METHOD

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

Hussein A Hejase[1], Natalie VandePol[2], Gregory M Bonito[2] and Kevin J Liu[3*]

Abstract

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? 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 the 2016 performance study of Hejase and Liu, state-of-the-art methods fell 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; genetic drift; incomplete lineage sorting; coalescent; divide and conquer

Recent advances in biomolecular sequencing [1] and phylogenomic modeling and inference [2, 3] have revealed that interspecific gene flow has played a major role in the evolution of many different organisms across

*Correspondence: kjl@msu.edu

[3]Department of Computer Science and Engineering, Michigan State University, 428 S. Shaw Lane, 48824 East Lansing, MI, USA

Full list of author information is available at the end of the article

the Tree of Life [4–6], including humans and ancient hominins [7, 8], butterflies [9], mice [10], and fungi [11]. 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 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. Solís-Lemus and Ané proposed SNaQ [12], a new statistical method which seeks to address the computational efficiency of species network inference using a pseudo-likelihood approximation. The method of Yu and Nakhleh [13] (referred to here as MPL, which stands for maximum pseudo-likelihood) substitutes pseudo-likelihoods in place of the full model likelihoods used by the methods of Yu et al. [14] (referred to here as MLE, which stands for maximum likelihood estimation, and MLE-length, which differ based upon whether or not gene tree branch lengths contribute to model likelihood). Two of us recently conducted a performance study which demonstrated the scalability limits of SNaQ, MPL, MLE, MLE-length, and other state-of-the-art phylogenetic methods in the context of phylogenetic network inference [15]. 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 [3]. The most accurate phylogenetic network inference methods performed statistical inference under phylogenomic models [12, 14, 16] that extended the multi-species coalescent model [17, 18]. MPL and SNaQ were among the fastest of these methods while MLE and MLE-length were the most accurate. None of the statistical phylogenomic inference methods completed analyses of datasets with 30 taxa or more after many weeks of CPU runtime – not even the pseudo-likelihood-based methods which were devised to address the scalability limitations of other statistical approaches. The remaining methods fell into two categories: split-based methods [19, 20] and the parsimony-based inference method of Yu et al. [21] (which we refer to as MP in this study). Both categories of methods were faster than the statistical phylogenomic inference methods but less accurate.

The methodological gap remains: how can species networks be accurately and efficiently inferred using large-scale genomic sequence datasets? 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. We focus on dataset size in terms of the number of taxa and the number of reticulations 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.

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 [22–24]. We and others have successfully applied divide-and-conquer approaches to enable scalable inference in the context of species tree estimation [24–26].

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. Following the instantaneous admixture model used by Durand et al. [27], each reticulation node contributes a parameter γ, where one incoming edge has admixture frequency γ and the other has admixture frequency $1 - \gamma$. 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 tree 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 [12, 14], 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 [12, 14], 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 $x \in X$.

Methods

The FastNet algorithm

We now describe our new divide-and-conquer algorithm, which we refer to as FastNet. A flowchart of the algorithm is shown in Figure 1.

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 experiments in our study utilized either true or inferred gene trees as input to summary-based inference methods, including FastNet (see below for details). We used FastTree [28] to perform maximum likelihood estimation of local gene trees. Our study made use of an outgroup, and the unrooted gene trees inferred by FastTree were rooted on the leaf edge corresponding to the outgroup.

Step one: obtaining a guide phylogeny. The subsequent subproblem decomposition step requires a rooted guide phylogeny $N^{(0)}$. 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.

Based on these criteria, we utilized two different methods to obtain guide phylogenies. We used the parsimony-based algorithm proposed by Yu et al. [21] to infer a rooted species network. The algorithm is implemented in the PhyloNet software package [29]. We refer to this method as MP. In a previous simulation study [15], we found that MP offers a significant runtime advantage relative to other state-of-the-art species network inference methods, but had relatively lower topological accuracy. We also used ASTRAL [30, 31], 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 consideration is that ASTRAL effectively infers an unrooted and undirected species tree. We rooted the species tree using outgroup rooting.

Step two: subproblem decomposition. The rooted and directed species network $N^{(0)}$ is then used to produce a subproblem decomposition D. The decomposition D consists of a "bottom-level" component and a "top-level" component, which refers to the subproblem decomposition technique. The bottom-level component is comprised of disjoint subsets D_i for $1 \leq i \leq q$ which partition the set of taxa X such that $\bigcup_{1 \leq i \leq q} D_i = X$. We refer to each subset D_i as a bottom-level subproblem. The top-level component consists of a top-level subproblem D_0 which overlaps each bottom-level subproblem D_i where $1 \leq i \leq q$.

The bottom-level component of the subproblem decomposition is obtained using the following steps. First, for each reticulation node in the network $N^{(0)}$, we delete the incoming edge with lower admixture frequency. Since the resulting phylogeny $T^{(0)}$ contains no reticulation edges and is therefore a tree, 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. Let S be an open set of nodes in the guide phylogeny $T^{(0)}$. Each node $s \in S$ induces a corresponding subproblem D_i for $1 \leq i \leq q$ which consists of the taxa corresponding to the leaves that are reachable from s in $T^{(0)}$. 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. [24] 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 $T^{(0)}$ and the parameter c_m, which specifies the maximum subproblem size. Our decomposition procedure also stipulates a minimum number of subproblems of 2. Initially the open set S consists of the root node $r(T^{(0)})$. The open set S is iteratively updated as follows: each iteration greedily selects a node $s \in S$ with maximal corresponding subproblem size, the node s is removed from the set S 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 [15]. The output of the search algorithm is effectively a search tree $T^{(0)}_{\text{top}}$ with a root corresponding to $r(T^{(0)})$, leaves corresponding to $s \in S$, and the subset of edges in $T^{(0)}$ which connect the root $r(T^{(0)})$ to the nodes $s \in S$ in $T^{(0)}$. The decomposition is obtained by deleting the search tree's corresponding edge structure in $T^{(0)}$, resulting in q sub-trees which induce bottom-level subproblems as before.

The top-level component augments the subproblem decomposition with a single top-level subproblem D_0 which overlaps each bottom-level subproblem. Phylogenetic structure inferred on D_0 represents ancestral evolutionary relationships among bottom-level subproblems. Furthermore, overlap between the top-level subproblem D_0 and bottom-level subproblems is necessary for the subsequent merge procedure (see "Step four" below). The top-level subproblem D_0 contains representative taxa taken from each bottom-level subproblem D_i for $1 \leq i \leq q$: for each bottom-level subproblem D_i, we choose the leaf in $T^{(0)}$ that is closest to the corresponding open set node $s \in S$ to represent D_i, and the corresponding taxon is included in the top-level subproblem D_0.

Step three: subproblem decomposition graph optimization. Tree-based divide-and-conquer approaches reduce evolutionary divergence within subproblems by effectively partitioning the inference problem based on phylogenetic relationships. Within each part of the true phylogeny corresponding to a subproblem, the space of possible unrooted 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 subproblem.

We address the issue of "inter-subproblem" reticulations through the use of an abstraction which we refer to as a subproblem decomposition graph. A subproblem decomposition graph $G_D = (V_D, E_D)$ is a bipartite graph where the vertices V_D can be partitioned into two sets: a set of source vertices V_D^{src} and a set of destination vertices V_D^{dst}. There is a source vertex $C_i^{\text{src}} \in V_D^{\text{src}}$ for each distinct subproblem $D_i \in D$ where $0 \leq i \leq q$, and similarly for destination vertices $C_i^{\text{dst}} \in V_D^{\text{dst}}$. An undirected edge $e_{ij} \in E_D$ connects a source vertex C_i^{src} to a destination vertex C_j^{dst} where $i \leq j$ and has a weight $w(e_{ij}) \in \mathcal{N}^+$. If an edge e_{ii} connects nodes C_i^{src} and C_i^{dst} that correspond to the same subproblem $D_i \in D$, then the edge weight $w(e_{ii}) > 0$ specifies the number of reticulations in the phylogenetic network to be inferred on subproblem D_i; otherwise, a phylogenetic tree is to be inferred on subproblem D_i. If an edge e_{ij} connects nodes C_i^{src} and C_j^{dst} where $i < j$, then the edge weight $w(e_{ij}) > 0$ specifies the number of "inter-subproblem" reticulations between the subproblems D_i and D_j (where an inter-subproblem reticulation is a reticulation with one incoming edge which is incident from the phylogeny to be inferred on subproblem D_i and the other incoming edge which is incident from the phylogeny to be inferred on D_j); otherwise, no reticulations are to be inferred between the two subproblems. A subproblem decomposition graph is constrained to have a total number of reticulations such that $\sum_{e \in E_D} w(e) = C_r$.

Given a subproblem decomposition D, FastNet's search routines make use of the correspondence between a subproblem decomposition graph G_D and a multiset with cardinality C_r that is chosen from $\binom{q+1}{2} + (q+1)$ elements. Enumeration over corresponding multisets is feasible when the number of subproblems and C_r are sufficiently small (Algorithm 1); oth-

erwise, perturbations of a corresponding multiset can be used as part of a local search heuristic.

A subproblem decomposition graph G_D facilitates phylogenetic inference given a subproblem decomposition D. The resulting inference is evaluated with respect to a pseudo-likelihood-based criterion. Pseudocode for the pseudo-likelihood calculation is shown in Algorithm 2.

The first step is to analyze each individual subproblem $D_i \in D$ where $0 \leq i \leq q$. If an edge e_{ii} exists, then a phylogenetic network with $w(e_{ii})$ reticulations is inferred on the corresponding subproblem D_i; otherwise, a phylogenetic tree is inferred. We utilized MLE, a summary-based MLE method, to perform phylogenetic inference on subproblems, which we refer to as a base method. We note that alternative optimization-based approaches (e.g., other likelihood-based approaches such as MLE-length or pseudo-likelihood-based approaches such as MPL) can be substituted in a straightforward manner.

Next, reticulations are inferred "between" pairs of subproblems as follows. Let N_i and N_j where $i \neq j$ be the networks inferred on subproblems D_i and D_j, respectively, using the above procedure. Construct the cherry $(N_i\colon b_i, N_j\colon b_j)$ANC; which consists of a new root node ANC with children $r(N_i)$ and $r(N_j)$, where N_i and N_j are respectively retained as subphylogenies. Then, infer branch lengths b_i and b_j and add $w(e_{ij})$ reticulations under the maximum likelihood criterion used by the base method. For pairs of subproblems not involving the top-level subproblem D_0, we used the base method to perform constrained optimization. For pairs of subproblems involving the top-level subproblem D_0, we used a greedy heuristic: initial placements were chosen arbitrarily for each reticulation, the source node for each reticulation edge was exhaustively optimized, and then the destination node for each reticulation edge was exhaustively optimized.

Inferred phylogenies and likelihoods were cached to ensure consistency among individual and pairwise subproblem analyses, which is necessary for the subsequent merge procedure. Caching also aids computational efficiency.

Finally, the subproblem decomposition graph and associated phylogenetic inferences are evaluated using a pseudolikelihood criterion:

$$\prod_{0\leq i\leq q} \delta[i, w(G_D, i, i)] \prod_{\substack{0\leq i\leq q \\ i<j\leq q}} \psi[i, j, w(G_D, i, j), w(G_D, i, i), w(G_D, j, j)] \quad (1)$$

where $w(G_D, i, j)$ is the weight of edge e_{ij} if it exists in $E(G_D)$ or 0 otherwise, $\delta[i, w(G_D, i, i)]$ is the cached likelihood for an individual subproblem D_i, and $\psi[i, j, w(G_D, i, j), w(G_D, i, i), w(G_D, j, j)]$ is the cached likelihood for a pair of subproblems D_i and D_j where $i < j$. The pseudo-likelihood calculation effectively assumes that subproblems are independent, although they are correlated through connecting edges in the model phylogeny. The choice of optimization criterion in this context represents a tradeoff between efficiency and accuracy, and several other state-of-the-art phylogenetic inference methods also use pseudo-likelihoods to analyze subsets of taxa (e.g., MPL and SNaQ). Other choices are possible. For example, an alternative would be to merge subproblem inferences into a single network hypothesis and calculate its likelihood under the MSNC model.

We optimize subproblem decomposition graphs under the pseudo-likelihood criterion. Exhaustive enumeration of subproblem decomposition graphs is possible for the datasets in our study. Pseudocode to obtain a global optimum is shown in Algorithm 3. For larger datasets with more reticulations, heuristic search techniques can be used to obtain local optima as a more efficient alternative.

Step four: merge subproblem phylogenies into a phylogeny on the full set of taxa. Given an op-

Algorithm 1 Subproblem decomposition graph initialization and iteration

```
static variable C_r                                                    ▷ Number of reticulations

procedure InitializeSubproblemDecompositionGraph(D)
    Construct subproblem decomposition graph G_D such that E_D = {e_00} and w(e_00) = C_r
    return (G_D)

procedure NextSubproblemDecompositionGraph(G_D)
    Based on edge weights in G_D, construct corresponding multiset s of size C_r from (q+1 choose 2) + (q+1) elements
    if s is last multiset in enumeration over (((q+1 choose 2)+(q+1)+C_r−1) choose C_r) possible multisets then
        return false
    else
        Enumerate next multiset s'
        Update edge weights in G_D based on multiset s'
        return true
```

timal subproblem decomposition graph G'_D returned by the previous step, the final step of the FastNet algorithm merges the "top-level" phylogenetic structure inferred on D_0 and "bottom-level" subproblem phylogenies D_i for $1 \leq i \leq q$ (Algorithm 4). First, the phylogeny inferred on the top-level subproblem D_0 serves as the top-level of the output phylogeny N'. Next, the ith taxon in N' is replaced with the phylogeny inferred on bottom-level subproblem D_i, which was cached during the evaluation of G'_D. Finally, each "inter-subproblem" reticulation that was inferred for a pair of subproblems D_i and D_j where $i < j$ is added to the output phylogeny N', which is compatible by construction of the decomposition D and the optimal subproblem decomposition graph G'_D. The result of the merge procedure is an output phylogeny N' on the full set of taxa X.

Simulation study

We conducted a simulation study to evaluate the performance of FastNet and existing state-of-the-art methods for phylogenetic network inference. The performance study utilized the following procedures. Detailed commands and software options are given in the Supplementary Material (see Additional File 1).

Simulation of model networks. For each model condition, random model networks were generated using the following procedure. First, r8s version 1.7 [32] was used to simulate random birth-death trees with n taxa where $n \in \{15, 20, 25, 30\}$, which served as in-group taxa during subsequent analysis. The height of each tree was scaled to 5.0 coalescent units. Next, a time-consistent level-r rooted phylogenetic network [33] was obtained by adding r reticulations to each tree, where $r \in [1, 4]$. The procedure for adding a reticulation consists of the following steps: based on a consistent timing of events in the tree, (1) choose a time t_M uniformly at random between 0 and the tree height, (2) randomly select two tree edges for which corresponding ancestral populations existed during time interval $[t_A, t_B]$ such that $t_M \in [t_A, t_B]$, and (3) add a reticulation to connect the pair of tree edges. Finally, an outgroup was added to the resulting network at time 15.0.

Reticulations in our study have the same interpretation as in the study of Leaché et al. [34]. Gene flow is modeled using an isolation-with-migration model, where each reticulation is modeled as a unidirectional migration event with rate 5.0 during the time interval $[t_A, t_B]$. We focus on paraphyletic gene flow as described by Leaché et al.; their study also investi-

Algorithm 2 Calculate optimization score for a subproblem decomposition graph and infer subproblem solutions

```
static variable C_r                                                    ▷ Number of reticulations
static variable G                                                      ▷ Set of gene trees

procedure ComputeOptimizationScoreAndInferSubproblemSolutions(G_D, Δ, δ, Ψ, ψ)
    for i = 0 to q do
        InferSubnetwork(G_D, i, Δ, δ)                                  ▷ Caches inferred network in Δ
                                                                       ▷ Caches inferred network likelihood in δ
    for i = 0 to q do
        for j = i + 1 to q do
            InferSubnetwork(G_D, i, j, Δ, δ, Ψ, ψ)                     ▷ Caches inferred network in Ψ
                                                                       ▷ Caches inferred network likelihood in ψ
    score = ∏_{0≤i≤q} δ[i, w(G_D, i, i)] ∏_{0≤i≤q, i<j≤q} ψ[i, j, w(G_D, i, j), w(G_D, i, i), w(G_D, j, j)]   ▷ Pseudolikelihood score
                                                                       ▷ Note that w(G_D, i, j) = (e_ij ∈ E(G_D)) ? weight of edge e_ij : 0
    return (score)

procedure InferSubnetwork(G_D, i, Δ, δ)
    k = (e_ii ∈ E(G_D)) ? w(e_ii) : 0                                  ▷ e_ii ∈ E(G_D)
    if defined(Δ[i, k]) then
        return
    (N_i, score) = F(G_|D_i, k)                                        ▷ "Boosted" method F(·, ·)
                                                                       ▷ G_|D_i is restriction of G to subproblem taxa D_i
    Δ[i, k] = N_i                                                      ▷ Return value by reference to mutable cache Δ
    δ[i, k] = score                                                    ▷ Return value by reference to mutable cache δ

procedure InferSubnetwork(G_D, i, j, Δ, δ, Ψ, ψ)
    k = (e_ij ∈ E(G_D)) ? w(e_ij) : 0
    l = (e_ii ∈ E(G_D)) ? w(e_ii) : 0
    m = (e_jj ∈ E(G_D)) ? w(e_jj) : 0
    if defined(Ψ[i, j, k, l, m]) then
        return
    if not defined(Δ[i, l]) then
        InferSubnetwork(G_D, i, Δ, δ)
    if not defined(Δ[j, m]) then
        InferSubnetwork(G_D, j, Δ, δ)
    N^cherry = ConstructCherry(Δ[i, l], Δ[j, m])                       ▷ Returns (Δ[i, l]:b_i, Δ[j, m]:b_j);
                                                                       ▷ where b_i and b_j are inferred using "boosted method" F(·, ·)
    (N_ij, score) = AddReticulations(N^cherry, k)                      ▷ Using "boosted method" F(·, ·) to perform constrained search
    Ψ[i, j, k, l, m] = N_ij                                            ▷ Return value by reference to mutable cache Ψ
    ψ[i, j, k, l, m] = score                                           ▷ Return value by reference to mutable cache ψ
```

Algorithm 3 Search for optimal subproblem decomposition graph

```
static variable C_r                                                    ▷ Number of reticulations
static variable G                                                      ▷ Set of gene trees

procedure ExhaustiveSearchForOptimalSubproblemDecompositionGraph(D)
    q = GetNumberOfSubproblems(D)
    (Δ, δ, Ψ, ψ) = InitializeCaches(q, C_r)   ▷ Mutable caches persist during subproblem decomposition graph search
    G_D = InitializeSubproblemDecompositionGraph(D)
    G_D^best = G_D
    score^best = ComputeOptimizationScoreAndInferSubproblemSolutions(G_D, Δ, δ, Ψ, ψ)
    while NextSubproblemDecompositionGraph(G_D) do
        score = ComputeOptimizationScoreAndInferSubproblemSolutions(G_D, Δ, δ, Ψ, ψ)
        if score > score^best then
            G_D^best = G_D                                             ▷ Update
            score^best = score
    return(G_D^best, score^best, Δ, δ, Ψ, ψ)
```

Algorithm 4 Merge

```
static variable C_r                                                    ▷ Number of reticulations

procedure Merge(G_D, Δ, Ψ)
    q = GetNumberOfSubproblems(G_D)
    w00 = (e_00 ∈ E(G_D)) ? w(e_00) : 0
    N = Δ[0, w00]                                                      ▷ "Top-level" subproblem phylogeny
                                                ▷ Note that w(e) = (e ∈ E(G_D)) ? weight of edge e : 0
    for i = 1 to q do
        k = (e_ii ∈ E(G_D)) ? w(e_ii) : 0
        ReplaceLeafWithSubnetwork(N, G_D, i, Δ[i, k])   ▷ Replace ith taxon in "top-level" subproblem D_0
    for i = 1 to q do
        for j = i + 1 to q do
            k = (e_ij ∈ E(G_D)) ? w(e_ij) : 0
            l = (e_ii ∈ E(G_D)) ? w(e_ii) : 0
            m = (e_jj ∈ E(G_D)) ? w(e_jj) : 0
            if k > 0 then
                AddCompatibleReticulations(N, G_D, i, j, Ψ[i, j, k, l, m])
    return(N)
```

gated two other classes of gene flow – isolation-with-migration and ancestral gene flow – both of which involve gene flow between two sister species after divergence. Our simulation study omits these two classes since several existing methods (i.e., MLE and MPL) have issues with identifiability in this context. We note that FastNet makes no assumptions about the type of gene flow to be inferred, and identifiability depends on the model used for inference by FastNet's base method.

As in the study of Leaché et al., we further classify simulation conditions based on whether gene flow is "non-deep" or "deep" based on topological constraints. Non-deep reticulations involve leaf edges only, and all other reticulations are considered to be deep. Similarly, model conditions with non-deep gene flow have model networks with non-deep reticulations only; all other model conditions include deep reticulations and are referred to as deep.

Simulation of local genealogies and DNA sequences. We used ms [35] to simulate local gene trees for independent and identically distributed (i.i.d.) loci under an extended multi-species coalescent model, where reticulations correspond to migration events as described above. Each coalescent simulation sampled one allele per taxon. The primary experiments in our study simulated 1000 gene trees for each random model network. Our study also investigated data requirements of different methods by including additional datasets where either 200 or 100 gene trees were simulated for each random model network.

Sequence evolution was simulated using seq-gen [36], which takes the local genealogies generated by ms as input and simulates sequence evolution along each genealogy under a finite-sites substitution model. Our simulations utilized the Jukes-Cantor substitution model [37]. We simulated 1000 bp per locus, and the resulting multi-locus sequence alignment had a total length of 1000 kb.

Replicate datasets. A model condition in our study consisted of fixed values for each of the above model parameters. For each model condition, the simulation procedure was repeated twenty times to generate twenty replicate datasets.

Species network inference methods. Our simulation study compared the performance of FastNet against existing methods which were among the fastest and most accurate in our previous performance study of state-of-the-art species network inference methods [15]. Like FastNet, these methods perform summary-based inference – i.e., the input consists of gene trees inferred from sequence alignments for multiple loci, rather than the sequence alignments themselves. The methods are broadly characterized by their statistical optimization criteria: either maximum likelihood or maximum pseudo-likelihood under the multi-species network coalescent (MSNC) model [16]. The maximum likelihood estimation methods consisted of two methods proposed by Yu et al. [14] which are implemented in PhyloNet [29]. One method utilizes gene trees with branch lengths as input observations, whereas the other method considers gene tree topologies only; we refer to the methods as MLE-length and MLE, respectively. Our study also included the pseudo-likelihood-based method of [21], which we refer to as MPL. For each analysis in our study, all species network inference methods – MLE, MLE-length, MPL, and FastNet – were provided with identical inputs.

Our study included two categories of experiments. The "boosting" experiments in our simulation study compared the performance of FastNet against its base method; we refer to all other experiments in our study as "non-boosting". To make boosting comparisons explicit, each boosting experiment will refer to "FastNet(BaseMethod)" which is FastNet run with a specific base method "BaseMethod" – either MLE-length, MLE, or MPL. The input for each boosting experiment

consisted of either true or inferred gene trees for all loci. The inferred gene trees were obtained using FastTree [28] with default settings to perform maximum likelihood estimation under the Jukes-Cantor substitution model [37]. The inferred gene trees were rooted using the outgroup. The non-boosting experiments focused on the performance of FastNet using MLE as a base method and inferred gene trees as input, where gene trees were inferred using the same procedure as in the boosting experiments.

Performance measures. The species network inference methods in our study were evaluated using two different criteria. The first criterion was topological accuracy. For each method, we compared the inferred species phylogeny to the model phylogeny using the tripartition distance [38], which counts the proportion of tripartitions that are not shared between the inferred and model network. The second criterion was computational runtime. All computational analyses were run on computing facilities in Michigan State University's High Performance Computing Center. We used compute nodes in the intel16 cluster, each of which had a 2.5 GHz Intel Xeon E5-2670v2 processor with 64 GiB of main memory. All replicates completed with memory usage less than 16 GiB.

Empirical datasets

Yeast dataset. We re-analyzed the 23 yeast genomes from the phylogenomic study conducted by Salichos and Rokas [39]. We briefly summarize the procedures that Salichos and Rokas followed to process genomic sequence data. First, orthologs were identified using the Yeast Gene Order Browser (YGOB) [40] and Candida Gene Order Browser (CGOB) [41, 42] synteny databases. Nucleotide sequences were translated into amino acid sequences. The unaligned amino acid sequences were aligned using MAFFT with default settings [43]. Loci were filtered based on alignment length and quality, where the minimum alignment length was 150 bases and at least half of an alignment's sites consisted of bases only (i.e., had zero indels). The resulting dataset consisted of 1,070 loci.

Unrooted gene trees were inferred using the same procedures as in the study of Salichos and Rokas. For each locus, RAxML [44] was used to perform maximum likelihood estimation of a gene tree under a model of amino acid substitution. The substitution model was selected using ProtTest [45]. Following the procedures of Yu and Nakhleh [13], unrooted gene trees were rooted under the MDC criterion [21] using the species tree reported in [39].

FastNet(MPL) was used to estimate a species network topology using the rooted gene trees as input. We performed fixed-topology optimization of continuous model parameters associated with the species network (i.e. branch length parameters and admixture frequency parameters) using MPL to perform maximum pseudo-likelihood estimation. The corresponding pseudo-likelihoods were compared as part of a slope analysis. As in the study of Yu and Nakhleh, the inferred species networks contained 0, 1, or 2 reticulations.

Mosquito dataset. Our empirical study included a re-analysis of a phylogenomic dataset from the study of Neafsey et al. [46]. We obtained the original version of the dataset from the study authors (D. Neafsey, personal comm.). (An updated version of the dataset can be downloaded from the OrthoDBmoz2 database [46, 47] (`http://cegg.unige.ch/orthodbmoz2`)). The dataset consists of genomic sequence data for single copy orthologs that are present in 18 anopheline species. A total of 5,099 loci are included. The dataset was known as the "SC Universal" dataset in the study of Neafsey et al.. For each locus, RAxML was used to infer an unrooted gene tree under the General Time Reversible (GTR) [48] model of nucleotide substitution with a GAMMA+I model of rate heterogeneity

across sites [49]. Unrooted gene trees were rooted using the same procedure as in the yeast dataset re-analysis. Phylogenetic support was evaluated using RAxML to conduct a standard bootstrap analysis with 100 replicates. Gene tree edges with bootstrap support below a threshold of 90% were contracted. Summary-based inference of species networks followed the procedure used in the yeast dataset re-analysis. The inferred species networks contained between 0 and 2 reticulations.

Results

Simulation study

FastNet's use of phylogenetic divide-and-conquer is compatible with a range of different methods for inferring rooted species networks on subproblems, which we refer to as "base" methods. From a computational perspective, FastNet can be seen as a general-purpose framework for boosting the performance of base methods. We began by assessing the relative performance boost provided by FastNet when used with two different state-of-the-art network inference methods. We evaluated two different aspects of performance: topological error as measured by the tripartition distance [38] between an inferred species network and the model network, and computational runtime. The initial set of boosting experiments focused on species network inference in isolation of upstream inference accuracy by providing true gene trees as input to all of the summary-based inference methods.

In the performance study of Hejase and Liu [15], 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. MPL utilized a pseudo-likelihood-based approximation for increased computational efficiency compared with full likelihood methods [13]. However, the tradeoff netted efficiency that was well short of current phylogenomic dataset sizes [15].

Table 1 shows the performance of FastNet(MPL) relative to MPL on model conditions with increasing numbers of taxa and non-deep reticulations. On model conditions with dataset sizes ranging from 15 to 30 taxa and from 1 to 4 reticulations, FastNet(MPL)'s improvement in topological error relative to its base method was statistically significant (one-sided pairwise t-test with Benjamini-Hochberg correction for multiple tests [50]; $\alpha = 0.05$ and $n = 20$) and substantial in magnitude – an absolute improvement that amounted to as much as 41%. Furthermore, the improvement in topological error grew as datasets became larger and involved more reticulations: the largest improvements were seen on the 30-taxon 4-reticulation model condition. Runtime improvements were also statistically significant and represented speedups which amounted to as much as a day and a half of runtime.

Next, we evaluated FastNet's performance when boosting MLE-length, the most accurate state-of-the-art method from the performance study of Hejase and Liu [15]. On model conditions with non-deep reticulations, FastNet(MLE-length) had a similar boosting effect as compared to FastNet(MPL) (Table 2). On the 15-taxon single-reticulation model condition, FastNet's average improvement in topological error was greater when MLE-length was used as a base method rather than MPL. An even greater improvement in computational runtime was seen: FastNet(MLE-length)'s runtime improvement over MLE-length was over an order of magnitude greater than FastNet(MPL)'s improvement over MPL. As the number of taxa increased from 15 to 20 (but the number of reticulations was fixed to one), FastNet(MLE-length)'s advantage in topological error and runtime relative to its base method nearly doubled. In all cases, FastNet(MLE-length)'s performance improvements were statistically significant (Benjamini-Hochberg-corrected one-sided pairwise t-test; $\alpha = 0.05$

and $n = 20$). Although FastNet(MLE-length) successfully completed analysis of larger datasets (i.e., model conditions with more than 20 taxa and/or more than one reticulation), we were unable to quantify FastNet(MLE-length)'s performance relative to its base method due to MLE-length's scalability limitations.

We further evaluated FastNet's performance in the context of additional experimental and methodological considerations. On model conditions with deep gene flow (Table 3), FastNet returned significant improvements in topological accuracy and runtime relative to its base method – either MPL or MLE-length – with one exception: on the 15-taxon single-reticulation model condition, FastNet(MPL) returned a small and statistically insignificant improvement in topological error over MPL. Otherwise, FastNet's performance boost was robust to the choice of base method. As dataset sizes increased, the average performance boost increased when MPL was the base method; a similar finding applied to runtime improvements when MLE-length was the base method, whereas topological error improvements were largely unchanged. We note that FastNet's performance boost was somewhat smaller on model conditions involving deep gene flow as opposed to non-deep gene flow. When maximum-likelihood-estimated gene trees were used as input to summary-based inference in lieu of true gene trees (Table 4), FastNet boosted the topological accuracy and runtime of its base method in all cases and the improvements were statistically significant. As dataset sizes increased, FastNet's improvement in topological accuracy and runtime grew when MPL was its base method; runtime improvements grew and topological error improvements were largely unchanged when MLE-length was the base method. Finally, we conducted an additional experiment to evaluate FastNet's statistical efficiency when given a finite number of observations in terms of the number of loci (Table 5). As the number of loci ranged from genome-scale (i.e., on the order of 1000 loci) to sizes that were smaller by up to an order of magnitude, FastNet's average topological error increased by less than 0.02.

Empirical study

Yeast dataset. The yeast dataset in our empirical study was originally published by Salichos and Rokas [39] and re-analyzed by Yu and Nakhleh [13]. Comparing and contrasting the three studies reveals several important areas of strict and majority consensus.

As shown in Supplementary Figure S1, a slope analysis based on pseudo-likelihoods clearly preferred species networks with either one or two reticulations over a species tree hypothesis. Pseudo-likelihoods were calculated using either FastNet's approach (expression 1) or MPL's approach. Furthermore, a species network with two reticulations was preferred over a species network with one reticulation. The two-reticulation phylogeny inferred by FastNet(MPL) is shown in Figure 2. (The one-reticulation phylogeny inferred by FastNet(MPL) is shown in Supplementary Figure S2).

The species phylogenies inferred by the three studies largely agreed in terms of topology, with two subtle differences. First, FastNet(MPL) inferred tree edges in the Candida clade (i.e., the Candida lineages described in Figure 1 of [39]) that were identical to those inferred by Salichos and Rokas using concatenated analysis; these tree edges were nearly identical to those inferred by Yu and Nakhleh except for the placement of *Candida guiliermondii* and *Debaryomyces hansenii*. Tree edges in the other clade (i.e., the Saccharomyces lineages described in Figure 1 of [39]) were identical across all three studies. Second, the species networks inferred in our study and the study of Yu and Nakhleh largely agreed on the placement of reticulations. Both studies placed one reticulation within the Candida clade. As noted by Yu and Nakhleh, the tree

edges spanned by the reticulation had low support in the original study of Salichos and Rokas. In our study, the reticulation is consistent with gene flow involving an unsampled and divergent taxon. Both studies placed another reticulation within the other clade of the species phylogeny. The exact placement, orientation, and admixture frequency of reticulations differed somewhat between our study and the study of Yu and Nakhleh. In particular, the reticulations in our study were relatively deeper within the species phylogeny as compared to the study of Yu and Nakhleh. Based on pseudo-likelihoods calculated using MPL's optimization criterion, FastNet's two-reticulation species phylogeny had a larger pseudo-likelihood compared to the MPL-inferred two-reticulation topology reported by Yu and Nakhleh (Supplementary Table S2); a similar outcome was observed when comparing single-reticulation phylogenies. Of course, we note that FastNet addresses a different pseudo-likelihood optimization criterion compared to MPL.

Mosquito dataset. We re-analyzed genomic sequence data that was originally published by Neafsey et al. [46], which is a superset of the data studied by Fontaine et al. [51] and Yu and Nakhleh [13]. The Neafsey et al. dataset contains 18 in-group taxa in total – 5 of which are represented in the Fontaine et al. dataset, which has 6 in-group taxa in total.

Consistent with the studies of the smaller six-taxon dataset, a slope analysis of pseudo-likelihoods consistently preferred a species network hypothesis to a species tree hypothesis, and a two-reticulation network was preferred to a single-reticulation network (Supplementary Figure S3). Pseudo-likelihoods were calculated under either FastNet's optimization criterion (expression 1) or MPL's optimization criterion.

The two-reticulation species phylogeny inferred by FastNet(MPL) is shown in Figure 3, where reticulations are visualized in a manner similar to [13]. (The single-reticulation species phylogeny inferred by FastNet(MPL) is shown in Supplementary Figure S4.) Based on this interpretation, the FastNet-inferred species phylogeny is largely consistent with the other studies in terms of topology. Specifically, the FastNet-inferred topology encodes a tree that refines the species tree reported by Neafsey et al., which is fully resolved except for the clade corresponding to the Gambiae complex. Focusing on the 5 taxa within the Gambiae complex, the species phylogeny inferred by FastNet on the Neafsey et al. dataset has tree edges which agree with the species phylogeny reported by Fontaine et al. (shown as Figure 1C in [51]) and the MLE-inferred species phylogeny reported by Wen et al. [52], both of which were inferred using the 6-taxon Fontaine et al. dataset. We note that other interpretations are possible (*cf.* Figure 7 in [52]). Similar to the studies of Wen et al. and Fontaine et al., FastNet(MPL) infers gene flow within the clade corresponding to the Gambiae complex. The above interpretation indicates that the FastNet-inferred reticulation involves the *A. gambiae* lineage and the MRCA of *A. quadriannulatus* and *A. arabiensis*. The FastNet-inferred species phylogeny has an additional reticulation that is ancestral to those reported on the smaller Fontaine et al. dataset. This finding is consistent with the study of Wen et al., which inferred an ancestral reticulation involving an unsampled basal taxon based on analysis of the X chromosome (see Figures 4D and 7B in [52]). Due to the larger set of taxa used in our study, FastNet(MPL) inference was able to pinpoint the source endpoints for this reticulation within the expanded Anopheles phylogeny.

Discussion

Simulation study. Relative to the state-of-the-art methods that served as base methods, FastNet consistently returned sizeable and statistically significant improvements in topological error and computational runtime across a range of dataset scales and gene

flow scenarios. There was only a single experimental condition where comparable error without statistically significant improvements was seen. This exception occurred when FastNet was used to boost a relatively inaccurate base method (MPL) on the smallest dataset sizes in our study and with deep gene flow; even still, large and statistically significant runtime improvements were seen in this case. In contrast, with a more accurate base method (i.e., MLE-length), large and statistically significant performance improvements were seen throughout our simulation study.

FastNet's boosting effect on topological error and runtime were robust to several different experimental and design factors. The boosting performance obtained using different base methods – one with lower computational requirements but higher topological error relative to a more computationally intensive alternative – suggests that, while accuracy improvements can be obtained even using less accurate subproblem inference, even greater accuracy improvements can be obtained when reasonably accurate subproblem phylogenies can be inferred. We note that the base methods were run in default mode. More intensive search 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 a base method. In this sense, improvements to the latter accrue to the former in a straightforward and modular manner. Furthermore, FastNet's performance effect was robust to gene tree error and varying numbers of observed loci.

The biggest performance gains were observed on the largest, most challenging datasets – dataset scales which are becoming increasingly common in systematic studies. The findings in our earlier performance study [15] suggest that, given weeks of computational runtime, even the fastest statistical methods (including MPL) would not complete analysis of datasets with more than 50 taxa or so and several reticulations. In comparison to MPL, FastNet(MPL) was faster by more than an order of magnitude on the largest datasets in our study, and we predict that FastNet(MPL) would readily scale to datasets with many dozens of taxa and multiple reticulations.

Yeast dataset re-analysis. FastNet analyses of the genomic sequence dataset from the study of Salichos and Rokas [39] indicated that a yeast phylogeny involving gene flow and incomplete lineage sorting (ILS) is a more plausible hypothesis than one involving ILS alone. Our finding is consistent with the study of Yu and Nakhleh [13], and contrasts with the conclusions of Salichos and Rokas. Specifically, the distribution of observed local genealogies better reflects a multi-species network coalescent model [16] as opposed to a basic multi-species coalescent model. The former model's local genealogical distribution can be seen as a specific distortion of the latter model's distribution that is obtained using one or a few reticulations. Furthermore, our study and the study of Yu and Nakhleh found that a species phylogeny with two reticulations is preferred to one with a single reticulation.

The topologies of the species phylogenies in the three studies are largely identical, with two main differences. First, all three studies agree on tree edges, except for the placement of *C. guiliermondii* and *D. hansenii*. Our study agrees with the study of Salichos and Rokas regarding the placement of these two taxa, and differs from the study of Yu and Nakhleh. Second, there were some differences between our study and the study of Yu and Nakhleh regarding the placement of reticulations. In both studies, one reticulation involves Candida lineages and the other involves Saccharomyces lin-

eages. The two reticulations were deeper in the species phylogeny as compared to those inferred by Yu and Nakhleh. Our findings are consistent with the deeper hybridization events hypothesized by another recent phylogenomic study [53]. We note that neither our study nor the study of Yu and Nakhleh reconstructed recent hybridizations which have been described in the literature. Some of the putative hybridizations involve species which are not sampled in our dataset (e.g., *S. bayanus* is thought to be a hybrid of *S. cerevisiae* and two other taxa that are not sampled in our dataset [54]) and present identifiability issues noted above. It's also possible that other hybridization and/or introgression events occurred during yeast genome evolution, requiring the exploration of more complex phylogenetic hypotheses (i.e., networks with more reticulations) than those explored in either study. We also note that Salichos and Rokas filtered the loci used in their study to account for hidden paralogy as well as horizontal gene transfer and other types of gene flow (*cf.* "Data matrix construction" in Methods section of [39]). Salichos and Rokas assert that hybridization and introgression involving the filtered loci would require a specific set of evolutionary outcomes which is thought to be relatively rare. The findings from our study and the study of Yu and Nakhleh show strong support for an alternative hypothesis that runs contrary to this conventional wisdom. As suggested by Wolfe [55], a more nuanced understanding of the interplay between gene flow and other complex evolutionary events such as whole-genome duplication [56–58] awaits further phylogenomic study.

Finally, FastNet-inferred species networks were more optimal than MPL-inferred species networks when compared under MPL's statistical criterion (with the caveat that FastNet's subproblem decomposition graph optimization makes use of a different statistical criterion). These findings suggest that phylogenetic divide-and-conquer may prove to be a useful technique for network search under a variety of different optimization criteria.

Mosquito dataset re-analysis. Our re-analysis of the 18-taxon dataset published by Neafsey et al. [46] confirmed the historical introgression in the Anopheles phylogeny that was detected by Fontaine et al. [51] and Wen et al. [52] by analyzing a smaller 6-taxon dataset. The larger dataset was nearly a superset of the smaller dataset – only a single taxon in the Gambiae complex was present in the latter but not in the former. The FastNet analysis returned a network with tree edges which are compatible with the species tree reported in Neafsey et al.'s study. When restricted to shared taxa within the Gambiae complex, the FastNet phylogeny has an interpretation which agrees with the consensus tree reported by Fontaine et al. and Wen et al.'s findings. The reticulation scenarios inferred by the three studies were generally in agreement. FastNet analysis recovered a reticulation within the Gambiae complex, which agrees with the findings of Fontaine et al. and Wen et al.; the FastNet-inferred species phylogeny also includes another reticulation ancestral to the Gambiae clade, which is consistent with one of the hypotheses explored in the study of Wen et al. (*cf.* Figures 4D and 7B in [52]). The expanded superset of in-group taxa used in our study enables us to place all source nodes of the additional reticulation within the expanded Anopheles phylogeny. We note that Fontaine et al.'s geography-based subset analysis suggests both recent and ancient signatures of introgression. Furthermore, there is evidence of ancient introgression based on low-read-depth re-sequencing studies of natural populations. We attribute minor variations among inferred reticulations to methodological differences among the three studies. The taxa examined by the three studies are different: our study examines three times the number of taxa than are

present in the other studies. Apart from providing a case study for large-scale phylogenetic network estimation, greater taxon sampling in phylogenomic studies potentially provides additional phylogenetic signal relative to small-scale sampling. Other methodological differences include gene tree rooting and the approach used for selection and/or filtering genomic loci.

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 network inference methods. We found that FastNet improves upon existing methods in terms of computational efficiency and topological accuracy. On the largest datasets explored in our study, the use of the FastNet algorithm as a boosting framework enabled runtime speedups that were over an order of magnitude faster than standalone analysis using a state-of-the-art method. Furthermore, FastNet returned comparable or typically improved topological accuracy compared to the state-of-the-art-methods that were used as its base method.

Future enhancements to FastNet's algorithmic design are anticipated to yield additional performance improvements. Here we highlight several possibilities. First, recursive application of phylogenetic divide-and-conquer will allow "load balancing" of subproblem size and divergence, enabling further scalability gains. Second, the use of a better guide phylogeny should yield an improved subproblem decomposition and better subproblem decomposition graph optimization. The FastNet-inferred species phylogeny would be a suitable choice in this regard, as demonstrated by the topological error comparisons in our performance study. This insight naturally suggests an iterative approach: the output phylogeny from one iteration of the FastNet algorithm would be used as the guide phylogeny for a subsequent iteration of the algorithm. Iteration would continue until convergence under a suitable criterion (e.g., FastNet's pseudo-likelihood criterion). Third, we note that the FastNet algorithm is pleasantly parallelizable (as are some of the other state-of-the-art methods explored in our study).

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 which can 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. 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 [14]). We believe that FastNet's subproblem decomposition graph optimization suggests a way forward via alternative phylogenetic representations that summarize gene flow within "regions" of a phylogeny. One possibility would be to generalize tree-based concordance factors [59] for this purpose.

Availability of data and materials

Software source code and datasets generated and/or analyzed in our study are publicly available under open access licenses. A snapshot of the data and electronic materials is available from figshare under the GNU General Public License version 3 (GPLv3); the associated DOI is 10.6084/m9.figshare.5785479. Updated versions of the FastNet software as well as datasets associated with our study can be found in a GitLab repository that is hosted by Michigan State University. The repository can be accessed at `https://gitlab.msu.edu/liulab/FastNet.data.scripts`.

Repository files are distributed on an open access basis under the terms of the GNU General Public License version 3 (https://www.gnu.org/licenses/gpl-3.0.txt) and the Creative Commons Attribution-ShareAlike 4.0 International license (https://creativecommons.org/licenses/by-sa/4.0/).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceived and designed the experiments: HAH KJL. Implemented software tools: HAH. Performed simulation study experiments: HAH. Performed empirical study experiments: HAH NVP. Analyzed the data: GMB HAH KJL NVP. Wrote the paper: GMB HAH KJL NVP. All of the authors have read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge the following support: National Science Foundation (NSF) grants no. CCF-1565719 (to KJL), CCF-1714417 (to KJL), and DEB-1737898 (to GMB and KJL), grants from the BEACON Center for the Study of Evolution in Action (NSF STC Cooperative Agreement DBI-093954) to GMB and KJL, and Michigan State University (MSU) faculty startup funds (to GMB and to KJL). Computational analyses were performed on computing facilities maintained by MSU's High Performance Computing Center (HPCC). We would also like to acknowledge Daniel Neafsey for kindly sending us a processed version of the genomic sequence dataset from [46].

Author details

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Road, 11724 Cold Spring Harbor, NY, USA. [2]Department of Plant, Soil and Microbial Sciences, Michigan State University, 1066 Bogue Street, 48824 East Lansing, MI, USA. [3]Department of Computer Science and Engineering, Michigan State University, 428 S. Shaw Lane, 48824 East Lansing, MI, USA.

References

1. Michael L Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.
2. Scott V Edwards. Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19, 2009.
3. Luay Nakhleh. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, 28(12):719 – 728, 2013.
4. James O McInerney, James A Cotton, and Davide Pisani. The prokaryotic tree of life: past, present... and future? *Trends in Ecology & Evolution*, 23(5):276–281, 2008.
5. Patrick J Keeling and Jeffrey D Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8):605–618, 2008.
6. Richard J Abbott and Loren H Rieseberg. Hybrid speciation. In *Encyclopaedia of Life Sciences*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2012.
7. Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.
8. David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Paabo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.
9. The Heliconious Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94–98, 2012.
10. Kevin J. Liu, Ethan Steinberg, Alexander Yozzo, Ying Song, Michael H. Kohn, and Luay Nakhleh. Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences*, 112(1):196–201, 2015.
11. Emile Gluck-Thaler and Jason C Slot. Dimensions of horizontal gene transfer in eukaryotic microbial pathogens. *PLoS Pathogens*, 11(10): e1005156, 2015.
12. Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3):1–21, 03 2016.
13. Yun Yu and Luay Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(Suppl 10):S10, 2015.
14. Yun Yu, Jianrong Dong, Kevin J Liu, and Luay Nakhleh. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453, 2014.
15. Hussein A Hejase and Kevin J Liu. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, 17(1):422, 2016.
16. Yun Yu, James H. Degnan, and Luay Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, 8(4):e1002660, 04 2012.
17. John Frank Charles Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.
18. Jotun Hein, Mikkel Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*. Oxford

University Press, Oxford, 2004.

19. Hans-Jürgen Bandelt and Andreas WM Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92(1):47–105, 1992.
20. David Bryant and Vincent Moulton. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265, 2004.
21. Yun Yu, Cuong Than, James H Degnan, and Luay Nakhleh. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2):138–149, 2011.
22. John P Huelsenbeck and David M Hillis. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42(3):247–264, 1993.
23. Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410, 1978.
24. K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564, 2009.
25. Kevin Liu, Tandy J. Warnow, Mark T. Holder, Serita M. Nelesen, Jiaye Yu, Alexandros P. Stamatakis, and C. Randal Linder. SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic Biology*, 61(1):90–106, 2012.
26. Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, 22(5):377–386, 2015.
27. Eric Y. Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.
28. M. Price, P. Dehal, and A. Arkin. FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3): e9490, March 2010.
29. Cuong Than, Derek Ruths, and Luay Nakhleh. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(1):322, 2008.
30. Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.
31. Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Theo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17): i541–i548, 2014.
32. Michael J Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, 2003.
33. Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, Cambridge, United Kingdom, 2010.
34. Adam D Leaché, Rebecca B Harris, Bruce Rannala, and Ziheng Yang. The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, page syt049, 2013.
35. Richard R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.
36. A. Rambaut and N. C. Grassly. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13:235–238, 1997.
37. T.H. Jukes and C.R. Cantor. *Evolution of Protein Molecules*, pages 21–132. Academic Press, New York, NY, USA, 1969.
38. Luay Nakhleh, Jerry Sun, Tandy Warnow, C Randal Linder, Bernard ME Moret, and Anna Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Pacific Symposium on Biocomputing*, volume 8, pages 315–326. World Scientific, 2003.
39. Leonidas Salichos and Antonis Rokas. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449): 327–331, May 2013.
40. Kevin P Byrne and Kenneth H Wolfe. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, 15(10):1456–1461, 2005.
41. Sarah L Maguire, Seán S ÓhÉigeartaigh, Kevin P Byrne, Markus S Schröder, Peadar O'Gaora, Kenneth H Wolfe, and Geraldine Butler. Comparative genome analysis and gene finding in Candida species using CGOB. *Molecular Biology and Evolution*, 30(6):1281–1291, 2013.
42. David A Fitzpatrick, Peadar O'Gaora, Kevin P Byrne, and Geraldine Butler. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. *BMC Genomics*, 11(1):290, 2010.
43. Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
44. Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313, 2014.
45. Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164–1165, 2011.
46. Daniel E. Neafsey, Robert M. Waterhouse, Mohammad R. Abai, Sergey S. Aganezov, Max A. Alekseyev, James E. Allen, James Amon, Bruno Arcà, Peter Arensburger, Gleb Artemov, Lauren A. Assour, Hamidreza Basseri, Aaron Berlin, Bruce W. Birren, Stephanie A. Blandin, Andrew I. Brockman, Thomas R. Burkot, Austin Burt, Clara S. Chan, Cedric Chauve, Joanna C. Chiu, Mikkel Christensen, Carlo Costantini, Victoria L. M. Davidson, Elena Deligianni, Tania Dottorini, Vicky Dritsou, Stacey B. Gabriel, Wamdaogo M. Guelbeogo, Andrew B. Hall, Mira V. Han, Thaung Hlaing, Daniel S. T. Hughes, Adam M. Jenkins, Xiaofang Jiang, Irwin Jungreis, Evdoxia G. Kakani, Maryam Kamali, Petri Kemppainen, Ryan C. Kennedy, Ioannis K. Kirmitzoglou, Lizette L. Koekemoer, Njoroge Laban, Nicholas Langridge, Mara K. N. Lawniczak, Manolis Lirakis, Neil F. Lobo,

Ernesto Lowy, Robert M. MacCallum, Chunhong Mao, Gareth Maslen, Charles Mbogo, Jenny McCarthy, Kristin Michel, Sara N. Mitchell, Wendy Moore, Katherine A. Murphy, Anastasia N. Naumenko, Tony Nolan, Eva M. Novoa, Samantha O'Loughlin, Chioma Oringanje, Mohammad A. Oshaghi, Nazzy Pakpour, Philippos A. Papathanos, Ashley N. Peery, Michael Povelones, Anil Prakash, David P. Price, Ashok Rajaraman, Lisa J. Reimer, David C. Rinker, Antonis Rokas, Tanya L. Russell, N'Fale Sagnon, Maria V. Sharakhova, Terrance Shea, Felipe A. Simão, Frederic Simard, Michel A. Slotman, Pradya Somboon, Vladimir Stegniy, Claudio J. Struchiner, Gregg W. C. Thomas, Marta Tojo, Pantelis Topalis, José M. C. Tubio, Maria F. Unger, John Vontas, Catherine Walton, Craig S. Wilding, Judith H. Willis, Yi-Chieh Wu, Guiyun Yan, Evgeny M. Zdobnov, Xiaofan Zhou, Flaminia Catteruccia, George K. Christophides, Frank H. Collins, Robert S. Cornman, Andrea Crisanti, Martin J. Donnelly, Scott J. Emrich, Michael C. Fontaine, William Gelbart, Matthew W. Hahn, Immo A. Hansen, Paul I. Howell, Fotis C. Kafatos, Manolis Kellis, Daniel Lawson, Christos Louis, Shirley Luckhart, Marc A. T. Muskavitch, José M. Ribeiro, Michael A. Riehle, Igor V. Sharakhov, Zhijian Tu, Laurence J. Zwiebel, and Nora J. Besansky. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522, 2015.

47. Evgeny M Zdobnov, Fredrik Tegenfeldt, Dmitry Kuznetsov, Robert M Waterhouse, Felipe A Simão, Panagiotis Ioannidis, Mathieu Seppey, Alexis Loetscher, and Evgenia V Kriventseva. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, 45(D1):D744–D749, 2016.

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

49. Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

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

51. 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. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217): 1258522, 2015.

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

53. Marina Marcet-Houben and Toni Gabaldón. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology*, 13(8): e1002220, 2015.

54. Huu-Vang Nguyen, Jean-Luc Legras, Cécile Neuvéglise, and Claude Gaillardin. Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of *Saccharomyces bayanus* strains NBRC1948 and CBS380T. *PLoS One*, 6(10):e25821, 2011.

55. Kenneth H Wolfe. Origin of the yeast whole-genome duplication. *PLoS Biology*, 13(8):e1002221, 2015.

56. Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624, 2004.

57. Fred S. Dietrich, Sylvia Voegeli, Sophie Brachat, Anita Lerch, Krista Gates, Sabine Steiner, Christine Mohr, Rainer Pöhlmann, Philippe Luedi, Sangdun Choi, Rod A. Wing, Albert Flavier, Thomas D. Gaffney, and Peter Philippsen. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, 304(5668):304–307, 2004.

58. Bernard Dujon, David Sherman, Gilles Fischer, Pascal Durrens, Serge Casaregola, Ingrid Lafontaine, Jacky de Montigny, Christian Marck, Cécile Neuvéglise, Emmanuel Talla, Nicolas Goffard, Lionel Frangeul, Michel Aigle, Véronique Anthouard, Anna Babour, Valérie Barbe, Stéphanie Barnay, Sylvie Blanchin, Jean-Marie Beckerich, Emmanuelle Beyne, Claudine Bleykasten, Anita Boisramé, Jeanne Boyer, Laurence Cattolico, Fabrice Confanioleri, Antoine de Daruvar, Laurence Despons, Emmanuelle Fabre, Cécile Fairhead, Hélène Ferry-Dumazet, Alexis Groppi, Florence Hantraye, Christophe Hennequin, Nicolas Jauniaux, Philippe Joyet, Rym Kachouri, Alix Kerrest, Romain Koszul, Marc Lemaire, Isabelle Lesur, Laurence Ma, Héloïse Muller, Jean-Marc Nicaud, Macha Nikolski, Sophie Oztas, Odile Ozier-Kalogeropoulos, Stefan Pellenz, Serge Potier, Guy-Franck Richard, Marie-Laure Straub, Audrey Suleau, Dominique Swennen, Fredj Tekaia, Micheline Wésolowski-Louvel, Eric Westhof, Bénédicte Wirth, Maria Zeniou-Meyer, Ivan Zivanovic, Monique Bolotin-Fukuhara, Agnès Thierry, Christiane Bouchier, Bernard Caudron, Claude Scarpelli, Claude Gaillardin, Jean Weissenbach, Patrick Wincker, and Jean-Luc Souciet. Genome evolution in yeasts. *Nature*, 430:35 EP –, Jul 2004.

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

60. Toni Gabaldón, Tiphaine Martin, Marina Marcet-Houben, Pascal Durrens, Monique Bolotin-Fukuhara, Olivier Lespinet, Sylvie Arnaise, Stéphanie Boisnard, Gabriela Aguileta, Ralitsa Atanasova, Christiane Bouchier, Arnaud Couloux, Sophie Creno, Jose Almeida Cruz, Hugo Devillers, Adela Enache-Angoulvant, Juliette Guitard, Laure Jaouen, Laurence Ma, Christian Marck, Cécile Neuvéglise, Eric Pelletier, Amélie Pinard, Julie Poulain, Julien Recoquillay, Eric Westhof, Patrick Wincker, Bernard Dujon, Christophe Hennequin, and Cécile Fairhead. Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics*, 14(1):623, Sep 2013.

61. Daniel H Huson and Celine Scornavacca. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biology*, 61(6):1061–1067, 2012.

Figures

Tables

Additional Files

Additional file 1 — Supplementary Material

This file contains supplementary text, supplementary figures, and supplementary tables.

[scale=0.25]figures/high-level-illustration-of-basic-algorithm.pdf

Figure 1 A high-level illustration of the FastNet algorithm. First, a guide phylogeny $N^{(0)}$ is inferred on the full set of taxa X. Next, the guide phylogeny $N^{(0)}$ is used to decompose X into subproblems $\{D_0, D_1, D_2, \ldots, D_{q-1}, D_q\} = D$. Then, the subproblem decomposition D is used to construct a bipartite graph $G_D = (V_D, E_D)$, which is referred to as the subproblem decomposition graph. The set of vertices V_D consist of two partitions: source vertices $V_D^{\text{src}} = \{C_0^{\text{src}}, C_1^{\text{src}}, \ldots, C_q^{\text{src}}\}$ where each subproblem D_i has a corresponding source vertex C_i^{src}, and destination vertices $V_D^{\text{dst}} = \{C_0^{\text{dst}}, C_1^{\text{dst}}, \ldots, C_q^{\text{dst}}\}$ similarly. The subproblem decomposition graph G_D is optimized to infer subproblem phylogenies and reticulations, where the latter are inferred based on the placement of weighted edges $e \in E_D$. Finally, the subproblem phylogenies are merged using the phylogeny inferred on D_0 as the "top-level" structure.

[scale=0.45]figures/yeast2-crop.pdf

Figure 2 The two-reticulation species phylogeny inferred by FastNet(MPL) on the yeast dataset. The yeast dataset was originally published by Salichos and Rokas [39], and Yu and Nakhleh [13] re-analyzed the dataset using MPL. The phylogeny is interpreted and visualized in a manner similar to Figure 3 in [13]. Branch lengths (blue text) are given in coalescent units. Reticulation edges (dashed red lines) are annotated with admixture frequencies (red text). The two-reticulation species network was preferred to phylogenetic hypotheses involving fewer reticulations (see Supplementary Figures S1 and S2 concerning the latter). For reference, the putative whole genome duplication event described by Gabaldón et al. (*cf.* Figure 1 in [60]) would be placed on the branch incident to the MRCA of the sampled Saccharomyces taxa and *Candida glabrata* (i.e., the branch with length 0.22). The Dendroscope software package [61] was used in the process of producing the illustration.

[scale=0.45]figures/mosquito2.pdf

Figure 3 The two-reticulation species phylogeny inferred by FastNet(MPL) on the mosquito dataset. The mosquito dataset was originally published by Neafsey et al. [46]. The two-reticulation species network was preferred to phylogenetic hypotheses involving fewer reticulations (see Supplementary Figures S3 and S4 concerning the latter). Otherwise, figure layout and description are identical to Figure 2.

Table 1 FastNet(MPL) "boosts" MPL's runtime and topological accuracy, where a greater performance boost occurs as dataset sizes increase. The relative performance of FastNet(MPL) and MPL is compared on model conditions with 15-30 taxa and 1-4 non-deep reticulations. The performance measures consisted of topological error as measured by the tripartition distance between an inferred species network and the model network and computational runtime in hours. Average ("Avg") and standard error ("SE") of FastNet(MPL)'s performance improvement over MPL is reported ($n = 20$). All methods were provided with true gene trees as input. The statistical significance of FastNet(MPL)'s performance improvement over MPL was assessed using a one-sided t-test. Corrected q-values are reported where multiple test correction was performed using the Benjamini-Hochberg method [50].

Number of taxa	Number of reticulations	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	Corrected q-value	Avg	SE	Corrected q-value
15	**1**	0.087	0.036	3.3×10^{-2}	2.8	0.3	7.2×10^{-5}
20	**2**	0.346	0.036	1.1×10^{-5}	9.6	0.1	1.1×10^{-2}
25	**3**	0.281	0.024	7.9×10^{-5}	35.6	5.6	8.5×10^{-4}
30	**4**	0.413	0.001	8.8×10^{-12}	30.3	6.5	2.8×10^{-2}

Table 2 FastNet(MLE-length) "boosts" MLE-length's runtime and topological accuracy, where a greater performance boost occurs as dataset sizes increase. The relative performance of FastNet(MLE-length) and MLE-length is compared on model conditions with 15-20 taxa and 1-2 non-deep reticulations. Note that, for the model condition with 20 taxa and 2 reticulations, MLE-length did not finish analysis of any replicates after a week of runtime. Otherwise, table layout and description are identical to Table 1.

Number of taxa	Number of reticulations	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	Corrected q-value	Avg	SE	Corrected q-value
15	**1**	0.103	0.021	8.8×10^{-4}	49.4	6.9	9.1×10^{-7}
20	**1**	0.195	0.024	6.1×10^{-5}	114.3	14.7	3.3×10^{-7}
20	**2**	Base method DNF					

Table 3 Boosting experiments on model conditions with deep gene flow. The performance improvement of FastNet over its base method (either MPL or MLE-length) is reported for two different performance measures: topological error as measured by tripartition distance and computational runtime in hours. The simulation conditions involved either 15 or 20 taxa and a single deep reticulation. Otherwise, table layout and description are identical to Table 1.

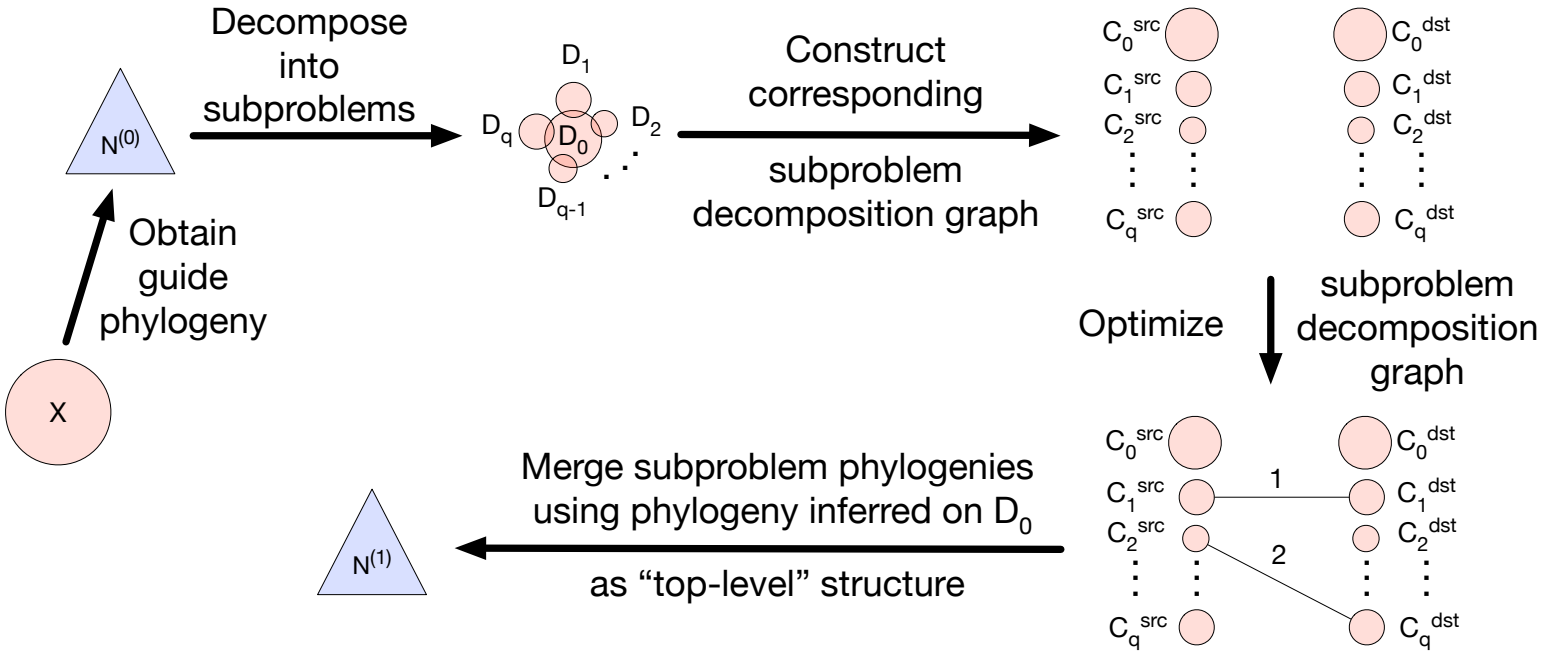
Number of taxa	Boosted method	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	q-value	Avg	SE	q-value
15	MPL	0.015	0.017	3.8×10^{-1}	2.3	0.2	5.1×10^{-4}
20	MPL	0.166	0.035	3.2×10^{-3}	8.0	1.5	3.2×10^{-3}
15	MLE-length	0.066	0.001	1.5×10^{-2}	35.0	4.1	1.3×10^{-7}
20	MLE-length	0.070	0.014	1.1×10^{-2}	71.1	7.7	8.7×10^{-8}

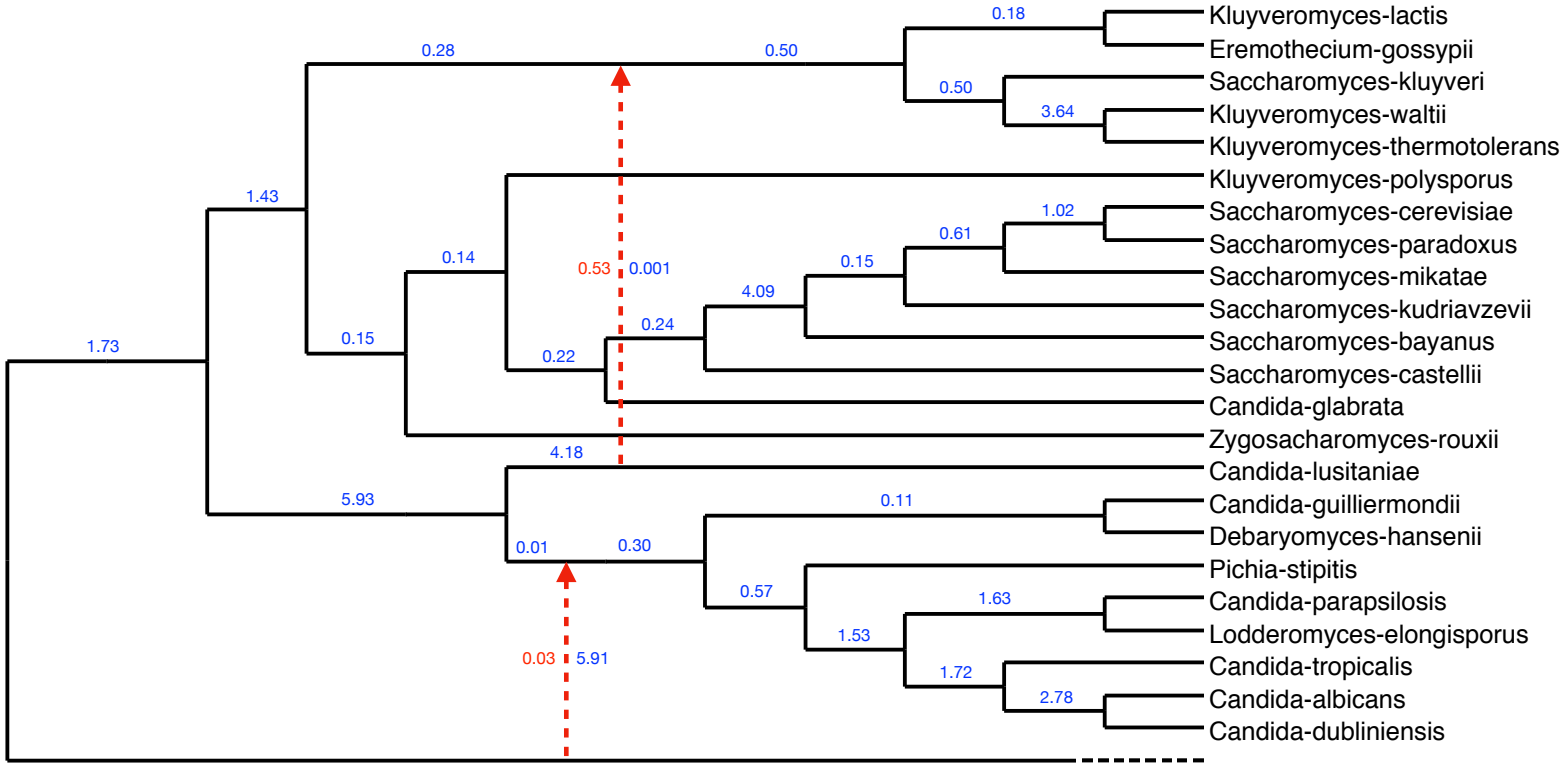
Table 4 Boosting experiments using inferred gene trees. The performance improvement of FastNet over its base method (either MPL or MLE-length) is reported for two different performance measures: topological error as measured by tripartition distance and computational runtime in hours. For each replicate dataset, all summary-based methods were provided with the same input: a set of rooted gene trees that was inferred using FastTree and outgroup rooting (see Methods section for more details). The simulation conditions involved either 15 or 20 taxa and 1-2 non-deep reticulations. Otherwise, table layout and description are identical to Table 1.

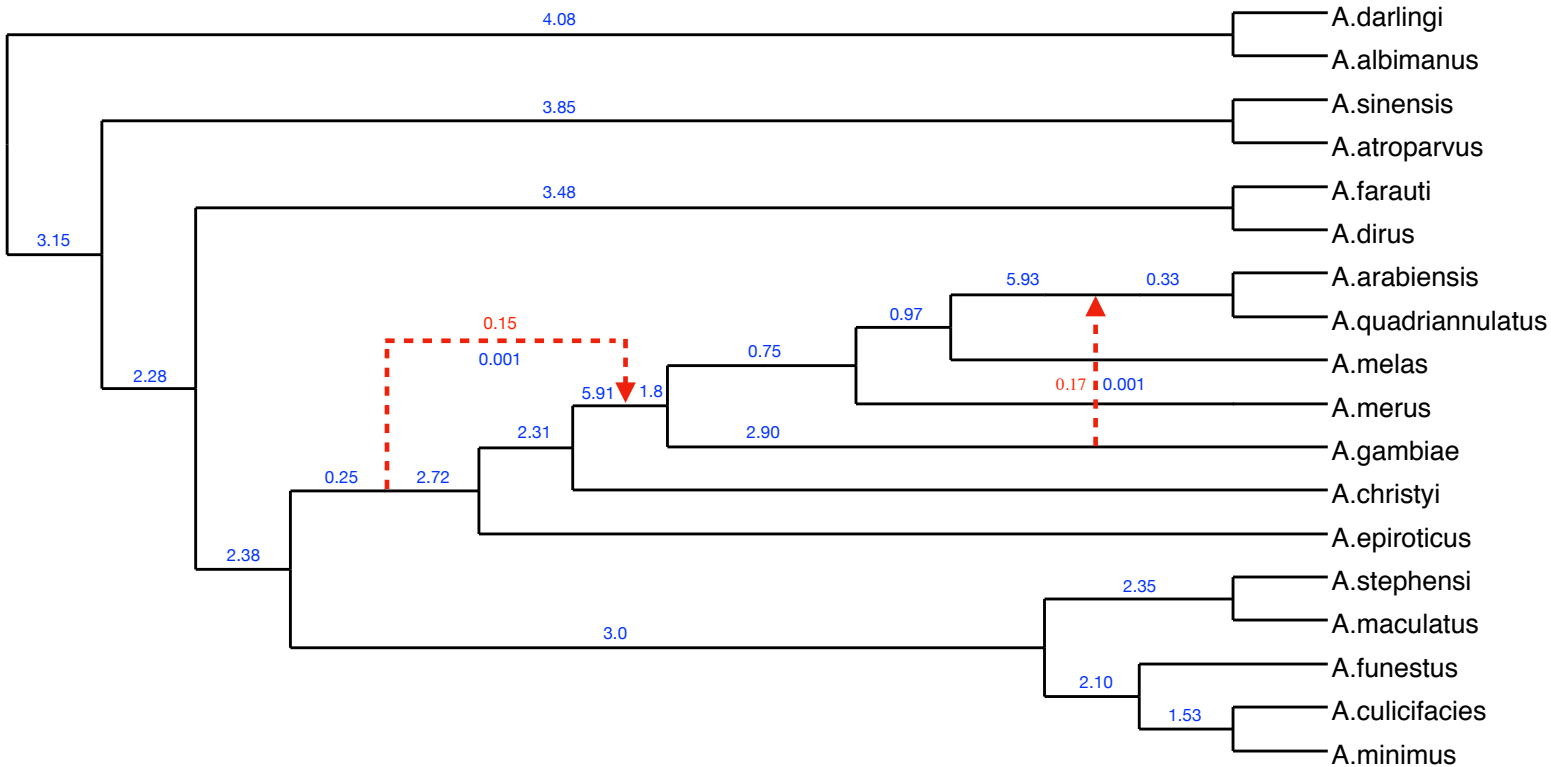
Number of taxa	Number of reticulations	Boosted method	Improvement in topological error			Improvement in runtime (h)		
			Avg	SE	q-value	Avg	SE	q-value
15	**1**	MPL	0.071	0.021	1.2×10^{-2}	3.8	0.5	7.7×10^{-5}
20	**2**	MPL	0.134	0.017	1.4×10^{-2}	15.1	1.7	6.9×10^{-6}
15	**1**	MLE-length	0.231	0.002	1.3×10^{-4}	15.4	2.0	6.7×10^{-7}
20	**1**	MLE-length	0.195	0.005	5.8×10^{-5}	43.2	7.3	1.7×10^{-5}

Table 5 The impact of the number of observed loci on FastNet(MLE)'s topological error. The inputs to FastNet(MLE) consisted of gene trees that were inferred using FastTree and outgroup rooting (see Methods section for more details). The simulations sampled between 100 and 1000 loci for a single 20-taxon 1-reticulation model condition involving non-deep gene flow. Topological error was evaluated based upon the tripartition distance between the model phylogeny and the species phylogeny inferred by FastNet(MLE); average ("Avg") and standard error ("SE") are shown ($n = 20$).

Number of loci	Topological error	
	Avg	SE
100	0.094	0.028
200	0.078	0.024
1000	0.075	0.027


Decompose into subproblems
$N^{(0)}$
D_1
D_2
D_q
D_0
D_{q-1}
Construct corresponding subproblem decomposition graph
C_0^{src}
C_1^{src}
C_2^{src}
C_q^{src}
C_0^{dst}
C_1^{dst}
C_2^{dst}
C_q^{dst}
Optimize
subproblem decomposition graph
1
2
Merge subproblem phylogenies using phylogeny inferred on D_0 as "top-level" structure
$N^{(1)}$
Obtain guide phylogeny
X


Kluyveromyces-lactis
Eremothecium-gossypii
Saccharomyces-kluyveri
Kluyveromyces-waltii
Kluyveromyces-thermotolerans
Kluyveromyces-polysporus
Saccharomyces-cerevisiae
Saccharomyces-paradoxus
Saccharomyces-mikatae
Saccharomyces-kudriavzevii
Saccharomyces-bayanus
Saccharomyces-castellii
Candida-glabrata
Zygosacharomyces-rouxii
Candida-lusitaniae
Candida-guilliermondii
Debaryomyces-hansenii
Pichia-stipitis
Candida-parapsilosis
Lodderomyces-elongisporus
Candida-tropicalis
Candida-albicans
Candida-dubliniensis
0.18
0.28
0.50
0.50
3.64
1.43
1.02
0.61
0.14
0.53
0.001
0.15
4.09
0.24
1.73
0.15
0.22
4.18
5.93
0.11
0.01
0.30
0.57
1.63
1.53
0.03
5.91
1.72
2.78


4.08
3.85
3.48
3.15
2.28
2.38
0.15
0.001
5.91
1.8
0.75
0.97
5.93
0.33
0.17
0.001
2.31
2.90
0.25
2.72
3.0
2.35
2.10
1.53
A.darlingi
A.albimanus
A.sinensis
A.atroparvus
A.farauti
A.dirus
A.arabiensis
A.quadriannulatus
A.melas
A.merus
A.gambiae
A.christyi
A.epiroticus
A.stephensi
A.maculatus
A.funestus
A.culicifacies
A.minimus