Syst. Biol. 73(5):807–822, 2024

https://doi.org/10.1093/sysbio/syae031
Advance Access Publication June 28, 2024

The Influence of the Number of Tree Searches on Maximum Likelihood Inference in Phylogenomics

Chao Liu[1,2], Xiaofan Zhou[3], Yuanning Li[4,5], Chris Todd Hittinger[6], Ronghui Pan[7], Jinyan Huang[8], Xue-xin Chen[1], Antonis Rokas[5], Yun Chen[1], and Xing-Xing Shen[1,2,*]

[1]*Department of Plant Protection, Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Zhejiang University, Hangzhou 310058, China*
[2]*Centre for Evolutionary & Organismal Biology, Zhejiang University, Hangzhou 310058, China*
[3]*Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, South China Agricultural University, Guangzhou 510642, China*
[4]*Institute of Marine Science and Technology, Shandong University, Qingdao 266237, China*
[5]*Department of Biological Sciences and Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA*
[6]*Laboratory of Genetics, Wisconsin Energy Institute, Center for Genomic Science Innovation, DOE Great Lakes Bioenergy Research Center, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53706, USA*
[7]*ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, 310027, China*
[8]*Zhejiang Provincial Key Laboratory of Pancreatic Disease, Zhejiang University School of Medicine First Affiliated Hospital, Hangzhou 310003, China*
Chao Liu and Xiaofan Zhou contributed equally to this article.
[*]*Correspondence to be sent to: Department of Plant Protection, Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Zhejiang University, Hangzhou 310058, China; E-mail: xingxingshen@zju.edu.cn.*

Received 4 November 2022; reviews returned 20 June 2024; accepted 26 June 2024
Associate Editor: Olivier Gascuel

Abstract.—Maximum likelihood (ML) phylogenetic inference is widely used in phylogenomics. As heuristic searches most likely find suboptimal trees, it is recommended to conduct multiple (e.g., 10) tree searches in phylogenetic analyses. However, beyond its positive role, how and to what extent multiple tree searches aid ML phylogenetic inference remains poorly explored. Here, we found that a random starting tree was not as effective as the BioNJ and parsimony starting trees in inferring the ML gene tree and that RAxML-NG and PhyML were less sensitive to different starting trees than IQ-TREE. We then examined the effect of the number of tree searches on ML tree inference with IQ-TREE and RAxML-NG, by running 100 tree searches on 19,414 gene alignments from 15 animal, plant, and fungal phylogenomic datasets. We found that the number of tree searches substantially impacted the recovery of the best-of-100 ML gene tree topology among 100 searches for a given ML program. In addition, all of the concatenation-based trees were topologically identical if the number of tree searches was ≥10. Quartet-based ASTRAL trees inferred from 1 to 80 tree searches differed topologically from those inferred from 100 tree searches for 6/15 phylogenomic datasets. Finally, our simulations showed that gene alignments with lower difficulty scores had a higher chance of finding the best-of-100 gene tree topology and were more likely to yield the correct trees. [Heuristic tree search; hill-climbing; local optima; maximum likelihood; phylogenomics; species tree estimation.]

Reconstructing the evolutionary relationships among organisms is important for understanding the patterns and mechanisms of genetic and phenotypic diversity (Hillis et al. 1996; Felsenstein 2003; Hamilton 2014; Yang 2014; Li et al. 2022). Because the number of possible trees grows exponentially with the number of taxa (Felsenstein 1978), modern phylogenetic inference relies on heuristic search algorithms (e.g., hill-climbing algorithms) to infer a nearly optimal tree (Chor and Tuller 2005) in the space that consists of all possible unrooted binary trees. Under the maximum likelihood (ML) framework (Felsenstein 1981, 2003), for instance, tree searching is typically an iterative process that begins with a starting tree (e.g., a tree inferred by parsimony or distance methods, or a random tree), from which a set of candidate trees is generated by rearrangement operations such as Nearest-Neighbor-Interchange (NNI) (Robinson 1971), Subtree-Pruning-and-Regrafting (SPR) (Swofford et al. 1996), and Tree-Bisection-and-Reconnection (TBR) (Allen and Steel 2001). If a candidate tree has a higher log-likelihood score than the starting tree, it will replace the starting tree to initiate a new iteration. The tree search process finishes when no tree with a higher log-likelihood score can be found and the tree with the highest score is deemed to be the nearly optimal or ML tree.

Popular programs for ML phylogenetic inference mainly differ in the rearrangement operations. For example, RAxML-NG implements a SPR-based hill-climbing search strategy; at each iteration of tree search, it identifies promising re-grafting positions within a certain radius from the pruning position and applies multiple SPR rearrangements simultaneously to speed up the inference. IQ-TREE initially maintains a pool of 20 best-candidate trees, determined from 1 BioNJ starting tree (Gascuel 1997) and 99 parsimony starting trees. Each iteration starts with a tree that is selected at random from the pool, the tree topology is stochastically perturbed and used as the starting tree for a NNI-based hill-climbing tree search. The analysis finishes if no better tree can be found for multiple iterations. PhyML utilizes both types of rearrangement operations successively; it first performs an SPR-based hill-climbing tree search, and the resulting tree is further improved by NNI-based hill-climbing. Because the search strategy is heuristic, it is not guaranteed that this ML tree is the one with the globally highest score. A standard solution to increasing the chance of finding a better ML tree is to conduct multiple independent tree searches, each from a different starting tree or a different random seed, in current fast ML-based programs such as IQ-TREE (Nguyen et al. 2015; Minh et al. 2020), MEGA (Tamura et al. 2011; Kumar et al. 2016), PhyML (Guindon and Gascuel 2003; Guindon et al. 2010), and RAxML/RAxML-NG (Stamatakis 2014; Kozlov et al. 2019).

Several previous studies have extensively examined the efficacy of different rearrangement operations on ML tree inference (e.g., Vinh and Haeseler 2004; Morrison 2007; Money and Whelan 2012). For example, Money and Whelan (2012) found that NNI performed poorly in finding the nearly optimal tree compared to SPR. In light of this result, IQ-TREE (Nguyen et al. 2015; Minh et al. 2020), one of the state-of-the-art ML-based programs, overcomes the weakness of NNI-based tree search by implementing a broad sampling of initial starting trees and random perturbation of current best trees. In addition, many previous studies have examined the efficacy of different fast ML-based programs on ML tree inference (Liu et al. 2011; Nguyen et al. 2015; Zhou et al. 2018; Kozlov et al. 2019; Park et al. 2021). For example, a recent analysis (Zhou et al. 2018) of 19 empirical phylogenomic datasets showed that IQ-TREE, PhyML, and RAxML had comparable performance when conducting 10 tree searches on each alignment.

A potential drawback of almost all of the previous studies is that they used no more than 20 tree searches, leaving the effect of varying the number of tree searches on ML phylogenetic inference underinvestigated. To address this gap, we performed 100 tree searches for each of the 19,414 single-gene alignments in 15 animal, plant, and fungal phylogenomic datasets and 20,000 simulated gene alignments (Shen et al. 2020; Höhler et al. 2022a). Then, we asked 2 questions: i) How does the number of tree searches affect the performance of finding the best ML tree? ii) Are extensive tree searches in ML phylogenetic inference necessary? Our results reveal that variation in number of tree searches can substantially influence ML tree inference and that the difficulty score (Haag et al. 2022) could be a useful predictor for roughly estimating the necessary number of tree searches for ML inference.

RESULTS

Different ML phylogenetic programs could have different tree search algorithms, resulting in varying numbers of tree searches during a single default run. For example, RAxML-NG's default run initiates with 10 parsimony starting trees and 10 random starting trees, followed by conducting one tree search on each starting tree. Ultimately, it produces the best ML tree from a total of 20 tree searches. IQ-TREE's default run begins with one BioNJ starting tree and 99 parsimony starting trees, followed by iteratively conducting tree searches on a pool of 20 best candidate trees throughout the analysis. In the end, it produces the best ML tree, determined from a stochastic number of tree searches. In this study, following the strategy of a recent study (Kozlov et al. 2019), we defined one tree search as using one starting tree for both RAxML-NG and PhyML and using one run for IQ-TREE. Note that the effect of varying the number of starting trees on ML phylogenetic inference is not directly comparable across different ML programs due to the variation in the number of tree searches performed by different ML programs. In addition, we had no intention of comparing likelihood scores or topological accuracies of different ML programs in this study. Therefore, our assessment solely focused on the effect of the number of tree searches on ML phylogenetic inference within a given ML program.

The Effect of Different Starting Trees on Single-Gene Tree Inference

As ML phylogenetic inference begins with a starting tree, we first investigated the effect of different starting trees on single-gene ML tree inferences. For each gene alignment, we conducted one independent tree search from a BioNJ tree, a parsimony tree, or a random tree using 1 CPU on a single compute node (AMD EPYC 7662 @ 2.0 GHz processor with 128 threads). We sampled 200 genes from each of 15 animal, plant, and fungal phylogenomic datasets since executing all of the tree searches on the same node was computationally expensive (Table 1). Thus, for each of 3000 gene alignments, we used 3 different starting trees to infer 3 ML gene trees and denoted the ML gene tree with the highest log-likelihood score as the best-of-3 ML gene tree topology within a given ML program. We then examined the fractions of the 3000 single-gene alignments for which the best-of-3 gene tree topology was found, which we refer to as the "recovery rate." Note that the recovery rate was not comparable between IQ-TREE, RAxML-NG, and PhyML, as the best-of-3 ML gene tree topology was determined independently for each ML program.

Overall, we found that in terms of recovery rates, the random starting tree was less efficient in finding the best-of-3 ML gene tree topology for IQ-TREE, as compared with the BioNJ starting tree and the parsimony starting tree (Fig. 1a). RAxML-NG and PhyML were less sensitive to different starting trees (Fig. 1a). Among the 15 phylogenomic datasets, different starting trees exhibited varying recovery rates for a given ML program. Notably, IQ-TREE displayed greater variation

TABLE 1. Summary of 15 phylogenomic datasets examined in this study.

Study ID	Dataset	Taxon level	No. taxa	No loci	Sampling method	Data type	Study reference
Bee	Animal: Bees	Genus	190	753	UCE	DNA	Blaimer et al., Evolution, 2018 (Blaimer et al. 2018)
Bird	Animal: Birds	Class	200	259	AHE	DNA	Prum et al., Nature, 2015 (Prum et al. 2015)
Butterfly	Animal: Butterflies	Order	207	352	AHE	DNA	Espeland et al., Current Biology, 2018 (Espeland et al. 2018)
Lizard	Animal: Lizards	Genus	29	1361	Exon-Capture	DNA	Blom et al., Syst Biol, 2017 (Blom et al. 2017)
Marine-fish	Animal: Marine fishes	Superorder	120	1001	UCE	DNA	Alfaro et al., Nat. Ecol. Evol. 2018 (Alfaro et al. 2018)
Rodent	Animal: Rodents	Family	37	1245	Exon-Capture	DNA	Roycroft et al., Syst Biol, 2019 (Roycroft et al. 2020)
Cardueae	Plant: Cardueae	Family	85	570	UCE	DNA	Herrando-Moraira et al., Mol Phyloge Evol, 2018 (Herrando-Moraira et al. 2018)
Caryophyllales	Plant: Caryophyllales	Order	95	1122	Transcriptome	AA	Yang et al., Mol Biol Evol, 2015 (Yang et al. 2015)
Green-Plants	Plant: Green plants	Phylum	1178	410	Transcriptome	AA	1KP Initiative, Nature, 2019 (One Thousand Plant Transcriptomes Initiative 2019)
Jaltomata	Plant: Jaltomata	Genus	15	6431	Transcriptome	DNA	Wu et al., Mol Ecol, 2018 (Wu et al. 2018)
Protea	Plant: Protea	Genus	65	498	AHE	DNA	Mitchell et al., American Journal of Botany, 2017 (Mitchell et al. 2017)
Aspergillaceae	Fungi: Aspergillaceae	Order	93	1668	Genome	DNA	Steenwyk et al., mBio, 2019 (Steenwyk et al. 2019b)
Saccharomycotina-Cell	Fungi: Budding yeasts	Subphylum	343	2408	Genome	AA	Shen et al., Cell, 2018 (Shen et al. 2018)
Hanseniaspora	Fungi: Hanseniaspora	Family	29	1033	Genome	AA	Steenwyk et al., PloS Biol, 2019 (Steenwyk et al. 2019a)
Rhizoplaca	Fungi: Rhizoplaca	Genus	31	303	Genome	DNA	Leavitt et al., Sci Rep, 2016 (Leavitt et al. 2016)

in recovery rate across the three different starting trees, especially for phylogenomic datasets with larger numbers of taxa (e.g., Bees, Birds, Butterflies, Green plants, and Budding yeasts) (see Supplementary Fig. S1). This observation is likely attributed to the fact that IQ-TREE employs an NNI-based hill-climbing tree search, which explores a less extensive tree searching space compared to the SPR-based hill-climbing tree search (Zhou et al. 2018).

In addition, as we executed all tree searches using 1 CPU on the same compute node, we can fairly compare the runtimes of inferring ML gene trees across different starting trees and different ML programs. Overall, we found that the use of a random starting tree had slightly longer runtimes than the uses of BioNJ starting tree and parsimony starting tree within a given ML program (Fig. 1b and Supplementary Fig. S2). RAxML-NG on average ran the fastest, followed by IQ-TREE and PhyML (Fig. 1b and Supplementary Fig. S2).

The Number of Tree Searches Substantially Influences the Identification of Single-Gene Trees With the Highest Log-Likelihood Scores

We investigated the effect of varying the number of tree searches on ML phylogenetic inference through an extensive analysis of all 19,414 single-gene alignments from 15 animal, plant, and fungal phylogenomic datasets (Table 1). For each of 19,414 single-gene alignments, we conducted 100 tree searches using 100 runs for IQ-TREE (v1.6.12) and using 50 parsimony starting trees and 50 random starting trees for RAxML-NG (v0.9.0), following the tree search strategy of a recent study (Kozlov et al. 2019). Because executing 100 tree searches from a single command line is computationally expensive, we chose to partition them into 5 sets. Each set involved running 20 tree searches, with IQ-TREE utilizing the option "-runs 20 -seed random number" and RAxML-NG utilizing the option "--tree pars{10},rand{10} --seed random number" (see Supplementary Text for details). In brief, RAxML-NG performs one tree search on each of the 20 starting trees using the SPR rearrangement operation. For each of 20 runs in IQ-TREE, it starts with one BioNJ starting tree and 99 parsimony starting trees and then maintains a pool of 20 best candidate trees to conduct one tree search using the NNI rearrangement operation. Due to differences in tree search algorithms between IQ-TREE and RAxML-NG, it is important to note that the results regarding the impact of varying the number of tree searches on ML phylogenetic inference are not directly comparable between the two programs.

After obtaining 100 single-gene ML trees labeled with R1 to R100 for each of 19,414 gene alignments within a given ML program, we considered the ML gene tree with the highest log-likelihood score as the best-of-100 ML gene tree topology and asked which runs achieved the best-of-100 ML gene tree topologies. Among the 19,414 gene alignments, we observed that 938 (4.8%) for IQ-TREE and 781 (4.0%) for RAxML-NG produced the best-of-100 ML gene trees that had equal highest log-likelihood scores but different topologies. In theory, distinct gene topologies should not share identical log-likelihood scores; however, it is well known that

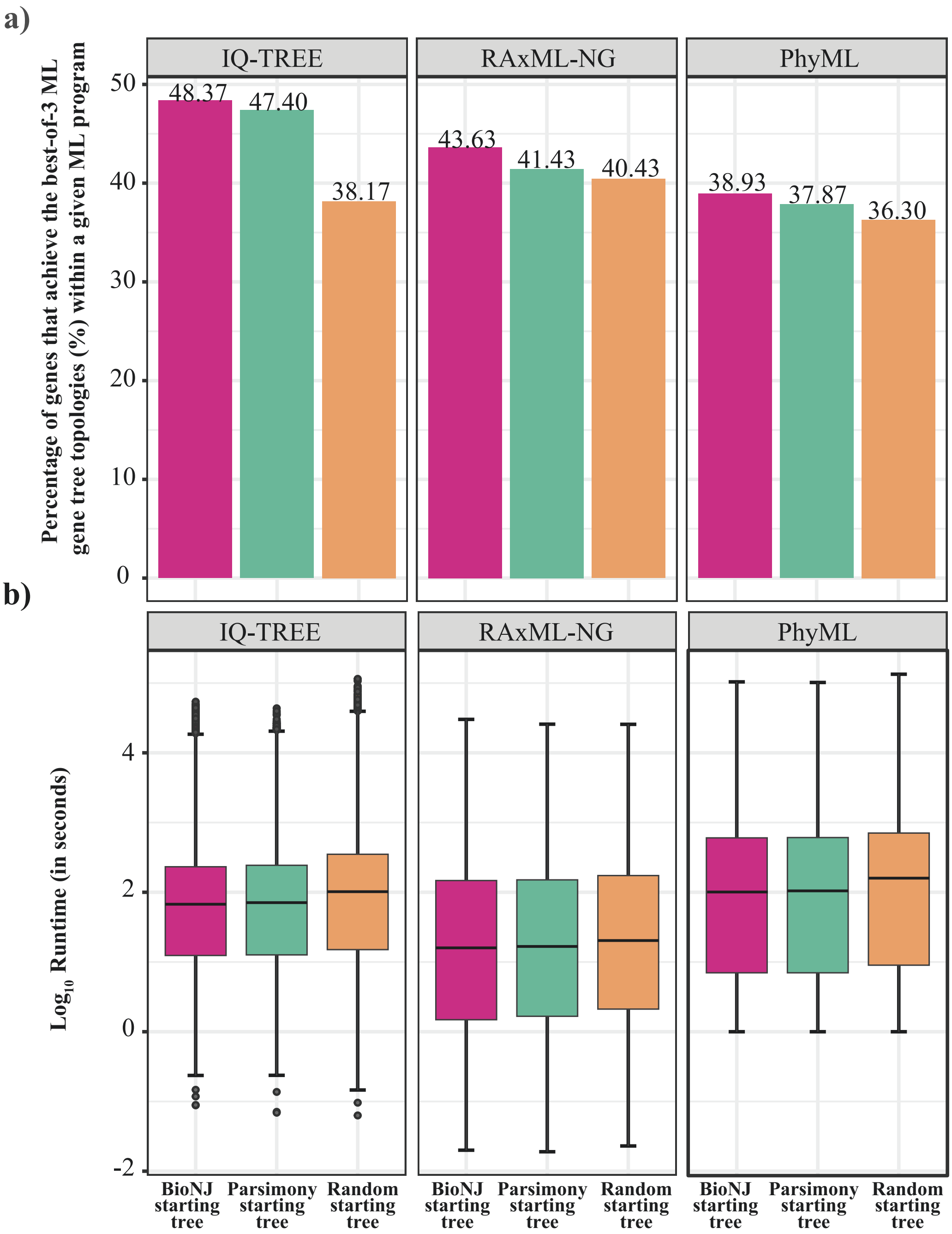

FIGURE 1. Effect of different starting trees on ML gene tree inferences. We conducted 1 independent ML tree search from BioNJ tree, parsimony tree, or random tree using 1 CPU on a single compute node (see "Methods" section for details). For each of 3000 gene alignments, we defined the ML gene tree with the highest log-likelihood score as the best-of-3 ML gene tree topology found by a given ML program. a) Percentage of genes that achieved the best-of-3 ML gene tree topologies within a given ML program. b) Runtime of one independent ML tree search using 1 CPU on the same node. The runtimes (in seconds) are shown in logarithm base 10. Horizontal bar in the boxplot denotes the median value. The individual results for each of 15 phylogenomic studies are given in Supplementary Figs. S1 and S2.

ML programs have limited numerical precision for log-likelihood score calculation. This could lead to different gene topologies having identical log-likelihood scores in the output files (Haag et al. 2023). Given that the proportion of gene alignments with identical highest log-likelihood scores but different topologies was relatively small, we included them in subsequent analyses.

Next, we examined the recovery rate, that is the fraction of the 19,414 single-gene alignments for which the best-of-100 ML gene tree topologies were found for a given number of tree searches. Overall, the recovery rates were ~51% for the 19,414 IQ-TREE-inferred gene trees and ~42% for the 19,414 RAxML-NG-inferred gene trees when using one tree search (Fig. 2 and

Supplementary Table S1). The recovery rate increased to 69% for the 19,414 IQ-TREE-inferred gene trees and ~64% for the 19,414 RAxML-NG-inferred gene trees when using ten tree searches, which is computationally tractable for most phylogenomic data matrices (Fig. 2). Among the 15 phylogenomic datasets examined, the recovery rates at ten tree searches varied between 10% and 99.5% with an average value of 54.7% for IQ-TREE and between 8% and 99% with an average value of 49.5% for RAxML-NG (Fig. 2). In addition to the recovery rate metric, we also calculated the probability (p) of finding the best-of-100 ML gene tree topology for a given number of tree searches (n) for each gene alignment. The probability (p) is $1-(1-f)^n$, where f is the fraction of the best-of-100 ML gene tree topologies observed out of 100 tree searches. It is important to note that the recovery rate metric (f) and the probability (p) serve different purposes. The former is an actual

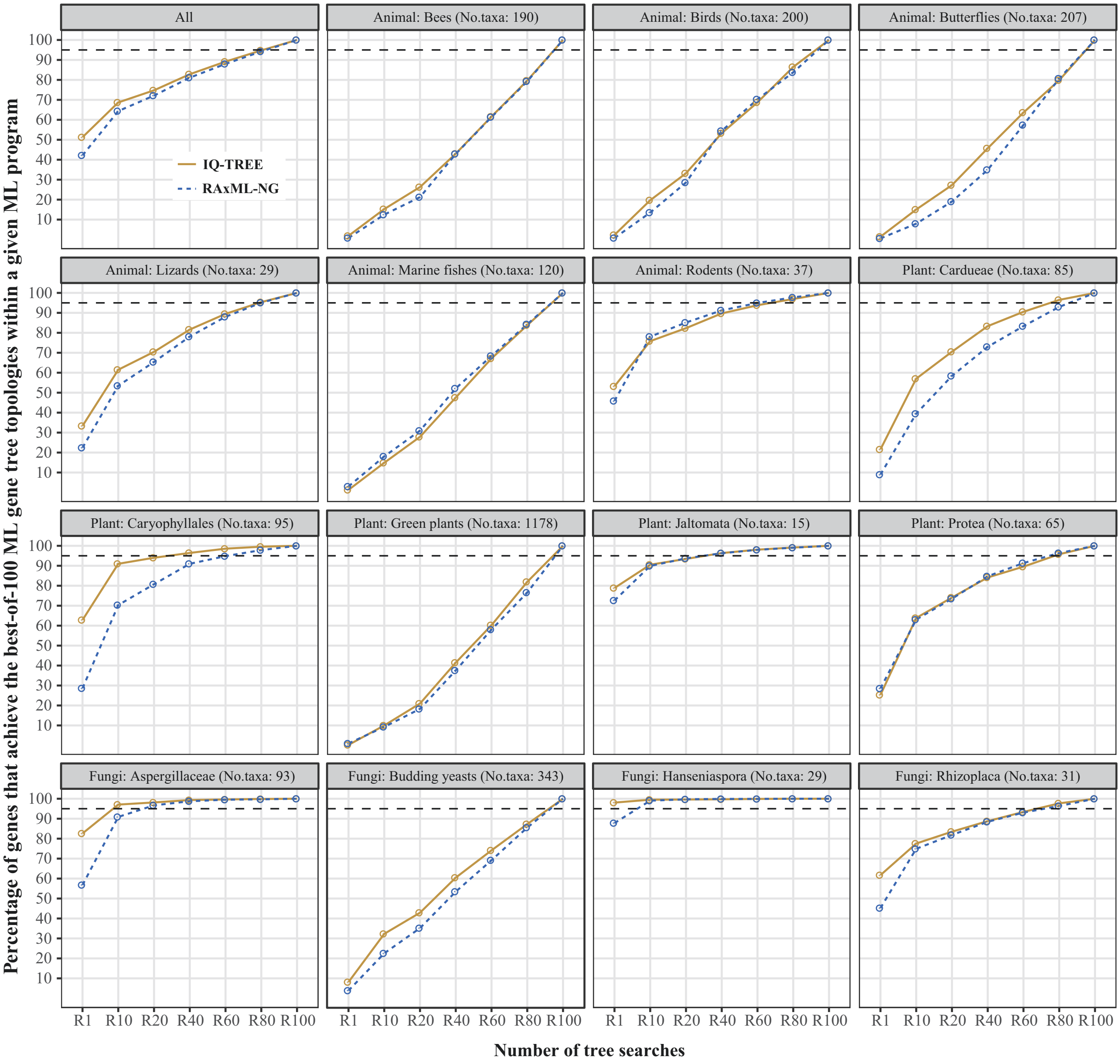

FIGURE 2. Effect of varying numbers of tree searches on finding the best-of-100 ML gene tree topology. For each of the 19,414 gene alignments from 15 diverse phylogenomic datasets (Table 1), we used 100 extensive tree searches with 2 maximum likelihood (ML) programs IQ-TREE and RAxML-NG. The ML gene tree topology with the highest log-likelihood score was defined as the best-of-100 ML gene tree topology found among 100 tree searches for a given ML program. To assess the effect of varying numbers of tree searches on finding the best-of-100 ML gene tree topology, we asked whether the best-of-100 ML gene tree topology was encountered when using 1, 10, 20, 40, 60, 80, and 100 tree searches for a given ML program, respectively. The dot plot at the upper left is based on all 19,414 gene alignments from 15 phylogenomic data sets. The rest of the dot plots show the individual results for each of the 15 phylogenomic data sets. Horizontally dashed lines denote 95%. The number of taxa for each of 15 phylogenomic data sets was included on the top of each panel title.

observed value and the latter is a predicted value based on the observed fraction. Among the 15 phylogenomic datasets, both the probability (*p*) and the recovery rate metric (*f*) demonstrated similar trends in evaluating the chance of finding the best-of-100 ML gene tree topologies (Supplementary Fig. S3), while they exhibited variation in the chance of finding the best-of-100 ML gene tree topologies at different numbers of tree searches.

As expected, we found that, with an increasing number of tree searches, both IQ-TREE- and RAxML-NG-inferred gene trees increased log-likelihood scores and were topologically more similar to the best-of-100 ML gene tree topologies (Fig. 3a,b and Supplementary Figs. S4 and S5). To examine whether the best-observed gene tree topology found from 1 to 80 significantly differed from the best-of-100 gene tree topology, we used the approximately unbiased (AU) test (Shimodaira 2002) to evaluate whether the best-observed gene tree topology and the best-of-100 gene tree topology could equally explain the gene alignment (null hypothesis H0). We found that the number of the best-observed gene tree topologies that had significantly lower log-likelihood scores than the best-of-100 gene tree topologies (AU test; *P* value ≤ 0.05) decreased with an increasing number of tree searches for the 19,414 single-gene alignments (Fig. 3c and Supplementary Fig. S6). When evaluating the changes in log-likelihood scores and gene tree topological similarity across different numbers of tree searches, we observed the biggest improvement from R1 to R10 (Fig. 3).

To explore the underlying causes of the varying chances of finding the best-of-100 ML gene tree topology, we divided the 19,414 single-gene alignments from 15 phylogenomic studies into 11 groups according to the number of the best-of-100 ML gene tree topologies observed out of 100 tree searches (Fig. 4a and Supplementary Fig. S7). Next, for each of the 11 groups of gene alignments, we examined 5 characteristics: difficulty score of the gene alignment predicted by Pythia (v1.1.2) (Haag et al. 2022), which integrates 8 features such as parsimony trees, entropy, and alignment attributes to quantify the degree of difficulty for analyzing a gene alignment *prior* to initiating ML tree inference; parsimony-informative sites in gene alignment; average bootstrap support across the best-of-100 ML gene tree topology; percentage of internal branches with high bootstrap support values; and percentage of internal branches with near-zero lengths. Overall, we found that gene alignments with lower chances of finding the best-of-100 ML gene tree topologies tended to have higher difficulty scores (Fig. 4b), lower numbers of parsimony-informative sites (Fig. 4c), lower average bootstrap support values (Fig. 4d), lower percentages of internal branches with high bootstrap support values (Fig. 4e), and higher percentages of internal branches with near-zero lengths (Fig. 4f). We observed similar trends within each of 15 phylogenomic datasets (Supplementary Figs. S8–S12). Furthermore, we examined the correlation between the chance of finding the best-of-100 ML gene tree topology and each of the 5 characteristics and found that the difficulty score exhibited the strongest correlation with the chance of finding the best-of-100 ML gene tree topology (Supplementary Fig. S13 and Supplementary Table S2). Finally, we found that the majority (IQ-TREE: 67.44%; RAxML-NG: 62.77%) of gene alignments with a ≤10% chance of finding the best-of-100 ML gene tree topology were characterized as difficult datasets (Supplementary Fig. S14a). In contrast, the majority (IQ-TREE: 65.49%; RAxML-NG: 76.61%) of gene alignments with a ≥90% chance of finding the best-of-100 ML gene tree topology were characterized as easy datasets (Supplementary Fig. S14b). Although predicting the number of tree searches required to find the best-of-100 ML gene tree topology is quite challenging (Vinh and Haeseler 2004; Haag et al. 2022; Höhler et al. 2022a), our results along with a recent finding (Togkousidis et al. 2023) suggested that the difficulty score could be a useful predictor for roughly estimating the necessary number of tree searches for ML inference.

The Effect of the Number of Tree Searches on Concatenation- and Quartet-Based Species Tree Inferences

We next assessed the effect of the number of tree searches on concatenation- and quartet-based species tree inferences. Given the difference in algorithm between concatenation- and quartet-based approaches we used, we did not directly compare the concatenation-based species trees with the quartet-based species trees inferred with different numbers of tree searches.

We first assessed the effect of the number of tree searches on concatenation-based species tree estimations. For each phylogenomic dataset, we used 100 tree searches to reconstruct concatenation-based ML trees with IQ-TREE and RAxML-NG (see Supplementary Text for details), respectively. The green plant phylogenomic data set was left out due to its very large number of taxa (1178) because inferring concatenation-based ML tree searches for the green plant phylogenomic data set using IQ-TREE and RAxML-NG on a 48-CPU node failed to finish after 5 months. Overall, the recovery rate of the best-of-100 concatenation-based ML tree topology was ~79% (11/14 datasets) for IQ-TREE and ~71% (10/14 datasets) for RAxML-NG when using 1 tree search (Fig. 5a), and 100% for both IQ-TREE and RAxML-NG when the number of tree searches was equal to or greater than 10 (Fig. 5a, Supplementary Fig. S15, and Supplementary Table S3). These results suggest that the use of 10 tree searches is sufficient to generate the best-of-100 concatenation-based ML tree topology.

We then examined the effect of varying numbers of tree searches on quartet-based species tree estimations for each of the 15 phylogenomic datasets. For each phylogenomic dataset, we inferred the quartet-based species phylogeny on the set of all the best-observed gene tree topologies inferred from 1 to 100 tree searches by weighting the branch support and branch length with the wASTRAL-h v1.3 (Zhang and Mirarab 2022). Here, we considered the ASTRAL tree

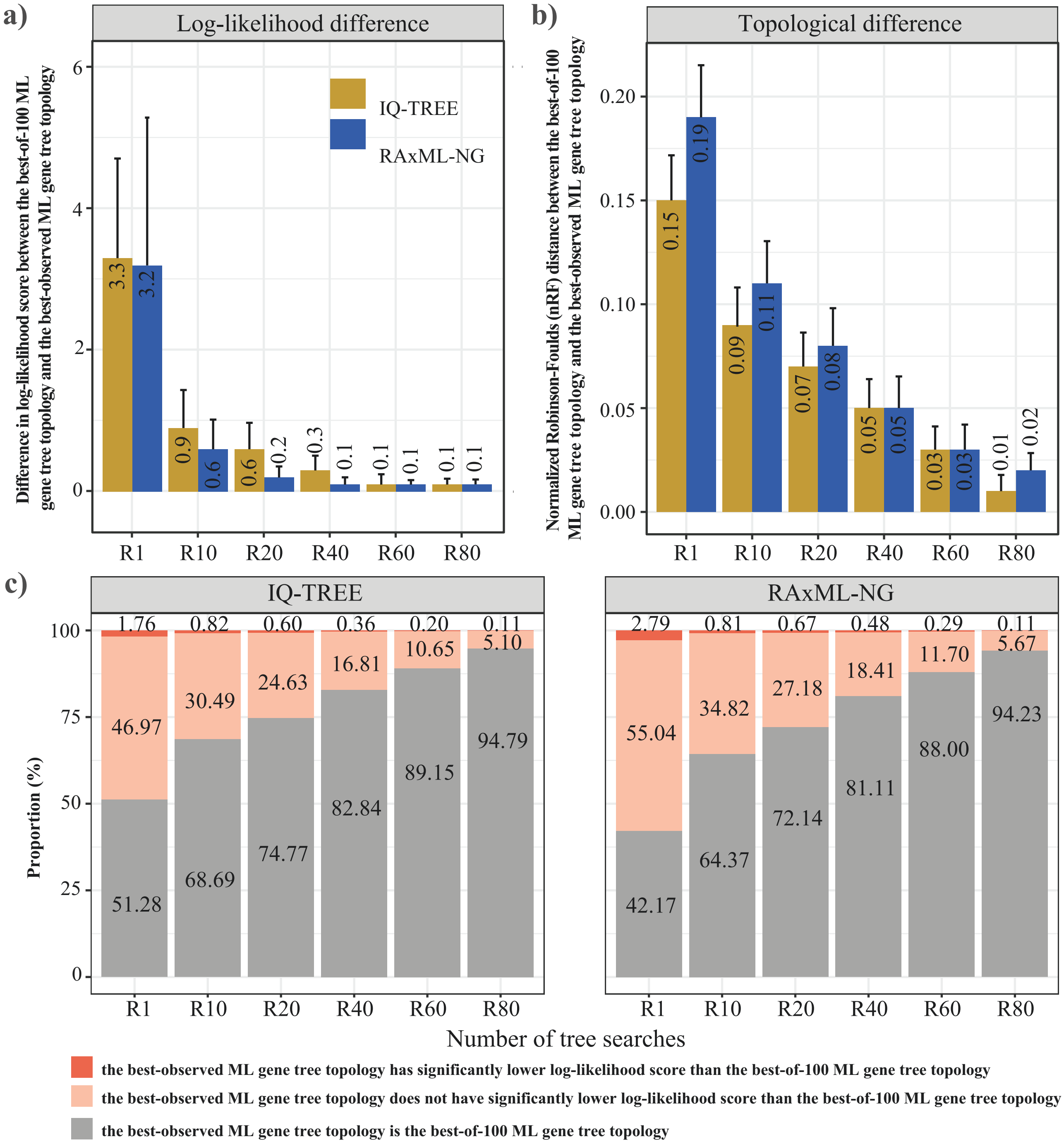

FIGURE 3. Comparisons of the best-of-100 ML gene tree topologies and the best-observed ML gene tree topologies were found at varying numbers of tree searches. Here, we examined all 19,414 single-gene alignments when using varying numbers of tree searches. a) Difference in log-likelihood score between the best-of-100 ML gene tree topology found from R100 and the best-observed ML gene tree topology found from R1, R10, R20, R40, R60, or R80 for a given ML program. Each bar denotes the mean value with standard deviation. b) Topological difference between the best-of-100 ML gene tree topology and the best-observed ML gene tree topology. The topological difference was the normalized Robinson–Foulds (nRF) distance between the best-of-100 gene tree topology and the best-observed gene tree topology. Each bar denotes the mean value with standard deviation. c) Compositions of all 19,414 inferred ML gene trees when using varying numbers of tree searches for a given ML program. Comparing with the best-of-100 ML gene tree topology found from R100, we assigned the best-observed ML gene tree topology found from R1, R10, R20, R40, R60, or R80 into each of three categories: i) the best-observed ML gene tree has significantly lower log-likelihood score than the best-of-100 ML gene tree topology; ii) the best-observed ML gene tree topology does not have significantly lower log-likelihood score than the best-of-100 ML gene tree topology; iii) the best-observed ML gene tree topology is the best-of-100 ML gene tree topology. We used the AU test to evaluate whether the best-observed ML gene tree topology has a significantly lower log-likelihood score than the best-of-100 ML gene tree topology or not.

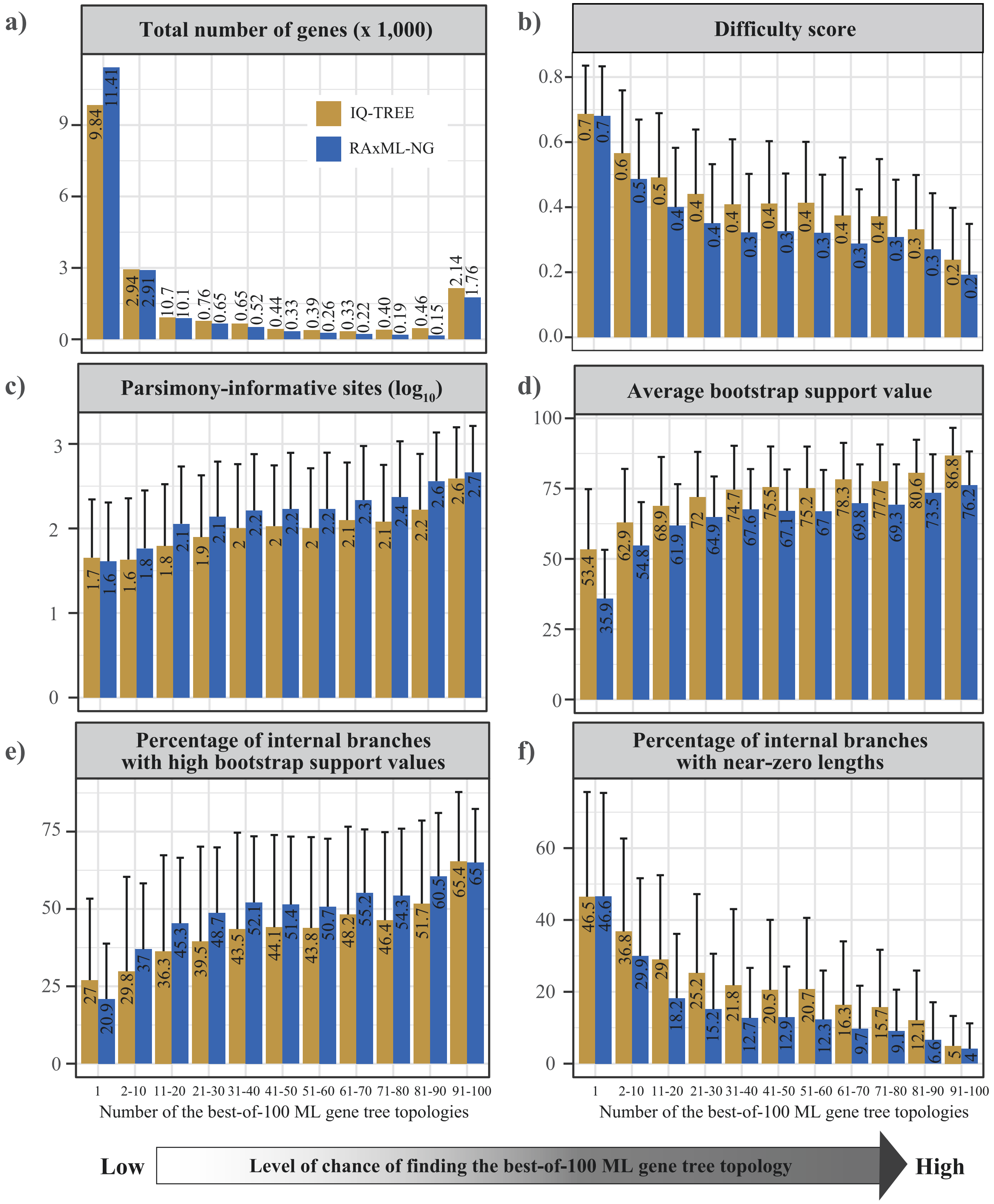

FIGURE 4. Characteristics of gene alignments that had different chances to achieve the best-of-100 ML gene tree topologies. We assigned all 19,414 single-gene alignments from 15 phylogenomic studies into eleven groups according to the number of the best-of-100 ML gene tree topologies observed out of 100 tree searches. a) Total number of genes that achieved the best-of-100 ML gene tree topologies in each of 11 groups. b) Difficulty score of gene alignment in each of 11 groups. c) Number of parsimony-informative sites in gene alignment in each of 11 groups. d) Average bootstrap support across the best-of-100 ML gene tree topologies in each of 11 groups. Note that IQ-TREE introduces an ultrafast bootstrap, while RAxML-NG introduces a standard nonparametric bootstrap. e) Percentage of internal branches with high bootstrap support values across the best-of-100 ML gene tree topology (IQ-TREE: > 90 ultrafast bootstrap value; RAxML-NG: > 70 standard bootstrap value) in each of 11 groups. f) Percentage of internal branches with near-zero lengths (that is the internal branch length of <0.0001) in each of 11 groups. Each bar denotes the mean value with standard deviation in panels b–f.

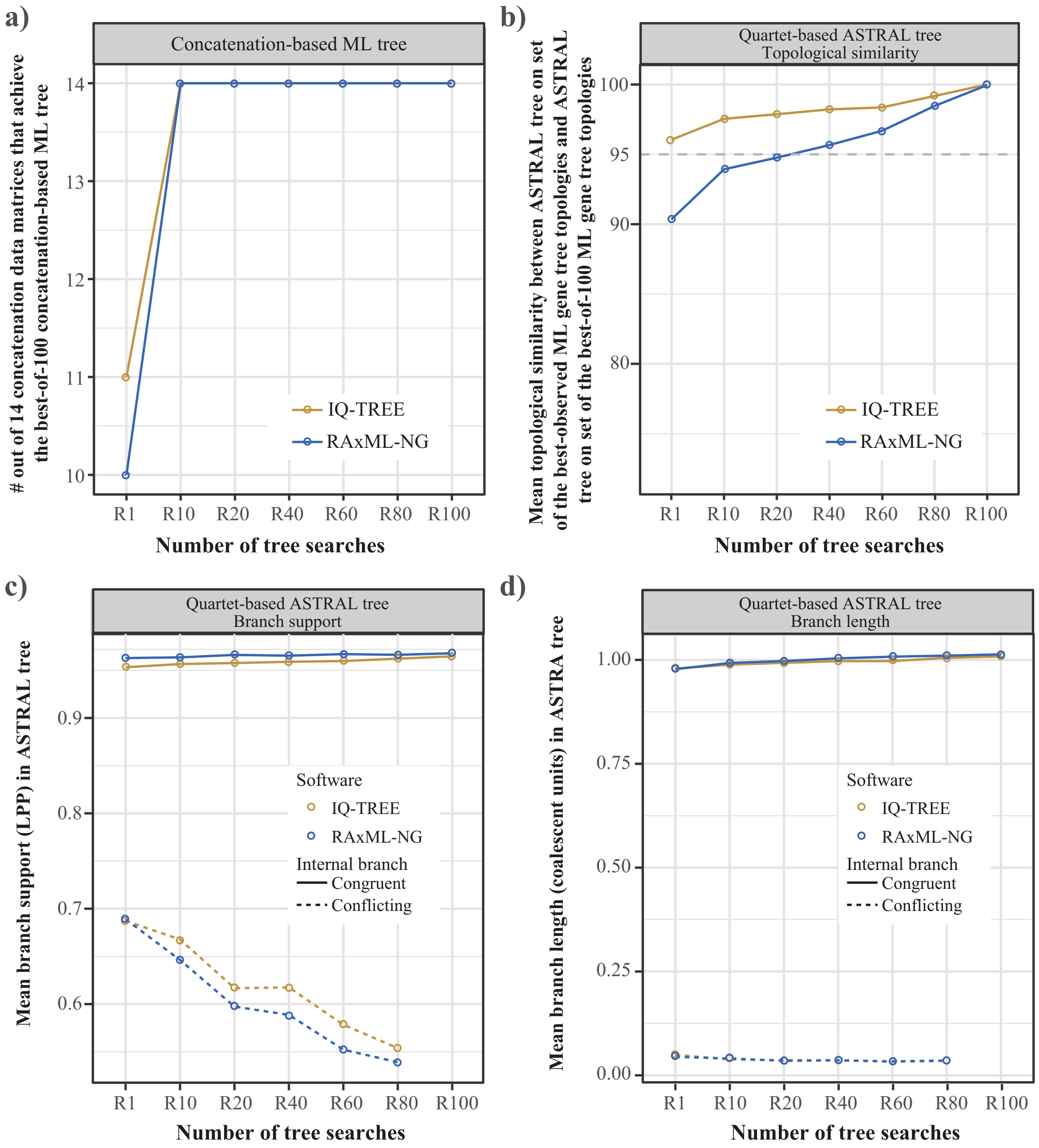

FIGURE 5. Effect of varying numbers of tree searches on concatenation- and quartet-based species tree estimations. a) For concatenation-based phylogenetic inference, we used 100 tree searches with IQ-TREE and RAxML-NG and compared the best-of-100 concatenation-based ML tree to the best-observed concatenation-based ML tree inferred from R1, R10, R20, R40, R60, or R80. b) For quartet-based phylogenetic inference, the quartet-based species tree was reconstructed from all the best-observed ML gene trees inferred using 1, 10, 20, 40, 60, 80, or 100 tree searches, respectively. We performed this analysis for each of 15 phylogenomic datasets with wASTRAL-h (Zhang and Mirarab 2022). To assess the effect of varying numbers of tree searches on the ASTRAL tree in terms of topology (panel b), mean branch support (panel c), and mean branch length in coalescent units (panel d), we compared the ASTRAL tree on the best-observed ML gene trees inferred from 1, 10, 20, 40, 60, or 80 tree searches to the ASTRAL tree on the best-of-100 ML gene trees inferred from 100 tree searches, respectively. In panels **c** and **d**, congruent internal branches denote bipartitions in ASTRAL trees on the best-observed ML gene trees inferred from a given number of tree searches that match those in ASTRAL trees on the best-of-100 ML gene trees inferred from 100 tree searches. Conflicting internal branches indicate bipartitions in ASTRAL trees on the best-observed ML gene trees inferred from a given number of tree searches that differ from those in ASTRAL trees on the best-of-100 ML gene trees inferred from 100 tree searches.

reconstructed from the set of single-gene trees inferred by 100 tree searches as the ASTRAL reference tree for each dataset. Note that the ASTRAL phylogenies were not directly comparable between IQ-TREE and RAxML-NG. This is because the single-gene trees used as input for ASTRAL estimations were independently inferred by each respective ML program. Overall, ASTRAL trees inferred using R1, R10, R20, R40, R60, and R80 showed increasing topological similarities to the ASTRAL reference trees (from 96% to 99% for IQ-TREE and from 91% to 98% for RAxML-NG) (Fig. 5b, Supplementary Fig. S16, and Supplementary Table S4). Notably, 6 of the 15 phylogenomic datasets (e.g., Bees and Green plants) had larger numbers of taxa than the remaining nine datasets (e.g., Rodents and Budding yeasts) (on average, the former has 327 taxa and the latter has 84 taxa), all ASTRAL trees inferred from 1 to 80 tree searches differed topologically from the ASTRAL reference trees inferred from 100 tree searches (Supplementary Fig. S16).

To further examine whether more tree searches would benefit the branch support and branch length estimations in quartet-based species phylogeny, we compared each inferred ASTRAL tree (using R1, R10, R20, R40, R60, or R80) with the ASTRAL reference tree (using R100) and examined the supports and lengths of congruent and conflicting internal branches, respectively. We found that the mean support values and mean branch lengths of all congruent internal branches increased as the number of tree searches increased (Fig. 5c,d and Supplementary Figs. S17 and S18), while the mean support values and mean branch lengths of all conflicting internal branches decreased as the number of tree searches increased. These results suggest that increasing the number of tree searches in single-gene tree inferences could benefit the branch support and branch length estimations in the ASTRAL species phylogeny.

Genes With Lower Difficulty Scores Had a Higher Chance of Finding the Best-of-100 ML Gene Tree Topologies and Were More Likely to Recover the Correct Trees

Since the true single-gene phylogenies for the 15 empirical phylogenomic datasets are unknown, it is impossible to precisely assess whether the best-observed gene tree topologies achieved with increasing numbers of tree searches would be more accurate. To address this issue, we adopted 20,000 simulated deoxyribonucleic acid (DNA) sequence alignments from a previous study (Höhler et al. 2022a), in which each gene alignment was simulated on the empirical data-derived gene tree and the model parameters in the RAxMLGrove database (Höhler et al. 2022b). For each of 20,000 simulated DNA sequence alignments, we conducted 100 tree searches using 100 runs for IQ-TREE, 50 parsimony starting trees and 50 random starting trees for RAxML-NG, and 1 BioNJ starting tree, 50 parsimony starting trees, and 49 random starting trees for PhyML (see Supplementary Text for details), respectively.

Analysis of these 20,000 simulated DNA sequence alignments showed that the recovery rate (that is the fraction of the 20,000 simulated gene alignments that recovered their best-of-100 ML gene tree topologies for a given number of tree searches) increased with increasing number of tree searches (Fig. 6a and Supplementary Table S5). Consistent with the findings from the 19,414 empirical gene alignments, we also found that the difficulty score of the simulated sequence alignment also exhibited the strongest correlation with the chance of finding the best-of-100 ML gene tree topology (Supplementary Fig. S19 and Supplementary Table S6). Therefore, following a previous study (Togkousidis et al. 2023), we divided the 20,000 simulated DNA gene alignments into 3 groups: the easy alignments (9560 genes with difficulty scores below 0.3), the intermediate alignments (8926 genes with difficulty scores between 0.3 and 0.7), and the difficult alignments (1514 genes with difficulty scores above 0.7). As expected, we found that the easy alignments tended to have the highest recovery rates of finding the best-of-100 ML gene tree topologies than the intermediate and difficult alignments (Fig. 6a).

Next, we examined the difference in log-likelihood scores between the best-observed gene tree topologies inferred from R1, R10, R20, R40, R60, R80, and R100 and the true gene trees, respectively. We found that the vast majority (83.9% for IQ-TREE, 83.7% for RAxML-NG, and 77.8% for PhyML) of 20,000 inferred gene trees had higher log-likelihood scores than the true gene trees (Fig. 6b). Furthermore, we noted a rise in the count of gene trees with higher log-likelihood scores than the true gene trees as the number of tree searches increased for RAxML-NG and PhyML. However, for IQ-TREE, the number of gene trees with higher log-likelihood scores than the true gene trees remained relatively constant across 100 tree searches (Fig. 6b). These results suggest that RAxML-NG and PhyML outperform IQ-TREE in terms of likelihood optimization. In addition, we used the AU test to evaluate whether the best-observed gene tree topology had a significantly higher log-likelihood score than the true tree or not. Our results show that among 20,000 inferred gene trees, 3650 (18.25%) IQ-TREE-inferred gene trees, 3346 (16.73%) RAxML-NG-inferred gene trees and 3040 (15.20%) PhyML-inferred gene trees had significantly higher log-likelihood scores than the true gene trees (AU test; P value ≤ 0.05) when using one tree search (Supplementary Fig. S20). As expected, with an increase in the number of tree searches, the percentage of the best-observed gene tree topologies that had significantly higher log-likelihood scores than the true gene trees increased, albeit weaker in magnitude (Supplementary Fig. S20).

To further investigate the topological accuracies of the best-observed gene tree topologies inferred from R1, R10, R20, R40, R60, R80, and R100, we examined the accuracy as measured by the quartet similarity between the inferred best-observed tree topology and the true tree using the R package Quartet (v1.2.5) (Smith 2019). Overall, we found that the mean topological

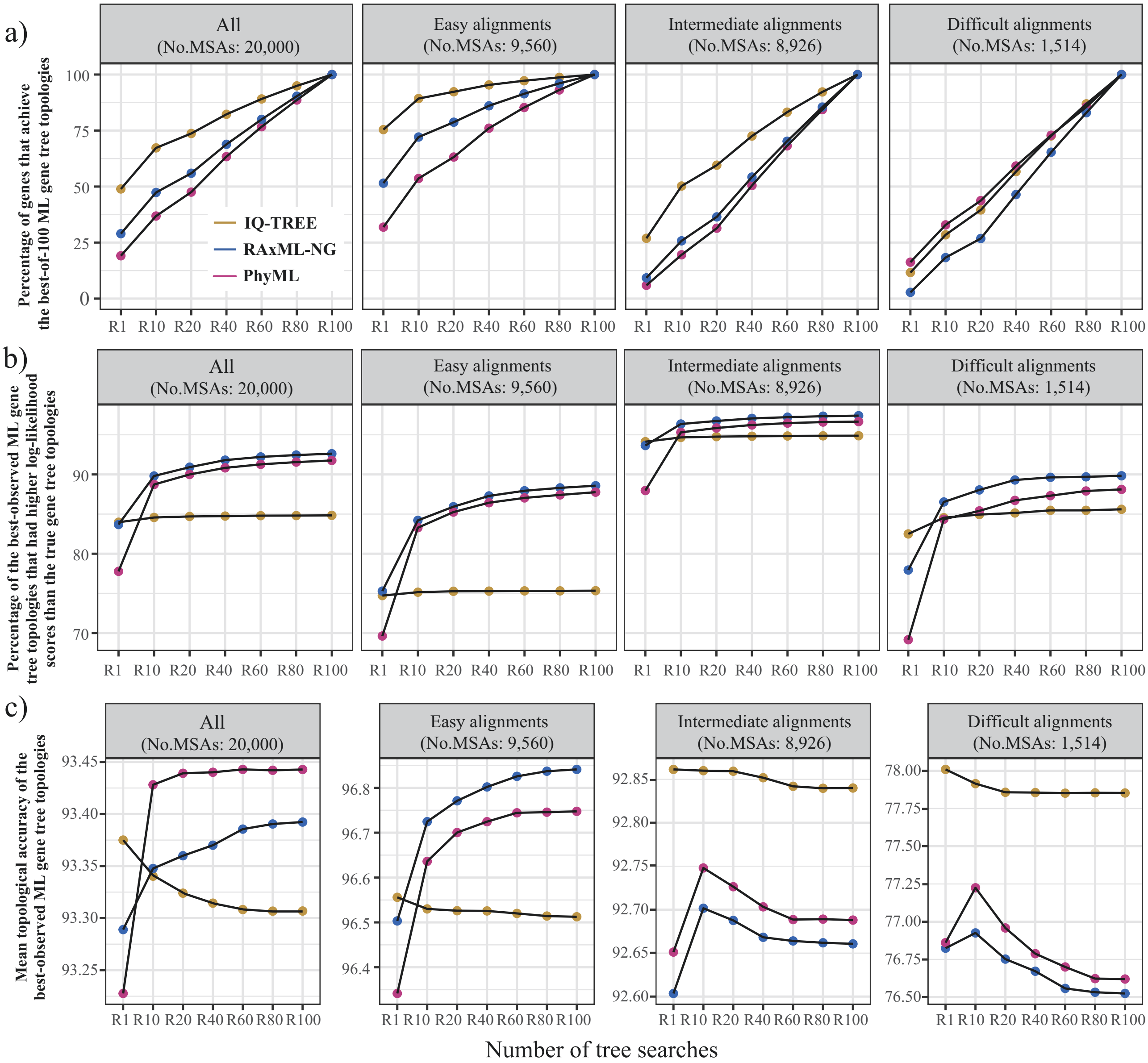

FIGURE 6. Gene alignments with lower difficulty scores had a higher chance of finding the best-of-100 ML gene tree topologies and are more likely to yield correct tree topologies. For each of 20,000 simulated DNA sequence alignments from the previous study (Höhler et al. 2022a), we conducted 100 tree searches using 100 runs for IQ-TREE, 50 parsimony starting trees and 50 random starting trees for RAxML-NG, and one BioNJ starting tree, 50 parsimony starting trees and 49 random starting trees for PhyML, respectively. a) Percentage of genes that achieved the best-of-100 ML gene tree topologies when using varying numbers of tree searches. b) Percentage of the best-observed ML gene tree topologies found at varying numbers of tree searches that had higher log-likelihood scores than true gene tree topologies. c) Mean topological similarity between the best-observed ML gene tree topologies found at varying numbers of tree searches and the true gene tree topologies. The topological similarity was quantified by the quartet distance between the best-of-100 ML gene tree topology and the true gene tree topology using the R package Quartet (v1.2.5) (Smith 2019). Following the previous study (Togkousidis et al. 2023), we divided the 20,000 simulated gene alignments into three groups: the easy alignments (9560 genes with a difficulty score below 0.3), the intermediate alignments (8926 genes with a difficulty score between 0.3 and 0.7), and the difficult alignments (1514 genes with a difficulty score above 0.7).

accuracy was 93.37% for the IQ-TREE-inferred gene trees, 93.29% for the RAxML-NG-inferred gene trees, and 93.23% for the PhyML-inferred gene trees when using one-tree search (Fig. 6c) for all 20,000 simulated alignments. As the number of tree searches increased, RAxML-NG and PhyML demonstrated an increase in mean topological accuracy, whereas IQ-TREE exhibited a gradual decrease in mean topological accuracy (albeit with a weaker magnitude of change). This trend is consistent with observations in the easy datasets but not in the intermediate and difficult datasets (Fig. 6c). In the intermediate and difficult datasets, RAxML-NG and PhyML displayed an increase in mean topological accuracy from R1 to R10, but they gradually decreased their mean topological accuracies after R10. IQ-TREE gradually decreased the mean topological accuracies from R1 to R100, although it outperforms RAxML-NG and PhyML in terms of mean topological accuracy. Finally, we found that inferred gene trees on the easy alignments tended to be topologically more similar to the true trees than those on the intermediate alignments and the difficult alignments (Fig. 6c). Collectively, our findings indicate that: 1) easy gene alignments have a higher chance of finding the best-of-100 ML gene tree topology; 2) these alignments tend to generate more topologically accurate gene trees; and 3) the accuracies of their inferred gene trees increase with increasing number of tree searches, although this improvement is program-dependent. For more difficult gene alignments, there is a noticeable pattern of overfitting, particularly with IQ-TREE at R1 and RAxML-NG and PhyML at R10, where these methods perform slightly better than at R100. The overfitting tendencies observed are relatively small, with differences in mean topological accuracy being less than 1%.

Discussion

This study aimed to uncover the effect of the number of tree searches on ML phylogenetic inference in phylogenomics. To achieve this goal, we carried out computationally extensive analyses of 19,414 single-gene alignments from 15 phylogenomic datasets. These 15 phylogenomic datasets, with the number of taxa ranging from 15 to 343, along with the green plant dataset containing 1178 taxa, represent different data types (non-coding DNA, coding DNA, and amino acid) and cover a broad taxonomic range from genus to phylum. In addition, we used 20,000 simulated DNA sequence alignments, with the number of taxa ranging from 4 to 469 and an average of 69, to examine the effect of varying the number of tree searches on the accuracy of single-gene tree estimation. To the best of our knowledge, our work represents the most comprehensive investigation on this topic to date in terms of both breadth and depth.

Do different starting trees affect ML phylogenetic inference? Modern ML phylogenetic inference typically begins with a starting tree and then executes an iterative, hill-climbing process using the rearrangement operations, to infer a nearly optimal tree (Chor and Tuller 2005). Our results show that when one independent tree search was executed using 1 CPU for a given ML program, the random starting tree was less efficient in finding the best-of-3 ML gene tree than the BioNJ and parsimony starting trees. Furthermore, the use of random starting tree had slightly longer runtimes than the uses of BioNJ and parsimony starting trees. Finally, the random starting tree generally is more different from the ML tree than the BioNJ and parsimony starting trees. Therefore, we suggest that considering BioNJ as one of the starting trees in current fast ML-based programs, such as RAxML-NG, would be helpful.

What is the general impact of the number of tree searches on ML phylogenetic inference? Our study found that ~69% of single-gene trees achieved their highest log-likelihood scores with ten tree searches (R10), which is computationally tractable for most phylogenomic studies. However, with R10, it is very rare for single-gene trees to reach their highest log-likelihood scores for 4 animal datasets (Bees, Birds, Butterflies, and Marine fishes), 1 plant dataset (Green plants), and 1 fungal dataset (Budding yeasts), all of which contain > 100 taxa. As single-gene trees are commonly used as input for species tree estimations, such as quartet-based ASTRAL species phylogeny (Mirarab et al. 2014; Zhang et al. 2018; Zhang and Mirarab 2022), we assessed the influence of the number of tree searches on the ASTRAL estimation. We found that all ASTRAL species phylogenies reconstructed on single-gene trees inferred using 1, 10, 20, 40, 60, or 80 tree searches differed topologically from the ASTRAL species phylogeny reconstructed on single-gene trees inferred using 100 tree searches for 6/15 phylogenomic datasets, although the differences in topology were usually small. Given that quartet-based species tree estimation relies on the accuracy of single-gene tree estimations, our results suggest that quartet-based analyses could take into account the effect of varying numbers of tree searches on single-gene ML tree estimations. Specifically, conducting more tree searches would at least benefit the branch support and branch length estimations in quartet-based species phylogeny. In addition, our results suggest that the use of 10 tree searches is sufficient to generate a robust ML tree for concatenation-based analysis.

Is conducting extensive tree searches in ML phylogenetic inference necessary? Our results suggest that increasing the number of tree searches improves log-likelihood score. At the same time, the marginal return of conducting more than 10 tree searches varied substantially among 15 phylogenomic data sets. Furthermore, both empirical and simulated datasets showed that the difficulty score of gene alignments exhibited the strongest correlation with the chance of finding the best-of-100 ML gene tree topologies. Easy gene alignments are more likely to discover the best-of-100 ML gene tree topologies and produce more accurate phylogenies compared to intermediate and difficult gene alignments. A recent phylogenetic study demonstrated that the difficulty score can directly reflect the amount of phylogenetic signal in the input gene alignment (Togkousidis et al. 2023). Based on the results from simulation datasets, for easy gene alignments (e.g., difficulty score ≤ 0.3) increasing the number of tree searches is beneficial for ML phylogenetic inference in terms of likelihood optimization and topological accuracy. Conversely, for intermediate and difficult gene alignments, extensive tree searches may be unnecessary or detrimental.

In summary, our study solely investigated how the number of tree searches influences ML phylogenetic inference within each program, rather than comparing across different ML programs. Since different ML phylogenetic programs involve varying numbers of tree searches during a single default run, fairly comparing the effects of the number of tree searches on ML phylogenetic inference across different programs would require ensuring an equal amount of running time. In addition, we found that difficulty score could serve as a useful predictor for estimating the necessary number of tree searches (Togkousidis et al. 2023). If computational resources permit, conducting at least 20 tree searches is recommended for IQ-TREE, and at least 10 tree searches for RAxML-NG and PhyML.

MATERIAL AND METHODS

Empirical Phylogenomic Datasets

We downloaded all 19,414 gene alignments from 15 phylogenomic studies in animals (6), plants (5), and fungi (4) as of 10 March 2021 (https://figshare.com/articles/dataset/Irreproducibility_of_maximum_likelihood_phylogenetic_inference/11917770?file=24764333) from a recent phylogenomic study (Shen et al. 2020) (Table 1). These 15 phylogenomic datasets were constructed using 5 different gene sampling approaches, namely Ultraconserved Element (UCE) capture, Anchored Hybrid Enriched (AHE) capture, conserved exon capture, transcriptome sequencing, and whole genome sequencing. They also represented a wide range of data types (non-coding DNA (DNA), exon (DNA), and amino acid (AA)) and a broad taxonomic range from genus to phylum. All 19,414 gene alignments in FASTA form can be found on the figshare repository (http://dx.doi.org/10.6084/m9.figshare.17086259).

Assessment of Effect of Different Starting Trees on Maximum Likelihood Gene Tree Inferences

To investigate the effect of different starting trees on the single-gene ML tree inferences, we conducted one independent tree search from BioNJ tree, parsimony tree, or random tree for IQ-TREE (version 1.6.12) (Nguyen et al. 2015), RAxML-NG (version 0.9.0) (Kozlov et al. 2019), and PhyML (version 3.3.20220408) (Guindon et al. 2010), each using 1 CPU on the same compute node (AMD EPYC 7662 @ 2.0 GHz processor with 128 threads). Since executing all tree searches on a single node was computationally expensive, we sampled 200 genes from each of 15 animal, plant, and fungal phylogenomic datasets (Table 1). The total number of gene alignments is 3000.

For each gene alignment, we first generated 3 different starting trees including BioNJ tree, parsimony tree, and random tree. Next, for each of the 3 different starting trees, we executed one independent ML tree search using IQ-TREE, RAxML-NG, and PhyML, respectively. Thereby we obtained 3 ML trees inferred from three starting trees for each of 3000 gene alignments. Last, for a given gene alignment and a given ML program, we denoted the ML gene tree with the highest log-likelihood score as the best-of-3 ML gene tree topology within a given ML program. The specific command line instructions and parameter settings for generating 3 different starting trees and executing ML gene tree inferences are given in the Supplementary text.

Assessment of Effect of Varying Numbers of Tree Searches on the Performance of Finding Gene Tree with the Highest Log-Likelihood Score

To take into account variation that may stem from different tree rearrangement algorithms used in heuristic search, we used both the NNI-based IQ-TREE and the SPR-based RAxML-NG to assess the effect of varying numbers of tree searches on the performance of finding gene tree with the highest log-likelihood score.

Given that the true highest likelihood score is unknown and that increasing the number of tree searches in heuristic searches is computationally very expensive, we limited the number of tree searches to 100, which is much higher than the number of tree searches used in all 15 original phylogenomic studies. To conduct 100 tree searches for all 19,414 alignments, we divided them into 5 sets, with each set running 20 tree searches. Following the tree search strategy of a recent study (Kozlov et al. 2019), we ran 20 tree searches with IQ-TREE utilizing the option "-runs 20 -seed random number" and RAxML-NG utilizing the option "--tree pars{10},rand{10} --seed random number." Two examples of the specific command line instructions and parameter settings for running 100 tree searches in IQ-TREE and RAxML-NG for a DNA sequence alignment and an amino acid alignment are given in the Supplementary text.

Overall, we executed 194,140 jobs (19,414 alignments × 5 times × 2 ML programs). Each job was run on a single node with 2 threads and 2 GB RAM on the Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison and the Center for Engineering and Scientific Computation (CESC) at Zhejiang University

Note that, the log-likelihood scores typically differ among distinct ML inference tools due to different round-off error propagation or subtle differences in the numerical implementation of model parameter optimization routines. In order to avoid a possible bias, we chose to reevaluate log-likelihood scores of 100 IQ-TREE-inferred gene trees and 100 RAxML-NG-inferred gene trees for each gene alignment with IQ-TREE. For a given gene alignment, once 100 extensive tree searches were completed, their log-likelihood scores were re-calculated using iqtree (iqtree -safe -nt 2 -seed 369284957 -quiet -me 0.0001 -m $model -s $fas_id.fasta -te $in.tre -pre $pre_id; note that the "-me" option specifies that the log-likelihood scores are calculated from a precision of 4 decimal places) on a laboratory server. We then labeled the 100 runs with R1 to R100 and recorded which runs achieved the highest log-likelihood scores (we referred them as best-of-100 ML gene tree). Specifically, we compared each of the gene trees (R1 to R100) against the best-of-100 ML tree topology in terms of topology and log-likelihood score. For a given run, we considered the gene tree as the best-of-100 tree topology if i) its topology was identical to that of the best-of-100 ML tree topology; or ii) if its log-likelihood score was identical to the highest log-likelihood score, despite of any topological differences. Note that ML programs have limited numerical precision for log-likelihood score calculation, which could lead to different gene topologies having identical log-likelihood scores in output files (Haag et al. 2023).

Assessment of Effect of Varying Numbers of Tree Searches on Concatenation- and Quartet-Based Species Tree Estimations

Since running 100 tree searches for the concatenation-based ML inference is computationally very expensive, we sampled the first 200 genes from each of the 15 phylogenomic datasets. For each of the 15 phylogenomic studies, we first concatenated each study's 200 genes into a supermatrix and then ran 100 tree searches for inferring concatenation-based species phylogeny using IQ-TREE and RAxML-NG.

The command line instructions and parameter settings for running 100 tree searches for a supermatrix in IQ-TREE and RAxML-NG are exactly same to those used for a single-gene alignment, except for the number of threads. Running 100 tree searches for a supermatrix used 16 threads ("-nt 16" in IQ-TREE; "--threads 16" in RAxML-NG) instead of 2 threads as for single-gene alignment analyses ("-nt 2" in IQ-TREE; "--threads 2" in RAxML-NG). Two examples of the specific command line instructions and parameter settings for running 100 tree searches in IQ-TREE and RAxML-NG for DNA and amino acid supermatrices are given in the Supplementary text.

Note that only 14 concatenation-based species phylogenies were successfully used to investigate the effect of varying numbers of tree searches on concatenation-based species tree estimations, because inferring 100 concatenation-based ML tree searches for the green plant phylogenomic data set (200 genes and 1178 taxa—by far the largest in its number of taxa) using IQ-TREE and RAxML-NG on a 48-CPU node failed to finish after 5 months. All analyses of concatenation-based species trees were executed on four laboratory servers. All 14 supermatrices and their concatenation-based species trees are available on the figshare repository.

For each of the 15 phylogenomic studies, we also reconstructed their quartet-based species trees from all individual gene trees with wASTRAL-h v1.3 (Zhang and Mirarab 2022), a weighted ASTRAL (Mirarab et al. 2014; Zhang et al. 2018) program that takes into account phylogenetic uncertainty by integrating signals from branch length and branch support in the set of input gene trees to improve quartet-based species tree inference (Zhang and Mirarab 2022). To investigate the effect of varying numbers of tree searches on quartet-based species tree estimations, we created 7 sets of the best-observed gene tree topologies inferred using 1, 10, 20, 40, 60, 80, or 100 tree searches and then reconstructed their quartet-based species trees with wASTRAL-h. All 105 sets of gene trees (15 studies ×7 sets) and their quartet-based species trees are available on the figshare repository. The specific command line instructions and parameter settings for inferring their quartet-based ASTRAL species trees are given in the Supplementary text.

Using simulated data to examine the accuracy of gene tree estimation in relation to the number of tree searches

To examine the difference between the best-of-100 tree topologies that had different chances to be found among 100 tree searches, we used 20,000 simulated DNA sequence alignments from a previous study (Höhler et al. 2022a), in which each gene alignment was simulated on the empirical data-derived gene tree and the model parameters in the RAxMLGrove database (Höhler et al. 2022b), which contains RAxML and RAxML-NG users' phylogenetic data on 2 web-servers (https://github.com/angtft/RAxMLGrove and https://www.phylo.org/index.php). Note that all taxon names in the simulation data sets were changed to artificial taxon IDs (e.g., taxon1).

For each of 20,000 simulated DNA sequence alignments, we conducted 100 tree searches using 100 runs for IQ-TREE, 50 parsimony starting trees, and 50 random starting trees for RAxML-NG, and 1 BioNJ starting tree, 50 parsimony starting trees, and 49 random starting trees for PhyML (see Supplementary text for details). For a given number of tree searches, the accuracy of gene tree estimation was calculated as the topological similarity between its single-gene tree and the reference true tree using the quartet distance (topological similarity = 1 − quartet distance) in the R package Quartet (v1.2.5) (Smith 2019).

Topological hypothesis testing

We used the AU test (Shimodaira 2002) in IQ-TREE to determine whether 2 gene trees were significantly different or not. Specifically, i) for each of 19,414 empirical gene alignments, we assessed whether the best-observed gene tree topology inferred from using R1, R10, R20, R40, R60, or R80 had significantly lower log-likelihood score than the best-of-100 gene tree topology found among 100 runs; ii) for each of 20,000 simulated gene alignments, we assessed whether the best-observed gene tree topology inferred from using R1, R10, R20, R40, R60, R80, or R100 had significantly higher log-likelihood score than the true gene tree (i.e., reference tree). The specific command line instructions and parameter settings for running AU tests are given in the Supplementary text.

Statistical analyses

All statistical analyses were performed in R (v. 3.6.3). Pearson's correlation coefficient was used to test for correlations among different variables. All bar or dot plots were generated using the ggplot2 package (Wickham 2009) in R.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: https://dx.doi.org/10.5061/dryad.rv15dv4b7.

ACKNOWLEDGMENTS

We thank Dimitri Höhler for kindly providing the simulation gene alignments. We also thank the Associate Editor Olivier Gascuel and the reviewer Alexis Stamatakis as well as 3 anonymous reviewers for their insightful comments that improved this work.

CONFLICT OF INTEREST

A.R. is a scientific consultant for LifeMine Therapeutics, Inc. The other authors declare no conflict of interest.

FUNDING

This work was conducted in part using the resources of the Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison and of the Center for Engineering and Scientific Computation (CESC) at Zhejiang University. X.X.S. was supported by the National Key R&D Program of China (2022YFD1401600), the National Science Foundation for Distinguished Young Scholars of Zhejiang Province (LR23C140001), the Fundamental Research Funds for the Central Universities (226-2023-00021), and the key research project of Zhejiang Lab (2021PE0AC04). X.Z. was supported by the Key-Area Research and Development Program of Guangdong Province (2018B020205003 and 2020B0202090001). C.T.H. was supported by the National Science Foundation (DEB-1442148 and DEB- 2110403), the USDA National Institute of Food and Agriculture (Hatch Project Nos. 1020204 and 7005101), in part by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science No. DE-SC0018409), and the Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (H. I. Romnes Faculty Fellow). Y.C. was supported by the National Natural Science Foundation of China (31922074) and the Fundamental Research Funds for the Central Universities (2021FZZX001-31). R.P. was supported by the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (grant no. 2019R01002). X.X.C. was supported by the Key International Joint Research Program of National Natural Science Foundation of China (No. 31920103005). Research in A.R.'s lab was supported by grants from the National Science Foundation (DEB-1442113 and DEB-2110404), the National Institutes of Health/ National Institute of Allergy and Infectious Diseases (R56 AI146096 and R01 AI153356), and the Burroughs Wellcome Fund.

DATA AVAILABILITY

All gene alignments, gene trees, and command lines, as well as summary statistics of the runs, are available on the figshare repository http://dx.doi.org/10.6084/m9.figshare.17086259.

REFERENCES

Alfaro M.E., Faircloth B.C., Harrington R.C., Sorenson L., Friedman M., Thacker C.E., Oliveros C.H., Černý D., Near T.J. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nat. Ecol. Evol. 2:688–696.

Allen B.L., Steel M. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Ann. Combinatorics. 5:1–15.

Blaimer B.B., Mawdsley J.R., Brady S.G. 2018. Multiple origins of sexual dichromatism and aposematism within large carpenter bees. Evol. Int. J. Org. Evol. 72:1874–1889.

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. Syst. Biol. 66:352–366.

Chor B., Tuller T. 2005. Maximum likelihood of evolutionary trees: hardness and approximation. Bioinformatics. 21:i97–106.

Espeland M., Breinholt J., Willmott K.R., Warren A.D., Vila R., Toussaint E.F.A., Maunsell S.C., Aduse-Poku K., Talavera G., Eastwood R., Jarzyna M.A., Guralnick R., Lohman D.J., Pierce N.E., Kawahara A.Y. 2018. A comprehensive and dated phylogenomic analysis of butterflies. Curr. Biol. 28:770–778.e5.

Felsenstein J. 1978. The number of evolutionary trees. Syst. Zool. 27:27.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

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.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Haag J., Höhler D., Bettisworth B., Stamatakis A. 2022. From easy to hopeless—predicting the difficulty of phylogenetic analyses. Mol. Biol. Evol. 39:msac254.

Haag J., Hübner L., Kozlov A.M., Stamatakis A. 2023. The Free Lunch is not over yet—systematic exploration of numerical thresholds in maximum likelihood phylogenetic inference. Bioinforma. Adv. 3:vbad124.

Hamilton A. 2014. The evolution of phylogenetic systematics (species and systematics). Berkeley (CA): University of California Press.

Herrando-Moraira S., Calleja J.A., Carnicero P., Fujikawa K., Galbany-Casals M., Garcia-Jacas N., Im H.T., Kim S.C., Liu J.Q., López-Alvarado J., López-Pujol J., Mandel J.R., Massó S., Mehregan I., Montes-Moreno N., Pyak E., Roquet C., Sáez L., Sennikov A., Susanna A., Vilatersana R. 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). Mol. Phylogenet. Evol. 128:69–87.

Hillis D., Moritz C., Mable B.K. 1996. Molecular systematics. 2nd ed. Sunderland (MA): Sinauer Associates.

Höhler D., Haag J., Kozlov A.M., Stamatakis A. 2022a. A representative performance assessment of maximum likelihood based phylogenetic inference tools. bioRxiv. doi:10.1101/2022.10.31.514545

Höhler D., Pfeiffer W., Ioannidis V., Stockinger H., Stamatakis A. 2022b. RAxML Grove: an empirical phylogenetic tree database. Bioinformatics. 38:1741–1742.

Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 35:4453–4455.

Kumar S., Stecher G., Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870–1874.

Leavitt S.D., Grewe F., Widhelm T., Muggia L., Wray B., Lumbsch H.T. 2016. Resolving evolutionary relationships in lichen-forming fungi using diverse phylogenomic datasets and analytical approaches. Sci. Rep. 6:22262.

Li Y., Liu Z., Liu C., Shi Z., Pang L., Chen C., Chen Y., Pan R., Zhou W., Chen X., Rokas A., Huang J., Shen X.-X. 2022. HGT is widespread in insects and contributes to male courtship in lepidopterans. Cell 185:2975–2987.e10.

Liu K., Linder C.R., Warnow T. 2011. RAxML and fasttree: comparing two methods for large-scale maximum likelihood phylogeny estimation. PLoS One 6:e27731.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.

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.

Mitchell N., Lewis P.O., Lemmon E.M., Lemmon A.R., Holsinger K.E. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *protea* L. Am. J. Bot. 104:102–115.

Money D., Whelan S. 2012. Characterizing the phylogenetic treesearch problem. Syst. Biol. 61:228–239.

Morrison D.A. 2007. Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. Syst. Biol. 56:988–1010.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574:679–685.

Park M., Zaharias P., Warnow T. 2021. Disjoint tree mergers for largescale maximum likelihood tree estimation. Algorithms. 14:148.

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.

Robinson D.F. 1971. Comparison of labeled trees with valency three. J. Comb. Theory B 11:105–119.

Roycroft E.J., Moussalli A., Rowe K.C. 2020. Phylogenomics uncovers confidence and conflict in the rapid radiation of australo-papuan rodents. Syst. Biol. 69:431–444.

Shen X.-X., Li Y., Hittinger C.T., Chen X., Rokas A. 2020. An investigation of irreproducibility in maximum likelihood phylogenetic inference. Nat. Commun. 11:6096.

Shen X.-X., Opulente D.A., Kominek J., Zhou X., Steenwyk J.L., Buh K.V., Haase M.A.B., Wisecaver J.H., Wang M., Doering D.T., Boudouris J.T., Schneider R.M., Langdon Q.K., Ohkuma M., Endoh R., Takashima M., Manabe R., Čadež N., Libkind D., Rosa C.A., DeVirgilio J., Hulfachor A.B., Groenewald M., Kurtzman C.P., Hittinger C.T., Rokas A. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Cell. 175:1533–1545.e20.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Smith M.R. 2019. Quartet: comparison of phylogenetic trees using quartet and bipartition measures. Compr. R Arch. Netw. 10:1. doi:10.5281/zenodo.2536318.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Steenwyk J.L., Opulente D.A., Kominek J., Shen X.-X., Zhou X., Labella A.L., Bradley N.P., Eichman B.F., Čadež N., Libkind D., DeVirgilio J., Hulfachor A.B., Kurtzman C.P., Hittinger C.T., Rokas A. 2019a. Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol. 17:e3000255.

Steenwyk J.L., Shen X.-X., Lind A.L., Goldman G.H., Rokas A. 2019b. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. MBio 10:1–25.

Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. 1996. Phylogenetic inference. In: Hillis D.M., Moritz C., Mable B.K., editors. Molecular systematics. Sunderland (MA): Sinauer. p. 407–514.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Togkousidis A., Kozlov O.M., Haag J., Höhler D., Stamatakis A. 2023. Adaptive RAxML-NG: accelerating phylogenetic inference under maximum likelihood using dataset difficulty. Mol. Biol. Evol. 40:msad227.

Vinh L.S., Haeseler A.V. 2004. IQPNNI: moving fast through tree space and stopping in time. Mol. Biol. Evol. 21:1565–1571.

Wickham H. 2009. ggplot2. NY: Springer New York.

Wu M., Kostyun J.L., Hahn M.W., Moyle L.C. 2018. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. Mol. Ecol. 27:3301–3316.

Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K.-S., Carpenter E.J., Zhang Y., Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N., Smith S.A. 2015. Dissecting molecular evolution in the highly diverse plant clade caryophyllales using transcriptome sequencing. Mol. Biol. Evol. 32:2001–2014.

Yang Z. 2014. Molecular evolution: a statistical approach. Oxford: Oxford University Press.

Zhang C., Mirarab S. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. Mol. Biol. Evol. 39:msac215.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19:153.

Zhou X., Shen X.-X., Hittinger C.T., Rokas A. 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. Mol. Biol. Evol. 35:486–503.