Reconstructing Phylogenies Using Branch-Variable Substitution Models and Unaligned Biomolecular Sequences: A Performance Study and New Resampling Method

Rei Doko
Michigan State University
Computer Science and Engineering
East Lansing, Michigan, USA

Kevin Liu
kjl@msu.edu
Computer Science and Engineering
Ecology, Evolution, and Behavior Program
Genetics and Genome Sciences
Michigan State University
East Lansing, Michigan, USA

ABSTRACT

In many clades in the Tree of Life, nucleotide substitution rates and base frequencies are hypothesized to have changed as genome evolution unfolded over time. Rigorous testing of this hypothesis relies on accurate phylogenetic reconstruction under suitable models of biomolecular sequence evolution. By far the most common approach for phylogenetic reconstruction is a "two-phase" analysis, where unaligned biomolecular sequence data are first aligned, and the resulting multiple sequence alignment (MSA) is used as input to downstream phylogenetic reconstruction. For a traditional "homogeneous" substitution model that is fixed across a species phylogeny, it has long been established that accurate phylogenetic inference and learning requires accurate upstream multiple sequence alignments. But the same question has not been carefully studied for "heterogeneous" models of substitution processes that can vary across the branches of a phylogeny.

We therefore conducted a comprehensive performance study to quantify the impact of upstream MSA estimation error on downstream phylogenetic inference and learning under branch-variable models of nucleotide substitution. Across model conditions with either 10 or 20 taxa and spanning a range of evolutionary divergence, we find a consistent and significantly positive association between upstream and downstream estimation error. The relationship is robust to the choice of MSA estimation method as well as substitution model mis-specification. We further quantify the relatively large contribution of upstream MSA estimation error to downstream phylogenetic reconstruction quality, compared to other experimental factors. We also conducted an empirical study of flowering monocots. Phylogenetic analyses of orthologous genes in the clade confirm the simulation study findings, and species tree estimation using branch-variable substitution models reveals new insights into sequence evolution heterogeneity. Our findings underscore several key gaps in the state of the art, including the need for MSA-aware phylogenetic inference and learning methods under heterogeneous models of sequence evolution.

To this end, we introduce a new computational method, NoHTS ("Non-Homogeneous Tree Support"), to directly assess phylogenetic estimation uncertainty due to MSA estimation error and other factors. The new method uses sequence-aware statistical resampling to place confidence intervals on a phylogeny estimated under a branch-variable substitution model. We demonstrate its superior type I and type II error versus a de facto standard in phylogenetic and phylogenomic studies – the phylogenetic bootstrap method.

CCS CONCEPTS

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

KEYWORDS

multiple sequence alignment, maximum likelihood estimation, branch-variable substitution model, phylogenetic tree, simulation study, Poales, phylogenetic support, bootstrap, resampling

ACM Reference Format:
Rei Doko and Kevin Liu. 2023. Reconstructing Phylogenies Using Branch-Variable Substitution Models and Unaligned Biomolecular Sequences: A Performance Study and New Resampling Method . 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/10.1145/3584371.3613011

1 INTRODUCTION

Throughout the Tree of Life, interspecific genomic variation in base composition and other features are thought to signify heterotachy and other shifts in biomolecular sequence evolution over time. Important models of this phenomenon include grasses [5], insects [4], and birds [26]. Knowing when and where these compositional biases arise in the evolutionary history of these organisms will help shed light into its functional significance. For example, nucleotide composition changes are hypothesized to result in concomitant downstream changes in transcription and translation (e.g., codon usage bias and cellular distributions of different amino acids) [18].

Rigorous statistical testing of this evolutionary hypothesis relies on statistical models and computational methods to reconstruct phylogenies using biomolecular sequence data, where the latter

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

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3613011

include models of sequence evolution that can vary across the branches of a phylogeny. By far the most widely used approach is a "two-phase" method: (1) a multiple sequence alignment (MSA) is first estimated on the input set of unaligned biomolecular sequences, and (2) a phylogeny is then reconstructed using the estimated MSA as input.

Many methods have been developed for addressing the initial phase of multiple sequence alignment, which is a classical and well-studied computational problem in computational biology and bioinformatics. The problem is also known to be NP-Complete [37]. For this reason, a variety of MSA heuristics have been developed. One such heuristic is progressive multiple sequence alignment, where an input guide tree is used to successively perform pairwise alignment of alignments. Among the most accurate and most popular MSA methods are MAFFT [16], MUSCLE [8], Clustal Omega [34] and its predecessor Clustal W [20], and FSA [2].

The second phase in a two-phase analysis consists of phylogenetic reconstruction using an estimated MSA as input. Statistical inference and learning methods for this task typically model sequence evolution as a branch-homogeneous substitution process (i.e., a substitution process that does not vary across the branches of a phylogeny). More general branch-variable substitution models have been proposed, and phylogenetic software packages for performing statistical inference and learning under these models have also been developed. These software are broadly classified by their statistical optimization criteria. One class consists of maximum likelihood estimation (MLE) methods. PAML [43] is a widely used method in this class, as well as nhPhyML [1]. Other popular phylogenetic MLE methods include limited support for branch-variable substitution models (e.g., RAxML [35] supports a leaf-versus-internal-edge substitution model). Another class of statistical methods utilize Bayesian inference and learning. BEAST [7] is widely used for Bayesian phylogenetic estimation, and it supports relaxed molecular clock models of rate heterogeneity among branches. MrBayes [33] is another option in this class, but it only offers limited support for covarion models [14]. We focus on PAML as a representative state-of-the-art method that is scalable to relatively large datasets.

A variety of factors contribute to phylogenetic reconstruction accuracy. Beyond the question of branch-variable sequence evolution, a large body of studies has demonstrated the central importance of estimated MSA quality in traditional phylogenetic analyses using branch-homogeneous substitution models [21, 22, 25]. But this question has not been well studied for phylogenetic estimation under branch-variable substitution models. The same question arises in the context of statistical estimation problems that are unique to branch-variable rate heterogeneity: these include shift edge inference, rooting phylogenetic trees under non-stationary models, and continuous parameter estimation (i.e. substitution rates, branch lengths, and base frequencies) for more complex substitution models. Another practical consideration is whether methodological guidance from earlier studies is applicable to inference and learning under more complex models.

Critically, new tools are needed to perform data-driven assessment of the relationship between upstream MSA estimation error and downstream phylogenetic estimation error. Such path-breaking tools promise to convert an "unknown unknown" – to paraphrase Donald Rumsfeld's infamous quote – into a quantifiable and surmountable challenge; a sufficient critical mass of evidence, as provided by such tools, can set the stage for further research progress.

In this study, we directly address both gaps. A comprehensive performance study using simulated and empirical datasets is conducted to assess the effect of multiple sequence alignment quality on phylogenetic estimation when evolution is non-homogeneous and non-stationary. We then apply RAWR ("RAndom Walk Resampling") [38], our recently introduced sequence-aware statistical resampling technique, to a new task: confidence interval estimation for phylogenetic tree reconstruction under branch-variable substitution models when unaligned biomolecular sequence data are used as input.

2 MATERIALS AND METHODS

We begin with the following notation and definitions. Let $T = (V, E)$ be a rooted tree with labeled leaves $X \subset V$ and root $\rho \in V$. An unrooted version of a tree T can be obtained by "omitting" the root ρ (i.e., deleting ρ and "connecting" its incident edges). Each edge $e = (u, v) \in E$ where $u, v \in V$ has a length $d(e)$. An edge (u, v) is a leaf edge if either u or v is a leaf, otherwise it is an internal edge. Deleting an edge e from a tree T gives two subtrees $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$. The vertex sets V_1 and V_2 are disjoint, and $V_1 \cup V_2 = V$. The same can be said for their respective leaf sets, so $\{X_1, X_2\}$ is a bipartition of X. Let this be denoted as $b(e) = \{X_1, X_2\}$.

2.1 Methods under study

Multiple sequence alignment. Our study included a range of the most commonly used and/or most accurate multiple sequence alignment methods. We aligned simulated and empirical datasets using MAFFT [16] version 7.475, MUSCLE [8] version 5.0.1428, Clustal Omega [34] version 1.2.4, Clustal W [20] version 2.1, and FSA [2] version 1.15.9. Each method was run using their respective default settings.

Phylogenetic estimation. We use the General Time Reversible (GTR) model of finite-sites nucleotide substitution for phylogenetic estimation. The GTR model is parameterized by base frequencies $\pi_T, \pi_C, \pi_A, \pi_G$ and substitution rate parameters a, b, c, d, e, f. We use the same conventions as used by [42], where a corresponds to $T \leftrightarrow C$, b corresponds to $T \leftrightarrow A$, c to $T \leftrightarrow G$, d to $C \leftrightarrow A$, e to $C \leftrightarrow G$, and f to $G \leftrightarrow A$. f is canonically fixed to 1 and the remaining free rate parameters are specified relative to f. The substitution rate matrix Q is then defined as follows:

$$Q = \begin{bmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{bmatrix}$$

with the diagonals set to $Q_{ii} = -\sum_{i \neq j} Q_{ij}$. The transition probability matrix is given by $P(t) = \exp(-Qt)$ and is used to calculate model likelihood for a phylogenetic tree. Typically in phylogenetic estimation using Markov models of substitution, the rate matrix is assumed to be constant over the whole tree. We refer to models under this assumption as homogeneous or "no-shift" models.

More general models are needed to account for substitution process variation across a phylogeny. To this end, the homogeneity

assumption can be relaxed by associating each edge e with a set of parameters $\theta(e)$ that defines the rate matrix for that edge. We use a GTR model for the branch models. The traditional homogeneous model is the case where $\theta(e)$ is fixed, i.e. $\theta(e_i) = \theta(e_j)$ for all $e_i, e_j \in E$. For heterogeneous models, we considered two different classes that we refer to as "single-shift" and "all-shift". For all-shift, $\theta(e)$ is independent for each edge. For single-shift, there are exactly two sets of parameters, θ_{shift} and $\theta_{\text{background}}$ and some restrictions on which edges they apply to. There is a shift edge, $e_{\text{shift}} \in E$, and all edges descending from it all have $\theta(e) = \theta_{\text{shift}}$. Any remaining edges are $\theta_{\text{background}}$. In nonhomogeneous models, rooting can impact likelihood values since these models are not time-reversible, so rooted trees are used. We note that the no-shift, single-shift, and all-shift models are nested in order of increasing model complexity. The all-shift model approaches the no common mechanism model, under which MLE is known to be statistically inconsistent [36].

RAxML [35] version 8.2.12 was used to perform maximum likelihood estimation under a homogeneous GTR model. PAML [43] version 4.9j was used to perform maximum likelihood estimation under fixed tree topologies using a branch model. PAML supports maximum likelihood fixed-tree-topology optimization of continuous parameters under nonhomogeneous substitution models, but does not support full tree search under nonhomogeneous models. We therefore implemented custom software to perform tree search and thereby maximize likelihood under the all-shift substitution model, where PAML was used to evaluate model likelihood of a tree topology during search. MLE under the single-shift model was a special case of all-shift model-based search, where the shift and background parameters are estimated and all tree edges in a given topology are evaluated to find a shift edge that maximizes single-shift model likelihood.

NoHTS, a new phylogenetic support estimation method. RAWR (or "RAndom Walk Resampling") [38] is a recently introduced method for sequence-aware statistical resampling of biomolecular sequence data. RAWR resampling takes the form of a random walk conducted directly on biomolecular sequences (see Supplementary Methods for pseudocode). The first application of RAWR resampling was to the task of phylogenetic support estimation under homogeneous substitution models [38]. In this study, we apply RAWR to a new task: phylogenetic support estimation under branch-variable substitution models (Figure 1). The new application naturally generalizes the original application since the branch-variable substitution models under study are a superset of traditional homogeneous substitution models. We refer to the resulting phylogenetic support estimation method as NoHTS (or "Non-Homogeneous Tree Support"). Each resampled replicate dataset consists of a set of unaligned sequences and is used to perform MSA and phylogenetic re-estimation. The resulting set of re-estimated trees is then used to calculate support for the annotation estimated tree, where the support for an edge in the annotation tree is the proportion of re-estimated trees that also display that edge. We note that other support estimation tasks are possible (e.g., support calculations for substitution process shifts along a phylogeny, tree rooting, and others), although we leave these to be explored in future research.

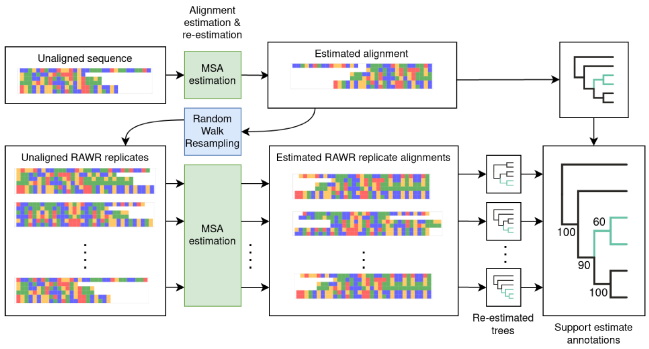

Figure 1: *Graphical overview of NoHTS (or "Non-Homogeneous Tree Support"), the new phylogenetic support estimation method.* RAWR [38] is used to perform sequence-aware resampling of an input set of biomolecular sequences. Whereas Wang et al. [38] originally applied RAWR resampling to phylogenetic support estimation under homogeneous substitution models, our new application of RAWR resampling focuses on phylogenetic support estimation under branch-variable substitution models. In the new application, each RAWR replicate consists of a set of unaligned sequences, and is used to perform MSA and phylogenetic tree re-estimation where the latter utilizes branch-variable substitution model-based MLE. The resulting re-estimates are used to calculate branch support for the annotation tree (i.e., the tree estimated on the original input dataset): the support value for a branch in the annotation tree is the proportion of re-estimated trees that also display that branch.

For phylogenetic support estimation, we compared NoHTS versus the phylogenetic bootstrap method [11] on the 10-taxon simulation study model conditions. NoHTS utilized RAWR resampling with reversal probability parameter $\gamma = 0.1$ and either 10 or 100 resampled replicates. The phylogenetic bootstrap analyses were run using 100 bootstrap replicates. All replicates were aligned using MAFFT, and alignment re-estimation was performed with MAFFT as well. Phylogenetic estimation and re-estimation were performed using single-shift maximum likelihood estimation.

Performance assessments. We use Robinson-Foulds distance [32] to assess topological difference between trees. Let $S(T) = \{b(e) | e \in E, e \text{ is an internal edge}\}$. The Robinson-Foulds distance between two unrooted trees T and T' is the symmetric difference of $S(T)$ and $S(T')$. For identifying root placement, we say two trees T and T' have identical roots, ρ and ρ' respectively, if the leaf sets of the subtrees induced by deleting the respective root nodes are identical.

We take the L1 norm of the relative errors for substitution model parameters to assess model parameter estimation performance in the simulation study. For the base frequencies, this would be $\sum_{i \in ACGT} \left| \frac{\pi_i - \hat{\pi}_i}{\pi_i} \right|$.

To assess how well the shift subtree is being predicted in the single-shift model, we use the size of the maximum agreement subtree (MAST) between the true and estimated shift subtrees. The MAST problem is to find a subtree given a set of trees $\mathscr{T}$ with the largest subset of leaves that also agrees with all the trees in $\mathscr{T}$.

For evaluating alignment quality, we use sum-of-pairs false positive and false negative proportions, denoted SP-FP and SP-FN respectively. SP-FP is calculated as the proportion of homologous nucleotide pairs in the estimated alignment and not in the true alignment. SP-FN is defined vice versa.

We performed linear regression analyses to quantify the relationship between upstream MSA estimation error and downstream phylogenetic estimation error. Python was used with the package scikit-learn [29] to perform the linear regression analyses.

For evaluating phylogenetic support estimation methods (i.e., NoHTS and the phylogenetic bootstrap method), we used precision-recall (PR) and receiver operating characteristic (ROC) curves and the area under curves (PR-AUC and ROC-AUC, respectively). Confusion matrix entries used to construct the PR and ROC curves were calculated using the same procedures as in [38].

To quantify the relative importance of each factor in the simulation study, we utilized the random forest approach in the study of Lanier and Knowles [19]. Likewise, we performed random forest analysis in R using the package randomForest [6]. We fit 1000 regression trees with the following factors: model condition (indel probability and model tree height), MLE method, alignment type (true or estimated), and number of taxa. Relative importance of a factor was measured as the increase in mean squared error (MSE) when excluding that factor over the highest increase in MSE.

2.2 Simulated datasets

Supplementary Figure S1 provides a graphical overview of our study's simulation procedures.

Model tree generation. Model trees were sampled using INDELible [12] under a random birth-death process. Non-ultrametricity was introduced using the procedure described in [27] with deviation factor $c = 2$. First, a rooted model tree is generated using INDELible under a birth-death process. For every branch, sample x from a uniform distribution $x \sim U(-\ln(2), \ln(2))$ and scale the branch length by x. The tree height is then rescaled so that the maximum root-to-tip distance is the height specified by the model condition. Finally, a subtree containing as close to half of the leaves is selected to evolve under the shift substitution model.

Simulating sequence evolution. Our simulations utilized model conditions that were based on the study of Wang et al. [38] (Table 2). Following the rationale of Wang et al. [38], the model condition parameter settings reflect a range of evolutionary divergence that are often encountered in modern-day phylogenetic systematics and related research topics. To simulate sequence evolution on the model trees, INDELible was used to perform finite-sites simulations under a GTR-based branch model and the indel model of [12]. Based on Nabholz et al. [26]'s report of GC content variation in the avian phylogeny, the GTR model parameters were empirically estimated using single-copy orthologs from [15] for the subset of species (*Calypte anna*, *Alligator mississippiensis*, *Melopsittacus undulatus*, *Corvus brachyrhynchos*, and *Manacus vitellinus*) included in Nabholz et al. [26]'s study. To estimate these parameters, we aligned the single-copy orthologs using MAFFT with the default settings. MLE under the single-shift model was used to estimate parameters on each individual aligned sequence. Then, we looked at the two sets of estimated substitution rates, and we observed that the ratio between them was bimodal. We chose GTR model parameters based on the estimated parameters in the lower rate mode (Table 1). The simulation outputs consisted of a true multiple sequence alignment, the corresponding set of unaligned sequences, a model tree with branch lengths, and the true substitution model instance. Table 3 lists summary statistics for true and estimated MSAs.

Table 1: *Simulation study: GTR model parameters used for the single-shift substitution model in our simulations.* The single-shift substitution model is comprised of two sets of model parameters: one corresponding to a "background" substitution process, and the other to a "shift" substitution process. Settings for each set of base frequency parameters $\pi_T, \pi_C, \pi_A, \pi_G$ are listed, followed by the settings for each set of substitution rate parameters a, b, c, d, e, f. (See Methods section for details.)

Parameter	π_T	π_C	π_A	π_G	C↔T	A↔T	G↔T	A↔C	C↔G	A↔G
Shift	0.216	0.237	0.317	0.230	5.847	3.186	1.214	3.437	1.307	1.0
Background	0.183	0.226	0.058	0.534	1.505	0.367	0.141	0.412	0.094	1.0

Table 2: *Simulation study model conditions.* The 10-taxon model conditions are named 10.A through 10.E in order of generally increasing evolutionary divergence, and the 20-taxon model conditions are named 20.A through 20.E similarly. As noted in the Methods section, model condition parameter settings were based on the study of Wang et al. [38] and reflect a range of evolutionary divergence. The model tree height and indel model parameter are listed for each model condition. Each model condition consists of settings for these two parameters and the single-shift substitution model parameters. (See Methods section for details.)

Model condition	Number of taxa	Tree height	Indel probability
10.A	10	0.47	0.13
10.B	10	0.7	0.1
10.C	10	1.2	0.06
10.D	10	2	0.031
10.E	10	4.4	0.013
20.A	20	0.47	0.13
20.B	20	0.7	0.1
20.C	20	1.2	0.06
20.D	20	2	0.031
20.E	20	4.4	0.013

Experimental replication. For each model condition, the simulation procedure was repeated to obtain 30 experimental replicates. Results are reported on average (along with standard errors) across all experimental replicates in each model condition.

2.3 Empirical datasets

Flowering monocot dataset. The distribution of GC content in the Poales, an order of flowering monocots, is bimodal [5]. This pattern is notably strong in rice. We applied nonhomogeneous substitution model-based phylogenetic tree estimation to a set of 7 taxa from the Poales – *Oryza sativa japonica* [28], *Sorghum bicolor* [23], *Carex cristatella*, *C. scoparia*, *Juncus effusus*, *Juncus inflexus* [30], and *Ananas comosus* [24] – and one additional taxon from the order Zingiberales – *Musa balbisiana* [39]. We identified 1900 single-copy orthologs using OrthoFinder [10] with default settings. Average sequence length across all taxa and single-copy orthologs was 1377.

Table 3: *Simulation study: summary statistics for ground truth and estimated MSAs.* MAFFT, Muscle, Clustal W, Clustal Omega, and FSA were used to estimated MSAs on each simulation study dataset. For each model condition, each MSA method's average alignment SP-FN and SP-FP error ("SP-FN" and "SP-FP", respectively) are reported across all replicate datasets ($n = 30$). Alignment length ("length"), average normalized Hamming distance ("ANHD") for aligned sequence pairs, and proportion of MSA cells that consist of indels ("Gappiness") are reported for each method as an average across all replicates in a model condition ($n = 30$).

Statistic	Alignment	Model condition									
		10.A	10.B	10.C	10.D	10.E	20.A	20.B	20.C	20.D	20.E
SP-FN	MAFFT	0.531	0.667	0.751	0.831	0.890	0.392	0.539	0.775	0.873	0.948
	MUSCLE	0.526	0.637	0.716	0.788	0.851	0.355	0.480	0.702	0.812	0.908
	CLUSTALW	0.715	0.765	0.806	0.855	0.892	0.630	0.736	0.846	0.890	0.943
	CLUSTALO	0.710	0.769	0.813	0.854	0.884	0.640	0.728	0.843	0.889	0.937
	FSA	0.680	0.751	0.820	0.886	0.927	0.550	0.706	0.864	0.923	0.961
SP-FP	MAFFT	0.526	0.663	0.750	0.833	0.893	0.364	0.519	0.770	0.872	0.949
	MUSCLE	0.513	0.633	0.715	0.792	0.858	0.331	0.465	0.700	0.814	0.913
	CLUSTALW	0.702	0.759	0.803	0.856	0.896	0.595	0.715	0.843	0.891	0.945
	CLUSTALO	0.667	0.741	0.795	0.847	0.886	0.566	0.679	0.827	0.884	0.939
	FSA	0.394	0.501	0.599	0.695	0.763	0.181	0.313	0.535	0.529	0.782
Length	TRUE	2123.8	2315.8	2315.8	2313.2	2063.0	2410.3	2585.1	2895.8	2696.2	2723.6
	MAFFT	1478.6	1477.1	1484.5	1461.8	1529.2	1643.7	1670.7	1683.0	1691.4	1804.3
	MUSCLE	1518.5	1570.1	1573.1	1561.9	1590.5	1790.2	1863.0	1907.7	1890.0	1979.5
	CLUSTALW	1191.4	1186.0	1170.6	1143.0	1146.5	1278.6	1261.2	1227.8	1181.6	1162.8
	CLUSTALO	1247.1	1248.7	1237.6	1222.9	1234.0	1306.3	1318.2	1303.2	1287.4	1283.8
	FSA	3609.3	4471.2	4992.4	5729.6	6196.2	4394.5	5959.0	8231.8	10177.9	11566.5
ANHD	TRUE	0.306	0.364	0.364	0.465	0.649	0.301	0.374	0.484	0.581	0.667
	MAFFT	0.389	0.435	0.484	0.528	0.569	0.368	0.433	0.516	0.569	0.608
	MUSCLE	0.413	0.449	0.499	0.542	0.589	0.380	0.445	0.530	0.585	0.630
	CLUSTALW	0.466	0.496	0.537	0.573	0.614	0.449	0.509	0.573	0.615	0.652
	CLUSTALO	0.484	0.504	0.541	0.565	0.602	0.471	0.518	0.570	0.602	0.637
	FSA	0.280	0.319	0.377	0.420	0.477	0.289	0.349	0.429	0.486	0.525
Gappiness	TRUE	0.528	0.564	0.564	0.566	0.510	0.581	0.606	0.649	0.622	0.629
	MAFFT	0.326	0.321	0.328	0.310	0.342	0.390	0.394	0.400	0.403	0.444
	MUSCLE	0.342	0.361	0.364	0.353	0.366	0.439	0.456	0.470	0.466	0.493
	CLUSTALW	0.165	0.155	0.148	0.119	0.124	0.216	0.198	0.178	0.148	0.140
	CLUSTALO	0.202	0.198	0.194	0.177	0.187	0.233	0.232	0.226	0.218	0.221
	FSA	0.715	0.770	0.795	0.820	0.836	0.763	0.819	0.874	0.899	0.913

We aligned the sequences individually using MAFFT, MUSCLE, Clustal Omega, Clustal W, and FSA using the same settings as those used in the simulation study. We performed phylogenetic estimation under a single-shift model for every individual gene.

2.4 Data availability statement

Data and scripts used are available under an open copyleft license at https://gitlab.msu.edu/liulab/nonhomogeneous-substitution-model-study-data-scripts.

3 RESULTS

3.1 Simulation study

MSA estimation error and topological error of single-shift MLE. We focus first on phylogenetic reconstruction with no model misspecification. On the least divergent 10.A model condition, MLE under the single-shift model returned the lowest average topological error when the true MSA was provided as input, followed by estimated MSAs excluding FSA (i.e., ClustalOmega, MAFFT, MUSCLE, and ClustalW), and FSA-estimated MSAs returned the worst accuracy overall. As evolutionary divergence increased across the 10-taxon model conditions – with 10.A being the least divergent and 10.E the most divergent – topological error returned by the different methods also increased to differing extents. The smallest effect was seen on true alignments; a relatively stronger effect seen on estimated alignments, with FSA exhibiting the greatest effect among all MSA estimation methods. Across the 10-taxon model conditions, MLE(TrueAln), MLE(ClustalOmega), MLE(MAFFT), MLE(Muscle), MLE(ClustalW, and MLE(FSA) returned average topological error of 0.116, 0.232, 0.245, 0.277, 0.253, and 0.386, respectively; the corresponding estimated MSAs had average SP-FN (SP-FP) error of 0.778 (0.816), 0.706 (0.743), 0.677 (0.723), 0.776 (0.816), and 0.782 (0.321), respectively.

All MSA estimation methods returned some degree of alignment error, and comparison of phylogenetic MLE using estimated MSAs vs. true MSAs demonstrates the clear impact of upstream estimation error on downstream phylogenetic reconstruction. However, relative comparisons among MSA methods were not a perfect predictor of resulting topological error, at least at the coarse granularity of per-model-condition averages. Muscle and MAFFT were generally among the more accurate MSA methods in terms of SP-FN and SP-FP error – with Muscle outperforming MAFFT slightly, FSA consistently returned lower SP-FP error compared to all other methods, and ClustalW-estimated alignments were generally least accurate. We note that our findings are consistent with other related studies involving traditional homogeneous model-based MLE [21, 22].

We performed linear regression analyses to examine the relationship between upstream MSA error and downstream topological error at a finer granularity. Per-replicate scatterplots and fitted linear regression models are shown in Figure 3. The correlation between alignment SP-FN error and topological error was observed to be positive and statistically significant across all model conditions. A similar outcome was observed between alignment SP-FP error and topological error, except on 2 of the least divergent 10-taxon model conditions (Supplementary Figure S4 in the SOM Appendix). Correlation coefficients tended to increase as evolutionary divergence increased, with the strongest associations observed on the most divergent model conditions.

Similar outcomes were observed on the 20-taxon model conditions. Topological error on the 20-taxon model conditions were somewhat higher than on the 10-taxon model conditions, as expected due to the combinatorially larger solution spaces required by the former compared to the latter. Across the 20-taxon model conditions, MLE(TrueAln), MLE(ClustalOmega), MLE(MAFFT), MLE(Muscle), MLE(ClustalW, and MLE(FSA) returned average topological error of 0.121, 0.329, 0.352, 0.310, 0.325, and 0.408, respectively; the corresponding MSAs had average SP-FN (SP-FP) error of 0.792 (0.814), 0.696 (0.712), 0.643 (0.675), 0.791 (0.822), and 0.778 (0.255), respectively. (Per-method regression analyses also yielded consistent results, as shown in Supplementary Figures S11 through S15.)

MSA estimation error and topological error: role of model mis-specification. Across the different model conditions in our study, single-shift MLE using a given MSA as input (i.e., the true MSA or one of the estimated MSAs) returned the best (or among the best) average topological error, compared to the zero-shift and all-shift MLE methods (Figure 2). In fact, on the most divergent 10.E and 20.E model conditions, zero-shift model-based MLE on the true MSA returned topological error that was comparable to estimated MSA-based MLE under any of the substitution models. For MLE on the true, MUSCLE, Clustal Omega, ClustalW, MAFFT, and FSA alignments, the respective average topological error for single-shift phylogenetic estimation was 0.094, 0.258, 0.261, 0.269, 0.289, and 0.380; the respective average topological error for all-shift phylogenetic estimation was 0.120,

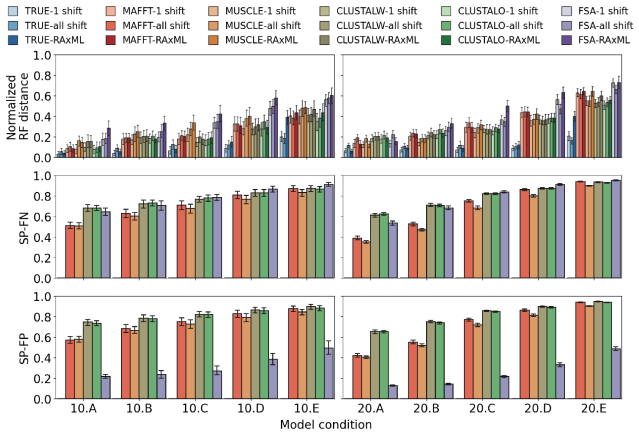

Figure 2: *Simulation study: estimated MSA and tree error returned by each method on each model condition.* Phylogenetic trees were estimated using the true and estimated MSAs as input, where the latter were estimated using a range of different MSA methods: MAFFT, MUSCLE, Clustal W, Clustal Omega, and FSA. Phylogenetic trees were estimated using MLE under one of three nested models: no-shift, single-shift, and all-shift. Topological error was measured using normalized Robinson-Foulds distance between a model tree and estimated tree. An estimated MSA was compared against the true MSA based on alignment SP-FN and SP-FP error. For each model condition and method, average and standard error of each performance assessment are reported ($n = 30$).

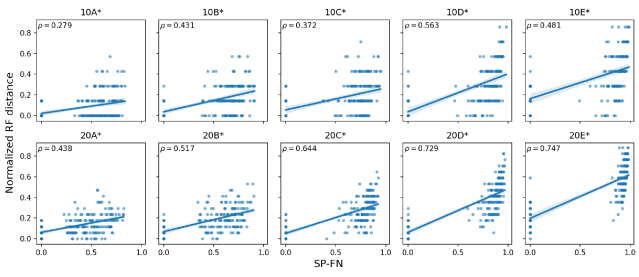

Figure 3: *Simulation study: the relationship between upstream MSA estimation error and downstream phylogenetic estimation error.* The former was assessed based on sum-of-pairs false negative rate, and the latter was assessed based on normalized Robinson-Foulds distance. We performed linear regression analyses to quantify the relationship between the two assessments on each model condition. The fitted linear regression model is shown as a blue line along with a confidence interval shown in a light blue shade. MSA and tree estimation methods are as described in Figure 2. Results are aggregated for all alignment types ($n = 180$).

0.312, 0.292, 0.290, 0.306, and 0.386; finally, the respective average topological error for zero-shift phylogenetic estimation was 0.149, 0.333, 0.286, 0.301, 0.297, and 0.445. On both true and estimated MSAs, the difference in average topological error between MLE under single-shift vs. all-shift models was smaller than single-shift vs. zero-shift comparison, with the exception of Clustal Omega on the 10-taxon model conditions and MAFFT and MUSCLE on the 20-taxon model conditions. We hypothesize that extraneous model parameters and possible overfitting is not appreciably degrading topological accuracy of phylogenetic reconstruction. Finally, the impact of model mis-specification on resulting topological error was less apparent on estimated MSAs than on the true MSA.

In our random forest analysis, relative importance of the evolutionary divergence, alignment type, number of taxa, and maximum likelihood method used were 1.0, 0.48, 0.09, and 0.03 for predicting normalized RF-distance.

Performance evaluation of NoHTS versus bootstrap support estimates. For phylogenetic support estimation, NoHTS with both 100 and 10 resampled replicates consistently yielded an improvement over the phylogenetic bootstrap method based on both PR-AUC and ROC-AUC assessments (Table 4). Model conditions with increased evolutionary divergence, and consequently increased MSA and phylogenetic tree estimation error, trended towards lower PR-AUC and ROC-AUC. However, both methods were not equally affected by evolutionary divergence. Based on AUC assessments, the performance advantage returned by NoHTS over the phylogenetic bootstrap method grew larger as model conditions became more divergent, suggesting that NoHTS is especially well suited to sequence inputs that pose a greater challenge in terms of MSA and phylogenetic estimation. This pattern is also visible in the curves (Supplementary Figure S16). Additional simulation experiments indicated that NoHTS was largely robust to the effects of varying biomolecular sequence length (Supplementary Figure S8).

MSA estimation error and phylogenetic tree rooting. Overall, single-shift-model-based phylogenetic estimation on true MSAs returned the most accurate rooting (Table 5). Among estimated MSAs, MAFFT and Muscle analyses under the single-shift model returned the most accurate rooting, followed by Clustal Omega and Clustal W; FSA returned the least accurate rooting. As evolutionary divergence increased across the 10-taxon model conditions, rooting error generally increased as well; smaller increases were seen on true MSAs and comparatively larger increases were seen on estimated MSAs, with the largest increases seen on FSA alignments. A similar effect concerning evolutionary divergence was seen on the 20-taxon model conditions.

Table 4: *Simulation study: AUC for the precision-recall and ROC curves of NoHTS versus bootstrap support estimation methods on the 10-taxon model conditions.* For the bootstrap method, 100 resampled replicates were utilized for each analysis. For NoHTS, either 100 or 10 resampled replicates were utilized for each analysis.

	PR-AUC			ROC-AUC		
Model condition	Bootstrap	NoHTS		Bootstrap	NoHTS	
	100 reps	10 reps	100 reps	100 reps	10 reps	100 reps
10.A	0.988	0.992	0.996	0.875	0.916	0.957
10.B	0.972	0.984	0.988	0.876	0.926	0.944
10.C	0.945	0.978	0.983	0.748	0.899	0.926
10.D	0.889	0.978	0.977	0.749	0.950	0.947
10.E	0.874	0.945	0.957	0.791	0.910	0.922

Table 5: *Simulation study: proportion of correct root placements by MSA, MLE method, and model condition.* Results are reported as an average across all replicates in each model condition ($n = 30$).

Model condition	MLE method	Correct root rate					
		TRUE	MAFFT	MUSCLE	CLUSTALW	CLUSTALO	FSA
10A	single shift	56.7%	23.3%	43.3%	26.7%	26.7%	23.3%
	all shift	23.3%	20.0%	20.0%	23.3%	6.7%	20.0%
10B	single shift	60.0%	16.7%	23.3%	13.3%	6.7%	20.0%
	all shift	30.0%	6.7%	10.0%	3.3%	6.7%	0.0%
10C	single shift	60.0%	16.7%	16.7%	10.0%	10.0%	16.7%
	all shift	30.0%	13.3%	16.7%	3.3%	0.0%	6.7%
10D	single shift	40.0%	20.0%	26.7%	6.7%	16.7%	6.7%
	all shift	23.3%	3.3%	16.7%	0.0%	6.7%	3.3%
10E	single shift	30.0%	10.0%	10.0%	6.7%	10.0%	3.3%
	all shift	6.7%	3.3%	3.3%	3.3%	6.7%	0.0%
20A	single shift	33.3%	20.0%	30.0%	13.3%	16.7%	36.7%
	all shift	31.0%	6.9%	7.1%	6.7%	16.7%	17.2%
20B	single shift	37.9%	17.2%	20.7%	6.9%	10.3%	24.1%
	all shift	24.1%	13.8%	10.3%	0.0%	7.1%	10.3%
20C	single shift	46.7%	20.0%	13.3%	13.3%	3.3%	0.0%
	all shift	13.3%	6.7%	0.0%	6.7%	0.0%	3.3%
20D	single shift	50.0%	13.3%	6.7%	13.3%	6.7%	3.3%
	all shift	10.3%	3.3%	0.0%	3.3%	0.0%	0.0%
20E	single shift	46.7%	16.7%	3.3%	6.7%	6.7%	0.0%
	all shift	6.7%	0.0%	3.3%	3.3%	0.0%	0.0%

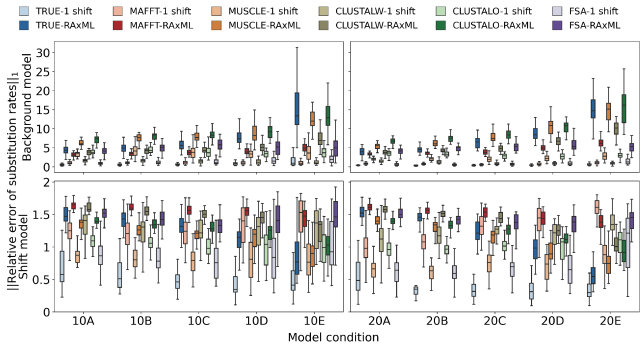

Figure 4: *Simulation study: substitution rate estimation error returned by the methods under study.* Substitution rate estimation error is measured by L1-norm of the relative errors for rate parameter estimates. Results for all replicates in a model condition are reported as a box-and-whisker plot ($n = 30$).

Model mis-specification also served to increase rooting error. Still, the relative comparison among different MSAs was largely similar under the all-shift model, as compared to the single-shift model.

For each model and method on each simulation condition, rooting error is generally higher than overall topological error. This finding suggests that rooting is more difficult than estimation of other aspects of topological reconstruction, which is consistent with experimental outcomes for traditional approaches to phylogenetic rooting (e.g., outgroup rooting, midpoint rooting, etc.) and phylogenetic estimation under homogeneous substitution models (cf. section 2.2.6 in [41]).

MSA estimation error and continuous parameter estimation. Looking beyond topological error, we next examined the consequences of upstream MSA estimation error on downstream estimation of substitution rates. Results for MLE under the single-shift model are shown in Figure 4. Single-shift MLE on the true MSA consistently returned the lowest average error across the 10- and 20-taxon model conditions, compared to single-shift MLE on estimated MSAs. We found that single-shift MLE on FSA-estimated MSAs was among the most accurate relative to the other estimated MSA-based methods. We also found that the other estimated MSA-based methods (excluding single-shift MLE on FSA-estimated MSAs) often returned relatively high substitution rate estimation error on background edges and lower substitution rate estimation error on shift edges, or vice versa, on each model condition. Model mis-specification also played a role. Compared to single-shift MLE, zero-shift MLE returned higher substitution rate estimation error (Figure 4), and all-shift MLE had higher error still (Supplementary Figure S4). The differences were stark – amounting to as much as an order of magnitude or more when comparing single-shift vs. all-shift MLE. Relatively high rate estimation errors were seen under mis-specified zero-shift and all-shift models, and the role of MSA quality became more difficult to discern.

Results for base frequency estimation are reported for each model condition and method in Figure 5. Similar to substitution rate estimation outcomes, single-shift MLE returned the most accurate base frequency estimates on true MSAs, compared to estimated MSAs. Single-shift MLE on FSA MSAs was typically least accurate, and single-shift MLE on the other estimated MSAs were intermediate in accuracy. The error difference between neighboring rankings became more pronounced as evolutionary divergence increased, whereas the different methods returned generally comparable base frequency estimation error on the least divergent model conditions. Also mirroring the substitution rate estimation outcomes, model mis-specification tended to inflate base frequency estimation error. Compared to single-shift MLE, zero- and all-shift MLE returned higher error by as much as several factors. Finally, all-shift MLE consistently returned higher error variance compared to zero- and single-shift MLE.

We also assessed branch length estimation error in our experiments. Leaf branch lengths were consistently over-estimated across our study, regardless of substitution model and input MSA (Supplementary Figure S5 in SOM Appendix). A more comprehensive examination of both leaf and internal branch lengths would be more revealing; while the branch-score metric used was proposed for this purpose [17], assessing internal branch lengths in isolation from other factors – particularly topological error – remains a challenge.

MSA error and shift edge estimation. For all single-shift methods, we evaluated shift clade accuracy based on maximal MAST size. Shift clade estimation was most accurate using MLE on the true MSA (Figure 6). Shift clade estimation using estimated MSAs was less accurate, with FSA MSAs resulting in comparable or lower accuracy versus the other MSA methods. Greater evolutionary divergence served to enhance these relative rankings, with the greatest differences between methods occurring on the most divergent 10- and 20-taxon model conditions.

3.2 Empirical study

Gene trees were reconstructed on each locus using the MSA and phylogenetic MLE methods under study. Per-locus topological disagreement among the different methods averaged ~ 5% or so based on traditional R-F distance and ~ 20% for the rooted version of

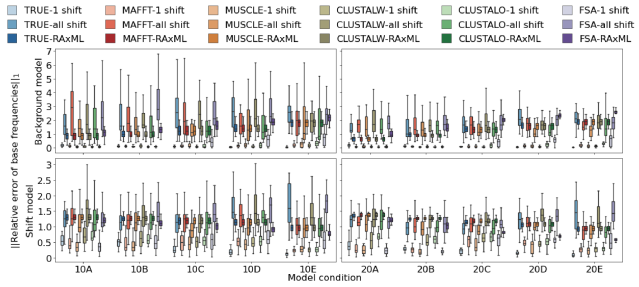

Figure 5: *Simulation study: base frequency estimation error returned by the methods under study.* Base frequency estimation error is measured by the L1 norm of the relative errors for base frequency parameter estimates. Results for all replicates in a model condition are reported as a box-and-whisker plot ($n = 30$).

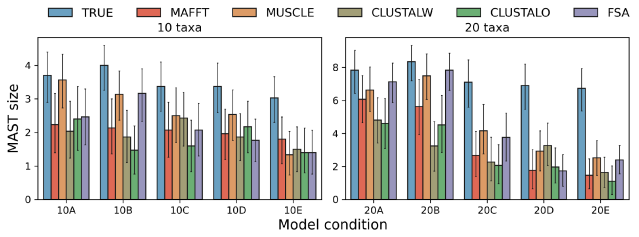

Figure 6: *Simulation study: shift edge assessment of single-shift MLE.* Shift edge accuracy is assessed based on the size of the MAST of the predicted shift subtree and the true subtree. Average and standard error bars are shown for all replicates in each model condition ($n = 30$).

R-F distance. On average across loci, the largest topological disagreements were observed among phylogenies reconstructed using ClustalW-estimated MSAs versus any other estimated MSA; all other pairwise comparisons returned relatively less topological disagreement (Table 6).

For continuous parameter estimates, the median absolute difference of estimated base frequencies under the background vs. shift models was in agreement across different single-shift model-based analyses and loci; observed differences were at most 0.10, 0.12, 0.11, and 0.10, respectively (Supplementary Figure S18). A similar outcome was observed for estimated substitution rates, with observed differences of at most 0.41. On the other hand, a relatively large amount of variation around the median was observed across loci for these estimates.

4 DISCUSSION

Simulation study. We observed a consistent impact of upstream MSA error on topological error of downstream phylogenetic MLE under variable-across-phylogeny substitution models. Regression analyses consistently returned positive and significant correlation between upstream and downstream estimation error across all of the model conditions in our simulation study. Upstream MSA error was also observed to have major consequences for continuous

Table 6: *Empirical study: topological disagreement among gene trees estimated using different MSA methods and single-shift MLE.* For each gene, topological discordance was calculated based on pairwise Robinson-Foulds distance between estimated gene trees returned by different methods. Average ("Avg") and standard error ("SE") are reported across all genes ($n = 1377$).

	Muscle		Clustal Omega		Clustal W		FSA	
nRF	Avg	SE	Avg	SE	Avg	SE	Avg	SE
MAFFT	0.044	0.088	0.052	0.097	0.059	0.106	0.041	0.087
Muscle			0.053	0.100	0.058	0.108	0.044	0.090
Clustal Omega					0.059	0.105	0.052	0.095
Clustal W							0.061	0.103

parameter estimation. The outcomes observed for discrete phylogenetic relationships were qualitatively different from those of continuous parameter estimates, which suggests that the latter is more difficult to reconstruct versus the former. Our random forest analysis attributed MSA type to have the second highest relative contribution to both topological error and branch length error, after evolutionary divergence.

We found that NoHTS support estimates consistently performed better than traditional bootstrap support estimates, even with as few as 10 resampled replicates. We attribute this performance advantage to a key difference between the two methods: NoHTS support estimation accounts for both MSA and tree estimation uncertainty, while bootstrap support only accounts for tree estimation uncertainty. The former is possible due to RAWR's ability to preserve sequential ordering information of resampled sites, which is necessary for meaningful MSA re-estimation. The experimental outcome in the NoHTS experiment reinforces our study finding concerning the impact of upstream MSA error on downstream phylogenetic reconstruction, and new computational methods like NoHTS will allow direct quantification of these effects.

Phylogenetic estimation using a branch model matching the number of shifts during sequence evolution returned the best estimation accuracy, as expected. Furthermore, model mis-specification effects were attenuated when estimated MSAs were used as input rather than true MSAs. We attribute the difference to the higher topological error observed on the former versus the latter, which may swamp effects of other methodological factors like model mis-specification. Interestingly, phylogenetic inference under the all-shift model returned topological error that was comparable to the no-shift model for topology estimation, except on the 20.E model condition, where it performs slightly better than the no-shift model. Overall, the impact of MSA error on downstream topological error was qualitatively larger than that of substitution model mis-specification.

Generally speaking, evolutionary divergence served to amplify upstream and downstream estimation error and their association. Increasing dataset size had similar effects, which we anticipate will become even more pronounced on datasets with hundreds or thousands of taxa.

Empirical study. Different MSA methods with varying MSA estimation error resulted in observable differences among downstream

single-shift MLE phylogenetic estimates. The genome-wide distributions of substitution rate estimates were also variable between alignment methods. We caution that direct comparison between the simulation study and empirical genomic sequence analyses are complicated by differences in data type: the former involves single-locus data and all loci were simulated i.i.d., whereas the latter utilized multi-locus data where different loci may well evolved under more complex and non-i.i.d. evolutionary processes (e.g., genetic recombination). A new generation of species-tree-aware phylogenomic inference methods can help better account for complex evolution of multi-locus biomolecular sequences [3, 9], but we anticipate that the influence of upstream MSA estimation error will become even greater under more complex evolutionary models.

5 CONCLUSIONS

Our performance study assessed the impact of upstream MSA estimation error on downstream phylogenetic MLE under branch-heterogeneous substitution models. Throughout the simulation study, MSA estimation error was consistently and significantly associated with downstream topological error for single-shift MLE. Greater correlations were observed as evolutionary divergence increased across model conditions. We introduced NoHTS, a random walk resampling method that assesses phylogenetic support for branch-variable model-based estimation using unaligned biomolecular sequence inputs. For all simulated model conditions, we found that NoHTS support estimation performed better for phylogenetic support estimation, compared to the phylogenetic bootstrap method. Estimation of continuous parameters, tree rooting, and shift edge were also impacted by MSA estimation error. Additional MLE experiments with under- and over-parameterized models indicate that our study findings are robust to model mis-specification. An empirical study of the order Poales revealed findings consistent with the simulation study experiments.

Several recommendations follow from our study's findings. First, NoHTS can and should be used to directly quantify the effects of upstream MSA estimation error on downstream phylogenetic MLE under branch-variable substitution models. We also recommend that NoHTS-estimated phylogenetic confidence intervals be reported alongside any phylogenies that are reconstructed using heterogeneous model-based MLE. Second, our study reinforces a through-line in computational phylogenetics: there is a great need for alignment-aware phylogenetic inference and learning – both under homogeneous [21, 22, 25] and branch-heterogeneous substitution models. Ideally, an MSA and phylogenetic tree would be co-estimated under the same evolutionary model that generated sequence observations. Such co-estimation methods exist for homogeneous models [22, 25, 31] but not for branch-heterogeneous models. In the interim, mitigation via best practices when reconstructing species trees offers a stop-gap workaround. We recommend that species trees be reconstructed using multi-locus data and the latest statistical methods for phylogenomic inference and learning, such as ASTRAL [44]. We caution that this mitigation measure is no silver bullet. More deliberate decision making is needed during upstream phylogenetic study and analysis design [3, 9]. For example, denser taxon sampling may help ameliorate long branch attraction [13], but our experiments suggest that the impact of upstream MSA error on downstream phylogenetic inference becomes more pronounced as both dataset sizes and divergence increase.

We conclude with thoughts on future work. Our study examined the relationship between upstream MSA estimation error and downstream phylogenetic reconstruction where exactly one substitution process shift occurred along a phylogeny. Even stronger effects are anticipated where multiple evolutionary shifts occur along a phylogeny, as is expected to be the case where larger and more divergent clades are sampled within Tree of Life. This hypothesis merits future study with an expanded set of multiple-shift simulations. Finally, a fundamental issue is model mis-specification due to the use of substitution-only model-based analysis of aligned biomolecular sequence data. This simplifying assumption is pervasive throughout phylogenetics and phylogenomics. Phylogenetic MLE under traditional homogeneous substitution models is known to be statistically inconsistent where sequences evolved under insertion and deletion processes [40]; a similar theoretical limitation is also expected for heterogeneous substitution models. As noted above, statistical co-estimation methods are needed to reconstruct MSAs and phylogenetic trees under a variable-across-phylogeny model of substitutions, insertions, and deletions.

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their detailed and constructive feedback. We also thank Kevin Childs and Yu-ya Liang for help with the flowering monocot dataset. This work is supported in part by the National Science Foundation (2144121 and 1740874 to KJL), the National Science Foundation Research Traineeship Program (DGE-1828149) to RD, and the Institute for Cyber-Enabled Research at Michigan State University.

REFERENCES

[1] Bastien Boussau and Manolo Gouy. 2006. Efficient Likelihood Computations with Nonreversible Models of Evolution. *Systematic Biology* 55, 5 (Oct. 2006), 756–768.
[2] Robert K. Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. 2009. Fast Statistical Alignment. *PLOS Computational Biology* 5, 5 (May 2009), e1000392.
[3] Gustavo A. Bravo, Alexandre Antonelli, Christine D. Bacon, Krzysztof Bartoszek, Mozes P. K. Blom, Stella Huynh, Graham Jones, L. Lacey Knowles, Sangeet Lamichhaney, Thomas Marcussen, Hélène Morlon, Luay K. Nakhleh, Bengt Oxelman, Bernard Pfeil, Alexander Schliep, Niklas Wahlberg, Fernanda P. Werneck, John Wiedenhoeft, Sandi Willows-Munro, and Scott V. Edwards. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ* 7 (2019), e6399.
[4] Stephen L. Cameron. 2014. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny. *Annual Review of Entomology* 59, 1 (2014), 95–117.
[5] Yves Clément, Margaux-Alison Fustier, Benoit Nabholz, and Sylvain Glémin. 2015. The Bimodal Distribution of Genic GC Content Is Ancestral to Monocot Species. *Genome Biology and Evolution* 7, 1 (Jan. 2015), 336–348.
[6] 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.
[7] Alexei J. Drummond and Andrew Rambaut. 2007. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolutionary Biology* 7, 1 (Dec. 2007), 1–8.
[8] Robert C. Edgar. 2004. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32, 5 (March 2004), 1792–1797.
[9] Scott V Edwards. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1 (2009), 1–19.
[10] David M. Emms and Steven Kelly. 2019. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biology* 20, 1 (Nov. 2019), 238.
[11] Joseph Felsenstein. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 4 (July 1985), 783–791.
[12] William Fletcher and Ziheng Yang. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution* 26, 8 (Aug. 2009),

1879–1888.
[13] David M Hillis. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* 47, 1 (1998), 3–8.
[14] John P. Huelsenbeck. 2002. Testing a Covariotide Model of DNA Substitution. *Molecular Biology and Evolution* 19, 5 (May 2002), 698–707.
[15] Erich D. Jarvis, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, Brant C. Faircloth, Benoit Nabholz, Jason T. Howard, Alexander Suh, Claudia C. Weber, Rute R. da Fonseca, Jianwen Li, Fang Zhang, Hui Li, Long Zhou, Nitish Narula, Liang Liu, Ganesh Ganapathy, Bastien Boussau, Md. Shamsuzzoha Bayzid, Volodymyr Zavidovych, Sankar Subramanian, Toni Gabaldón, Salvador Capella-Gutiérrez, Jaime Huerta-Cepas, Bhanu Rekepalli, Kasper Munch, Mikkel Schierup, Bent Lindow, Wesley C. Warren, David Ray, Richard E. Green, Michael W. Bruford, Xiangjiang Zhan, Andrew Dixon, Shengbin Li, Ning Li, Yinhua Huang, Elizabeth P. Derryberry, Mads Frost Bertelsen, Frederick H. Sheldon, Robb T. Brumfield, Claudio V. Mello, Peter V. Lovell, Morgan Wirthlin, Maria Paula Cruz Schneider, Francisco Prosdocimi, José Alfredo Samaniego, Amhed Missael Vargas Velazquez, Alonzo Alfaro-Núñez, Paula F. Campos, Bent Petersen, Thomas Sicheritz-Ponten, An Pas, Tom Bailey, Paul Scofield, Michael Bunce, David M. Lambert, Qi Zhou, Polina Perelman, Amy C. Driskell, Beth Shapiro, Zijun Xiong, Yongli Zeng, Shiping Liu, Zhenyu Li, Binghang Liu, Kui Wu, Jin Xiao, Xiong Yinqi, Qiuemei Zheng, Yong Zhang, Huanming Yang, Jian Wang, Linnea Smeds, Frank E. Rheindt, Michael Braun, Jon Fjeldsa, Ludovic Orlando, F. Keith Barker, Knud Andreas Jønsson, Warren Johnson, Klaus-Peter Koepfli, Stephen O'Brien, David Haussler, Oliver A. Ryder, Carsten Rahbek, Eske Willerslev, Gary R. Graves, Travis C. Glenn, John McCormack, Dave Burt, Hans Ellegren, Per Alström, Scott V. Edwards, Alexandros Stamatakis, David P. Mindell, Joel Cracraft, Edward L. Braun, Tandy Warnow, Wang Jun, M. Thomas P. Gilbert, and Guojie Zhang. 2014. Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds. *Science* 346, 6215 (Dec. 2014), 1320–1331.
[16] 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 (April 2013), 772–780.
[17] M K Kuhner and J Felsenstein. 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11, 3 (May 1994), 459–468.
[18] Abigail L Labella, Dana A Opulente, Jacob L Steenwyk, Chris Todd Hittinger, and Antonis Rokas. 2019. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genetics* 15, 7 (2019), e1008304.
[19] Hayley C. Lanier and L. Lacey Knowles. 2012. Is Recombination a Problem for Species-Tree Analyses? *Systematic Biology* 61, 4 (July 2012), 691–701.
[20] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X Version 2.0. *Bioinformatics* 23, 21 (Nov. 2007), 2947–2948.
[21] Kevin Liu, Serita Nelesen, Sindhu Raghavan, C. Randal Linder, and Tandy Warnow. 2009. Barking Up The Wrong Treelength: The Impact of Gap Penalty on Alignment and Tree Accuracy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6, 1 (Jan. 2009), 7–21.
[22] 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 (Jan. 2012), 90–90.
[23] Ryan F. McCormick, Sandra K. Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock D. Weers, Brian McKinley, Ashley Mattison, Daryl T. Morishige, Jane Grimwood, Jeremy Schmutz, and John E. Mullet. 2018. The Sorghum Bicolor Reference Genome: Improved Assembly, Gene Annotations, a Transcriptome Atlas, and Signatures of Genome Organization. *The Plant Journal* 93, 2 (2018), 338–354.
[24] Ray Ming, Robert VanBuren, Ching Man Wai, Haibao Tang, Michael C. Schatz, John E. Bowers, Eric Lyons, Ming-Li Wang, Jung Chen, Eric Biggers, Jisen Zhang, Lixian Huang, Lingmao Zhang, Wenjing Miao, Jian Zhang, Zhangyao Ye, Chenyong Miao, Zhicong Lin, Hao Wang, Hongye Zhou, Won C. Yim, Henry D. Priest, Chunfang Zheng, Margaret Woodhouse, Patrick P. Edger, Romain Guyot, Hao-Bo Guo, Hong Guo, Guangyong Zheng, Ratnesh Singh, Anupma Sharma, Xiangjia Min, Yun Zheng, Hayan Lee, James Gurtowski, Fritz J. Sedlazeck, Alex Harkess, Michael R. McKain, Zhenyang Liao, Jingping Fang, Juan Liu, Xiaodan Zhang, Qing Zhang, Weichang Hu, Yuan Qin, Kai Wang, Li-Yu Chen, Neil Shirley, Yann-Rong Lin, Li-Yu Liu, Alvaro G. Hernandez, Chris L. Wright, Vincent Bulone, Gerald A. Tuskan, Katy Heath, Francis Zee, Paul H. Moore, Ramanjulu Sunkar, James H. Leebens-Mack, Todd Mockler, Jeffrey L. Bennetzen, Michael Freeling, David Sankoff, Andrew H. Paterson, Xinguang Zhu, Xiaohan Yang, J. Andrew C. Smith, John C. Cushman, Robert E. Paull, and Qingyi Yu. 2015. The Pineapple Genome and the Evolution of CAM Photosynthesis. *Nature Genetics* 47, 12 (Dec. 2015), 1435–1442.
[25] 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 (May 2015), 377–386.
[26] Benoit Nabholz, Axel Künstner, Rui Wang, Erich D. Jarvis, and Hans Ellegren. 2011. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Molecular Biology and Evolution* 28, 8 (Aug. 2011), 2197–2210.
[27] Luay Nakhleh, Bernard M. E. Moret, Usman Roshan, Katherine St John, Jerry Sun, and Tandy Warnow. 2002. The Accuracy of Fast Phylogenetic Methods for Large Datasets. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2002), 211–222.
[28] Shu Ouyang, Wei Zhu, John Hamilton, Haining Lin, Matthew Campbell, Kevin Childs, Françoise Thibaud-Nissen, Renae L. Malek, Yuandan Lee, Li Zheng, Joshua Orvis, Brian Haas, Jennifer Wortman, and C. Robin Buell. 2007. The TIGR Rice Genome Annotation Resource: Improvements and New Features. *Nucleic Acids Research* 35, suppl_1 (Jan. 2007), D883–D887.
[29] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. [n. d.]. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* ([n. d.]).
[30] Jose Planta, Yu-Ya Liang, Haoyang Xin, Matthew T. Chansler, L. Alan Prather, Ning Jiang, Jiming Jiang, and Kevin L. Childs. 2022. Chromosome-Scale Genome Assemblies and Annotations for Poales Species Carex Cristatella, Carex Scoparia, Juncus Effusus, and Juncus Inflexus. *G3 (Bethesda, Md.)* 12, 10 (Sept. 2022), jkac211.
[31] Benjamin D. Redelings and Marc A. Suchard. 2005. Joint Bayesian Estimation of Alignment and Phylogeny. *Systematic Biology* 54, 3 (June 2005), 401–418.
[32] D. F. Robinson and L. R. Foulds. 1981. Comparison of Phylogenetic Trees. *Mathematical Biosciences* 53, 1 (Feb. 1981), 131–147.
[33] Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* 61, 3 (May 2012), 539–542.
[34] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. 2011. Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology* 7, 1 (Jan. 2011), 539.
[35] Alexandros Stamatakis. 2014. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30, 9 (May 2014), 1312–1313.
[36] Chris Tuffley and Mike Steel. 1997. Links between Maximum Likelihood and Maximum Parsimony under a Simple Model of Site Substitution. *Bulletin of Mathematical Biology* 59, 3 (May 1997), 581–607.
[37] Lusheng Wang and Tao Jiang. 1994. On the Complexity of Multiple Sequence Alignment. *Journal of Computational Biology* 1, 4 (1994), 337–348.
[38] Wei Wang, Ahmad Hejasebazzi, Julia Zheng, and Kevin J Liu. 2021. Build a Better Bootstrap and the RAWR Shall Beat a Random Path to Your Door: Phylogenetic Support Estimation Revisited. *Bioinformatics* 37, Supplement_1 (2021), i111–i119.
[39] Zhuo Wang, Hongxia Miao, Juhua Liu, Biyu Xu, Xiaoming Yao, Chunyan Xu, Shancen Zhao, Xiaodong Fang, Caihong Jia, Jingyi Wang, Jianbin Zhang, Jingyang Li, Yi Xu, Jiashui Wang, Weihong Ma, Zhangyan Wu, Lili Yu, Yulan Yang, Chun Liu, Yu Guo, Silong Sun, Franc-Christophe Baurens, Guillaume Martin, Frederic Salmon, Olivier Garsmeur, Nabila Yahiaoui, Catherine Hervouet, Mathieu Rouard, Nathalie Laboureau, Remy Habas, Sebastien Ricci, Ming Peng, Anping Guo, Jianghui Xie, Yin Li, Zehong Ding, Yan Yan, Weiwei Tie, Angélique D'Hont, Wei Hu, and Zhiqiang Jin. 2019. Musa Balbisiana Genome Reveals Subgenome Evolution and Functional Divergence. *Nature Plants* 5, 8 (Aug. 2019), 810–821.
[40] Tandy Warnow. 2012. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Currents* 4 (2012).
[41] Tandy Warnow. 2017. *Computational phylogenetics: an introduction to designing methods for phylogeny estimation.* Cambridge University Press.
[42] Ziheng Yang. 1994. Estimating the Pattern of Nucleotide Substitution. *Journal of Molecular Evolution* 39, 1 (July 1994), 105–111.
[43] Ziheng Yang. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24, 8 (Aug. 2007), 1586–1591.
[44] Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. 2018. ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees. *BMC Bioinformatics* 19, 6 (May 2018), 15–30.

Received 11 June 2023; revised 03 August 2023; accepted

Supplementary Online Materials

Rei Doko
Michigan State University
Computer Science and Engineering
East Lansing, Michigan, USA

Kevin Liu
kjl@msu.edu
Computer Science and Engineering
Ecology, Evolution, and Behavior Program
Genetics and Genome Sciences
Michigan State University
East Lansing, Michigan, USA

ACM Reference Format:
Rei Doko and Kevin Liu. 2023. Supplementary Online Materials. 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, 7 pages. https://doi.org/10.1145/3584371.3613011

S1 SUPPLEMENTARY METHODS

Additional assessments of phylogenetic tree estimates. A way to extend RF distance to rooted trees is to consider the bipartition representation for labeled nodes (i.e. $X \cup \{\rho\}$). So for the two subtrees T_1, T_2 induced by deleting an edge $e \in E$, if $\rho \in V_1$ then the edge representation becomes $b'(e) = \{X_1 \cup \{\rho\}, X_2\}$ and vice versa if $\rho \in V_2$.

For assessing branch length estimation error, we used the branch score developed by Kuhner and Felsenstein [4].

Pseudocode for NoHTS phylogenetic support estimation procedure. Pseudocode for NoHTS is shown in Algorithm 1.

Software commands used. INDELible [2] version 1.03 was run using the following settings to simulate model tree evolution

```
[TYPE] NUCLEOTIDE 1
[MODEL] mymodel
    [submodel] JC
[TREE] mytree
    [rooted] <# taxa>
[PARTITIONS] mypartition [mytree mymodel 1]
[EVOLVE] mypartition <# replicates> output
```

The trees are rescaled to have non-ultrametric branch lengths and the following control settings to simulate sequence evolution:

```
[TYPE] NUCLEOTIDE 1
[SETTINGS]
  [output] PHYLIP
[MODEL] background
  [submodel] GTR <CT> <AT> <GT> <AC> <CG>
  [statefreq] <T> <C> <A> <G>
  [indelmodel] USER <path to indel distribution>
  [indelrate] <indel rate>
```

Algorithm 1 NoHTS procedure and support estimation algorithm.

procedure RESAMPLE(A, γ)
  $X \leftarrow \langle\rangle$
  $i \sim U(1, \ldots, |A|)$
  direction $\sim U(-1, 1)$
  for $j \in \{1 \ldots |A|\}$ **do**
    $X[j] \leftarrow A[i]$
    $r \sim U[0, 1]$
    if $r < \gamma \vee i + \text{direction} \notin \{1, \ldots, |A|\}$ **then**
      direction $\leftarrow -1 \times$ direction
    end if
    $i \leftarrow i + \text{direction}$
  end for
  remove all indels from X
  return X
end procedure

procedure CALCULATESUPPORT(A, γ, n, T, MSA, MLE)
  $\forall e \in T, \epsilon(e) \leftarrow 0$
  for $i \in \{1 \ldots n\}$ **do**
    $X_i \leftarrow \text{Resample}(A, \gamma)$
    $T_i \leftarrow \text{MLE}(\text{MSA}(X_i))$
    for $e \in \{e | e \in T \wedge e \in T_i\}$ **do**
      $\epsilon(e) \leftarrow \epsilon(e) + \frac{1}{n}$
    end for
  end for
  return ϵ
end procedure

```
[MODEL] shift
  [submodel] GTR <CT> <AT> <GT> <AC> <CG>
  [statefreq] <T> <C> <A> <G>
  [indelmodel] USER <path to indel distribution>
  [indelrate] <indel rate>
[TREE] mytree <tree>
  [treedepth] <specified tree height>
[BRANCHES] mymodel <tree with model placements defined>
[PARTITIONS] mypartition [mytree mymodel <sequence length>]
[EVOLVE] mypartition 1 sequence
```

The following command was used to perform MSA estimation with MAFFT [3] version 7.475

```
mafft <input sequence> <output alignment>
```

MUSCLE [1] version 5.0.1428 was run with the following:

```
muscle -align <input sequence> -output <output alignment>
```


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

ACM ISBN 979-8-4007-0126-9/23/09…$15.00
https://doi.org/10.1145/3584371.3613011

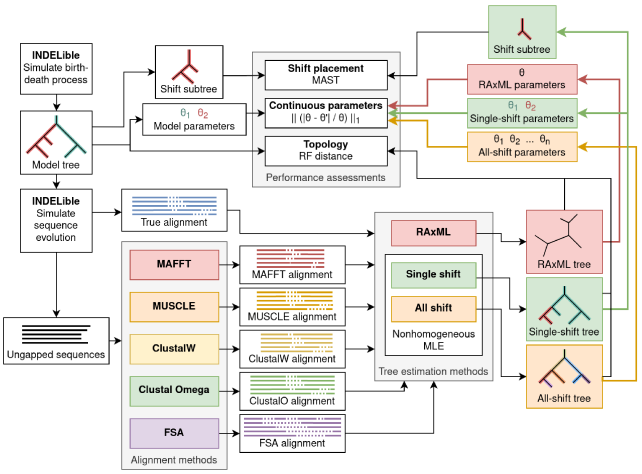

Figure S1: *Flowchart of simulation study steps.* A model tree is first generated under a birth-death process using INDELible, and then nucleotide sequence evolution is simulated along the model tree under a nonhomogeneous substitution model. Unaligned sequences are aligned using MAFFT, MUSCLE, ClustalW, Clustal Omega, and FSA. Every estimated alignment, as well as the true alignment, is then used to perform tree estimation. Tree estimation is performed using MLE under three different classes of models. MLE under the homogeneous no-shift model is performed using RAxML. MLE under the nonhomogeneous single-shift and all-shift models are performed using local optimization coupled with PAML to calculate likelihood scores and optimize continuous parameters. Finally, the resulting trees from each alignment and tree estimation method pair is compared against the model tree. For the single shift model, shift placement is evaluated by computing the size of the maximum agreement subtree (MAST) for the true and estimated shift subtrees.

Clustal Omega [6] version 1.2.4 was run with the following

```
clustalo -i <input sequence> -t DNA –threads 1 ><output alignment>
```

ClustalW [5] version 2.1 was run with the following

```
clustalw2 <input sequence> -type=DNA -outfile=<output alignment>
```

FSA [6] version 1.15.9 was run with the following

```
fsa <input sequence –maxram 8192 ><output alignment>
```

For single-shift search, PAML [8] was run with the following control file:

```
     seqfile = <sequence_path>
    treefile = <tree path>
     outfile = <result output path>
       noisy = 3
     verbose = 3
     runmode = 0
       model = 7   * GTR model
       Mgene = 0
       ndata = 1
       nhomo = 5
   fix_kappa = 2
       clock = 0
   fix_alpha = 1
       alpha = 0.
       getSE = 0
RateAncestor = 0
   cleandata = 0
      method = 0
 fix_blength = 0
```

The control file for the all-shift model is identical, except:

```
       nhomo = 3
   fix_kappa = 0
```

RAxML [7] was run with the following command:

```
raxml -s <msa path> -n <name> -m GTRCAT -V -p <random number>
```

Flowering monocot dataset processing. To obtain single copy orthologs for the flowering monocot dataset, we ran orthofinder using the following command:

```
orthofinder -f <path containing sequence files>
```

Table S1: Mean normalized Robinson-Foulds distances for each model condition, alignment method, and maximum likelihood estimation method.

Model condition	TRUE			MAFFT			MUSCLE		
	0-shift	1-shift	all-shift	0-shift	1-shift	all-shift	0-shift	1-shift	all-shift
10A	0.043	0.033	0.057	0.086	0.086	0.105	0.148	0.081	0.162
10B	0.057	0.043	0.090	0.190	0.181	0.190	0.252	0.171	0.229
10C	0.081	0.067	0.129	0.205	0.181	0.214	0.338	0.219	0.276
10D	0.152	0.086	0.124	0.310	0.329	0.329	0.400	0.281	0.381
10E	0.390	0.200	0.186	0.438	0.405	0.386	0.486	0.395	0.481
20A	0.061	0.065	0.118	0.129	0.137	0.191	0.129	0.125	0.183
20B	0.095	0.075	0.110	0.227	0.207	0.237	0.183	0.148	0.187
20C	0.088	0.073	0.120	0.294	0.296	0.339	0.320	0.239	0.288
20D	0.124	0.090	0.108	0.445	0.435	0.445	0.425	0.361	0.373
20E	0.400	0.208	0.163	0.645	0.629	0.618	0.643	0.551	0.551
Model condition	ClustalW			Clustal Omega			FSA		
	0-shift	1-shift	all-shift	0-shift	1-shift	all-shift	0-shift	1-shift	all-shift
10A	0.157	0.100	0.157	0.105	0.076	0.095	0.286	0.176	0.186
10B	0.210	0.190	0.205	0.181	0.186	0.210	0.333	0.195	0.252
10C	0.181	0.148	0.200	0.190	0.167	0.181	0.424	0.338	0.348
10D	0.319	0.276	0.300	0.295	0.276	0.348	0.581	0.476	0.500
10E	0.471	0.410	0.419	0.438	0.352	0.386	0.605	0.562	0.581
20A	0.204	0.184	0.204	0.192	0.180	0.218	0.157	0.135	0.225
20B	0.223	0.213	0.243	0.237	0.223	0.275	0.331	0.260	0.296
20C	0.275	0.276	0.273	0.278	0.259	0.288	0.503	0.367	0.347
20D	0.369	0.369	0.361	0.384	0.375	0.384	0.633	0.563	0.473
20E	0.600	0.525	0.539	0.559	0.512	0.535	0.732	0.727	0.662

S2 SUPPLEMENTARY RESULTS AND DISCUSSION

Simulation study runtime and memory usage. On the 10-taxon simulated datasets, single-shift MLE runtimes were modest – amounting to less than an hour for each dataset analysis in almost all cases (Supplementary Figure S2). Single-shift MLE runtimes on 20-taxon datasets were considerably higher and amounted to multiple hours and as much as a day on the most divergent datasets. For each simulation condition, single-shift MLE on different MSAs yielded comparable runtimes, with the exception of FSA which returned higher runtime by several factors. Main memory requirements were reasonable – at most 1 GiB in all cases, which is well within the scope of modern personal computers.

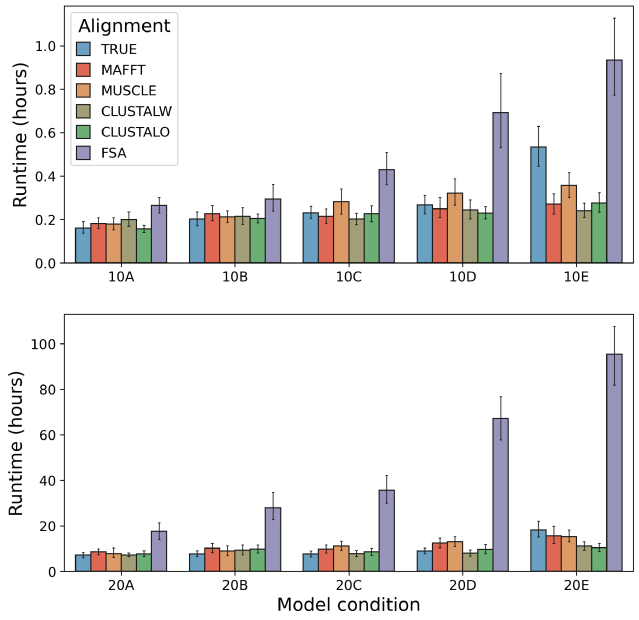

Figure S2: Simulation study: single shift search runtime. Averages and standard error bars are reported for each model condition ($n = 30$).

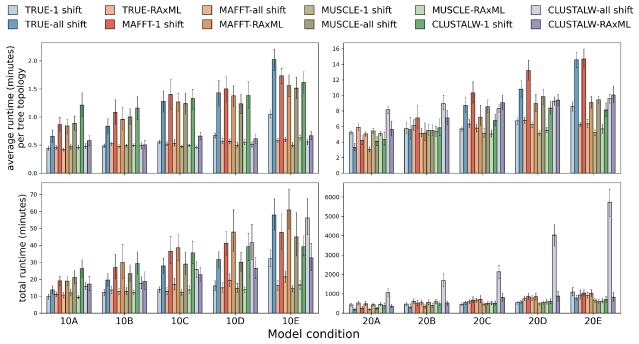

Figure S3: Runtime for nonhomogeneous tree search. The first row represents how long it takes for PAML to do continuous parameter optimization and likelihood calculation for a fixed tree topology, and the second row represents the total time used for non-homogeneous MLE.

Estimation error. For branch lengths, there wasn't a clean way to quantify error, in part because tree estimation and branch length estimation are heavily intertwined. Comparing leaf-edge distances only captures some of the estimated branch lengths, but comparing pairwise distances does not consider the topology. Kuhner-Felsenstein distance [4], shown in the middle panels in S5, takes topology into consideration, but does not easily allow examination of one in isolation of the other.

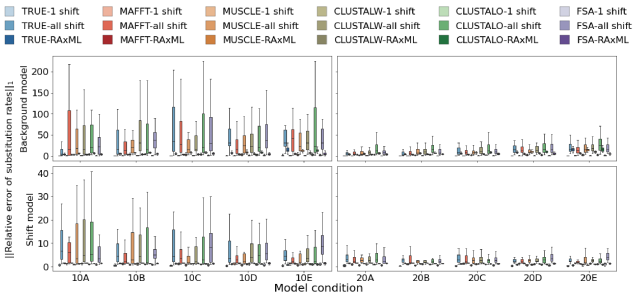

Figure S4: Substitution rate error for no-shift, single-shift, and all-shift based estimation. Figure layout and description are otherwise identical to Figure 4 in the main manuscript.

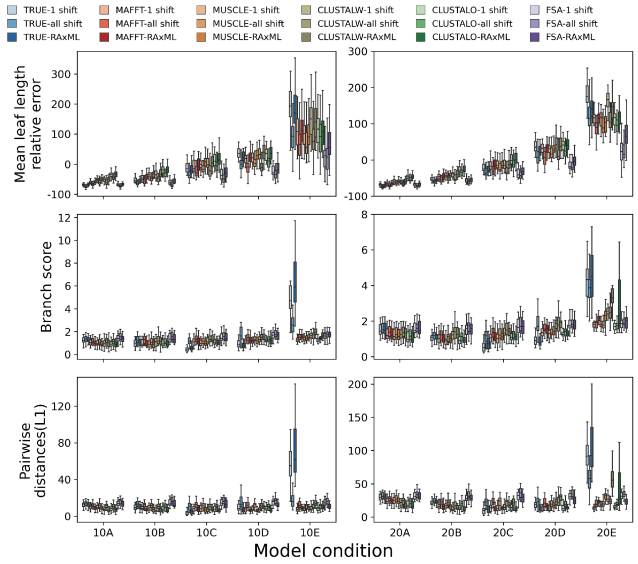

Figure S5: Various measures for branch length estimation error. The first row reports mean leaf length error for each replicate, the second row reports branch score, the measure described in [4], and the third row reports the L1 norm of the pairwise distance error between each pair of leaves.

	muscle		clustalo		clustalw		fsa	
nRF (rooted)	Mean	Std	Mean	Std	Mean	Std	Mean	Std
mafft	0.171	0.191	0.188	0.200	0.191	0.198	0.191	0.205
muscle			0.192	0.204	0.191	0.203	0.198	0.206
clustalo					0.192	0.204	0.213	0.205
clustalw							0.217	0.207

Table S2: Empirical study: topological disagreement among gene trees estimated using different MSA methods and single-shift MLE, accounting for rooting. Values are calculated similarly to Table 6 in the main manuscript, where values reported are a rooted extension of Robinson-Foulds distance.

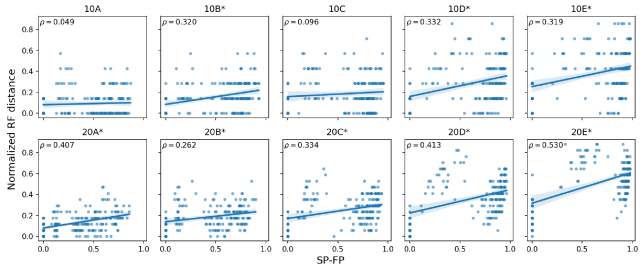

Figure S6: *Simulation study: the relationship between upstream MSA estimation error and downstream phylogenetic estimation error.* The former was assessed based on sum-of-pairs false positive rate, and the latter was assessed based on normalized Robinson-Foulds distance. Figure layout and description are otherwise identical to Figure 3 in main manuscript.

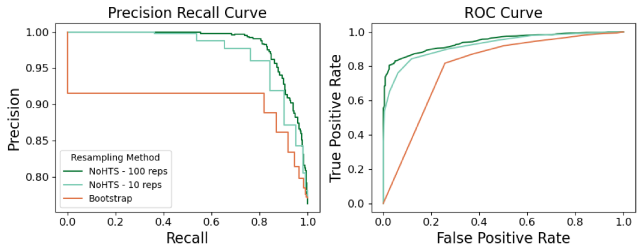

Figure S7: Simulation study: NoHTS and bootstrap support estimation precision-recall and ROC-curves. Results are aggregated over all 10-taxon model conditions.

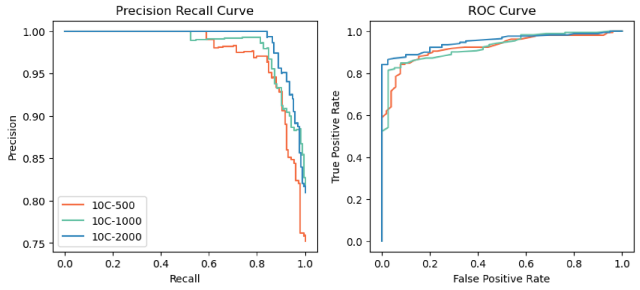

Figure S8: NoHTS support estimation precision-recall and ROC curves for 10C model conditions with varying sequence length of 500, 1000, and 2000 base pairs long. The PR-AUC and ROC-AUC values were 0.976 (0.924), 0.983 (0.926), 0.989 (0.952) for the 500 bp, 1000 bp, and 2000 bp model conditions, respectively.

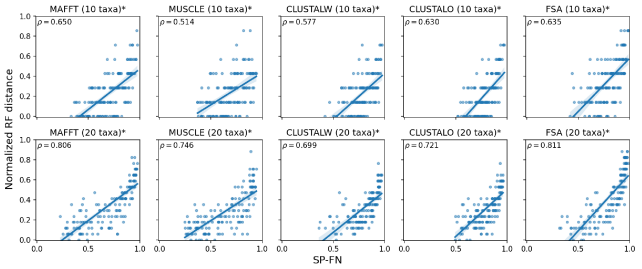

Figure S9: SP-FN vs normalized Robinson-Foulds distance linear by alignment method, aggregated across all 10 and 20 taxa model conditions.

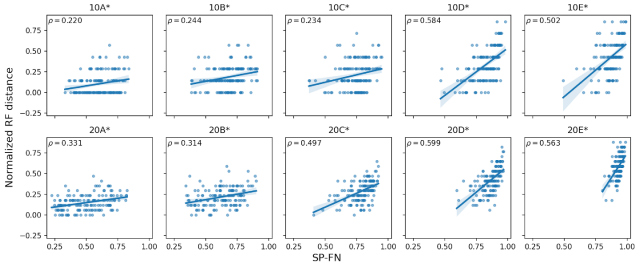

Figure S10: *SP-FN vs normalized Robinson-Foulds distance linear regression when excluding true alignments.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

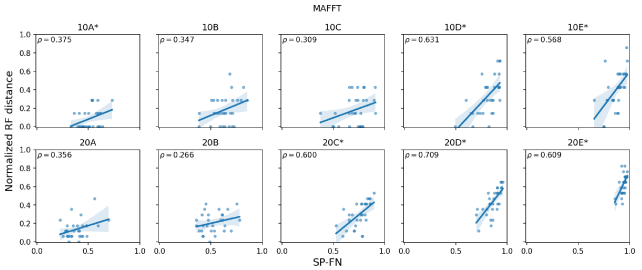

Figure S11: *SP-FN vs normalized Robinson-Foulds distance linear regression using MAFFT-based analyses.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

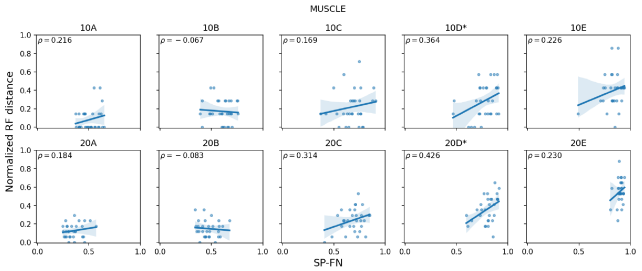

Figure S12: *SP-FN vs normalized Robinson-Foulds distance linear regression using MUSCLE-based analyses.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

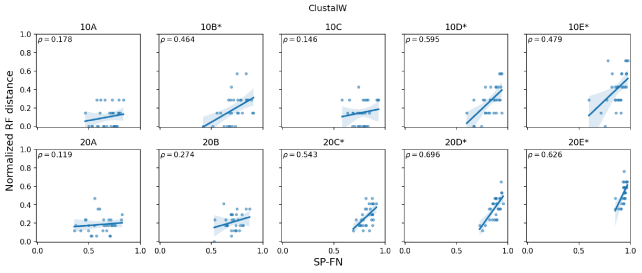

Figure S13: *SP-FN vs normalized Robinson-Foulds distance linear regression using ClustalW-based analyses.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

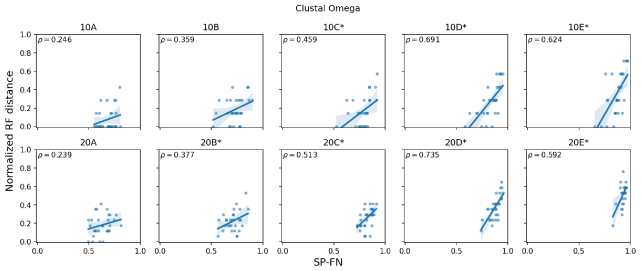

Figure S14: *SP-FN vs normalized Robinson-Foulds distance linear regression using Clustal Omega-based analyses.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

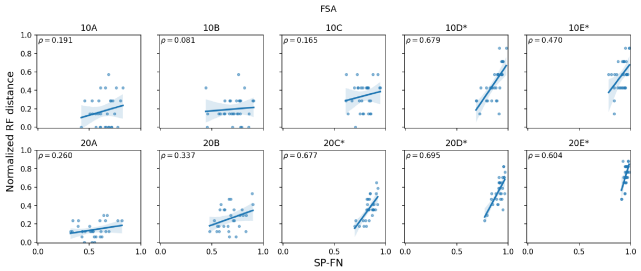

Figure S15: *SP-FN vs normalized Robinson-Foulds distance linear regression using FSA-based analyses.* Figure layout and description are otherwise identical to Figure 3 in the main manuscript.

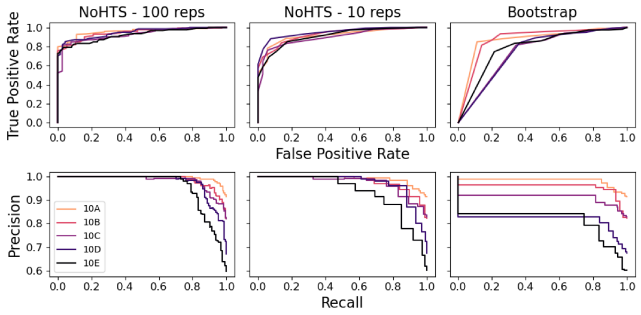

Figure S16: NoHTS and bootstrap support estimation precision-recall and ROC curves, comparing model conditions. For model conditions 10A through 10E, the PR-AUC and ROC-AUC values of NoHTS with 100 replicates were 0.996 (0.957), 0.988 (0.944), 0.983 (0.926), 0.977 (0.947), and 0.957 (0.922) respectively. For NoHTS with 10 replicates, these values were 0.992 (0.916), 0.984 (0.926), 0.978 (0.899), 0.978 (0.950), and 0.945 (0.910) respectively. For bootstrap, these values were 0.988 (0.875), 0.972 (0.876), 0945 (0.748), 0.889 (0.749), and 0.874 (0.791) respectively.

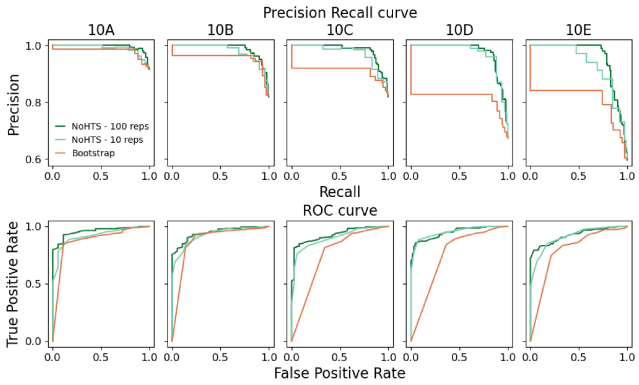

Figure S17: NoHTS and bootstrap support estimation precision-recall and ROC curves, comparing methods. In every model condition, NoHTS both with 100 and 10 replicates yielded a better PR-AUC and ROC-AUC value over bootstrap with 100 replicates.

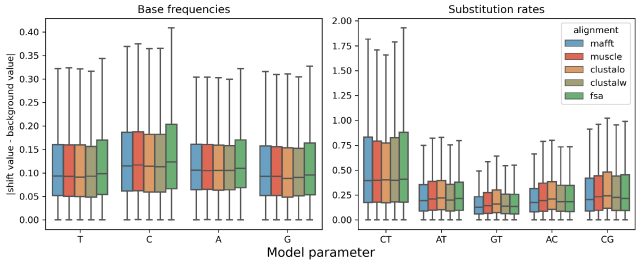

Figure S18: *Empirical study: box-and-whisker plot of the difference between background and shift model parameters estimated on single copy orthologs.* Median and interquartile range are shown, whiskers are 1.5IQR, and values outside of the whiskers are not shown ($n = 1377$).

REFERENCES

[1] Robert C. Edgar. 2004. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32, 5 (March 2004), 1792–1797.

[2] William Fletcher and Ziheng Yang. 2009. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution* 26, 8 (Aug. 2009), 1879–1888.

[3] 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 (April 2013), 772–780.

[4] M K Kuhner and J Felsenstein. 1994. A Simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution* 11, 3 (May 1994), 459–468.

[5] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X Version 2.0. *Bioinformatics* 23, 21 (Nov. 2007), 2947–2948.

[6] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. 2011. Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology* 7, 1 (Jan. 2011), 539.

[7] Alexandros Stamatakis. 2014. RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30, 9 (May 2014), 1312–1313.

[8] Ziheng Yang. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24, 8 (Aug. 2007), 1586–1591.