Assessing the potential of gene tree parsimony for microbial phylogenomics

Samson Weiner[1], Yutian Feng[2], J. Peter Gogarten[2,3], and Mukul S. Bansal[1,3]

[1] School of Computing, University of Connecticut, Storrs, USA
[2] Department of Molecular and Cell Biology, University of Connecticut, Storrs, USA
[3] Institute for Systems Genomics, University of Connecticut, Storrs, USA
samson.weiner@uconn.edu, yutian.feng@uconn.edu, gogarten@uconn.edu, mukul.bansal@uconn.edu

Abstract. A key challenge in microbial phylogenomics is that microbial gene families are often affected by extensive horizontal gene transfer (HGT). As a result, most existing methods for microbial phylogenomics can only make use of a small subset of the gene families present in the microbial genomes under consideration, potentially biasing their results and affecting their accuracy. One well-known approach for truly genome-scale phylogenomics is gene tree parsimony (GTP), which takes as input a collection of gene trees and finds a species tree that most parsimoniously reconciles with the input gene trees. While GTP based methods are widely used for phylogenomic studies of non-microbial species, their underlying reconciliation models are not designed to handle HGT and, therefore, they cannot be meaningfully applied to microbes. No GTP based methods have yet been developed for microbial phylogenomics.
In this work, we (i) design and implement the first GTP based approach, *PhyloGTP*, for microbial phylogenomics, (ii) use an extensive simulation study to systematically assess the accuracies of PhyloGTP and two other recently developed methods, SpeciesRax and ASTRAL-Pro-2, under a range of different conditions, and (iii) analyze two real microbial datasets with different characteristics. We find that PhyloGTP and SpeciesRax are more accurate than ASTRAL-Pro-2 across nearly all tested conditions, that PhyloGTP and SpeciesRax have similar accuracies overall, but there are conditions under which PhyloGTP consistently outperforms SpeciesRax, and that both PhyloGTP and SpeciesRax can sometimes yield incorrect, misleading phylogenies on complex real datasets.

Keywords: Phylogenomics · Microbial evolution · Gene tree parsimony

1 Introduction

The accurate inference of phylogenetic relationships between different microbes is an important problem in evolutionary biology. A key difficulty in estimating such phylogenies is the presence of extensive horizontal gene transfer (HGT) in microbial evolutionary histories. This can result in markedly different evolutionary histories for different gene families, obfuscating the underlying species-level

or strain-level phylogeny. As a result, the traditional approach for reconstructing microbial phylogenies is to use only "well-behaved" gene families resistant to HGT. This includes the use of small-subunit ribosomal RNA genes (e.g., [48,65]) or of a concatenated alignment of a few core genes from the genomes of interest (e.g., [14,35,41]). Both these approaches, however, are known to be error-prone. For instance, ribosomal RNA genes are known to engage in horizontal transfer [24, 66, 68] and to yield histories that are inconsistent with those inferred using other core genes [17, 18, 29, 30]. Furthermore, ribosomal RNA genes often cannot be used when studying closely related species due to excessive sequence similarity. Similarly, concatenation based approaches, such as the widely used multilocus sequence analysis (MLSA) technique [23], essentially ignore horizontal gene transfer and aggregate the phylogenetic signal from several gene families with potentially distinct evolutionary histories [22,44]. Indeed, the tree resulting from the concatenation might represent neither the organismal phylogeny nor any of the genes included in the concatenation [36].

To overcome these limitations, several genome-scale methods have also been proposed for microbial phylogeny inference. These include methods such as Phylo SI that are based on gene order information [54, 55], supertree-based methods such as SPR supertrees [64] and MRP [8, 68] that allow for the use of multiple orthologous gene families, and methods based on average nucleotide identity (ANI) of genomes [26, 28, 33]. Such genome-scale methods are inherently preferable to methods that base phylogeny reconstruction on only a single gene or a small set of concatenated genes [44]. However, while these above methods all represent useful approaches for microbial phylogenomics, they are either targeted at analyzing closely related strains or species (gene order and ANI based methods), or are limited to using single-copy gene families or orthologous groups and do not model key evolutionary events affecting microbial gene family evolution (supertree based methods). Recently, truly genome-scale approaches for microbial phylogenomics, capable of using thousands of complete (multi-copy) gene families, have also been developed. The two most prominent such methods are ASTRAL-Pro 2 [67] and SpeciesRax [46], both of which take as input a collection of unrooted gene family trees, where each gene family tree may contain zero, one, or multiple genes from any species/strain under consideration. ASTRAL-Pro 2 is based on quartets and seeks a species tree that maximizes a quartet based score [67]. While ASTRAL-Pro 2 does not directly model any specific evolutionary processes, such as HGT or gene duplication, responsible for gene tree discordance, it can handle complete (multi-copy) gene families and previous research suggests that it's quartet based approach should be robust to HGT [16]. SpeciesRax uses an explicit Duplication-Transfer-Loss model of gene family evolution in microbes and seeks a species tree that maximizes the reconciliation likelihood of observing the input gene trees under that model [46].

In this work, we propose a new approach for microbial phylogenomics and systematically compare its performance with ASTRAL-Pro 2 and SpeciesRax using simulated and real datasets. The new approach, called *PhyloGTP*, is based on *gene tree parsimony* (GTP), a well-known technique for phylogenomic in-

ference. GTP provides a framework for inferring species trees from a collection of gene trees impacted by complex evolutionary processes. Specifically, GTP seeks a species tree that most parsimoniously reconciles all the input gene trees under an appropriately chosen model of gene-tree/species-tree reconciliation. Facilitated by effective software implementations [13, 62], GTP is widely used for phylogenomic studies of multicellular eukaryotes (e.g., [12, 27, 40, 42, 43]), where the most appropriate reconciliation model is often the duplication-loss (DL) model [25]. To apply GTP to microbes, one must account for HGT by using the more complex Duplication-Transfer-Loss (DTL) reconciliation model. Despite its promise, GTP has not yet been implemented with DTL reconciliation and has therefore not yet been applied to microbial genomes. PhyloGTP addresses this gap, allowing for the first systematic assessment of GTP's potential for microbial phylogenomics. We note that PhyloGTP is conceptually similar to SpeciesRax since both methods are based on explicit DTL models of microbial gene family evolution, and both methods seek species trees that best reconcile the input gene trees under their DTL models. However, there are two key differences between PhyloGTP and SpeciesRax: First, PhyloGTP uses a standard, widely-used parsimony-based DTL model [3, 60] while SpeciesRax uses a different, probabilistic DTL model [46]. And second, PhyloGTP and SpeciesRax use different heuristic search strategies to find their best species tree estimates, as we discuss later.

We use an extensive simulation study to evaluate the accuracies of PhyloGTP, ASTRAL-Pro 2, and SpeciesRax, focusing especially on the impact of number of input gene trees, DTL rates, and input gene tree error rates. We find that PhyloGTP and SpeciesRax are more accurate than ASTRAL-Pro-2 across nearly all tested conditions, that PhyloGTP often substantially outperforms SpeciesRax when the number of input gene trees is small or when DTL rates are high, and that SpeciesRax generally outperforms PhyloGTP on datasets with high gene tree error but low DTL rates. We also used PhyloGTP and SpeciesRax to analyze two real microbial datasets; a more complex 174-taxon Archaeal dataset exhibiting extreme divergence and compositional biases, and a less complex dataset of 44 Frankiales exhibiting low divergence. While both PhyloGTP and SpeciesRax perform well on these real datasets, they do result in a few clearly incorrect placements for the Archaeal dataset. This suggests that both PhyloGTP and SpeciesRax are potentially susceptible to biases present in complex datasets.

While our prototype implementation of PhyloGTP is considerably slower than SpeciesRax, our results establish GTP as a promising approach for microbial phylogenomics, and show that PhyloGTP is capable of yielding more accurate microbial species trees for many datasets. At the same time, our results show that even reconciliation-based phylogenomic approaches like PhyloGTP and SpeciesRax may not produce accurate results for certain complex microbial datasets and that their results should be interpreted with caution. An open-source prototype implementation of PhyloGTP is available from `https://github.com/samsonweiner/PhyloGTP`.

2 Basic Definitions and Preliminaries

Let T be a leaf-labeled tree with node, edge, and leaf sets denoted by $V(T)$, $E(T)$, and $Le(T)$. If T is rooted, we denote it's root by $rt(T)$. For any node $v \in V(T)$, where T is a rooted tree, the (maximal) subtree rooted at v is denoted T_v. Unless otherwise specified, all trees are binary and unrooted.

We use the term *species tree* for the tree depicting evolutionary relationships for the taxa (e.g., species, strains, etc.) under consideration. Given a gene family from the taxa under consideration, a *gene tree* is a tree that depicts the evolutionary relationships of the genes in the gene family. We assume that each leaf in a gene tree is labeled with the taxon from which that leaf (i.e., gene sequence) was taken. Note that a gene tree may have zero, one, or multiple genes from the same taxon.

Throughout this work, we assume that the taxon set under consideration is denoted by Ω and that the species tree, denoted S, depicts the evolutionary relationships for taxa in Ω, i.e., $Le(S) = \Omega$. We use $\mathcal{G}$ to denote a collection of gene trees $\{G_1, ..., G_k\}$, where each G_i, $1 \leq i \leq k$, describes the evolutionary history of a different gene family present in the taxon set Ω. We also implicitly assume that $Le(S) = \cup_{i=1}^{k} Le(G_i)$.

DTL reconciliation. The DTL reconciliation model allows for the reconciliation of a given rooted gene tree with a given rooted species tree by postulating gene duplication, HGT, and gene loss events. DTL reconciliation is often performed in a maximum parsimony framework, in which each event type has an associated (user-defined) cost and the objective is to find a reconciliation of minimum total cost [3,15,19,58–60]. In the current work, we specifically use the DTL reconciliation model first developed in [3,60], for which optimal (most parsimonious) DTL reconciliations can be computed in $O(mn)$ time, where m and n denote the number of leaves in the gene tree and species tree being reconciled, respectively. Importantly, an unrooted gene tree can be reconciled with a rooted species tree within the same $O(mn)$ time complexity [3].

In the following, we denote the event costs for gene duplications, HGTs, and gene losses by P_d, P_t, and P_l, respectively. Given a gene tree $G \in \mathcal{G}$, species tree S, and event costs P_d, P_t, and P_l, we denote by $\mathcal{R}_{P_d,P_t,P_l}(G, S)$ the reconciliation cost of an optimal DTL reconciliation of G and S under the event costs P_d, P_t, and P_l.

Definition 1 (Total DTL Reconciliation Cost). *Given a species tree S, a collection of gene trees $\mathcal{G} = \{G_1, ..., G_k\}$, and event costs P_d, P_t, and P_l, the* total DTL reconciliation cost *of $\mathcal{G}$ with S is the sum of the DTL reconciliation costs of each $G \in \mathcal{G}$ with S, i.e., $\sum_{i=1}^{k} \mathcal{R}_{P_d,P_t,P_l}(G_i, S)$.*

GTP-based Problem formulation. To compute accurate genome-scale microbial phylogenies, we use a gene tree parsimony formulation based on DTL reconciliation. Specifically, given as input a collection of hundreds or thousands of gene trees, we seek a species tree that minimizes the total DTL reconciliation cost against the collection of input gene trees. More formally,

Problem 1 (Most Parsimonious Species Tree (MPST)) *Given a collection of gene trees $\mathcal{G}$ and event costs P_d, P_t, and P_l, find a species tree S that minimizes the total DTL reconciliation cost with $\mathcal{G}$.*

The MPST problem can be shown to be NP-hard, W[2]-hard, and inapproximable to within log factor through a reduction from the NP-hard gene duplication problem [6, 38]. The gene duplication problem is a special case of MPST problem defined in this manuscript and seeks a species tree minimizing just the total number of gene duplications. Details of the reduction are straightforward and omitted for brevity. Given the NP-hardness of the MPST problem, PhyloGTP uses a local search heuristic to solve the problem, as described in the next section.

3 Description of PhyloGTP

The local search heuristic implemented in PhyloGTP is similar to those use for many other NP-hard phylogeny inference problems, including those used for other popular variants of gene tree parsimony [13, 39, 49, 62]. The local search heuristic starts with an initial candidate rooted species tree and iteratively improves it using local search. Specifically, in each local search iteration, the heuristic finds a minimum reconciliation cost tree in the "local neighborhood" of the current species tree. The best tree found in that local neighborhood then becomes the starting point for the next local search iteration. The heuristic terminates when a lower cost tree cannot be found in the local neighborhood of the current species tree. Next, we describe how PhyloGTP computes the initial candidate species tree and how it defines the local neighborhood for each subsequent local search iteration.

Construction of initial candidate species tree. If an estimated user-defined initial species tree is unavailable, PhyloGTP uses a stepwise taxon-addition algorithm to compute a reasonable initial species tree for the local search. The stepwise taxon-addition algorithm works by starting from a two-taxon rooted species tree and iteratively placing taxa, one at a time, onto the species tree topology along the branch that minimizes the total DTL reconciliation cost. In our implementation, the taxa are added in order of decreasing coverage, where the coverage of taxon s is the number of gene trees that include a gene from s. At each iteration, each gene tree is pruned to reflect only the taxa present in the current (incomplete) species tree. Once all taxa have been added, the resulting rooted species tree is used as the starting species tree for the subsequent local search. We found that using this stepwise taxon-addition algorithm results in an average reduction of 93% in the number of local search iterations until convergence when compared to using a random species tree topology as the initial starting tree (detailed results not shown).

Description of local search iterations. PhyloGTP implements a constrained (rooted) subtree prune and regraft (SPR) [9] based local search using the initial

tree as a starting point. SPR is the most commonly used tree edit operation for phylogenetic local search and induces a local neighborhood of $\Theta(n^2)$ trees, where n is the number of leaves in the species tree [56]. Rather than always evaluating all trees in the full SPR neighborhood at each iteration, PhyloGTP first considers only the restricted set of trees obtained by regrafting a single pruned subtree S_v, rooted at a some node $v \in V(S)/\, rt(S)$, onto each possible edge in the current species tree S. It finds the lowest cost tree S' within that restricted neighborhood and, if S' has lower cost than S, then S is replaced by S' and PhyloGTP proceeds to the next local search iteration. If no improvement was found in the restricted neighborhood using S_v, then a new node $u \neq v \in V(S)$ is chosen and the restricted local search step is repeated using the pruned subtree S_u. Thus, PhyloGTP is initially constrained to a small subset of the full SPR search space, but will incrementally expand the set of trees under consideration until an improvement is found, or until the full SPR neighborhood is explored. In the latter case, if no improvement is found then the search is determined to have converged. Note that the order in which subtrees are considered for pruning is randomized at the beginning of each local search iteration. In addition, if there are multiple species trees with minimum reconciliation cost within a restricted neighborhood, then the new species tree S' is selected uniformly at random among them.

Observe that we use a search strategy based on restricted SPR local neighborhoods instead of exploring the full SPR local neighborhood at each local search iteration. This is motivated by the underlying computational complexity of the computation. If n denotes the number of taxa in the analysis and k the number of input gene trees then, assuming most of the k gene trees have $\Theta(n)$ leaves, the time complexity of naively evaluating all candidate species trees in a single SPR local neighborhood becomes $\Theta(n^2) \times \Theta(n^2) \times \Theta(k)$ which is $\Theta(k \cdot n^4)$. This does not scale well with increasing n. Furthermore, many local search iterations have to be performed during a single execution of the heuristic. By using a search strategy based on restricted SPR local neighborhoods, the number of candidate species trees evaluated during most local search iterations reduces to $\Theta(n)$, reducing the time complexity of most local search iterations to a more reasonable $\Theta(k \cdot n^3)$. Importantly, this approach retains the key advantage of using a full SPR-based search since the heuristic search only terminates if a better tree is not found in the full SPR local neighborhood. Previous work on a simpler GTP problem suggests that heuristics based on restricted SPR local neighborhoods perform as well as those based on using full SPR neighborhoods during each local search iteration [63].

DTL event costs assignment. By default, PhyloGTP uses event costs of 2, 3, and 1 for gene duplications, HGTs, and gene losses, respectively (i.e., $P_d = 2$, $P_t = 3$, and $P_l = 1$). These are standard costs used in the DTL reconciliation literature and have been previously observed to work well in practice for microbial datasets [5, 7, 15]. All experimental results reported in this manuscript are based on these default event costs for PhyloGTP.

Parallelization. PhyloGTP implements parallelization to further improve its scalability and enable application to large-scale datasets. The parallelization strategy works by dynamically distributing the computation associated with obtaining the reconciliation costs of candidate species trees in the local search neighborhood across a user-defined number of cores. Thus, when using c cores, the running time of the heuristic is reduced by roughly a factor of c.

4 Results

We use both simulated and real biological datasets to carefully assess the reconstruction accuracy of PhyloGTP. We also compare the accuracy of PhyloGTP against two recently developed state-of-the-art methods: SpeciesRax [46] and ASTRAL-Pro 2 [67]. SpeciesRax first uses a novel distance-based method, miniNJ, which estimates leaf-leaf distances based on the input gene trees, to construct an initial species tree using Neighbor Joining, and then executes a lightweight local search heuristic to optimize the initial species tree based on a probabilistic DTL reconciliation model. ASTRAL-Pro 2 first constructs a constrained search space of candidate species trees based on greedily optimizing a quartet similarity score, and then uses dynamic programming to find the best tree within that constrained search space. Both SpeciesRax and ASTRAL-Pro 2 were run using default parameter settings as provided in their respective manuals.

4.1 Results on simulated data

Dataset description. We used simulated datasets with known ground truth species trees to assess the impact of three key parameters on reconstruction accuracy: Number of input gene trees, rates of gene duplication, HGT, and gene loss (or DTL rates for short), and estimation error in the input gene trees.

Simulated datasets were created using a three-step pipeline: (1) simulation of a ground-truth species tree and corresponding true gene trees with varying DTL rates, (2) simulation of sequence alignments of different lengths for each gene tree, and (3) reconstruction of estimated gene trees from the sequence alignments. In the first step, we used SaGePhy [34] to first simulate ground-truth species trees, each with exactly 50 leaves (taxa) and a height (root to tip distance) of 1, under a probabilistic birth-death framework. We then used these species trees to simulate multiple gene trees under the probabilistic duplication-transfer-loss model implemented in SaGePhy. This resulted in 9 different datasets of simulated true gene trees, each corresponding to a different number of input gene trees (10, 100, or 1000), and a different DTL rate (low, medium, or high; see Table 1). Each dataset comprised of 10 replicates. The chosen DTL rates are based on the relative rates and frequencies of gene duplication and HGT events in real microbial datasets [7]. In each case, the gene loss rate is assigned to be equal to the gene duplication rate plus the additive HGT rate, so as to balance the number of gene gains with the number of gene losses (Table 1).

Basic statistics on these simulated true gene trees, including average sizes and numbers of gene duplication and HGT events, are provided in Table 2.

In the second step, we used AliSim [37] to simulate DNA sequence alignments along each simulated gene tree under the General Time-Reversible (GTR) model (using default AliSim GTR model settings) with three different sequence lengths: 400, 100, and 50 bp. In the third and final step, maximum-likelihood gene trees were inferred using IQ-TREE 2 [45] from the simulated sequence alignments under the Jukes-Cantor (JC) model. We use the simpler JC model when estimating gene trees, instead of the GTR model used to generate the sequences, since this better captures the biases of applying substitution models to real sequences. Thus, from each dataset of true gene trees, we derive 3 additional datasets of estimated gene trees corresponding to the three sequence lengths. The purpose of the second and third steps above is to generate error-prone gene trees that reflect the reconstruction/estimation error present in real gene trees. We found that the estimated gene trees had average normalized Robinson-Foulds distances [53] (defined below) of 0.08, 0.22, and 0.35 for sequence lengths 400, 100, and 50 bp, respectively, to the corresponding true gene trees.

Table 1. Key parameters used in the simulation study. The table lists the main parameters and their values explored in the simulation study. All 36 ($= 3 \times 3 \times 4$) combinations of these three parameters were evaluated at 10 replicates each. DTL rates are specified in the form (d, t, l), where d is the gene duplication rate, t is the HGT rate (split evenly between additive and replacing HGTs), and l is the gene loss rate. The number of species was fixed at 50 for these datasets.

Parameter	**Values**
Number of gene trees	$10, 100, 1000$
DTL rates	low $= (0.3, 0.6, 0.6)$ med $= (0.6, 0.12, 0.12)$ high $= (0.12, 0.24, 0.24)$
Sequence length (nucleotides)	$400, 100, 50$, and true gene trees

Table 1 summarises the specific ranges of parameter values we explored for the number of gene trees, DTL rates, and sequence lengths. We evaluated all combinations of these parameter values, resulting in a total of 36 simulated datasets, with each dataset comprising of 10 replicates created using that specific assignment of parameter values. We also created some additional datasets with 10 and 100 taxa for the runtime analysis.

Evaluating reconstruction accuracy. To evaluate the species tree reconstruction accuracies of the different methods, we compare the species tree estimated by each method with the corresponding ground truth species tree. To perform this comparison we utilize the widely used (unrooted) normalized Robinson-Foulds distance (NRFD) [53] between the reconstructed and ground truth species trees.

Table 2. Basic statistics for simulated gene trees. Average number of leaves, duplications, and HGTs, and losses in the simulated low, medium, and high DTL gene trees. For each DTL rate, the number of losses is roughly equal to the number of duplications plus half the number of HGTs. Results were averaged over all 10 replicates of the 100 gene tree datasets.

DTL rate	Leaves	Duplications	HGTs
Low	53.618	3.408	6.586
Med	55.121	6.15	11.125
High	59.718	10.077	18.37

For any reconstructed species tree, the NRFD reports the fraction of non-trivial splits in that species tree that do not appear in the corresponding ground truth species tree. For ease of interpretation, we report results in terms of *percentage accuracy*, defined to be the percentage of non-trivial splits in the reconstructed species tree that also appear in the ground truth species tree. Thus, percent accuracy is simply $(1 - \text{NRFD}) \times 100$. Thus, for example, a percentage accuracy of 87% is equivalent to an NRFD of 0.13.

Accuracy on true (error-free) gene trees. We first evaluate the accuracy of the species tree reconstruction methods when given true (error-free) gene trees as input (effectively skipping steps 2 and 3 of the simulation pipeline). While error-free gene trees do not capture the complexities of real data, this analysis helps us understand how the different methods perform in a controlled, ideal setting. Figure 1 shows the results for low, medium, and high DTL rates with varying numbers of gene trees for 50-taxon datasets. Unsurprisingly, we find that both DTL rates and number of input gene trees are highly impactful parameters. The performance of all three methods worsens as DTL rates increase, and improves as the numbers of input gene trees increase. Both PhyloGTP and SpeciesRax substantially outperform ASTRAL-Pro 2, especially on the medium and high DTL datasets. In particular, we find that ASTRAL-Pro 2 is highly susceptible to high DTL rates, and that it also shows poor performance when the number of input gene tree is small. Interestingly, the accuracy of Astral-pro 2 improves rapidly as the number of gene trees increases, with the method performing equivalently to PhyloGTP and SpeciesRax on the low and medium DTL datasets when the input consists of 1000 gene trees. Between PhyloGTP and SpeciesRax, we find that PhyloGTP shows higher accuracy when the number of gene trees is small (100 or fewer), particularly when DTL rates are medium or high. For the remaining datasets, both PhyloGTP and SpeciesRax show nearly identical accuracies.

Accuracy on estimated (erroneous) gene trees. We next assess the accuracy of the reconstructed species trees when the input consists of estimated (erroneous) gene trees. Figure 2 shows the results of this analysis for all 27 combinations of number of input gene trees, DTL rates, and sequence lengths (or

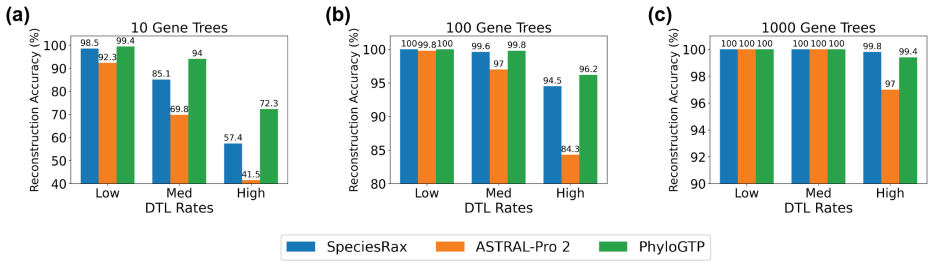

Fig. 1. Accuracy on true gene trees. Tree reconstruction accuracies are shown for PhyloGTP, SpeciesRax, and ASTRAL-Pro 2 when applied to error-free or 'true' gene trees. Results are shown for increasing numbers of input gene trees (10, 100, and 1000) and for low, medium, and high DTL rates. The number of taxa (i.e., number of leaves in the species tree) is fixed at 50. Higher percentages (y-axis) imply greater accuracy.

gene tree estimation error rates). As expected, the accuracy of all three methods is substantially affected by the quality of the estimated gene trees, with higher accuracies achieved using gene trees estimated from longer sequences. We also find that an increased number of input gene trees can partly make up for error in the input gene trees. For example, compared to using true input gene trees (Figure 1), PhyloGTP shows a $5-21\%$ reduction in accuracy with 10 estimated gene trees but only a $1-4\%$ reduction with 1000 estimated gene trees, depending on sequence length. Similar trends are observed with SpeciesRax and ASTRAL-Pro2. Overall, we find that PhyloGTP and SpeciesRax still outperform ASTRAL-Pro-2 across most datasets and that ASTRAL-Pro 2 continues to be more susceptible to high DTL rates than the other methods. As before, the performance of ASTRAL-Pro 2 improves rapidly with increasing number of input gene trees, even sometimes outperforming SpeciesRax and PhyloGTP when DTL rates are low or medium. This suggests that ASTRAL-Pro 2 may be appropriate for microbial phylogenomics on datasets with lots of gene trees and relatively low prevalence of HGT. Comparing PhyloGTP with SpeciesRax, we find that both methods have similar performance overall, with PhyloGTP and SpeciesRax showing average percent accuracies of 88.36% and 86.87%, respectively, when averaged across all 27 datasets. However, PhyloGTP consistently outperforms SpeciesRax on datasets with high DTL rates (showing better accuracy, sometimes substantially better, in all by one high DTL dataset), as well as on datasets with 10 input gene trees. We also find that SpeciesRax tends to outperform PhyloGTP on datasets with high gene tree error and low DTL rates. This suggests that PhyloGTP may be especially useful for analyzing datasets with high levels of HGT or with a small number of gene trees.

Runtimes. We compare the runtimes of the three methods when varying the number of taxa (10, 50, and 100) over low, medium, and high DTL rates. In addition, we also evaluate the impact of the number of input gene trees (100 and 1000) using the 50-taxon dataset. These runtimes are shown in Table 3. All methods have parallel implementations and were allocated 12 cores on a 2.8

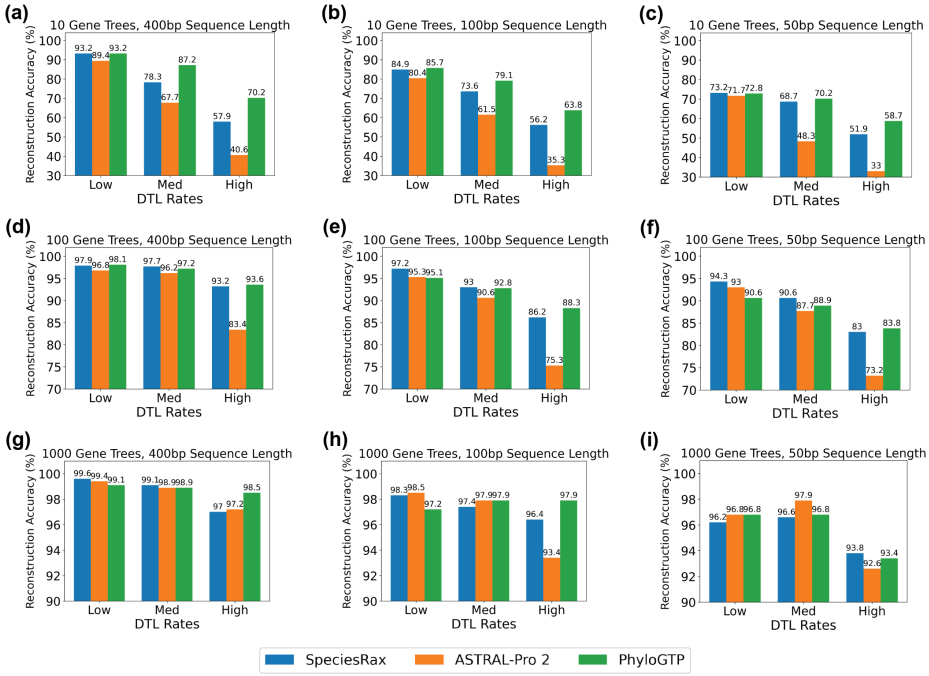

Fig. 2. Accuracy on estimated gene trees. Tree reconstruction accuracies are shown for PhyloGTP, SpeciesRax, and ASTRAL-Pro 2 when applied to estimated gene trees. Results are shown for all 27 combinations of number of input gene trees, sequence lengths (shorter sequence lengths imply greater gene tree estimation error), and DTL rates. The first, second, and third rows correspond to datasets with 10, 100, and 1000 gene families, respectively, and the first, second, and third columns correspond to 400, 100, and 50 base pair sequence lengths, respectively. The number of taxa (i.e., number of leaves in the species tree) is fixed at 50. Higher percentages imply greater accuracy.

GHz × 4 Intel i7 processor with 16 GB of RAM. We find that ASTRAL-Pro 2 is, by far, the fastest method, requiring only about 5 seconds on the 50-taxon 1000 gene tree datasets and less than 10 seconds on the 100-taxon 100 gene tree datasets. SpeciesRax is also extremely fast, requiring only about 60 seconds and 50 seconds, respectively, on those datasets. PhyloGTP is much slower than the other two methods, requiring about 3.5 hours and 10.5 hours on those same datasets. This is expected since this prototype implementation of PhyloGTP has a time complexity that is quartic (n^4) in the number of species. Unlike PhyloGTP, ASTRAL-Pro 2 does not rely on local search heuristics, instead using highly efficient algorithms for computing quartet similarity scores and for finding an optimal species tree within a constrained search space. SpeciesRax does implement a local search heuristic and uses DTL reconciliation, but it's heuristic is light-weight and searches over a smaller search space. SpeciesRax

Table 3. Impact of number of taxa and gene trees on running time. Runtimes in seconds are shown for the three methods for datasets with 10, 50, and 100 taxa and low medium, and high rates of DTL. For the 10- and 100-taxon datasets, the number of input gene trees is 100. For 50-taxon datasets, results are shown for both 100 and 1000 gene trees. The runtimes are based on simulated true input gene trees and are averaged over 10 replicate runs. Each method was allocated 12 cores on a 2.8 GHz × 4 Intel i7 processor with 16 GB of RAM.

Dataset size	DTL rate	SpeciesRax	ASTRAL-Pro 2	PhyloGTP
10 taxa, 100 gene trees	low	1.45	0.08	4.02
	med	1.36	0.08	4.45
	high	1.35	0.09	4.82
50 taxa, 100 gene trees	low	5.69	1.11	1,299.56
	med	6.25	1.14	1,374.92
	high	8.9	1.43	2,015.03
50 taxa, 1000 gene trees	low	50.34	5.38	10,011.79
	med	52.33	5.29	11,433.19
	high	59.95	5.55	13,137.93
100 taxa, 100 gene trees	low	22.05	3.61	19,871.48
	med	28.90	4.84	32,606.87
	high	47.19	7.15	38,259.04

also uses a fast distance-based approach to compute a good initial species tree, which greatly reduces the number of local search steps needed. It may be possible to use some of these techniques to speed up PhyloGTP as well, without sacrificing accuracy.

4.2 Results on biological data

We assembled two previously used biological datasets of different size, composition, and complexity to assess the accuracy and consistency of species trees inferred by PhyloGTP as compared to SpeciesRax and traditional non-DTL cognizant methods such as MLSA and tANI [26] (Table 4). To examine the effect of extreme divergence and genome complexity variation on species tree inference, we used a dataset composed of 176 Archaea, which was drawn from [21]. The Archaea included in the dataset span 2-3 kingdoms (or superphylums), and radically different lifestyles (from extremophiles inhabiting Antarctic lakes to mammal gut constituents). Because the pan-genome of an entire domain would be immeasurably large and computationally infeasible to accurately infer, we have reduced the number of gene families in this dataset to 282 core genes, which are shared by all members. This also allows direct comparison of the PhyloGTP species tree to previously calculated phylogenies in [21] which used the

same loci. It should be noted that the 282 gene families used in the PhyloGTP analysis have been expanded to include all homologs (paralogs, xenologs, etc.) found in each genome, while only orthologs were used in [21].

To examine the impact of low sequence divergence on PhyloGTP species tree inference, we used a dataset of 44 Frankiales genomes, drawn from [26]. These included taxa are all closely related members of the order Frankiales, and as such the entire pan-genome (8,862 gene families with at least 4 sequences) was used for inference in PhyloGTP and SpeciesRax. The order Frankiales are composed of nitrogen-fixing symbionts of pioneer flora [61], and although they demonstrate variation in GC content and genome size these factors were previously shown to not bias phylogenetic inference [26].

Table 4. Summary of the two biological datasets.

Dataset	**Number of gene families**	**Potential biases**	**Previous methods used to infer species tree**
176 Archaea (domain)	282	Extreme divergence, long branch attraction, compositional bias	tANI, MLSA, single gene
44 Frankiales (order)	8,862	Low divergence, contamination, genome size difference	tANI, MLSA

Archaeal dataset. A myriad of controversies surround the phylogeny of Archaea. These controversies include the monophyly of the DPANN superphylum [2, 10, 21, 47, 52], the placement of extreme halophiles [1, 21, 47, 57], and the root of the Archaea [51]. These differences in phylogenetic inference are driven by many factors including, but not limited to compositional bias, long branch attraction, extremely small genomes, numerous HGT events, and biased sampling of metagenome-assembled genomes. Thus, it is interesting to evaluate the performance of PhyloGTP in the face of these factors.

Using 282 unrooted input gene trees, both PhyloGTP and SpeciesRax inferred Archaeal species trees with small inaccuracies with respect to commonly accepted placements of groups in previous analyses. These inaccuracies should be interpreted in the context that for several Archaeal clades (mostly halophiles) there is no consistent, consensus position that is universally accepted amongst Archaeaologists. For example, the monophyly of the DPANN superphylum is considered by some to be an artifact (driven by long branch attraction or biased genome sampling) [2, 21, 69]. Both species trees (Figure 3) recover a monophyletic DPANN superphylum and successfully resolve the TACK clade. PhyloGTP successfully recovers a monophyletic Euryarchaea kingdom (Fig 3a), whereas SpeciesRax has misplaced the Methanomada and Thermococcales (both euryarchaeotes) onto the branches leading to the TACK group. One major point

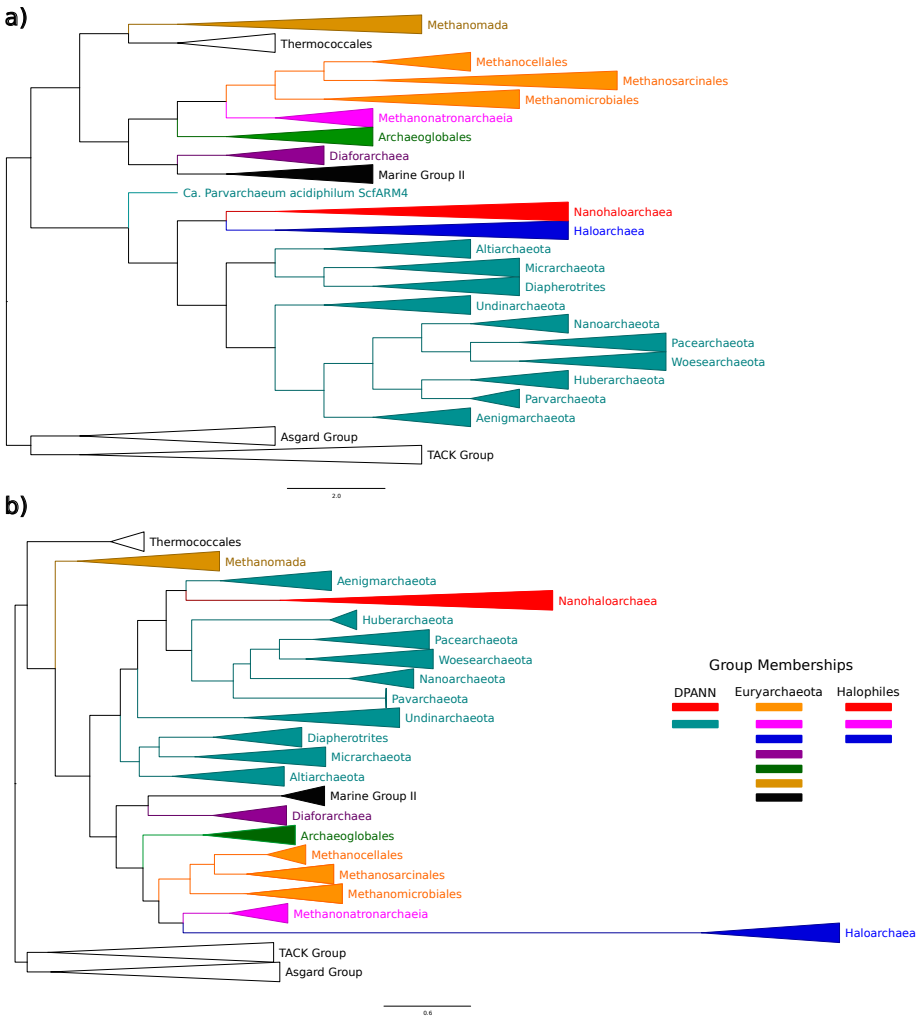

Fig. 3. Archaeal species tree reconstructions. Individual taxa on both trees have been collapsed into clades and are colored corresponding to higher level classifications (clades with the same color are part of the same class or phylum). The legend shows previously reported Kingdom memberships of these collapsed clades, and also the halophiles which may group together as a result of compositional bias. Part a) Unrooted Archaeal tree inferred by PhyloGTP; to be read as a cladogram since PhyloGTP does not infer branch lenghs. Part b) Unrooted Archaeal tree inferred by SpeciesRax.

of disagreement between the two methods is the placement of the Haloarchaea. SpeciesRax correctly recovers the Haloarchaea within the euryarchaeota (Figure 3b), while this group has moved inside the DPANN superphylum, to be the sister group of the Nanohaloarchaea, in the PhyloGTP phylogeny (Figure 3a). In addition, the position of the Methanonatronarchaeia (another halophile) in both trees is recovered as later branching euryarchaeota (Figure 3), in contrast to previous studies which report them as basal to the Methanotecta + Archaeoglobales superclass [1,21]. Incorrect placements of the Nanohaloarchaea, Haloarchaea and Methanonatronarchaeia are often attributed to compositional bias [21]. These halophiles prefer acidic amino acid residues (such as aspartate and glutamate), on account of their survival strategies in hypersaline environments, and these acidified proteomes attract the placement of these groups together in phylogenetic reconstructions.

Overall, these results demonstrate that PhyloGTP can produce a mostly accurate Archaeal tree, even in the face of the many biases present in the dataset (Table 4). At the same time, these results also show that PhyloGTP and SpeciesRax are both susceptible to the presence of problematic groups (such as the extreme halophiles) and other biases in complex datasets, potentially limiting their accuracy in some cases.

Frankiales dataset. In the case of the Frankiales, reconstructions with PhyloGTP and SpeciesRax yield identical relationships between the major clades (Figure 4). This suggest that both programs have comparable efficacy when the dataset analyzed is less complex and less divergent. Since this analysis used the entire pan-genome of the Frankiales, a possible concern is that small gene families (such as those that are only found in 4-8 genomes) may negatively impact these reconciliation based methods. To assess the impact of small gene families on species tree reconstruction, a subset of 1,702 genes families present in at least 20 genomes and in the smallest Frankia genome (*Frankia* sp. DG2) was used for inference using PhyloGTP and SpeciesRax. The trees produced from this subset recovered the same topologies for major clades as those in the full complement, indicating that the smaller gene families are not a problem for either method.

In comparison to previous trees inferred on the same genomes using previous non-DTL methods, such as those shown in [26], there are a few rearrangements of early branching clades in the backbone of the Frankiales. In phylogenies inferred using tANI and MLSA sequence methods, Group 1 (Figure 4) is basal to the rest of the Frankiales. In the PhyloGTP and SpeciesRax trees, Group 3 is basal to the other Frankiales, with Group 1 as a later branching basal group. In addition to the movement of these clades, *Frankia* sp. NRRLB16219 and *Frankia* sp. CgIS1 have swapped positions, where *Frankia* sp. CgIS1 has moved from Group 2 to Group 5. These rearrangements may be attributed to the additional genomic data used to reconstruct the PhyloGTP and SpeciesRax trees. Only 24 loci were used in [26], and the inclusion of thousands of additional gene families have painted a slightly different picture of evolution throughout the Frankiales. This suggests that truly genome-scale methods like PhyloGTP could lead to more accurate phylogenomic inference on real datasets compared to other methods.

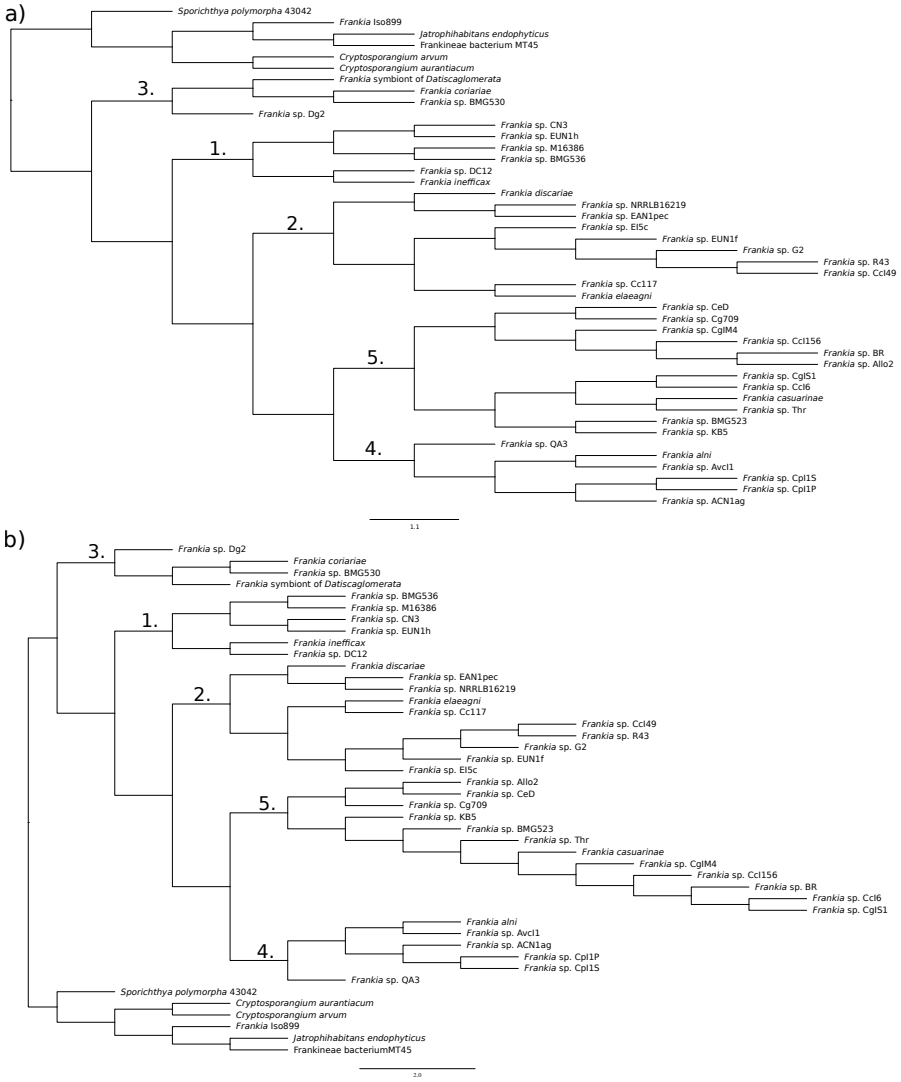

Fig. 4. Cladograms of the Frankiales. Clades on both trees are categorized and sorted based on the group designations described in [26]. Note that both trees show identical relationships among the labeled clades, but not necessarily within those clades. Part a) Frankiales cladogram inferred by PhyloGTP. Part b) Cladogram inferred by SpeciesRax.

Comparison of total DTL reconciliation costs. We also compared the total DTL reconciliation costs of the PhyloGTP and SpeciesRax species trees for these biological datasets. We find that, for both datasets, PhyloGTP species trees show considerably lower reconciliation costs. Specifically, for the Archaeal dataset, PhyloGTP and SpeciesRax species trees have total DTL reconciliation costs of 58,291 and 59,140, respectively. For the Frankiales dataset, these reconciliation costs are 148,898 and 156,376, respectively. These numbers show that, unlike PhyloGTP, speciesRax does not necessarily minimize the total DTL reconciliation cost. This is likely due to the different objective function used by SpeciesRax.

Details of dataset assembly. Annotated genomes of 176 Archaea used in [21] were collected. The 282 core gene loci described in [21] were used as amino acid query sequences to search every collected genome, using blastp [11] with default parameters (-evalue was changed to 1e-10). All significant sequence for every loci across all genomes were collected (provided they met a length threshold of 50% in reference to the average gene family sequence size to filter partial sequences). Each gene family was then aligned using mafft-linsi [32] with default parameters and used as the basis for gene tree inference in IQ-Tree 2 [45], where the best substitution model for each gene family was determined using Bayesian Inference Criterion [31].

Annotated proteomes of the 44 Frankiales used in [26] were collected. Protein sequences were clustered into gene families and using the OrthoFinder2.4 pipeline [20] with default parameters (the search algorithm was changed to blast). Briefly, all-vs-all blastp (evalue of 1e-3) was used to find the best hits between input species. The set of query-matches were then clustered into gene families using the MCL algorithm, and the subsequent gene families were aligned using mafft-linsi with default parameters. Resulting alignments were used to create gene trees using FastTree [50] using the JTT model and default parameters.

5 Discussion and Conclusion

In this work, we introduced PhyloGTP, a new method for microbial species tree inference using GTP. PhyloGTP searches for the most parsimonious species tree under the DTL reconciliation model, making it the first GTP-based method suitable for microbial phylogenomics. Our simulation study shows that PhyloGTP can substantially outperform SpeciesRax when the number of input gene trees is small or when DTL rates are high. However, PhyloGTP is not consistently better than SpeciesRax and SpeciesRax tends to outperform PhyloGTP on datasets with high gene tree error but low DTL rates. We also find that both PhyloGTP and SpeciesRax almost always outperform ASTRAL-Pro 2, a highly scalable but HGT-naive method. Our results on the two biological datasets suggest that PhyloGTP works very well on real datasets overall, but also that both PhyloGTP and SpeciesRax can sometimes be misled by problematic taxa and compositional and other biases present in complex datasets.

While our experiments with PhyloGTP have yielded promising results, the prototype implementation of PhyloGTP is far slower, and hence far less scalable, than SpeciesRax or ASTRAL-Pro 2. However, we expect future work on improved algorithms and heuristics for GTP under DTL reconciliation to result in software implementations that are much faster and more accurate than the current PhyloGTP prototype. There are several possible directions for future research. First, PhyloGTP will likely benefit from differential weighting of the input gene trees. For example, the current implementation of PhyloGTP does not take into account the level of inference uncertainty in the input gene trees. Such measures of uncertainty are often readily available, such as bootstrap support values, and they could be used to distinguish between more reliable and less reliable gene trees. A simple multiplicative weight between 0 and 1 could then be assigned to each gene tree, reflecting confidence in that gene tree. It may also make sense to normalize reconciliation costs based on gene tree size and to down-weight gene trees exhibiting very high reconciliation costs. Exploring and carefully evaluating such weighting schemes is a promising direction for future research.

Second, PhyloGTP could be substantially sped up using alternative tree search strategies or improved algorithms for reconciliation cost computations. For example, it may be possible to constrain the search space of candidate species trees without sacrificing accuracy, or design algorithms to quickly approximate the total DTL reconciliation cost of candidate species trees to guide the local search heuristic. Third, given our findings on the Archaeal dataset, it would be useful to characterize the performance of PhyloGTP and SpeciesRax more carefully using more nuanced simulated datasets exhibiting some of the complications and biases observed in complex biological datasets. Finally, it may also be possible to devise asymptotically faster algorithms to compute the lowest DTL reconciliation cost tree within an entire SPR local neighborhood, as has been previously accomplished for simpler reconciliation models [4].

Acknowledgments. This work was supported in part by a University of Connecticut Research Excellence Program award to JPG and MSB.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. M. Aouad, G. Borrel, C. Brochier-Armanet, and S. Gribaldo. Evolutionary placement of methanonatronarchaeia. *Nature microbiology*, 4(4):558–559, 2019.
2. M. Aouad, N. Taib, A. Oudart, M. Lecocq, M. Gouy, and C. Brochier-Armanet. Extreme halophilic archaea derive from two distinct methanogen class ii lineages. *Molecular phylogenetics and evolution*, 127:46–54, 2018.
3. M. S. Bansal, E. J. Alm, and M. Kellis. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12):283–291, 2012.

4. M. S. Bansal and O. Eulenstein. Algorithms for genome-scale phylogenetics using gene tree parsimony. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(4):939–956, 2013.
5. M. S. Bansal, M. Kellis, M. Kordi, and S. Kundu. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216, 04 2018.
6. M. S. Bansal and R. Shamir. A note on the fixed parameter tractability of the gene-duplication problem. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(3):848–850, 2011.
7. M. S. Bansal, Y.-C. Wu, E. J. Alm, and M. Kellis. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, 31(8):1211–1218, 2015.
8. R. G. Beiko, T. J. Harlow, and M. A. Ragan. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14332–14337, 2005.
9. M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8:409–423, 2004.
10. C. Brochier-Armanet, P. Forterre, and S. Gribaldo. Phylogeny and evolution of the archaea: one hundred genomes later. *Current opinion in microbiology*, 14(3):274–281, 2011.
11. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10:1–9, 2009.
12. M. A. Cerón-Romero, M. M. Fonseca, L. de Oliveira Martins, D. Posada, and L. A. Katz. Phylogenomic Analyses of 2,786 Genes in 158 Lineages Support a Root of the Eukaryotic Tree of Life between Opisthokonts and All Other Lineages. *Genome Biology and Evolution*, 14(8):evac119, 07 2022.
13. R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernandez-Baca, and O. Eulenstein. iGTP: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics*, 11(1):574, 2010.
14. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287, 2006.
15. L. A. David and E. J. Alm. Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469:93–96, 2011.
16. R. Davidson, P. Vachaspati, S. Mirarab, and T. Warnow. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics*, 16 (Suppl 10):S1, 2015.
17. W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.
18. W. F. Doolittle, Y. Boucher, C. L. Nesbo, C. J. Douady, J. O. Andersson, and A. J. Roger. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1429):39–58, 2003.
19. J.-P. Doyon, C. Scornavacca, K. Y. Gorbunov, G. J. Szöllosi, V. Ranwez, and V. Berry. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In E. Tannier, editor, *RECOMB-CG*, volume 6398 of *Lecture Notes in Computer Science*, pages 93–108. Springer, 2010.
20. D. M. Emms and S. Kelly. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, 20:1–14, 2019.

21. Y. Feng, U. Neri, S. Gosselin, A. S. Louyakis, R. T. Papke, U. Gophna, and J. P. Gogarten. The evolutionary origins of extreme halophilic archaeal lineages. *Genome biology and evolution*, 13(8):evab166, 2021.
22. S. R. Gadagkar, M. S. Rosenberg, and S. Kumar. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304B(1):64–74, 2005.
23. S. P. Glaeser and P. Kämpfer. Multilocus sequence analysis (mlsa) in prokaryotic taxonomy. *Systematic and Applied Microbiology*, 38(4):237–245, 2015. Taxonomy in the age of genomics.
24. J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19(12):2226–2238, 2002.
25. M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage. a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–163, 1979.
26. S. Gosselin, M. S. Fullmer, Y. Feng, and J. P. Gogarten. Improving phylogenies based on average nucleotide identity, incorporating saturation correction and nonparametric bootstrap support. *Systematic Biology*, 71(2):396–409, 2022.
27. R. E. Green, E. L. Braun, J. Armstrong, D. Earl, et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science*, 346(6215), 2014.
28. S. R. Henz, D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics*, 21(10):2329–2335, 05 2004.
29. E. Hilario and J. P. Gogarten. Horizontal transfer of {ATPase} genes – the tree of life becomes a net of life. *Biosystems*, 31(2â€“3):111 – 119, 1993.
30. R. P. Hirt, J. M. Logsdon, B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. Microsporidia are related to fungi: Evidence from the largest subunit of rna polymerase ii and other proteins. *Proceedings of the National Academy of Sciences*, 96(2):580–585, 1999.
31. S. Kalyaanamoorthy, B. Q. Minh, T. K. Wong, A. Von Haeseler, and L. S. Jermiin. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589, 2017.
32. K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.
33. K. T. Konstantinidis and J. M. Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, 102(7):2567–2572, 2005.
34. S. Kundu and M. S. Bansal. SaGePhy: an improved phylogenetic simulation framework for gene and subgene evolution. *Bioinformatics*, 02 2019.
35. J. M. Lang, A. E. Darling, and J. A. Eisen. Phylogeny of bacterial and archaeal genomes using conserved genes: Supertrees and supermatrices. *PLoS ONE*, 8(4):e62510, 04 2013.
36. P. O. Lewis, M.-H. Chen, L. Kuo, L. A. Lewis, K. Fucikova, S. Neupane, Y.-B. Wang, and D. Shi. Estimating bayesian phylogenetic information content. *Systematic Biology*, 65(6):1009–1023, 2016.
37. N. Ly-Trong, S. Naser-Khdour, R. Lanfear, and B. Q. Minh. AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era. *Molecular Biology and Evolution*, 39(5):msac092, 05 2022.
38. B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM J. Comput.*, 30(3):729–752, 2000.

39. W. P. Maddison and D. Maddison. Mesquite: a modular system for evolutionary analysis. version 2.6. http://mesquiteproject.org, 2009.
40. M. Marcet-Houben and T. Gabaldon. Horizontal acquisition of toxic alkaloid synthesis in a clade of plant associated fungi. *Fungal Genetics and Biology*, 86:71 – 80, 2016.
41. V. M. Markowitz, I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N. N. Ivanova, and N. C. Kyrpides. Img 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*, 42(D1):D560–D567, 2014.
42. F. Marlétaz, E. de la Calle-Mustienes, R. Acemel, et al. The little skate genome and the evolutionary emergence of wing-like fins. *Nature*, 616:495–503, 2023.
43. C. G. P. McCarthy and D. A. Fitzpatrick. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *mSphere*, 2(2), 2017.
44. J. O. McInerney, J. A. Cotton, and D. Pisani. The prokaryotic tree of life: past, presentâ€¦and future? *Trends in Ecology & Evolution*, 23(5):276 – 281, 2008.
45. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 02 2020.
46. B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, and A. Stamatakis. SpeciesRax: A Tool for Maximum Likelihood Species Tree Inference from Gene Family Trees under Duplication, Transfer, and Loss. *Molecular Biology and Evolution*, 39(2):msab365, 01 2022.
47. P. Narasingarao, S. Podell, J. A. Ugalde, C. Brochier-Armanet, J. B. Emerson, J. J. Brocks, K. B. Heidelberg, J. F. Banfield, and E. E. Allen. De novo metagenomic assembly reveals abundant novel major lineage of archaea in hypersaline microbial communities. *The ISME journal*, 6(1):81–93, 2012.
48. G. J. Olsen, C. R. Woese, and R. Overbeek. The winds of (evolutionary) change: breathing new life into microbiology. *Journal of Bacteriology*, 176(1):1–6, 1994.
49. R. D. M. Page. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics (Oxford, England)*, 14(9):819–820, 1998.
50. M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.
51. K. Raymann, C. Brochier-Armanet, and S. Gribaldo. The two-domain tree of life is linked to a new root for the archaea. *Proceedings of the National Academy of Sciences*, 112(21):6670–6675, 2015.
52. K. Raymann, P. Forterre, C. Brochier-Armanet, and S. Gribaldo. Global phylogenomic analysis disentangles the complex evolutionary history of dna replication in archaea. *Genome biology and evolution*, 6(1):192–212, 2014.
53. D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981.
54. G. Sevillya, D. Doerr, Y. Lerner, J. Stoye, M. Steel, and S. Snir. Horizontal Gene Transfer Phylogenetics: A Random Walk Approach. *Molecular Biology and Evolution*, 37(5):1470–1479, 12 2019.
55. A. Shifman, N. Ninyo, U. Gophna, and S. Snir. Phylo si: a new genome-wide approach for prokaryotic phylogeny. *Nucleic Acids Research*, 42(4):2391–2404, 2014.
56. Y. S. Song. On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, 7(3):365–379, 2003.

57. D. Y. Sorokin, K. S. Makarova, B. Abbas, M. Ferrer, P. N. Golyshin, E. A. Galinski, S. Ciorda, M. C. Mena, A. Y. Merkel, Y. I. Wolf, et al. Reply to 'evolutionary placement of methanonatronarchaeia'. *Nature Microbiology*, 4(4):560–561, 2019.
58. M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):409–415, 2012.
59. A. Tofigh. *Using Trees to Capture Reticulate Evolution : Lateral Gene Transfers and Cancer Progression*. PhD thesis, KTH Royal Institute of Technology, 2009.
60. A. Tofigh, M. T. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 8(2):517–535, 2011.
61. M. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kampfer, F. Rainey, and W. Whitman. *Bergey's manual of systematics of archaea and bacteria*. Wiley Online Library, 2021.
62. A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein. Duptree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13), 2008.
63. A. Wehe and J. Burleigh. Scaling the gene duplication problem towards the tree of life. In *2nd International Conference on Bioinformatics and Computational Biology 2010, BICoB 2010*, pages 133–138, 01 2010.
64. C. Whidden, N. Zeh, and R. G. Beiko. Supertrees based on the subtree prune-and-regraft distance. *Systematic Biology*, 2014.
65. C. R. Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–271, 1987.
66. W. H. Yap, Z. Zhang, and Y. Wang. Distinct types of rrna operons exist in the genome of the actinomycete thermomonospora chromogena and evidence for horizontal transfer of an entire rrna operon. *Journal of Bacteriology*, 181(17):5201–5209, 1999.
67. C. Zhang and S. Mirarab. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*, 38(21):4949–4950, 09 2022.
68. O. Zhaxybayeva, W. F. Doolittle, R. T. Papke, and J. P. Gogarten. Intertwined evolutionary histories of marine synechococcus and prochlorococcus marinus. *Genome Biology and Evolution*, 1:325–339, 2009.
69. O. Zhaxybayeva, R. Stepanauskas, N. R. Mohan, and R. T. Papke. Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*, 17:265–275, 2013.