Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

FastNet: Fast and accurate inference of phylogenetic networks using large-scale genomic sequence data

Hussein A Hejase [1], Natalie VandePol [2], Gregory A Bonito [2] and Kevin J Liu [1,*]

[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, 48824, MI, USA and
[2]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, 48824, MI, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Advances in next-generation sequencing technologies and phylogenomics have reshaped our understanding of evolutionary biology. One primary outcome is the emerging discovery that interspecific gene flow has played a major role in the evolution of many different organisms across the Tree of Life. To what extent is the Tree of Life not truly a tree reflecting strict "vertical" divergence, but rather a more general graph structure known as a phylogenetic network which also captures "horizontal" gene flow?
Results: The answer to this fundamental question not only depends upon densely sampled and divergent genomic sequence data, but also computational methods which are capable of accurately and efficiently inferring phylogenetic networks from large-scale genomic sequence datasets. Recent methodological advances have attempted to address this gap. However, in a recent performance study, we demonstrated that the state of the art falls well short of the scalability requirements of existing phylogenomic studies.
The methodological gap remains: how can phylogenetic networks be accurately and efficiently inferred using genomic sequence data involving many dozens or hundreds of taxa? In this study, we address this gap by proposing a new phylogenetic divide-and-conquer method which we call FastNet. Using synthetic and empirical data spanning a range of evolutionary scenarios, we demonstrate that FastNet outperforms state-of-the-art methods in terms of computational efficiency and topological accuracy.
We predict an imminent need for new computational methodologies that can cope with dataset scale at the next order of magnitude, involving thousands of genomes or more. We consider FastNet to be a next step in this direction. We conclude with thoughts on the way forward through future algorithmic enhancements.
Keywords: phylogenomics; phylogenetic network; gene flow; inference; divide and conquer; simulation study; computational runtime; topological accuracy
Contact: kjl@msu.edu
Supplementary information: Supplementary data are available at https://gitlab.msu.edu/liulab/FastNet.data.scripts.

1 Introduction

Recent advances in biomolecular sequencing (Metzker, 2010) and phylogenomic modeling and inference (Edwards, 2009; Nakhleh, 2013) have revealed that interspecific gene flow has played a major role in the evolution of many different organisms across the Tree of Life (McInerney *et al.*, 2008; Keeling and Palmer, 2008; Abbott and Rieseberg, 2012), including humans and ancient hominins (Green *et al.*, 2010; Reich *et al.*, 2010), butterflies (The Heliconious Genome Consortium, 2012), and mice (Liu *et al.*, 2015). These findings point to new directions for phylogenetics and phylogenomics: to what extent is the Tree of Life not truly a tree reflecting strict vertical divergence, but rather a more general graph structure known as a phylogenetic network where reticulation edges and nodes capture gene flow? And what is the evolutionary role of gene

flow? In addition to densely sampled and divergent genomic sequence data, one additional ingredient is critically needed to make progress on these questions: computational methods which are capable of accurately and efficiently inferring phylogenetic networks on large-scale genomic sequence datasets.

Recent methodological advances have attempted to address this gap. In particular, Solís-Lemus and Ané proposed SNaQ (Solís-Lemus and Ané, 2016), a new statistical method which seeks to address the computational efficiency of species network inference using a pseudo-likelihood approximation, and the method of Yu and Nakhleh (2015) (referred here as MPL, which stands for maximum pseudo-likelihood) substitutes pseudo-likelihoods in the optimization criterion used by MLE (which stands for maximum likelihood estimation). We recently conducted a performance study which demonstrated the scalability limits of SNaQ, MPL, MLE, and other state-of-the-art phylogenetic methods in the context of phylogenetic network inference (Hejase and Liu, 2016). The scalability of the state of the art falls well short of that required by current phylogenetic studies, where many dozens or hundreds of divergent genomic sequences are common (Nakhleh, 2013). The most accurate phylogenetic network inference methods performed statistical inference under phylogenomic models (Yu *et al.*, 2014a, 2012; Solís-Lemus and Ané, 2016) that extended the multi-species coalescent model (Kingman, 1982; Hein *et al.*, 2004). SNaQ was among the fastest of these methods while the probabilistic multi-locus inference method of Yu *et al.* (2014b) that performs full likelihood calculations was the most accurate. None of the aforementioned statistical phylogenomic inference methods completed analyses of datasets with 30 taxa or more after many weeks of CPU runtime. The other methods fell into two categories: split-based methods (Bandelt and Dress, 1992; Bryant and Moulton, 2004) and the parsimony-based inference method of Yu *et al.* (Yu *et al.*, 2011) (which we refer to as MP in this study). Both classes of methods were faster than the statistical phylogenomic inference methods but less accurate.

The methodological gap remains: how can phylogenetic networks be accurately and efficiently inferred using genomic sequence data involving many dozens or hundreds of taxa? In this study, we address this question and propose a new method for this problem. We investigate this question in the context of two constraints. First, we focus on dataset size in terms of the number of taxa in the species phylogeny. We note that scalability issues arise due to other dataset features as well, including population-scale allele sampling for each taxon in a study and sequence divergence. Second, we follow a trend in emerging methodologies and performance studies (Davidson *et al.*, 2015; Solís-Lemus and Ané, 2016; Leaché *et al.*, 2013) and focus our attention on a specific classes of gene flow events. We focus on paraphyletic and monophyletic gene flow, which have been the subject of several recent high-profile studies (Liu *et al.*, 2015; The Heliconious Genome Consortium, 2012), and ancestral gene flow deeper in the species phylogeny.

2 Approach

One path forward is through the use of divide-and-conquer. The general idea behind divide-and-conquer is to split the full problem into smaller and more closely related subproblems, analyze the subproblems using state-of-the-art phylogenetic network inference methods, and then merge solutions on the subproblems into a solution on the full problem. Viewed this way, divide-and-conquer can be seen as a computational framework that "boosts" the scalability of existing methods (and which is distinct from boosting in the context of machine learning). The advantages of analyzing smaller and more closely related subproblems are two-fold. First, smaller subproblems present more reasonable computational requirements compared to the full problem. Second, the evolutionary divergence of taxa in a subproblem is reduced compared to the full set of taxa, which has been shown to improve accuracy for phylogenetic tree inference (Huelsenbeck and Hillis, 1993; Felsenstein, 1978; Liu *et al.*, 2009). We and our collaborators have successfully applied divide-and-conquer approaches to enable scalable inference in the context of species tree estimation (Liu *et al.*, 2009, 2012; Mirarab *et al.*, 2015).

Here, we consider the more general problem of inferring species phylogenies that are directed phylogenetic networks. A directed phylogenetic network $N = (V, E)$ consists of a set of nodes V and a set of directed edges E. The set of nodes V consists of a root node $r(N)$ with in-degree 0 and out-degree 2, leaves $\mathcal{L}(N)$ with in-degree 1 and out-degree 0, tree nodes with in-degree 1 and out-degree 2, and reticulation nodes with in-degree 2 and out-degree 1. A directed edge $(u, v) \in E$ is a tree edge if and only if v is a tree node, and is otherwise a reticulation edge. The edges in a network N can be labeled by a set of branch lengths ℓ. A directed phylogenetic tree is a special case of a directed phylogenetic network which contains no reticulation nodes (and edges). An unrooted topology can be obtained from a directed tree by ignoring edge directionality.

The phylogenetic network inference problem consists of the following. One input is a partitioned multiple sequence alignment $\boldsymbol{A}$ containing data partitions a_i for $1 \leq i \leq k$, where each partition corresponds to the sequence data for one of k genomic loci. Each of the n rows in the alignment $\boldsymbol{A}$ is a sample representing taxon $x \in X$, and each taxon is represented by one or more samples. Similar to other approaches (Yu *et al.*, 2014a; Solís-Lemus and Ané, 2016), we also require an input parameter c_r which specifies the number of reticulation nodes in the output phylogeny. Under the evolutionary models used in our study and others (Yu *et al.*, 2014a; Solís-Lemus and Ané, 2016), we note that increasing c_r for a given input alignment $\boldsymbol{A}$ results in a solution with either better or equal model likelihood. For this reason, inference to address this and related problems is coupled with standard model selection techniques to balance model complexity (as determined by c_r) with model fit to the observed data. The output consists of a directed phylogenetic network N where each leaf in $\mathcal{L}(N)$ corresponds to a taxon $x \in X$.

3 Methods

3.1 The FastNet algorithm

We now describe our new divide-and-conquer algorithm, which we refer to as FastNet.

Step zero: obtaining local gene trees. FastNet is a summary-based method for inferring phylogenetic networks. Each subsequent step of the FastNet algorithm therefore utilizes a set of gene trees G as input, where a gene tree $g_i \in G$ represents the evolutionary history of each data partition a_i. The simulation study includes "boosting" experiments to evaluate the performance of FastNet relative to the base method that is being "boosted" (see below); all methods in these experiments (including FastNet) make use of true gene trees. In all other simulation study experiments and empirical analyses, we use FastTree (Price *et al.*, 2010) to estimate gene trees as input to the summary-based methods (including FastNet). See below for details regarding FastTree inference.

Step one: obtaining a guide phylogeny. The subsequent subproblem decomposition step requires a guide phylogeny N_0. With the goal of constraining evolutionary divergence of taxa in a subproblem, the phylogenetic relationships in the guide phylogeny are used to measure evolutionary relatedness. The phylogenetic relationships need not be completely accurate. Rather, the guide tree needs to be sufficiently accurate to inform subsequent divide-and-conquer steps. Another essential requirement is that the method used for inferring the guide phylogeny must have reasonable computational requirements.

For these reasons, we utilized ASTRAL (Mirarab and Warnow, 2015; Mirarab *et al.*, 2014a), a state-of-the-art phylogenomic inference method that infers species trees, to infer a guide phylogeny that was a tree

rather than a network. A primary reason for the use of species tree inference methods is their computational efficiency relative to state-of-the-art phylogenetic network inference methods. While ASTRAL accurately infers species trees for evolutionary scenarios lacking gene flow, the assumption of tree-like evolution is generally invalid for the computational problem that we consider. As we show in our performance study, our divide-and-conquer approach can still be applied despite this limitation, suggesting that FastNet is robust to guide phylogeny error. Another limitation of using ASTRAL in this context is that it effectively infers an unrooted and undirected species tree. To obtain a rooted guide phylogeny, our study made use of an outgroup taxon and the ASTRAL-inferred tree was rooted on the leaf edge corresponding to the outgroup taxon.

Step two: subproblem decomposition. The rooted and directed guide phylogeny N_0 is then used to produce a subproblem decomposition D. The decomposition D induces a partitioning of the set of taxa X into disjoint subsets D_i for $1 \leq i \leq q$ where $\bigcup_{1 \leq i \leq q} D_i = D$. We refer to each subset D_i as a "bottom-level" subproblem, which refers to the subproblem decomposition technique.

Our study made use of guide phylogenies that were strictly trees, which simplifies the subproblem decomposition procedure. Since the guide phylogeny is a tree and contains no reticulation edges, removal of any single edge will disconnect the phylogeny into two subtrees; the leaves of the two subtrees will form two subproblems. We extend this observation to obtain decompositions with two or more subproblems. The subproblem decomposition D is a set of nodes in the guide phylogeny N_0, where each node $d \in D$ induces a subproblem consisting of the taxa corresponding to the leaves which are reachable from d in N_0. Of course, not all decompositions are created equal. In this study, we explore the use of two criteria to evaluate decompositions: the maximum subproblem size c_m and a lower bound on the number of subproblems. We addressed the resulting optimization problem using a greedy algorithm. The algorithm is similar to the Center-Tree-i decomposition used by Liu *et al.* (Liu *et al.*, 2009) in the context of species tree inference. The main difference is that we parameterize our divide-and-conquer based upon a different set of optimization criteria. The input to our decomposition algorithm is the rooted directed tree N_0 and the parameter c_m, which specifies the maximum subproblem size. Our decomposition procedure also stipulates a minimum number of subproblems of 2. The initial decomposition consists of the root node $r(N_0)$. A current decomposition D is iteratively updated as follows: each iteration greedily selects a node $d \in D$ for which no larger subproblem exists in D, the node d is removed from the set D and replaced by its children. Iteration terminates when both decomposition criteria (the maximum subproblem size criterion and the minimum number of subproblems) are satisfied. If no decomposition satisfies the criteria, then the search is restarted using a maximum subproblem size of $c_m - 1$. In practice, the parameter c_m is set to an empirically determined value which is based upon the largest datasets that state-of-the-art methods can analyze accurately within a reasonable timeframe (Hejase and Liu, 2016). The output of the search algorithm is effectively a search tree N_0^{top} with a root corresponding to $r(N_0)$, leaves corresponding to $d \in D$, and the subset of edges in N_0 which connect the root $r(N_0)$ to the nodes $d \in D$ in N_0. The decomposition is obtained by deleting the search tree's corresponding edge structure in N_0, resulting in q sub-trees which induce subproblems as before.

Step three: gene flow detection and inferring phylogenies on subproblems. Assuming that a reasonable subproblem decomposition can be obtained, tree-based divide-and-conquer approaches reduce evolutionary divergence within subproblems by effectively partitioning the inference problem within parts of the true phylogeny. Within each part of the true phylogeny corresponding to a subproblem, the space of possible sub-tree topologies contributes a smaller set of distinct bipartitions (each corresponding to a possible tree edge) that need to be evaluated during search as compared to the full inference problem. The same insight can be applied to reticulation edges as well, except that a given reticulation is not necessarily restricted to a single "bottom-level" subproblem. Another possibility is that reticulation edges can span two distinct subproblems. We therefore construct additional "pairwise" subproblems C_j for $1 \leq j \leq \binom{q}{2}$, where each pairwise subproblem C_j contains the taxa from a distinct pair of bottom-level subproblems D_k and D_l (i.e., $C_j = D_k \cup D_l$). Furthermore, to account for reticulations that are ancestral to bottom-level subproblems, we also form a "top-level" subproblem C_{top} by randomly sampling a single taxon from each bottom-level subproblem D_i. The augmented decomposition D' consists of $D \cup \{C_j\} \cup \{C_{\text{top}}\}$.

The following inference problem uses the augmented decomposition D' to detect gene flow (or lack thereof) within and/or between subproblems. Additional inputs to the problem include the guide phylogeny N_0 and the parameter c_r which specifies the number of reticulation nodes in the final phylogeny. Furthermore, we include the gene trees G inferred using alignment $\boldsymbol{A}$ as input rather than the sequence data itself, since we utilize summary-based approaches to address this problem. Let G_s be the restriction of the gene trees G to the set of subproblem taxa for $s \in D'$. The output consists of a subproblem subset $D'' \subseteq D'$, for use in the subsequent subproblem inference and merge steps of the FastNet algorithm, and a function $\Delta(\cdot)$ that assigns each subproblem s to an integer in $[0, c_r]$, where $\Delta(s)$ is the estimated number of reticulation nodes for the phylogeny corresponding to subproblem s (i.e., the number of reticulations that were detected in a subproblem in the augmented decomposition) and the total number of reticulation nodes is subject to the constraint $\sum_{s \in D'} \Delta(s) = c_r$.

To address this problem, we utilize a greedy algorithm which makes use of existing statistical summary-based methods for phylogenetic network inference. Let Ξ be such a method which infers species networks under an evolutionary model with parameters Θ. In this study, the choices for Ξ consist of several state-of-the-art summary-based methods. One suitable choice which was shown to be accurate in our previous performance study (Hejase and Liu, 2016) is the maximum likelihood estimation method of Yu *et al.* (2014a), which we refer to as PhyloNet-MLE. Let $F_{\Xi,\Theta}(G_s, k)$ be the species network with k reticulation nodes which is inferred by method Ξ under its model with parameters Θ and subproblem input G_s. Let $L_{\Xi,\Theta}(G_s, k)$ be the corresponding model likelihood of the inferred network.

The greedy algorithm requires a ranking of all valid assignments Δ, which we compute as follows. First, we enumerate all valid assignments Δ given the augmented decomposition D' and the number of reticulation nodes c_r. For each possible assignment Δ, we use method Ξ to analyze each subproblem $s \in D'$ to obtain the subproblem network $F_{\Xi,\Theta}(G_s, \Delta(s))$ and the corresponding likelihood $L_{\Xi,\Theta}(G_s, \Delta(s))$. For more speed, we constrained the number of network topology searches for Ξ to 1000 for subproblems containing more than 8 taxa; for subproblems containing more than 14 taxa, we used pseudo-likelihood approximations to full model likelihood calculations given that datasets of this size exceed the scalability limits of full likelihood models (Hejase and Liu, 2016). We then rank the Δ assignments using a probabilistic criterion which is based on a full model likelihood $L_{\Xi,\Theta}(\cdot, \cdot)$. For more speed, we relaxed the full likelihood calculation and instead optimized the product of subproblem likelihoods $\prod_{s \in D'} L_{\Xi,\Theta}(G_s, \Delta(s))$. This calculation is an approximation since it effectively assumes that subproblems are independent, although they are correlated through connecting edges in the model phylogeny.

Given the ranked Δ assignments, we perform the following greedy search starting from the top-ranked Δ assignment. Let D'' be the current candidate set of subproblems for use in the subsequent subproblem

inference and merge steps of the FastNet algorithm; initially, D'' is the empty set $\emptyset$. Furthermore, let N'' be a "greedy pairwise merge" phylogeny, initially set to a star tree on the full set of taxa X. (1) We greedily select the subproblem $s' = \arg\max_s \Delta(s)$ that has the largest number of estimated reticulations according to the current Δ assignment and that has not yet been processed during this step. We add the greedily chosen subproblem s' to D'' if s' has a corresponding estimated phylogeny $F_{\Xi,\Theta}(G'_s, \Delta(s'))$ that can be "merged" with N'', in the sense that there exists a network N''' defined on taxa $(\bigcup_{d\in D''} d) \cup \{s'\}$ such that N''' displays $F_{\Xi,\Theta}(s', \Delta(s'))$; we correspondingly update N'' to be identical to N''' in this case. (The top-level subproblem C_{top} is merged differently compared to the other subproblems; see below for details). In principle, finding N''' is possible using brute force but inefficient; in practice (and by design), the subproblem decomposition $\{D_i\}$ obtained in step two of FastNet tends to result in an augmented decomposition with corresponding subproblem phylogenies which are (nearly) monophyletic for each D_i. (2) We repeat the first step until either $\bigcup_{d\in D''} d = X$ or no suitable candidates s' remain (i.e., all subproblems in the augmented decomposition have been processed). (3) If $\bigcup_{d\in D''} d = X$, then terminate and return D'', Δ, and N'' as output. Otherwise, repeat the greedy search from the first step with the next-ranked Δ assignment.

In the experiments in our study, a satisfying Δ assignment was always found. In principle, a (very) complicated model phylogeny may necessitate relaxing some search constraints, such as reducing c_r.

Step four: merge subproblem phylogenies into a phylogeny on the full set of taxa. The ranking criterion in step three effectively assumes that the phylogenetic relationships connecting the bottom-level subproblem phylogenies $N_{C_i} = F_{\Xi,\Theta}(G_{D_i}, \Delta(D_i))$ consist of a star tree $(N_1 : \delta, N_2 : \delta, \ldots, N_q : \delta)$; with branch length δ past saturation. The FastNet algorithm resolves the "top-level" structure of the output phylogeny using the phylogeny inferred on C_{top} and the other outputs of step three, including the optimal assignment Δ. For each bottom-level subproblem D_i, the rooted network $F_{\Xi,\Theta}(G_{D_i}, \Delta(D_i))$ replaces the corresponding leaf in the "top-level" phylogeny $F_{\Xi,\Theta}(G_{C_{\text{top}}}, \Delta(C_{\text{top}})$. For each pairwise subproblem C_j formed from two distinct bottom-level subproblems D_k and D_l, we utilized the following merge procedure which is designed to account for the (relatively rare) case that the corresponding subproblem phylogeny $F_{\Xi,\Theta}(G_{C_j}, \Delta(C_j)$ lacks monophyly for D_k or D_l. We retain the two outgoing edges incident upon the root node of the pairwise subproblem phylogeny $F_{\Xi,\Theta}(G_{C_j}, \Delta(C_j)$ but replace the root node with two "dangling" nodes (i.e., nodes with in-degree zero and out-degree one) – one for each outgoing edge; each outgoing edge will "attach" to the "top-level" phylogeny $F_{\Xi,\Theta}(G_{C_{\text{top}}}, \Delta(C_{\text{top}})$ by replacing one leaf edge corresponding to either bottom-level subproblem D_k or D_l. The criteria for replacement is based upon the bottom-level subproblem (either D_k or D_l) that has the greatest amount of overlap with the set of taxa that are reachable from a given dangling node via tree edges only. The result of the merge procedure is a phylogeny N on the full set of taxa.

3.2 Performance study

Below we describe the steps used in the performance study. Detailed commands and software options are given in the Appendix.

Simulation of model networks. Our goal was to simulate model networks that reflected recent (paraphyletic and monophyletic) and ancestral gene flow deeper in the species phylogeny, similar to the corresponding simulations in the study of Leaché *et al.* (2014). We first simulated random model trees using r8s version 1.7 (Sanderson, 2003) for 20, 30, and 50 taxa. Twenty random model tree replicates were generated for each model condition (recent and deep gene flow) where the height of each tree replicate was scaled to 5. One, two, or three reticulations were added to each model tree using the following two steps. First, select two taxa or clades. Second, add unidirectional migration occurring from 0 to t with a rate of 5.0 between the two selected taxa or clades. After generating a random model network, an outgroup was added at coalescent time 10. 1000 gene trees were simulated for each random model network using ms (Hudson, 2002). In our simulation study, we sampled one allele per taxon.

Simulation of DNA sequences. The evolution of sequences was simulated using seq-gen (Rambaut and Grassly, 1997), which takes the gene trees generated by ms as input and simulates the evolution of sequences according to a finite-sites model. Using the Jukes-Cantor mutation model (Jukes and Cantor, 1969), we simulated the evolution of DNA sequences for each local genealogy generated by ms. The simulated sequence had a total length of 1000 kb, which was equally distributed across all local genealogies (1000 bp per local genealogy).

Gene tree inference. FastTree (Price *et al.*, 2009, 2010) was used for local gene tree inference. We used the Jukes-Cantor model (Jukes and Cantor, 1969) to infer a maximum-likelihood gene tree for each sequence alignment. The inferred gene trees were rooted using the outgroup.

Species phylogeny estimation methods. The performance of FastNet was evaluated relative to each state-of-the-art method that it "boosted". We used a likelihood-based approach based on the MLE method of Yu and Nakhleh (Yu *et al.*, 2011), which calculates model likelihood while factoring in gene tree branch lengths (Bryant *et al.*, 2012). We refer to this method as MLE-length. We also applied a pseudo-likelihood inference method based on the method of Yu and Nakhleh (Yu and Nakhleh, 2015), which uses pseudo-likelihood approximations to the full model likelihood.

Performance measures. We evaluated the inference methods using multiple criteria. The first evaluation criterion is topological accuracy. We compared the inferred phylogeny to the model phylogeny using the tripartition distance (Nakhleh *et al.*, 2003), which counts the proportion of tripartitions that are not shared between the inferred and model network. The second evaluation criterion is the computational requirements of the method, which was measured in terms of CPU runtime. All computational analyses were run on Michigan State University's High-Performance Computing Center. We used compute nodes in the intel14 cluster which had 2.5 GHz Intel Xeon E5-2670v2 processor with 64 GiB of main memory per node.

3.3 Empirical data

Bacteria belonging to the Burkholderiaceae are of interest given their importance in human and plant disease, but also given their role as plant and fungal endosymbionts and their metabolic capacity to degrade xenobiotics. Fully sequenced (closed) genomes belonging to Burkholderiaceae were selected and downloaded from the PATRIC web portal (www.patricbrc.org). We chose to maximize phylogenetic and ecological diversity in this sampling, so we included available genomes belonging to free-living, pathogenic, and endosymbiotic species spanning across the genera *Burkholderia*, *Ralstonia*, *Pandoraea*, *Cupriavidus*, *Mycoavidus*, and *Polynucleobacter*. Genomes ranged in size from 2048 (1.56 MB) and 9172 (9.70 MB) coding DNA sequences (CDS). The software package Proteinortho was used to select for single copy orthologs across selected genomes based upon amino acid similarity and using default parameters (Lechner *et al.*, 2011). Proteinortho inferred 549 orthologs where the multiple sequence alignment of each ortholog contained 57 taxa. To obtain estimated gene trees for each ortholog, FastTree under the JTT+CAT model was used to infer the maximum-likelihood unrooted gene tree for each sequence alignment. Each gene tree was rooted using the following outgroup taxon: *Polynucleobacter necessarius subsp asymbioticus*.

4 Results

4.1 Simulation study

Runtime. We report the runtimes of FastNet and the other phylogenetic inference methods in Table 1. Recall that the probabilistic network inference methods were found to be the most accurate among state-of-the-art methods, and MPL was among the fastest methods in this class (Hejase and Liu, 2016). Furthermore, MPL was designed to tradeoff optimization under a pseudo-likelihood-based approximation for increased computational efficiency compared with full likelihood methods (Yu and Nakhleh, 2015). However, the tradeoff netted efficiency that was well short of current phylogenomic dataset sizes (Hejase and Liu, 2016). In comparison to MPL, FastNet was faster by orders of magnitude. For dataset sizes of 30 taxa or less, FastNet completed analysis in less than six hours. For the largest 50 taxon datasets in our study, FastNet completed analysis in approximately a day. MPL was slower than FastNet by several factors. MPL completed analysis on dataset sizes of 20 and 30 taxa in approximately one and three days, respectively. MPL did not complete the analysis of datasets with 50 taxa after one week of runtime. We predict that MPL's computational runtime requirements render it infeasible for analysis of any datasets larger than 50 taxa in our study. MP was faster than FastNet by several factors. As we have shown in our previous study (Hejase and Liu, 2016), the speed of this method comes at a cost in terms of accuracy, which is consistent with other studies examining problems of phylogenetic tree estimation (Mirarab *et al.*, 2014b) and statistical inconsistency of MP approaches in this context (Felsenstein, 1978). The runtime of each method increased as the number of taxa increased.

FastNet as a boosting method. The purpose of FastNet is to boost existing methods, much like previous tree-based divide-and-conquer methods (Liu *et al.*, 2009). Using true gene trees as input, we evaluated the performance boost of FastNet using MLE-length as a base method compared to the base method itself (Table 2). We observed a boost in terms of accuracy and runtime of 0.146 (as measured by the tripartition distance) and 64 hours, respectively. FastNet, using MLE-length as a base method, was significantly more accurate and faster than the boosted method (Benjamini-Hochberg-corrected pairwise t-test; $\alpha = 0.05$ and $n = 20$). Similarly, we compared the performance of FastNet, using MPL as a base method, relative to the base method itself (Table 2). We observed a performance boost of 0.165 as measured by the tripartition distance and 20 hours for the accuracy and runtime, respectively. The performance advantage of FastNet over the boosted method was significant using a Benjamini-Hochberg-corrected pairwise t-test ($\alpha = 0.05$ and $n = 20$). The aforementioned two boosting experiments demonstrated the relative improvement in accuracy and time for FastNet versus the boosted method (MLE-length or MPL).

Topological accuracy of FastNet: one-reticulation-node model conditions. We next evaluated the topological accuracy of the phylogenies inferred by FastNet across a range of evolutionary scenarios and using inferred gene trees as input. Figure 1 shows the topological accuracy of FastNet on one-reticulation-node model conditions. We found that FastNet's topological error tended to increase from 0.02 to 0.05 as dataset sizes increased from 20 to 50 taxa for the recent gene flow model conditions. For the ancestral gene flow model conditions, the accuracy ranged between 0.06 and 0.15. Overall, FastNet was more accurate in the recent gene flow model conditions compared to the ancestral gene flow model conditions across all taxa with an average accuracy improvements of 0.08, 0.12, and 0.02, for dataset sizes of 20, 30, and 50, respectively. This observation suggests that detecting ancestral gene flow deeper in the species phylogeny is more difficult compared to detecting recent (paraphyletic and monophyletic) gene flow.

On the twenty taxon datasets, FastNet perfectly recovered the topology of the model phylogeny for 16 and 6 replicates for the recent and ancestral gene flow model conditions, respectively, and was nearly perfectly accurate on the remaining replicates. On the thirty and fifty taxon datasets, FastNet perfectly recovered the topology of the model phylogeny for 10 and 8 replicates for the recent gene flow model conditions, and 0 and 4 replicates for the ancestral deeper in the species phylogeny gene flow model conditions. Therefore, we observed more difficulty in perfectly inferring a phylogeny as we increase the number of taxa. One minor exception to this observation is the performance of FastNet on thirty and fifty taxon datasets using ancestral gene flow model conditions.

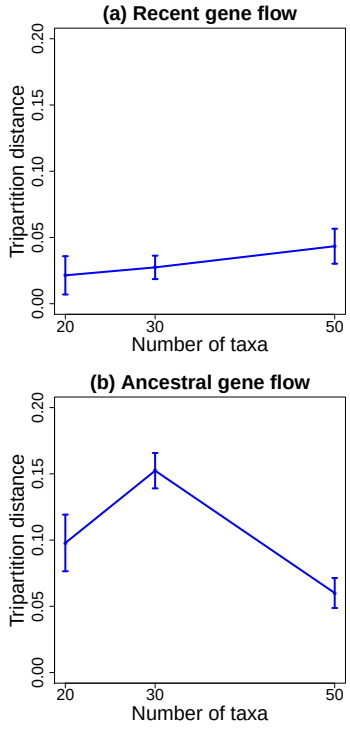

Fig. 1. Topological accuracy of FastNet on one-reticulation-node model conditions on dataset sizes ranging from 20 to 50. The topological accuracy of each inferred phylogeny with respect to the model phylogeny is evaluated using the tripartition distance (See Methods for details). The panels show results for model conditions with either: (a) recent gene flow or (b) ancestral gene flow. Average distances and standard error bars are shown ($n = 20$).

We evaluated the performance of FastNet as we varied the number of loci from 100 to 1000 using twenty taxon datasets on recent gene flow model conditions (Figure 2). As we increase the number of loci from 100 to 1000, the topological error as measured by the tripartition distance decreased from 0.06 to 0.02.

Topological accuracy of FastNet: two-reticulation-node and three-reticulation-node model conditions. On model conditions where the model phylogeny contained two reticulation nodes rather than one, FastNet inferred a rooted, directed phylogeny with an average tripartition distance of 0.04 (Figure 3). Similarly, on model conditions where the model phylogeny contained three reticulation nodes, FastNet had relatively high accuracy with an average topological distance of 0.08. FastNet perfectly recovered the topology of the model phylogeny in 16, 13, and 7 replicates for the one-reticulation-node, two-reticulation-node and three-reticulation-node model conditions, respectively. As the number of reticulation nodes increased from one to three, FastNet's error tended to increase from 0.02 to 0.08.

Table 1. **Runtime in hours for FastNet and the other phylogenetic inference methods.** The model conditions involved dataset sizes ranging from 20 to 50 and model phylogenies that contained one reticulation node. Average runtime ("Avg") and standard errors ("SE") are listed ($n = 20$). MPL was unable to finish analysis of datasets with fifty taxa after one week of runtime.

Number of taxa	Runtime in hours											
	Recent gene flow						Ancestral gene flow					
	FastNet		MPL		MP		FastNet		MPL		MP	
	Avg	SE	Avg	SE	Avg	SE	Avg	SE	Avg	SE	Avg	SE
20	3.7	1.3	23.7	4.3	<1	<0.1	2.5	0.6	21.4	4.5	<1	<0.1
30	6.1	2.4	60.3	14.7	<1	<0.1	3.8	1.0	66.1	11.8	<1	<0.1
50	25.6	3.6	-	-	<1	<0.1	19.2	2.8	-	-	<1	<0.1

Table 2. **Average distances and runtimes for the performance boost of FastNet (with either MLE-length or MPL as base methods) over base method (either MLE-length or MPL) using model conditions that contained 20 taxa.** The topological distance between the inferred and model phylogenies was measured using the tripartition distance. The model conditions involved model phylogenies that contained one reticulation node. True gene trees were used as input to FastNet, MLE-length, and MPL. Average ("Avg") and standard errors ("SE") for the performance improvement of topological distances and runtimes are listed ($n = 20$). A one-sided t-test comparing the performance advantage of FastNet over the boosted method (MLE-length or MPL) for the evaluation criteria (i.e. topological distance and runtime) was conducted. Corrected q-values are reported where multiple test correction was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

Boosted Method	Topological distance			Runtime in hours		
	Avg	SE	q value	Avg	SE	q value
MLE-length	0.146	0.025	1.78×10^{-5}	64.21	5.887	2.41×10^{-9}
MPL	0.165	0.01	7.93×10^{-3}	20.501	5.44	6.99×10^{-4}

4.2 Empirical study

To estimate a phylogeny on the empirical dataset, we coupled FastNet analysis with standard model selection approaches to choose the number of reticulation nodes in the output phylogeny. One approach consisted of standard information criteria (Akaike, 1974; Schwarz, 1978). As shown in Supplementary Table S4 in the Appendix, the FastNet-inferred phylogeny with two reticulation nodes is preferred to the FastNet-inferred phylogeny with one reticulation node, and both are preferred to the ASTRAL-inferred tree which served as FastNet's guide phylogeny. The preferred phylogeny (i.e., the FastNet-inferred phylogeny with two reticulation nodes) is shown in Supplementary Figure S2 in the Appendix. The preferred phylogeny contains the FastNet-inferred phylogeny with one reticulation node (Supplementary Figure S1 panel i in the Appendix); the latter phylogeny contained the same reticulation node and edges, all of which were subsumed by the preferred phylogeny. The FastNet-inferred phylogeny with two reticulation nodes contained 53 out of the 53 internal tree edges which formed the ASTRAL-inferred tree. A standard slope analysis (similar to the study of Solís-Lemus and Ané (Solís-Lemus and Ané, 2016)) suggests that the true phylogeny may contain additional reticulation edges and nodes (Supplementary Figure S3 in the Appendix).

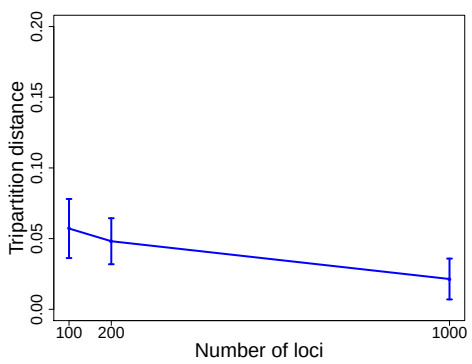

Fig. 2. Topological accuracy of FastNet on one-reticulation-node model conditions where the number of loci per replicate dataset ranged between 100 and 1000 loci. The model conditions consisted of recent gene flow events. Figure layout and description are otherwise identical to Figure 1.

5 Discussion

Performance study. Using both simulated and empirical data, we evaluated the performance of FastNet across a range of evolutionary scenarios. We evaluated performance based upon computational runtime and topological accuracy.

FastNet was roughly an order of magnitude faster and more accurate than MPL, a pseudo-likelihood-based inference method. FastNet analyses required only slightly longer runtime compared to the fastest method in our study: MP, which was among the least accurate methods in our previous study (Hejase and Liu, 2016). On the largest datasets in our study which had 50 taxa, FastNet analyses completed within a day – well within the range of feasibility. We predict that FastNet will almost certainly scale to

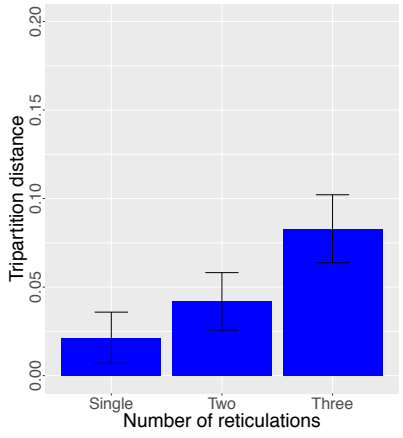

Fig. 3. Topological accuracy of FastNet on one-reticulation-node, two-reticulation-node and three-reticulation-node model conditions. Figure layout and description are otherwise identical to Figure 1.

datasets with hundred taxa. We note that the FastNet algorithm is pleasantly parallelizable (as are some of the other methods in this study).

We used FastNet to "boost" the performance of two state-of-the-art methods, resulting in improved topological accuracy and computational runtime. We note that the base methods (MLE-length and MPL) were run in default mode. More intensive settings for each base method's optimization procedures may allow a tradeoff between topological accuracy and computational runtime. We stress that our goal was *not* to make specific recommendations about the nuances of running the base methods. Rather, FastNet's divide-and-conquer framework can be viewed as orthogonal to the specific algorithmic approaches utilized by the base method to be boosted. In this sense, improvements to the latter accrue to the former in a straightforward and modular manner.

We explored the impact of multiple factors upon FastNet's topological accuracy. For evolutionary scenarios involving recent gene flow, FastNet's accuracy was relatively robust to the dataset sizes explored in our study (in terms of the number of taxa as well as the number of loci). A relatively greater impact on topological accuracy was seen in the presence of ancestral gene flow as well as a greater number of reticulations in the model phylogeny.

To our knowledge, our study contains the first data-driven survey of horizontal gene transfer (HGT) among the Burkholderiaceae which utilizes the latest phylogenomic models and methods. We note that the literature suggest pervasive horizontal gene transfer among the Burkholderiaceae (Boyd *et al.*, 2009). These findings echo the common sentiment among microbiologists that horizontal gene transfer must be widespread throughout the prokaryotic Tree of Life. Many corollary questions follow. For example, is horizontal gene transfer uniformly distributed across the prokaryotic Tree of Life, or not? Concrete and quantitative answers to these questions crucially depend upon computational methodologies that scale to large and divergent datasets. We view FastNet as the first step in this direction. The empirical analysis in our study points to multiple horizontal gene transfer events with unknown evolutionary roles. Fine-scale intra-genomic detection of alleles and loci involved in HGT will shed more light into the role of HGT in the evolution of the Burkholderiaceae. Analyses using PhyloNet-HMM (Liu *et al.*, 2014) or related methods (Mailund *et al.*, 2012; Durand *et al.*, 2011) would be suitable for this purpose.

The algorithmic design of FastNet. We consider the procedures used in FastNet's step one (inferring a guide phylogeny), step two (subproblem decomposition), and step four (merge) to be reasonable approaches, but more sophisticated alternatives can (and should) be proposed. Rather, we decided to focus our effort on the part of the phylogenetic network inference problem that we hypothesized to both have a first-order impact on inference accuracy and that substantially differentiates the problem from species tree inference: namely, step three – gene flow detection within "regions" of a guide phylogeny. Despite these methodological limitations and the added challenge of moderate gene tree error relative to other studies (see Appendix), we were able to obtain consistent improvements in topological accuracy and computational runtime when FastNet was used to boost the performance of a base method. Taken together, these results suggest that FastNet is robust to the choice of guide phylogeny, subproblem decomposition, merge technique, and gene tree error. We attribute FastNet's performance advantage to the techniques used in step two of the algorithm. Inference accuracy tended to diminish as dataset sizes increased, which is likely due to larger subproblem sizes and more complex correlation structure (due to additional edge structure) "above" subproblems. Recursive application of divide-and-conquer should help to address these issues.

6 Conclusion

In this study, we introduced FastNet, a new computational method for inferring phylogenetic networks from large-scale genomic sequence datasets. FastNet utilizes a divide-and-conquer algorithm to constrain two different aspects of scale: the number of taxa and evolutionary divergence. We evaluated the performance of FastNet in comparison to state-of-the-art phylogenetic inference methods. We found that FastNet improves upon existing methods in terms of computational efficiency and topological accuracy. FastNet was an order of magnitude faster than the most accurate state-of-the-art phylogenetic network inference method. Furthermore, FastNet's topological accuracy was comparable to or typically better than all other methods in our study.

Future enhancements to FastNet's algorithmic design are anticipated to yield further performance improvements. Here we highlight several possibilities. First, the greedy search used in step three of FastNet will suffice for small values of c_r, but dynamic programming based upon Δ assignments will likely be necessary to retain computational efficiency as the number of reticulation nodes increases. Second, as noted above, more sophisticated techniques for gene tree inference, inferring a guide phylogeny, subproblem decomposition, and merging phylogenetic networks inferred on subproblems can be substituted for the approaches used in our study. Third, the use of a guide phylogeny naturally invites

iteration: the output phylogeny from one iteration of the FastNet algorithm would be used as the guide phylogeny for a subsequent iteration of the algorithm. This requires modifying step one of FastNet to utilize a guide network in lieu of a guide tree. We provide specific recommendations below.

To provide added flexibility in algorithmic design, we briefly return to the question of subproblem decomposition using a guide phylogeny that is not necessarily a tree. In general, the use of a phylogenetic network for subproblem decomposition presents two challenges. First, if a phylogeny contains no reticulation edges and is therefore a tree, removal of any single edge will break the phylogeny into two subtrees and the leaves of the two subtrees will form two subproblems. However, when a phylogeny contains reticulation edges, this observation no longer applies: removal of a single edge (reticulation or tree) does not necessarily disconnect a phylogenetic network. To deal with this challenge, we choose a random member from the set of rooted tree topologies encoded in the network topology of N_0 and use the rooted tree topology T_0 for subproblem decomposition. The greedy algorithm in step one can then be applied to the rooted tree topology T_0.

Several aspects of our performance study can also be revisited in the future to better understand the performance of FastNet and related methods. Recall that the gene tree topology distributions explored in our study became more "tree-like" as more taxa and reticulation nodes were added to the model phylogeny. This observation depends in part upon the simulation conditions explored in our study. One contributing factor is that the gene flow parameter values used in our study were uniform across the phylogeny (and different model conditions). In general, not all reticulation nodes are equal in this sense, and we predict that different reticulation nodes will have differential effects upon coalescent histories. For example, rare gene flow events will have less effect compared to more common gene flow events. Our performance study focused on topological comparisons and sets the stage for future evaluations that also evaluate branch length accuracy. More work is needed to devise suitable phylogenetic network distances, possibly building upon the related work of Kuhner and Felsenstein (1994) and Solís-Lemus and Ané (2016).

We conclude with some parting thoughts about the computational problem of phylogenetic network inference. In today's post-genomic era, current trends in biomolecular sequence technologies suggest that even the scalability advance set forth in this study will not suffice for near future studies. There is a critical need for new phylogenomic methodologies to infer species networks involving thousands of taxa or more. Another important issue involves appropriate representations for phylogenies involving both vertical divergence and horizontal gene flow. In our view, inference of vertical divergence and inference of horizontal gene flow really represent two orthogonal questions. We recommend the use of orthogonal measures to separately evaluate the two (despite differences in scale in terms of the number of edges of each type). Furthermore, we would argue that, in explicit phylogenetic network representations, exact placement of the endpoints of reticulation edges may be difficult in some cases, whereas a summary-based localization may be tractable and almost as informative (cf. Figure 3 in Yu *et al.* (Yu *et al.*, 2014a)). We believe that step three of FastNet suggests a way forward (i.e., alternative phylogenetic representations that summarize gene flow within "regions" of a phylogeny). One possibility would be to generalize tree-based concordance factors (Baum, 2007) for this purpose.

Acknowledgements

Funding

We gratefully acknowledge the following support: National Science Foundation Grant CCF-1565719 (to KJL), a grant from the BEACON Center for the Study of Evolution in Action (NSF STC Cooperative Agreement DBI-093954) to KJL and GAB, and Michigan State University faculty startup funds (to KJL and to GAB).

References

Abbott, R. J. and Rieseberg, L. H. (2012). Hybrid speciation. In *Encyclopaedia of Life Sciences*. John Wiley & Sons, Ltd, Hoboken, NJ, USA.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Bandelt, H.-J. and Dress, A. W. (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, **92**(1), 47–105.

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, pages 417–426.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**(1), 289–300.

Boyd, E. F., Almagro-Moreno, S., and Parent, M. A. (2009). Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends in Microbiology*, **17**(2), 47–53.

Bryant, D. and Moulton, V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**(2), 255–265.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, **29**(8), 1917–1932.

Davidson, R., Vachaspati, P., Mirarab, S., and Warnow, T. (2015). Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics*, **16**(Suppl 10), S1.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**(8), 2239–2252.

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, **63**(1), 1–19.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, **27**(4), 401–410.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, **328**(5979), 710–722.

Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*. Oxford University Press, Oxford.

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

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

Huelsenbeck, J. P. and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, **42**(3), 247–264.

Jukes, T. and Cantor, C. (1969). *Evolution of Protein Molecules*, pages 21–132. Academic Press, New York, NY, USA.

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

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, **13**(3), 235–248.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, **11**(3), 459–468.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2013). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, page syt049.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2014). The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, **63**(1), 17–30.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics*, **12**(1), 1.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**(5934), 1561–1564.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P., and Linder, C. R. (2012). SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, **61**(1), 90–106.

Liu, K. J., Dai, J., Truong, K., Song, Y., Kohn, M. H., and Nakhleh, L. (2014). An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, **10**(6), e1003649.

Liu, K. J., Steinberg, E., Yozzo, A., Song, Y., Kohn, M. H., and Nakhleh, L. (2015). Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences*, **112**(1), 196–201.

Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., Andersen, L. N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A., and Schierup, M. H. (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet*, **8**(12), e1003125.

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

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**(1), 31–46.

Mirarab, S. and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, **31**(12), i44–i52.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014a). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**(17), i541–i548.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2014b). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, page syu063.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, **22**(5), 377–386.

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

Nakhleh, L., Sun, J., Warnow, T., Linder, C. R., Moret, B. M., and Tholse, A. (2003). Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Pacific Symposium on Biocomputing*, volume 8, pages 315–326. World Scientific.

Price, M., Dehal, P., and Arkin, A. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, **26**(7), 1641–1650.

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

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M. V., Derevianko, A. P., Hublin, J.-J., Kelso, J., Slatkin, M., and Paabo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**(7327), 1053–1060.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Solís-Lemus, C. and Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*, **12**(3), 1–21.

The Heliconious Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**(7405), 94–98.

Yu, Y. and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, **16**(Suppl 10), S10.

Yu, Y., Than, C., Degnan, J. H., and Nakhleh, L. (2011). Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, **60**(2), 138–149.

Yu, Y., Degnan, J. H., and Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, **8**(4), e1002660.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014a). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, **111**(46), 16448–16453.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014b). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, **111**(46), 16448–16453.