Received: 1 April 2021 | Revised: 2 August 2021 | Accepted: 30 August 2021

DOI: 10.1111/zsc.12512

Zoologica Scripta | WILEY

ORIGINAL ARTICLE

Gene tree quality affects empirical coalescent branch length estimation

Michael Forthman[1,2] | **Edward L. Braun**[3] | **Rebecca T. Kimball**[3]

[1]Department of Entomology & Nematology, University of Florida, Gainesville, FL, USA

[2]California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food & Agriculture, Sacramento, CA, USA

[3]Department of Biology, University of Florida, Gainesville, FL, USA

Correspondence
Michael Forthman, California State Collection of Arthropods, Plant Pest Diagnostics Branch, California Department of Food & Agriculture, 3294 Meadowview Road, Sacramento, CA 95832, USA.
Email: michael.forthman@cdfa.ca.gov

Funding information
Division of Environmental Biology, Grant/Award Number: 1118823 and 1655683

Abstract

Assessing effects of gene tree error in coalescent analyses have widely ignored coalescent branch lengths (CBLs) despite their potential utility in estimating ancestral population demographics and detecting species tree anomaly zones. However, the ability of coalescent methods to obtain accurate estimates remains largely unexplored. Errors in gene trees should lead to underestimates of the true CBL, and for a given set of comparisons, longer CBLs should be more accurate. Here, we furthered our empirical understanding of how error in gene tree quality (i.e., locus informativeness and gene tree resolution) affect CBLs using four datasets comprised of ultraconserved elements (UCE) or exons for clades that exhibit wide ranges of branch lengths. For each dataset, we compared the impact of locus informativeness (assessed using number of parsimony-informative sites) and gene tree resolution on CBL estimates. Our results, in general, showed that CBLs were drastically shorter when estimates included low informative loci. Gene tree resolution also had an impact on UCE datasets, with polytomous gene trees producing longer branches than randomly resolved gene trees. However, resolution did not appear to affect CBL estimates from the more informative exon datasets. Thus, as expected, gene tree quality affects CBL estimates, though this can generally be minimized by using moderate filtering to select more informative loci and/or by allowing polytomies in gene trees. These approaches, as well as additional contributions to improve CBL estimation, should lead to CBLs that are useful for addressing evolutionary and biological questions.

KEYWORDS

coalescent branches, exons, gene trees, locus informativeness, polytomy, ultraconserved elements

1 | INTRODUCTION

With advancements in coalescent methods and availability of genome-scale datasets, it is becoming increasingly possible for investigators to accurately estimate coalescent branch lengths (CBLs) when inferring species trees. Coalescent branches lengths can provide a means to estimate historical population demographic parameters (Costa et al., 2016; Degnan & Rosenberg, 2009; Kubatko & Degnan, 2007; Weber et al., 2014), as has been done in, for example, hominoid (Rannala & Yang, 2003; Yang, 2002) and avian (Ericson et al., 2019; Houde et al., 2020; Jennings & Edwards, 2005) studies. Such information can give insights into modes of speciation and selection. Furthermore, under the multispecies coalescent framework, empirical tests to detect if parts of a species tree are

likely in the anomaly zone (Degnan & Rosenberg, 2006) not only rely on accurate topological estimates but also CBLs (Cloutier et al., 2019; Léveillé-Bourret et al., 2020; Linkem et al., 2016; Springer & Gatesy, 2019). While CBLs can be estimated from, for example, Bayesian species tree co-estimation approaches (Flouri et al., 2020; Shi & Yang, 2018; Thawornwattana et al., 2018), these methods are often computationally intensive and impractical for large phylogenomic datasets comprised of thousands of loci from several tens to hundreds of taxa. In contrast, some summary coalescent approaches, which estimate gene trees and then combine those gene trees to yield an estimate of the species tree (Liu et al., 2010; Mirarab et al., 2016; Mirarab, Reaz, et al., 2014), are practical for phylogenomic datasets. Although some summary coalescent methods estimate CBLs in species trees, most studies using summary methods have largely focused on topology rather than CBLs. Thus, while CBLs have promising utility in historical population genetic studies and species tree evaluations, the ability of summary coalescent methods to obtain accurate estimates remains largely unexplored.

Under multispecies coalescent theory, the length of coalescent branches in the species tree reflects the amount of genuine conflict among gene trees due to incomplete lineage sorting (Degnan & Rosenberg, 2009; Liu et al., 2009; Mirarab et al., 2016; Sayyari & Mirarab, 2016). However, gene tree estimation error can be misinterpreted as genuine discordance in summary coalescent methods; truly random errors in gene trees might not have an impact on the estimate of topology, but they would be expected to reduce estimated species tree branch lengths (Huang & Knowles, 2009; Mirarab et al., 2016; Sayyari & Mirarab, 2016; Yang, 2002). Several simulation and empirical studies have shown that gene trees estimated using short sequence lengths (a proxy for locus informativeness) can result in high levels of gene tree estimation error (Bayzid & Warnow, 2013; Betancur-R et al., 2014; Molloy & Warnow, 2018), and this in turn can underestimate CBLs in the species tree (Mirarab, Bayzid, et al., 2014; Sayyari et al., 2017; Sayyari & Mirarab, 2016).

Recently, Peng et al. (2021) made the same point regarding the impact of gene tree estimation error on CBL estimation. Specifically, they pointed out that while the most common method for CBL estimation in this framework (Sayyari & Mirarab, 2016) is statistically consistent, it also makes a strong assumption: that the gene trees that are summarized represent an unbiased sample of true gene trees. This problem with gene tree summary methods motivated Peng et al. (2021) to develop a new method of CBL estimation. However, their method is currently limited to the Jukes–Cantor model (Jukes & Cantor, 1969) of sequence evolution. We would like to address the problem of CBL estimation in a different way by asking whether there are reasonable empirical guidelines that allow practising phylogeneticists to obtain useful CBLs within the gene tree summary framework. One approach might be to limit species tree inferences to the most informative loci, which should be subject to less gene tree estimation error and will therefore yield longer, more accurate CBL estimates (as long as the sample of "informative gene trees" remains large enough).

Gene tree resolution should also affect CBL estimation given it is known to have an impact on species tree topological inferences (Zhang et al., 2017; but see Blom et al., 2017). Loci with limited variation to resolve all nodes in a phylogeny will result in polytomous estimates of gene trees whereas more variable loci are likely to produce better resolved estimates of gene trees (McCormack et al., 2012; Zhang et al., 2017). Resolving these polytomies in a manner that is independent of the data (often arbitrary) can introduce error by producing artefactual relationships, which in turn should yield underestimated species tree CBLs because those incorrectly resolved gene trees will be misinterpreted as discordance (Gatesy & Springer, 2014; Simmons & Kessenich, 2020; Springer & Gatesy, 2014). Thus, rather than introducing artefactual relationships due to arbitrary resolution of polytomous gene trees, using polytomous estimates of gene trees when conducting species tree analyses may also improve gene tree quality of less informative loci (Zhang et al., 2017). Despite these recent studies highlighting how to improve topological estimates in summary methods, we still know little about the limits of currently available coalescent methods to accurately estimate CBLs for biological and analytical purposes. More importantly, we do not know the degree to which factors like gene tree error and resolution affect CBL estimation, and thus how to limit CBL underestimation across various types of molecular data.

2 | PREDICTIONS

We make several predictions regarding the estimation of CBLs from gene trees. If the number of estimated gene trees is sufficient to overcome sampling error, CBLs will always be underestimated if the gene trees exhibit unbiased errors. By *unbiased error*, we mean that there is no tendency for a specific incorrect topology to be overrepresented (in contrast to systematic error, such as genome-wide long-branch attraction). Since there are more ways to be discordant with a species tree than concordant, unbiased error is more likely to result in topologies discordant with the species tree, and this discordance reduces CBL estimates. Given that we expect a negative relationship between the information content of the loci used to generate gene trees and error in gene tree estimates, accurate

estimation of CBLs will be positively related to locus informativeness. Finally, the shape of the tree should have an impact on the accuracy of the gene trees, with long gene tree branches (in this case, measured in expected substitutions per site, not coalescent units) typically being recovered accurately and short gene tree branches being associated with more error (even when using high informative loci). These properties lead to our predictions shown in Figure 1, where the estimated CBLs for the shortest and longest branches in the species tree will be fairly accurate, and intermediate branch lengths will be less accurate. Because uninformative loci are expected to have greater gene tree estimation error (which will be misinterpreted as discordance for the purposes of coalescent branch length calculations), the height of the curve should be greater for less informative loci (Figure 1, blue line) than is expected for more informative loci (Figure 1, brown line).

Although the use of more informative loci in analyses is one way to improve the accuracy of gene trees, another approach is to collapse branches not supported by data (i.e., zero-length branches) to yield polytomies. Many methods of phylogenetic inference that yield completely resolved trees always resolve zero-length branches in an arbitrary manner, which should have a negative impact on the CBL estimate. Resolution of zero-length branches in gene trees should primarily occur on short branches in the species tree and when using less informative loci. Thus, we predict that allowing polytomies will reduce the maximum of the curve in Figure 1, similar to the use of more informative loci.

Our predictions (Figure 1) represent a conceptual model; we expect the exact shape of the curve for empirical datasets to depend on the details of the tree topology and the patterns of sequence evolution for the loci used to generate the gene trees. Moreover, the pattern shown in Figure 1 cannot be tested in a simple and direct manner since the true CBLs are unknown. However, it is possible to compare CBL estimates from various conditions to a reference set of CBLs. Regardless of the CBLs chosen as the reference set, the difference in CBL values (Δ_{CBL}) will typically be positive if the reference set tends to underestimate CBLs (i.e., if the gene tree quality is poor) and negative if the opposite is true. Using this approach, we address how gene tree quality—as a function of locus informativeness and gene tree resolution—affects CBL estimates. We analyse four empirical vertebrate datasets using either ultraconserved element (UCE) loci or exons and ask whether we can find evidence that error in CBL estimates follows the predicted pattern (Figure 1) when the Δ_{CBL} values are plotted against a set of reference CBLs.

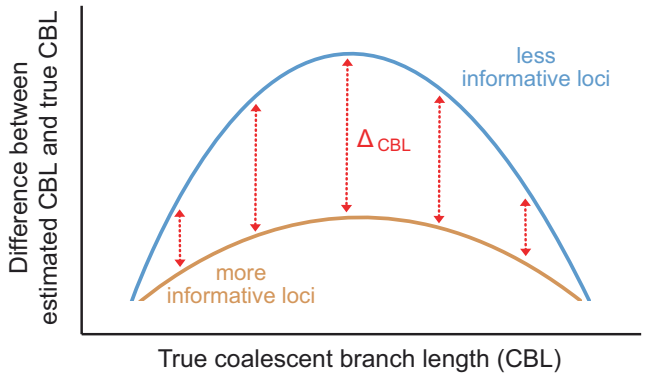

FIGURE 1 Predicted trends in differences between estimated and true coalescent branch lengths (CBLs) (Δ_{CBL}) when comparing less and more informative loci

3 | MATERIALS AND METHODS

3.1 | Empirical datasets

We selected the following four vertebrate datasets that vary in data type and taxonomic group (Table 1): UCE alignments for the "ocellated clade" (Chen et al., 2018) and Laurasiatheria (Esselstyn et al., 2017) taxa and nuclear protein-coding (exons) loci for oscine passerines (Prum et al., 2015) and Euarchontoglires (Song et al., 2012) taxa. We selected these datasets because they include types of loci commonly used in empirical phylogenomic analyses, as well as to have representation of different taxonomic groups with differing evolutionary rates. We did not select datasets based on the total number of loci sampled or average parsimony-informative sites per branch. As such, our selection criteria provided us with the ability to investigate the impacts of our variables on CBL estimation across a range of empirical datasets rather than just the "best" or "worst" datasets.

We subsampled taxa in each dataset to include clades with a variety of branch lengths. Taxa were excluded based on duplicate higher-level representation (e.g., to reduce the number of taxa within a family) and/or relatively low locus availability.

3.2 | Gene tree estimation and summary coalescent species tree inference

We recognize that some phylogenetic programs that produce fully resolved gene trees may not always resolve zero-length branches randomly but in a deterministic manner (Goloboff & Simmons, 2014). However, we use the phrase "randomly resolved" to indicate that zero-length branches are resolved by the phylogenetic program in some manner other than changes in the data analysed.

TABLE 1 Summary data of the four subsampled vertebrate datasets analysed in this study

Dataset	Data type	No. of taxa	No. of loci	Mean locus length	Mean locus informativeness	Min locus informativeness	Max locus informativeness
Ocellated clade	UCE	25	2,007	511.56	21.92	0	131
Laurasiatheria	UCE	38	1,101	583.80	115.45	8	277
Oscine passerines	Exon	25	231	1,372.62	195.61	7	509
Euarchontoglires	Exon	25	413	3,128.16	878.60	107	9,302

For each locus, we used RAxML v8.2.8 (Stamatakis, 2014) to perform 100 maximum likelihood (ML) search replicates with random starting trees and the GTR+Γ model of sequence evolution to generate optimal gene trees with polytomies randomly resolved (hereafter referred to as RAX gene trees). We also generated two sets of optimal gene trees in GARLI v2.01 (Zwickl, 2006) by performing 20 ML searches using one of the 24 MrBayes models selected by MrAIC v1.4.6 (Nylander, 2004) using PhyML v3.1 (Guindon et al., 2010) and the Akaike information criterion with a correction for small sample sizes (Hurvich & Tsai, 1989). Optimal GARLI gene trees could either have zero-length branches collapsed (collapsebranches = 1; GARpoly) or randomly resolved (collapsebranches = 0; GARres). For each gene tree resolution strategy, we then subsampled all gene trees based on locus informativeness using a direct measure (i.e., the number of parsimony-informative sites): 5% (UCE datasets only), 10%, and 25% most parsimony-informative loci and 75% least parsimony-informative loci (leading to analysis of 12 sets of gene trees for each of the UCE datasets and 10 for each of the exon datasets). Although there are other methods for locus filtering (e.g., filtering for "question-specific" loci; Chen et al., 2015), we feel that focusing on numbers of parsimony-informative sites make our approach very general because it is independent of specific clades. We measured the extent to which GARpoly branches were collapsed by calculating symmetric differences between a completely unresolved tree and GARpoly gene trees in PAUP* v4.0a152 (Swofford, 2003).

Species tree inferences were performed using the summary coalescent program ASTRAL-II v4.10.5 (Mirarab, Reaz, et al., 2014; Mirarab & Warnow, 2015). ASTRAL-II finds the species tree with the maximum number of shared quartet trees for a given set of gene trees, estimates internal branch lengths in coalescent units (terminal branch lengths are not estimated by this program) and allows unrooted gene trees to have unresolved branches (Mirarab, Reaz, et al., 2014; Mirarab & Warnow, 2015; also see Zhang et al., 2017). First, to get the best estimate of the species tree for each taxonomic group, we first inferred species trees from the 25% most informative GARpoly optimal gene trees, with nodal support measured using local posterior probabilities (LPP) (Sayyari & Mirarab, 2016). We selected the 25% most informative optimal gene trees because (a) this informative threshold has been shown to improve or recover similar topological support compared to estimates based on datasets with more uninformative or informative gene trees, respectively (Hosner et al., 2016; see also Meiklejohn et al., 2016), and (b) Hosner et al. (2016) showed that results from the 25% most informative gene trees were largely congruent with other inference methods, for example SVD quartets and concatenation, with few differences recovered compared to higher informativeness thresholds; we used GARpoly optimal gene trees because species tree topology has been shown to improve with the use of polytomous gene trees in summary coalescent methods (Zhang et al., 2017). To control for topological differences, we then used the 25% most informative GARpoly species tree as a constrained topology to estimate branch lengths and LPP support for all other sets of gene trees for each taxonomic group using the -q option in ASTRAL-II. To determine whether the use of a single topology introduced any biases, we also compared overall species tree lengths, which were calculated by summing the internal CBLs for all optimal ASTRAL trees.

3.3 | Evaluation of gene tree quality on CBLs

Differences in coalescent branch length estimates among our various gene tree sets and the reference set were $\Delta_{CBL} = CBL_{est} - CBL_{ref}$, where the CBL_{est} values are those for all CBL estimates from the GARpoly, GARres and RAX species trees, except the RAX tree inferred using all loci, which we used as the reference CBL set (CBL_{ref}). We selected the total evidence RAX species tree as the reference given this included all loci (including low informative loci) and randomly resolved gene trees, leading us to predict that it would exhibit the shortest branch lengths (i.e., greatest underestimation of CBLs) and typically yield positive Δ_{CBL} values in comparisons with other sets of gene trees.

4 | RESULTS

Across taxonomic groups, gene tree resolution improved as the number of parsimony-informative sites within loci increased, exhibiting a logarithmic pattern (Figure S1). Both UCE datasets contained some loci with fewer parsimony-informative sites than the number of internal branches, indicating that gene trees were certainly unresolved; an exceptionally limited number of gene trees were fully resolved in the GARpoly analyses of the UCE datasets (one [ocellated clade] and three [Laurasiatheria] gene trees). For the exon datasets, loci had relatively larger numbers of fully resolved gene trees (42% [oscine passerines] and 58% [Euarchontoglires] gene trees). Overall, gene trees from exon datasets were mostly resolved, while those from UCE datasets exhibited a wide range of variation in resolution.

Our fundamental hypothesis is that errors in gene trees, which are more likely to be present in gene trees based on less informative loci, will be misinterpreted as incomplete lineage sorting when CBLs are calculated. This makes a fundamental prediction: the sum of internal CBLs (the tree length) will increase as less informative loci are removed until the tree length reaches some plateau. This should be true for an estimate of the species tree that is relatively close to the true species tree. To determine whether this is true, we calculated the sum of CBLs in the optimal species tree of each dataset and examined when increases in the sum of CBLs reach a plateau as datasets were filtered. The species tree length for the UCE datasets reached a plateau with the 25% most informative loci (Table 2), although we note that further filtering (to the 5% most informative UCE loci) resulted in further tree length increases (albeit a modest one). In contrast, the tree length was maximal for the 25% most informative exon loci (Table 2), suggesting that further filtering was detrimental for exons. These results corroborate our fundamental hypothesis and indicate that comparisons of CBLs estimated using a single topology will yield meaningful results.

When we estimated CBLs on the 25% most informative GARpoly constraint topology, filtering gene trees by locus informativeness had the greatest impact on CBL estimation regardless of data type (i.e., UCE and exons). For many branches across our datasets, CBLs were often about >1.5× longer when analysing the 5%–25% most informative gene trees compared to more uninformative gene tree sets (Table S1), with some branches seeing as much as a sixfold increase in length. In general, there was a positive trend for higher Δ_{CBL} at longer branches when more informative UCE gene trees were compared (Figure 2, Figures S2–S4; Table 3). In contrast, exonic gene trees typically exhibited weak or no trends in Δ_{CBL} (Figure 3, Figure S2, S5, and S6), though there was a distinct negative trend when analyses were restricted to the 10% most informative gene trees (Figure 3, Figure S2). For both

TABLE 2 Sum of internal coalescent branch lengths across species trees (i.e., species tree length) inferred from each dataset.

Taxon	Locus informativeness	GARpoly	GARres	RAX
Ocellated clade	5% MI	27.0782	24.9169	24.8986
	10% MI	26.8650	23.7631	24.1428
	25% MI	23.7729	19.2653	19.7033
	TE	17.2751	8.6685	8.8798
	75% LI	15.0918	6.5285	6.7301
Laurasiatheria	5% MI	42.6645	39.6612	41.0539
	10% MI	42.5790	38.4256	40.1555
	25% MI	41.9837	37.6909	37.9926
	TE	32.9439	25.2602	25.6694
	75% LI	30.0455	22.2995	22.7527
Oscine passerines	10% MI	34.6420	34.3692	33.8722
	25% MI	41.7889	40.896	41.2547
	TE	36.4553	33.874	34.1420
	75% LI	33.3779	30.6555	30.9547
Euarchontoglires	10% MI	42.9453	42.6880	42.1396
	25% MI	44.6101	44.3177	44.0270
	TE	35.8123	34.7953	34.5908
	75% LI	32.7058	31.6655	31.5479

Abbreviations: GARpoly, GARLI gene trees with polytomies; GARres, GARLI gene trees with polytomies randomly resolved; LI, least informative loci; MI, most informative loci.

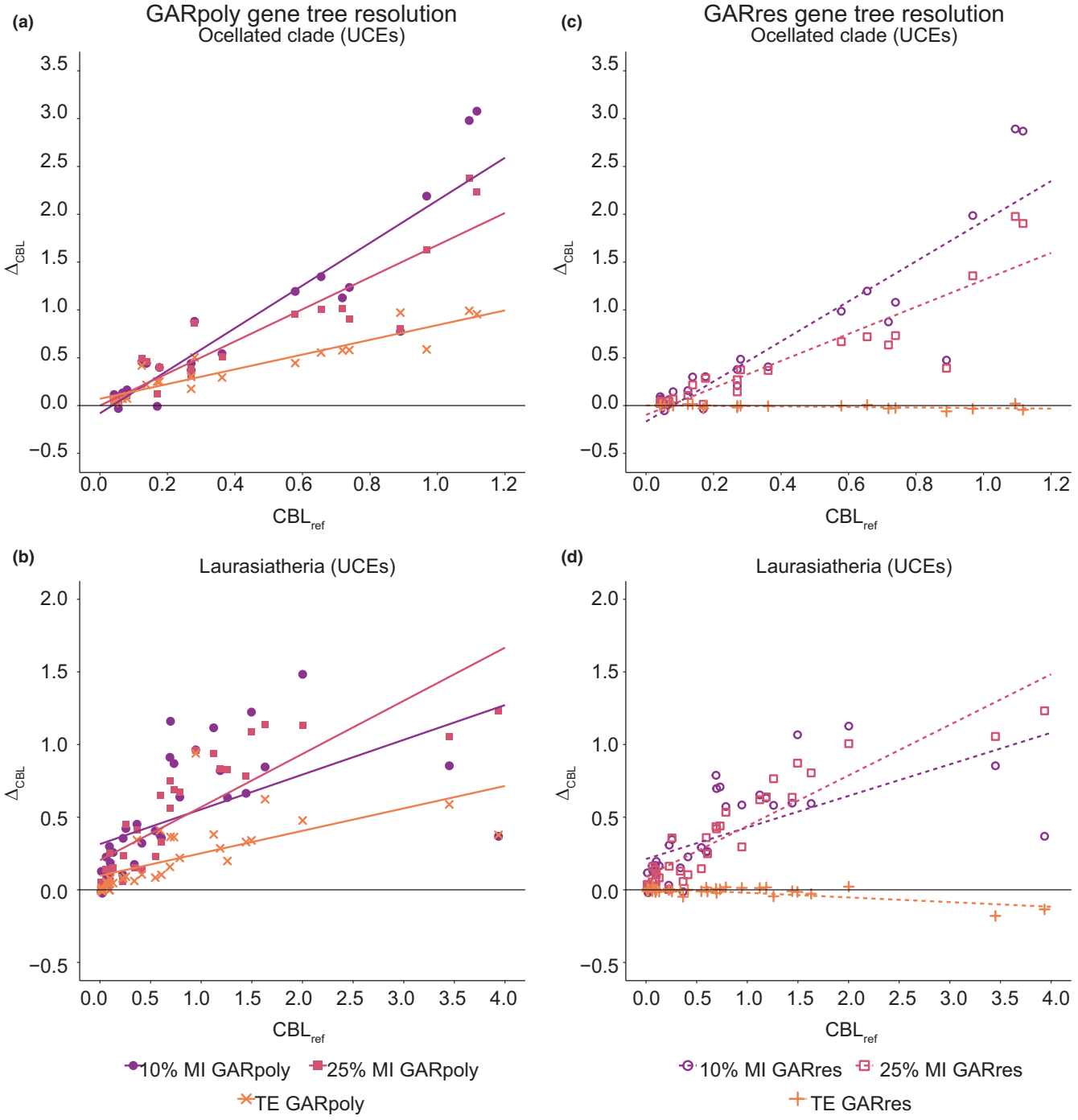

FIGURE 2 Differences in coalescent branch lengths (CBLs) (Δ_{CBL}) between the (a, c) ocellated clade and (b, d) Laurasiatheria ultraconserved element (UCE) locus informative datasets and a reference species tree (shown in a solid black line) generated from all RAxML gene trees for each taxonomic group. Species tree CBLs estimated from GARLI gene trees with polytomies (GARpoly) are shown on the left side of the figure (a, b) whereas GARLI gene trees with polytomies randomly resolved (GARres) are on the right (c, d). See Figure S2 for additional datasets not shown here. LI, least informative loci; MI, most informative loci; TE, total evidence

UCEs and exons, more informative gene trees generally estimated longer CBLs, but there was variation with respect to which informative threshold produced the longest CBL for a given branch; for example, the 25% most informative gene trees resulted in similar or longer CBLs than the more informative gene tree sets for most taxonomic groups (Figures 2 and 3, Figure S2).

When using the constraint topology, gene trees with polytomies only had a positive impact on CBL estimates when datasets included more uninformative gene trees,

TABLE 3 Slopes and r^2 values when plotting differences in coalescent branch lengths (Δ_{CBL}) between different locus informative datasets and the total evidence (TE) loci with RAxML gene trees with polytomies randomly resolved (RAX) (i.e., CBL_{ref}) for each taxonomic group. See Table 2 for additional abbreviations

Taxon	Locus informativeness	GARpoly		GARres		RAX	
		Slope	r^2	Slope	r^2	Slope	r^2
Ocellated clade	5% MI	2.1367	.8114	2.0928	.8116	2.0913	.7747
	10% MI	2.2267	.8357	2.0953	.8003	2.1617	.7899
	25% MI	1.6783	.8511	1.4116	.8157	1.4582	.8111
	TE	0.7703	.8639	−0.0272	.2374	—	—
	75% LI	0.5486	.6190	−0.2446	.9595	−0.2163	.8820
Laurasiatheria	5% MI	0.1375	.0682	0.1191	.0681	0.1223	.0688
	10% MI	0.2393	.2998	0.2172	.3999	0.2435	.3711
	25% MI	0.3657	.6974	0.3482	.8514	0.3369	.7665
	TE	0.1544	.4128	−0.0321	.5264	—	—
	75% LI	0.0859	.1708	−0.1083	.9287	−0.0739	.8283
Oscine passerines	10% MI	−0.2141	.3547	−0.2112	.3463	−0.2350	.3932
	25% MI	−0.0138	.0020	−0.0063	.0005	−0.0189	.0037
	TE	0.0322	.2187	−0.0325	.1766	—	—
	75% LI	−0.0285	.1755	−0.0962	.7101	−0.0574	.5809
Euarchontoglires	10% MI	−0.2739	.2224	−0.2687	.2160	−0.2646	.2253
	25% MI	−0.0478	.0139	−0.0419	.0106	−0.0414	.0113
	TE	0.0286	.2359	0.0140	.0858	—	—
	75% LI	−0.0233	.0934	−0.0404	.2644	−0.0551	.5860

that is the UCE datasets (compare Figure 2a to b and c to d). However, inclusion of uninformative gene trees still led to drastically shorter CBLs in the species tree compared with analyses including just more informative gene trees, even with polytomous trees (Figure 2c,d). The inclusion of polytomous gene trees did not appear to have a noticeable effect on CBL estimates in exons (compare Figure 3a to b and c to d). We also found the sum of CBLs in the optimal species tree for a given filtered dataset (Table 2) was nearly the same as the sum of CBLs from the same dataset when branch lengths were estimated using our reference topology (i.e., the 25% most informative GARpoly constraint topology; see Table S1).

5 | DISCUSSION

The use of CBLs to infer historical population demographic parameters or to test whether regions of the species tree are in the anomaly zone is likely to become more common in future studies. Although coalescent methods assume that conflict among gene trees is due to incomplete lineage sorting, factors such as low informative loci, poor taxon sampling, and model misspecifications have been shown to negatively affect the accuracy of gene tree estimation and the resulting species tree topology (Bayzid & Warnow, 2013; Gatesy & Springer, 2014; Molloy & Warnow, 2018; Patel et al., 2013; Sayyari et al., 2017; Springer & Gatesy, 2014, 2016). Thus, it is critical that future investigations on the effects of systematic error in species tree inferences also include evaluations of CBLs given our results suggest that CBLs are also greatly affected.

Overall, our study provides strong evidence that more informative loci may reduce CBL estimation error, as might be expected based on studies that demonstrate low informative loci to be problematic for gene tree estimation along short branches (Gatesy & Springer, 2014; Xi et al., 2015). In empirical datasets, there is some degree of error in gene tree estimates that may result in minority resolutions when the correct resolution would be the congruent (majority) resolution. This gene tree estimation error can be incorrectly interpreted as discordance by summary coalescent methods and result in underestimated CBLs in the species tree (Gatesy et al., 2017; Sayyari & Mirarab, 2016; Springer & Gatesy, 2014). Thus, low informative loci will negatively impact gene tree quality (i.e., increases gene tree estimation error) and increase error in CBL estimation, which is consistent with our results on our UCE datasets (which included many low informative loci).

However, we also observed that restriction to the 5% or 10% most informative loci did not improve CBL estimation

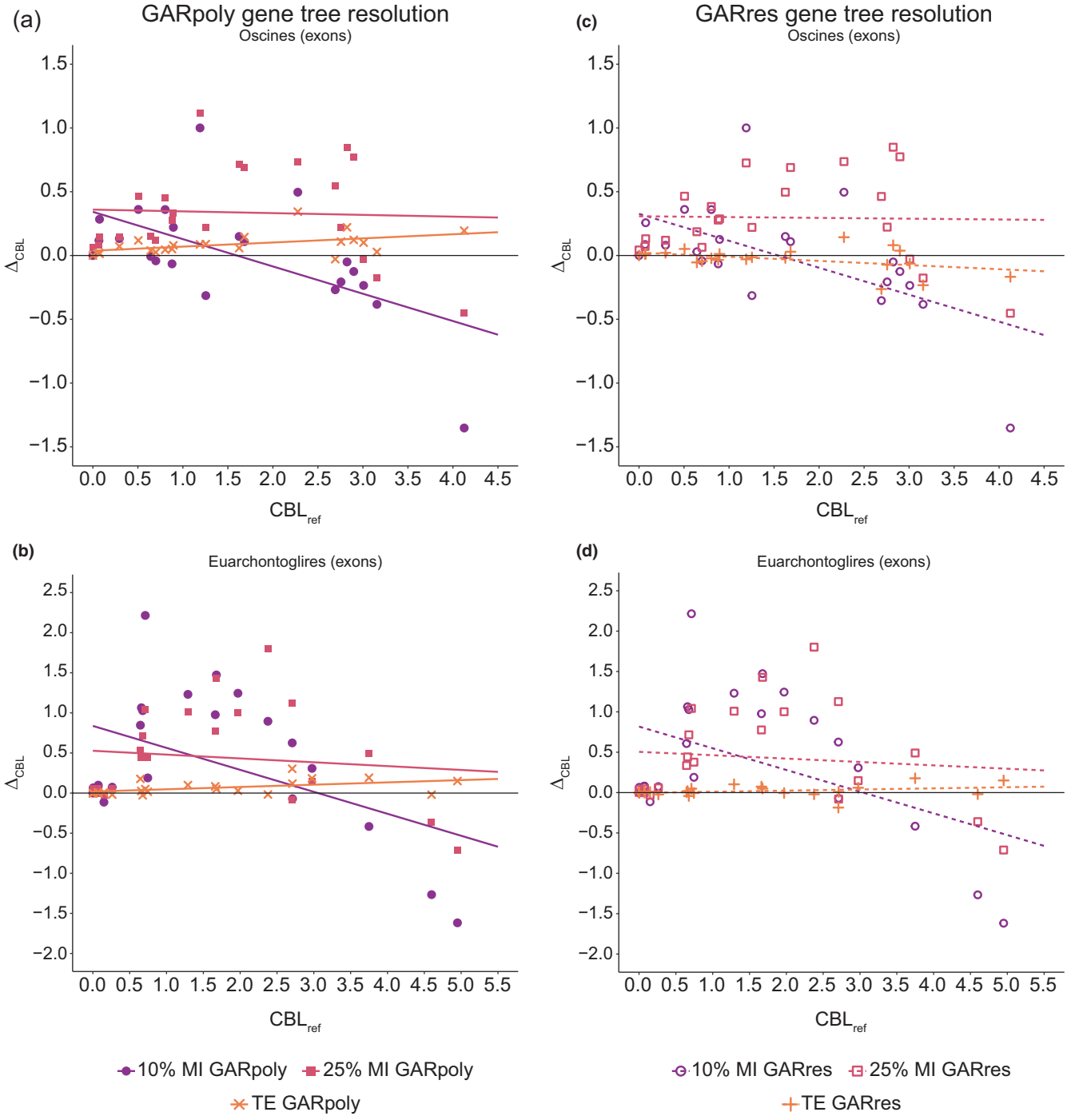

FIGURE 3 Differences in coalescent branch lengths (CBLs) (Δ_{CBL}) between the (a, c) oscine passerines and (b, d) Euarchontoglires exon informative datasets and a reference species tree (shown in a solid black line) generated from all RAxML gene trees for each taxonomic group. Species tree CBLs estimated from GARLI gene trees with polytomies (GARpoly) are shown on the left side of the figure (a, b) whereas GARLI gene trees with polytomies randomly resolved (GARres) are on the right (c, d). See Figure S2 for additional datasets not shown here. LI, least informative loci; MI, most informative loci; TE, total evidence

for certain datasets. Several simulation studies have demonstrated that species tree topology and CBL estimation improves in accuracy as more loci/gene trees are included in analyses (Leaché & Rannala, 2011; Liu et al., 2010; Mirarab et al., 2016; Mirarab, Bayzid, et al., 2014; Mirarab, Reaz, et al., 2014; Sayyari & Mirarab, 2016). This leads to a trade-off between the number of gene trees included in an analysis and the topological accuracy of those gene trees; too many topologically inaccurate gene trees will lead to CBL underestimation, but too few gene trees will inflate the variance of the CBL estimates. Thus, implementing very restrictive filtering can drastically reduce the number

of gene trees available, as in the case of our 10% most informative loci in the exon datasets (e.g., 23 [oscine passerines] and 41 [Euarchontoglires] loci; Table S2), which should also negatively affect CBL estimates in empirical studies (as well as estimates of the species tree topology). The impact of filtering can have differential impacts on CBL estimates. Since the relationship between CBLs and the proportion of congruent gene trees is logarithmic, small changes in the proportion of congruent gene trees (as may occur when filtering) will have a large impact on CBL estimates for nodes where most gene trees are congruent with the species tree, while small changes in the proportion of congruent gene trees will have a much more limited impact on CBL estimates for nodes with a lower proportion of congruent gene trees. Thus, care should be taken to avoid excessive filtering of loci/gene trees.

Avoiding excessive filtering requires the articulation of a criterion that can be used to decide whether loci should be retained or excluded. For datasets with a large sample of gene trees based on characteristically uninformative loci, like our ocellated clade UCE dataset, very restrictive filtering (5%–10% of the most parsimony-informative loci) appears to result in the retention of a sufficient sample of loci to yield more accurate CBL estimates. However, in cases where there are fewer loci, such as the Laurasiatheria UCE dataset or the exon datasets, such restrictive filtering is unlikely to provide much improvement compared with the 25% most informative loci and may even have a negative impact. Indeed, it is possible to imagine two classes of extreme datasets: (a) a dataset where all loci are highly informative and (b) a dataset where all loci have very little information. In the former, filtering is unlikely to be helpful (arguably, our exonic datasets are in this category). In the latter, there is unlikely to be any way to obtain accurate CBLs; none of our datasets appear to fall into this category but any such datasets would be characterized by the universal absence of long CBLs combined with little or no impact of filtering. Since all empirical datasets differ in their properties, it is advisable to explore how filtering based on locus informativeness affects CBL estimation in empirical studies given that the number of loci used in coalescent analysis are an important variable. We suggest two specific ways to avoid over-filtering. First, one can examine the type of plots we present here, which are a useful empirical tool in this context. Large negative Δ_{CBL} values associated with at least some of the large reference CBLs (which should be relatively accurate given a sufficient number of gene trees) provide evidence of over-filtering; this is what we observed for gene trees based on the 10% most informative exons (Figure 3), and we believe that using the 25% most informative exons is appropriate for this reason. Second, one can simply examine the sum of internal branch lengths for the optimal species tree given various datasets; the point at which the sum reaches a plateau is a good empirical estimate of the best filtering level, with decreases after a maximum (as observed for our exons) as evidence of over-filtering.

Our prediction that there would be few differences in CBL estimates for longer branches—where even low informative gene trees are expected to be resolved—was largely not supported. The Laurasiatheria UCE dataset was the only one that exhibited patterns similar to our prediction, with Δ_{CBL} values relatively low for the longer branches. However, we cannot exclude the possibility that this result may be biased by a limited sample of long branches. In general, our UCE results likely reflect the left half of Figure 1 where the differences in CBL estimates obtained using high and low informative loci are still increasing. Thus, for these datasets, the low informative loci appear to be so uninformative as to frequently fail to recover even long branches, and it is possible that a dataset including loci more informative than those included here would have matched our prediction. For the exon data, differences in CBL estimates were relatively small for most of our locus informative subsets across all branch lengths, with results often showing only weak trends. This is likely to indicate that uninformative exons still contain sufficient parsimony-informative sites that filtering for more informative loci offers little improvement in CBL estimates. Overall, our results suggest that it may be desirable to do at least limited gene filtering (while still retaining sufficiently large numbers of loci) to obtain more accurate CBL estimates.

An alternative to filtering for more informative loci is to use optimal parsimony or ML gene trees that include polytomies for branches without support from the data (i.e., zero-length branches). Although true polytomies in gene trees have been considered unlikely (Edwards, 2009), real-world phylogenetic datasets often contain loci with limited variation that cannot support resolution of all nodes in the gene tree (McCormack et al., 2012; Zhang et al., 2017). Leaving unsupported branches (i.e., zero-length branches) collapsed avoids the inclusion of randomly resolved relationships that are most likely artefactual, so it should never be problematic to invoke a simple polytomy option whenever it is possible. Indeed, our results show that CBLs were often shorter when randomly resolved gene trees were used for species tree inference compared to polytomous gene trees, but this only applied to the more uninformative UCE datasets. Thus, these results suggest that, as we (Figure 1) and others (Simmons & Kessenich, 2020) predicted, allowing polytomies in gene trees may improve gene tree quality by addressing the issue of artificial resolution for less informative loci. We further note that the use of polytomous gene trees and limiting analyses to highly informative loci are not mutually

exclusive and that CBL estimates (and likely also species tree topologies) may improve when both approaches are combined. This may be particularly important in cases where erroneous branches may remain in uninformative gene trees even when polytomies are allowed (e.g., see figure 3 from Meiklejohn et al., 2016).

There are other approaches that have been applied to prevent unbiased errors in gene trees from negatively affecting species trees topologies and/or CBL estimates. These alternative approaches may be necessary since some phylogenetic programs may yield imperfect branch length estimates, making it difficult to determine when a gene tree branch is actually zero versus a very small positive value (e.g., 0.000001; see Simmons & Kessenich, 2020. Several studies have collapsed branches based on support levels (Sayyari & Mirarab, 2016; Simmons & Gatesy, 2021; Simmons & Kessenich, 2020), although the best way to determine an appropriate bootstrap cut-off (as in Houde et al., 2019) is unclear. A better approach is to ask whether there is any evidence that a branch is resolved in a particular gene tree; this could be accomplished as we did here (by collapsing near-zero length branches) or by collapsing branches with a Shimodaira-Hasegawa-like approximate likelihood ratio (SH-like aLRT) of 0%. Indeed, Simmons and Gatesy (2021) showed that collapsing using the 0% SH-like aLRT criterion generally produces longer CBLs compared to those estimated when gene tree branches are collapsed based on bootstrap support. Alternatively, one can use a strict consensus of optimal trees (Simmons & Gatesy, 2021), although that is most appropriate when the maximum parsimony criterion is used to estimate gene trees. This reflects the fact that there are seldom multiple equally optimal trees in maximum likelihood searches, and the alternative strategy of obtaining the optimal tree and many trees with very similar scores is often impractical. It is important to note that the use of more stringent criteria for collapsing nodes or filtering gene trees may lead to collapsing correctly resolved nodes and/or filtering of loci that accurately recovered the gene tree, both of which increase the variance of CBL estimates. Ultimately, the goal of any effort to reduce the impact of inaccurate gene trees on CBL estimation will involve finding a good balance between minimizing the impact of errors in estimation while still minimizing the variance of the estimates, perhaps by using a mix of strategies (versus one very stringent criterion) such as we did in this study.

The benefits of collapsing branches in gene trees may seem counter-intuitive given that true polytomies in gene trees are highly unlikely (Edwards, 2009; Hudson, 1990; Slowinski, 2001). However, the conservative practice of collapsing branches in gene trees that have no empirical support is appropriate if the method used to estimate CBLs simply ignores gene tree polytomies, just as ASTRAL does. In fact, this finding suggests another method that might improve CBL estimation: using rare genomic changes like long indels (Houde et al., 2020), transposable element insertions (Springer & Gatesy, 2019), or microinversions (Braun et al., 2011). Each rare genomic change can be viewed as a polytomous gene tree with a single resolved branch. However, the benefit of homoplasy-free (or virtually homoplasy-free) rare genomic changes is that the single branch in each of those is very likely to be a correct branch in the true gene tree; as long as the polytomies are not considered in CBL calculations, those highly unresolved gene trees should yield more accurate CBLs than a collection of fully resolved gene trees in which some of the internal branches are erroneous.

6 | CONCLUSION

Population genetic studies, species tree evaluations and other evolutionary investigations stand to benefit from coalescent methods that estimate CBLs, but the conditions in which coalescent methods can obtain accurate CBLs in empirical datasets have not been well explored. Reducing CBL estimation error will have positive effects on such analyses that may use CBLs for other inferences. CBL estimation can be improved in several ways. First, it is desirable to limit analysed gene trees to those based on more informative loci (e.g., using number of parsimony-informative sites, which is one type of parameter that provides information for phylogenetic analysis). Second, one can eliminate branches likely to be erroneous by collapsing them to form polytomies. Finally, one can combine both approaches. Alternatively, one might focus on rare genomic changes that reliably define a single branch or even transversions at individual sites within highly conserved regions (Tiley et al., 2020) (although the last approach is only appropriate for shallow divergences where transversion homoplasy is limited). Collapsing branches definitely appears to be beneficial for loci that are generally characterized by low information content (e.g., UCEs). Other sources of error known to affect topological estimates (e.g., recombination and missing data) also remain to be investigated for their impacts on empirically estimated CBLs. It is our hope that this study stimulates further contributions to improve CBL estimation and their reliable utility in various subdisciplines of evolutionary biology.

ACKNOWLEDGEMENTS

We thank Mark P. Simmons and an anonymous reviewer for their valuable comments that improved our manuscript. This work was supported by the U.S. National Science Foundation (grant number DEB 1118823 and DEB 1655683 to R.T.K. and E.L.B.).

ORCID

Michael Forthman https://orcid.org/0000-0002-6987-8503

REFERENCES

Bayzid, M. S., & Warnow, T. (2013). Naive binning improves phylogenomic analyses. *Bioinformatics*, *29*, 2277–2284. https://doi.org/10.1093/bioinformatics/btt394

Betancur-R, R., Naylor, G. J. P., & Ortí, G. (2014). Conserved genes, sampling error, and phylogenomic inference. *Systematic Biology*, *63*, 257–262. https://doi.org/10.1093/sysbio/syt073

Blom, M. P. K., Bragg, J. G., Potter, S., & Moritz, C. (2017). Accounting for uncertainty in gene tree estimation: Summary-coalescent species tree inference in a challenging radiation of Australian lizards. *Systematic Biology*, *66*, 352–366. https://doi.org/10.1093/sysbio/syw089

Braun, E. L., Kimball, R. T., Han, K.-L., Iuhasz-Velez, N. R., Bonilla, A. J., Chojnowski, J. L., Smith, J. V., Bowie, R. C. K., Braun, M. J., Hackett, S. J., Harshman, J., Huddleston, C. J., Marks, B. D., Miglia, K. J., Moore, W. S., Reddy, S., Sheldon, F. H., Witt, C. C., & Yuri, T. (2011). Homoplastic microinversions and the avian tree of life. *BMC Evolutionary Biology*, *11*, 141. https://doi.org/10.1186/1471-2148-11-141

Chen, D., Braun, E. L., Forthman, M., Kimball, R. T., & Zhang, Z. (2018). A simple strategy for recovering ultraconserved elements, exons, and introns from low coverage shotgun sequencing of museum specimens: Placement of the partridge genus *Tropicoperdix* within the galliformes. *Molecular Phylogenetics and Evolution*, *129*, 304–314. https://doi.org/10.1016/j.ympev.2018.09.005

Chen, M.-Y., Liang, D., & Zhang, P. (2015). Selecting question-specific genes to reduce incongruence in phylogenomics: A case study of jawed vertebrate backbone phylogeny. *Systematic Biology*, *64*, 1104–1120. https://doi.org/10.1093/sysbio/syv059

Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., & Edwards, S. V. (2019). Whole-genome analyses resolve the phylogeny of flightless birds (Palaeognathae) in the presence of an empirical anomaly zone. *Systematic Biology*, *68*, 937–955. https://doi.org/10.1093/sysbio/syz019

Costa, I. R., Prosdocimi, F., & Jennings, W. B. (2016). *In silico* phylogenomics using complete genomes: A case study on the evolution of hominoids. *Genome Research*, *26*, 1257–1267. https://doi.org/10.1101/gr.203950.115

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLOS Genetics*, *2*, e68. https://doi.org/10.1371/journal.pgen.0020068

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*, 332–340. https://doi.org/10.1016/j.tree.2009.01.009

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, *63*, 1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x

Ericson, P. G. P., Qu, Y., Rasmussen, P. C., Blom, M. P. K., Rheindt, F. E., & Irestedt, M. (2019). Genomic differentiation tracks earth-historic isolation in an Indo-Australasian archipelagic pitta (Pittidae; Aves) complex. *BMC Evolutionary Biology*, *19*, 151. https://doi.org/10.1186/s12862-019-1481-5

Esselstyn, J. A., Oliveros, C. H., Swanson, M. T., & Faircloth, B. C. (2017). Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. *Genome Biology and Evolution*, *9*, 2308–2321. https://doi.org/10.1093/gbe/evx168

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2020). A Bayesian Implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution*, *37*, 1211–1223. https://doi.org/10.1093/molbev/msz296

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*, 296–332. https://doi.org/10.1111/cla.12170

Gatesy, J., & Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, *80*, 231–266. https://doi.org/10.1016/j.ympev.2014.08.013

Goloboff, P. A., & Simmons, M. P. (2014). Bias in tree searches and its consequences for measuring group supports. *Systematic biology*, *63*, 851–861.

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. *Systematic Biology*, *59*, 307–321. https://doi.org/10.1093/sysbio/syq010

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2016). Avoiding missing data biases in phylogenomic inference: An empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, *33*, 1110–1125. https://doi.org/10.1093/molbev/msv347

Houde, P., Braun, E. L., Narula, N., Minjares, U., & Mirarab, S. (2019). Phylogenetic signal of indels and the neoavian radiation. *Diversity*, *11*, 108. https://doi.org/10.3390/d11070108

Houde, P., Braun, E. L., & Zhou, L. (2020). Deep-time demographic inference suggests ecological release as driver of neoavian adaptive radiation. *Diversity*, *12*, 164. https://doi.org/10.3390/d12040164

Huang, H., & Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Systematic Biology*, *58*, 527–536. https://doi.org/10.1093/sysbio/syp047

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, *7*, 1–44.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307. https://doi.org/10.1093/biomet/76.2.297

Jennings, W. B., & Edwards, S. V. (2005). Speciational history of australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution*, *59*, 2033–2047. https://doi.org/10.1554/05-280.1

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro (Ed.), *Mammalian protein metabolism* (pp. 21–132). Academic Press.

Kubatko, L. S., & Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, *56*, 17–24. https://doi.org/10.1080/10635150601146041

Leaché, A. D., & Rannala, B. (2011). The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology*, *60*, 126–137. https://doi.org/10.1093/sysbio/syq073

Léveillé-Bourret, É., Chen, B.-H., Garon-Labrecque, M.-È., Ford, B. A., & Starr, J. R. (2020). RAD sequencing resolves the phylogeny, taxonomy and biogeography of Trichophoreae despite a recent rapid radiation (Cyperaceae). *Molecular Phylogenetics and Evolution*, *145*, 106727. https://doi.org/10.1016/j.ympev.2019.106727

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). *Systematic Biology*, *65*, 465–477. https://doi.org/10.1093/sysbio/syw001

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302. https://doi.org/10.1186/1471-2148-10-302

Liu, L., Yu, L., Kubatko, L., Pearl, D. K., & Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution*, *53*, 320–328. https://doi.org/10.1016/j.ympev.2009.05.033

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, *22*, 746–754. https://doi.org/10.1101/gr.125864.111

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: Evidence for a bias in some multispecies coalescent methods. *Systematic Biology*, *65*, 612–627. https://doi.org/10.1093/sysbio/syw014

Mirarab, S., Bayzid, M. S., Boussau, B., & Warnow, T. (2014). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, *346*, 1250463. https://doi.org/10.1126/science.1250463

Mirarab, S., Bayzid, M. S., & Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, *65*, 366–380. https://doi.org/10.1093/sysbio/syu063

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*, i541–i548. https://doi.org/10.1093/bioinformatics/btu462

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. https://doi.org/10.1093/bioinformatics/btv234

Molloy, E. K., & Warnow, T. (2018). To include or not to include: The impact of gene filtering on species tree estimation methods. *Systematic Biology*, *67*, 285–303. https://doi.org/10.1093/sysbio/syx077

Nylander, J. A. A. (2004). *MrAIC.pl. Program distributed by the author*. Evolutionary Biology Centre, Uppsala University.

Patel, S., Kimball, R. T., & Braun, E. L. (2013). Error in phylogenetic estimation for bushes in the tree of life. *Journal of Phylogenetics & Evolutionary Biology*, *1*, 110. https://doi.org/10.4172/2329-9002.1000110

Peng, J., Swofford, D. L., & Kubatko, L. (2021). Estimation of speciation times under the multispecies coalescent. *bioRxiv*, https://doi.org/10.1101/681023

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*, 569–573. https://doi.org/10.1038/nature15697

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*, 1645–1656. https://doi.org/10.1093/genetics/164.4.1645

Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, *33*, 1654–1668. https://doi.org/10.1093/molbev/msw079

Sayyari, E., Whitfield, J. B., & Mirarab, S. (2017). Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Molecular Biology and Evolution*, *34*, 3279–3291. https://doi.org/10.1093/molbev/msx261

Shi, C.-M., & Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Molecular Biology and Evolution*, *35*, 159–179. https://doi.org/10.1093/molbev/msx277

Simmons, M. P., & Gatesy, J. (2021). Collapsing dubiously resolved gene-tree branches in phylogenomic coalescent analyses. *Molecular Phylogenetics and Evolution*, *158*, 107092. https://doi.org/10.1016/j.ympev.2021.107092

Simmons, M. P., & Kessenich, J. (2020). Divergence and support among slightly suboptimal likelihood gene trees. *Cladistics*, *36*, 322–340. https://doi.org/10.1111/cla.12404

Slowinski, J. B. (2001). Molecular polytomies. *Molecular Phylogenetics and Evolution*, *19*, 114–120. https://doi.org/10.1006/mpev.2000.0897

Song, S., Liu, L., Edwards, S. V., & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 14942–14947. https://doi.org/10.1073/pnas.1211733109

Springer, M. S., & Gatesy, J. (2014). Land plant origins and coalescence confusion. *Trends in Plant Science*, *19*, 267–269. https://doi.org/10.1016/j.tplants.2014.02.012

Springer, M. S., & Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33. https://doi.org/10.1016/j.ympev.2015.07.018

Springer, M. S., & Gatesy, J. (2019). Retroposon insertions within a multispecies coalescent framework suggest that ratite phylogeny is not in the 'anomaly zone'. *bioRxiv*, 643296. https://doi.org/10.1101/643296

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4*. Sinauer Associates.

Thawornwattana, Y., Dalquen, D., & Yang, Z. (2018). Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Molecular Biology and Evolution*, *35*, 2512–2527. https://doi.org/10.1093/molbev/msy158

Tiley, G. P., Pandey, A., Kimball, R. T., Braun, E. L., & Burleigh, J. G. (2020). Whole genome phylogeny of *Gallus*: Introgression and data-type effects. *Avian Research*, *11*, 7. https://doi.org/10.1186/s40657-020-00194-w

Weber, C. C., Boussau, B., Romiguier, J., Jarvis, E. D., & Ellegren, H. (2014). Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biology*, *15*, 549. https://doi.org/10.1186/s13059-014-0549-1

Xi, Z., Liu, L., & Davis, C. C. (2015). Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, *92*, 63–71. https://doi.org/10.1016/j.ympev.2015.06.009

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, *162*, 1811–1823. https://doi.org/10.1093/genetics/162.4.1811

Zhang, C., Sayyari, E., & Mirarab, S. (2017). ASTRAL-III: Increased scalability and impacts of contracting low support branches. In J. Meidanis, & L. Nakhleh (Eds.), *Comparative genomics. RECOMB-CG 2017. Lecture notes in computer science* (Vol. *10562*, pp. 53–75). Springer. https://doi.org/10.1007/978-3-319-67979-2_4

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion* (Thesis). https://repositories.lib.utexas.edu/handle/2152/2666

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Forthman, M., Braun, E. L., & Kimball, R. T. (2022). Gene tree quality affects empirical coalescent branch length estimation. *Zoologica Scripta*, 51, 1–13. https://doi.org/10.1111/zsc.12512