The Impact of Species Tree Estimation Error on Cophylogenetic Reconstruction

Julia Zheng
Michigan State University
East Lansing, Michigan, USA

Yuya Nishida
Michigan State University
East Lansing, Michigan, USA

Alicja Okrasińska
University of Warsaw
Warsaw, Poland

Gregory M. Bonito
Michigan State University
East Lansing, Michigan, USA

Elizabeth A.C. Heath-Heckman
Michigan State University
East Lansing, Michigan, USA

Kevin J. Liu*
kjl@msu.edu
Michigan State University
East Lansing, Michigan, USA

ABSTRACT

Just as a phylogeny encodes the evolutionary relationships among a group of organisms, a cophylogeny represents the coevolutionary relationships among symbiotic partners. Both are primarily reconstructed using computational analysis of biomolecular sequence data. The most widely used cophylogenetic reconstruction methods utilize an important simplifying assumption: species phylogenies for each set of coevolved taxa are required as input and assumed to be correct. Many studies have shown that this assumption is rarely – if ever – satisfied, and the consequences for cophylogenetic studies are poorly understood.

To address this gap, we conduct a comprehensive performance study that quantifies the relationship between species tree estimation error and downstream cophylogenetic estimation accuracy. We study the performance of state-of-the-art methods for cophylogenetic reconstruction using *in silico* model-based simulations. Our investigation also includes assessments of cophylogenetic reproducibility using genomic sequence datasets sampled from two important models of symbiosis: soil-associated fungi and their endosymbiotic bacteria, and bobtail squid and their bioluminescent bacterial symbionts.

Our findings conclusively demonstrate the major impact that upstream phylogenetic estimation error has on downstream cophylogenetic reconstruction quality. Relative to other experimental factors such as cophylogenetic estimation method choice and coevolutionary event costs, phylogenetic estimation error ranked highest in importance based on a random forest-based variable importance assessment. We conclude with practical guidance and future research directions. In particular, among the many considerations needed for accurate cophylogenetic reconstruction – choice of cophylogenetic reconstruction method and method settings, sampling design, and others – just as much attention must be paid to careful species phylogeny estimation using modern best practices.

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3612964

CCS CONCEPTS

• **Applied computing** → *Computational genomics*; *Computational biology*; *Molecular sequence analysis*; *Molecular evolution*; *Computational genomics*; *Bioinformatics*; *Population genetics.*

KEYWORDS

cophylogeny, cophylogenetic reconciliation, species tree, simulation study, *Mortierella*, bobtail squid, symbiont, symbiosis

ACM Reference Format:
Julia Zheng, Yuya Nishida, Alicja Okrasińska, Gregory M. Bonito, Elizabeth A.C. Heath-Heckman, and Kevin J. Liu. 2023. The Impact of Species Tree Estimation Error on Cophylogenetic Reconstruction. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3584371.3612964

1 INTRODUCTION

A cophylogeny represents the evolutionary and coevolutionary relationships among multiple sets of coevolved taxa, and cophylogenies are widely used to study fundamental and applied topics throughout biology and the life sciences [4, 26]. For example, untangling coevolutionary histories is essential to reconstructing the web of life [49], as symbiosis and coevolution have played an important role in evolution at different scales – from genes to proteins, biomolecular pathways, organisms, populations, and beyond [22].

As is the case in phylogenetic estimation, cophylogenies are principally reconstructed using computational analyses of biomolecular sequences – increasingly abundant thanks to next-generation biomolecular sequencing technologies [43] – as well as other types of biological data [12]. The most widely used computational approach for cophylogenetic estimation consists of a multi-stage pipeline where: (1) a species tree is independently estimated for each coevolved set of taxa using the same approaches as in a traditional phylogenetic study, and (2) cophylogenetic analysis proceeds using the estimated species trees as input, alongside the known host and symbiont associations (Figure 1).

Many cophylogenetic methods have been developed and they fall into two broad categories. (1) Global-fit methods [4] evaluate overall congruence between host and symbiont tree topologies, and examples include PARAFIT [21], PACo [1], and MRCAlink [42]. (2) Event-based methods perform phylogenetic reconciliation using either parsimony-based optimization or, less commonly, model-based statistical optimization. These optimization problems are known

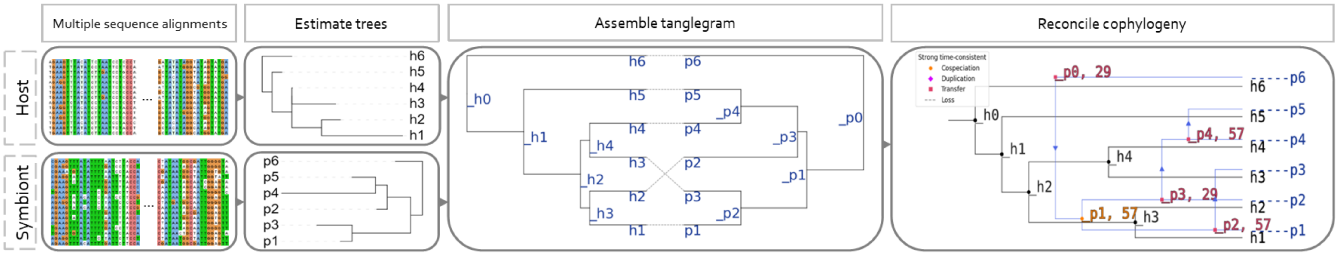


Figure 1: *A typical workflow for cophylogenetic reconstruction.* (1) Biomolecular sequence data for host taxa and symbiont taxa are aligned. (2) A host species tree and symbiont species tree are independently estimated using each multiple sequence alignment as input. (3) The input to cophylogenetic reconstruction consists of the estimated host tree, estimated symbiont tree, and known host/symbiont associations. All three can be visualized using a tanglegram. (4) Finally, a cophylogeny is reconstructed using the two species trees and host/symbiont associations as input. The cophylogeny maps topological structure in the host tree to corresponding topological structure in the symbiont tree based on shared coevolutionary history, where each element in the mapping corresponds to a coevolutionary event (e.g., a cospeciation event, a host shift event, etc.). Example dataset from [14].

to be computationally difficult [32]. Substantial research effort has been devoted to developing scalable and practical algorithms for this problem, which include eMPRess [41], Jane [9], Treemap [7], COALA [2], CoRe-PA [28], and TALE [27]. Event-based methods typically account for multiple types of coevolutionary events [6]: cospeciation (or codivergence or codifferentiation) involving both host and symbiont lineages, duplication of a symbiont lineage within a host lineage, loss of a symbiont lineage within a host lineage, and host shift (or host switch or host transfer) where a symbiont lineage's association transfers to a different host lineage. In this study, we focus on event-based cophylogenetic reconstruction methods to investigate a finer granularity of evolutionary and coevolutionary event reconstructions.

The multi-stage pipeline design requires a critically important assumption: the estimated species trees in the first stage are used directly in the second stage under the assumption that they are correct. However, it is well understood in traditional phylogenetics that many factors can cause phylogenetic estimation methods to return some degree of estimation error, and estimation errors introduced in upstream computational tasks are important factors to consider. For example, numerous studies have investigated the strong impact that upstream multiple sequence alignment error can have on subsequent gene tree estimation [23, 29]. But this insight conflicts with the prevailing assumption made by cophylogenetic reconstruction pipelines, as noted by [4].

The implications of this conflict must be carefully assessed. A rigorous examination of inter-related estimation error across a multi-stage cophylogenetic reconstruction pipeline is needed, and there is a lack of relevant experimental studies today [12]. The high-level findings can be qualitative (e.g., whether major effects occur or if downstream estimation is largely robust to upstream estimation error) or, more usefully, quantitative assessments (e.g., quantification of the relationship between estimation error in different pipeline stages.) Going further, more nuanced implications arise from context dependence (e.g., the extent to which the combination of estimation errors is modulated by evolutionary divergence) and other experimental factors (e.g., dataset size, taxon sampling density, etc.). Finally, these outcomes will point to important practical consequences for systematists and other researchers that study cophylogenies. As with other topics in phylogenetics, recommendations regarding best practices are needed to reconstruct accurate phylogenies and cophylogenies. Any reconstruction errors invite misinterpretation and incorrect conclusions in dependent analyses. For example, incorrectly estimated cophylogenetic relationships among a set of model organisms may yield spurious conclusions about the evolutionary processes under study (e.g., the relative frequency and significance of different coevolutionary processes). Another practical matter is empirical study design. Ideally, computational resources and research effort should concentrate on computational and experimental bottlenecks. Clarification of the above questions will help illuminate whether some, all, or none of traditional cophylogeny reconstruction pipelines merit careful consideration relative to other aspects of study design (e.g., taxon sampling, sequencing technology, sequencing effort, etc.).

To address this gap, we have undertaken a study to examine the relationship between upstream phylogenetic estimation error and downstream cophylogeny reconstruction accuracy. Our performance study utilizes both simulated and empirical datasets that span a range of dataset sizes, evolutionary divergence, and evolutionary scenarios.

2 METHODS

Our performance study included a comprehensive suite of simulated benchmarking datasets that spanned a range of evolutionary conditions. The simulation conditions differed in terms of number of taxa, sequence length, evolutionary divergence, and distribution of coevolutionary event types.

The simulation study experiments utilized two different procedures to simulate synthetic datasets (Supplementary Figure S1). First, the "mixed" simulations utilized an empirically estimated cophylogeny and its constituent species trees and host/symbiont associations as the models for *in silico* simulation of biomolecular sequence evolution. Second, a fully *in silico* set of simulations were

run using the forward-time cophylogeny model proposed by [13], which we refer to as the "forward" simulations. Cophylogenetic and phylogenetic method performance on each simulated dataset was then assessed with respect to model/reference cophylogenies and phylogenies.

We also performed comparative analyses of two empirical genomic sequence datasets. One empirical dataset consists of cephalopod hosts and their bacterial symbionts, which serve as a well-studied model of open symbiosis (i.e., partnerships arising from horizontal transmission between hosts and/or the environment); the other dataset was sampled from fungal hosts and their bacterial endosymbionts, which are an emerging model of closed symbiosis (i.e., partnerships whose coevolution involves strictly vertical descent over time). The two systems thus provide a comparative contrast along a spectrum of symbiotic partnership flexibility [34].

The combination of experimental approaches is a design choice in our study. Taken together, the simulation study and empirical dataset experiments represent an array of natural symbiotic systems – by design and by definition, respectively. Some differences between the experimental approaches are worth noting. The forward simulations provide ground truth coevolutionary histories that enable the analysis of cophylogenetic reconciliation accuracy, whereas the mixed simulation experiments use an estimated cophylogeny reconciliation as reference to analyze cophylogenetic reconciliation precision in the context of phylogenetic inference error. On the other hand, our study uses empirical datasets to assess cophylogenetic reconciliation reproducibility without prior knowledge of the true coevolutionary history.

2.1 Definitions

We now introduce mathematical background needed to describe the experimental procedures. Some of the notation and definitions follow [50].

A rooted phylogenetic tree $T_{\mathcal{N}} = (V_{\mathcal{N}}, E_{\mathcal{N}})$ is a rooted evolutionary history for a set of taxa $\mathcal{N}$. We note that many cophylogenetic reconstruction algorithms require rooted binary phylogenetic trees as input. The rooted binary tree $T_{\mathcal{N}}$ has a root ρ with in-degree zero and out-degree two, leaves $\mathcal{N} \subseteq V_{\mathcal{N}}$ where each leaf has out-degree zero and in-degree one, and inner nodes $v \in V_{\mathcal{N}} \backslash \mathcal{N}$ where each inner node has out-degree two and in-degree one. For each directed edge $(u, v) \in E_{\mathcal{N}}$, v is a child of u. Each edge is also denoted by e_v with branch length $bl(e_v) \in \mathbb{R}^+$. For vertices $u, v \in V_n$, u is an ancestor of v, $u \in anc(v)$, v is a descendent of u, and $u \in desc(v)$ if and only if u lies on the unique path from root ρ to v.

For a pair of rooted phylogenetic trees T_H and T_S denoting the evolutionary history of a set H of hosts and a set S of symbionts, respectively, T_H is the host tree and T_S is the symbiont tree. A mapping function $\phi(s, h) : S \times H \rightarrow \{0, 1\}$ denotes known interactions between the extant species of T_H and T_S, where $\phi(s, h) = 1$ means a symbiont is associated with a host, and otherwise $\phi(s, h) = 0$. The tuple (T_H, T_S, ϕ) serves as the input to cophylogenetic methods, and can be nicely visualized using a tanglegram. A cophylogenetic reconciliation or reconstruction is defined as the set of event associations $\mathcal{R} \subset V_S \times V_H$ between the internal nodes of the symbiont tree T_S and the internal nodes of the host tree T_S. For a symbiont s, an event association $(s, h) \in \mathcal{R}$ means h is one of the host species known to have been associated with s.

The unrooted version $U_{\mathcal{N}}$ of a rooted phylogenetic tree $T_{\mathcal{N}}$ can be obtained by converting all directed edges into undirected edges, deleting the root, and connecting its two outgoing edges into a single remaining edge. Equivalently, an unrooted binary tree $U_{\mathcal{N}}$ on the leaf set $\mathcal{N}$ has internal nodes with degree three and leaves with degree one, and each leaf represents a distinct taxon in the taxon set $\mathcal{N}$.

In our study, tree topology differences were evaluated with normalized Robinson-Fould (nRF) distances [38]. For two unrooted trees U_1 and U_2 with the same set of leaf nodes $\mathcal{N}$ and having bipartition sets B_1 and B_2 respectively, the Robinson-Fould (RF) metric is the cardinality of the symmetric difference between the sets of bipartitions that appear in U_1 and U_2, which is $|B_1 - B_2| + |B_2 - B_1|$. (Note that bipartitions corresponding to leaf edges are trivial since the latter must always appear, and trivial bipartitions do not contribute meaningfully to the RF calculation.) The normalized RF distance is calculated by dividing RF distance by the maximum RF distance between two trees with $|\mathcal{N}|$ taxa, which is $\frac{|B_1 - B_2| + |B_2 - B_1|}{2|\mathcal{N}| - 6}$. We note that the RF distance is a de facto standard for topological comparisons of phylogenetic trees involving the same set of taxa. Generalizations of the RF distance have been proposed for comparing phylogenetic trees with overlapping but non-identical sets of taxa (e.g., [24]), although we note that the issue does not arise in the context of our study due to the nature of our simulation and empirical dataset analysis procedures.

Reconciled cophylogenies were compared based on the calculation proposed by [50], which we refer to as cophylogenetic precision. We now define this calculation. Let $\mathcal{R}_A$ and $\mathcal{R}_B$ be the reconstructed event associations of all internal vertices from cophylogenetic reconciliations A and B, respectively. Then, the proportion of reconciled events in $\mathcal{R}_A$ that were also found in $\mathcal{R}_B$ is $\frac{|\mathcal{R}_B \cap \mathcal{R}_A|}{|\mathcal{R}_A|}$. Cophylogenetic precision factors in all coevolutionary event types that are accounted for by the cophylogenetic reconstruction methods in this study – i.e., cospeciation, duplication, loss, and host switch events.

2.2 Simulation study

Mixed simulations. The mixed simulations utilized empirically-based phylogenetic estimates to perform parametric sampling of synthetic biomolecular sequence data. The simulation procedure begins with the former: obtaining a pair of species trees and cophylogeny via empirical dataset analysis. Six empirical datasets were obtained from literature to sample a range of evolutionary scenarios and dataset types: from single-locus datasets with sequence length under 1 kb to next-generation-sequencing (NGS) multi-locus datasets with sequence length well over 1 Mb (Table 1). The sequence data were preprocessed and aligned using MAFFT v7.221 with default settings [19]. Species phylogenies were reconstructed from concatenated multiple sequence alignments under the General Time Reversible (GTR) model of nucleotide substitution with Γ model of rate heterogeneity [51] and midpoint rooted using RAxML v8.2.12 [45]. As methods eMPRess [41] and COALA [2] were limited to one-to-one host/symbiont associations; symbiont

taxa were subsampled as needed to address this limitation. Cophylogenetic events were estimated with eMPRess [41] from the host and symbiont phylogenies and host-symbiont associations.

Next, the resulting phylogenetic estimates for each empirical dataset served as the statistical model for downstream *in silico* simulation of biomolecular sequence data. Specifically, the reconstructed species trees (including branch lengths) and associated substitution model parameter estimates served as generative models from which multiple sequence alignments were simulated using Seq-Gen [37]; accordingly, we refer to the species trees as model trees. Note that the above cophylogeny was used for assessing methodological performance (see "Phylogenetic and cophylogenetic reconstruction and assessment" below) but was not directly used for simulations.

As noted above, our study is motivated by more nuanced questions beyond establishing the impact of upstream phylogenetic estimation error on downstream cophylogeny reconstruction. We also investigated how this relationship is modulated by two key contextual factors – the evolutionary divergence and number of taxa under study – via two additional simulation experiments. In simulations with varying evolutionary divergence, model tree branch lengths were multiplied by a scaling parameter h. We explored a range of settings for the parameter h where each set of experiments selected a setting from the set $\{0.1, 0.5, 1, 2, 5, 10\}$. The simulations with varying dataset size were conducted by modifying alignment lengths (as listed in Table 1) to 400,228 bp and 1,455,978 bp for host and symbiont, respectively. The modified lengths were adapted from the concatenated MSA lengths of the avian host dataset [35] and the avian feather lice parasite dataset [11].

Forward simulations. The forward simulations utilized the R-based [36] implementation of the Treeducken [13] version 1.1.0 software and its forward-time coalescent model to sample a model cophylogeny, along with its associated species trees and host/symbiont associations. The model cophylogeny and model trees served as the reference cophylogeny and reference trees, respectively, during subsequent performance assessments (see "Performance assessments" below). Model parameter settings (Table 2) were based on estimates from selected empirical datasets. The resulting five model conditions included a range of dataset sizes (i.e., number of taxa and sequence length), substitution rates, base frequency distributions, and coevolutionary event distributions (Table 3). Model trees were deviated away from ultrametricity using Moret et al. [30]'s approach with deviation factor $c = 2.0$ [31]. We used custom scripts to perform the ultrametricity deviation calculations. Sequence evolution was then simulated on each model tree using the same approach as in the mixed simulation procedure, resulting in host and symbiont MSAs.

Additional experiments varying evolutionary divergence were performed with the forward simulation procedure, where the scaling parameter h was assigned a value from $\{0.1, 0.5, 1, 2, 5, 10\}$.

Experimental replication. For each model condition, the procedure to simulate biomolecular sequence evolution was repeated to obtain 100 replicate datasets. Results are reported across all replicate datasets in each model condition.

Phylogenetic and cophylogenetic reconstruction and assessment. On each simulated dataset, phylogenetic trees were reconstructed under the GTR+Γ model and midpoint rooted using RAxML v8.2.12. The resulting phylogenetic estimates and host/symbiont associations were used by eMPRess [41] to perform cophylogenetic reconciliation using default settings. We also conducted additional eMPRess analyses using alternative cophylogenetic event costs that were estimated using COALA [2] and CoRe-PA [28]; the additional estimated cophylogenies were used in the random forest-based variable importance analyses described below (and additional experiments in the Supplementary Online Materials). (Also see Supplementary Online Materials section S7 for an additional experiment that uses TALE to perform statistical cophylogenetic reconstruction).

In each simulation study experiment, the topological error of an estimated tree was compared to its corresponding model tree based on nRF distance. Each estimated cophylogeny was compared to the reference cophylogeny based on [50]'s precision calculation. Scatterplots and linear regression analysis were used to characterize the relationship between upstream phylogenetic estimation error and downstream cophylogenetic reconstruction error, where phylogenetic estimation error was assessed based on average topological error of host and symbiont trees, and cophylogenetic reconstruction error was assessed using cophylogenetic precision. The linear regression analyses were performed using R version 4.2.2 [36].

Variable importance analysis. In mixed and forward simulation experiments, the relative importance of species tree topology and other factors that can impact cophylogenetic reconciliation accuracy was assessed using the randomForest package [10] implemented in R [36]. The following variables were assessed for their impact on cophylogenetic reconciliation: tree topology (true species trees versus reconstructed trees in nRF distance), cophylogenetic software (eMPRess versus CoRe-PA), dataset size (default versus modified alignment lengths), event cost parameter (default versus alternative), and evolutionary divergence (tree height scaling factor $h = 0.1$ versus 10).

Random forests are used in machine learning to perform regression, classification, and other statistical analyses. To evaluate the relative importance of each variable, the out-of-bag (OOB) data for the tested variable was randomly shuffled and then this shuffled OOB data was used to construct 1000 regression trees. The original OOB data was used to construct another 1000 regression trees. On each regression tree, a mean squared error (MSE) is calculated based on the regression tree's prediction error rate. The variable importance is the difference in MSE between the random forest constructed on original OOB values and the random forest constructed on the shuffled OOB values, divided by standard error [10]. We scaled the importance of each factor to the most important variable to generate partial dependence plots.

2.3 Empirical study of soil-associated fungi and their bacterial endosymbionts

Sample acquisition and sequencing. A total of 13 metagenomic samples of *Mortierella spp.* and their associated endobacteria were collected and sequenced. Next-generation sequencing reads were assembled into contigs, which were then used to call single-nucleotide polymorphism variants (SNVs). The SNV MSAs for fungi and their bacterial endosymbionts had total length of 4,607,802 bp and 215,165 bp, respectively.

Model conditions	Source	Taxa	# taxa	Aln length	ANHD Avg	ANHD SE	Height Avg	Height SE	# cospec	# dup	# switch	# loss
mixed-gopher	[14]	Host	15	379	0.2241	0.0007	0.4024	0.0042	8	0	8	2
		Symbiont	17	379	0.5249	0.0007	3.0598	0.0359				
mixed-stinkbug	[17]	Host	7	1,745	0.2371	0.0016	0.2651	0.0016	5	5	1	0
		Symbiont	12	1,583	0.0661	0.0006	0.1349	0.0011				
mixed-primate	[46]	Host	55	696	0.2599	0.0002	0.6079	0.0046	24	0	14	22
		Symbiont	41	425	0.3376	0.0004	0.8169	0.0050				
mixed-damselfly	[25]	Host	24	1,051	0.1734	0.0004	0.4919	0.0036	4	3	15	4
		Symbiont	23	3,297	0.1327	0.0004	0.2643	0.0010				
mixed-moth	[52]	Host	82	1,404	0.1021	0.0001	0.2147	0.0013	13	0	27	12
		Symbiont	53	4,326	0.0250	0.0000	0.0486	0.0003				
mixed-bird	[35]	Host	37	5,000	0.1087	0.0001	0.1526	0.0009	15	12	29	17
	[11]	Symbiont	57	5,000	0.3562	0.0001	0.5459	0.0011				

Table 1: *Summary statistics for mixed simulation datasets.* Each mixed simulation condition ("Model conditions") is based on a previously published cophylogenetic study ("Source"). For each dataset type (either host or symbiont, as denoted by "Taxa"), the number of taxa ("# taxa"), true MSA length ("Aln length"), average and standard error of normalized Hamming distance of true MSAs ("ANHD Avg" and "ANHD SE", respectively), and average and standard error of model tree height ("Height Avg" and "Height SE", respectively) are reported. The number of cospeciation, duplication, host switch, and loss events in the reference cophylogeny are reported as "# cospec, "# dup", "# switch", and "# loss", respectively.

Model condition	H_{tips}	S_{tips}	λ_H	λ_C	λ_S	μ_H	μ_S	time
forward-gopher	35	55	0.3104	1.2000	0.0290	0	0	2.2
forward-stinkbug	35	55	0.2104	1.2000	0.0290	0	0	2.0
forward-primate	203	50	0.3374	0.6246	0.0452	0	0	4.8
forward-damselfly	25	25	0.1843	0.8846	0.2920	0	0	2.0
forward-bird	27	134	0.0544	0.6000	0.4520	0	0	4.0

Table 2: *Treeducken parameters used in forward simulations.* Treeducken was used to simulate cophylogenies and their constituent species phylogenies under a forward-time coalescent-based model [13]. Treeducken's model specifies the following parameters: the symbiont speciation rate λ_S, the symbiont extinction rate μ_S, the cospeciation rate λ_C, the host speciation rate λ_H, the host extinction rate μ_H, the expected number of host taxa H_{tips}, and the expected number of symbiont taxa S_{tips}.

Reconstruction and comparison of phylogenies and cophylogenies. Maximum likelihood tree estimation was performed using RAxML v8.2.12 [45] under finite-sites models of nucleotide sequence evolution. The latter consisted of the GTR+Γ [48] and nested models – specifically the HKY [15], K80 [20], and Jukes-Cantor [18] models; these substitution models span a range of model complexity from simplest (in the case of Jukes-Cantor) to more complex (i.e., GTR, HKY, and K80). PAUP* [47] was used to conduct additional phylogenetic reconstructions using neighbor-joining (NJ) [39] and the unweighted pair group method with arithmetic mean (UPGMA) algorithms [44]. Multispecies coalescent model-based species tree reconstruction was performed using SVDquartet [8]. If SVDquartet produced a tree with polytomies, the matrix rank was set to 1, 4, and 5 to produce three different tree topologies. Reconstructed phylogenetic trees were midpoint rooted. Finally, the estimated phylogenetic trees were reconciled to obtain a cophylogeny using either CoRe-PA [28] or eMPRess [41].

For each dataset, phylogenetic and cophylogenetic estimates obtained using any phylogenetic estimation method and eMPRess, respectively, were compared on a pairwise basis using the calculations described below; CoRe-PA-based results were evaluated similarly. For each pairwise comparison, phylogenetic tree estimation agreement was assessed using the average of the nRF distance between the two host trees and the nRF distance between the two symbiont trees. Then, for each pairwise comparison, cophylogenetic estimation agreement was assessed using the precision of [50]. Linear regression analyses were also performed to assess the relationship between phylogenetic tree estimation agreement and cophylogenetic estimation agreement, using the same procedures as in the simulation study experiments.

2.4 Empirical study of bobtail squids and their symbiotic bioluminescent bacteria

Sample acquisition and sequencing. Genomic sequence data for 22 samples of bobtail squids from the study of Sanchez et al. [40], and metadata for 37 *Vibrio* samples from the study of Bongrand et al. [5] were downloaded. The concatenated squid MSA had total length of 37,512 bp. Sanchez et al. [40] sequenced the former via genome skimming to identify more than 5000 ultraconserved loci. Host-symbiont association data came from the study of Bongrand et al. [5].

Reconstruction and comparison of phylogenies and cophylogenies. We reconstructed a phylogenetic tree for host taxa using the same approach as in the fungal/endobacterial dataset analysis. The bacterial symbiont phylogeny consisted of the *Vibrio* phylogeny reported by Bongrand et al. [5]. Cophylogenetic reconciliation and comparison of estimated phylogenies and cophylogenies followed the same procedures as in the other empirical dataset analysis.

3 RESULTS

3.1 Simulation study

The impact of upstream phylogenetic estimation error on downstream cophylogenetic reconciliation accuracy. Across the mixed simulation conditions, phylogenetic tree estimation returned average

Model conditions	Source	Taxa	# taxa	Aln len	ANHD Avg	ANHD SE	Height Avg	Height SE	# cosp	# dup	# switch	# loss
forward-gopher	[14]	Host	17	300	0.5664	0.0010	2.3260	0.0313	16	0	1	0
		Symbiont	16	300	0.5426	0.0009	2.5639	0.0403				
forward-stinkbug	[17]	Host	16	1,000	0.5672	0.0012	4.2617	0.0707	14	0	2	0
		Symbiont	14	1,000	0.5825	0.0016	3.9159	0.0326				
forward-primate	[46]	Host	48	400	0.6030	0.0002	8.0586	0.0791	31	3	17	0
		Symbiont	34	400	0.7017	0.0004	10.7577	0.2931				
forward-damselfly	[25]	Host	24	1,000	0.3437	0.0003	0.5804	0.0031	12	9	12	0
		Symbiont	21	1,000	0.4233	0.0007	1.1334	0.0066				
forward-bird	[35] [11]	Host	31	5,000	0.6953	0.0004	4.1329	0.0023	21	33	10	0
		Symbiont	54	5,000	0.7125	0.0002	5.0964	0.0027				

Table 3: *Summary statistics for forward simulation datasets.* For each model condition ("Model conditions"), Treeducken was used to perform forward simulations based on a previously published cophylogenetic study ("Source"). Each simulated dataset consisted of a model cophylogeny, its constituent model species trees and host/symbiont associations, and true MSAs. Table layout and description are otherwise identical to Table 1.

topological error of 7% and cophylogenetic reconstruction returned average precision of 66%. (Supplementary Figure S2 reports average topological errors of estimated species trees and cophylogenies for each model condition.)

Random forest-based variable importance analyses confirmed that tree topology inference error was the most important contributor to cophylogenetic reconciliation accuracy, while the second most important was evolutionary divergence at 70% of the variable importance of tree topology (Table 4). In our experiments, the choice of cophylogenetic reconciliation software and the choice of default versus statistically estimated event cost vectors contributed the least to cophylogenetic reconciliation accuracy.

The relationship between phylogenetic and cophylogenetic estimation error was examined using linear regression: Figure 2a shows the regression models fitted to observed topological errors across replicate datasets in each model condition. The regression analyses were statistically significant in all cases ($\alpha = 0.05; n = 100$), as shown in Supplementary Table S1. Increasing topological error during upstream estimation was clearly associated with reduced cophylogenetic accuracy, as evidenced by consistently negative regression coefficients and average regression coefficient of -1.96 across model conditions. We also observed varying scatter around fitted models: the coefficient of determination was highest in the mixed-gopher, mixed-stinkbug, and mixed-primate model conditions – ranging between 0.47 and 0.89 – and lower in others.

As in the mixed simulations, the partial dependence scores from random forest-based variable importance analysis showed that tree topology inference error was the most important contributor to cophylogenetic reconciliation accuracy in forward simulations, with evolutionary divergence having 82% of the relative importance of tree topology (Table 4). Similar to mixed simulations, the choice of cophylogenetic reconciliation software and the choice of default versus statistically estimated event cost vectors contributed the least to cophylogenetic reconciliation accuracy in forward simulation experiments. Topological error of estimated phylogenies and cophylogenies varied among forward simulation conditions. The observation is due in part to heterogeneity among the empirical estimates that served as the basis for the forward simulation conditions. On the other hand, topological errors were somewhat higher than in the other simulation experiments: the forward simulation experiments returned average tree topology error of 13% and average cophylogenetic precision of 35% (Supplementary Figure S4). As shown in Figure 2b, correlation between upstream tree estimation error and downstream cophylogeny reconstruction precision yielded similar findings as in the rest of simulation study. We observed significant and negative correlation in all forward simulation conditions (Supplementary Table S2). Furthermore, the coefficient of determination varied across forward simulation conditions in a similar pattern to the mixed simulation conditions, based on shared empirical dataset estimates. The largest values were seen on forward-gopher, forward-stinkbug, and forward-primate model conditions – ranging between 0.585 and 0.744; smaller values were seen on the other model conditions.

The impact of evolutionary divergence on the relationship between phylogenetic and cophylogenetic reconstruction accuracy. For each set of forward simulation conditions (Figure 3b), we found that phylogenetic and cophylogenetic estimation error was negatively and significantly correlated as the tree height parameter h varied between 0.1 and 10. Regression analysis returned regression coefficients between -0.899 and -0.220, and coefficients of determination between 0.668 and 0.222 (Supplementary Table S4). Both upstream and downstream topological error was lowest for the smallest h settings (i.e., 0.1, 0.5, and 1.0). As the height h increased, both topological errors increased in tandem, and both were largest on simulations with height $h = 10$. The latter was likely at saturation, as topological errors tended to be maximal. Similar outcomes were observed in the corresponding mixed simulation experiments with varying tree height h, as shown in Figure 3a. The effect of increasing h on topological error was more complicated and non-linear in some cases. This was in part due to heterogeneity of empirical estimates used for parametric resampling, unlike the fully *in silico* simulations used elsewhere in the simulation study.

3.2 Empirical study

Soil-associated fungi and their bacterial endosymbionts. Topological disagreements among estimated phylogenies were higher than in the simulation study (Supplementary Figure S5); a similar outcome was observed among estimated cophylogenies. This is by

Simulations	Tree topology	Evolutionary divergence	Dataset size	Cophylogenetic software	Event cost parameter
Mixed	1.0000	0.7029	0.5511	0.3513	0.0611
Forward-time	1.0000	0.8160	N/A	0.7786	0.3144

Table 4: *Simulation study: variable importance assessment for mixed and forward simulations.* A random forest model was used to determine the mean importance of each variable. Results are reported as an average across all mixed simulation conditions and scaled relative to the most importance variable ($n = 100$), and similarly for the forward simulation conditions.

design: the empirical study utilized a wide array of phylogenetic reconstruction methods with varying estimation accuracy. The design choice provides an indirect means to vary the topological accuracy of input phylogenies and then observe its effects on downstream cophylogenetic estimation, in contrast to the direct control and model/reference comparisons enabled by *in silico* simulations. We analyzed the relationship between phylogenetic and cophylogenetic estimation error using linear regression (Figure 4a). Consistent with the simulation study, we observed that greater topological agreement in the former set of inputs was significantly associated with greater topological agreement of the latter output, as assessed using an F-test with Benjamini-Hochberg [3] correction for multiple tests ($\alpha = 0.05$; $n = 114$). The full assembly dataset analysis returned a regression coefficient of -2.067 and coefficient of determination of 0.672, which is also in line with the simulation study.

Bobtail squids and their symbiotic bioluminescent bacteria Topological disagreements among species cophylogenies and resulting cophylogenetic reconciliations were somewhat smaller than those observed on the fungal/endosymbiont dataset (Supplementary Figure S6). Another key difference concerns host/symbiont associations: relatively few squid hosts were associated with most bacterial symbionts. Still, we observed a similar relationship between upstream phylogenetic estimation agreement and downstream cophylogeny precision (Figure 4b). Linear regression analyses returned significant and negative correlation ($\alpha = 0.05$; $n = 216$), along with a regression coefficient of -0.449, intercept of 0.841, F-test p-value $< 10^{-12}$, coefficient of determination of 0.213, and residual standard error of 0.109.

4 DISCUSSION

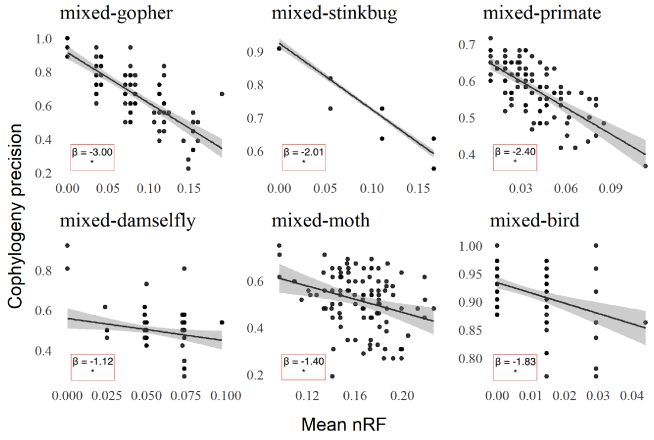
Across all forward simulation experiments, correlation between upstream phylogenetic estimation error and downstream cophylogenetic estimation accuracy was significant and consistently negative. As the former increased, the latter would degrade. The mixed simulation experiments and empirical dataset analyses involving eMPRess-estimated cophylogenies (as well as a supplementary simulation experiment involving TALE, as described in the Supplementary Online Materials) also returned a consistent outcome: namely, a significant and negatively correlated relationship between upstream phylogenetic reconstruction error and downstream cophylogenetic estimation reproducibility. The expanded simulation experiments that focused on varying evolutionary divergence (while fixing other experimental factors) refined our study's primary finding and demonstrated that evolutionary divergence plays a key role in modulating upstream and downstream estimation error in tandem. Of course, other factors also play a role (e.g., taxon sampling, choice of phylogenetic and cophylogenetic reconstruction method(s), coevolutionary event distribution, evolutionary and coevolutionary model mis-specification, etc.), and the relationship between phylogenetic and cophylogenetic reconstruction is complex [12]. Heterogeneity among simulation conditions due to these factors helps to explain some of the more minor differences among experimental outcomes. Nevertheless, our primary finding – that phylogenetic estimation error strongly impacts downstream cophylogenetic reconciliation accuracy – was robust to these factors. Furthermore, variable importance analyses revealed that phylogenetic tree estimation error was the most important experiment factor associated with cophylogenetic reconciliation accuracy, compared to the other factors.

We note a key difference between the simulation study and the empirical study. A primary advantage of the former is the ability to benchmark against model/reference phylogenies and cophylogenies. But the latter is inherently more complex and nuanced than the former. For example, the two systems in our empirical study are models sampled along a continuum of symbiotic coevolution modes [34]: from open – as in the case of bobtail squids and their bioluminescent symbionts [34] – to mixed to closed – as in the case of early diverging fungi and their endosymbionts [33]. Where a system exists along this continuum is thought to strongly influence the probabilities of different coevolutionary events: for example, host shifts occur more frequently in an open system, and cospeciation predominates in a closed system. Depending on the taxa under study, it is plausible that symbiotic coevolution may switch between different modes along a phylogeny (e.g., from closed to mixed). But we are not aware of any suitable non-homogeneous cophylogenetic models and we also lack a basic understanding of their theoretical properties (e.g., statistical identifiability). The gap between natural symbiotic coevolution and current statistical cophylogenetic modeling represents an opportunity for advanced model and methods development; for now, this study is constrained by the limitations of the state of the art.

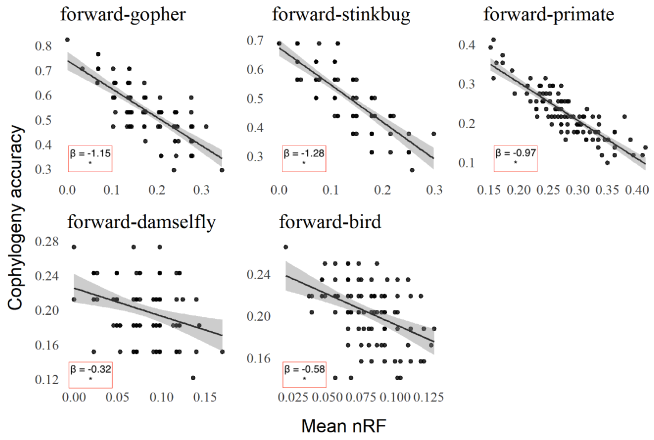
5 CONCLUSIONS

This study demonstrated the major effect that phylogenetic estimation error has on downstream cophylogenetic reconstruction accuracy. The finding was consistently observed throughout the simulation study experiments. Empirical analyses of two genomic sequence datasets for models of symbiosis also revealed that variable phylogenetic tree estimation quality decreased reproducibility of cophylogenetic estimation.

We propose the following strategies to put the key findings of our study into practice. One ideal solution would be to develop and utilize a new generation of cophylogenetic reconstruction methods

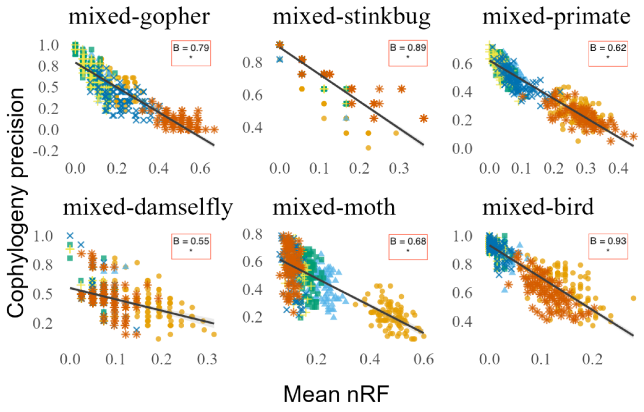


(a) Mixed simulation conditions.

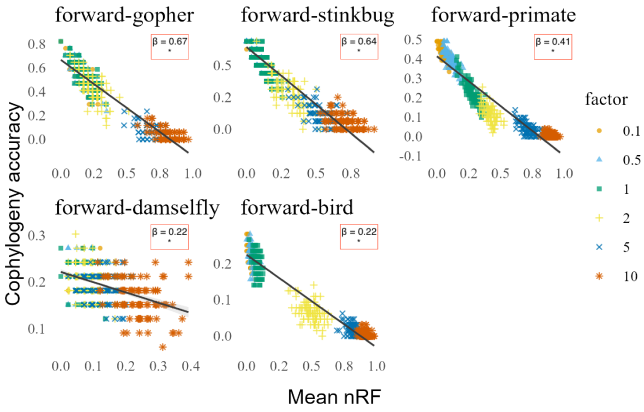


(b) Forward simulation conditions.

Figure 2: *Simulation study: the relationship between phylogenetic and cophylogenetic estimation error.* For each model condition, the topological error returned by phylogenetic tree estimation (averaged across the pair of host and symbiont datasets) and the precision returned by cophylogenetic reconstruction are shown for each replicate dataset ($n = 100$). A fitted linear regression model is shown for each model condition as well, and the 95% confidence interval is shown in grey around the regression line. A red box inside each plot shows the regression coefficient β and an asterisk (*) denoting statistical significance ($\alpha = 0.05$; $n = 100$) using an F-test with Benjamini-Hochberg multiple test correction [3]. The linear regression analyses were statistically significant in all cases. (Supplementary Tables S1 and S2 provide additional regression analysis results.)



(a) Mixed simulation experiments.



(b) Forward simulation experiments.

Figure 3: *Simulation study: the impact of evolutionary divergence on phylogenetic and cophylogenetic estimation error.* Estimation error was assessed based upon average topological error of estimated trees (averaged across the pair of host and symbiont datasets) and cophylogenetic precision. Model tree branch lengths were scaled by height parameter h ("factor"); data points for a given setting of h are distinguished by a distinct color. A fitted linear regression model is shown for each simulation condition. A red box inside each plot shows the regression coefficient β and an asterisk (*) denoting statistical significance ($\alpha = 0.05$; $n = 600$) using an F-test with Benjamini-Hochberg multiple test correction [3]. The linear regression analyses were statistically significant in all cases. (Supplementary Tables S3 and S4 provide additional regression analysis results.)

that account for upstream phylogenetic estimation error and perform statistical inference and learning directly from biomolecular sequence data inputs. To our knowledge, the choices are very limited for now. We are aware of one option that represents a partial first step towards this goal: a new method called TALE [27] which accepts distributions of symbiont species trees and gene trees as input, but only a point estimate for the host species tree (as of this writing). However, given the outcome of a supplementary experiment involving TALE as well as other considerations regarding TALE's design (see Supplementary Online Materials section S7), our study underscores the need for continued research, modeling, and computational methods development in this direction. By far the most widely used options for cophylogenetic reconstruction remain the current generation of methods which require fixed species trees

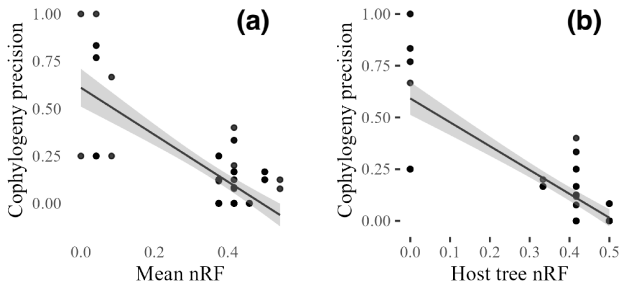


Figure 4: *Empirical study: topological discordance among phylogenetic and cophylogenetic estimates.* The scatterplots show topological discordance between each pair of different phylogenetic tree estimation methods – averaged across the host dataset and symbiont dataset – versus disagreement among the resulting cophylogenetic reconciliations produced using either eMPRess or CoRe-PA. (a) In the fungal dataset, a range of different phylogenetic tree estimation methods were used to estimate phylogenetic trees on the host dataset and the symbiont dataset. Along with the known host/symbiont associations, each estimated host tree and symbiont tree pair returned by a given phylogenetic tree estimation method was reconciled into a cophylogeny using either eMPRess or CoRe-PA. Topological discordance between different host trees estimated by different methods was assessed on a pairwise basis using nRF distance. Disagreement among cophylogenies estimated using eMPRess reconciliation of different phylogenetic tree estimates was assessed on a pairwise basis using cophylogenetic precision; disagreement among CoRe-PA estimates was evaluated similarly. A fitted linear regression model is shown ($n = 114$). (b) The squid dataset analyses used a symbiont tree that was fixed to the estimate of [40] and $n = 216$; analyses and results are otherwise reported similarly to the fungal dataset.

as input. In lieu of an ideal solution, we provide the following practical guidance as temporary workarounds. First, we propose that researchers adopt more intensive species tree reconstruction as best practices in a cophylogenetic study. For example, we recommend that researchers select more intensive local optimization heuristic settings for addressing the computationally difficult tree reconstruction problems in this study and in the state of the art. Second, more intensive sequencing effort to obtain additional high-quality biomolecular sequence data can also help, assuming that suitable methods can be used to account for the complex interplay of evolutionary processes – substitutions, sequence insertion and deletion, genetic drift and incomplete lineage sorting, and more – that arises in this setting. A new generation of phylogenomic inference and learning methods are now used to better address species phylogeny reconstruction using large-scale multi-locus and/or genomic sequence data, and they may also pay dividends when reconstructing cophylogenies using genomic sequence analysis.

We conclude with thoughts on future research directions. First, we have already mentioned the need for richer coevolutionary models. Our study's empirical models of open symbiosis (Hawaiian bobtail squid and its bioluminescent bacterial symbiont) and closed symbiosis (soil-associated fungi and its bacterial endosymbionts) bookend a rich spectrum of symbiotic lifestyles and coevolution modes. Richer statistical models are urgently needed to better account for the dynamic interplay of different coevolutionary processes that can shift over time. Second, new methods that jointly reconstruct species trees, gene trees, and a cophylogeny from multi-locus sequence data are needed. While TALE represents an important partial step in this direction, more methodological research and development is needed. But important prerequisites must be addressed first: realistic models of coevolution that also permit tractable statistical calculations, as well as statistically efficient inference and learning algorithms under the new models. Scalability-enhancing algorithmic techniques such as phylogenetic divide-and-conquer [16, 23, 29] may prove fruitful here.

6 DATA AVAILABILITY

Updated versions of the study data and software scripts underlying this article are available in the public GitLab repository at https://gitlab.msu.edu/liulab/cophylogeny-species-tree-quality-performance-study-data-scripts. An archival snapshot of the study data and software scripts has been uploaded to Figshare and can be accessed at https://doi.org/10.6084/m9.figshare.21713996.v1.

ACKNOWLEDGMENTS

The authors would like to thank three anonymous reviewers for their detailed and constructive feedback. This research has been supported in part by the National Science Foundation (2144121, 2214038, 1737898 to KJL), the National Science Foundation Research Traineeship Program (DGE-1828149) to JZ, and MSU (EEB summer fellowship to JZ). All computational experiments and analyses were performed on the MSU High Performance Computing Center, which is part of the MSU Institute for Cyber-Enabled Research.

REFERENCES

[1] Juan Antonio Balbuena, Raúl Míguez-Lozano, and Isabel Blasco-Costa. 2013. PACo: a novel procrustes application to cophylogenetic analysis. *PloS One* 8, 4 (2013), e61048.
[2] Christian Baudet, Béatrice Donati, Blerina Sinaimeri, Pierluigi Crescenzi, Christian Gautier, Catherine Matias, and M-F Sagot. 2015. Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology* 64, 3 (2015), 416–431.
[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.
[4] Isabel Blasco-Costa, Alexander Hayward, Robert Poulin, and Juan A Balbuena. 2021. Next-generation cophylogeny: unravelling eco-evolutionary processes. *Trends in Ecology & Evolution* 36, 10 (2021), 907–918.
[5] Clotilde Bongrand, Silvia Moriano-Gutierrez, Philip Arevalo, Margaret McFall-Ngai, Karen L Visick, Martin Polz, and Edward G Ruby. 2020. Using colonization assays and comparative genomics to discover symbiosis behaviors and factors in *Vibrio fischeri*. *mBio* 11, 2 (2020), e03407–19.
[6] MA Charleston. 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.* 149, 2 (1998), 191–223.
[7] MA Charleston and RDM Page. 2002. TreeMap 2. A Macintosh program for cophylogeny mapping. https://sites.google.com/site/cophylogeny/
[8] Julia Chifman and Laura Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30, 23 (2014), 3317–3324.
[9] Chris Conow, Daniel Fielder, Yaniv Ovadia, and Ran Libeskind-Hadas. 2010. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology* 5, 1 (2010), 1–10.

[10] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. 2007. Random forests for classification in ecology. *Ecology* 88, 11 (2007), 2783–2792.

[11] Robert S de Moya, Julie M Allen, Andrew D Sweet, Kimberly KO Walden, Ricardo L Palma, Vincent S Smith, Stephen L Cameron, Michel P Valim, Terry D Galloway, Jason D Weckstein, et al. 2019. Extensive host-switching of avian feather lice following the Cretaceous-Paleogene mass extinction event. *Communications Biology* 2, 1 (2019), 445.

[12] Wade Dismukes, Mariana P Braga, David H Hembry, Tracy A Heath, and Michael J Landis. 2022. Cophylogenetic methods to untangle the evolutionary history of ecological interactions. *Annual Review of Ecology, Evolution, and Systematics* 53 (2022), 275–298.

[13] Wade Dismukes and Tracy A Heath. 2021. treeducken: An R package for simulating cophylogenetic systems. *Methods Ecol. Evol.* 12, 8 (2021), 1358–1364.

[14] Mark S Hafner, Philip D Sudman, Francis X Villablanca, Theresa A Spradling, James W Demastes, and Steven A Nadler. 1994. Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science* 265, 5175 (1994), 1087–1090.

[15] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22, 2 (1985), 160–174. Publisher: Springer.

[16] Hussein A Hejase, Natalie VandePol, Gregory M Bonito, and Kevin J Liu. 2018. FastNet: fast and accurate statistical inference of phylogenetic networks using large-scale genomic sequence data. In *Comparative Genomics: 16th International Conference, RECOMB-CG 2018, Magog-Orford, QC, Canada, October 9-12, 2018, Proceedings 16*. Springer, 242–259.

[17] Takahiro Hosokawa, Yoshitomo Kikuchi, Naruo Nikoh, Masakazu Shimada, and Takema Fukatsu. 2006. Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria. *PLoS Biology* 4, 10 (2006), e337.

[18] T.H. Jukes and C.R. Cantor. 1969. *Evolution of Protein Molecules*. Academic Press, New York, NY, USA, 21–132.

[19] Kazutaka Katoh and Daron M Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30, 4 (2013), 772–780.

[20] Motoo Kimura. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16, 2 (1980), 111–120. Publisher: Springer.

[21] Pierre Legendre, Yves Desdevises, and Eric Bazin. 2002. A statistical test for host–parasite coevolution. *Systematic Biology* 51, 2 (2002), 217–234.

[22] Ran Libeskind-Hadas, Yi-Chieh Wu, Mukul S Bansal, and Manolis Kellis. 2014. Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics* 30, 12 (2014), i87–i95.

[23] Kevin Liu, Tandy J Warnow, Mark T Holder, Serita M Nelesen, Jiaye Yu, Alexandros P Stamatakis, and C Randal Linder. 2012. SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology* 61, 1 (2012), 90.

[24] Mercè Llabrés, Francesc Rosselló, and Gabriel Valiente. 2021. The generalized Robinson-Foulds distance for phylogenetic trees. *Journal of Computational Biology* 28, 12 (2021), 1181–1195.

[25] MO Lorenzo-Carballa, Y Torres-Cambas, K Heaton, GDD Hurst, S Charlat, TN Sherratt, H Van Gossum, A Cordero-Rivera, and CD Beatty. 2019. Widespread Wolbachia infection in an insular radiation of damselflies (Odonata, Coenagrionidae). *Scientific Reports* 9, 1 (2019), 1–13.

[26] Andrés Martínez-Aquino. 2016. Phylogenetic framework for coevolutionary studies: a compass for exploring jungles of tangled trees. *Current Zoology* 62, 4 (2016), 393–403.

[27] Hugo Menet, Alexia Nguyen Trung, Vincent Daubin, and Eric Tannier. 2023. Host-symbiont-gene phylogenetic reconciliation. *Peer Community Journal* 3, Article e47 (2023).

[28] Daniel Merkle, Martin Middendorf, and Nicolas Wieseke. 2010. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC bioinformatics* 11, 1 (2010), 1–10.

[29] Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. 2015. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology* 22, 5 (2015), 377–386.

[30] Bernard ME Moret, Usman Roshan, and Tandy Warnow. 2002. Sequence-length requirements for phylogenetic methods. *Lecture Notes in Computer Science* 2452, 343–356.

[31] S Nelesen, Kevin Liu, Donggao Zhao, C Randal Linder, and Tandy Warnow. 2008. The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. In *Biocomputing 2008*. World Scientific, 25–36.

[32] Yaniv Ovadia, Daniel Fielder, Chris Conow, and Ran Libeskind-Hadas. 2011. The cophylogeny reconstruction problem is NP-complete. *Journal of Computational Biology* 18, 1 (2011), 59–65.

[33] Teresa E. Pawlowska, Maria L. Gaspar, Olga A. Lastovetsky, Stephen J. Mondo, Imperio Real-Ramirez, Evaniya Shakya, and Paola Bonfante. 2018. Biology of fungi and their bacterial endosymbionts. *Annual Review of Phytopathology* 56, 1 (2018), 289–309.

[34] Julie Perreau and Nancy A Moran. 2022. Genetic innovations in animal–microbe symbioses. *Nature Reviews Genetics* 23, 1 (2022), 23–39.

[35] Richard O Prum, Jacob S Berv, Alex Dornburg, Daniel J Field, Jeffrey P Townsend, Emily Moriarty Lemmon, and Alan R Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526, 7574 (2015), 569–573.

[36] R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[37] Andrew Rambaut and Nicholas C Grass. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* 13, 3 (1997), 235–238.

[38] David F Robinson and Leslie R Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53, 1-2 (1981), 131–147.

[39] Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 4 (1987), 406–425.

[40] Gustavo Sanchez, Fernando Á Fernández-Álvarez, Morag Taite, Chikatoshi Sugimoto, Jeffrey Jolly, Oleg Simakov, Ferdinand Marlétaz, Louise Allcock, and Daniel S Rokhsar. 2021. Phylogenomics illuminates the evolution of bobtail and bottletail squid (order Sepiolida). *Communications Biology* 4, 1 (2021), 819.

[41] Santi Santichaivekin, Qing Yang, Jingyi Liu, Ross Mawhorter, Justin Jiang, Trenton Wesley, Yi-Chieh Wu, and Ran Libeskind-Hadas. 2021. eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics* 37, 16 (2021), 2481–2482.

[42] Christopher L Schardl, Kelly D Craven, S Speakman, Arnold Stromberg, A Lindstrom, and Ruriko Yoshida. 2008. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses. *Systematic Biology* 57, 3 (2008), 483–498.

[43] Barton E Slatko, Andrew F Gardner, and Frederick M Ausubel. 2018. Overview of next-generation sequencing technologies. *Current protocols in molecular biology* 122, 1 (2018), e59.

[44] Robert R Sokal. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* 38 (1958), 1409–1438.

[45] Alexandros Stamatakis. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 9 (2014), 1312–1313. Publisher: Oxford University Press.

[46] William M Switzer, Marco Salemi, Vedapuri Shanmugam, Feng Gao, Mian-er Cong, Carla Kuiken, Vinod Bhullar, Brigitte E Beer, Dominique Vallet, Annie Gautier-Hion, et al. 2005. Ancient co-speciation of simian foamy viruses and primates. *Nature* 434, 7031 (2005), 376–380.

[47] David L. Swofford. 2003. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates.

[48] Simon Tavaré. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci* 17, 2 (1986), 57–86.

[49] John N Thompson. 2010. Four central points about coevolution. *Evolution: Education and Outreach* 3, 1 (2010), 7–13.

[50] Nicolas Wieseke, Tom Hartmann, Matthias Bernt, and Martin Middendorf. 2015. Cophylogenetic reconciliation with ILP. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12, 6 (2015), 1227–1235.

[51] Ziheng Yang. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* 11, 9 (1996), 367–372.

[52] Yongjie Zhang, Shu Zhang, Yuling Li, Shaoli Ma, Chengshu Wang, Meichun Xiang, Xin Liu, Zhiqiang An, Jianping Xu, and Xingzhong Liu. 2014. Phylogeography and evolution of a fungal–insect association on the Tibetan Plateau. *Molecular Ecology* 23, 21 (2014), 5337–5355.

Received 11 June 2023; revised 03 August 2023; accepted

Supplementary Online Materials: The Impact of Species Tree Estimation Error on Cophylogenetic Reconstruction

JULIA ZHENG, Michigan State University, USA
YUYA NISHIDA, Michigan State University, USA
ALICJA OKRASIŃSKA, University of Warsaw, Poland
GREGORY M. BONITO, Michigan State University, USA
ELIZABETH A.C. HEATH-HECKMAN, Michigan State University, USA
KEVIN J. LIU*, Michigan State University, USA

ACM Reference Format:
Julia Zheng, Yuya Nishida, Alicja Okrasińska, Gregory M. Bonito, Elizabeth A.C. Heath-Heckman, and Kevin J. Liu. 2023. Supplementary Online Materials: The Impact of Species Tree Estimation Error on Cophylogenetic Reconstruction. In *14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '23), September 3–6, 2023, Houston, TX, USA.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3584371.3612964

BCB '23, September 3–6, 2023, Houston, TX, USA

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3612964

S1 SUPPLEMENTARY METHODS

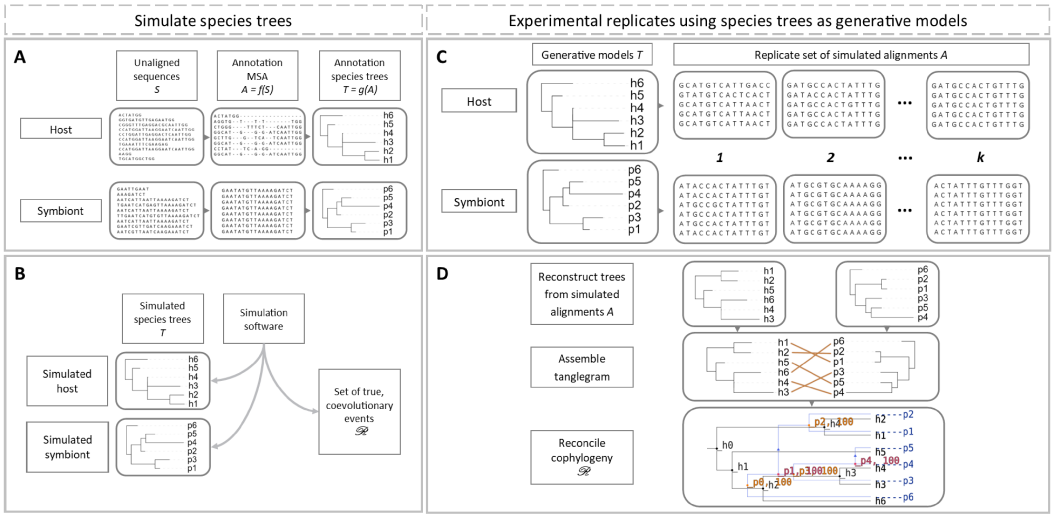


Fig. S1. *Illustrated overview of simulation study experiments.* Two simulation procedures were used to simulate datasets. The procedures differ in the cophylogeny source and simulation software that they utilized. (A) The "mixed" simulation experiments utilized a cophylogeny and constituent species trees that were based on empirical dataset analyses. (B) The fully in-silico "forward"-in-time simulations sampled cophylogenies and constituent species trees under Treeducken's forward-in-time coalescent-based cophylogeny model [12]. Since both serve as generative parametric models in our simulations, we refer to them as "model cophylogenies" and "model trees", respectively. (C) For each model condition, sequence evolution along each constituent species tree was simulated under finite-sites models, resulting in a multiple sequence alignment (MSA). The simulation procedure was repeated to obtain k experimental replicates. (D) Phylogenetic and cophylogenetic reconstruction was performed using a computational pipeline. For each replicate dataset, a phylogenetic tree was reconstructed for host taxa using their corresponding MSA as input, and similarly for symbionts. The estimated host tree and estimated symbiont tree were combined with host/symbiont association data to produce a tanglegram. The two species trees and host/symbiont associations were then used as input to reconstruct a cophylogeny.

S2 SUPPLEMENTARY RESULTS

Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
mixed-gopher	0.9146	-2.9996	0.6406	0.1061	$< 10^{-16}$	$< 10^{-23}$
mixed-stinkbug	0.9254	-2.0067	0.8903	0.0331	$< 10^{-16}$	$< 10^{-48}$
mixed-primate	0.6704	-2.3987	0.4732	0.0511	$< 10^{-15}$	$< 10^{-14}$
mixed-damselfly	0.5590	-1.1198	0.0564	0.0928	0.0173	0.0173
mixed-moth	0.7460	-1.4036	0.1010	0.1146	0.0012	0.0025
mixed-bird	0.9341	-1.8328	0.1663	0.0408	$< 10^{-5}$	$< 10^{-5}$

Table S1. **Linear regression results for mixed simulation experiments.** The fitted model's intercept ("intercept"), regression coefficient ("B coefficient"), coefficient of determination ("R^2"), and residual standard error ("RSE") are shown. Statistical significance was assessed using the F-test, and uncorrected p-values ("p-value") and corrected q-values ("q-value") based on Benjamini-Hochberg multiple test correction [5] are reported ($n = 100$).

Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
forward-gopher	0.7385	-1.1485	0.5854	0.0680	$< 10^{-16}$	$< 10^{-20}$
forward-stinkbug	0.6729	-1.2848	0.6171	0.0632	$< 10^{-16}$	$< 10^{-21}$
forward-primate	0.4968	-0.9702	0.7442	0.0312	$< 10^{-16}$	$< 10^{-30}$
forward-damselfly	0.2252	-0.3232	0.1035	0.0326	0.0011	0.0011
forward-bird	0.2495	-0.5780	0.1129	0.0141	$< 10^{-6}$	$< 10^{-6}$

Table S2. **Linear regression results for forward-time simulation experiments.** Table layout and description are otherwise identical to Table S1.

Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
mixed-gopher	0.7901	-1.4661	0.7906	0.1216	$< 10^{-8}$	$< 10^{-7}$
mixed-stinkbug	0.8930	-1.6693	0.7860	0.0543	$< 10^{-16}$	$< 10^{-46}$
mixed-primate	0.6218	-1.3590	0.8797	0.0570	$< 10^{-16}$	$< 10^{-19}$
mixed-damselfly	0.5514	-0.9679	0.1880	0.1067	$< 10^{-5}$	$< 10^{-5}$
mixed-moth	0.6783	-0.9971	0.6026	0.1090	$< 10^{-5}$	$< 10^{-5}$
mixed-bird	0.9329	-2.2698	0.7975	0.0706	$< 10^{-16}$	$< 10^{-16}$

Table S3. **Linear regression results for mixed simulation experiments with varying evolutionary divergence.** Table layout and description are otherwise identical to Table S1.

Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
forward-gopher	0.6677	-0.8078	0.9091	0.0738	$< 10^{-16}$	$< 10^{-313}$
forward-stinkbug	0.6429	-0.8991	0.9091	0.0777	$< 10^{-16}$	$< 10^{-313}$
forward-primate	0.4133	-0.5121	0.8796	0.0584	$< 10^{-16}$	$< 10^{-276}$
forward-damselfly	0.2217	-0.2200	0.1693	0.0344	$< 10^{-16}$	$< 10^{-26}$
forward-bird	0.2241	-0.2553	0.9317	0.0257	$< 10^{-16}$	0

Table S4. **Linear regression results for forward-time simulation experiments with varying evolutionary divergence.** Table layout and description are otherwise identical to Table S1.

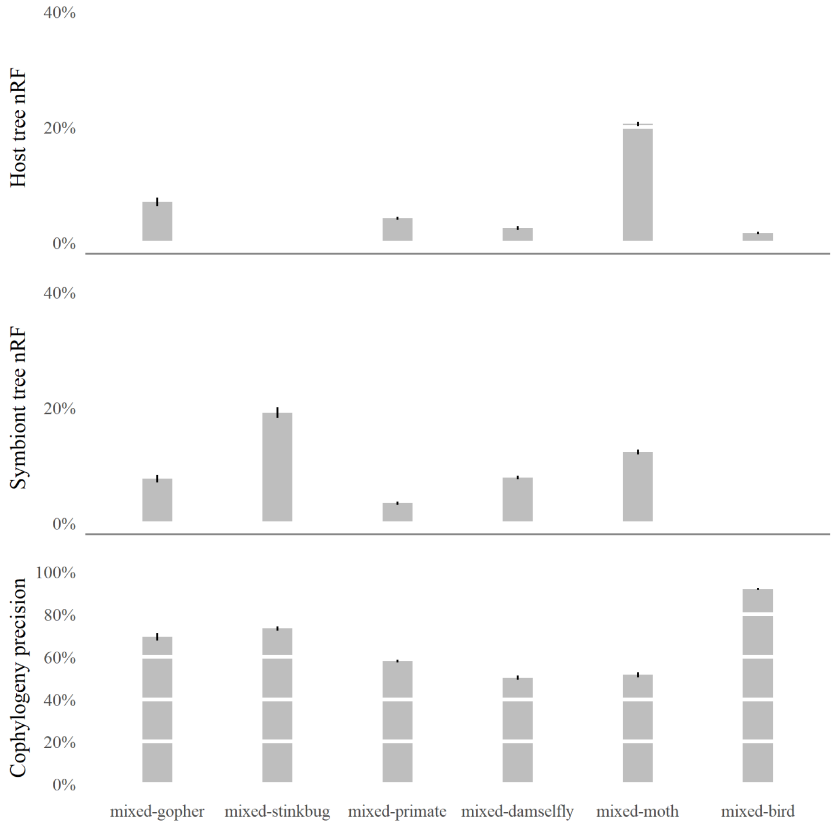


Fig. S2. **For each mixed simulation condition, host tree topology error, average symbiont tree topology error, and cophylogenetic precision are shown.** Averages are reported across all experimental replicate for each model condition ($n = 100$). Standard error bars are shown.

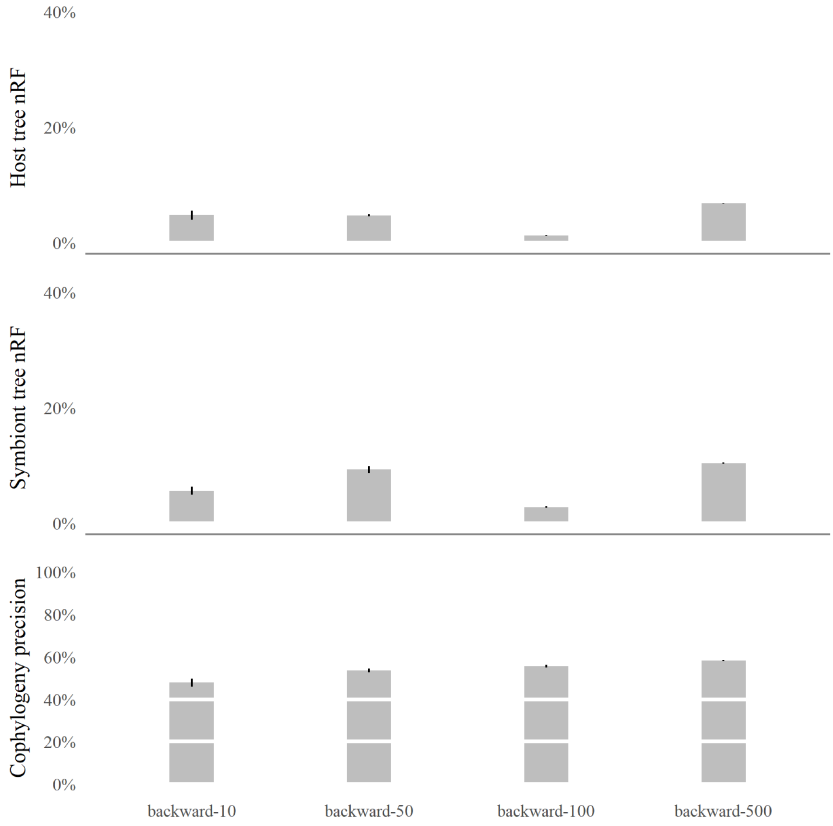


Fig. S3. **Backward simulation bar graphs for average host tree topology error, average symbiont tree topology error, and average cophylogenetic precision.** Averages are reported across all experimental replicate for each model condition ($n = 100$). Error bars visualize standard error.

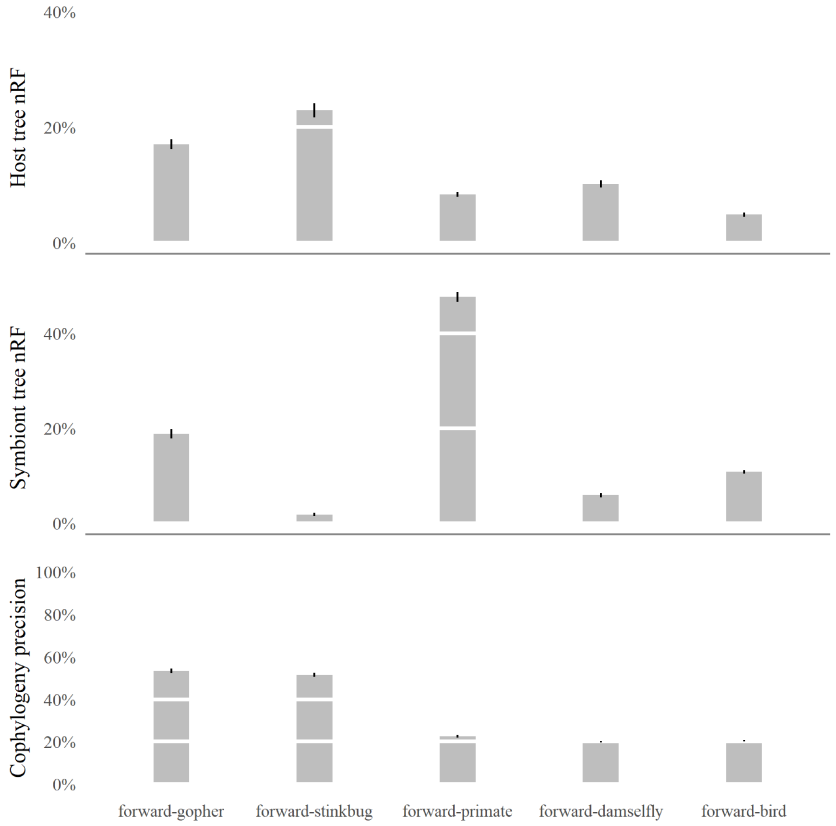


Fig. S4. **Forward simulation bar graphs for average host tree topology error, average symbiont tree topology error, and average cophylogenetic precision.** Error bars visualize standard error. Averages are reported across all experimental replicate for each model condition ($n = 100$).

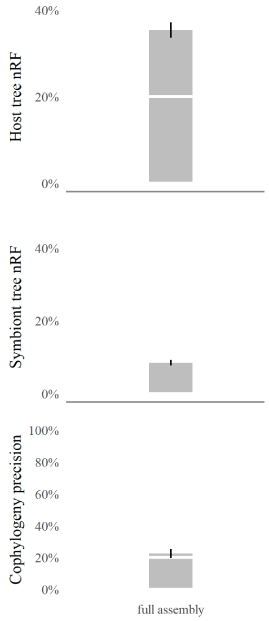


Fig. S5. **Bar graphs for *Mortierella spp.* and endobacteria dataset.** Top to bottom: Average host tree error, average symbiont tree error, and average cophylogenetic precision ($n = 100$). Standard error bars are also shown.

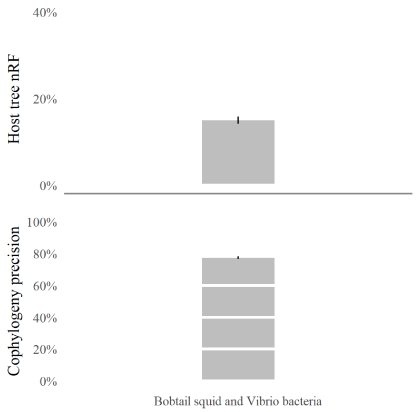


Fig. S6. **Bar plots for bobtail squid and *Vibrio* dataset for average host tree error and average cophylogenetic precision.** Averages are reported across all experimental replicates ($n = 100$). Standard error bars are also shown.

S3 METAGENOMIC PROCESSING AND CONTIG ASSEMBLY FOR *MORTIERELLA SPP.* AND ENDOSYMBIONT

Sample acquisition and sequencing. Isolates were collected and also sourced from established culture collections. Modified versions of the soil plate [32] and selective-baiting method [28] were used to isolate Mortierellomycotina from soil. The techniques described in [6] were used to isolate Mortierellomycotina from pine and spruce roots.

In total, thirteen metagenomic samples of *Mortierella spp.* and their associated endobacteria were collected and sequenced (Table S5). Ten samples were sequenced using Illumina HiSeq 2500 short-read sequencing and three samples were sequenced using PacBio long-read sequencing.

Illumina-sequenced metagenomic reads were trimmed with BBDuk (ftl=5 minlen=90) [7] to remove Illumina adapters, trim five leftmost bases, and discard reads shorter than 90 bp after trimming. The quality of trimmed reads was assessed by FastQC [1]. De novo assembly of the metagenomic samples was conducted with SPAdes (-k 21,33,55,77,99,127) [3] to produce contigs. BBMap [7] was used to calculate summary statistics on assembled contigs. BUSCO [29] was used with the mucoromycota_odb10 and burkholderiales_odb10 databases to assess the completeness of de novo assembly and confirm the presence of endobacteria, respectively (Table S6).

The PacBio-sequenced metagenomic reads were de novo assembled with CANU [16], with the exception of sample AV005: its draft assembly was obtained directly from JGI (Project ID: 1203140). Completeness and summary statistics were assessed in the same manner as for Illumina-sequenced assemblies (Table S6).

Sample ID	BioProject	BioSample	SRA accession	GOLD JGI ID	Instrument	Geographic location	Specimen Scope	Fungal organism
AD022	PRJNA367465	SAMN06267312	SRR5822949	Gp0136994	Illumina HiSeq 2500	Bryce Canyon, UT, USA	Rhizosphere	Mortierella elongata
AD045	PRJNA340843	SAMN05720529	SRR5190920	Gp0154302	Illumina HiSeq 2500	East Lansing, MI, USA	Rhizosphere	Mortierella gamsii
AD051	PRJNA370772	SAMN06297100	SRS2351483	Gp0136990	PacBio RS II	Laingsburg, MI, USA	Rhizosphere	Mortierella minutissima
AD058	PRJNA340839	SAMN05720441	SRR5190916	Gp0154298	Illumina HiSeq 2500	Laingsburg, MI, USA	Rhizosphere	Podila epicladia
AD073	PRJNA364919	SAMN06265150	SRR5822802	Gp0136992	Illumina HiSeq 2500	Michigan, USA	Rhizosphere	Mortierella elongata
AD086	PRJNA365031	SAMN06264397	SRR5822800	Gp0136991	Illumina HiSeq 2500	Coatesville, PA, USA	Soil	Mortierella humilis
AD266	PRJNA713069	SAMN18261529	NA	Gp0397541	PacBio Sequel	Oregon, USA	Soil	Mortierella alpina
AM1000	PRJNA340828	SAMN05720794	SRS1930920	Gp0154287	Illumina HiSeq 2500	Illinois, USA	Monoisolate	Mortierella clonocystis
AM980	PRJNA340833	SAMN05720525	SRR5190941	Gp0154292	Illumina HiSeq 2500	NA	Monoisolate	Mortierella elongata
AV005	PRJNA713068	SAMN18259510	NA	Gp0397540	PacBio Sequel	Camuy, Puerto Rico	Soil	Mortierella capitata
CK281	PRJNA364924	SAMN06266091	SRR5823416	Gp0136997	Illumina HiSeq 2500	North Carolina, USA	Soil	Mortierella minutissima
NVP60	PRJNA340844	SAMN05720530	SRR5192043	Gp0154303	Illumina HiSeq 2500	Cassopolis, MI, USA	Monoisolate	Linnemannia gamsii
TTC192	PRJNA410574	SAMN07687234	SRR6257765	Gp0154326	Illumina HiSeq 2500	North Carolina, USA	Soil	Mortierella verticillata

Table S5. List of *Mortierella spp.* and endobacteria used in this study.

	Metagenomic assembly summary statistics					BUSCO Marker Percentage (*Mortierella spp.*)				BUSCO Marker Percentage (endobacteria)			
Sample ID	# Contig	Mbp	L50	N50	GC %	Full	Single	Duplicate	Fragment	Full	Single	Duplicate	Fragment
AD022	14019	50.92	9866	1486	48.64	93.3	92.0	1.3	2.4	89.2	88.5	0.7	1.2
AD045	4647	49.84	23855	618	47.70	94.5	93.4	1.1	1.4	90.0	89.4	0.6	1.2
AD051	577	49.90	487613	29	48.90	97.4	92.3	5.1	0.2	88.9	82.7	6.2	1.2
AD058	7618	41.20	9691	1226	48.35	82.6	81.2	1.4	5.8	86.4	85.8	0.6	1.2
AD073	2797	50.79	113421	125	48.27	97.5	96.0	1.5	0.5	89.7	89.0	0.7	1.2
AD086	6417	45.46	85097	158	48.60	96.7	94.4	2.3	0.8	85.1	84.4	0.7	1.9
AD266	471	41.25	150867	77	50.13	90.0	88.0	2.0	1.7	89.8	89.1	0.7	0.6
AM1000	5069	41.99	16545	784	48.39	94.3	92.6	1.7	2.2	81.9	81.2	0.7	4.1
AM980	27840	23.86	2648	655	47.76	1.6	1.4	0.2	0.3	93.3	89.4	3.9	0.4
AV005	151	39.25	647500	21	49.35	92.9	92.3	0.6	1.9	89.3	88.7	0.6	1.0
CK281	3629	45.73	29152	448	48.54	96.6	94.7	1.9	2.5	90.4	89.4	1.0	1.3
NVP60	12396	50.25	7755	1896	48.13	86.0	84.9	1.1	5.7	89.6	89.2	0.4	1.2
TTC192	6909	42.60	11619	1075	48.95	85.6	84.2	1.4	5.2	90.7	90.1	0.6	1.0

Table S6. **Summary statistics for *Mortierella spp.* and endobacterial assemblies.**

Variant calling. The all-genomic-loci dataset was processed using the following steps. Contigs were extracted using the draft genome *Linnemannia elongata* AD073 v1.0 (JGI Project ID: 1203123) as the reference genome for fungus and draft genome *Mycoavidus cysteinexigens* B1-EB (Genome ID: 1553431.3) from the PATRIC database as a reference for endobacteria. The reference fungal genome was processed using RepeatMasker [9]. BLASTN (-outfmt 6 -max_target_seqs 200) [8] was used to identify fungus and endobacteria in the de novo assembly against the corresponding reference genomes. Seqtk (subseq -l 60) [17] analyzed BLAST hits to recover a draft fungal genome and a draft endobacteria genome from the de novo assembly. Variant calling was performed with the MUMmer package [11] using the draft genomes against the reference genomes. Within the MUMmer suite [11], NUCmer was used to align the draft genome against the reference and show-snps identified the single nucleotide variants (SNV). Then, the MUMmerSNPs2VCF software was used to convert SNVs into a VCF-formatted file (software downloaded from https://github.com/liangjiaoxue/PythonNGSTools). Sequences with greater than 99.95% sequence similarity were pruned. The SNV MSA for *Mortierella spp.* was 4,607,802 bp long with 81.9% gappiness and 0.03% average normalized Hamming distance (ANHD); whereas the associated endobacteria had SNV alignments of length 215,165 bp with 47.4% gappiness and 0.22% ANHD.

S4 BACKWARDS-IN-TIME SIMULATIONS AND EXPERIMENTS

Simulation and experimental procedures. The backward-time model of [2] was used to simulate coevolution among n host taxa and n symbiont taxa, as well as host/symbiont associations. Our simulations explored varying numbers of taxa $n \in \{10, 50, 100, 500\}$. The simulations made use of a custom-modified Python program that was originally implemented by Avino et al. [2]. The simulation program takes a host tree as input and simulates a symbiont tree backward-in-time along the host tree by randomly drawing wait times to determine the timing and type of coevolutionary event(s) on a particular host tree branch. We used INDELible to sample host trees under a random birth-death model (see Supplementary Materials for more details). Model trees were deviated away from ultrametricity using Moret et al. [22]'s approach with deviation factor $c = 2.0$ [23]. We used custom scripts to perform the ultrametricity deviation calculations. We note that the Avino et al. [2]'s simulation software does not directly provide the model cophylogeny as output. Instead, a reference cophylogeny was obtained using eMPRess estimation on the true model trees for host and symbiont taxa as input. The choice of reference cophylogeny allows comparison of cophylogenetic estimation when ground truth inputs are provided (i.e., true model trees) versus cophylogenetic estimation when estimated trees are used as input.

Simulation of sequence evolution along model phylogenies followed the same procedure as in the mixed simulations. The substitution model parameters were based on empirical estimates from our re-analysis of the dataset from [10]'s study. Model condition parameter values and simulated dataset summary statistics are listed in Table S7.

As with the mixed simulations, additional experiments with varying evolutionary divergence were performed using the backward-time simulation procedure. The scaling parameter h was similarly set to a value from $\{0.1, 0.5, 1, 2, 5, 10\}$.

Results and discussion. Similar outcomes were observed in the backward-time simulation experiments, as compared to the mixed simulation experiments. Upstream tree estimation returned topological error of around 10% or less (Supplementary Figure S3). Estimated cophylogeny precision was also similar – ranging around 50% to 60%. Negative and significant correlation between upstream tree error and downstream cophylogeny precision was observed on all model conditions ($\alpha = 0.05$; $n = 100$), as shown in Supplementary Figure S7. Regression coefficients ranged between -0.644 and -0.848 (Table S8). Scatter around linear regression models was smaller than in the

Model conditions	Taxa	# taxa	Aln length	ANHD Avg	ANHD SE	Height Avg	Height SE	# cospec	# dup	# switch
backward-10	Host	10	1,000	0.6298	0.0008	2.6711	0.0191	5	1	2
	Symbiont	10	1,000	0.6820	0.0011	4.4742	0.0466			
backward-50	Host	50	1,000	0.7060	0.0002	8.8000	0.0465	15	13	12
	Symbiont	50	1,000	0.7232	0.0001	8.9585	0.1965			
backward-100	Host	100	10,000	0.7281	0.0000	8.1247	0.0439	34	32	47
	Symbiont	100	10,000	0.7283	0.0000	8.6243	0.0448			
backward-500	Host	500	10,000	0.7951	0.0039	4.6108	0.0077	157	177	271
	Symbiont	500	10,000	0.7894	0.0039	5.6020	0.0474			

Table S7. **Summary statistics for backward-time simulation datasets.** Each backward-time simulation condition ("Model conditions") varied the number of host and symbiont taxa ("# taxa") simulated under Avino et al. [2]'s backward-time coevolutionary model. The simulations included cospeciation, duplication, and host switch events, but not loss events.

backward-time simulations, with coefficient of determination between 0.653 and 0.938. One minor difference between backward-time simulation experiments and mixed simulation experiments is that former the returned more consistent regression analysis results compared to the latter. We attribute the difference in part to the relative heterogeneity of the mixed simulation conditions compared to the backward-time simulation conditions.

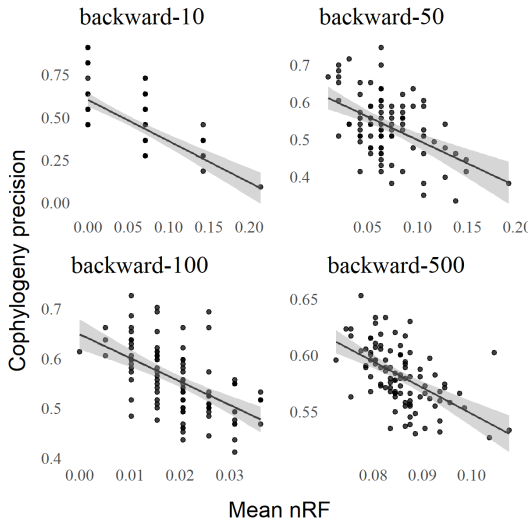


Fig. S7. **The relationship between phylogenetic and cophylogenetic estimation error on the backward-time simulation conditions.**

Simple Linear Regression						
Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
backward-10	0.6018	-0.6870	0.6525	0.1644	$< 10^{-14}$	$< 10^{-13}$
backward-50	0.6236	-0.7010	0.9074	0.0817	$< 10^{-7}$	$< 10^{-7}$
backward-100	0.6482	-0.6438	0.9379	0.0545	$< 10^{-9}$	$< 10^{-9}$
backward-500	0.7793	-0.8475	0.8950	0.0968	$< 10^{-9}$	$< 10^{-9}$

Table S8. **Linear regression results for backward-time simulation experiments.**

Simple Linear Regression						
Model conditions	intercept	B coefficient	R^2	RSE	p-value	q-value
backward-10	0.5458	-0.6163	0.7227	0.1541	$< 10^{-16}$	$< 10^{-183}$
backward-50	0.6049	-0.6578	0.9253	0.0783	$< 10^{-16}$	0
backward-100	0.5647	-0.6028	0.9566	0.0530	$< 10^{-16}$	0
backward-500	0.7152	-0.7807	0.9189	0.0936	$< 10^{-16}$	0

Table S9. **Linear regression results for backward-time simulation experiments with varying evolutionary divergence.**

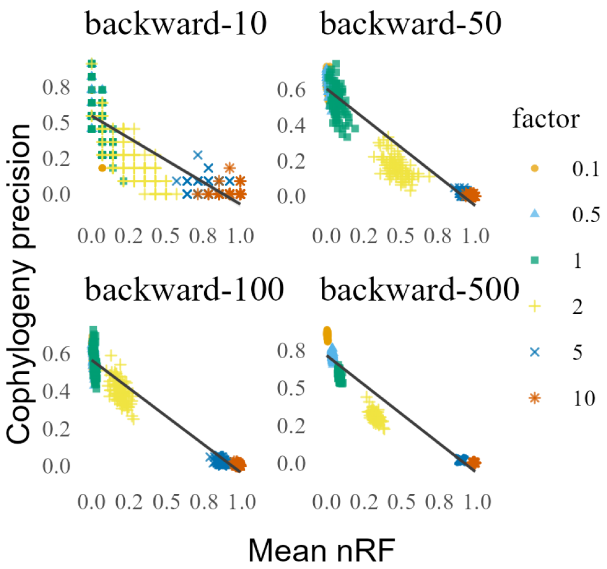


Fig. S8. **Backward-time simulation experiments: the impact of evolutionary divergence on phylogenetic and cophylogenetic estimation error.**

S5 COMPARISON BETWEEN DEFAULT EVENT COST PENALTY AND ALTERNATIVE EVENT PENALTIES

Experiments on event costs used for co-phylogenetic reconciliation. Reconciliations were assessed with different event costs estimated by COALA and CoRe-PA. On all forward-time model conditions, we found that the alternative event costs did not outperform the default event costs used by eMPRess (Figure S10). A similar outcome was observed on the mixed simulation conditions (Figure S9). For this reason, our performance study primarily utilizes default event costs to perform eMPRess analyses.

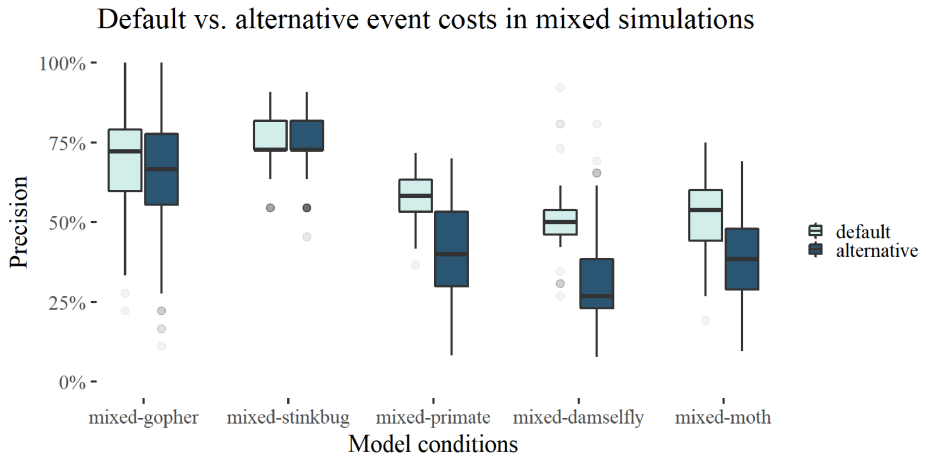


Fig. S9. **Effect of using default event cost versus COALA and CoRe-PA-estimated event frequencies in eMPRess reconciliations for mixed simulations.** Co-phylogenetic precision is reported across all model condition, each with $n = 100$ experimental replicates.

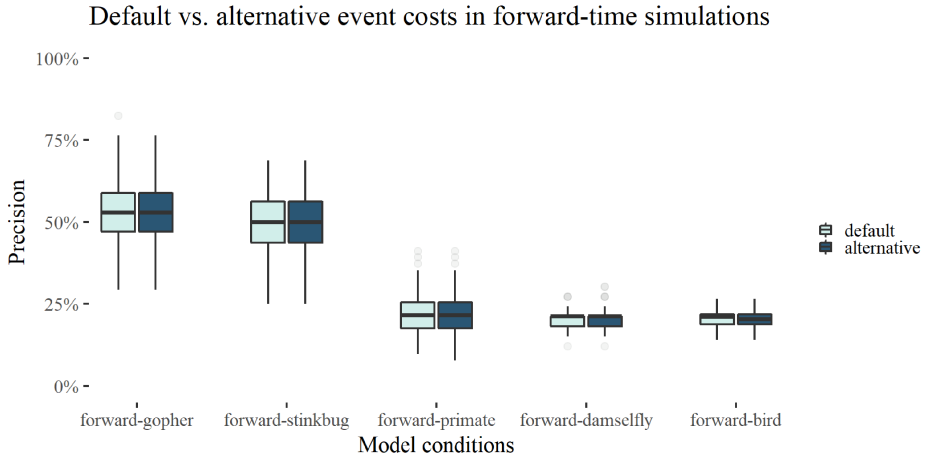


Fig. S10. **Effect of using default event cost versus COALA and CoRe-PA-estimated event frequencies in eMPRess reconciliations for forward simulations.** Co-phylogenetic accuracy is reported across all model condition, each with $n = 100$ experimental replicates.

S6 EXPERIMENTS WITH CORE-PA

Following the simulation methods section in the main paper, we used the same experimental conditions and reconstructed the cophylogenies using CoRe-PA [20] instead of eMPRess. In general, we obtained similar findings in CoRe-PA experiments as in the eMPRess experiments, thus confirming our findings in the main manuscript.

S6.1 Mixed simulation results with CoRe-PA

We obtained similar results using CoRe-PA as we did with eMPRess. There exists a negative correlation between cophylogeny precision and average host and symbiont tree topology error. The confidence band around the simple linear regressions were tight, indicating the data points clustered around the regression line.

Contrary to eMPRess results, the mixed-stinkbug model condition obtained a nearly horizontal regression line, showing that for this dataset, 15% perturbance in the tree topology did not result in appreciable change to the cophylogenetic precision, which remained low at under 5% cophylogenetic precision. The original annotation cophylogeny reconstruction was estimated using eMPRess, which predicted 5 cospeciations, 5 duplications, and 1 host switch event. On the other hand, CoRe-PA reconstructions on the replicate simulations on average predicted 2 cospeciations and 2 duplications. We attribute the finding to CoRe-PA's low cophylogenetic precision, which was among the lowest observed in our study. The topological error returned by its cophylogenetic reconstructions may overshadow the influence of upstream estimation error and other factors.

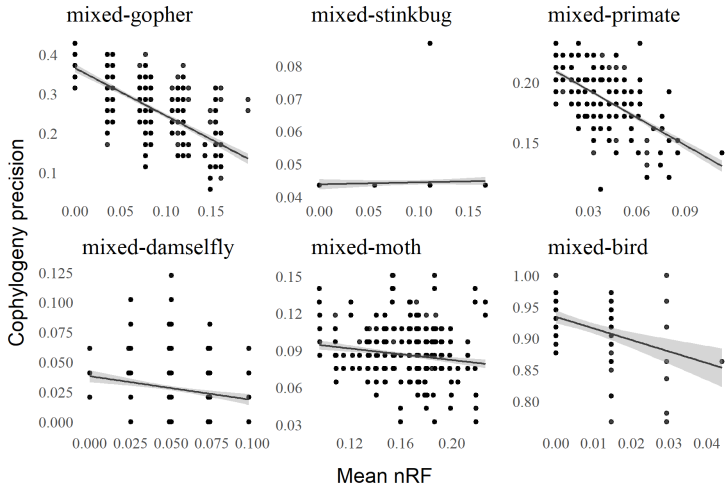


Fig. S11. Mixed simulation datasets: precision of CoRe-PA reconciliations compared with averaged host and symbiont tree normalized Robinson-Fould (nRF) distances. For each height scaling factor, a set of 100 replicates were simulated. Co-phylogenetic reconciliation precision was calculated as the aggregate statistic for events found in all of the replicate cophylogeny reconstructions and their respective, original annotation cophylogeny reconstruction.

S6.2 Backward-time simulation results with CoRe-PA

In backward-time simulations, we obtained similar results using CoRe-PA as we did with eMPRess such that there exists a negative correlation between cophylogeny precision and average host and symbiont tree topology error. The data points clustered around the regression line as indicated by the tight confidence band around the simple linear regressions line.

S6.3 Forward-time simulation results with CoRe-PA

In forward-time simulations, we obtained similar results using CoRe-PA as we did with eMPRess. We found a negative correlation between cophylogeny precision and average host and symbiont tree topology error. The confidence band around the simple linear regressions were tight, indicating

Simple Linear Regression					
Model conditions	intercept	B coefficient	R^2	RSE	p-value
mixed-gopher	0.3655	-1.2081	0.4621	0.0606	$< 10^{-16}$
mixed-stinkbug	0.0438	0.0061	0.0022	0.0061	0.4176
mixed-primate	0.2161	-0.7561	0.3726	0.0196	$< 10^{-16}$
mixed-damselfly	0.0381	-0.1989	0.0218	0.0264	$< 10^{-5}$
mixed-moth	0.1056	-0.1167	0.0194	0.0225	$< 10^{-6}$
mixed-bird	0.9341	-1.8328	0.1663	0.0408	$< 10^{-5}$

Table S10. Linear regression results for mixed simulation experiments involving CoRe-PA.

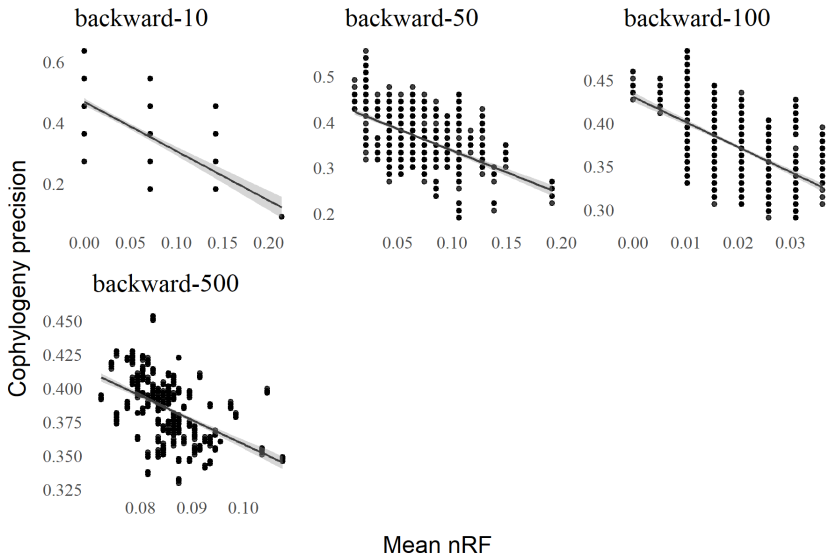


Fig. S12. Backward-time simulation datasets: precision of CoRe-PA reconciliations compared with averaged host and symbiont tree normalized Robinson-Fould (nRF) distances. For each height scaling factor, a set of 100 replicates were simulated. Co-phylogenetic reconciliation precision was calculated as the aggregate statistic for events found in all of the replicate cophylogeny reconstructions and their respective, original annotation cophylogeny reconstruction.

Simple Linear Regression					
Model conditions	intercept	B coefficient	R^2	RSE	p-value
backward-10	0.4689	-1.6189	0.3565	0.1031	$< 10^{-16}$
backward-50	0.4327	-0.9491	0.2813	0.0481	$< 10^{-16}$
backward-100	0.4305	-2.9033	0.3227	0.0333	$< 10^{-16}$
backward-500	0.5380	-1.7934	0.2201	0.0210	$< 10^{-16}$

Table S11. Linear regression results for backward-time simulation experiments involving CoRe-PA.

the data points clustered around the regression line. The forward-damselfly model condition corresponded with the mixed-damselfly model condition in mixed simulations, which also demonstrated

a linear regression line slope that was smaller in magnitude in CoRe-PA results than in eMPRess results. Similarly, forward-bird model condition corresponded with the mixed-bird model condition in mixed simulations, and it also demonstrated a linear regression line slope that was smaller in magnitude in CoRe-PA results than in eMPRess results. Contrary to mixed simulations, forward-stinkbug obtained a trendline closer to model conditions mixed-stinkbug and forward-stinkbug from eMPRess results.

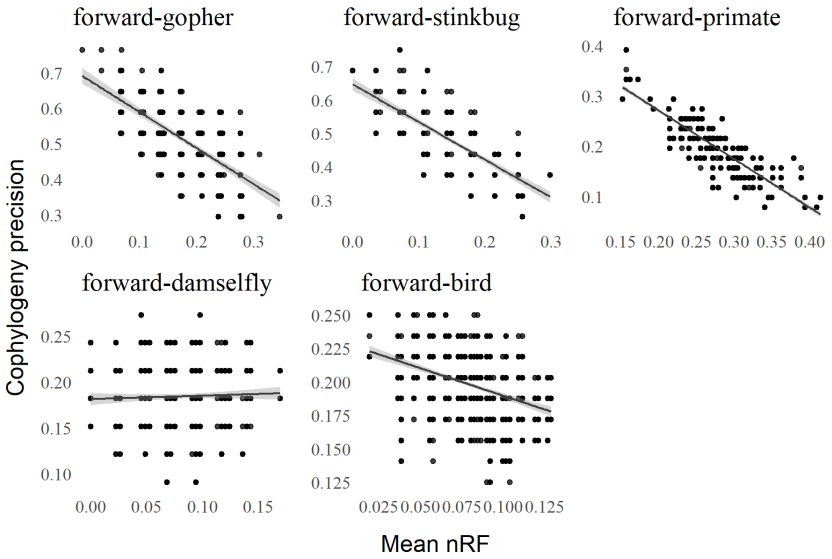


Fig. S13. Forward-time simulation datasets: accuracy of CoRe-PA reconciliations compared with averaged host and symbiont tree normalized Robinson-Fould (nRF) distances. Co-phylogenetic reconciliation accuracy was calculated as the aggregate statistic for events found in the 100 replicate cophylogeny reconstructions that were also found in the true coevolutionary history.

Simple Linear Regression					
Model conditions	intercept	B coefficient	R^2	RSE	p-value
forward-gopher	0.6913	-1.0173	0.4635	0.0707	$< 10^{-16}$
forward-stinkbug	0.6470	-1.1315	0.5401	0.0641	$< 10^{-16}$
forward-primate	0.4654	-0.9690	0.7348	0.0315	$< 10^{-16}$
forward-damselfly	0.1813	0.0374	0.0014	0.0346	0.3090
forward-bird	0.2309	-0.4118	0.1380	0.0230	$< 10^{-16}$

Table S12. Linear regression results for forward-time simulation experiments involving CoRe-PA.

S7 MULTILOCUS SIMULATION EXPERIMENT WITH TALE

TALE [19] is a new, multilocus, probabilistic DTL-based cophylogenetic reconstruction method. We conducted an experiment using multilocus data simulations to assess how reproducible TALE's cophylogenetic reconciliations are in the presence of phylogenetic uncertainty.

Experimental design. The model host and symbiont species trees and true MSA for gopher-lice match that of our main manuscript's mixed-gopher model condition summary statistics. From an empirical MSA obtained from [14], we reconstructed host and symbiont species trees using maximum likelihood estimation under the GTRGAMMA model implemented in RAxML v8.2.12

[30]. We simulated 100 gene trees under the model symbiont species tree using SimPhy [18]. We followed the SimPhy simulation procedure from [21], where the height of the SimPhy input tree was adjusted to 2 million generations tall, and used the same SimPhy parameters as [21] (Table S13.)

Parameter	Description	Value
rl	Number of locus per replicate	100
rg	Number of gene trees per locus tree	1
sp	Population size n	200,000
su	Global substitution rate	1,000,000
hs	Species branch rate heterogeneity modifiers	Log normal (1.5,1)
hl	Locus rate heterogeneity modifiers	Log normal (1.2,1)
hg	Gene-tree-branch rate heterogeneity modifiers	Log normal (1.4,1)
cs	Seed for random number generator	22

Table S13. SimPhy parameters used in to simulate gene trees under the three-tree model.

Cophylogenetic reconstruction with TALE. TALE's input consisted of the host species tree, symbiont species tree, the set of 100 simulated gene trees, host-symbiont mapping, and symbiont-gene mapping. We used TALE to perform cophylogenetic reconstruction under its sequential heuristic algorithm, which was shown by the original authors to provide similar recall and precision as the more theoretically more robust Monte Carlo algorithm [19].

Experimental replication. We repeated the procedure to obtain 10 replicates.

Phylogenetic and cophylogenetic reconstruction and assessment. The phylogenetic inference methods matched that of the simulation study in the main manuscript. The TALE reconciliations on each replicate dataset were compared against the reference TALE reconciliation. We followed the main manuscript and assessed phylogenetic uncertainty alongside of cophylogenetic reconciliation precision using linear regression. Note that we calculated the tree topology error by comparing the species trees as specified by TALE to correspond to each of its output reconciliations.

Results and discussion. In this experiment (Figure S14), we observe similar correlation between species tree topology uncertainty and cophylogenetic reconciliation precision, compared to the experiments described in the main manuscript. Similar to parsimony-based cophylogenetic reconstruction methods, TALE's cophylogenetic reconstruction accuracy is impacted by phylogenetic estimation error. The experiment therefore confirms the main finding in our study.

A secondary finding was that TALE returned lower cophylogenetic reconstruction precision in this simulation experiment compared to eMPRess in the main simulation study, as well as in our own testing of TALE using datasets from the study of Menet et al. [19] (results not shown). Several factors help to explain the discrepancy. First, we note that TALE only partially accounts for phylogenetic estimation error on the symbiont side (since symbiont gene trees are assumed to be correct) and not at all on the host side (since host species trees are wholly assumed to be correct). Phylogenetic estimation error is present in our simulation experiment and in real-world practice, and its presence conflicts with common simplifying assumptions about input tree correctness. Another key difference is that Menet et al. [19]'s simulation experiments utilized true species trees and true gene trees as input to TALE while our simulation experiments utilized TALE inputs that included some phylogenetic estimation error. Finally, TALE, eMPRess, and other state-of-the-art cophylogenetic reconstruction methods do not explicitly account for genetic drift and incomplete

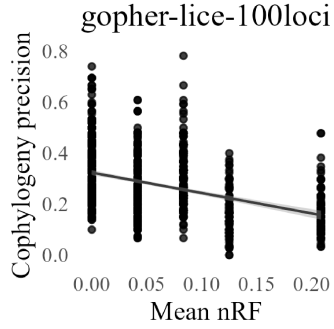


Fig. S14. TALE experiment on gopher-lice dataset with 100 simulated symbiont gene trees.

lineage sorting. The latter are omnipresent in this simulation experiment and in natural genome evolution. More work is needed to address this model misspecification.

Finally, we note that Menet et al. [19]'s main finding that 3-level DTL reconciliation (e.g., DTL modeling at the host species and symbiont species level and then nested DTL modeling at the symbiont species and symbiont gene level) outperforms 2-level DTL modeling (e.g., traditional DTL modeling at the species and gene level) complements our main study finding. The general theme emerging from both studies is the need for integrated computational frameworks that account for upstream and downstream estimation error and uncertainty in a unified manner.

TALE command.

```
python3 TALE-main/src/main.py {symbiont directory}/ {genes directory}/
-mf {symbiont-gene-mapping} -tl {host directory}/ -imf {host-symbiont-mapping}
-b -o {out directory}/ -ncpu 8
```

S8 COMMANDS TO RUN COPHYLOGENETIC RECONCILIATION SOFTWARE

EMPRess v1.2.1 [25] was used to reconcile cophylogenies in two ways. First, we ran eMPRess v1.2.1 with default cost scheme.

```
python empress_cli.py reconcile {host tree file} {symbiont tree file}
{extant species associations} --csv {out file name}.csv
```

Second, we ran eMPRess v1.2.1 with modified event cost schemes.

```
python empress_cli.py reconcile {host tree file} {symbiont tree file}
{extant species associations} {event cost frequencies} --csv {out file}.csv
```

CoRe-PA version 0.5.2 [20] was used to reconcile cophylogenies and to generate alternative event cost schemes.

```
java -jar core-pa_cli_0.5.2.jar -i {CoRe-PA's nexus format file} -o {out file}
```

COALA version 1.2.1 [4] was used to calculate alternative event cost schemes.

```
java -Xms4056M -Xms8g -jar Coala-1.2.1.jar -input {nexus format file}
-cluster -threads 16
```

S9 COMMANDS USED IN EMPIRICAL EXPERIMENTS

Note that texts inside curly brackets {} indicate files and inputs the user passes into the software, thus they are not part of the command.

BBtools version 37.62 [7] was invoked to run BBMap, BBDuk, and Reformat. The following BBDuk command was used to filter and trim Illumina short reads to reduce artifacts and contaminants.

```
# run if lanes 1 and 2 are separate files
bbduk.sh in1={lane1 reads} in2={lane2 reads} out1={paired reads 1}
    out2={paired reads 2} ref=bbmap_adaptor.fa forcetrimleft=5 minlen=90

# run if you have interleaved reads
bbduk.sh in={interleaved reads} out1={lane1 reads} out2={lane2 reads}
reformat.sh in1={lane1 reads} in2={lane2 reads} out1={paired reads 1}
    out2={paired reads 2} ref=bbmap_adaptor.fa forcetrimleft=5 minlen=90

# produce summary statistics for assembly
statswrapper.sh {assembly} format=4 >> {out file}
```

SPAdes version 3.15.5 [3] was used to assemble paired short reads.

```
spades.py -k 21,33,55,77,99,127 -o {directory} -1 {paired reads 1}
    -2 {paired reads 2} -t 16
```

BUSCO version 5.3.2 [29] was used to assess the completeness of the assemblies.

```
busco -i $fungi -l burkholderiales_odb10 -o {out directory} -m genome -c 4
    --force #bacteria
busco -i $endobac -l mucoromycota_odb10 -o {out directory} -m genome -c 4
    --force #fungi
```

CANU version 2.2 [16] was used to assemble PacBio long reads.

```
canu -p {assembly prefix} -d {directory} genomeSize={size in bases} -pacbio {pacbio reads}
```

BLAST+ version 2.2.31 was used to query *Mortierella spp.* and endobacterial assembled contigs from their respective *de novo* assemblies. Seqtk version 1.3 was used to extract contigs from assembly using the blasted bed file to produce fasta format contigs.

```
blastn -query {assembly} -outfmt 6 -max_target_seqs 200 -db {reference} -out {blast file}
 awk '!_[$1]++' {blast file} > {bed file}
 seqtk subseq -l 60 {blast file} {bed file} > {fasta file}
```

MUMmer version 3.23 [11] was used to variant call the extracted *Mortierella spp.* and endobacterial contigs against their respective reference genomes. SAMtools version 1.15 was used to index and retrieve the VCF file.

```
nucmer --prefix={prefix name} {blasted contigs} {reference genome}
show-snps -Clr -x 1 -T {SNPs prefix}.delta > {SNPs prefix}.snps
MUMmerSNPs2VCF.py {SNPs prefix}.snps {SNPs prefix}.vcf
bgzip -c {SNPs prefix}.vcf > {SNPs prefix}.vcf.gz
tabix -p vcf {SNPs prefix}.vcf.gz
```

Barrnap version 0.9 [27] was used to extract rRNA genes from *Mortierella spp.* assembly.

```
barrnap --kingdom euk --threads 8 -o {out directory} < {assembly} > {extract rRNA genes}
```

PROKKA version 1.14.6 [26] was used to extract rRNA genes from *Mortierella's* endobacterial assembly.

```
prokka {assembly} --centre X --compliant --force
```

RAxML version 8.2.12 [30] was used to reconstruct phylogenies under specified software (GTR, HKY85, JC69, and K80).

```
raxmlHPC -m GTRGAMMA -s {unrooted tree} --{software} -p {random number}
-n {out file suffix}
```

RAxML version 8.2.12 [30] was used to bootstrap alignments.

```
raxmlHPC -f j -b {random number} -# {number of samples} -m GTRGAMMA
-s {alignment} -n {out file suffix}
```

RAxML version 8.2.12 [30] was used to midpoint root the phylogenies.

```
raxmlHPC -f I -m GTRCAT -t {unrooted tree} -n {rooted tree file suffix}
-p {random number}
```

PAUP* 4.0 [31] was used to reconstruct phylogenies under NJ, UPGMA, and SVDquartet.

```
paup4a168_centos64
exe {alignment file}
{lower case model name}
savetree file={out tree file} brlen=yes
quit
```

Linear regression was performed using base R version 4.2.2 with the following code.

```
lm(precision ~ avg_nRF, df)
```

S10 COMMANDS USED IN SIMULATION EXPERIMENTS

Note that texts inside curly brackets {} indicate files and inputs the user passes into the software, thus they are not part of the command.
MAFFT v7.490 [15] was used to align sequences in empirical datasets that provided unaligned sequence data.

```
mafft {unaligned sequence file} > {alignment file}
```

Seq-Gen v1.3.4 [24] was used to simulate gap-less alignments under model species trees from parameters obtained from running RAxML v8.2.12 [30] on the original empirical alignments.

```
seq-gen -mGTR -r{GTR rate parameters} -z {random number} -or
    -l{simulated alignment length} -f{nucleotide frequencies}
    < {model species tree file} > {simulated alignment file}
```

Seq-Gen v1.3.4 [24] was used to simulate gap-less alignments under model species trees from parameters obtained from running RAxML v8.2.12 [30] on the original empirical alignments.

RAxML version 8.2.12 [30] was used to reconstruct phylogenies under the GTR model.

```
raxmlHPC -m GTRGAMMA -s {alignment file} -p {random number} -n {tree file suffix}
```

RAxML version 8.2.12 [30] was used to midpoint root the phylogenies.

```
raxmlHPC -f I -m GTRCAT -t {unrooted tree} -n {rooted tree file suffix}
-p {random number}
```

INDELible version 1.03 [13] was used to simulate n-taxa trees that serve as input to reverse-time simulator originally from [2]. To run INDELible, use the following command in the same folder as a INDELible control file called "control.txt".

```
indelible
```

We used the following code in INDELible control file to sample an n-taxa tree topology under a birth-death model with birth rate 2.4, death rate 1.1, sampling fraction 0.2566, and mutation rate 0.34.

```
[TYPE] NUCLEOTIDE 1
[TREE] tree1
```

```
  [unrooted] 10 2.4 1.1 0.2566 0.34
```

We used the following code in INDELible control file to assign branch lengths using the GTR parameter rates and nucleotide frequencies from the original annotation of the empirical dataset [10] on avian feather lice.

```
[TYPE] NUCLEOTIDE 1
[MODEL] GTRmodel
  [submodel] GTR 1.475477 4.831617 1.410614 1.732842 7.069432
  [statefreq] 0.319 0.192 0.223 0.266
[TREE] tree1 {newick format tree topology from previous INDELible step}
  [branchlengths] NON-ULTRAMETRIC
[PARTITIONS] taxapartition
  [tree1 GTRmodel 1000]
[EVOLVE] taxapartition 1 species_tree
```

Relative variable importance was calculated using randomForest package in R version 4.2.2 with the following code. Linear regression was performed using base R version 4.2.2 with the following code.

```
randomForest(study ~ .,data=df,ntree=1000,importance=TRUE)
```

Linear regression was performed using base R version 4.2.2 with the following code.

```
lm(precision ~ avg_nRF, df)
```

S11 COMMANDS TO RUN SIMULATOR SOFTWARE

A modified version of the reverse-time nested coalescent simulator by [2] was used to simulate host tree, symbiont tree, and output the true coevolutionary history. To the best of our knowledge, this simulator was not published under copyleft license, therefore we could not include the modified scripts used in this performance study. The following command was used to run the original reverse-time cophylogeny simulator.

```
python nestedCoalescent.py {rooted host tree file} 0.8 0.3 0.4 {symbiont tree file}
```

Treeducken v1.1.0 [12] R software was used to simulate the host tree, the symbiont tree, and the extant species associations. We modified Treeducken data structures to additionally output the true coevolutionary history for the pair of trees in the next section. The following R code was used to run Treeducken v1.1.0 software.

```
library(treeducken)
lambda_H <- {see Treeducken parameters table}
mu_H <- {see Treeducken parameters table}
lambda_C <- {see Treeducken parameters table}
lambda_S <- {see Treeducken parameters table}
mu_S <- {see Treeducken parameters table}
time <- {see Treeducken parameters table}
cophy_obj <- sim_cophylo_bdp(hbr = lambda_H,
                             hdr = mu_H,
                             sbr = lambda_S,
                             sdr = mu_S,
                             cosp_rate =lambda_C,
                             host_exp_rate = 0.0,
                             time_to_sim = time,
                             numbsim = 1)
```

S12 MODIFIED TREEDUCKEN CODE

The following R code was used to modify Treeducken's data structures post simulation to rename coevolution events and output the desired format trees with internal node labeling as well as the true, coevolutionary history.

```
library(treeducken)
library(ape)
library(geiger)

# Run Treeducken as normal
lambda_H <- {see Treeducken parameters table}
mu_H <- {see Treeducken parameters table}
lambda_C <- {see Treeducken parameters table}
lambda_S <- {see Treeducken parameters table}
mu_S <- {see Treeducken parameters table}
time <- {see Treeducken parameters table}
cophy_obj <- sim_cophylo_bdp(hbr = lambda_H,
                             hdr = mu_H,
                             sbr = lambda_S,
                             sdr = mu_S,
                             cosp_rate =lambda_C,
                             host_exp_rate = 0.0,
                             time_to_sim = time,
                             numbsim = 1)

# Start modifying phylo and associations data objects
# to output the coevolutionary history with the event types we want

### label internal nodes ###
label_internal_nodes <- function(tree){ #where tree is a phylo object
  tot_internal_nodes<-tree$Nnode # total number of nodes
  start_internal_nodes<-length(tree$tip.label)+1
  end_internal_nodes<-start_internal_nodes+tot_internal_nodes-1
  labels<-list()
  for (i in start_internal_nodes:end_internal_nodes){
    # nodes start incrementing from number of tips
    name<-paste(tips(tree,i),collapse = "_")
    labels <- append(labels, name)
  }
  tree$node.label <- labels
  new_tree <- write.tree(tree)
  return(new_tree)
}
output_unlabeled_tree<-function(tree){
  print(tree)
  new_tree <- write.tree(tree)
  return(new_tree)
}
```

```
  #host
  write.table(output_unlabeled_tree(cophy_obj[[1]]$host_tree), file_host,
              append = FALSE, sep = " ",
              row.names = FALSE, col.names = FALSE,
              quote=FALSE)
write.table(label_internal_nodes(cophy_obj[[1]]$host_tree), file_host_labeled,
              append = FALSE, sep = " ",
              row.names = FALSE, col.names = FALSE,
              quote=FALSE)
  #symb
  write.table(output_unlabeled_tree(cophy_obj[[1]]$symb_tree), file_symb,
              append = FALSE, sep = " ",
              row.names = FALSE, col.names = FALSE,
              quote=FALSE)
write.table(label_internal_nodes(cophy_obj[[1]]$symb_tree), file_symb_labeled,
              append = FALSE, sep = " ",
              row.names = FALSE, col.names = FALSE,
              quote=FALSE)

  ### relabel event history to format: event host_node symb_node) ###
  #where tree is a phylo object
relabel_treeducken_event_history <- function(event_history, hosttree, symbtree){
    #host trees
    tot_internal_nodes_h<-hosttree$Nnode # total number of nodes
    num_leaf_host<-length(hosttree$tip.label)
    start_internal_nodes_h<-num_leaf_host+1
    end_internal_nodes_h<-start_internal_nodes_h+tot_internal_nodes_h-1
    labels_host<-list()
    for (i in start_internal_nodes_h:end_internal_nodes_h){
      # nodes start incrementing from number of tips
      name<-paste(tips(hosttree,i),collapse = "_")
      labels_host <- c(labels_host, name)
    }
    hosttree$node.label <- labels_host
    #symb trees
    tot_internal_nodes_s<-symbtree$Nnode # total number of nodes
    num_leaf_symb<-length(symbtree$tip.label)
    start_internal_nodes_s<-num_leaf_symb+1
    end_internal_nodes_s<-start_internal_nodes_s+tot_internal_nodes_s-1
    labels_symb<-list()
    for (i in start_internal_nodes_s:end_internal_nodes_s){
      # nodes start incrementing from number of tips
      name<-paste(tips(symbtree,i),collapse = "_")
      labels_symb <- c(labels_symb, name)
    }
    symbtree$node.label <- labels_symb
    num_events<-nrow(event_history)
    events<-c()
```

```
    hosts<-c()
    symbs<-c()
    prefix_host<-"H"  # H for host, S for symb
    prefix_symb<-"S"

  # update event names in Treeducken to the known 4 events that works with cophy software
  # https://github.com/wadedismukes/treeducken/blob/main/src/Simulator.cpp#L682
  treeducken_events=c("SX", "HX", "SSP", "HSP", "AG", "AL", "CSP", "DISP","EXTP", "SHE", "
    known_events=c("loss", "loss", "duplication", "host_switch",
                   "duplication", "loss", "cospeciation", "cospeciation",
                   "loss", "host_switch","host_Switch")
    event_renaming=data.frame(treeducken_events, known_events)
    # mapping to known format event history
    for (i in 1:num_events){
      print(i)
      if (event_history$Event_Type[i] == "I"){
        print("Initialized")
        #skip this one, "I" stands for initialize event vector.
        next
      }
      else{
      new_event<-event_renaming$known_events[event_renaming$treeducken_events
                                             ==event_history$Event_Type[i]]
        events <- c(events, new_event) # events
      }
      if (event_history$Host_Index[i] > num_leaf_host){ #hosts
       hosts <- c(hosts, labels_host[event_history$Host_Index[i]-num_leaf_host])
      }
      else{
        hosts <- c(hosts, paste0(prefix_host,event_history$Host_Index[i]))
      }
      if (event_history$Symbiont_Index[i] > num_leaf_symb){ #symbs
     symbs <- c(symbs, labels_symb[event_history$Symbiont_Index[i]-num_leaf_symb])
      }
      else{
        symbs <- c(symbs, paste0(prefix_symb,event_history$Symbiont_Index[i]))
      }
    }
    new_event_history<-data.frame(events, paste(hosts, sep=" "),
              data.frame("symbs" = paste(symbs, sep=" ")))
              colnames(new_event_history) <- c("events", "hosts", "symbs")
    print(new_event_history)
    return(new_event_history)
  }
 new_event_history<-relabel_treeducken_event_history(cophy_obj[[1]]$event_history,
              cophy_obj[[1]]$host_tree, cophy_obj[[1]]$symb_tree)
  write.table(new_event_history, file_event_history,
              append = FALSE, sep = " ",
```

```
            row.names = FALSE, col.names = FALSE,
            quote=FALSE)
### output nexus and empress association links ###
Which.names <- function(DF, value, file_empress_link, file_nexus_link){
  ind <- as.data.frame(which(DF==value, arr.ind=TRUE, useNames =TRUE))
  print(ind)
  num_links<-length(colnames(DF))
  links_empress<-""
  links_nexus<-""
  for (i in 1:num_links){
    symb<-colnames(association_mat)[ind$col[i]]
    host<-rownames(association_mat)[ind$row[i]]
    links_empress<-paste(links_empress,paste(symb, host ,sep=":"), sep="\n")
  links_nexus<-paste(links_nexus,paste0("'",symb,"':'",host,"',",collapse=""), sep="\n
  }
  links_empress<-sub(".", "", links_empress) # remove first character \n
  links_nexus<-sub(".", "", links_nexus)
  cat(links_empress)
links_nexus <- gsub(".{1}$", ";",links_nexus) # replace last character with ";"
  cat(links_nexus)
  write(links_empress, file_empress_link)
  write(links_nexus, file_nexus_link)
}
association_mat<-cophy_obj[[1]]$association_mat
# where cell value is 1 means association exists
Which.names(association_mat, 1, links_empress, links_nexus)
cophy_obj[[1]]$host_tree$Nnode
cophy_obj[[1]]$symb_tree$Nnode
length(new_event_history$events)
sum(new_event_history$events=="cospeciation")
sum(new_event_history$events=="duplication")
sum(new_event_history$events=="host_switch")
num_links<-length(colnames(association_mat))
ind <- as.data.frame(which(association_mat==1, arr.ind=TRUE, useNames =TRUE))
all_symb=c()
# the following only matters if the cophylogenetic software doesn't allow
# a symbiont to associate with multiple hosts. eMPRess and CoRe-PA don't mind.
for (i in 1:num_links){
  symb<-colnames(association_mat)[ind$col[i]]
  if(sum(all_symb==symb) < 1){
    all_symb<-append(all_symb,symb)
  }
  else{
    print("symb lineage on multiple hosts.")
    break
  }
}
```

REFERENCES

[1] Simon Andrews. FastQC: a quality control tool for high throughput sequence data, 2010.
[2] Mariano Avino, Garway T Ng, Yiying He, Mathias S Renaud, Bradley R Jones, and Art FY Poon. Tree shape-based approaches for the comparative study of cophylogeny. *Ecology and Evolution*, 9(12):6756–6771, 2019.
[3] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.
[4] Christian Baudet, Béatrice Donati, Blerina Sinaimeri, Pierluigi Crescenzi, Christian Gautier, Catherine Matias, and M-F Sagot. Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431, 2015.
[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.
[6] Gregory Bonito, Khalid Hameed, Rafael Ventura, Jay Krishnan, Christopher W Schadt, and Rytas Vilgalys. Isolating a functionally relevant guild of fungi from the root microbiome of Populus. *Fungal Ecology*, 22:35–42, 2016.
[7] B Bushnell. BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. 2018.
[8] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):1–9, 2009. Publisher: Springer.
[9] Nansheng Chen. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, 5(1):4–10, 2004.
[10] Robert S de Moya, Julie M Allen, Andrew D Sweet, Kimberly KO Walden, Ricardo L Palma, Vincent S Smith, Stephen L Cameron, Michel P Valim, Terry D Galloway, Jason D Weckstein, et al. Extensive host-switching of avian feather lice following the Cretaceous-Paleogene mass extinction event. *Communications Biology*, 2(1):445, 2019.
[11] Arthur L Delcher, Steven L Salzberg, and Adam M Phillippy. Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics*, pages 10–3, 2003.
[12] Wade Dismukes and Tracy A Heath. treeducken: An R package for simulating cophylogenetic systems. *Methods Ecol. Evol.*, 12(8):1358–1364, 2021.
[13] William Fletcher and Ziheng Yang. INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, 2009.
[14] Mark S Hafner, Philip D Sudman, Francis X Villablanca, Theresa A Spradling, James W Demastes, and Steven A Nadler. Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science*, 265(5175):1087–1090, 1994.
[15] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
[16] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5): 722–736, 2017. Publisher: Cold Spring Harbor Lab.
[17] Heng Li. seqtk, 2018.
[18] Diego Mallo, Leonardo de Oliveira Martins, and David Posada. SimPhy: phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, 65(2):334–344, 2016.
[19] Hugo Menet, Alexia Nguyen Trung, Vincent Daubin, and Eric Tannier. Host-symbiont-gene phylogenetic reconciliation. *Peer Community Journal*, 3:e47, 2023.
[20] Daniel Merkle, Martin Middendorf, and Nicolas Wieseke. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC bioinformatics*, 11(1):1–10, 2010.
[21] Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.
[22] Bernard ME Moret, Usman Roshan, and Tandy Warnow. Sequence-length requirements for phylogenetic methods. volume 2452, pages 343–356. Springer, 2002.
[23] S Nelesen, Kevin Liu, Donggao Zhao, C Randal Linder, and Tandy Warnow. The effect of the guide tree on multiple sequence alignments and subsequent phylogenetic analyses. In *Biocomputing 2008*, pages 25–36. World Scientific, 2008.
[24] Andrew Rambaut and Nicholas C Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.
[25] Santi Santichaivekin, Qing Yang, Jingyi Liu, Ross Mawhorter, Justin Jiang, Trenton Wesley, Yi-Chieh Wu, and Ran Libeskind-Hadas. eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, 37(16):2481–2482, 2021.
[26] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014. Publisher: Oxford University Press.
[27] Torsten Seemann. Barrnap, 2018.

[28] T Shirouzu, D Hirose, and S Tokumasu. Biodiversity survey of soil-inhabiting mucoralean and mortierellalean fungi by a baiting method. *T Mycol Soc Jpn*, 53:33–39, 2012.
[29] Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015. Publisher: Oxford University Press.
[30] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014. Publisher: Oxford University Press.
[31] David L. Swofford. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, 2003.
[32] JH Warcup. The soil-plate method for isolation of fungi from soil. *Nature*, 166(4211):117–118, 1950.