pseudoreplicates (F. Delsuc, C. Linkem, Z. Xi, pers. comms.).

Xi sampled 310 nuclear genes for 46 vascular-plant species, with a focus on resolving the extant sister group to the remaining extant angiosperms—*Amborella* alone or the clade of (*Amborella*, *Nuphar*). They concluded that their data best support the clade (*Amborella*, *Nuphar*) as sister to the remaining extant angiosperms based on their MP-EST and STAR coalescent analyses, which assigned 99% (MP-EST) and 97% (STAR) bootstrap support to the clade. Xi applied Seo's (2008) approach, and hence bootstrapped both sites within genes as well as the genes themselves. In a reanalysis of Xi's dataset, Mirarab and Warnow (2015) instead performed site-wise bootstrap resampling analyses and reported 100% MP-EST bootstrap support for (*Amborella*, *Nuphar*), but 75% ASTRAL bootstrap support for *Amborella* alone as sister. The high MP-EST bootstrap support for (*Amborella*, *Nuphar*) is quite surprising given that an MP-EST ver. 1.5 analysis of Xi's optimal gene trees actually supports *Amborella* alone as sister and the two topologies differ by just 4.37 log pseudolikelihood units (−944121.38 vs. −944125.74; Simmons and Gatesy, 2015; Simmons, 2017a). We used Mirarab and Warnow's (2015) site-wise bootstrapped gene trees (200 pseudoreplicates) for our analyses.

A discrepancy in the literature is that Mirarab and Warnow (2015) reported that their MP-EST analysis of Xi's 310 genes resolved (*Amborella*, *Nuphar*) as sister whereas Simmons and Gatesy (2015) reported that their MP-EST analysis of Xi's optimal gene trees resolved *Amborella* alone as sister. The cause for this discrepancy is that Mirarab and Warnow (2015) reanalyzed Xi's dataset to derive optimal gene trees whereas Simmons and Gatesy (2015) used Xi's original gene trees. The original and re-inferred gene trees differ in topology in some cases. For example, 103 of Xi's gene trees resolve the bipartition (*Selaginella*, gymnosperms, *Amborella* (*Nuphar*, all other angiosperms)) whereas only 98 of Mirarab and Warnow's (2015) gene trees do. We used RAxML ver. 8.2.1's -f e function with the GTRGAMMA model to calculate log likelihoods for both sets of gene trees. Taken across all 310 genes, Xi's trees have 22.4 higher log likelihood than Mirarab and Warnow's (2015) trees. Therefore, we used Xi's original gene trees for our analyses.

The bootstrap pseudoreplicates we used for all three empirical studies were conducted using site-wise rather than gene + site resampling. Bootstrapping genes in addition to bootstrapping characters would add greater variance among pseudoreplicates, but should have a minimal affect our reported averages across all pseudoreplicates. For example, Mirarab et al. (2016) did not find any statistically significant differences in species-tree-estimation error when they compared site-wise vs. gene + site bootstrapping and also reported obtaining generally similar support percentages. Hence our average site-wise bootstrap results, though not necessarily our confidence intervals, should generally extend to gene + site resampling.

2.2. Conflicts among gene trees

We used three complementary methods to quantify the greater conflict among inferred gene-trees in bootstrap pseudoreplicates relative to gene trees inferred from the original sequence data. First, we used the RFdistances.twoFiles.v2.py script from RF Distances Filter (https://github.com/dbsloan/rf_distances_filter; Simmons et al., 2016) to calculate pairwise Robinson-Foulds distances (RF distances; Robinson and Foulds, 1981) among all of the optimal gene trees for a dataset. For the first 100 bootstrap pseudoreplicates from a particular study, we then calculated average pairwise RF Distances among bootstrapped gene trees for each pseudoreplicate. Lower pairwise RF distances indicate greater congruence among gene trees. The decrease in accuracy caused by site-wise bootstrapping was then quantified by subtracting the pairwise RF distance (scaled from 0 to 1 following Rosenberg and Kumar, 2001) for the optimal gene trees from the average scaled RF distance among the 100 bootstrap pseudoreplicates.

Short coalescent genes may have insufficient phylogenetic signal and hence higher tree-estimation error than longer loci (Hudson, 1990; Doyle, 1995). Hence we tested for a correlation between alignment length and pairwise RF distances by assigning each gene to one of three bins (shortest third, middle third, and longest third) and calculating the average pairwise RF distance (and ± 95% confidence intervals) for genes in the shortest third and longest third bins. Our *a priori* hypothesis was that the genes in the longest bin would, on average, have significantly greater congruence (and hence lower pairwise RF distances)

Table 1
Average pairwise RF distances among genes sampled from each of the three empirical studies after application of site-wise bootstrap resampling.

Dataset	Gene number [b]	Alignment length (average) bp	Pairwise RF distances	
			Optimal trees [c]	Site-wise bootstrap trees [d]
Chiari				
all genes	248	348–1899 (7 5 4)	0.513 (± 0.015)	0.609 (± 0.002)
shortest genes	83	348–606 (5 2 6)	0.532 (± 0.026)	0.634 (± 0.019)
longest genes	84	804–1899 (1036)	0.483 (± 0.023)	0.582 (± 0.015)
Linkem				
all genes	429	338–1070 (6 4 4)	0.865 (± 0.006)	0.910 (± 0.001)
shortest genes	145	338–606 (5 4 0)	0.876 (± 0.011)	0.918 (± 0.006)
longest genes	143	688–1070 (7 5 2)	0.849 (± 0.010)	0.899 (± 0.006)
Xi[a]				
all genes	310	306–1641 (7 7 3)	0.560 (± 0.009)	0.670 (± 0.001)
shortest genes	103	306–651 (4 8 3)	0.595 (± 0.016)	0.700 (± 0.011)
longest genes	103	888–1641 (1076)	0.527 (± 0.013)	0.641 (± 0.009)

[a] Using Xi et al.'s (2014) original gene trees and Mirarab and Warnow's (2015) bootstrap trees.
[b] The reasons for different numbers of genes sampled in the shortest and longest bins for Chiari and Linkem are that fractions of genes could not be assigned to different bins and genes with the same length were not assigned to different bins.
[c] ± 95% confidence interval calculated across optimal gene trees.
[d] ± 95% confidence interval calculated across all 100 bootstrap pseudoreplicates sampled.

with other genes than would genes in the shortest bin. The rationale is that longer genes tend to have more parsimony-informative characters. The number, range, and average alignment length in each bin for each study are presented in Table 1.

Second, we used the RFdistances.twoFiles.v2.py script to check whether a series of reference clades were resolved or contradicted in each of the inferred gene trees. The reference clades were selected as those that were present in the optimal summary-coalescent tree for each study, also were supported by concatenation analysis in each study, and have been well-corroborated by independent phylogenetic studies. These uncontroversial reference clades were not the primary phylogenetic focus of the sampled studies. The five reference clades for Chiari are Aves, Crocodylia, Mammalia, Squamata, and Testudines (Fig. 2A). The four reference clades for Linkem are Lygosominae, Scincinae + Lygosominae, *Sphenomorphus* group (*Lobulia, Sphenomorphus, Tytthoscincus*), and (*Emoia*, *Lygosoma*, *Mabuya*) (Fig. 2B). Following Simmons and Gatesy (2015), the four reference clades for Xi are angiosperms, angiosperms except *Amborella* and *Nuphar*, monocots, and eudicots (Fig. 2C). The decrease in accuracy caused by site-wise bootstrapping was then quantified by comparing the average number of times that a reference clade was contradicted among the 100 bootstrap replicates to the number of optimal gene trees that contradicted the same clade.

Third, for each study, we quantified how many times that each of three alternative resolutions of the primary focal clades were resolved in the set of inferred gene trees. The three alternative resolutions differ from each other by a single nearest-neighbor-interchange swap [e.g., ((A, B)(C, D)), ((A, C)(B, D)), and ((A, D)(B, C))]. In each study, two of the alternative resolutions have been well supported by previous phylogenetic analyses and/or other results presented in the same study (e.g., concatenation-based phylogenetic trees), whereas the third alternative resolution had not. Hereafter the two resolutions that have been well supported are referred to as *plausible* and the third alternative resolution is referred to as *implausible*. Having stated that, we acknowledge that the third alternative resolution may be represented in a minority of gene trees because of lineage sorting (e.g., Maddison, 1997).

For Chiari the two plausible alternative resolutions are (Testudines (Crocodylia, Aves)) and (Aves (Crocodylia, Testudines)). The implausible resolution is (Crocodylia (Testudines, Aves)), although this resolution was recovered in a phylogenetic analysis of amino-acid sequences for 69 vertebrate mitochondrial genomes (Pollock et al., 2000) and a gene-tree-parsimony analysis of 118 nuclear genes (Cotton and Page, 2002). Note that for the position of Testudines we did not evaluate two other hypotheses that were addressed by Chiari. First, that Testudines is the sister taxon to all diapsid reptiles including archosaurs (Aves, Crocodylia) and lepidosaurs (lizards, snakes [=Squamata], tuatara). This hypothesis has been supported by cladistic studies of morphological characters (e.g., Gauthier et al., 1988), but is not supported by concatenated analyses of molecular data sets (Iwabe et al., 2005; Shen et al., 2011; Crawford et al., 2012, 2015; Field et al., 2014). Second, that Testudines is the sister taxon of lepidosaurs. The primary evidence for this hypothesis is based on Lyson et al.'s (2012) analysis of microRNAs. However, Field et al.'s (2014) analysis of an expanded microRNA presence/absence data set with more rigorous criteria for microRNA annotation provided strong support for a (Testudines, archosaurs) clade. Field et al. (2014:194) demonstrated that the apparent incongruence of Lyson et al.'s (2012) turtle-plus-lepidosaur clade with other molecular studies that support Testudines plus archosaurs (Iwabe et al., 2005; Shen et al., 2011; Chiari et al., 2012; Crawford et al., 2012, 2015) is the result of "misrecognition of primary homologies" by Lyson et al. (2012).

For Linkem the two plausible resolutions are (other Scincinae (*Brachymeles*, Lygosominae)) and (Lygosominae (*Brachymeles*, other Scincinae)). The implausible resolution is (*Brachymeles* (other Scincinae, Lygosominae)). For Xi the two plausible alternative resolutions are (*Amborella* (*Nuphar* (other angiosperms))) and ((other angiosperms) (*Amborella*, *Nuphar*)). The implausible resolution is (*Nuphar* (*Amborella* (other angiosperms))). Simmons and Gatesy (2015) demonstrated that the (gymnosperms, *Amborella*, *Nuphar*) clade, caused by Xi using a highly divergent outgroup (*Selaginella*), is a mis-rooting artifact that biased their phylogenetic inference that (*Amborella, Nuphar*) is sister to the other angiosperms. Hence, we included (gymnosperms, *Amborella*, *Nuphar*) as a fourth, artifactual clade in our analyses.

When calculating the percentage of the estimated gene trees with the plausible and implausible resolutions, only gene trees with one of the three alternative resolutions [i.e., ((A, B)(C, D)), ((A, C)(B, D)), and ((A, D)(B, C))] were considered. We compared the percentage of gene trees with the implausible (and artifactual for Xi) and plausible resolutions between the optimal gene trees with the average percentage among the 100 bootstrap pseudoreplicates.

2.3. Gene-wise resampling

In addition to running site-wise bootstrap analyses in ASTRAL, MP-EST, NJst, and STAR, we created MSC (multispecies coalescent) Tree

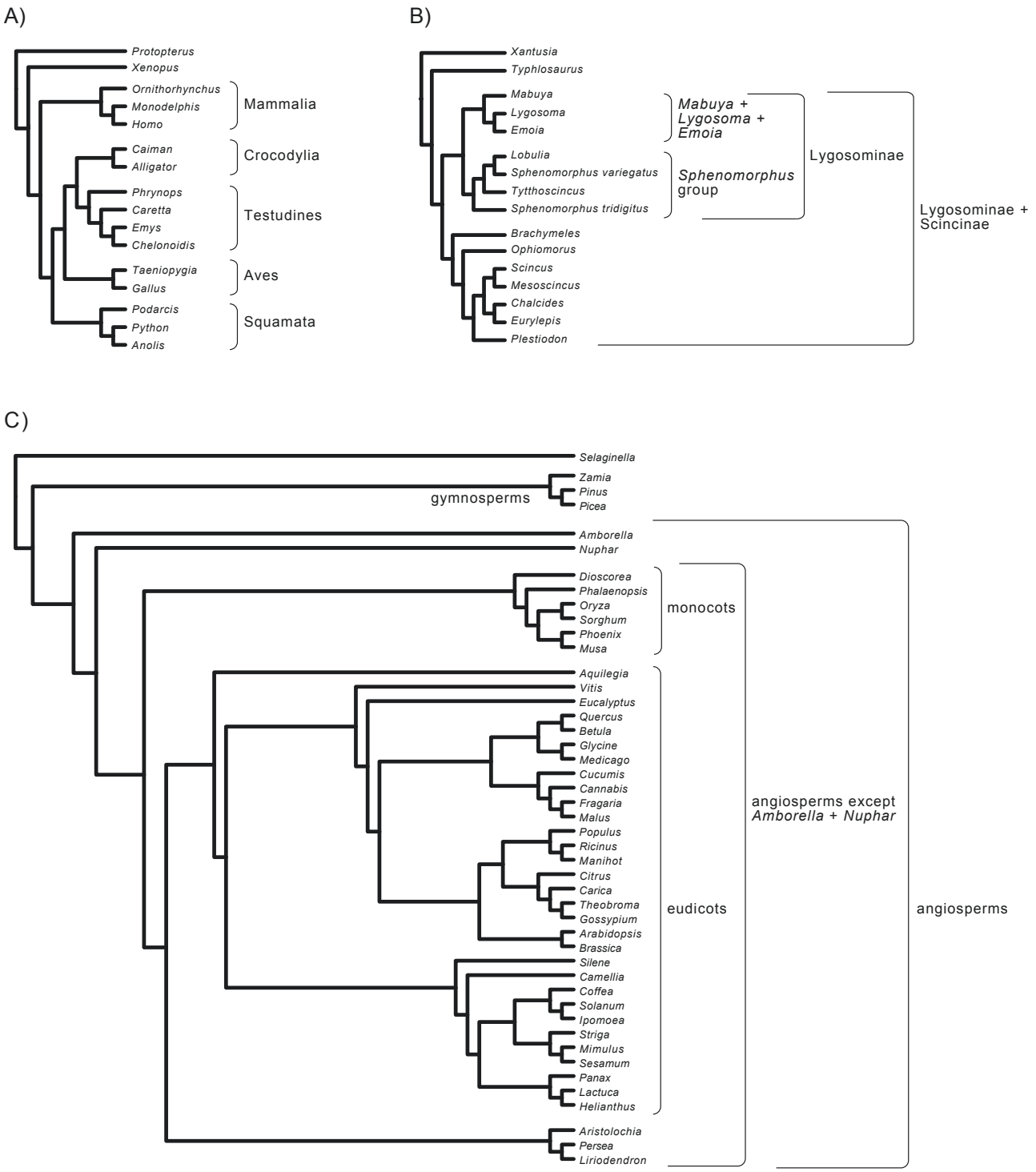

Fig. 2. Thirteen 'reference clades' that were used to assess the decay in congruence among inferred gene trees when sites within genes are bootstrapped: Chiari Amniota (A), Linkem Scincidae (B), and Xi angiosperms (C) datasets. The three topologies shown are based on MP-EST analysis of optimal gene trees derived from the original sequence data. For each dataset, the reference clades are robustly supported by both concatenation and summary coalescence methods. The gymnosperms clade is also labeled.

Resampling (https://github.com/dbsloan/msc_tree_resampling), which is a Perl script that implements gene-wise bootstrap and jackknife resampling and automates the calling of ASTRAL, MP-EST, NJst, or STAR to analyze the generated pseudoreplicates. This script allows users to set the number of pseudoreplicates, choose bootstrapping or jackknifing, and set the jackknife deletion probability applied to each gene. The input file consists of the optimal gene trees in Newick format. MSC Tree Resampling requires that the user specify the outgroup taxon for STAR analyses; it assumes that the gene trees are consistently rooted for the MP-EST analyses. MSC Tree Resampling enables the user to specify different numbers of search replicates within each pseudoreplicate for MP-EST analyses, but will only output the first tree found with the highest likelihood if multiple searches are performed.

For the three sampled datasets, we applied the gene-wise bootstrap as well as two alternative gene-wise jackknife deletion probabilities: 0.5, following Felsenstein (1985, 2004), and e^{-1} (~0.3679), following

Farris et al. (1996). For all three studies, we used the original authors' optimal gene trees. One thousand pseudoreplicates were performed for all gene-wise resampling analyses, after which extended majority-rule consensus trees were calculated using Phyutility ver. 2.2 (Smith and Dunn, 2008). Resampling support values were then mapped onto the preferred coalescent-based species tree reported by the original authors using TreeGraph 2 ver. 2.10.0–637 (Stöver and Müller, 2010).

Gene-wise bootstrapping, wherein resampling with replacement is performed, entails duplicating some gene trees in each pseudoreplicate and deleting others. Alternatively jackknifing resampling is performed without replacement, and entails deleting gene trees in each pseudoreplicate. The gene-tree duplication entailed in bootstrapping may therefore artifactually raise the average gene-tree congruence relative to the optimal gene trees. To quantify this effect we compared the average pairwise RF distance among the optimal gene trees with average pairwise RF distances in the first 100 bootstrap pseudoreplicates and the first 100 jackknife pseudoreplicates (e^{-1} deletion probability).

2.4. Summary coalescence analyses

We performed summary coalescent analyses using ASTRAL ver. 4.11.2, MP-EST ver. 1.5, and NJst and STAR using Phybase ver. 1.5. Gene-wise jackknife analyses were performed using a single MP-EST search per pseudoreplicate while bootstrap analyses were alternately performed using either one search or ten searches per pseudoreplicate in MP-EST.

Supplementary data, including gene trees from both the site-wise and gene-wise resampling pseudoreplicates, the species trees inferred for each resampling pseudoreplicate, trees with reference clades, and a Microsoft Excel file containing the raw and summarized data, are posted at: https://figshare.com/articles/Supplemental_data_for_Gene-wise_resampling_outperforms_site-wise_resampling_in_phylogenetic_coalescence_analyses_/4476188.

3. Results

3.1. Conflicts among gene trees

The average pairwise RF distances among genes before and after application of site-wise bootstrap resampling are presented in Table 1. Taken across all of the optimal gene trees, the average incongruence among Linkem's gene trees (scaled RF = 0.865, equivalent to 11.2 of 13 possible clades; all gene trees have complete taxon sampling) was far higher than that for Chiari (0.513, equivalent to 6.7 of 13 possible clades in gene trees with complete taxon sampling) or Xi (0.560, equivalent to 24.1 of 43 possible clades in gene trees with complete taxon sampling). For all three studies, the incongruence was significantly greater (non-overlapping ± 95% confidence intervals) among site-wise bootstrap trees than among the optimal gene trees. The additional incongruence among site-wise bootstrap trees averaged 0.096 (equivalent to 1.2 clades) for Chiari, 0.045 (equivalent to 0.6 clades) for Linkem, and 0.110 (equivalent to 4.7 clades) for Xi.

Pairwise gene-tree incongruence was significantly greater (non-overlapping ± 95% confidence intervals) for the gene trees based on the shortest subset of genes relative to the gene trees based on the longest subset of genes for all three studies (Table 1). The additional incongruence for the shortest subset of genes averaged 0.049 for Chiari, 0.027 for Linkem, and 0.068 for Xi.

Gene trees that conflict with reference clades (Fig. 2) from each of the three studies are presented in Table 2, with the results from Chiari highlighted in Fig. 3. When applying site-wise bootstrapping rather than using the optimal gene trees the number of conflicts increased anywhere from 6.8% to 100%. In all 13 of these cases, the number of optimal gene trees that conflicted with the reference clade was below the 95% confidence interval for the 100 bootstrap pseudoreplicates sampled.

Alternative resolutions of the primary focal clades are presented in Table 3, wherein plausible, implausible, and artifactual clades are listed. When applying site-wise bootstrapping rather than using the optimal gene trees, the number of gene trees with either of the plausible resolutions decreased substantially (by 12.1–50.1%) for all three studies. For each of the six plausible resolutions, the number of optimal gene trees that resolved the clade was above the 95% confidence interval for the 100 bootstrap pseudoreplicates sampled.

More diverse results were obtained for the implausible and artifactual clades. The implausible resolution for Chiari increased by 30.9% and the artifactual resolution for Xi increased by 52.8% when applying site-wise bootstrapping (Table 3). Alternatively, the implausible resolution for Linkem decreased by 36.4%, and the implausible resolution for Xi decreased by 40.2%.

3.2. Gene-wise resampling

The average pairwise RF distances among genes before and after application of gene-wise resampling are presented in Table 4. Although there was slightly greater gene-tree congruence among the gene-wise bootstrap pseudoreplicates than among the optimal gene trees for all three datasets, in all cases the average for the optimal gene trees (and jackknife pseudoreplicates) was within the 95% confidence interval for the 100 bootstrap pseudoreplicates sampled.

3.3. Gene-wise versus site-wise bootstrap

Gene-wise bootstrap and jackknife percentages as well as site-wise bootstrap percentages for the primary focal clades from each of the three studies are presented in Fig. 4. The complete trees are presented in Figs. S1 (Chiari), S2 (Linkem), S3 (Xi ASTRAL and MP-EST), and S4 (Xi NJst and STAR). The relevant comparison between gene-wise and site-wise resampling is the bootstrap because site-wise jackknife support was not estimated in this study.

All four coalescence methods show decreased support for the (Crocodylia, Testudines) clade when applying the gene-wise bootstrap rather than the site-wise bootstrap to Chiari's gene trees (Fig. 4A). Support for the (Crocodylia, Aves) clade increased from 55% to 62% for ASTRAL, support for the (Crocodylia, Testudines) clade decreased from 86% to 56% for MP-EST, while NJst and STAR switched from providing > 50% support for the (Crocodylia, Testudines) clade to providing > 50% support for the (Crocodylia, Aves) clade (61% to 69% for NJst, 67% to 54% for STAR).

All four coalescence methods show decreased support for the Scincinae clade when applying the gene-wise bootstrap rather than the site-wise bootstrap to Linkem's gene trees (Fig. 4B). ASTRAL changed from providing 58% support for Scincinae to providing 56% support for (*Brachymeles*, Lygosominae). Support for the Scincinae clade dropped from 68% to 57% for MP-EST and 91% to 81% for both NJst and STAR.

All four coalescence methods show decreased support for the (*Amborella*, *Nuphar*) clade when applying the gene-wise bootstrap rather than the site-wise bootstrap to Xi's gene trees (Fig. 4C). ASTRAL support for *Amborella* alone as sister to the remaining angiosperms increased from 74% to 100%. MP-EST changed from providing 98% support for the (*Amborella*, *Nuphar*) clade to providing 60% support for *Amborella* alone as sister. Support for the (*Amborella*, *Nuphar*) clade dropped from 100% to 97% by NJst and 99% to 94% by STAR.

When considering clades that were not the primary focus, there were multiple cases wherein > 15% differences were obtained using gene-wise bootstrapping rather than site-wise bootstrapping from all three studies (Figs. S1–S4). Large differences in percent support were obtained for all four coalescent methods and these large differences included clades with > 70% gene-wise and/or site-wise resampling support. Some of the more extreme cases are as follows: ASTRAL (58% site-wise vs. −81% [i.e., majority support for a contradictory clade]

Table 2
Gene trees that conflict with reference clades from each of the three empirical studies.

Dataset/clade	# optimal gene trees	# site-wise bootstrap gene trees	# applicable genes	Percentage increase for bootstrap trees
Chiari				
Aves	25	42.37 (± 0.80)	248	69.5
Crocodylia	0	0.75 (± 0.15)	23	100
Mammalia	61	90.24 (± 1.04)	248	47.9
Squamata	11	19.16 (± 0.61)	134	74.2
Testudines	37	55.23 (± 0.83)	182	49.3
Linkem				
Lygosominae	287	338.66 (± 0.98)	429	18
Scincinae + Lygosominae	314	335.46 (± 1.41)	429	6.8
Sphenomorphus group	123	186.33 (± 1.43)	429	51.5
Emoia, Lygosoma, Mabuya	220	283.74 (± 1.20)	429	29
Xi				
angiosperms	72	115.46 (± 1.23)	286	60.4
angiosperms excluding *Amborella* and *Nuphar*	144	210.65 (± 1.29)	310	46.3
monocots	57	100.88 (± 1.28)	300	77
eudicots	112	174.69 (± 1.47)	310	56

gene-wise, −78% vs. 69%, 44% vs. 92%), MP-EST (86% site-wise vs. 56% gene-wise, 48% vs. 98%, 66% vs. 90%), NJst (−45% site-wise vs. 89% gene-wise, 56% vs. −80%, 68% vs. 96%), and STAR (59% site-wise vs. 94% gene-wise, 57% vs. 87%, 78% vs. 96%). Neither gene-wise nor site-wise bootstrapping percentages were consistently higher than the other.

3.4. Gene-wise bootstrap versus gene-wise jackknife versus posterior probabilities

Gene-wise bootstrap and jackknife resampling percentages for the primary focal clades from each of the three studies generally differed by only a minor extent (Fig. 4). The percentage jackknife support calculated using the e^{-1} deletion probability was 0–7 greater than that calculated using the 0.5 deletion probability. When they differed, the jackknife support calculated using the 0.5 deletion probability was generally more similar to the bootstrap support than was jackknife support calculated using the e^{-1} deletion probability. The percentage of MP-EST bootstrap support using 10 tree searches per pseudoreplicate ranged from four lower to five greater than that generated using one tree search per pseudoreplicate.

No consistent differences in levels of support were observed when comparing ASTRAL posterior probabilities with ASTRAL site-wise or gene-wise bootstrap percentages (Figs. S1–S3).

3.5. Strongly conflicting support scores when different methods are applied

In several instances, large differences in bootstrap and jackknife percentages were recorded at primary focal clades when different summary coalescent methods were applied to the same data set (Fig. 4). For Chiari, MP-EST provided relatively high (86%) site-wise bootstrap support for (Testudines, Crocodylia) even though all other coalescent methods instead supported (Aves, Crocodylia) with gene-wise resampling support of 52% to 73%. The (Aves, Crocodylia) clade also was assigned an ASTRAL local posterior probability of 1.0 and concatenation bootstrap of 100%.

For Linkem, ASTRAL and concatenation both supported (*Brachymeles*, Lygosominae) but contrasted with MP-EST, NJst, and STAR that instead supported monophyly of Scincinae. Concatenation (100% bootstrap) strongly conflicted with NJst and STAR coalescence (site-wise bootstrap = 91% for both methods).

For Xi, analysis of optimal gene trees supported *Amborella* alone as sister to the remaining angiosperms for two coalescent methods (ASTRAL, MP-EST) as well as concatenation; bootstrap, jackknife, and posterior probabilities generally were high for ASTRAL and concatenation. Remarkably, site-wise bootstrapping of gene trees using MP-EST indicated 98% support for the conflicting resolution where (*Amborella*, *Nuphar*) is sister to the remaining angiosperms. High support for (*Amborella*, *Nuphar*) also was recorded for STAR and NJst (jackknife and bootstrap scores ranged from 93% to 100%). But when

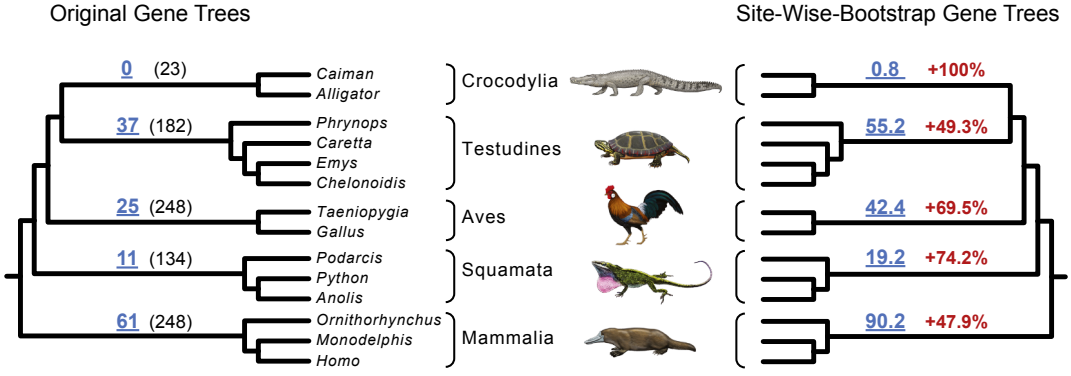

Fig. 3. Increased conflicts among inferred gene trees that emerge in site-wise bootstrap pseudoreplicates relative to optimal ML gene trees for the Chiari Amniota dataset. The number of gene trees that conflict with each of the five reference clades for the original set of 248 optimal gene trees (left) and the average number of conflicts per site-wise bootstrap pseudoreplicate (right) are shown in blue and underlined. The percent increase in conflicts for the bootstrap gene trees is indicated in red. The number in parentheses above each branch indicates the number of gene trees that include at least two representative species for that reference clade. The species tree shown is based on MP-EST analysis of optimal ML gene trees derived from the original sequence data. *Anolis* illustration is by R. Meredith; remaining illustrations are by C. Buell.

Table 3
Alternative resolutions of the primary focal clades from each of the three empirical studies.

Dataset/resolution	Optimal gene trees		Site-wise bootstrap gene trees		Percentage decrease in plausible resolutions for bootstrap trees
	Number	Percentage[c]	Number	Percentage[c]	
Chiari					
plausible: (Crocodylia, Aves)	79	51.6	60.76 (± 1.02)	46.1	23.1
plausible: (Crocodylia, Testudines)	60	39.2	52.72 (± 1.07)	40.0	12.1
implausible[a]: (Testudines, Aves)	14	9.2	18.33 (± 0.68)	13.9	
Linkem					
plausible: (*Brachymeles*, other Scincinae)	12	18.5	6.47 (± 0.48)	16.7	46.1
plausible: (*Brachymeles*, Lygosominae)	33	50.8	19.55 (± 0.70)	50.5	40.6
implausible[a]: (other Scincinae, Lygosominae)	20	30.8	12.73 (± 0.60)	32.9	
Xi					
plausible: ((*Nuphar*, *Amborella*) (other angiosperms))	28	16.2	18.82 (± 0.74)	16.9	32.8
plausible: (*Amborella* (*Nuphar* (other angiosperms)))	82	47.4	40.92 (± 0.91)	36.7	50.1
implausible[a]: (*Nuphar* (*Amborella* (other angiosperms)))	48	27.7	28.69 (± 0.87)	25.8	
artifact[b]: (gymnosperms, *Amborella*, *Nuphar*)	15	8.7	22.92 (± 1.20)	20.6	

[a] Implausible in the context of phylogenetic relationships, not necessarily gene-tree topologies in the context of lineage sorting.
[b] Clear gene-tree misrooting artifact that is also implausible in the context of lineage sorting; only applicable for the 286 gene trees that included at least one gymnosperm (Simmons and Gatesy, 2015).
[c] Percentage of the alternative three or four resolutions shown here, not the percentage across all gene trees.

Table 4
Average pairwise RF distances among genes sampled from each of the three empirical studies after application of gene-wise resampling using the bootstrap and the jackknife (e^{-1} deletion probability).

Dataset	Optimal gene trees	Gene-wise bootstrap pseudoreplicates	Gene-wise jackknife pseudoreplicates
Chiari	0.5127	0.5108 (± 0.0028)	0.5118 (± 0.0021)
Linkem	0.8650	0.8647 (± 0.0012)	0.8655 (± 0.0010)
Xi	0.5596	0.5581 (± 0.0018)	0.5602 (± 0.0013)

gene-wise bootstrapping and jackknifing was executed for MP-EST, resampling support shifted to favor *Amborella* alone as sister to remaining angiosperms (53% to 60%), which is congruent with the MP-EST tree supported by analysis of optimal gene trees (Fig. 4).

4. Discussion

4.1. Conflicts among gene trees

There was much more gene-tree topological conflict in sets of site-wise-resampling bootstrap pseudoreplicates than there was among the optimal gene trees for all three studies (Table 1). This conflict ranged from an average of 0.6 (of 13) to 4.7 (of 43) additional contradicting clades per gene tree for gene trees that were already highly incongruent (6.7 of 13 clades, 11.2 of 13 clades, 24.1 of 43 clades) when inferred using the original sequence data. For the 13 reference clades, application of the site-wise bootstrap resulted in inferred gene trees that had 6.8% to 100% more conflicts than optimal gene trees inferred from the original data (Table 2; Fig. 3). Given that gene tree reconstruction error is the bugaboo of summary coalescence methods (Townsend et al., 2011; Bayzid and Warnow, 2013; Betancur-R et al., 2014; Mirarab et al., 2014; Gatesy and Springer, 2014; Simmons and Gatesy, 2015; Rivers et al., 2016; Simmons et al., 2016; Springer and Gatesy, 2016; Gatesy et al., 2017; Richards et al., 2018), the additional conflicts that come with site-wise bootstrapping is problematic. Instead of being representative of the original sequence data, site-wise bootstrapping pseudoreplicates were strongly biased towards increased conflict among gene trees relative to optimal gene trees derived from the original data. This pattern was exacerbated in genes that are short (Table 1), which corroborates concerns that gene-tree-estimation error can be an even more severe problem for short coalescent genes (Hudson, 1990; Doyle, 1995).

Aside from a general increase in incongruence among gene trees, the plausible resolutions of primary focal clades from each of the three studies dropped by 12.1% to 50.1% among the site-wise bootstrap gene trees relative to optimal gene trees inferred from the original data matrices (Table 3). The implausible and artifactual resolutions variously dropped by up to 40.2% or increased by up to 91%. But the summed percentage of the two alternative plausible resolutions for the primary focal clades decreased for all three studies. Taken together, these results demonstrate that there is reason to be concerned with additional gene-tree-estimation error caused by site-wise resampling for all three studies.

4.2. Gene-wise bootstrap resampling

Greater average gene-tree topological congruence among gene-wise bootstrap pseudoreplicates relative to the optimal gene trees was observed (Table 4). This pattern is likely caused by resampling a given gene tree two or more times in the bootstrapping procedure that samples with replacement, but the increase in congruence relative to the optimal gene trees was negligible for all three studies. Hence the inclusion of identical gene trees in gene-wise bootstrap pseudoreplicates did not cause a substantial bias in topological congruence for the three studies.

4.3. Gene-wise versus site-wise bootstrap

Gene-wise relative to site-wise bootstrapping indicated clear shifts in support for alternative resolutions of the primary focal clades for all four coalescent methods applied to all three studies (Figs. 2–4). For Chiari all four coalescent methods provided increased support for the (Crocodylia, Aves) clade and decreased support for the (Crocodylia, Testudines) clade when using gene-wise rather than site-wise bootstrapping (Table S1). These results make sense given that the (Crocodylia, Aves) clade was resolved in 32% (79 vs. 60) more of the optimal gene trees than the (Crocodylia, Testudines) clade, but an average of just 15% (60.76 vs. 52.72) more of the site-wise bootstrap gene trees (Table 3).

Seven reasons to prefer the (Crocodylia, Aves) clade over the (Crocodylia, Testudines) clade are as follows. First, the (Crocodylia, Aves) clade was resolved in 32% more of the optimal gene trees than the (Crocodylia, Testudines) clade. Second, Chiari resolved the (Crocodylia, Aves) clade with 1.0 posterior probability in both their

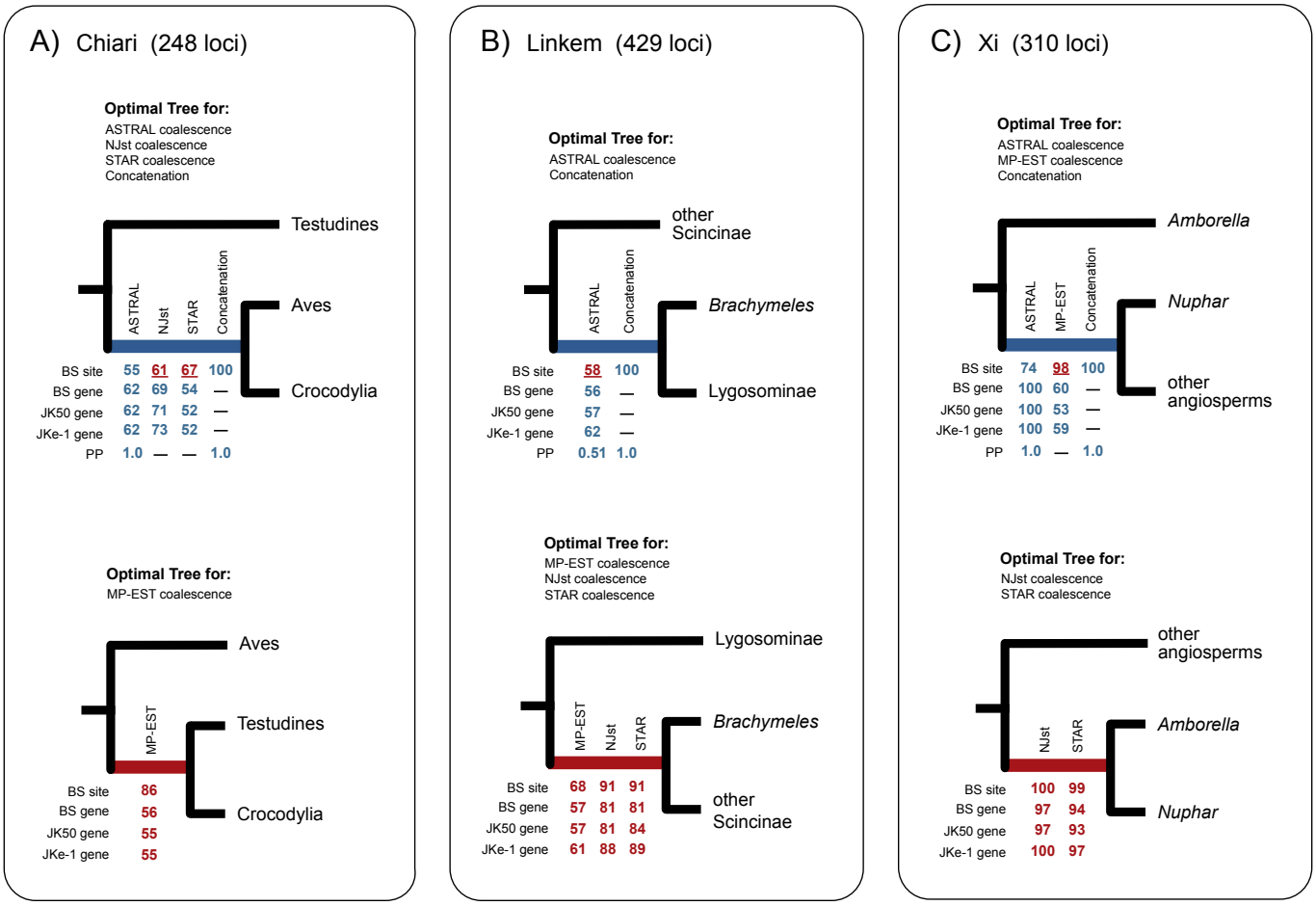

Fig. 4. Plausible resolutions of the 'primary focal clades' for Chiari Amniota (A), Linkem Scincidae (B), and Xi angiosperms (C) with bootstrap support, jackknife support, and Bayesian posterior probabilities. For each dataset, the optimal trees for five different phylogenetic methods are shown; our preferred phylogenetic hypotheses are shown above (blue branches) and alternative hypotheses supported by one or more coalescence methods are shown below (red branches). In addition to results for direct analysis of the original sequence data, support scores are at internal branches: site-wise bootstrap (BS site), gene-wise bootstrap (BS gene), jackknife with 50% probability of deletion (JK50 gene), jackknife with e^{-1} probability of deletion (JKe^{-1} gene), and Bayesian posterior probability (PP). In multiple cases, there is > 50% bootstrap support for a clade that is not supported by analysis of the original sequence data. These scores are underlined; red and blue colored font indicate which phylogenetic hypothesis is supported. Dashes indicate inapplicable values. For Chiari concatenation, BS site support is for their analysis in which their nucleotide exon characters were partitioned by codon position.

amino-acid and DNA-based concatenation analyses. Third, ASTRAL, which has generally outperformed other summary coalescent methods when applied to both simulated and empirical data (Mirarab et al., 2014; Mirarab and Warnow, 2015; Simmons and Gatesy, 2015; Meiklejohn et al., 2016 Simmons et al., 2016; Gatesy et al., 2017), resolved the same topology as the concatenation-based analysis. Fourth, Chiari resolved the (Crocodylia, Aves) clade with 99% site-wise bootstrap support in their MP-EST-based analysis of gene trees inferred using amino-acid characters. Given limited taxon sampling and the age of these lineages, it is reasonable to expect that AA characters may outperform DNA characters for this dataset (reviewed in Simmons et al., 2004; Simmons, 2017b). Sixth, the (Crocodylia, Aves) clade, Archosauria, is the traditionally recognized resolution and has been corroborated by both morphological and phylogenomic studies (e.g., Gauthier et al., 1988; Irisarri et al., 2017). Seventh, two of the gene matrices that strongly support the (Crocodylia, Testudines) clade are confounded by the inclusion of paralogs (Brown and Thomson, 2017). Taken together, we suggest that the (Crocodylia, Aves) clade is better supported than the alternative (Crocodylia, Testudines) clade for both Chiari's dataset as well as published phylogenetic evidence considered as a whole. Hence we suggest that gene-wise bootstrapping, which provided increased support for the (Crocodylia, Aves) clade and decreased support for the (Crocodylia, Testudines) clade for all four coalescent methods, outperformed site-wise bootstrapping for Chiari's dataset.

For Linkem, all four coalescent methods provided increased support for the (*Brachymeles*, Lygosominae) clade and reduced support for the Scincinae clade when using gene-wise rather than site-wise bootstrapping (Table S1). These results cannot be attributed primarily to the ratio of optimal vs. site-wise-bootstrap gene trees that support the alternative clades. The (*Brachymeles*, Lygosominae) clade was resolved in 175% (33 vs. 12) more of the optimal gene trees than the Scincinae clade and 202% (19.55 vs. 6.47) more of the site-wise bootstrap gene trees (Table 3). But given that these are just 10% (45/429) and 6% (26.02/429), respectively, of the 429 genes that Linkem sampled, the remaining gene trees have a large influence on the alternative resolutions of these lineages. In such cases, the impact of each of the gene-tree topologies on the alternative phylogenetic topologies may be quantified using partitioned coalescence support (Gatesy et al., 2017).

Three reasons to prefer the (*Brachymeles*, Lygosominae) clade over the Scincinae clade are as follows. First, the (*Brachymeles*, Lygosominae) clade was resolved in 175% more of the optimal gene trees than the Scincinae clade (Table 3). Second, Linkem resolved the (*Brachymeles*, Lygosominae) clade with 1.0 posterior probability in their concatenation analysis. Third, ASTRAL resolved the same topology as the concatenation-based analysis when optimal gene trees based on the

original sequence data were analyzed. Other relevant factors to consider are that there are no skull-morphology synapomorphies for the Scincinae (Linkem et al., 2016), Greer (1970) recognized Scincinae as ancestral to other skink subfamilies, and the Scincinae have been variously resolved as paraphyletic (Wiens et al., 2012; Lambert et al., 2015) or monophyletic (Pyron et al., 2013; Lambert et al., 2015) in previous phylogenetic analyses. Taken together, we suggest that the (*Brachymeles*, Lygosominae) clade is better supported than the alternative Scincinae clade for Linkem's dataset and hypothesize that the high (81–89%) NJst and STAR gene-wise resampling supports for the (Scincinae) clade are artifacts caused when these methods are applied to inaccurately inferred gene trees (also see Xi below). Hence we suggest that gene-wise bootstrapping, which provided increased support for the (*Brachymeles*, Lygosominae) clade and decreased support for the Scincinae clade for all four coalescent methods, outperformed site-wise bootstrapping for Linkem's dataset. Having stated that, we acknowledge that the evidence for the (*Brachymeles*, Lygosominae) clade is not as strong as that for the (Crocodylia, Aves) clade for Chiari or the (*Amborella* (*Nuphar* (other angiosperms))) topology for Xi (see below).

For Xi all four coalescent methods provided increased support for *Amborella* alone as sister to the remaining extant angiosperms and reduced support for the (*Amborella*, *Nuphar*) clade when using gene-wise rather than site-wise bootstrapping (Table S1). These results make sense given that the (*Amborella* (*Nuphar* (other angiosperms))) topology was resolved in 193% (82 vs. 28) more of the optimal gene trees than the ((*Nuphar*, *Amborella*) (other angiosperms)) topology, and an average of 117% (40.92 vs. 18.82) more of the site-wise bootstrap gene trees (Table 3). These are the two gene-tree topologies that directly support either of the two plausible phylogenetic hypotheses. But other gene-tree topologies also clearly have a large impact on the results given that, despite the (*Amborella* (*Nuphar* (other angiosperms))) topology being resolved in 117% more of the site-wise bootstrap trees, MP-EST, NJst, and STAR still provided 98%, 100%, and 99% site-wise bootstrap support for the (*Amborella*, *Nuphar*) clade (Fig. 4C).

Other relevant gene-tree topologies to consider are those wherein the clade (*Amborella*, *Nuphar*) is resolved but the topology ((*Amborella*, *Nuphar*) (other angiosperms)) is not. There are 42 such optimal gene trees and an average of 53.33 such gene trees in each site-wise bootstrap pseudoreplicate. By contrast, the clade (*Nuphar*, other angiosperms except *Amborella*) is resolved in 103 optimal gene trees and an average of just 64.87 gene trees in each site-wise bootstrap pseudoreplicate. So the number of otherwise implausible gene trees that resolve (*Amborella*, *Nuphar*) increased by 27% when applying site-wise bootstrapping whereas the number of gene trees that support *Amborella* alone as sister to the remaining angiosperms dropped by 37%. This sort of skewed gene-tree ratio, which can be explained by more severe gene-tree-inference errors when applying site-wise bootstrapping, can lead to site-wise bootstrapping underestimating support (74% ASTRAL site-wise bootstrap vs. 100% ASTRAL gene-wise bootstrap for *Amborella* alone as sister) as well as site-wise bootstrapping overestimating support (98% MP-EST site-wise bootstrap support for the (*Amborella*, *Nuphar*) clade vs. 38% MP-EST gene-wise bootstrap; Table S1).

Five reasons to prefer *Amborella* alone as sister rather than the clade (*Amborella*, *Nuphar*) as sister to the remaining angiosperms that Xi sampled are as follows. First, the (*Amborella* (*Nuphar* (other angiosperms))) topology is resolved in 193% (82 vs. 28) more gene trees than the ((*Nuphar*, *Amborella*) (other angiosperms)) topology (Table 3). Second, both Xi as well as Simmons and Gatesy (2015) resolved the (*Amborella* (*Nuphar* (other angiosperms))) topology in their concatenation analyses of all characters (see Simmons and Gatesy, 2016; Simmons, 2017a for the problems caused by Xi's tree-independent-character-subsampling analyses). Third, ASTRAL resolved the same topology as the concatenation-based analysis. Fourth, three recent phylogenomic studies that have argued for the ((*Nuphar*, *Amborella*) (other angiosperms)) topology (Goremykin et al., 2013, 2015; Xi et al., 2014) have been challenged by re-analyses and reinterpretation of the evidence (Drew et al., 2014; Simmons and Gatesy, 2015; Simmons et al., 2016; Simmons, 2017a; Zhong and Betancur-R, 2017). Fifth, multiple independent phylogenetic analyses have supported the (*Amborella* (*Nuphar* (other angiosperms))) topology (e.g., Mathews and Donoghue, 1999; Moore et al., 2007; Soltis et al., 2011; Wickett et al., 2014). Taken together, we suggest that *Amborella* alone as sister to the remaining extant angiosperms is better supported than the alternative (*Amborella*, *Nuphar*) clade for both Xi's dataset as well as published phylogenetic evidence considered as a whole. Hence we suggest that gene-wise bootstrapping, which provided increased support for *Amborella* alone as sister and decreased support for the (*Amborella*, *Nuphar*) clade for all four coalescent methods, outperformed site-wise bootstrapping for Xi's dataset.

Another relevant factor to consider is whether the site-wise bootstrap is more conservative than the gene-wise bootstrap. When considering all clades resolved in the coalescent trees from each study (Figs. S1–S4), several large differences (> 15%) between site-wise and gene-wise bootstrap support were observed for all four coalescent methods. Admittedly, this is a crude approach to quantifying large differences in support given that, for example, a shift from 10% to 40% support is not nearly as substantial as a shift from 70% to 100% support (Felsenstein, 1985; Farris et al., 1996; Simmons and Webb, 2006). But these differences were not restricted to weakly supported clades. Some of the differences were of sufficient magnitude (e.g., 48–98%, 59–94%) to change an investigator's inference about which clades are well supported by their data, and neither bootstrapping method consistently provided more conservative support than the other (Table 5).

Our results indicate that the gene-wise bootstrap outperformed the site-wise bootstrap for all 12 cases examined (three studies × four coalescent methods), site-wise bootstrapping was not consistently more conservative than gene-wise bootstrapping, and that the different resampling percentages provided by the two alternative methods can be enough to change an investigator's qualitative inference about which clades are well supported by their data. Based on these results we suggest that gene-wise resampling should be preferred over site-wise (or gene + site) resampling.

We identified some striking differences in support, for both site-wise and gene-wise resampling, among different summary coalescence methods. These differences occurred at the focal nodes of the three studies that we examined (Fig. 4). Given that all of these methods have the same underlying theoretical basis, the multispecies coalescent, these results are disconcerting. Similar strong conflicts among different coalescence methods have been noted in previous work (e.g., Gatesy et al., 2017).

Although we only compared site-wise vs. gene-wise resampling, we argue that our site-wise-resampling results are directly applicable to gene + site resampling (Seo et al., 2005; Seo, 2008). Incorporating gene resampling will slightly lower topological incongruence among genes sampled within a given bootstrap pseudoreplicate by excluding some genes and sampling others more than once (Table 4). But that does not change the fact that the gene trees inferred after application of site-resampling will, on average, both have higher topological conflict with

Table 5
Average local posterior probability (LPP), site-wise bootstrap and gene-wise bootstrap support among all clades shown in Figs. S1–S4.[a]

Dataset	ASTRAL			MP-EST		NJst		STAR	
	LPP	Site	Gene	Site	Gene	Site	Gene	Site	Gene
Chiari	96.4	92.8	93.5	95.3	94.8	97.0	97.2	95.6	92.9
Linkem	88.2	88.7	90.7	86.8	84.5	89.5	89.5	89.5	88.8
Xi	94.2	91.9	94.0	90.1	91.1	88.6	92.3	87.8	90.7

[a] Or clades with the highest contradictory bootstrap support in cases wherein the clade shown in Figs. S1–S4 was contradicted in the extended majority-rule bootstrap tree.

each other and be less accurately inferred than the optimal gene trees (Tables 1–3). Finally, Mirarab et al. (2016) did not find any statistically significant differences when they compared site-wise vs. gene + site bootstrapping.

4.4. Gene-wise bootstrap versus gene-wise jackknife

When considering resampling percentages for the primary focal clades from each of the three studies, bootstrap and jackknife using the 0.5 or e^{-1} deletion probabilities generally differed by only a small extent (Fig. 4). When they differed, the jackknife support calculated using the 0.5 deletion probability was generally more similar to the bootstrap support than was jackknife support calculated using the e^{-1} deletion probability. This result extends, in the context of summary coalescent analyses, Felsenstein's (2004) mathematical demonstration that delete-half jackknife values should better approximate bootstrap values in the context of character (or here, gene-tree) conflict. But these differences were generally minor and probably insufficient to change an investigator's qualitative interpretation of their results.

5. Conclusions

Genes, not sites, are the fundamental units of phylogenetic coalescence analyses (Maddison, 1997). When numerous genes are sampled and summary coalescent methods are applied, gene-wise resampling support should be favored over gene + site or site-wise resampling because site-wise resampling generates artifactual conflicts among gene trees (Fig. 3) that can bias support scores. As demonstrated in Sections 3.3 and 4.3, these topological artifacts in gene trees can cause resampling support in summary coalescent analyses to be alternatively over- or underestimated, often depending on which summary coalescence method the researcher applied. We expect these results to generalize to other empirical coalescent studies. But we acknowledge that our current results are based on just three studies and their generality should be further tested.

Sayyari and Mirarab (2016) introduced a method for calculating local posterior probabilities for clades in ASTRAL-II to address the extraneous conflicts among gene trees that are caused by bootstrap resampling sites. An alternative approach is to apply gene-wise resampling. In addition to excluding an artifactual source of gene-tree-estimation error, both of these methods have a time advantage given that the optimal gene-tree topologies do not need to be re-estimated to calculate branch support. This saved time can be better spent conducting more rigorous gene-tree searches (Springer and Gatesy, 2016; Gatesy et al., 2017), increasing precision of resampling percentages by analyzing more pseudoreplicates (Hedges, 1992), and conducting more thorough coalescent analyses when using MP-EST (Simmons et al., 2016; Gatesy et al., 2017).

Gene-wise bootstrapping is already available in ASTRAL-III by applying the –gene-only option (https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md#multi-locus-bootstrapping). Our MSC Tree Resampling Perl script (https://github.com/dbsloan/msc_tree_resampling) implements both bootstrap and jackknife gene-wise resampling and automates the calling of ASTRAL, MP-EST, NJst, or STAR to analyze the generated pseudoreplicates. For summary-coalescent-based phylogenetic analyses, we suggest that gene-wise resampling is preferable to gene + site or site resampling when numerous genes are sampled.

Acknowledgments

We thank Ed Braun and two anonymous reviewers for suggestions with which to improve the manuscript; Frederic Delsuc, Charles Linkem, and Siavash Mirarab for answering questions and/or sending data from the empirical studies that we re-analyzed; Pablo Goloboff for discussions about the jackknife; and Jessica Warren for help running the analyses. This research was funded by National Science Foundation grants MCB-1733227 (D.B.S) and DEB-1457735 (J.G., M.S.S.).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2018.10.001.

References

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.
Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29, 2277–2284.
Bayzid, M.S., Mirarab, S., Boussau, B., Warnow, T., 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS ONE 10, e0129183.
Betancur-R, R., Naylor, G.J.P., Ortí, G., 2014. Conserved genes, sampling error, and phylogenomics inference. Syst. Biol. 63, 257–262.
Brown, J.M., Thomson, R.C., 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66, 517–530.
Carpenter, J.M., 1996. Uninformative bootstrapping. Cladistics 12, 177–181.
Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10, 65.
Cotton, J.A., Page, R.D.M., 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. London B 269, 1555–1561.
Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K., Glenn, T.C., 2012. More than 100 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8, 783–786.
Crawford, N.G., Parham, J.F., Sellas, A.B., Faircloth, B.C., Glenn, T.C., Papenfuss, T.J., Henderson, J.B., Hansen, M.H., Simison, W.B., 2015. A phylogenomic analysis of turtles. Mol. Phylogenet. Evol. 83, 250–257.
Doyle, J.J., 1995. The irrelevance of allele tree topologies for species delimitation, and a non-topological alternative. Syst. Bot. 20, 574–588.
Drew, B.T., Ruhfel, B.R., Smith, S.A., Moore, M.J., Briggs, B.G., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., 2014. Another look at the root of the angiosperms reveals a familiar tale. Syst. Biol. 63, 368–382.
Edwards, S.V., 2016. Phylogenomic subsampling: a brief review. Zool. Scr. 45, 63–74.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.
Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates Inc., Sunderland, Mass.
Field, D.J., Gauthier, J.A., King, B.L., Pisani, D., Lyson, T.R., Peterson, K.J., 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. Evol. Dev. 16, 189–196.
Freudenstein, J.V., Davis, J.I., 2010. Branch support via resampling: an empirical study. Cladistics 26, 643–656.
Gatesy, J., Springer, M.S., 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80, 231–266.
Gatesy, J., Meredith, R.W., Janecka, J.E., Simmons, M.P., Murphy, W.J., Springer, M.S., 2017. Resolution of a concatenation/coalescence kerfuffle: partitioned coalescence support and a robust family-level tree for Mammalia. Cladistics 33, 295–332.
Gauthier, J., Kluge, A.G., Rowe, T., 1988. Amniote phylogeny and the importance of fossils. Cladistics 4, 105–209.
Giarla, T.C., Esselstyn, J.A., 2015. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. Syst. Biol. 64, 727–740.
Goloboff, P.A., Farris, J.S., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.
Goloboff, P.A., Pol, D., 2005. Parsimony and Bayesian phylogenetics. In: Albert, V.A. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 148–159.
Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.
Goremykin, V.V., Nikiforova, S.V., Biggs, P.J., Zhong, B., Delange, P., Martin, W., Woetzel, S., Atherton, R.A., McLenachan, P.A., Lockhart, P.J., 2013. The evolutionary root of flowering plants. Syst. Biol. 62, 50–61.
Goremykin, V.V., Nikiforova, S.V., Cavalieri, D., Pindo, M., Lockhart, P., 2015. The root of flowering plants and total evidence. Syst. Biol. 64, 879–891.
Greer, A.E., 1970. A subfamilial classification of scincid lizards. Bull. Mus. Comp. Zool. 139, 151–183.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Hedges, S.B., 1992. The number of replications needed for accurate estimation of the bootstrap *p* value in phylogenetic studies. Mol. Biol. Evol. 9, 366–369.
Hobolth, A., Dutheil, J.Y., Hawks, J., Schierup, M.H., Mailund, T., 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21, 349–356.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. 7, 1–44.
Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., Philippe, H., 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nature Ecol. Evol. 1, 1370–1378.
Iwabe, N., Hara, Y., Kumazawa, Y., Shibamoto, K., Saito, Y., Miyata, T., Katoh, K., 2005. Sister group relationship of turtles to the bird–crocodilian clade revealed by nuclear DNA–coded proteins. Mol. Biol. Evol. 22, 810–813.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: In: Munro, H.N. (Ed.), Mammalian Protein Metabolism, vol. 3. Academic Press, New York, pp. 21–132.
Kopuchian, C., Ramírez, M.J., 2010. Behavior of resampling methods under different weighting schemes, measures and variable resampling strengths. Cladistics 26, 86–97.
Lambert, S.M., Reeder, T.W., Wiens, J.J., 2015. When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny. Mol. Phylogenet. Evol. 82, 146–155.
Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.
Linkem, C.W., Minin, V.N., Leaché, A.D., 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). Syst. Biol. 65, 465–477.
Liu, L., Yu, L., 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60, 661–667.
Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.
Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.
Liu, L., Zhang, J., Rheindt, F.E., Lei, F., Qu, Y., Wang, Y., Zhang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F., Chen, J., Edwards, S.V., Meng, J., Wu, S., 2017. Claims of homology errors and zombie lineages do not compromise the dating of placental diversification. Proc. Natl. Acad. Sci. U.S.A. 114, E9433–E9434.
Lyson, T.R., Sperling, E.A., Heimberg, A.M., Gauthier, J.A., King, B.L., Peterson, K.J., 2012. MicroRNAs support a turtle + lizard clade. Biol. Lett. 8, 104–107.
Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.
Mathews, S., Donoghue, M.J., 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286, 947–950.
Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T., Braun, E.L., 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65, 612–627.
Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52.
Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T., 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346, 1337.
Mirarab, S., Bayzid, M.S., Warnow, T., 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65, 366–380.
Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548.
Moore, M.J., Bell, C.D., Soltis, P.S., Soltis, D.E., 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. U.S.A. 104, 19363–19368.
Nixon, K.C., Carpenter, J.M., 1996. On consensus, collapsibility, and clade concordance. Cladistics 12, 305–321.
Pollock, D.D., Eisen, J.A., Doggett, N.A., Cummings, M.P., 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol. Biol. Evol. 17, 1776–1788.
Pyron, R.A., Burbrink, F.T., Wiens, J.J., 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. 13, 93.
Rao, J.N.K., Wu, C.F.J., 1988. Resampling inference with complex survey data. J. Am. Stat. Assoc. 83, 231–241.
Richards, E.J., Brown, J.M., Barley, A.J., Chong, R.A., Thomson, R.C., 2018. Variation across mitochondrial gene trees provides evidence for systematic error: how much gene tree variation is biological? Syst. Biol. 67, 847–860.
Rivers, D.M., Darwell, C.T., Althoff, D.M., 2016. Phylogenetic analysis of RAD-seq data: examining the influence of gene genealogy conflict on analysis of concatenated data. Cladistics 32, 672–681.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Roch, S., Warnow, T., 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64, 663–676.
Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751–10756.
Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668.
Seo, T.-K., 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25, 960–971.
Seo, T.-K., Kishino, H., Thorne, J.L., 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc. Natl. Acad. Sci. USA 102, 4436–4441.
Shen, X.X., Liang, D., Wen, J.-Z., Zhang, P., 2011. Multiple genome alignments facilitate development of NPCL markers: a case study of tetrapod phylogeny focusing on the position of turtles. Mol. Biol. Evol. 28, 3237–3252.
Simmons, M.P., 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28, 208–222.
Simmons, M.P., 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80, 267–280.
Simmons, M.P., 2017a. Mutually exclusive phylogenomic inferences at the root of the angiosperms: *Amborella* is supported as sister and Observed Variability is biased. Cladistics 33, 488–512.
Simmons, M.P., 2017b. Relative benefits of amino-acid, codon, degeneracy, DNA, and purine-pyrimidine character coding for phylogenetic analyses of exons. J. Syst. Evol. 55, 85–109.
Simmons, M.P., Freudenstein, J.V., 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol. Phylogenet. Evol. 61, 177–191.
Simmons, M.P., Gatesy, J., 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol. Phylogenet. Evol. 91, 98–122.
Simmons, M.P., Gatesy, J., 2016. Biases of tree-independent-character-subsampling methods. Mol. Phylogenet. Evol. 100, 424–443.
Simmons, M.P., Goloboff, P.A., 2013. An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters. Mol. Phylogenet. Evol. 69, 265–275.
Simmons, M.P., Norton, A.P., 2013. Quantification and relative severity of inflated branch-support values generated by alternative methods: an empirical example. Mol. Phylogenet. Evol. 67, 277–296.
Simmons, M.P., Webb, C.T., 2006. Quantification of the success of phylogenetic inference in simulations. Cladistics 22, 249–255.
Simmons, M.P., Carr, T.G., O'Neill, K., 2004. Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. Mol. Phylogenet. Evol. 32, 913–926.
Simmons, M.P., Sloan, D.B., Gatesy, J., 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol. Phylogenet. Evol. 97, 76–89.
Smith, S.A., Dunn, C.W., 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716.
Soltis, D.E., Smith, S.A., Cellinese, N., Wurdack, K.J., Tank, D.C., Brockington, S.F., Refulio-Rodriguez, N.F., Walker, J.B., Moore, M.J., Carlsward, B.S., Bell, C.D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C.A., Gitzendanner, M.A., Sytsma, K.J., Qiu, Y.L., Hilu, K.W., Davis, C.C., Sanderson, M.J., Beaman, R.S., Olmstead, R.G., Judd, W.S., Donoghue, M.J., Soltis, P.S., 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Amer. J. Bot. 98, 704–730.
Song, S., Liu, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. U.S.A. 109, 14942–14947.
Springer, M.S., Gatesy, J., 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, 1–33.
Springer, M.S., Gatesy, J., 2018. Delimiting coalescence genes (c-genes) in phylogenomic data sets. Genes 9, 123.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stöver, B.C., Müller, K.F., 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinf. 11, 7.
Townsend, T.M., Mulcahy, D.G., Noonan, B.P., Sites, J.W., Kuczynski, C.A., Wiens, J.J., Reeder, T.W., 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.
Wickett, N., Mirarab, J.S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., et al., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. P. Natl. Acad. Sci. U.S.A. 111, E4859–E4868.
Wiens, J.J., Hutter, C.R., Mulcahy, D.G., Noonan, B.P., Townsend, T.M., Sites, J.W., Reeder, T.W., 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8, 1043–1046.
Xi, Z., Liu, L., Rest, J.S., Davis, C.C., 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63, 919–932.
Zhang, C., Sayyari, E., Mirarab, S., 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis, J., Nakhleh, L. (Eds.), RECOMB International Workshop on Comparative Genomics. Springer, London, pp. 53–75.
Zhong, B., Betancur-R, R., 2017. Expanded taxonomic sampling coupled with gene genealogy interrogation provides unambiguous resolution for the evolutionary root of angiosperms. Genome Biol. Evol. 9, 3154–3161.
Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation. The University of Texas at Austin.