The Impact of Gene Sequence Alignment and Gene Tree Estimation Error on Summary-Based Species Network Estimation

Meijun Gao
Computer Science and Engineering
Michigan State University
East Lansing, Michigan, USA

Wei Wang
Meta Corporation
Menlo Park, California, USA

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

ABSTRACT

Thanks in part to rapid advances in next-generation sequencing technologies, recent phylogenomic studies have demonstrated the pivotal role that non-tree-like evolution plays in many parts of the Tree of Life – the evolutionary history of all life on Earth. As such, the Tree of Life is not necessarily a tree at all, but is better described by more general graph structures such as a phylogenetic network. Another key ingredient in these advances consists of the computational methods needed for reconstructing phylogenetic networks from large-scale genomic sequence data. But virtually all of these methods either require multiple sequence alignments (MSAs) as input or utilize gene trees or other inputs that are computed using MSAs. All of the input MSAs and gene trees must be estimated on empirical data. The methods themselves do not directly account for upstream estimation error, and, apart from prior studies of phylogenetic tree reconstruction and anecdotal evidence, little is understood about the impact of estimated MSA and gene tree error on downstream species network reconstruction.

We therefore undertake a performance study to quantify the impact of MSA error and gene tree error on state-of-the-art phylogenetic network inference methods. Our study utilizes synthetic benchmarking data as well as genomic sequence data from mosquito and yeast. We find that upstream MSA and gene tree estimation error can have first-order effects on the accuracy of downstream network reconstruction and, to a lesser extent, its computational runtime. The effects become more pronounced on more challenging datasets with greater evolutionary divergence and more sampled taxa. Our findings highlight an important need for computational methods development: namely, scalable methods are needed to account for estimated MSA and gene tree error when reconstructing phylogenetic networks using unaligned biomolecular sequence data.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
BCB '22, August 7–10, 2022, Northbrook, IL, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9386-7/22/08…$15.00
https://doi.org/10.1145/3535508.3545559

CCS CONCEPTS

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

KEYWORDS

multiple sequence alignment, gene tree, phylogenetic network, species network, simulation study, mosquito, yeast

ACM Reference Format:
Meijun Gao, Wei Wang, and Kevin J. Liu. 2022. The Impact of Gene Sequence Alignment and Gene Tree Estimation Error on Summary-Based Species Network Estimation. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3535508.3545559

1 INTRODUCTION

Among the many advances enabled by next-generation genomic sequencing and large-scale genomic data are new discoveries concerning the prevalence of gene flow in the Tree of Life [1, 17, 22]. A corollary question immediately follows: to what extent is the Tree of Life not really a tree, but rather a more general graph structure known as a phylogenetic network [14]?

In addition to data, advances in computational methodologies are needed to resolve these major questions. New algorithms have been developed for reconstructing phylogenetic networks from genomic sequence data [14, 25]. The methods broadly fall into two classes: (1) statistical methods, and (2) parsimony-based methods. The former class of methods requires an explicit evolutionary model such as the multi-species network coalescent (MSNC) model [41, 44]; methods to perform statistical optimization under these models include maximum likelihood methods [36, 40, 42], maximum pseudo-likelihood methods [35, 43], and Bayesian methods [2, 39, 45]. The latter class of methods includes the method of Yu et al. [44]. We focus on statistical methods in our study since they been shown to be generally more accurate than the latter class, where a reasonable parametric model is available for statistical inference and learning [11].

But advances in phylogenetic network estimation alone are insufficient. Many widely-used phylogenetic network estimation methods and methodological pipelines either directly or indirectly utilize a multiple sequence alignment (MSA) as a fixed input [25]. For example, summary-based species phylogeny estimation methods

require a set of gene trees as input, and gene trees are typically estimated using a "two-phase" method: for each gene, unaligned sequence data are first aligned into an MSA, and then gene tree reconstruction is performed using the MSA as input. Upstream estimates are fixed in downstream phylogenetic estimation using multi-locus inputs, which effectively makes an implicit assumption about estimated MSA and estimated gene tree accuracy. (Throughout our study and the rest of this manuscript, "locus" refers to a gene unless otherwise noted.)

This assumption has long been known to be an over-simplification for phylogenetic tree reconstruction, and numerous studies have demonstrated the major impact of MSA quality on downstream phylogenetic tree estimation accuracy [20, 21, 28]. Since phylogenetic networks generalize trees, a similar conclusion seems natural for phylogenetic network estimation. But there remains a need for direct investigation to explicitly test this hypothesis.

On the other hand, anecdotal evidence does exist. An illustrative example can be found in Hagelberg et al. [8]'s phylogenetic analysis of human mitochondrial DNA (mtDNA) sequence data sampled from across the western Pacific. The outcome of their study was the putative discovery of genetic recombination in human mtDNA. Such a finding would provide a ground-breaking counterexample against the conventional wisdom that human mtDNA is matrilineally inherited, and would imply that the evolutionary history of human mtDNA takes the form of an ancestral recombination graph (ARG) rather than a tree. But it turned out that their phylogenetic analyses utilized an estimated MSA but did not account for alignment error: a single mis-alignment resulted in artificially elevated divergence at a key site and was interpreted as a rare mutation [9]. The resulting artifact led the authors to support a non-tree-like evolutionary hypothesis invoking genetic recombination due to paternal mtDNA leakage. Subsequent to the study's publication, the authors identified the MSA error and took the commendable action of issuing a major correction to their publication and rescinding their finding.

The example provides an important cautionary tale. What else could we be missing in phylogenetic and phylogenomic studies of gene flow and other claimed findings of non-tree-like biomolecular sequence evolution?

2 MATERIALS AND METHODS

To begin to resolve these questions, we conducted a performance study to investigate the impact of MSA accuracy and gene tree accuracy on state-of-the-art summary-based phylogenetic network estimation methods.

Simulated datasets. Our simulation study utilized single-reticulation model networks with either 4 or 8 in-group taxa. Model networks were obtained using the basic procedure from the study of Hejase and Liu [11]. First, a random tree with $n \in \{4, 8\}$ taxa and height h_0 was sampled under a random birth-death process using r8s version 1.8.1 [33]. The branch lengths were then rescaled by a factor $\frac{h}{h_0}$ so that the model phylogeny had height h. Next, a single reticulation was added using the following procedure: (1) a reticulation event time t_M was chosen uniformly at random from the interval $[0.01, \frac{h}{4}]$, two extant populations at time t_M were randomly selected, and a reticulation edge with random orientation was added to connect the corresponding pair of tree edges. Finally, an outgroup taxon was added to the resulting model network with GMRCA at time $2.0h$. (See Supplementary Figure S1 in Appendix for several examples of model networks from our simulation study.)

For each model network, ms [13] was used to conduct simulations under the multi-species coalescent and isolation-with-migration (MSC+IM) model. As in the study of Hejase and Liu [11], a reticulation with time t_M was modeled using a unidirectional migration event from time $t_M - 0.01$ to $t_M + 0.01$ with migration rate 5.0. Each MSC+IM simulation sampled 1000 local coalescent histories and gene trees.

INDELible version 1.03 [6] was used to simulate sequence evolution along each local gene tree. Local coalescent history times measured in coalescent units were converted into gene tree branch lengths measured in expected numbers of substitutions using equation 3.1 in [10] and scaled mutation rate θ. Gene tree branch lengths were then deviated away from ultrametricity under the model of [26] with deviation factor $c = 2.0$, resulting in non-ultrametric gene trees. Sequence evolution was simulated under a finite-sites model of nucleotide substitutions, insertions, and deletions. The General Time-Reversible (GTR) model was used for the former. GTR model parameter values were based on empirical nematode Tree of Life (NemAToL) estimates from the study of Liu et al. [21], where base frequencies $[\pi_T, \pi_C, \pi_A, \pi_G]$ were set to [0.3115, 0.1913, 0.3004, 0.1967] and substitution rates $[r_{TC}, r_{TA}, r_{TG}, r_{CA}, r_{CG}, r_{AG}]$ were set to $[1.2620, 0.1401, 0.2878, 0.3577, 0.3083, 1.0]$. The insertion/deletion model utilized the medium gap length distribution from the study of Liu et al. [21], where the probability distribution $[p_1, p_2, \ldots]$ that specifies the probability p_i of a gap with length i was set to [0.2012, 0.1600, 0.1280, 0.1024, 0.0819, 0.0655, 0.0524, 0.0419, 0.0336, 0.0268, 0.0215, 0.0172, 0.0137, 0.0110, 0.0088, 0.0070, 0.0056, 0.0045, 0.0036, 0.0029, 0.0023, 0.0018, 0.0015, 0.0012, 0.0009, 0.0008, 0.0006, 0.0005, 0.0004, 0.0003, 0.0002]. Mutation rates and insertion/deletion rates were chosen to span a range of sequence divergence observed in non-intronic and intronic loci, similar to the traditional phylogenetic marker-based benchmarking datasets in the Comparative RNA Website database [4] and the simulation study of Liu et al. [21]. As another point of reference, our intermediate model conditions have observed sequence divergence that is in line with subsets of the avian Tree of Life WGS dataset from the study of Jarvis et al. [15]. The sequence length at the root was set to 1 kb.

To obtain experimental replication, the simulation procedure was repeated to obtain 20 replicate datasets per model condition. Table 1 lists model condition parameter values and summary statistics for true MSAs.

Methods and performance measures used in simulation study. Our study focused on statistical methods for phylogenetic network reconstruction. As is the case with many of the most popular phylogenetic methods, the phylogenetic methods in our study are used as one stage of a methodological pipeline. MSA estimation and gene tree estimation are typically performed as upstream pipeline stages.

We focused on summary-based network inference under the MSNC model as implemented in the PhyloNet software package [36, 40], which has been shown to be among the most accurate and popular approaches for this problem [11]. Similar to widely-used "two-phase" methods in traditional phylogenetics that first estimate an MSA from biomolecular sequence data and then estimate a

phylogenetic tree from the inferred MSA, summary-based inference methods take the form of a methodological pipeline: (1) unaligned biomolecular sequence data for each locus is aligned into an MSA, (2) a gene tree is inferred using the input MSA for each locus, and (3) a species network is inferred using the set of inferred rooted gene trees, which act as "summaries" of biomolecular sequence data from the sampled loci.

In our simulation study, MSAs in the first pipeline stage consisted of either the true alignment generated by INDELible or an estimated alignment. We used either MAFFT [16], FSA [3], Clustal Omega [34], or ClustalW [18] for MSA estimation; these MSA methods have been shown to be among the leading methods in terms of alignment accuracy, downstream tree inference accuracy, and/or popularity [16, 19, 21]. (See Appendix for supplementary experiments with two additional MSA methods.) Default settings were used for the MAFFT L-INS-i algorithm as implemented in MAFFT version 7.450, Clustal Omega version 1.2.4, ClustalW version 2.1, and FSA version 1.15.9. Table 2 lists summary statistics for the estimated MSAs.

Gene trees in the second pipeline stage consisted of either the true gene trees or inferred gene trees that were obtained using the following procedure: on either the true MSA or an estimated MSA, we ran FastTree version 2.1.11 [29] with default settings to perform maximum likelihood estimation of an unrooted gene tree under the GTR+Γ model of nucleotide substitution [5, 27, 31]. To obtain rooted gene trees for input into the last pipeline stage, estimated gene trees in the second-to-last pipeline stage were rooted using outgroup rooting. The leaf edge to the outgroup taxon was then pruned from each rooted gene tree, since outgroups were used solely for rooting gene trees.

Finally, for each set of rooted gene trees – either true gene trees, or estimated gene trees that were inferred from true alignments or estimated alignments – PhyloNet [36, 40] was used to perform summary-based network inference under one of two different MSNC-based optimization criteria: model likelihood given gene tree topologies as input [43], or model pseudo-likelihood given gene tree topologies as input [43]. We refer the two summary-based inference methods as MLE and MPL, respectively; both were run using default settings and version 3.8.2 of the PhyloNet software package.

We evaluated method performance based on estimated MSA error with respect to the true MSA and topological error of an inferred phylogeny with respect to model phylogeny. We assessed both type I and type II error of estimated MSAs: the SP-FP proportion is the fraction of nucleotide-nucleotide homologies that appear in an estimated MSA but not the true MSA, and the SP-FN proportion is the fraction of nucleotide-nucleotide homologies that appear in the true MSA but not an estimated MSA, respectively. Topological error of gene trees was assessed using the Robinson-Foulds distance [30], which is the proportion of bipartitions that are present in an estimated tree but not the true tree or vice versa. Topological error of species networks was assessed using Nakhleh's [24] equivalence-based calculation which is a metric on the set of reduced phylogenetic networks. The calculation reflects the number of rooted subnetworks that appear in one network but not the other or vice versa.

Empirical datasets and methods. Our empirical study utilized genomic sequence datasets from two previous studies [32, 38]. Both studies included species for which non-tree-like evolution has been hypothesized.

The first dataset came from Fontaine et al. [7]'s study of adaptive introgression in mosquitoes and was later re-analyzed by Wen et al. [38]. The dataset samples a total of 7 species (including one out-group taxon) and 3019 genomic loci. The in-group taxa are *Anopheles gambiae*, *A. coluzzii*, *A. arabiensis*, *A. quadriannulatus*, *A. merus*, and *A. melas*, which we abbreviate as G, C, A, Q, R, and L, respectively; *A. christyi* serves as the out-group taxon. The number of taxa and their evolutionary divergence are generally within the scope of the simulation study, although the number of loci is greater by a factor of ∼3.

The second dataset came from the study of Salichos and Rokas [32]. The dataset includes genomic sequence data for 23 yeast species and 4435 loci in total. Species network analyses of a larger dataset – both in terms of number of taxa and sampled loci – complement the other experiments in our performance study and provide additional guidance on the impact of dataset scale on our study findings.

For both datasets, species networks were reconstructed using summary-based phylogenomic inference as in the simulation study. First, a multiple sequence alignment was estimated for each locus using ClustalW version 2.1, MAFFT version 7.222, or FSA version 1.15.9. Next, we used FastTree version 2.1.11 to infer an unrooted gene tree for each multiple sequence alignment under the GTR+Γ model of nucleotide substitution. Unrooted gene trees were then outgroup rooted, after which the outgroup taxon was pruned by deleting its pendant leaf edge. Finally, given the rooted gene trees as input, a species network with r reticulations was inferred using MPL optimization as implemented in PhyloNet version 3.6.0, where $r \in [0, 4]$.

Data availability and computational resources used for experiments. The software and datasets used in our study are available at https://gitlab.msu.edu/liulab/impact-of-msa-quality-on-network-inference.public. The study software and datasets are provided under permissive copyleft open licenses.

Detailed software commands are provided in the Supplementary Appendix. All computational experiments were performed on the MSU High Performance Computing Center, with hardware consisting of Intel Xeon CPUs running at 2.4 or 2.5 GHz.

3 RESULTS

Estimation accuracy of summary-based phylogenetic methods. Summary statistics calculated on the estimated MSAs are shown in Table 2. Estimated MSA error, topological error of estimated gene trees, and topological error of MLE-estimated species networks are shown in Figures 1, 2, and 3 respectively.

In general, summary-based MLE analyses returned highest accuracy when provided true trees as input, followed by FastTree run on true MSAs, and then FastTree run on estimated MSAs – where the latter were ranked in order of increasing topological error as follows: MLE(FastTree(ClustalOmega)), MLE(FastTree(MAFFT)), MLE(FastTree(ClustalW), and MLE(FastTree(FSA)). The relative method comparisons are as expected since the first two pipelines utilize ground truth and therefore represent theoretical baselines. The MLE-based method comparisons were consistent across dataset

Table 1: Simulation study: model parameters and true multiple sequence alignment (MSA) statistics. The 4-taxon model conditions are named 4.A through 4.E in order of generally increasing evolutionary divergence; the 8-taxon model conditions are named 8.A through 8.E similarly. Additional model condition parameters include the insertion/deletion rate and the model phylogeny height (see Methods section for details). Average normalized Hamming distance ("ANHD"), the percentage of true MSA cells that consist of indels ("Gappiness"), and the number of true MSA sites ("Length") are reported as an average for each model condition ($n = 20$).

Model condition	Insertion/ deletion rate	Model phylogeny height
4.A	0.12	0.3
4.B	0.08	0.5
4.C	0.06	0.7
4.D	0.04	1.0
4.E	0.03	1.4
8.A	0.06	0.5
8.B	0.05	0.6
8.C	0.03	1.0
8.D	0.02	1.7
8.E	0.013	2.4

Model condition	True MSA		
	ANHD	Gappiness	Length
4.A	0.373	0.312	1458.6
4.B	0.432	0.334	1505.7
4.C	0.470	0.343	1529.9
4.D	0.510	0.333	1505.5
4.E	0.545	0.344	1530.8
8.A	0.338	0.324	1487.2
8.B	0.360	0.325	1487.3
8.C	0.422	0.325	1487.6
8.D	0.477	0.353	1552.8
8.E	0.515	0.333	1506.8

sizes. On four-taxon model conditions, the MLE-based methods – MLE(TrueGeneTrees), MLE(FastTree(TrueAln)), MLE(FastTree(ClustalOmega)), MLE(FastTree(MAFFT)), MLE(FastTree(ClustalW)), and MLE(FastTree(FSA)) – returned average network errors (as measured using Nakhleh's reduction-based metric) of 1.71, 2.38, 3.05, 3.13, 3.56, and 4.24, respectively. On eight-taxon model conditions, higher absolute network estimation error was observed overall and average network errors of these methods was 3.27, 3.61, 3.71, 4.34, 4.50, and 6.13, respectively. Some per-model-condition variability was observed within these overall trends, which we attribute to stochasticity of the simulation study experiments. The study findings were somewhat sensitive to random reticulations, which experimental replication helps to mitigate. For each method, estimation error increased in a fairly consistent manner as evolutionary divergence increased in the 4-taxon simulation experiments, and a similar phenomenon was observed in the 8-taxon simulation experiments. The observation applied to MSA estimation (Figure 1), gene tree estimation (Figure 2), and species network estimation (Figure 3). However, the rate of increase in estimation error as evolutionary divergence increased was not the same across all methods. Some increased faster (e.g., FSA-based phylogenetic analyses), and others more slowly.

Overall, MSA error, gene tree error, and MLE network estimation error were qualitatively correlated across model conditions and methods. While the quality of input MSAs and gene trees – ranging from ground truth to relatively accurate estimates to relatively inaccurate estimates – tended to be reflected in topological error of downstream network estimation, a few minor exceptions were noted. On two of the least divergent model conditions in our study, FastTree analyses of some estimated MSAs return gene tree accuracy comparable to FastTree on true MSAs; these were the only model conditions where downstream MLE network inference using these estimated gene trees returned comparable accuracy to network inference using true gene trees as input. Another interesting anomaly concerned Clustal Omega MSAs. Clustal Omega was among the least accurate MSA methods in terms of alignment error as measured using SPFN and SPFN; FastTree(ClustalOmega) similarly returned less accurate gene trees compared to other gene tree estimation methods. And yet downstream MLE network estimation using FastTree(ClustalOmega)-estimated gene trees as input was often the most accurate among analyses using estimated gene trees. The finding is surprising since Clustal Omega was developed for MSA estimation (and gene tree estimation, to a lesser extent) but not for species network estimation. We hypothesize that Clustal Omega MSA estimation may be biased in a manner that is suited to the specific experimental settings in our study.

Figure 4 reports topological error of MPL-estimated networks. Overall, MPL analyses returned higher topological error compared to MLE analyses on a given model condition and set of inputs (i.e., MSAs and gene trees). This is as expected since MPL utilizes a pseudolikelihood criterion that is an approximation to MLE's full model likelihood criterion, and pseudolikelihoods were designed to tradeoff accuracy for speed during optimization [35, 43]. The finding is also consistent with prior performance studies [11]. The comparison among different MPL-based analyses returned smaller differences in topological error, as compared to MLE-based analyses. We attribute this finding to the higher overall topological error returned by MPL-based analysis compared to an otherwise equivalent MLE-based analysis. As overall estimation error approaches saturation, the impact of upstream factors (e.g., input MSA error and input gene tree error) likely becomes more difficult to distinguish from downstream network estimation error.

Runtime and memory usage of summary-based phylogenetic methods. MSA and gene tree quality also had secondary impacts on species network estimation runtime. Figures 5 and 6 show runtime results for MLE- and MPL-based analyses, respectively. FSA-based MSA estimation and downstream gene tree and species network estimation using FSA were least accurate compared to other MSA estimation methods; runtime of FSA-based analyses was also slowest, and increased evolutionary divergence had the most dramatic impact on FSA-based analysis runtime – ballooning by as much

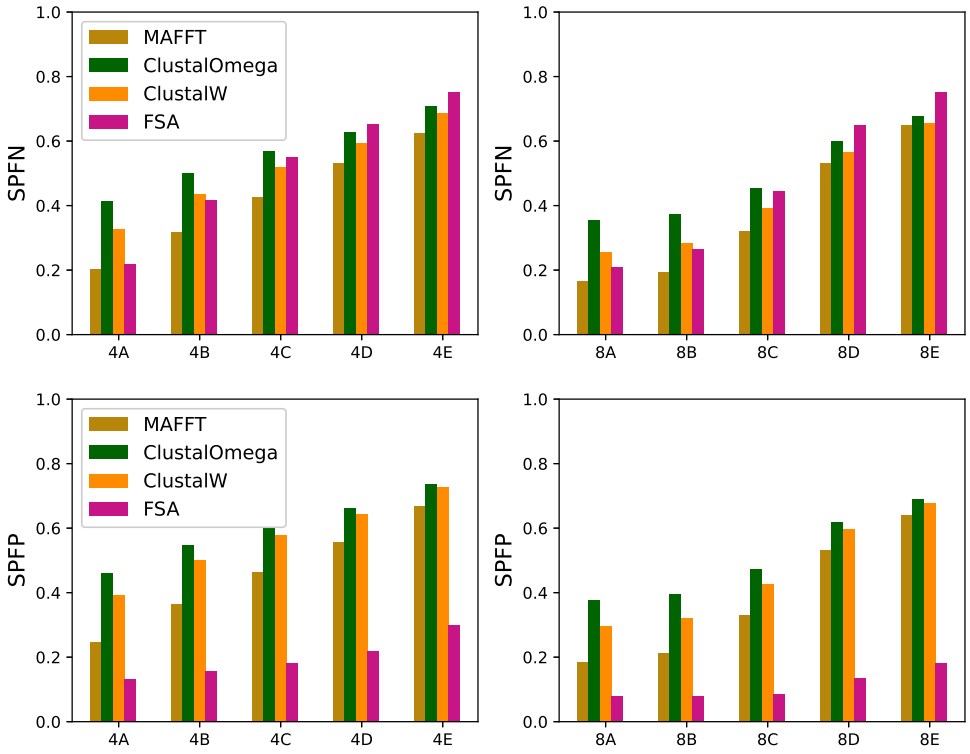

Figure 1: Simulation study: estimated MSA error. The MSA methods in our study consisted of MAFFT, Clustal Omega, ClustalW, and FSA. We assessed MSA estimation error based on type I and type II error: the former was assessed based on SP-FP proportion ("SPFP"), which is the proportion of nucleotide-nucleotide homologies that appear in the estimated alignment but not the true alignment, and the latter was assessed based on SP-FN proportion ("SPFN"), which is the proportion of nucleotide-nucleotide homologies that appear in the true alignment but not the estimated alignment. Average SPFN and SPFP are shown for each MSA method on each model condition ($n = 20$).

as multiple factors on the two most divergent 8-taxon model conditions in our study. The runtime increase is likely due to FSA's tendency to under-align input sequences, resulting in artificially high gappiness (and artificially low ANHD), and increased MSA length (i.e., number of MSA sites) inflates runtime. All other methods exhibited smaller increases in runtime as evolutionary divergence increased across model conditions. The impact of MSA error and gene tree error on network estimation runtime was also less pronounced for MSA methods other than FSA, with stronger differentiation between methods as evolutionary divergence increased (especially on more divergent 8-taxon model conditions). As expected, MPL-based analyses were faster than MLE analyses, with runtime difference of roughly an order of magnitude. Peak main memory usage was less than 600 MiB in most cases – well within the scope of modern PCs and other computing infrastructure. Relative differences in main memory usage were smaller than runtime comparisons as well.

3.1 Empirical Study

Mosquito dataset. We compared the topologies of estimated networks using different MSAs (including reference and estimated MSAs) and gene trees (Table 3). Single-reticulation phylogenetic network estimation using FastTree(ClustalW)-, FastTree(ClustalOmega)-, and FastTree(FSA)-estimated gene trees were topologically identical to estimation using reference MSAs and FastTree. In contrast, network analyses using FastTree(MAFFT)-estimated gene trees returned different topologies compared to analyses of the reference and other estimated MSAs. A similar outcome was observed for two-reticulation estimated networks, with one change: network estimation using FastTree(FSA)-estimated gene trees now returned different topologies compared to all other methods.

As network hypotheses become more complex (i.e., more reticulations were allowed in output networks), topological differences between the different methods increased as well. Estimation of 3- and 4-reticulation networks returned different topologies regardless of whether reference or estimated MSAs were used as input to gene tree estimation and species network estimation. No clear trend was observed in terms of MSA and gene tree estimation methods: estimated network topologies differed regardless of the methods under comparison. We note an important difference between the empirical study and simulation study: the reference MSAs used in the former are not the same as the true MSAs used in the latter,

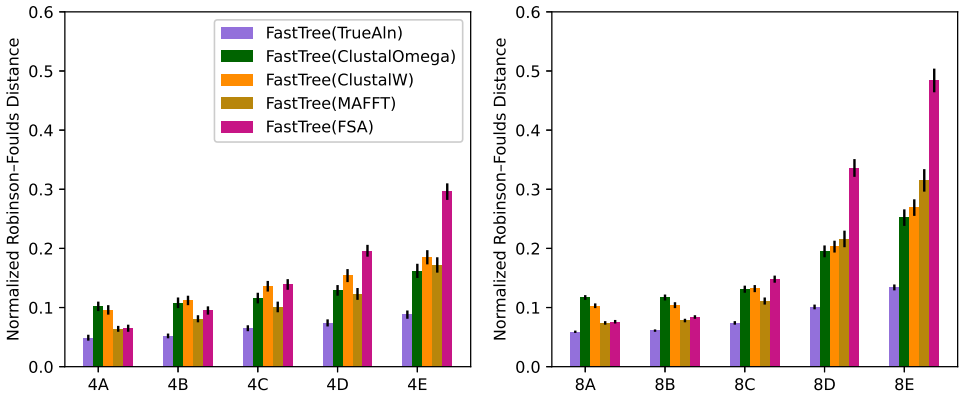

Figure 2: Simulation study: topological error of gene trees estimated using different MSAs. Topological error was assessed based on the normalized Robinson–Foulds distance between an estimated gene tree and the true gene tree. Gene trees were estimated using MLE analysis of five different input MSAs: (1) FastTree analysis of the true MSA ("FastTree(TrueAln)"), (2) FastTree analysis of a Clustal Omega-estimated MSA ("FastTree(ClustalOmega)"), (3) FastTree analysis of a ClustalW-estimated MSA ("FastTree(ClustalW)"), (4) FastTree analysis of a MAFFT-estimated MSA ("FastTree(MAFFT)"), or (5) FastTree analysis of an FSA-estimated MSA ("FastTree(FSA)"). Averages and standard error bars are shown for each method and model condition in the simulation study ($n = 20$).

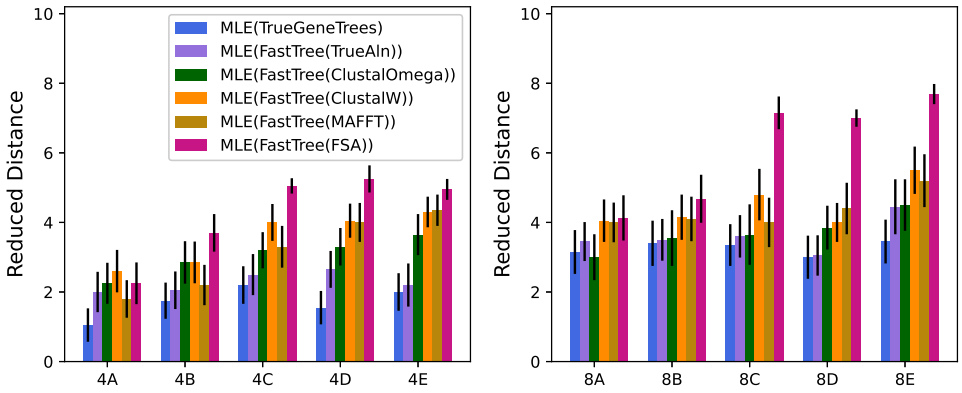

Figure 3: Simulation study: the impact of estimated MSA and gene tree error on topological error returned by MLE-based species network estimation. MLE was run on six different inputs: (1) true gene trees ("MLE(TrueGeneTrees)"), (2) gene trees estimated by FastTree analyses of true MSAs ("MLE(FastTree(TrueAln))"), (3) gene trees estimated by FastTree analyses of Clustal Omega-estimated MSAs ("MLE(FastTree(ClustalOmega))"), (4) gene trees estimated by FastTree analyses of ClustalW-estimated MSAs ("MLE(FastTree(ClustalW))"), (5) gene trees estimated by FastTree analyses of MAFFT-estimated MSAs ("MLE(FastTree(MAFFT))"), or (6) gene trees estimated by FastTree analyses of FSA-estimated MSAs ("MLE(FastTree(FSA))"). Topological error was measured using Nakhleh [24]'s equivalence-based network metric. Averages and standard error bars are shown for each method and model condition in the simulation study ($n = 20$).

since reference MSAs were estimated in part using computational approaches.

Yeast dataset. Similar outcomes were observed on the yeast dataset, with the exception that reference MSAs were not available. Topological differences were lowest for single-reticulation network estimations, and increased as network hypotheses became more complex in terms of the number of reticulations. Again, no clear trends were observed based on topological differences between a pair of estimates using different input MSAs and gene trees.

4 DISCUSSION

Our study demonstrated that MSA estimation error and gene tree estimation error can increase topological error of downstream species network estimation, and this effect becomes more pronounced as evolutionary divergence increases. We hypothesize that this finding is due in part to the tendency of MSA estimation methods to over-align, resulting in artificially elevated site divergence. The latter then contributes spurious signal for topologically incongruent gene trees and more divergent alleles originating via

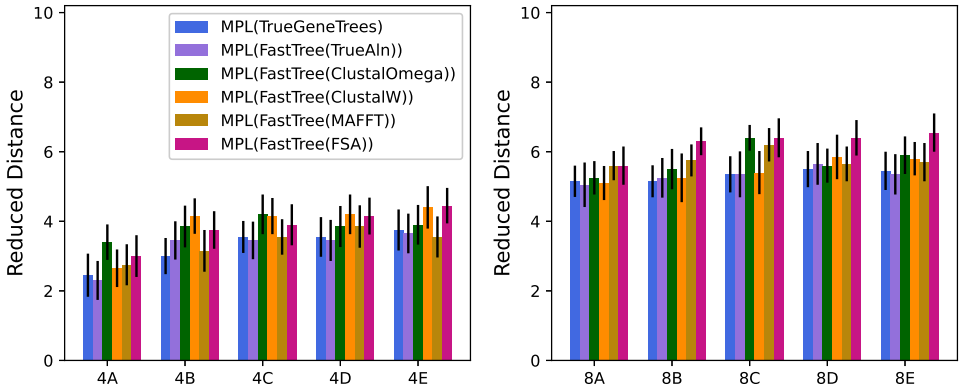

Figure 4: Simulation study: the impact of estimated MSA and gene tree error on topological error returned by MPL-based species network estimation. Figure description and layout are otherwise identical to Figure 3.

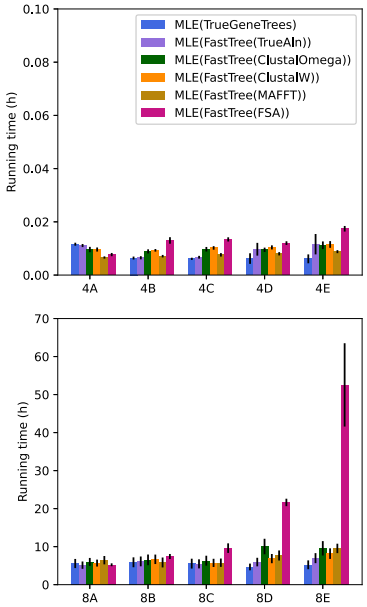

Figure 5: Simulation study: computational runtime requirements of MLE-based species network inference methods. Results on 4-taxon and 8-taxon model conditions are shown in top and bottom panels, respectively. Figure legend and layout are otherwise identical to Figure 3. Averages and standard error bars are shown for each method and model condition in the simulation study ($n = 20$).

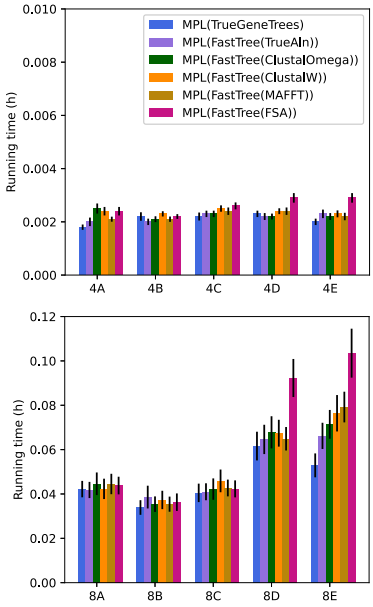

Figure 6: Simulation study: computational runtime requirements of MPL-based species network inference methods. Figure legend and layout are otherwise identical to Figure 5.

gene flow/reticulations during subsequent species network reconstruction. This findings mirrors Hagelberg et al. [8]'s correction, writ large. The simulation study experiments involving MAFFT-, Clustal Omega-, and ClustalW-estimated MSAs are largely consistent with this hypothesis. These methods estimated MSAs with consistently lower gappiness and higher ANHD as compared to true alignments. FSA had an opposite bias: it tended to return much gappier estimated MSAs compared with the other two MSA methods, resulting in typically lower site divergence. Under the finite-sites substitution models used for gene tree MLE in our study (and much of traditional phylogenetics), indels are treated either as missing data or an additional base. We note that phylogenetic tree MLE is not statistically consistent under substitution models where indels are treated as missing data or an additional base [37]. We suspect that this particular case of model mis-specification has the effect of underestimating substitution events, with corresponding impacts on downstream gene tree and species network inference error.

Table 2: Simulation study: estimated multiple sequence alignment (MSA) summary statistics. The model conditions are the same as in Table 1. Average normalized Hamming distance ("ANHD"), the percentage of estimated MSA cells that consist of indels ("Gappiness"), and the number of estimated MSA sites ("Length") are reported as an average for each model condition and MSA estimation method. (n = 20).

Model condition	MAFFT ANHD	MAFFT Gappiness	MAFFT Length
4.A	0.399	0.213	1274.2
4.B	0.464	0.217	1279.5
4.C	0.504	0.222	1287.6
4.D	0.537	0.226	1295.1
4.E	0.566	0.241	1321.1
8.A	0.373	0.234	1308.9
8.B	0.397	0.233	1306.1
8.C	0.465	0.238	1315.6
8.D	0.527	0.273	1380.8
8.E	0.554	0.291	1418.8
Model condition	**Clustal Omega ANHD**	**Clustal Omega Gappiness**	**Clustal Omega Length**
4.A	0.488	0.147	1174.7
4.B	0.533	0.147	1173.8
4.C	0.560	0.146	1172.2
4.D	0.582	0.143	1167.7
4.E	0.605	0.144	1169.9
8.A	0.462	0.161	1193.6
8.B	0.479	0.159	1190.9
8.C	0.529	0.155	1185.5
8.D	0.584	0.160	1191.9
8.E	0.608	0.159	1190.9
Model condition	**ClustalW ANHD**	**ClustalW Gappiness**	**ClustalW Length**
4.A	0.414	0.139	1163.6
4.B	0.477	0.128	1148.2
4.C	0.516	0.119	1136.4
4.D	0.549	0.108	1122.3
4.E	0.580	0.100	1112.4
8.A	0.397	0.161	1194.1
8.B	0.420	0.156	1185.5
8.C	0.492	0.137	1159.4
8.D	0.563	0.121	1138.6
8.E	0.596	0.107	1120.4
Model condition	**FSA ANHD**	**FSA Gappiness**	**FSA Length**
4.A	0.426	0.364	1598.6
4.B	0.493	0.512	2101.4
4.C	0.508	0.586	2481.9
4.D	0.509	0.633	2805.4
4.E	0.502	0.673	3147.5
8.A	0.364	0.458	1891.4
8.B	0.376	0.515	2109.9
8.C	0.391	0.634	2859.0
8.D	0.374	0.730	3941.5
8.E	0.356	0.769	4581.6

Table 3: Empirical study: topological comparison of different MLE-based species network estimation methods on the mosquito dataset. The MSAs consisted of either reference MSAs from the original study of Fontaine et al. [7] or MSAs estimated using ClustalW, MAFFT, or FSA. Gene trees were estimated on MSAs using FastTree, and the resulting gene trees were used as input to MLE. The method abbreviations "M(F(Clo))", "M(F(Clu))", "M(F(MAF))", "M(F(FSA))", and "M(F(Ref))" refer to MLE analyses of gene trees estimated using FastTree analyses of Clustal Omega, ClustalW, MAFFT, FSA, and reference MSAs, respectively. We also compared estimation of species networks with differing complexity, where MLE was used to estimate a species network with at most 1, 2, 3, or 4 reticulations. Topological distances between a pair of estimated networks were measured using Nakhleh [24]'s equivalence-based network metric. Only upper triangular entries in the pairwise distance matrix are shown.

1 reticulation	M(F(Clo))	M(F(Clu))	M(F(MAF))	M(F(FSA))	M(F(Ref))
M(F(Clo))	-	0	6	0	0
M(F(Clu))		-	6	0	0
M(F(MAF))			-	6	6
M(F(FSA))				-	0
M(F(Ref))					-
2 reticulations	M(F(Clo))	M(F(Clu))	M(F(MAF))	M(F(FSA))	M(F(Ref))
M(F((Clo))	-	0	4	7	0
M(F((Clu))		-	4	7	0
M(F((MAF))			-	7	4
M(F((FSA))				-	7
M(F((Ref)					-
3 reticulations	M(F(Clo))	M(F(Clu))	M(F(MAF))	M(F(FSA))	M(F(Ref))
M(F((Clo))	-	8	10	10	10
M(F((Clu))		-	10	10	10
M(F((MAF))			-	10	4
M(F((FSA))				-	9
M(F((Ref))					-
4 reticulations	M(F(Clo))	M(F(Clu))	M(F(MAF))	M(F(FSA))	M(F(Ref))
M(F((Clo))	-	11	12	13	11
M(F((Clu))		-	11	11	11
M(F((MAF))			-	14	12
M(F((FSA))				-	11
M(F((Ref))					-

Our experiments revealed another important practical point: MSA accuracy also had a secondary impact on the computational requirements of statistical phylogenetic network reconstruction. On more divergent model conditions, we observed an increase in species network estimation runtimes, especially those involving less accurate MSAs and gene trees. We attribute this finding to greater topological discordance among the inputs provided to summary-based species network inference. In reality, the dataset sizes and divergences in our study are modest compared to recent phylogenomic studies, where analyses of many dozens of genomic sequences or more are increasingly commonplace. The observed runtime and main memory requirements of the estimation methods under study were commensurate with small dataset scales: consistently less than 6 minutes and 1 GiB. But the trend towards increasing runtime as dataset scales grew suggest that MSA error

Table 4: Empirical study: topological comparison of different MLE-based species network estimation methods on the yeast dataset. MSAs were estimated using ClustalW, MAFFT, or FSA. Gene trees were estimated using FastTree, and species networks were estimated using MLE. The method abbreviations "M(F(Clu))", "M(F(MAF))", and "M(F(FSA))", refer to MLE analyses of gene trees estimated using FastTree analyses of ClustalW, MAFFT, and FSA MSAs, respectively. Table description and layout are otherwise identical to Table 3.

1 reticulation			
	M(F(Clu))	M(F(MAF))	M(F(FSA))
M(F(Clu))	-	5	8
M(F(MAF))		-	4
M(F(FSA))			-
2 reticulations			
	M(F(Clu))	M(F(MAF))	M(F(FSA))
M(F(Clu))	-	13	15
M(F(MAF))		-	15
M(F(FSA))			-
3 reticulations			
	M(F(Clu))	M(F(MAF))	M(F(FSA))
M(F(Clu))	-	15	20
M(F(MAF))		-	22
M(F(FSA))			-
4 reticulations			
	M(F(Clu))	M(F(MAF))	M(F(FSA))
M(F(Clu))	-	16	18
M(F(MAF))		-	21
M(F(FSA))			-

and gene tree error will have non-negligible effects on species network estimation runtime on WGS datasets that are much larger and/or more divergent than those considered in our study. And we note that in no realistic scenario involving empirical data should we expect to have access to MSA quality approaching ground truth. The resulting scalability challenge represents a road-block to wide uptake of statistical network inference by the systematics research community.

Here, the scalability issue prevented us from exploring larger dataset sizes, which are now becoming commonplace in today's post-genomic era. Nevertheless, our results suggest that the impact of MSA accuracy and gene tree accuracy will become even more pronounced as the scope of modern phylogenomic and phylogenetic studies grows in terms of number of taxa, number of sampled genomes, and evolutionary divergence. The scalability study of [11] provides a theoretical baseline: in experiments with 10-taxon datasets, MLE analyses using ground truth inputs (i.e., true MSAs and true gene trees) required more than a week of computational runtime and returned greater topological error than experiments on smaller datasets. Our findings suggest that MSA estimation error and gene tree estimation error will only amplify the impact of dataset scale on species network estimation error and runtime requirements.

Our study also underscores the need to move beyond performance studies of summary-based phylogenetic inference methods where experimental inputs primarily consist of true gene trees and/or true MSAs. Rather, future studies and algorithmic development efforts need to focus on the exact opposite setting: experiments where inputs do not include ground truth and consist only of estimated reconstructions for all upstream analysis tasks. We note that gene tree estimation errors in our study are actually fairly low. We observed average topological error of around 5-25% for the most accurate method on each model condition, which is well within the range seen in prior performance studies of gene tree estimation methods [20, 21, 23].

5 CONCLUSIONS

Our simulation study clearly demonstrated the impact of MSA estimation error and gene tree estimation error on downstream phylogenetic network inference and learning. The impact on both topological accuracy and computational runtime became more pronounced as evolutionary divergence increased. Topological comparisons in our empirical study were consistent with these findings, and were also influenced by network hypothesis complexity. Our findings point to the need to account for MSA error and gene tree error in phylogenetic network analyses, as well as the need to create new computational methodologies for scalable joint estimation of MSAs, gene trees, and a species network. In particular, mis-aligned MSA sites with artificially elevated site divergence can present spurious signal of gene tree incongruence and non-tree-like evolution, and species phylogeny reconstructions based on sequence homology information must factor in underlying uncertainty in upstream estimation tasks.

We conclude with comments on future work. New methods are needed for reconstructing MSAs, gene trees, and a species network from unaligned sequence inputs, but a formidable challenge remains: namely, scalability on ever larger and more divergent genomic datasets. Current network estimation methods are already computationally intensive relative to phylogenetic tree estimation methods. And yet, algorithmic enhancements to improve computational efficiency must not compromise the ability to extract sufficient phylogenetic signal from noisy data. One solution is to utilize phylogenetic divide-and-conquer to boost both estimation accuracy and computational efficiency. For example, FastNet [12], our previously developed phylogenetic divide-and-conquer framework, can be readily adapted to this task.

ACKNOWLEDGMENTS

We gratefully acknowledge the three anonymous reviewers of this manuscript. Their constructive feedback helped us to revise the study and manuscript. We would also like to acknowledge the support of the National Science Foundation (2144121, 1714417, 1737898, 1740874 to KJL) and MSU (faculty startup to KJL). All computational experiments were performed on the MSU High Performance Computing Center, which is part of the MSU Institute for Cyber-Enabled Research.

REFERENCES

[1] Richard J Abbott and Loren H Rieseberg. 2012. Hybrid Speciation. In *Encyclopaedia of Life Sciences*. John Wiley & Sons, Ltd, Hoboken, NJ, USA.

[2] Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 15, 4 (2019), e1006650.

[3] 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 (2009), e1000392.

[4] J.J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Muller, N. Pande, Z. Shang, N. Yu, and R.R. Gutell. 2002. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron and Other RNAs. *BMC Bioinformatics* 3, 15 (2002). http://www.rna.ccbb.utexas.edu.

[5] Joseph Felsenstein. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *Journal of Molecular Evolution* 53, 4-5 (2001), 447–455.

[6] William Fletcher and Ziheng Yang. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution* 26, 8 (2009), 1879–1888.

[7] Michael C. Fontaine, James B. Pease, Aaron Steele, Robert M. Waterhouse, Daniel E. Neafsey, Igor V. Sharakhov, Xiaofang Jiang, Andrew B. Hall, Flaminia Catteruccia, Evdoxia Kakani, Sara N. Mitchell, Yi-Chieh Wu, Hilary A. Smith, R. Rebecca Love, Mara K. Lawniczak, Michel A. Slotman, Scott J. Emrich, Matthew W. Hahn, and Nora J. Besansky. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347, 6217 (2015), 1258522.

[8] E Hagelberg, N Goldman, P Lio, Schiefenhovel Whelan, W Schiefenhöel, JB Clegg, and Donald Keith Bowden. 1999. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 266, 1418 (1999), 485–492.

[9] E Hagelberg, N Goldman, P Lio, S Whelan, W Schiefenhövel, JB Clegg, and DK Bowden. 2000. Evidence for mitochondrial DNA recombination in a human population of island Melanesia: correction. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267, 1452 (2000), 1595–1596.

[10] Jotun Hein, Mikkel Schierup, and Carsten Wiuf. 2004. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory.* Oxford University Press, Oxford.

[11] Hussein A Hejase and Kevin J Liu. 2016. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics* 17, 1 (2016), 422.

[12] 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*, Mathieu Blanchette and Aïda Ouangraoua (Eds.). Springer International Publishing, Cham, 242–259.

[13] Richard R. Hudson. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 2 (2002), 337–338.

[14] Daniel H Huson, Regula Rupp, and Celine Scornavacca. 2010. *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge University Press, Cambridge, United Kingdom.

[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 (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 (2013), 772–780.

[17] Patrick J Keeling and Jeffrey D Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9, 8 (2008), 605–618.

[18] Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. 2007. Clustal W and Clustal X version 2.0. *bioinformatics* 23, 21 (2007), 2947–2948.

[19] 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. ClustalW and ClustalX version 2.0. *Bioinformatics* 23, 21 (November 2007), 2947–2948.

[20] K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. 2009. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science* 324, 5934 (2009), 1561–1564.

[21] 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), 90–106.

[22] James O McInerney, James A Cotton, and Davide Pisani. 2008. The prokaryotic tree of life: past, present... and future? *Trends in Ecology & Evolution* 23, 5 (2008), 276–281.

[23] 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.

[24] Luay Nakhleh. 2009. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 7, 2 (2009), 218–222.

[25] Luay Nakhleh. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution* 28, 12 (2013), 719–728.

[26] Luay Nakhleh, Bernard ME Moret, Usman Roshan, Katherine St. John, Jerry Sun, and Tandy Warnow. 2001. The accuracy of fast phylogenetic methods for large datasets. In *Biocomputing 2002*. World Scientific, 211–222.

[27] Masatoshi Nei, Ranajit Chakraborty, and Paul A Fuerst. 1976. Infinite allele model with varying mutation rate. *Proceedings of the National Academy of Sciences* 73, 11 (1976), 4164–4168.

[28] S. Nelesen, K. Liu, D. Zhao, C. R. Linder, and T. Warnow. 2008. The Effect of the Guide Tree on Multiple Sequence Alignments and Subsequent Phylogenetic Analyses. In *Pacific Symposium on Biocomputing*, Vol. 13. 15–24.

[29] M. Price, P. Dehal, and A. Arkin. 2010. FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5, 3 (March 2010), e9490.

[30] D.F. Robinson and L.R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53 (1981), 131–147.

[31] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* 142 (1990), 485–501.

[32] Leonidas Salichos and Antonis Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497, 7449 (16 May 2013), 327–331.

[33] M. J. Sanderson. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 2 (2003), 301–302.

[34] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7, 1 (2011), 539.

[35] Claudia Solís-Lemus and Cécile Ané. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. *PLoS Genetics* 12, 3 (03 2016), 1–21.

[36] Cuong Than, Derek Ruths, and Luay Nakhleh. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9, 1 (2008), 322.

[37] Tandy Warnow. 2012. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Currents* 4 (2012).

[38] Dingqiao Wen, Yun Yu, Matthew W Hahn, and Luay Nakhleh. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology* 25, 11 (2016), 2361–2372.

[39] Dingqiao Wen, Yun Yu, and Luay Nakhleh. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genetics* 12, 5 (2016), e1006006.

[40] Dingqiao Wen, Yun Yu, Jiafan Zhu, and Luay Nakhleh. 2018. Inferring phylogenetic networks using PhyloNet. *Systematic Biology* 67, 4 (2018), 735–740.

[41] Yun Yu, James H Degnan, and Luay Nakhleh. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics* 8, 4 (2012), e1002660.

[42] Yun Yu, Jianrong Dong, Kevin J. Liu, and Luay Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences* 111, 46 (2014), 16448–16453.

[43] Yun Yu and Luay Nakhleh. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16, Suppl 10 (2015), S10.

[44] Yun Yu, Cuong Than, James H Degnan, and Luay Nakhleh. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology* 60, 2 (2011), 138–149.

[45] Chi Zhang, Huw A Ogilvie, Alexei J Drummond, and Tanja Stadler. 2018. Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution* 35, 2 (2018), 504–517.

Supplementary Appendix: The Impact of Gene Sequence Alignment and Gene Tree Estimation Error on Summary-Based Species Network Estimation

Meijun Gao[1], Wei Wang[4], and Kevin J. Liu[1,2,3]

[1] Department of Computer Science and Engineering
[2] Ecology, Evolution, and Behavior Program
[3] Genetics and Genome Sciences Program
Michigan State University
East Lansing, MI, USA
kjl@msu.edu
[4] Meta Corporation, Menlo Park, CA, USA

1 Supplementary Materials and Methods

Figure S1 shows example visualizations of model species networks with 4 and 8 taxa from our simulation study.

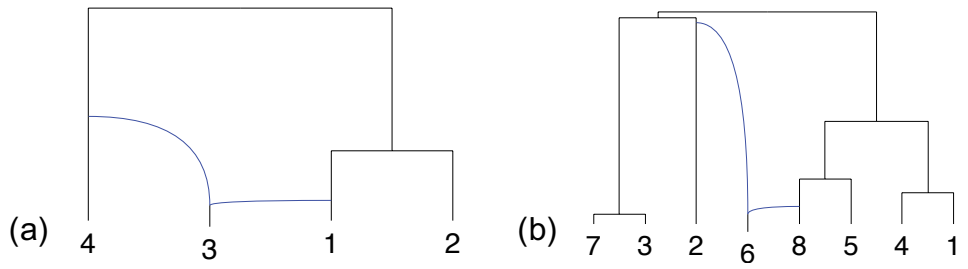

Fig. S1. Examples of model networks in our simulation study. (a) A 4-taxon model network from the first replicate of the 4.A model condition is shown. Branch lengths in coalescent units are visualized to scale. (b) An 8-taxon model network from the first replicate of the 8.A model condition is shown similarly.

Software commands used in our performance study are listed below.

1) The following r8s [9] script was used to simulate random birth-death model trees:

```
begin rates;
simulate diversemodel=bdback seed=<random seed> nreps=20 ntaxa=<n> T=0;
describe tree=0 plot=chrono_description;
end;
```

2) MSC+IM simulations were run using the following ms [5] command:

```
ms <n> 1000
   -T -I <n> <s_1, s_2, ... s_n>
   -ej <t_D> <i_D> <j_D> ...
   -em <t_1> <i> <j> 5.0
   -em <t_2> <i> <j> 0
```

The command options are as follows. The `-T` option is used to output sampled local gene trees. The `-I <n> <s_1 s_2 ... s_n>` option specifies n structured populations corresponding to n sampled taxa

where the ith taxon is represented by a single allele from the ith structured population by setting $s_i = 1$. In our simulations, we sampled one allele per taxon. Each `-ej <t_0> <i_0> <j_0>` option encodes an ancestral divergence event at time time t_0 that resulted in descendant populations i_0 and j_0. The `-em <t_1> <i> <j> 5.0 -em <t_2> <i> <j> 0` options specify a unidirectional migration event from population i to j that spans the time interval from t_1 to t_2.

3) INDELible [3] version 1.03 was run using the following command and configuration file:

```
./indelible

Parameters are defined in control.txt.
Example control.txt file:

[TYPE] NUCLEOTIDE 1
[MODEL] GTRexample
 [submodel] GTR 1.26195738509
                0.140055369456
                0.287783034615
                0.35766826674
                0.308267431018
                // GTR: a=TtoC, b=TtoA,
                // c=TtoG, d=CtoA, e=CtoG, f=AtoG=1
 [statefreq] 0.311475 0.191363 0.300414 0.196748
             // pi_T=0.1, pi_C=0.2, pi_A=0.3, pi_G=0.4
 [indelmodel] USER medium_gap.txt
 [indelrate] <indelRate>
[TREE] tree1
 <geneTree>
 [treedepth] <treeDepth>
[PARTITIONS] pGTR [tree1 GTRexample 1000]
             // tree 1, model GTRexample, root length of 1000
[SETTINGS]
 [output] FASTA // FASTA, NEXUS, PHYLIP or PHYLIPT
[EVOLVE]
 pGTR 1 GTRout
```

4) MAFFT [6] version 7.222 was run using the following command:

```
mafft <seqFile> > <estiAlnFile>
```

5) Clustal Omega [10] version 1.2.4 was run using the following command:

```
clustalo -i <seqFile> -o <estiAlnFile> --outfmt=fasta"
```

6) Clustalw [7] version 2.1 was run using the following command:

```
clustalw2 -INFILE=<seqFile> -ALIGN -TYPE=dna -outfile=<estiAlnFile> -output=FASTA
```

7) FSA [1] version 1.15.9 was run using the following command

```
fsa <seqFile> > <estiAlnFile>
```

8) PhyloNet [11, 12] version 3.6.0 was run using the following command and configuration file:

```
java -jar PhyloNet_3.6.0.jar <nexFile>  > <resultFile>

Example nexfile
#NEXUS
```

```
BEGIN TREES;
<rootedGeneTrees>
END;
BEGIN PHYLONET;
InferNetwork_ML (all) 1 -bl; #MLE-length
 InferNetwork_ML (all) 1; #MLE
 InferNetwork_MPL (all) 1; #MPL
 InferNetwork_MP (all) 1; #MP
END;
```

9) FastTree [8] version 2.1.11 was run using the following command:

```
FastTree -nt -nosupport -gtr < <estiAlnFile> > <infGeneTreeFile>
```

2 Supplementary Results

PhyloNet [11, 12] can also perform summary-based network inference under MDC criterion using parsimony-based method. We refer it as MDC in the following. Figure S2 shows the MDC results using true gene trees and estimated gene trees inferred from true MSAs and estimated MSAs. Compared with inference performance under MLE and MPL criteria, MDC analysis for all 8-taxon model conditions returned higher topological error. For 4-taxon model conditions, some of them returned higher, lower or comparable topological error. Different from the inference results of MLE and MDC, for each model condition, MDC using all different gene trees inferred from estimated MSAs returned comparable topological error and costed comparable running time, since the simple optimization criterion.

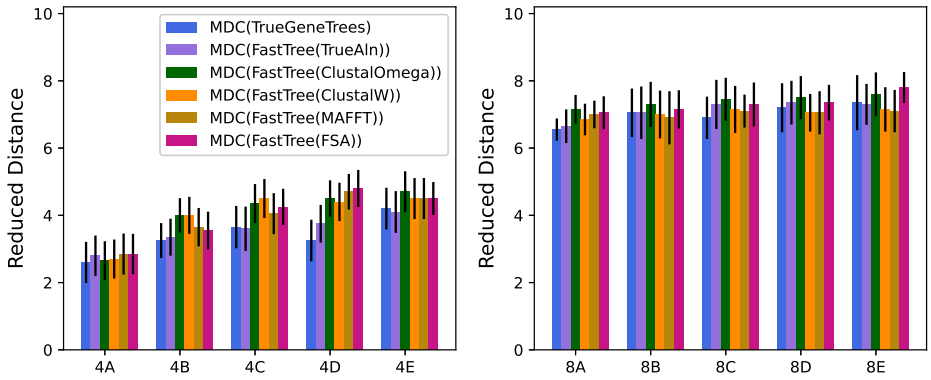

Fig. S2. Simulation study: the impact of estimated MSA and gene tree error on topological error returned by downstream MDC species network estimation. Figure description and layout are otherwise identical to Figure 2. Averages and standard error bars are shown for each model condition in the simulation study ($n = 20$).

We also investigated two other MSA methods that utilize a different approach compared to the other MSA methods in our performance study: MUSCLE [2] and POA [4]. Table S1 and S4 show summary statistics and MSA error for MUSCLE- and POA-estimated MSAs. Figure S5 shows topological error of MLE gene trees estimated on MUSCLE- and POA-estimated alignments. As with the other MSA methods under study, both the MSA error and topological distance between true and estimated gene trees increase with increasing species number and evolutionary divergence. As shown in Figure S6, network analyses using FastTree(POA)-estimated gene trees returned worse topological error compared to species network estimation using true MSAs and all other estimated MSAs, with a single exception: on model conditions 8.C and 8.D, species network analyses using FastTree(FSA)-estimated gene trees returned the worst topological error, followed by

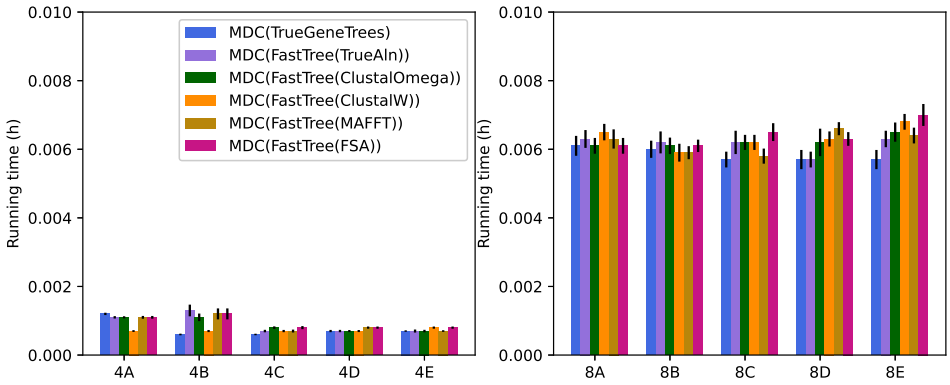

Fig. S3. Simulation study: computational runtime requirements of MDC summary-based species network inference methods. Figure legend and layout are identical to Figure 2. Averages and standard error bars are shown for each model condition in the simulation study ($n = 20$).

the FastTree(POA)-based analyses. Species network analyses using FastTree(MUSCLE)-estimated gene trees returned better topological error than FastTree(POA)-based species network estimation, but had comparable or typically worse error compared to the other species network methods under study.

Table S1. Simulation study: summary statistics for MUSCLE- and POA-estimated MSAs. See Table 1 in the main manuscript for a complete description of model conditions and summary statistics. Average ANHD, gappiness, and MSA length are reported as an average for each model condition and MSA estimation method ($n = 20$).

Model	MUSCLE			POA		
condition	ANHD	Gappiness	Length	ANHD	Gappiness	Length
4.A	0.398	0.197	1249.5	0.391	0.172	1210.2
4.B	0.460	0.204	1257.6	0.452	0.171	1207.5
4.C	0.497	0.209	1267.6	0.485	0.171	1207.1
4.D	0.529	0.212	1272.6	0.511	0.169	1205.1
4.E	0.559	0.223	1289.5	0.531	0.171	1208.2
8.A	0.364	0.239	1319.3	0.388	0.199	1251.6
8.B	0.384	0.240	1320.2	0.411	0.199	1250.1
8.C	0.443	0.252	1342.3	0.475	0.201	1253.3
8.D	0.497	0.286	1407.5	0.527	0.213	1272.1
8.E	0.523	0.298	1433.3	0.545	0.216	1277.6

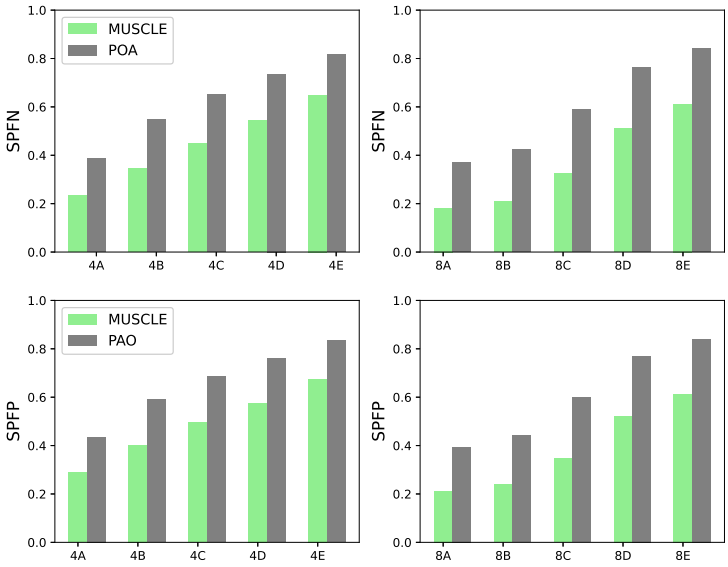

Fig. S4. Simulation study: estimated error of MUSCLE- and POA-estimated MSAs. Average SPFN and SPFP are shown for each MSA method on each model condition ($n = 20$).

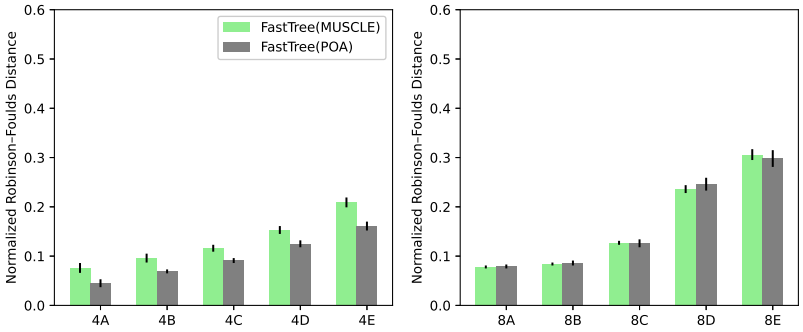

Fig. S5. Simulation study: topological error of gene trees estimated using MUSCLE- and POA-estimated MSAs. Topological error was assessed based on the normalized Robinson–Foulds distance between an estimated gene tree and the true gene tree. Gene trees were estimated using MLE analysis of two different input MSAs: (1) FastTree analysis of a MUSCLE-estimated MSA ("FastTree(MUSCLE)"), (2) FastTree analysis of a POA-estimated MSA ("FastTree(POA)"), Averages and standard error bars are shown for each method and model condition in the simulation study ($n = 20$).

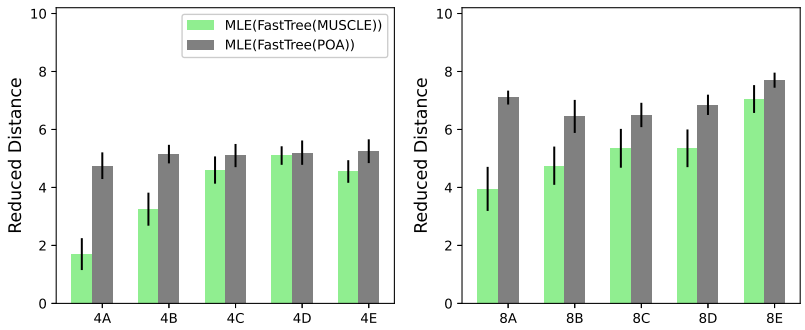

Fig. S6. Simulation study: the impact of MUSCLE- and POA-estimated MSA and gene tree error on topological error returned by MLE-based species network estimation. MLE was run on two different inputs: (1) gene trees estimated by FastTree analyses of MUSCLE-estimated MSAs ("MLE(FastTree(MUSCLE))"), (2) gene trees estimated by FastTree analyses of POA-estimated MSAs ("MLE(FastTree(POA))"). Averages and standard error bars are shown for each model condition in the simulation study ($n = 20$).

Bibliography

[1] Bradley, R.K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., Pachter, L.: Fast statistical alignment. PLoS Computational Biology **5**(5), e1000392 (2009)

[2] Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**(5), 1792–1797 (2004)

[3] Fletcher, W., Yang, Z.: INDELible: a flexible simulator of biological sequence evolution. Molecular Biology and Evolution **26**(8), 1879–1888 (2009)

[4] Grasso, C., Lee, C.: Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics **20**(10), 1546–1556 (2004)

[5] Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**(2), 337–338 (2002).

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

[7] Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al.: Clustal W and Clustal X version 2.0. Bioinformatics **23**(21), 2947–2948 (2007)

[8] Price, M., Dehal, P., Arkin, A.: FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE **5**(3), e9490 (March 2010).

[9] Sanderson, M.J.: r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**(2), 301–302 (2003)

[10] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology **7**(1), 539 (2011)

[11] Than, C., Ruths, D., Nakhleh, L.: PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics **9**(1), 322 (2008)

[12] Wen, D., Yu, Y., Zhu, J., Nakhleh, L.: Inferring phylogenetic networks using PhyloNet. Systematic Biology **67**(4), 735–740 (2018)