Molecular Phylogenetics and Evolution 97 (2016) 76–89

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

The effects of subsampling gene trees on coalescent methods applied to ancient divergences ☆

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

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
[b] Department of Biology, University of California, Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Received 9 October 2015
Revised 3 December 2015
Accepted 20 December 2015
Available online 6 January 2016

Keywords:
ASTRAL
Concatenation
MP-EST
Phylogeny
Shortcut coalescent methods
STAR

ABSTRACT

Gene-tree-estimation error is a major concern for coalescent methods of phylogenetic inference. We sampled eight empirical studies of ancient lineages with diverse numbers of taxa and genes for which the original authors applied one or more coalescent methods. We found that the average pairwise congruence among gene trees varied greatly both between studies and also often within a study. We recommend that presenting plots of pairwise congruence among gene trees in a dataset be treated as a standard practice for empirical coalescent studies so that readers can readily assess the extent and distribution of incongruence among gene trees. ASTRAL-based coalescent analyses generally outperformed MP-EST and STAR with respect to both internal consistency (congruence between analyses of subsamples of genes with the complete dataset of all genes) and congruence with the concatenation-based topology. We evaluated the approach of subsampling gene trees that are, on average, more congruent with other gene trees as a method to reduce artifacts caused by gene-tree-estimation errors on coalescent analyses. We suggest that this method is well suited to testing whether gene-tree-estimation error is a primary cause of incongruence between concatenation- and coalescent-based results, to reconciling conflicting phylogenetic results based on different coalescent methods, and to identifying genes affected by artifacts that may then be targeted for reciprocal illumination. We provide scripts that automate the process of calculating pairwise gene-tree incongruence and subsampling trees while accounting for differential taxon sampling among genes. Finally, we assert that multiple tree-search replicates should be implemented as a standard practice for empirical coalescent studies that apply MP-EST.

1. Introduction

Coalescent methods that allow for differential lineage sorting among genes are now often applied in phylogenetic analyses of both recently diverged and ancient lineages—even in cases where there is no reason to believe that the anomaly zone, wherein the most likely gene-tree topology contradicts the phylogenetic tree (Degnan and Rosenberg, 2006), may apply. Shortcut coalescent methods (i.e., those that do not co-estimate the phylogenetic tree with the gene trees; Gatesy and Springer, 2014) are statistically consistent if the gene trees are known without error (Liu et al., 2009, 2010; Mirarab et al., 2014). Gene-tree-estimation error can be caused by limited character variation among recently diverged lineages (Huang and Knowles, 2009), but for ancient lineages there are the additional potential problems of long-branch attraction (Felsenstein, 1978), mis-rooting (Rosenfeld et al., 2012), convergent nucleotide composition (Lockhart et al., 1992), and short coalescent genes (Hobolth et al., 2011; Gatesy and Springer, 2014; Springer and Gatesy, 2016). Indeed, high levels of gene-tree conflict are frequently reported in empirical studies that sampled ancient lineages (e.g., Betancur-R et al., 2013; Salichos and Rokas, 2013; Pyron et al., 2014).

Gene-tree incongruence that is caused by estimation errors can be difficult to distinguish from the biological process of lineage sorting (Yang, 2002; Leigh et al., 2008; Betancur-R et al., 2014). This is particularly true when applying coalescent methods to resolve rapid ancient radiations because resolution of such problems is dependent upon rapidly evolving genes to provide synapomorphies along these short branches, yet these same synapomorphies are likely to be obscured by subsequent mutations (Maddison and Knowles, 2006). Mis-estimated gene trees have been identified as a probable cause of artifacts in shortcut

☆ This paper was edited by the Associate Editor G. Orti.

* Corresponding author at: Department of Biology, 200 West Lake Street, Colorado State University, Fort Collins, CO 80523-1878, USA. Fax: +1 970 491 0649.
E-mail address: psimmons@lamar.colostate.edu (M.P. Simmons).

http://dx.doi.org/10.1016/j.ympev.2015.12.013
1055-7903/

coalescent analyses of some empirical datasets (Meredith et al., 2011; Townsend et al., 2011; Gatesy and Springer, 2014; Simmons and Gatesy, 2015; Springer and Gatesy, 2016), and different coalescent methods can produce mutually exclusive phylogenetic trees in these cases (e.g., Gatesy and Springer, 2014; Springer and Gatesy, 2014, 2016; Simmons and Gatesy, 2015).

A variety of approaches have been proposed to quantify and/or help minimize phylogenetic-inference artifacts caused by divergent gene trees. Leigh et al. (2011a) clustered gene trees based on their shared bipartitions, after which each cluster may be analyzed independently of the others. De Vienne et al. (2012) subsampled both genes and taxa with the most similar pairwise distances among taxa in their gene trees. Salichos and Rokas (2013) subsampled those gene trees with high bootstrap support (Felsenstein, 1985). Mirarab et al. (2014) excluded two outlier genes with high pairwise Robinson-Foulds distance (hereafter "RF distance;" Robinson and Foulds, 1981) relative to other gene trees (clades in gene trees with <75% bootstrap support were collapsed). Similarly, Pyron et al. (2014) quantified pairwise incongruence among gene trees and also between each gene tree and the coalescent phylogenetic tree using RF distances. The latter approach is implemented in STRAW (Shaw et al., 2013), which also takes into account differential taxon sampling among gene trees. Sharma et al. (2014) alternatively subsampled genes with the lowest percentage of missing terminals or those with the highest percent pairwise amino-acid identity. Zimmermann et al. (2014) used *BEAST (Heled and Drummond, 2010), a coalescent method that co-estimates gene trees and the phylogenetic tree, to improve gene-tree estimation prior to applying a shortcut coalescent method.

Of these various alternatives, we focused on the approach of subsampling those gene trees that have low average RF distance with other gene trees after correcting for the number of shared terminals (hereafter the "RF method"). By comparing gene trees pairwise, the RF method bypasses comparison to a species tree inferred from the gene trees and does not favor similarity to an initial species tree that may be biased by outlier gene trees. This same approach can be implemented using rooted triplets or unrooted quartets (Estabrook et al., 1985; Leigh et al., 2011b; Zwickl et al., 2014) instead of RF, but these methods, despite their expected greater stability to outlier terminals, have not performed well in practice (Kuhner and Yamato, 2015).

In this study, we sampled eight empirical studies of ancient lineages with diverse numbers of taxa and genes for which the original authors applied one or more coalescent methods. For each of these studies we quantified topological incongruence among gene trees, compared the relative performance of three shortcut coalescent methods (ASTRAL, MP-EST, and STAR) that are frequently applied to empirical datasets, tested the effectiveness of subsampling gene trees using the RF method for improving coalescent-based phylogenetic inference, quantified how heuristic MP-EST tree searches can affect the inferred phylogenetic tree, tested alternative character-coding and character-sampling approaches for two studies, and used the RF method to identify outlier gene trees. We did so in a particularly challenging context—ancient lineages for which we expect gene-tree-estimation error to be severe.

2. Materials and methods

2.1. Primary empirical studies sampled

For the core analyses of our study, we sampled the following eight empirical studies: Betancur-R et al. (2013; hereafter "Betancur"), Chiari et al. (2012; hereafter "Chiari"), McCormack et al. (2012; hereafter "McCormack"), Pyron et al. (2014; hereafter "Pyron"), Townsend et al. (2011; hereafter "Townsend"), Wiens et al. (2012; hereafter "Wiens"), Xi et al. (2014; hereafter "Xi") and Zhong et al. (2013; hereafter "Zhong"). These studies include diverse animal and plant lineages, numbers of taxa (16–214), and numbers of gene trees (20–333; Table 1).

Gene trees and concatenation-based phylogenetic trees were obtained from the authors, downloaded from Dryad, or manually re-created from the original authors' figures. In most cases, the original authors' gene trees and phylogenetic trees were used, though some gene trees required re-rooting (applicable to Betancur wherein gene trees were rooted using *Zeus* when the two outgroups [*Polymixia* and *Zeus*] were resolved as polyphyletic, and Zhong for which gene trees were rooted using *Micromonas* when the three Chlorophyte outgroups were not resolved as a clade) and/or re-naming a small minority of inconsistently named taxa. We used Simmons and Gatesy's (2015) concatenation-based phylogenetic trees for Xi, which were based on partitioned-by-gene RAxML tree searches using 100 search replicates, in contrast to the original authors' unpartitioned analysis from a single RAxML search replicate. We also used Simmons and Gatesy's (2015) gene trees that were based on subsamples of the characters because Xi were not able to provide these.

In cases where the original authors analyzed alternative datasets with different numbers of gene trees and/or taxa, we always selected the dataset with the higher number of taxa (applicable to McCormack, Townsend, and Wiens), even if the original authors did not apply coalescent analyses to these datasets (applicable to Townsend and Wiens). For Zhong, wherein the original authors analyzed the same number of taxa with different numbers of genes (184 or 289), we selected the larger dataset. All gene trees and concatenation-based phylogenetic trees used are posted as supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1615928.

Chiari performed their coalescent (and concatenation) analyses alternatively using gene trees inferred from amino-acid or nucleotide characters (hereafter Chiari AA and Chiari DNA). Both concatenation-based analyses and the Chiari AA coalescent analysis resolved the topology ((turtles)((crocodylians)(birds))),

Table 1
Characteristics of the eight empirical studies sampled.

Study	# Terminals	# Gene trees	Lineage[a]	Reference clade(s)
Betancur-R et al. (2013)	214	20	Percomorpha	Flatfishes
Chiari et al. (2012)	16	248	Sarcopterygii	((turtles)((crocodylians)(birds)))
McCormack et al. (2012)	29	183	Amniota	((Rodentia)(Lagomorpha))
Pyron et al. (2014)	33	333	Squamata	Caenophidia
Townsend et al. (2011)	76	29	Lepidosauria	((Leiosaurids)(Oplurids))
Wiens et al. (2012)	171	44	Amniota	(Dibamidae)(Gekkota)[b]
Xi et al. (2014)	46	310	Tracheophyta	(*Amborella*, (other flowering plants))
Zhong et al. (2013)	32	289	Viridiplantae	((Zygnematales)(land plants))

[a] Including outgroup(s).
[b] Together as a clade or as a paraphyletic group on successive branches.