Ancient Rapid Radiation Explains Most Conflicts Among Gene Trees and Well-supported Phylogenomic Trees of Nostocalean Cyanobacteria

Carlos J. Pardo-De la Hoz[1*], Nicolas Magain[2], Bryan Piatkowski[3], Luc Cornet[2,4], Manuela Dal Forno[5], Ignazio Carbone[6], Jolanta Miadlikowska[1], François Lutzoni[1].

[1]Department of Biology, Duke University, Durham, North Carolina, 27708, United States of America.

[2]Evolution and Conservation Biology, InBioS Research Center, Université de Liège, Liège, 4000, Belgium.

[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37830, United States of America.

[4]BCCM/IHEM, Mycology and Aerobiology, Sciensano, Brussels, Belgium.

[5]Botanical Research Institute of Texas, Fort Worth, Texas, 76107, United States of America.

[6]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, North Carolina, 27606, United States of America.

[*]Correspondence to be sent to: Department of Biology, Duke University, Durham, North Carolina, 27708, United States of America. E-mail: cjpardodelahoz@gmail.com.

Abstract

Prokaryotic genomes are often considered to be mosaics of genes that do not necessarily share the same evolutionary history due to widespread Horizontal Gene Transfers (HGTs). Consequently, representing evolutionary relationships of prokaryotes as bifurcating trees has long been controversial. However, studies reporting conflicts among gene trees derived from phylogenomic datasets have shown that these conflicts can be the result of artifacts or evolutionary processes other than HGT, such as incomplete lineage sorting, low phylogenetic signal, and systematic errors due to substitution model misspecification. Here, we present the results of an extensive exploration of phylogenetic conflicts in the cyanobacterial order Nostocales, for which previous studies have inferred strongly supported conflicting relationships when using different concatenated phylogenomic datasets. We found that most of these conflicts are concentrated in deep clusters of short internodes of the Nostocales phylogeny, where the great majority of individual genes have low resolving power. We then inferred phylogenetic networks to detect HGT events while also accounting for incomplete lineage sorting. Our results indicate that most conflicts among gene trees are likely due to incomplete lineage sorting linked to an ancient rapid radiation, rather than to HGTs. Moreover, the short internodes of this radiation fit the expectations of the anomaly zone, i.e., a region of the tree parameter space where a species tree is discordant with its most likely gene tree. We demonstrated that concatenation of different sets of loci can recover up to 17 distinct and well-supported relationships within the putative anomaly zone of Nostocales, corresponding to the observed conflicts among well-supported trees based on concatenated datasets from previous studies. Our findings highlight the important role of rapid radiations as a potential cause of strongly conflicting phylogenetic relationships when using phylogenomic datasets of bacteria. We propose that polytomies may be the most appropriate phylogenetic representation of these rapid radiations that are part of anomaly zones, especially when all possible genomic markers have been considered to infer these phylogenies.

Keywords: Anomaly zone, bacteria, horizontal gene transfer, incomplete lineage sorting, Nostocales, phylogenomic conflict, rapid radiation, Rhizonema.

INTRODUCTION

With the discovery of Archaea (Woese and Fox 1977), molecular phylogenetics based on 16S ribosomal RNA started to transform the field of microbiology. However, about two decades later, the first complete sequences of prokaryotic genomes revealed that horizontal gene transfers (HGTs) were a major force shaping genomes across the tree of life (Doolittle and Logsdon Jr 1998; Lawrence and Ochman 1998; Doolittle 1999a, 1999b; Nelson et al. 1999). This challenged the adequacy of bifurcating trees to describe prokaryotic evolutionary histories and prompted an intense debate. Some argued that the estimated rate of HGT among prokaryotic genomes implied the loss of the signal of vertical descent (Lawrence and Ochman 1998; Doolittle 1999b; Nesbø et al. 2001), while others claimed that a universal prokaryotic tree exists even in the presence of HGT (Daubin 2002; Kurland et al. 2003). Since then, many studies have shown that rates of HGT are not constant through time, across lineages, or among genes within a genome (Zhaxybayeva 2006; Sorek et al. 2007; Shi and Falkowski 2008; Coleman et al. 2021; Groussin et al. 2021). More importantly, there is growing evidence that most bacterial gene families evolved vertically most of the time, thus leaving a strong tree-like signal in bacterial genomes (Hernández-López et al. 2013; Murray et al. 2016; Avni and Snir 2020; Coleman et al. 2021).

Nevertheless, many phylogenetic studies of prokaryotes have used a single gene that is less likely to be subjected to HGT (e.g., 16S) to avoid potential HGT-related artifacts when inferring species trees (Janda and Abbott 2007; Tringe and Hugenholtz 2008; Yarza et al. 2008). These single gene phylogenies tend to have low statistical support as the number of taxa increases (Bremer et al. 1999; Rokas and Carroll 2005; Janda and Abbott 2007). For this reason, there is a need to use multiple genes to infer phylogenetic relationships of prokaryotes, especially now that genomes are readily available. However, the use of genome-scale datasets may increase the chance of HGT events lowering the accuracy of phylogenetic inferences (McInerney et al. 2008).

One of the most common ways to detect HGT events is to identify strongly supported phylogenetic conflicts among gene trees (Doolittle et al. 2003; McInerney et al. 2008). However,

phylogenetic conflicts can also result from artifacts and evolutionary processes other than HGT, including low phylogenetic signal (Huang et al. 2010; Philippe et al. 2011), systematic errors due to model misspecification (Wang et al. 2019; Redmond and McLysaght 2021), hidden paralogs (Zhang et al. 2020a), and incomplete lineage sorting (ILS; Rosenberg 2013; Meleshko et al. 2021). In recent years, many phylogenomic studies of eukaryotic organisms have revealed the pervasiveness of gene tree conflicts that result from the issues listed above (e.g., Smith et al. 2015; Shen et al. 2017; Richards et al. 2018; Cloutier et al. 2019; Li et al. 2020). This is in contrast with most phylogenomic studies of bacteria, which do not include any tests for conflicts and infer phylogenetic relationships based on concatenated datasets of single-copy core genes (Parks et al. 2017; Greenlon et al. 2019; Pérez-Carrascal et al. 2019; Taib et al. 2020; Kim et al. 2021). In addition, the few studies that have characterized phylogenetic incongruences among bacterial genes focus mostly on HGT and gene duplication and loss (Murray et al. 2016; Coleman et al. 2021; Cornet et al. 2021). Therefore, there is a need for a more comprehensive exploration of the methodological issues and biological processes that can contribute to phylogenetic conflicts in bacterial phylogenies.

An interesting case of rampant phylogenetic conflicts occurs in the cyanobacterial order Nostocales, an important group of photosynthetic and nitrogen-fixing bacteria (Komárek et al. 2014). Previous studies of this order have recovered at least five distinct and highly supported sets of relationships among higher-level lineages (e.g., Fig. 1; Sánchez-Baracaldo 2015; Gagunashvili and Andrésson 2018; Warshan et al. 2018; Gutiérrez-García et al. 2019; Nelson et al. 2019). Although these conflicting relationships have not spurred controversy, some studies have relied on different conflicting phylogenies to address questions about trait evolution in Nostocales (e.g., nitrogenases and symbiotic lifestyle; Gagunashvili and Andrésson 2018; Warshan et al. 2018; Nelson et al. 2019), and could lead to unstable and conflicting classifications at the family rank within this order (Komárek et al. 2014). This highlights the importance of understanding the challenges to accurate phylogenetic inference in this group when using multiple genes. Fossils and molecular dating evidence suggest that the crown of Nostocales is at least one billion years old (Tomitani et al. 2006; Sánchez-Baracaldo et

al. 2014; Sánchez-Baracaldo 2015; Demoulin et al. 2019). This deep phylogenetic history of nostocalean cyanobacteria adds another layer of complexity to the resolution of their relationships.

In this study, our aim was to determine the main sources of topological conflicts among published phylogenies of nostocalean cyanobacteria (Fig. 1), including potential methodological artifacts and evolutionary processes. First, we sought to identify the most problematic relationships by estimating the level of phylogenetic conflicts across topological bipartitions. Next, we determined if different methodological choices such as the number and set of loci, alignment trimming strategies, types of substitution models, and inference methods, contributed to the level of conflicts among gene trees. We then inferred a coalescent phylogenetic network with representatives from the major lineages of Nostocales to estimate the contributions of various evolutionary processes to phylogenetic conflicts. This approach models gene tree conflicts as a result of reticulated evolution and incomplete lineage sorting, while also accounting for phylogenetic uncertainty and substitution site-heterogeneity (Lartillot et al. 2009; Larget et al. 2010; Solís-Lemus and Ané 2016). We used the resulting coalescent phylogenetic network to test whether the internodes with highest conflict fell within the anomaly zone, i.e., a region of the tree parameter space where a species tree is discordant with its most likely gene tree (Degnan and Rosenberg 2006; Rosenberg 2013). And finally, we used phylogenomic jackknifing to show that concatenation of gene alignments with low phylogenetic signal, expected from the anomaly zone, is the likely cause of most conflicts among concatenated phylogenies of Nostocales.

MATERIALS AND METHODS

Genome Data Acquisition

Taxon sampling.—We retrieved all 199 genomes of Nostocales available in the NCBI RefSeq database as of April 20th, 2020. We also included 17 metagenome-assembled genomes (MAGs) of lichenized Nostocales: 12 *Nostoc* sp. from Cornet et al. (2021), as well as two *Nostoc* sp. and three *Rhizonema* sp. sequenced as part of this study (see below; Table S1). Finally, we included the genomes of four representative taxa of the sister order Chroococcidiopsidales as outgroup:

Chroococcidiopsis thermalis PCC 7203, *Gloeocapsa* sp. PCC 7428, *Synechocystis* sp. PCC 7509, and a MAG of the cyanobiont of the lichenized fungus *Peltula cylindrica* (McDonald et al. 2013; Shih et al. 2013). Thus, the total taxon sampling consisted of 220 genomes (Table S1).

Sequencing and assembly of new Nostocales genomes.— For the *Rhizonema* cyanobionts of *Dictyonema* sp. 3644, *Dictyonema* sp. 3668, and *Dictyonema* sp. 2651, healthy and clean-looking lichen fragments were selected under the stereoscope with sterile tweezers. DNA was isolated with the FastDNA™ SPIN Kit (MP Biomedicals) or the DNeasy PowerSoil Pro Kit (Qiagen). Samples were then quantified with Qubit ds HS Assay Kit (Invitrogen) using 5 µl of each DNA sample with 195 µl of the Qubit working solution. Library preparation was performed using the Nextera XT Library Prep Kit and the libraries were sequenced using an Illumina MiSeq Reagent Kit v2 500 cycle (2 × 250 bp). For the *Nostoc* cyanobionts of *Lepidocollema* sp. T6 and *Kroswia* sp. DP7, extraction of DNA followed a modified protocol from Zolan and Pukkila (1986). DNA libraries (500 bp insert) were sequenced with an Illumina NovaSeq 6000 S-Prime cell (2 × 150 bp). All newly sequenced metagenomes, as well as the metagenome of the cyanolichen *Peltula cylindrica* from McDonald et al. (2013), were assembled using metaSPAdes (Nurk et al. 2017) and binned using MetaBAT 2 with default parameters (Kang et al. 2019).

Quality assessment of genome assemblies.—To assess the completeness of each of the 220 genome assemblies, we ran BUSCO v4.1.3. (Simão et al. 2015) using the "cyanobacteria_odb10" as the reference database (Kriventseva et al. 2019). This database consists of 773 single-copy orthologs conserved across Cyanobacteria, and BUSCO quantifies the percentage of those genes that are complete and single copy in each assembly. We then removed nine genomes that had less than 89% of these genes from all phylogenetic analyses (Table S1). For the remaining 211 taxa, we aligned the amino acid sequences of the 773 BUSCO loci using MAFFT v7.475 (Katoh and Standley 2013) with default parameters. We then concatenated these alignments and inferred a maximum likelihood phylogeny with 1000 bootstrap replicates using RAxML v8.2.12 (Stamatakis 2014) with the GTRGAMMA model and default parameters. To assess conflicts among published phylogenies, we used the resulting phylogeny (Fig. S1) to allocate taxa used in previous studies to nine major groups

(Fig. 1). To reduce dataset complexity for subsequent analyses, we selected a subset of 55 taxa (subset 0; Fig. S1 and Table S1) that includes representatives from all major lineages that we identified on this initial phylogeny.

We also used Genome UNClutterer (GUNC; Orakov et al. 2021), a method that can detect both redundant and non-redundant contamination in genome assemblies (Orakov et al. 2021). We ran GUNC on all 220 genome assemblies that we sampled. Because all loci we used in this study are found in the bacterial chromosome, we excluded the plasmid sequences from the complete assemblies before running GUNC to prevent false positives. We confirmed that all 55 taxa part of subset 0 (Table S1) passed the GUNC filter of contamination (Supplementary File S1).

Sequence Datasets

Previous studies that included phylogenomic trees of Nostocales have used molecular datasets of different sizes (e.g., Fig. 1) including genes selected using different criteria. Therefore, we performed our phylogenetic analyses using multiple datasets to determine their potential role as sources of phylogenetic conflicts and whether some of these datasets should be preferred to others. Below is a description of the datasets and the criteria used to select genes for each of them.

Loci set L31.—31 housekeeping genes that were initially identified for the AMPHORA pipeline and proposed as suitable markers for phylogenomic studies in Bacteria (Wu and Eisen 2008). Most of these genes encode ribosomal proteins, and they are used routinely for phylogenetic and taxonomic studies of Cyanobacteria (e.g., Shih et al. 2013; Komárek et al. 2014; Walter et al. 2017; Gagunashvili and Andrésson 2018; Bell-Doyon et al. 2020).

Loci set L70.—70 genes that are part of 22 Collinear Orthologous Regions (CORs) found across Nostocales (Cornet et al. 2021). This set includes only genes that are part of the largest operon in each COR. Cornet et al. (2021) selected these markers using synteny, collinearity, and operon criteria under the assumption that genes that are present in the same order and orientation across a clade are less likely to have undergone HGT.

Loci set L746.—A subset of 746 genes from the 773 single-copy ortholog genes in the "cyanobacteria_odb10" database used by BUSCO, which are conserved across all Cyanobacteria. We removed the genes that were present as complete and single-copy in less than 90% of the 220 genomes sampled for this study (Table S1) and kept the remaining 746 loci. A similar set of loci from an earlier version of the database ('cyanobacteria_odb9') was used in a recent phylogenetic study on Nostocales (Nelson et al. 2019).

Loci set L1648.—A subset of 1648 genes from the 1899 single-copy ortholog genes in the 'nostocales_odb10' database used by BUSCO, which are conserved across the order Nostocales. We removed the genes that were present as complete and single-copy in less than 90% of the 220 genomes sampled for this study (Table S1) and kept the remaining 1648 loci. Note that the loci from all other sets used for this study are included in L1648.

Loci set L1082.—A subset of 1082 nucleotide gene alignments from the L1648 dataset. We generated this subset by removing loci with low phylogenetic signal, which has been shown to reduce incongruences among gene trees (Salichos and Rokas 2013; Zhang et al. 2020b). Specifically, we inferred maximum likelihood gene trees with 1000 UltraFast Bootstrap (UFBoot) replicates as described in the *Tree Inference* section below. Then, we fitted a curve to the relationship between the mean UFBoot support and the number of variable sites from each alignment. Finally, we inferred the y-value of the inflexion point of this curve (mean UFBoot = 73%) and kept the alignments with mean UFboot > 73% and >100 variable sites (Fig. S2a). In addition, we pruned the taxa that were detected as outlier long branches in the gene trees using TreeShrink v.1.2.0 with the 'per-species' mode and the -b 20 flag (Mai and Mirarab 2018). TreeShrink estimates the contribution of each taxon to the diameter of the gene trees and identifies loci for which that contribution is unusually high. For those outlier loci, TreeShrink removes the taxon from the alignment if it increases the tree diameter by more than 20% (-b 20). This strategy can reduce systematic errors in downstream analyses and was shown to decrease phylogenetic conflicts (Mai and Mirarab 2018).

Loci set L1233.—A subset of 1233 amino acid alignments from the L1648 dataset. We generated this subset using the same procedure as for the L1082 dataset above. In this case, we kept the alignments with mean UFBoot > 60% and > 100 variable sites (Fig. S2b). We also pruned taxa detected as outlier long branches using TreeShrink as described above.

Alignments

All loci sets included highly divergent sequences because our taxon sampling spans an entire order of Cyanobacteria that is estimated to be at least 1 billion years old (Tomitani et al. 2006; Sánchez-Baracaldo 2015). Therefore, we used a strategy that improves alignment accuracy for highly divergent sequences by incorporating homologous sequences and protein structural information into the alignment process (Nute et al. 2019; Rozewicki et al. 2019). More specifically, we aligned all loci at the amino acid level using the `mafft-homologs.rb` script with MAFFT v7.475 and the `--dash` option (i.e., MAFFT-DASH Homologs; Katoh and Standley 2013; Rozewicki et al. 2019). MAFFT-DASH Homologs incorporated homologous sequences from two external sources: i) a Database of Aligned Structural Homologs (DASH), and ii) a custom BLAST database that contained all BUSCO loci from the 'nostocales_odb10' found in all 211 genomes included in this study (Table S1). In all cases, we used the `--globalpair` algorithm with 100 refinement iterations (Katoh 2005). We then obtained nucleotide alignments by back-translating the amino acid alignments using PAL2NAL v14 (Suyama et al. 2006) and the unaligned nucleotide sequences as input.

In all sequence datasets, we used trimAl v1.2rev59 (Capella-Gutierrez et al. 2009) to remove ambiguously aligned regions by trimming all sites with gaps (i.e., +ng; Table 1). Because some alignment filtering methods can impact the accuracy of phylogenetic inference (Tan et al. 2015; Steenwyk et al. 2020), we tested whether different trimming strategies could result in different levels of phylogenetic conflict. We generated additional alignments from the L1648 datasets by trimming with three additional strategies: i) removing non-parsimony informative (nPSI) and gappy sites using the 'smart-gappy' algorithm implemented in ClipKit v1.1.0 (Steenwyk et al. 2020), and keeping constant and parsimony informative (PSI) sites (i.e., L1648+kcg); ii) as in L1648+kcg but removing

sites with more than 20% gaps instead of using the smart-gappy algorithm (i.e., L1648+kcg2); and iii) trimming highly variable and gappy sites using the 'strict' algorithm implemented in trimAl (i.e., L1648+strict; Capella-Gutierrez et al. 2009).

For the L1648 datasets, we used AMAS (Borowiec 2016) to calculate the proportion of missing, variable, and parsimony-informative (PSI) sites, as well as the length of each gene alignment across all the datasets that we used in the study. The resulting proportion of variable sites was very similar across all four trimming strategies (Fig. S3b, f). However, the differences in the proportion of PSI sites (Fig. S3c, g) indicate that the variable sites are not fully overlapping across trimming strategies, even if they represent similar proportions in the alignments.

Maximum Likelihood Gene Trees

Substitution model selection.—Phylogenetic inference of deep divergences can yield spurious relationships when using amino acid substitution models that assume a single replacement matrix across all sites in an alignment (i.e., site-homogeneous models; Wang et al. 2019). The reality is that each amino acid site is constrained to a specific subset of the 20 amino acids depending on its position and role in the protein's structure (Goldman et al. 1998; Le et al. 2008a, 2008b). One way to account for these site-specific biochemical constraints is to model amino acid substitutions with a mixture of profiles of stationary frequencies of the amino acid residues (i.e., site-heterogeneous models; Lartillot and Philippe 2004; Le et al. 2008a). We performed model selection analyses with ModelFinder within IQ-Tree v1.6.12 (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017) and allowed the program to test 20 site-heterogeneous models with empirical profile mixtures (Le et al. 2008a) in addition to all the site-homogeneous models available. To determine which type of model was a better fit, we calculated the difference in Bayesian Information Criterion (BIC) of the best site-homogeneous and the best site-heterogeneous model for each alignment. If the difference is positive, this indicates that the site-heterogeneous model is better. We found that site-heterogeneous models provide a much better fit for almost every protein alignment in the L1648+ng dataset according to the Bayesian Information Criterion (Fig. S4). However, almost none of the published phylogenies of Nostocales (e.g., Fig. 1a-b,

d) were inferred using site-heterogeneous models. Therefore, we determined whether accounting for site-heterogeneity could change the number of phylogenetic conflicts among single-gene maximum likelihood phylogenies. This was achieved by inferring trees using both the best site-homogeneous and the best site-heterogeneous model for each protein alignment in all L1648 datasets (Table 1). For the nucleotide datasets, we partitioned the alignments into 1st, 2nd, and 3rd codon position and searched for the best partition scheme and substitution models using ModelFinder and PartitionFinder2 as implemented in IQ-Tree v1.6.12 (-m MFP+MERGE option; Nguyen et al. 2015; Lanfear et al. 2016; Kalyaanamoorthy et al. 2017).

Tree inference.—We inferred all gene trees using IQ-Tree with 1000 UFBoot replicates followed by nearest-neighbor interchange refinement via the `-bnni` flag (Hoang et al. 2018). For the nucleotide alignments, we used the `-spp` flag which allows different partitions to have different evolutionary rates but linked branch lengths. In total, we inferred 20 sets of gene trees including eight sets from the nucleotide datasets, eight sets inferred with site-heterogeneous models from the amino acid datasets, and four sets inferred with site-homogeneous models from the amino acid alignments of the L1648 datasets (Table 1).

Species Trees

Concatenated dataset trees.—We generated concatenated datasets with all loci sets in Table 1 using AMAS (Borowiec 2016). As with the gene trees, we inferred eight trees with the concatenated nucleotide alignments, eight trees with the concatenated amino acid alignments using site-heterogeneous models, and four trees with the L1648 amino acid alignments using site-homogeneous models (Table 1) for a total of 20 species trees.

For the nucleotide alignments, we first partitioned them into codon positions of each gene and searched for the best partition scheme and substitution models using the fast relaxed clustering algorithm from PartitionFinder2 within IQ-Tree. We used the options `-rclusterf 10` and `-rcluster-max 100` for computational efficiency and inferred maximum likelihood trees with 1000 UFBoot replicates in IQ-Tree.

For the trees inferred with the amino acid alignments and site-heterogeneous models, we used a posterior mean site frequency (PMSF) approach to infer profile mixtures of amino acid stationary frequencies (Wang et al. 2018). The site-specific amino acid frequencies of these profiles are calculated based on a mixture model fitted to a preliminary guide tree. We inferred the guide trees with the model LG+C20+F+G4, and then fitted a mixture model with 60 categories (LG+PMSF.C60+F+G4) and used the PMSF profiles to infer maximum likelihood trees with 1000 UFBoot replicates in IQ-Tree. The PMSF approach accelerates the inference and bootstrapping for large datasets while yielding equal or superior accuracy compared to pre-defined profile mixtures and partitioning (Wang et al. 2018, 2019).

For the trees inferred with the amino acid alignments and site-homogeneous models, we first assigned each gene to a different subset in the concatenated alignment. We then used the same procedure as for the concatenated nucleotide alignments to find the best partition scheme and considered only site-homogeneous models. Finally, we inferred maximum likelihood trees with 1000 UFBoot replicates in IQ-Tree.

ASTRAL trees.—We inferred 20 coalescent species trees using each of the 20 sets of maximum likelihood gene trees described above as input. First, we collapsed all internodes with UFBoot < 10% in the gene trees using Newick Utilities v1.6 (Junier and Zdobnov 2010) following Zhang et al. (2018). Then, we obtained coalescent species trees using ASTRAL-III (Zhang et al. 2018).

Analyses of Phylogenetic Conflict

Gene trees vs. concatenated dataset trees.—We compared the gene trees from each of the 20 sets to the corresponding trees derived from the concatenated datasets. For example, we compared the gene trees generated with the L1648+ng+aa+site-het. dataset to the tree derived from the concatenation of the genes from that same dataset. We then used DiscoVista (Sayyari et al. 2018) to calculate the proportion of gene trees that strongly support, strongly reject, weakly support, and weakly reject each of the bipartitions in the corresponding concatenated dataset tree. We used 95%

UFBoot as the threshold to evaluate strong support. When a gene tree had missing taxa, the corresponding missing taxa were removed from the concatenated dataset tree before evaluating conflict. To test whether some datasets induce different levels of conflict compared to other datasets, we performed analyses of variance (ANOVAs) to compare the proportions of strongly supporting, strongly rejecting, and uninformative (i.e., weak support, weak conflict, and missing) gene trees across the different sets. Because these proportions were calculated for each bipartition in the concatenated trees, and all concatenated trees had 55 taxa (subset 0), all the datasets that we compared had 52 data points corresponding to the 52 bipartitions in the concatenated trees.

Recovery of key topological bipartitions across datasets and inference methods.—We used DiscoVista to explore the recovery or rejection of 22 key bipartitions across 40 species trees, i.e., 20 trees inferred from concatenated datasets and 20 ASTRAL trees. Bipartitions 1–8 define the monophyly of eight of the major lineages that we delimited in this study: the 1) *Aphanizomenon*, 2) *Nostoc* II, 3) *Nostoc* I, 4) *Nodularia*, 5) *Fortiea*, 6) *Rivularia*, 7) *Fischerella*, and 8) *Scytonema* clades (Figs. 2 and 3a). Bipartitions 9–15 (Fig. 3a) are the relationships among these lineages as shown in the tree of Figure 2. Bipartitions 16–22 (Fig. 3a) are alternative relationships among these lineages, including some that were inferred in previous studies of Nostocales: bipartitions 16 and 17 are concordant with the tree in Figure 1a, bipartition 18 is an alternative relationship not recovered by any analysis so far, and bipartitions 19–22 are concordant with the tree shown in Figure 4.

Phylogenetic Networks with SNaQ

We used the Species Networks applying Quartets (SNaQ) method within the PhyloNetworks package to infer phylogenetic networks, which account simultaneously for incomplete lineage sorting and reticulated evolution (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). For computational tractability and to maximize accuracy, we ran these analyses on a subset of 12 taxa that included representatives from the major lineages of Nostocales (subset 1; Table 2; Solís-Lemus and Ané 2016). SNaQ takes concordance factors (CFs) as input, which are the frequencies of the three possible unrooted topologies of each set of four taxa (i.e., quartets) in a sample of gene trees. To account for

uncertainty in gene tree estimation, we inferred CFs and CF credibility intervals with BUCKy v1.4.4 using as input a sample of the posterior distribution of gene trees obtained with PhyloBayes v4.1 (Larget et al. 2010; Lartillot et al. 2015). For these analyses, we used a subset of the L1648+ng+aa alignments that included all genes with no missing data for the taxa in subset 1 (i.e., 1293 genes; Table 2). For each locus, we used ModelFinder as implemented in IQ-Tree to search for the best substitution model among all protein matrices (LG, WAG, JTT, MTREV, MTZOA, and MTART) and pre-defined amino acid profile mixtures (C20-C60) available in PhyloBayes v4.1. Then, we ran two MCMC chains for 50,000 cycles (note that PhyloBayes cycles are not analogous to chain generations in other Bayesian samplers) and sampled every 5 cycles for a total of 10000 trees per chain. We discarded the first 2000 trees from each chain as burnin, and then assessed convergence by estimating the largest discrepancy observed across all post-burnin bipartitions (maxdiff) in both chains using `bpcomp` from PhyloBayes v4.1. Convergence was achieved in all runs with maxdiff $<$ 0.01 (Lartillot 2020). We used the mbsum script from the TICR pipeline to prepare the trees for the analysis in BUCKy (Stenz et al. 2015). We then ran SNaQ allowing the maximum number of reticulations to vary from $h = 0$ to $h = 4$ and selected the best h using a slope heuristic suggested by Solís-Lemus and Ané (2016). Each SNaQ run was performed with 10 optimization iterations. Finally, we did a BootSNaQ analysis with the best h, where 100 bootstrap replicates were sampled from the CF credibility intervals inferred with BUCKy (Solís-Lemus and Ané 2016).

Goodness-of-fit test of the network coalescent model

We used QuartetNetworkGoodnessFit.jl, which is a method that assesses the fit between observed quartet concordance factors (qCFs) and the qCFs expected under a candidate phylogenetic network (Cai and Ané 2021). The expected qCFs are obtained by simulating gene trees that evolve on the candidate phylogenetic network using the network coalescent model. With this method, we compared the fit of three different phylogenetic hypotheses to explain the observed gene tree conflicts: i) SNaQ tree with $h = 0$, where all conflicts are explained as a result of ILS; ii) SNaQ network with $h = 2$, where conflicts are explained as a result of both ILS and two reticulations (Fig. 4); and iii) SNaQ tree with $h = 0$, where all internodes that give rise to the *Aphanizomenon*, *Nostoc* I,

Nostoc II, *Nodularia*, and *Fortiea* clades (internodes III-VI in Fig. 4a) are collapsed to length 0, which represents ancestral panmixia. We ran QuartetNetworkGoodnessFit.jl with 10,000 sets of 1293 gene trees simulated under the network coalescent for each of the three phylogenetic hypothesis. The simulated datasets were then used to generate the expected distribution of the proportion of outlier quartets. Finally, the algorithm tests whether the observed proportion of outlier quartets deviates significantly from 0.05 by comparing it to the expected distribution obtained from simulations (Cai and Ané 2021).

Test of the Anomaly Zone

The anomaly zone is a region of the tree parameter space where species trees are discordant with their most likely gene tree (Rosenberg 2013). Equation 4 in Degnan and Rosenberg (2006) can be used to calculate the value of $a(x)$, which is the boundary of the anomaly zone for internode x that has a descendant internode y. If $y < a(x)$, then x and y are in the anomaly zone. To test for the potential presence of internodes in the anomaly zone in Nostocales, we calculated $a(x)$ for each internode x in the major edges of the network with $h = 2$ inferred with SNaQ rooted with the outgroup taxa (Table 2). We then compared $a(x)$ to the branch lengths of each descendant internal branch y in coalescent units.

Divergence Time Estimation

To determine whether rapid speciation is linked to phylogenetic conflicts and the anomaly zone, we inferred divergence times for the major lineages of Nostocales. We used the approximate likelihood calculation implemented in MCMCTree, which allows Bayesian estimation of divergence times for a fixed topology and large phylogenomic alignments (Yang 2007; dos Reis and Yang 2011). We used the major edge topology of the phylogenetic network with $h = 2$ inferred with SNaQ for taxa subset 1, and the corresponding concatenated alignment of 1293+ng amino acid loci. This allowed us to directly link the timing of speciations with the topology that was used to test for the presence of the anomaly zone. We used two calibrations: i) a maximum age for the root set to 2,700 Myr with default right tail probability p_R=0.025 (Yang 2020), which is based on geological evidence for the early origin

of oxygenic photosynthesis (Farquhar et al. 2011; Uyeda et al. 2016); and ii) a calibration for the crown age of Nostocales with a minimum age set to 1,600 Myr based on fossil evidence of akinete-like structures which have a single origin in Nostocales, and a maximum age set to 2,320 Myr which is the lower bound for the rise in atmospheric oxygen and must have predated the evolution of heterocysts (Bekker et al. 2004; Tomitani et al. 2006). We used LG+G5 as the substitution model, an uncorrelated relaxed clock model with default priors, and a birth (λ)-death (μ) prior on node ages with $\lambda = \mu = 1$ and sampling fraction ρ = 0.1. This combination of parameters implies equal rates of extinction and speciation with an incomplete sampling, which reflects the fact that our analysis included only representatives from the major extant lineages. We sampled from both the prior and posterior distribution of divergence times using three MCMC chains with 500000 generations, sampling every 100th generation, and discarded the first 500 samples as burnin. We assessed convergence by comparing the mean posterior node ages inferred with each of the three chains (Fig. S5) and checking that the effective sample size was > 200 using custom R scripts (R Core Team 2013).

Phylogenomic Jackknifing

To explore the behavior of concatenation in the anomaly zone using datasets with different numbers of loci, we subsampled the 1293 amino acid gene alignments (part of L1648+ng) that have all taxa from subset 1 (Table 2). We sampled alignments randomly and without replacement (jackknifing) to generate subsets of 31, 51, 71, 91, 111, 131, 331, 531, 731, and 1131 loci. Using each of these sample sizes, we sampled 50 times for a total of 500 datasets. We then inferred maximum likelihood trees for each of these 500 concatenated alignments as described above for the amino acid datasets with the site-heterogeneous model (LG+PMSF.C60+F+G4) and 1000 UFBoot replicates in IQ-Tree.

We pruned taxa from all resulting trees to generate two sets of subtrees. The first set of subtrees includes only taxa that span internodes that fall in the anomaly zone (i.e., *Cylindrospermum stagnale* PCC 7417, *Nodularia* sp. NIES-3585, *Nostoc* sp. cyanobiont of *Peltigera aphthosa* JL23,

Fortiea contorta PCC 7126, *Trichormus variabilis* ATCC 29413, and *Tolypothrix* sp. NIES-4075). The second set of subtrees includes taxa that span the internodes outside of the anomaly zone (i.e., *Trichormus variabilis* ATCC 29413, *Tolypothrix* sp. NIES-4075, *Scytonema* sp. H-K 05, *Fischerella* sp. PCC 9605, *Rivularia* sp. PCC 7116, cyanobiont of *Peltula cylindrica, and Chroococcidiopsis thermalis* PCC 7203). We then calculated the percentage of subtrees inferred for each of the ten sample sizes where all internodes were highly supported (i.e., UFBoot > 95 %). Finally, we counted the number of distinct fully supported topologies in each of the two sets of subtrees. Two topologies were considered distinct if the Robinson-Foulds (RF) distance (Robinson and Foulds 1981) between them was > 0. RF distances were computed with the R package *ape* v5.4-1 (Paradis et al. 2004)

RESULTS

Conflicting relationships within Nostocales are mostly associated with deep short internodes

Our analyses of conflict among gene trees and trees from concatenated datasets revealed heterogeneity in the amount of conflicts and support across the phylogeny of Nostocales. All relationships recovered with the concatenated L1648+ng+aa+site-het. dataset had 100% UFBoot support (Fig. 2). In this tree, shallow internodes were often strongly or weakly supported by large fractions of the gene trees (dark blue and cyan areas, respectively, of pie charts in Fig. 2). In contrast, deeper and shorter internodes were seldom recovered, in some cases by less than 5% of gene trees (e.g., bipartitions 6, 9, 11, 13, and 14; Fig. 2). Nevertheless, the concatenated L1648+ng+aa+site-het. dataset recovered these bipartitions with 100% UFBoot support.

When we assessed which of the 40 species trees recovered each of the 22 key bipartitions (Fig. 3a), we found that bipartitions 1–8 were strongly supported by most species trees, including those inferred with concatenation (with amino acid and nucleotide data; Fig. 3b), or with ASTRAL (with amino acid data; Fig. 3c). Bipartition 4, which defines the *Nodularia* clade (Figs. 2 and 3a), was an exception to this trend because six of the phylogenetic trees based on concatenated datasets strongly rejected the monophyly of the *Nodularia* clade (Fig. 3b). However, all these strong conflicts were due to the unstable position of *Nostoc* sp. NIES 4103, which is the first split within the

Nodularia clade in Figure 2. The three other taxa included in this clade were always monophyletic. For deeper bipartitions 9–15, support was more heterogeneous, although most conflicts were not highly supported (Fig. 3b). Topological bipartitions 9, 11, and 13 were the most problematic (Fig. 3b, c), which were also among the least frequently recovered bipartitions by gene trees (Fig. 2).

Overall, the results from both sets of analyses of conflict (Figs. 2–3) indicate that there are two main regions of the Nostocales phylogeny that are most problematic, and both regions span deep and short internodes. One of these regions involves the relationships among the *Aphanizomenon*, *Nostoc* I, *Nostoc* II, *Nodularia*, and *Fortiea* clades (bipartitions 9, 10 and 11; Figs. 2 and 3b, c). The other problematic region involves relationships among the *Fischerella*, *Rivularia*, and *Scytonema* clades, and connection to the outgroup (bipartitions 13 and 14; Figs. 2 and 3b, c). These areas are also consistent with the conflicting relationships observed among published phylogenies of Nostocales (Fig. 1).

The effect of gene sampling and methodological choices on phylogenetic conflicts

When we compared gene trees from each dataset (Table 1) to the corresponding concatenated tree, we found no significant differences in the proportion of strongly concordant, strongly conflicting, or uninformative relationships among different datasets (Figs. S6 and S7). This was true for both amino acid (Fig. S6) and nucleotide datasets (Fig. S7), and across different sets of loci (Figs. S6a-c and S7a-c), alignment trimming strategies (Figs. S6d-i and S7d-f), and substitution model types for the amino acid data (Fig. S6d-i).

However, when we compared maximum likelihood trees based on concatenated datasets and ASTRAL species trees to 22 key bipartitions, we found differences in the patterns of topological conflicts (Fig. 3). First, ASTRAL trees inferred from nucleotide datasets with 746 genes or more were in strong conflict with almost all bipartitions from the tree shown in Figure 2 (bipartitions 1 through 14; Fig. 3a, c) and with alternative relationships recovered in this or previous studies (bipartitions 16 through 22; Fig. 3a, c). This is likely linked to systematic errors in the inference of the nucleotide gene trees due to the high level of saturation that is expected at the nucleotide level for these highly

divergent taxa. Since ASTRAL relies on gene trees, this type of error greatly impacts the inference of the species tree (Bossert et al. 2021; Zhang and Mirarab 2022). Second, trees inferred from datasets with lower number of genes (i.e., L31 and L70 datasets) recovered more bipartitions with weak support (Fig. 3b, c). Third, trees inferred from datasets from which genes with low phylogenetic signal were removed (L1082 and L1233 datasets; Table 1) had similar conflict patterns to the trees inferred from the datasets where these loci where not removed (L1648 datasets; Fig. 3b, c). The exception to this was the ASTRAL tree inferred with the L1233+ng+aa+site-het. dataset, which did not recover the monophyly of the order Nostocales (bipartition 15; Fig. 3c) nor the monophyly of the *Nostoc* I, *Nostoc* II, *Nodularia, Fortiea, Rivularia,* and *Fischerella* clades (bipartitions 2–7; Fig. 3c). Finally, most concatenated dataset trees recovered the *Fischerella* clade as sister to the *Rivularia* clade (bipartition 13; Fig. 3a, b), while most ASTRAL trees inferred from amino acid data recovered the *Fischerella* clade as sister to the *Scytonema* clade (bipartition 22; Fig. 3a, c).

With both concatenation and ASTRAL, we found no supported conflicts among the trees inferred from the L1648 datasets generated with different alignment trimming strategies, except for ASTRAL trees inferred with nucleotide data (Table 1; notice same pattern of conflict for all L1648 dataset trees in Fig. 3b, c). Likewise, using the better-fitting site heterogeneous models (Fig. S4) did not result in any strong conflicts with trees inferred using site homogeneous models for the amino acid datasets (Fig. 3b, c).

Anomaly Zone Linked to an Ancient Rapid Radiation

We found three consecutive internodes (internodes III-IV, IV-V and V-VI; Fig. 4a and Table 3) that fall within the anomaly zone (sensu Rosenberg 2013). These three internodes are in one of the areas that we identified as most problematic with our analyses of phylogenetic conflict shown in Figures 2 and 3 (i.e., the relationships among the *Aphanizomenon*, *Nostoc* I, *Nostoc* II, *Nodularia*, and *Fortiea* clades). The short coalescent lengths of these internodes indicate that there is a high level of phylogenetic conflict among the posterior distribution of gene trees that we used to infer the network (Table 3 and Fig. 4a). We hypothesized that this high level of conflict is the result of rapid and

consecutive speciation events (i.e., a rapid radiation) that led to extensive incomplete lineage sorting, which in turn generated the anomaly zone that we detected (Table 3). The highly overlapping posterior distributions of the ages for nodes III-VI relative to the surrounding nodes is in strong agreement with this hypothesis (Fig. 4c). We also sampled from the marginal prior distribution of node ages to ensure that the posterior results are driven by the data and not the priors (Fig. 4b-c; Brown and Smith 2018).

Evidence of reticulated evolution in Nostocales

According to the goodness-of-fit test, all three phylogenetic hypotheses that we tested fit the data poorly. The best hypothesis involved reticulations in addition to ILS (Table 4), indicating that some reticulations are needed to explain the data. The phylogenetic network with $h = 2$ was recovered as the best fit to the data according to the pseudolikelihood score heuristic and the goodness-of-fit test (Table 4; Fig. 4a; Table S2). In this network, the first reticulation has a minor edge with a γ suggesting that 29.6% of the core genome of the *Nodularia* lineage shares a most recent common ancestor with the *Nostoc* I lineage (Fig. 4a). The γ value of the minor edge of the second reticulation indicates that almost half (45.3%) of the core genome of *Fischerella* shares a most recent common ancestor with *Rivularia*. Interestingly, the sister relationship between *Fischerella* and *Scytonema* was the only one that was not highly supported by bootstrap analyses of concordance factors (i.e., BootSNaQ < 95%).

Concatenation of Phylogenomic Data Strongly Supports Conflicting Topologies Linked to the Anomaly Zone

To determine if concatenation of gene alignments with low phylogenetic signal, expected from the anomaly zone, can explain most of the conflicts among published phylogenies of Nostocales, which were based on concatenated datasets of various combination and numbers of genes, we generated random sets of genes including various numbers of genes using jackknifing. We found a high level of topological heterogeneity in the trees inferred from the concatenation of jackknifed sets of genes. The concatenation of at least 331 genes resulted in more than 75% of all subtrees part of the

anomaly zone having high support for all internodes (Fig. 5a–b). The same was true for relationships among taxa outside the anomaly zone (Fig. 5d–e). Strikingly, for the six taxa derived from the anomaly zone (Fig. 5a), we found 17 distinct and fully supported topologies (Fig. 5c). Moreover, the most frequent topology was different for almost every different number of genes (Fig. 5c). The most frequent topology across number of genes (Tc10; Fig. S8a) was also different from the one inferred with SNaQ (Fig. 5a). In contrast, for the seven taxa outside the anomaly zone (Fig. 5d), we only found 12 distinct and fully supported topologies. In this case, a single topology (Tf1; Fig. S8b) was recovered at the highest frequency with every number of genes (Fig. 5f). However, topology Tf1 has *Fischerella* sister to *Rivularia* (Fig. S8b), which is concordant with one of the plausible major edge relationships inferred by SNaQ (γ = 0.547; Figs. 4a and 5d).

DISCUSSION

Our results indicate that an ancient rapid radiation is the most likely cause of strongly supported phylogenetic conflicts among major lineages of Nostocales (Figs. 1, 4, and 5). A series of at least four rapid and consecutive speciation events that occurred > 1,000 Myr ago likely resulted in extensive incomplete lineage sorting, with two pairs of internodes falling within the anomaly zone (Fig. 4 and Table 3). Previous studies with fossil-calibrated molecular phylogenies had already recovered the internodes that define the relationships among the lineages involved in this anomaly zone (i.e., *Aphanizomenon*, *Nostoc* I, *Nostoc* II, *Nodularia*, and *Fortiea* clades; Fig. 4a) as some of the shortest across Cyanobacteria (Sánchez-Baracaldo 2015) and within Nostocales (Warshan et al. 2018). However, there is ample levels of variation in the estimated

divergence times of these lineages, including this study (Fig. 4c; Sánchez-Baracaldo et al. 2014; Sánchez-Baracaldo 2015; Warshan et al. 2018). Therefore, it is difficult to link the radiation to specific events from the past. Despite the large uncertainty associated with these estimates, we found strong evidence for the contemporaneous diversification of these lineages (Fig. 4c).

Radiations as an Overlooked Cause of Phylogenetic Conflicts Among Bacterial Phylogenies

Previous reports of empirical anomaly zones are from animal and plant phylogenies, including skink lizards (Linkem et al. 2016), Ostariophysan fishes (Chakrabarty et al. 2017), flightless birds (Cloutier et al. 2019), the *Gila robusta* complex of fish species (Chafin et al. 2021), the Trichophorae tribe of Cyperaceae plants (Léveillé-Bourret et al. 2020), and amaranth plants (Morales-Briones et al. 2021). Most likely, many other major lineages of the tree of life, including bacterial lineages, have also experienced extensive incomplete lineage sorting throughout their evolution as a result of radiations. Although bacterial reproduction is clonal, there is increasing evidence that they frequently exchange genes with conspecific relatives through homologous recombination (Vos and Didelot 2009; Pérez-Carrascal et al. 2019). This recombinogenic dynamic is necessary for incomplete lineage sorting to occur (Barraclough 2019; Bryant and Hahn 2020). Furthermore, the dispersal ability coupled with the relatively short life cycle of many bacteria may facilitate frequent instances of rapid range expansion (Martiny et al. 2006; Louca 2021). These rapid expansions can lead to fast diversification through isolation or local adaptation, as documented for microbial pathogens such as *Bacillus anthracis* (Vergnaud et al. 2016; Lienemann et al. 2018; Timofeev et al. 2019), *Yersinia pestis* (Keller et al. 2019), and SARS-CoV-2 (Morel et al. 2021). However, most studies addressing the sources of phylogenetic conflicts in bacteria have not accounted for incomplete lineage sorting (e.g., Murray et al. 2016; Coleman et al. 2021; but see Hernández-López et al. 2013 and Zhu et al. 2019). Although this is certainly not the only process that can impede phylogenetic inference, an increasing number of studies are describing radiations as major causes of phylogenetic conflicts (Pouchon et al. 2018; Alda et al. 2019; Roycroft et al. 2020; Lopes et al. 2021; Meleshko et al. 2021). Therefore, we suggest that future phylogenetic studies of bacteria consider radiations and incomplete lineage sorting as potential causes of phylogenetic incongruences. This is particularly important with the increasing use of phylogenomic datasets, which can lead to strong support for incorrect relationships (Huang et al. 2021).

The Role of Reticulated Evolution

Our findings suggest that HGT also plays an important role as a cause of some of the conflicts, especially concerning the relationships among *Fischerella*, *Rivularia*, and *Scytonema* lineages (Fig. 4a). The gamma values for the reticulation between *Fischerella* and *Rivularia* imply that almost half of the core genome of *Fischerella* sp. PCC 9605 is of hybrid origin (Fig. 4a). Most of the concatenated dataset and ASTRAL trees that we inferred, as well as published phylogenies of Nostocales, recovered mostly two alternative relationships: *Fischerella* sister to *Rivularia* or *Fischerella* sister to *Scytonema* (Figs. 1–3 and S8). This bias towards two of the possible alternative topologies is a further indication that the conflict is largely the result of reticulated evolution.

The approach we used to infer reticulations can only infer level-1 networks, which assumes that there are no overlapping reticulation cycles (Solís-Lemus and Ané 2016). Thus, there may be more reticulations that we are unable to detect with current methods. Indeed, the results of the goodness-of-fit test indicate that even the best network we can infer is still a poor fit to the data. Nonetheless, a recent study by Coleman et al. (2021) estimated that roughly two thirds of bacterial gene families have evolved vertically most of the time. Our results are consistent with this result and other studies that have shown that reticulated evolution is often not the main source of phylogenetic conflicts among core genes in bacterial phylogenies (Hernández-López et al. 2013; Murray et al. 2016).

Artifacts of phylogenomic analyses when inferring relationships in the Anomaly Zone using concatenated datasets

Our phylogenomic analyses of jackknifed datasets revealed that many distinct and fully supported topologies can be recovered with maximum likelihood in the anomaly zone by changing the set of concatenated genes (Fig. 5c). Previously, Kubatko and Degnan (2007) had shown with simulations that maximum likelihood inferences based on concatenated datasets are statistically inconsistent when the true species tree has high levels of incomplete lineage sorting, including the extreme case of the anomaly zone. Therefore, the probability of inferring a wrong tree topology with

high confidence increases with the amount of sequence data. More recently, Mendes and Hahn (2016) showed that the failure of concatenation coupled with maximum likelihood to recover the correct tree is not directly linked to the anomaly zone, and that the length of the branches surrounding the short internodes has an effect on whether the correct tree can be recovered or not. However, both of those studies were conducted with simulations of an anomaly zone involving only four taxa (i.e., rooted quartet). Our results for the order Nostocales suggest that there are at least five lineages (i.e., three internodes) whose relationships lie within the anomaly zone (Table 3). Mendes and Hahn (2016) emphasized that with five taxa, the behavior of maximum likelihood phylogenetic inferences based on concatenated dataset are expected to be far more complex (Rosenberg and Tao 2008; Rosenberg 2013).

We also found that most individual genes do not resolve relationships linked to the anomaly zone with high bootstrap support (Fig. 2). This low phylogenetic signal is also a signature of loci evolving under the anomaly zone that was predicted to occur in empirical datasets (Huang and Knowles, 2009). These authors concluded that the risks of the anomaly zone may rarely be realized with empirical datasets, since anomalous gene trees would not be recovered with high support. However, it has been shown that when concatenated alignments are large enough and alternative topologies have similar likelihoods, alternative resolutions can receive high bootstrap support stochastically (Huang et al. 2021). Moreover, when a concatenated dataset contains discordant sites, likelihood-based methods may accommodate the discordance by favoring a topology that is only weakly supported by each of the discordant sites (e.g., bipartition 9 in Figs. 2 and 3; Lewis et al. 2016; Shen et al. 2021). This agrees with the results of our phylogenomic inferences based on concatenated jackknifed datasets: if the few informative sites of a gene stochastically support alternative topologies, maximum likelihood based on concatenation of different sets of genes can recover multiple conflicting resolutions with high support. This behavior is the likely explanation for the strong conflicts observed among previous studies which inferred trees with maximum likelihood and different concatenated datasets (Fig. 1).

In contrast, only one of the three bipartitions that fall in the anomaly zone was recovered with high support by most ASTRAL trees based on amino acid gene trees (i.e., bipartition 10; Fig. 3c). Furthermore, two of the three internodes that fall in the anomaly zone were also recovered with extremely short coalescent lengths by ASTRAL (e.g., Fig. S9), which provides an appropriate description of the high amount of incomplete lineage sorting along these branches.

A Phylogenetic Framework for Nostocales

Several methods to infer species trees and phylogenetic networks are designed to deal with issues that result from high levels of incomplete lineage sorting, including the extreme case of the anomaly zone. However, it is possible that such methods (e.g., SNaQ network; Fig. 4a) can infer incorrect relationships among lineages that arose from a rapid radiation. The results of the goodness-of-fit test (Table 4) are evidence that a more complex phylogenetic hypothesis is needed to explain all conflicts in Nostocales. One reason for this is the potential violation of the assumption that the true network is of level-1 (Solís-Lemus and Ané 2016). This is relevant because one of the inferred reticulations involves lineages derived from the radiation (i.e., *Nodularia* and *Nostoc* I; Fig 4a), which indicates that not all observed gene tree conflicts can be accommodated as the result of incomplete lineage sorting. Therefore, the true network might have reticulations and, more importantly, an order of topological splits that cannot be inferred even with genome-wide datasets and the best currently available methods. Furthermore, when a radiation is deep in the tree, multiple substitutions on branches derived from the radiation may further obscure the already low phylogenetic signal of the short parental branches and decrease the accuracy of coalescent-based methods (Philippe et al. 2011; Liu et al. 2015). This may also be relevant for Nostocales, since we estimated that the mean crown age of the lineages involved in the radiation is 1,56 million years old, which falls within the range of previous estimates of 800 to 1,5 Myr (Fig. 4c; Sánchez-Baracaldo et al. 2014; Sánchez-Baracaldo 2015; Warshan et al. 2018).

Given these limitations, we propose that the relationships among the *Aphanizomenon, Fortiea, Nodularia, Nostoc* I, and *Nostoc* II clades be represented as a polytomy in future studies (i.e.,

0 length for internodes in anomaly zone), or the three short internodes part of the anomaly zone be highlighted as such. This would convey the inability to infer the precise order of speciation among these lineages—not for lack of data, but because of an evolutionary process resulting in a rapid radiation (Hoelzer and Meinick 1994; Walsh et al. 1999; Keller et al. 2019). This would also signal that future studies aimed at understanding the evolution of traits in nostocalean cyanobacteria should evaluate alternative scenarios across the heterogeneous pool of plausible genealogies of the radiation (Hahn and Nakhleh 2016).

We also recommend that future taxonomic studies of Nostocales consider each of the lineages derived from this radiation as separate families. This would prevent establishing unstable classifications that result from spurious relationships inferred from different gene datasets (Fig. 5c). Currently, several families of Nostocales include taxa that belong to more than one of the lineages derived from the radiation. For example, the family Aphanizomenonaceae comprises genera such as *Aphanizomenon* and *Nodularia*, which are part of the *Aphanizomenon* and *Nodularia* clades that we delimited (Fig. 2 and Fig. S1; Kaštovský et al. 2014; Komárek et al. 2014). However, we rarely recovered the *Aphanizomenon* and *Nodularia* clades as forming a monophyletic group (bipartition 19 in Fig. 3; topologies Tc2, Tc3, Tc7, Tc13 in Fig. 5c and Fig. S8). Segregating *Nodularia* and related genera into a separate family would reconcile the taxonomy with the phylogeny and would help to avoid problems related to the phylogenetic anomaly zone linked to the radiation that we detected.

Our results also indicated that no gene dataset is significantly better than the others in terms of the number of phylogenetic conflicts that they induce (Fig. S6 and Fig. S7). However, our ability to detect and distinguish between different biological causes of phylogenetic conflicts (e.g., ILS vs. HGT) increases when the gene datasets are larger because the underlying evolutionary processes can be modeled more accurately (Solís-Lemus and Ané 2016). Therefore, we recommend that future phylogenetic studies of Nostocales utilize the largest set of orthologous single-copy genes available. To that end, we released the alignments and concatenated tree from the most comprehensive dataset we assembled (Fig. S10) in the Tree-Based Alignment Selector Toolkit web platform (T-BAS; Carbone et al. 2019). In T-BAS, users can add new taxa to curated alignments and place them onto a

phylogenetic tree by using either evolutionary placement algorithms, or *de novo* phylogenetic tree inference with maximum likelihood. This tree includes all 211 taxa that passed our initial genome quality control, and was inferred with the L1648+ng dataset, and with the addition of the 16S rDNA gene. We hope that this resource will be useful for unifying phylogenetic hypotheses for Nostocales as more data become available, especially because there are multiple major lineages of Nostocales for which no genomes have been sequenced yet (Komárek et al. 2014; Ward et al. 2021).

SUPPLEMENTARY MATERIAL

Supplementary materials, Supplementary File S1, data, and code are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.tht76hf1p, and temporarily available at https://datadryad.org/stash/share/2ygrqiQjQFUh285v5pNfhHT48Azom_5zELaCFHEGW5Y. The scripts and detailed workflow instructions are also available in the GitHub repository https://github.com/cjpardodelahoz/nostocales.

FUNDING

This study was funded by the National Science Foundation award BEE 1929994 to FL and JM.

REFERENCES

Alda F., Tagliacollo V.A., Bernt M.J., Waltz B.T., Ludt W.B., Faircloth B.C., Alfaro M.E., Albert J.S., Chakrabarty P. 2019. Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting. Syst. Biol. 68:573–593.

Avni E., Snir S. 2020. A New Phylogenomic Approach For Quantifying Horizontal Gene Transfer Trends in Prokaryotes. Sci. Rep. 10:12425.

Barraclough T.G. 2019. Species and speciation without sex. The Evolutionary Biology of Species. Oxford: Oxford University Press. p. 110–131.

Bekker A., Holland H. D., Wang P. L., Rumble D. III, Stein H. J., Hannah J. L., ... Beukes N. J. 2004. Dating the rise of atmospheric oxygen. Nature. 427:117-120

Bell- Doyon P., Laroche J., Saltonstall K., Villarreal Aguilar J.C. 2020. Specialized bacteriome uncovered in the coralloid roots of the epiphytic gymnosperm, *Zamia pseudoparasitica*. Environ. DNA:edn3.66.

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660.

Bossert S., Murray E.A., Pauly A., Chernyshov K., Brady S.G., Danforth B.N. 2021. Gene Tree Estimation Error with Ultraconserved Elements: An Empirical Study on *Pseudapis* Bees. Syst. Biol. 70:803–821.

Bremer B., Jansen R.K., Oxelman B., Backlund M., Lantz H., Kim K.J. 1999. More characters or more taxa for a robust phylogeny - Case study from the coffee family (Rubiaceae). Syst. Biol. 48:413–435.

Brown J.W., Smith S.A. 2018. The Past Sure is Tense: On Interpreting Phylogenetic Divergence Time Estimates. Syst. Biol. 67:340–353.

Bryant D., Hahn M.W. 2020. The concatenation question. In: Scornavacca C., Delsuc F., Galtier N.,

editors. Phylogenetics in the Genomic Era. No commercial publisher | Authors open access book. p. 3.4:1–3.4:23.

Cai R., Ané C. 2021. Assessing the fit of the multi-species network coalescent to multi-locus data. Bioinformatics. 37:634-641

Capella-Gutierrez S., Silla-Martinez J.M., Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Carbone I., White J.B., Miadlikowska J., Arnold A.E., Miller M.A., Magain N., U'Ren J.M., Lutzoni F. 2019. T-BAS Version 2.1: Tree-Based Alignment Selector Toolkit for Evolutionary Placement of DNA Sequences and Viewing Alignments and Specimen Metadata on Curated and Custom Trees. Microbiol. Resour. Announc. 8:1–5.

Chafin T.K., Douglas M.R., Bangs M.R., Martin B.T., Mussmann S.M., Douglas M.E. 2021. Taxonomic Uncertainty and the Anomaly Zone: Phylogenomics Disentangle a Rapid Radiation to Resolve Contentious Species (*Gila robusta* Complex) in the Colorado River. Genome Biol. Evol. 13:1–19.

Chakrabarty P., Faircloth B.C., Alda F., Ludt W.B., Mcmahan C.D., Near T.J., Dornburg A., Albert J.S., Arroyave J., Stiassny M.L.J., Sorenson L., Alfaro M.E. 2017. Phylogenomic Systematics of Ostariophysan Fishes: Ultraconserved Elements Support the Surprising Non-Monophyly of Characiformes. Syst. Biol. 66:881–895.

Cloutier A., Sackton T.B., Grayson P., Clamp M., Baker A.J., Edwards S. V. 2019. Whole-Genome Analyses Resolve the Phylogeny of Flightless Birds (Palaeognathae) in the Presence of an Empirical Anomaly Zone. Syst. Biol. 68:937–955.

Coleman G.A., Davín A.A., Mahendrarajah T.A., Szánthó L.L., Spang A., Hugenholtz P., Szöllősi G.J., Williams T.A. 2021. A rooted phylogeny resolves early bacterial evolution. Science. 372:eabe0511.

Cornet L., Magain N., Baurain D., Lutzoni F. 2021. Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: a case study with Cyanobacteria and cyanolichens. Mol. Phylogenet. Evol. 162:107100.

Daubin V. 2002. A Phylogenomic Approach to Bacterial Phylogeny: Evidence of a Core of Genes Sharing a Common History. Genome Res. 12:1080–1090.

Degnan J.H., Rosenberg N.A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet. 2:e68.

Demoulin C.F., Lara Y.J., Cornet L., François C., Baurain D., Wilmotte A., Javaux E.J. 2019. Cyanobacteria evolution: Insight from the fossil record. Free Radic. Biol. Med.

Doolittle W.F. 1999a. Lateral genomics. Trends Biochem. Sci. 24:M5–M8.

Doolittle W.F. 1999b. Phylogenetic Classification and the Universal Tree. Science. 284:2124–2128.

Doolittle W.F., Boucher Y., Nesbø C.L., Douady C.J., Andersson J.O., Roger A.J., Andersson S.G.E., Martin W., Raven J.A., Lane N., Whatley F.R. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos. Trans. R. Soc. B Biol. Sci. 358:39–58.

Doolittle W.F., Logsdon Jr J.M. 1998. Archaeal genomics: Do archaea have a mixed heritage? Curr. Biol. 8:R209–R211.

Farquhar J., Zerkle A. L., Bekker A. 2011. Geological constraints on the origin of oxygenic photosynthesis. Photosynthesis research. 107:11-36

Gagunashvili A.N., Andrésson Ó.S. 2018. Distinctive characters of *Nostoc* genomes in cyanolichens. BMC Genomics. 19:1–18.

Goldman N., Thorne J.L., Jones D.T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

Greenlon A., Chang P.L., Damtew Z.M., Muleta A., Carrasquilla-Garcia N., Kim D., Nguyen H.P.,

Suryawanshi V., Krieg C.P., Yadav S.K., Patel J.S., Mukherjee A., Udupa S., Benjelloun I., Thami-Alami I., Yasin M., Patil B., Singh S., Sarma B.K., Von Wettberg E.J.B., Kahraman A., Bukun B., Assefa F., Tesfaye K., Fikre A., Cook D.R. 2019. Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proc. Natl. Acad. Sci. U. S. A. 116:15200–15209.

Groussin M., Poyet M., Sistiaga A., Kearney S.M., Moniz K., Noel M., Hooker J., Gibbons S.M., Segurel L., Froment A., Mohamed R.S., Fezeu A., Juimo V.A., Lafosse S., Tabe F.E., Girard C., Iqaluk D., Nguyen L.T.T., Shapiro B.J., Lehtimäki J., Ruokolainen L., Kettunen P.P., Vatanen T., Sigwazi S., Mabulla A., Domínguez-Rodrigo M., Nartey Y.A., Agyei-Nkansah A., Duah A., Awuku Y.A., Valles K.A., Asibey S.O., Afihene M.Y., Roberts L.R., Plymoth A., Onyekwere C.A., Summons R.E., Xavier R.J., Alm E.J. 2021. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell. 184:2053-2067.e18.

Gutiérrez-García K., Bustos-Díaz E.D., Corona-Gómez J.A., Ramos-Aboites H.E., Sélem-Mojica N., Cruz-Morales P., Pérez-Farrera M.A., Barona-Gómez F., Cibrián-Jaramillo A. 2019. Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and *Caulobacter* spp. That Encode Niche-Specific Biosynthetic Gene Clusters. Genome Biol. Evol. 11:319–334.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution. 70:7–17.

Hernández-López A., Chabrol O., Royer-Carenzi M., Merhej V., Pontarotti P., Raoult D. 2013. To Tree or Not to Tree? Genome-Wide Quantification of Recombination and Reticulate Evolution during the Diversification of Strict Intracellular Bacteria. Genome Biol. Evol. 5:2305–2317.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. Mol. Biol. Evol. 35:518–522.

Hoelzer G.A., Meinick D.J. 1994. Patterns of speciation and limits to phylogenetic resolution. Trends Ecol. Evol. 9:104–107.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

Huang H., Knowles L.L. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58:527–536.

Huang J., Liu Y., Zhu T., Yang Z. 2021. The Asymptotic Behavior of Bootstrap Support Values in Molecular Phylogenetics. Syst. Biol. 70:774–785.

Janda J.M., Abbott S.L. 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J. Clin. Microbiol. 45:2761–2764.

Junier T., Zdobnov E.M. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 26:1669–1670.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

Kang D.D., Li F., Kirton E., Thomas A., Egan R., An H., Wang Z. 2019. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019:1–13.

Kaštovský J., Berrendero Gomez E., Hladil J., Johansen J.R. 2014. *Cyanocohniella calida* gen. et sp. nov. (Cyanobacteria: Aphanizomenonaceae) a new cyanobacterium from the thermal springs from Karlovy Vary, Czech Republic. Phytotaxa. 181:279.

Katoh K. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33:511–518.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Keller M., Spyrou M.A., Scheib C.L., Neumann G.U., Kröpelin A., Haas-Gebhard B., Päffgen B.,

Haberstroh J., Ribera i Lacomba A., Raynaud C., Cessford C., Durand R., Stadler P., Nägele K., Bates J.S., Trautmann B., Inskip S.A., Peters J., Robb J.E., Kivisild T., Castex D., McCormick M., Bos K.I., Harbeck M., Herbig A., Krause J. 2019. Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750). Proc. Natl. Acad. Sci. 116:12363–12372.

Kim J., Na S., Kim D., Chun J. 2021. UBCG2: Up-to-date bacterial core genes and pipeline for phylogenomic analysis. J. Microbiol. 59:609–615.

Komárek J., Kaštovský J., Mareš J., Johansen J.R. 2014. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. Preslia. 86:295–335.

Kriventseva E. V., Kuznetsov D., Tegenfeldt F., Manni M., Dias R., Simão F.A., Zdobnov E.M. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47:D807–D811.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Syst. Biol. 56:17–24.

Kurland C.G., Canback B., Berg O.G. 2003. Horizontal gene transfer: A critical view. Proc. Natl. Acad. Sci. 100:9658–9662.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2016. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. Mol. Biol. Evol. 34:msw260.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics. 26:2910–2911.

Lartillot N. 2020. Phylobayes: Bayesian phylogenetics using site-heterogeneous models. .

Lartillot N., Blanquart S., Lepage T. 2015. PhyloBayes. A Bayesian software for phylogenetic reconstruction using mixture models. PhyloBayes manual:1–21

Lartillot N., Lepage T., Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–2288.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lawrence J.G., Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. 95:9413–9417.

Le S.Q., Gascuel O., Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics. 24:2317–2323.

Le S.Q., Lartillot N., Gascuel O. 2008b. Phylogenetic mixture models for proteins. Philos. Trans. R. Soc. B Biol. Sci. 363:3965–3976.

Léveillé-Bourret É., Chen B.-H., Garon-Labrecque M.-È., Ford B.A., Starr J.R. 2020. RAD sequencing resolves the phylogeny, taxonomy and biogeography of Trichophoreae despite a recent rapid radiation (Cyperaceae). Mol. Phylogenet. Evol. 145:106727.

Lewis P.O., Chen M.-H., Kuo L., Lewis L.A., Fučíková K., Neupane S., Wang Y.-B., Shi D. 2016. Estimating Bayesian Phylogenetic Information Content. Syst. Biol. 65:1009–1023.

Li Y., Shen X.-X., Evans B., Dunn C.W., Rokas A. 2020. Rooting the animal tree of life. Mol. Biol. Evol. 38: 4322–4333.

Lienemann T., Beyer W., Pelkola K., Rossow H., Rehn A., Antwerpen M., Grass G. 2018. Genotyping and phylogenetic placement of *Bacillus anthracis* isolates from Finland, a country with rare anthrax cases. BMC Microbiol. 18:102.

Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the Anomaly Zone in Species Trees and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny (Squamata: Scincidae).

Syst. Biol. 65:465–477.

Liu L., Xi Z., Davis C.C. 2015. Coalescent Methods Are Robust to the Simultaneous Effects of Long Branches and Incomplete Lineage Sorting. Mol. Biol. Evol. 32:791–805.

Lopes F., Oliveira L.R., Kessler A., Beux Y., Crespo E., Cárdenas-Alayza S., Majluf P., Sepúlveda M., Brownell R.L., Franco-Trecu V., Páez-Rosas D., Chaves J., Loch C., Robertson B.C., Acevedo-Whitehouse K., Elorriaga-Verplancken F.R., Kirkman S.P., Peart C.R., Wolf J.B.W., Bonatto S.L. 2021. Phylogenomic Discordance in the Eared Seals is best explained by Incomplete Lineage Sorting following Explosive Radiation in the Southern Hemisphere. Syst. Biol. 70:786–802.

Louca S. 2021. The rates of global bacterial and archaeal dispersal. ISME J.:24–31.

Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics. 19:272.

Martiny J.B.H., Bohannan B.J.M., Brown J.H., Colwell R.K., Fuhrman J.A., Green J.L., Horner-Devine M.C., Kane M., Krumins J.A., Kuske C.R., Morin P.J., Naeem S., Øvreås L., Reysenbach A.-L., Smith V.H., Staley J.T. 2006. Microbial biogeography: putting microorganisms on the map. Nat. Rev. Microbiol. 4:102–112.

McDonald T.R., Mueller O., Dietrich F.S., Lutzoni F. 2013. High-throughput genome sequencing of lichenizing fungi to assess gene loss in the ammonium transporter/ammonia permease gene family. BMC Genomics. 14:225.

McInerney J.O., Cotton J.A., Pisani D. 2008. The prokaryotic tree of life: past, present…and future? Trends Ecol. Evol. 23:276–281.

Meleshko O., Martin M.D., Korneliussen T.S., Schröck C., Lamkowski P., Schmutz J., Healey A., Piatkowski B.T., Shaw A.J., Weston D.J., Flatberg K.I., Szövényi P., Hassel K., Stenøien H.K. 2021. Extensive Genome-Wide Phylogenetic Discordance Is Due to Incomplete Lineage Sorting

and Not Ongoing Introgression in a Rapidly Radiated Bryophyte Genus. Mol. Biol. Evol. 38:2750–2766.

Mendes F.K., Hahn M.W. 2016. Gene Tree Discordance Causes Apparent Substitution Rate Variation. Syst. Biol. 65:711–721.

Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling Sources of Gene Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in Amaranthaceae s.l. Syst. Biol. 70:219–235.

Morel B., Barbera P., Czech L., Bettisworth B., Hübner L., Lutteropp S., Serdari D., Kostaki E.-G., Mamais I., Kozlov A.M., Pavlidis P., Paraskevis D., Stamatakis A. 2021. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Mol. Biol. Evol. 38:1777–1791.

Murray G.G.R., Weinert L.A., Rhule E.L., Welch J.J. 2016. The Phylogeny of *Rickettsia* Using Different Evolutionary Signatures: How Tree-Like is Bacterial Evolution? Syst. Biol. 65:265–279.

Nelson J.M., Hauser D.A., Gudiño J.A., Guadalupe Y.A., Meeks J.C., Salazar Allen N., Villarreal J.C., Li F.-W. 2019. Complete Genomes of Symbiotic Cyanobacteria Clarify the Evolution of Vanadium-Nitrogenase. Genome Biol. Evol. 11:1959–1964.

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C., Fraser C.M. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature. 399:323–329.

Nesbø C.L., Boucher Y., Doolittle W.F. 2001. Defining the core of nontransferable prokaryotic genes: The euryarchaeal core. J. Mol. Evol. 53:340–350.

Nguyen L.T., Schmidt H.A., Von Haeseler A., Minh B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27:824–834.

Nute M., Saleh E., Warnow T. 2019. Evaluating Statistical Multiple Sequence Alignment in Comparison to Other Alignment Methods on Protein Data Sets. Syst. Biol. 68:396–411.

Orakov A., Fullam A., Coelho L.P., Khedkar S., Szklarczyk D., Mende D.R., Schmidt T.S.B., Bork P. 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol. 22:1–19.

Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.

Parks D.H., Rinke C., Chuvochina M., Chaumeil P.-A., Woodcroft B.J., Evans P.N., Hugenholtz P., Tyson G.W. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2:1533–1542.

Pérez-Carrascal O.M., Terrat Y., Giani A., Fortin N., Greer C.W., Tromas N., Shapiro B.J. 2019. Coherence of *Microcystis* species revealed through population genomics. ISME J. 13:2887–2900.

Philippe H., Brinkmann H., Lavrov D. V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9.

Pouchon C., Fernández A., Nassar J.M., Boyer F., Aubert S., Lavergne S., Mavárez J. 2018. Phylogenomic Analysis of the Explosive Adaptive Radiation of the *Espeletia* Complex (Asteraceae) in the Tropical Andes. Syst. Biol. 67:1041–1060.

R Core Team A. 2013. R: A language and environment for statistical computing.

Redmond A.K., McLysaght A. 2021. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. Nat. Commun. 12:1783.

dos Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian Estimation of Divergence Times. Mol. Biol. Evol. 28:2161–2172.

Richards E.J., Brown J.M., Barley A.J., Chong R.A., Thomson R.C. 2018. Variation Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much Gene Tree Variation Is Biological? Syst. Biol. 67:847–860.

Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas A., Carroll S.B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.

Rosenberg N.A. 2013. Discordance of Species Trees with Their Most Likely Gene Trees: A Unifying Principle. Mol. Biol. Evol. 30:2709–2713.

Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Syst. Biol. 57:131–140.

Roycroft E.J., Moussalli A., Rowe K.C. 2020. Phylogenomics Uncovers Confidence and Conflict in the Rapid Radiation of Australo-Papuan Rodents. Syst. Biol. 69:431–444.

Rozewicki J., Li S., Amada K.M., Standley D.M., Katoh K. 2019. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 47:W5–W10.

Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.

Sánchez-Baracaldo P. 2015. Origin of marine planktonic cyanobacteria. Sci. Rep. 5:17418.

Sánchez-Baracaldo P., Ridgwell A., Raven J.A. 2014. A neoproterozoic transition in the marine

nitrogen cycle. Curr. Biol. 24:652–657.

Sayyari E., Whitfield J.B., Mirarab S. 2018. DiscoVista: Interpretable visualizations of gene tree discordance. Mol. Phylogenet. Evol. 122:110–115.

Shen X.-X., Hittinger C.T., Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.

Shen X.-X., Steenwyk J.L., Rokas A. 2021. Dissecting Incongruence between Concatenation- and Quartet-Based Approaches in Phylogenomic Data. Syst. Biol. 70:997–1014.

Shi T., Falkowski P.G. 2008. Genome evolution in cyanobacteria: The stable core and the variable shell. Proc. Natl. Acad. Sci. 105:2510–2515.

Shih P.M., Wu D., Latifi A., Axen S.D., Fewer D.P., Talla E., Calteau A., Cai F., Tandeau de Marsac N., Rippka R., Herdman M., Sivonen K., Coursin T., Laurent T., Goodwin L., Nolan M., Davenport K.W., Han C.S., Rubin E.M., Eisen J.A., Woyke T., Gugger M., Kerfeld C.A. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc. Natl. Acad. Sci. 110:1053–1058.

Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E. V., Zdobnov E.M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:1–15.

Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLOS Genet. 12:e1005896.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: A Package for Phylogenetic Networks. Mol. Biol. Evol. 34:3292–3298.

Sorek R., Zhu Y., Creevey C.J., Francino M.P., Bork P., Rubin E.M. 2007. Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer. Science. 318:1449–1452.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313.

Steenwyk J.L., Buida T.J., Li Y., Shen X.-X., Rokas A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biol. 18:e3001007.

Stenz N.W.M., Larget B., Baum D.A., Ané C. 2015. Exploring Tree-Like and Non-Tree-Like Patterns Using Genome Sequences: An Example Using the Inbreeding Plant Species *Arabidopsis thaliana* (L.) Heynh. Syst. Biol. 64:809–823.

Suyama M., Torrents D., Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Taib N., Megrian D., Witwinowski J., Adam P., Poppleton D., Borrel G., Beloin C., Gribaldo S. 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. Nat. Ecol. Evol. 4:1661–1672.

Tan G., Muffato M., Ledergerber C., Herrero J., Goldman N., Gil M., Dessimoz C. 2015. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. Syst. Biol. 64:778–791.

Timofeev V., Bahtejeva I., Mironova R., Titareva G., Lev I., Christiany D., Borzilov A., Bogun A., Vergnaud G. 2019. Insights from *Bacillus anthracis* strains isolated from permafrost in the tundra zone of Russia. PLoS One. 14:e0209140.

Tomitani A., Knoll A.H., Cavanaugh C.M., Ohno T. 2006. The evolutionary diversification of cyanobacteria: Molecular-phylogenetic and paleontological perspectives. Proc. Natl. Acad. Sci. 103:5442–5447.

Tringe S.G., Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. Curr. Opin.

Microbiol. 11:442–446.

Uyeda J. C., Harmon L. J., Blank C. E. 2016. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. PloS one. 11:e0162539

Vergnaud G., Girault G., Thierry S., Pourcel C., Madani N., Blouin Y. 2016. Comparison of French and Worldwide *Bacillus anthracis* Strains Favors a Recent, Post-Columbian Origin of the Predominant North-American Clade. PLoS One. 11:e0146216.

Vos M., Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 3:199–208.

Walsh H.E., Kidd M.G., Moum T., Friesen V.L. 1999. Polytomies and the power of phylogenetic inference. Evolution. 53:932–937.

Walter J.M., Coutinho F.H., Dutilh B.E., Swings J., Thompson F.L., Thompson C.C. 2017. Ecogenomics and Taxonomy of Cyanobacteria Phylum. Front. Microbiol. 8.

Wang H.-C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. 67:216–235.

Wang H.C., Susko E., Roger A.J. 2019. The Relative Importance of Modeling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference. Syst. Biol. 68:1003–1019.

Ward R.D., Stajich J.E., Johansen J.R., Huntemann M., Clum A., Foster B., Foster B., Roux S., Palaniappan K., Varghese N., Mukherjee S., Reddy T.B.K., Daum C., Copeland A. 2021. Metagenome Sequencing to Explore Phylogenomics of Terrestrial Cyanobacteria. Microbiol. Resour. Announc. 10:e00258-21.

Warshan D., Liaimer A., Pederson E., Kim S.-Y., Shapiro N., Woyke T., Altermark B., Pawlowski

K., Weyman P.D., Dupont C.L., Rasmussen U. 2018. Genomic Changes Associated with the Evolutionary Transitions of *Nostoc* to a Plant Symbiont. Mol. Biol. Evol. 35:1160–1175.

Woese C.R., Fox G.E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Natl. Acad. Sci. 74:5088–5090.

Wu M., Eisen J.A. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 2020. User Guide PAML: Phylogenetic Analysis by Maximum Likelihood. Version 4.9j. .

Yarza P., Richter M., Peplies J., Euzeby J., Amann R., Schleifer K.-H., Ludwig W., Glöckner F.O., Rosselló-Móra R. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. 31:241–250.

Zhang C., Mirarab S. 2022. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. bioArxiv.:1–28.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:15–30.

Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2020a. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. Mol. Biol. Evol.

Zhang R., Wang Y.-H., Jin J.-J., Stull G.W., Bruneau A., Cardoso D., De Queiroz L.P., Moore M.J., Zhang S.-D., Chen S.-Y., Wang J., Li D.-Z., Yi T.-S. 2020b. Exploration of Plastid Phylogenomic Conflict Yields New Insights into the Deep Relationships of Leguminosae. Syst. Biol. 69:613–622.

Zhaxybayeva O. 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.

Zhu Q., Mai U., Pfeiffer W., Janssen S., Asnicar F., Sanders J.G., Belda-Ferre P., Al-Ghalith G.A., Kopylova E., McDonald D., Kosciolek T., Yin J.B., Huang S., Salam N., Jiao J., Wu Z., Xu Z.Z., Cantrell K., Yang Y., Sayyari E., Rabiee M., Morton J.T., Podell S., Knights D., Li W., Huttenhower C., Segata N., Smarr L., Mirarab S., Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat. Commun. 10:5477.

Zolan M.E., Pukkila P.J. 1986. Inheritance of DNA methylation in *Coprinus cinereus*. Mol. Cell. Biol. 6:195–200.

FIGURE CAPTIONS

FIGURE 1. Examples of conflicts among published phylogenies of Nostocales. Schematic representation of relationships among major lineages of Nostocales inferred from different sets of concatenated loci across four different studies: a) Gagunashvili & Andrésson (2018), b) Warshan et al. (2018), c) Sánchez-Baracaldo (2015), and d) Nelson et al. (2019). We allocated the strains included in these studies to each of the nine major clades following their position in our comprehensive phylogeny in Figure S1 (Available in the Dryad Digital repository https://doi.org/10.5061/dryad.tht76hf1p).

FIGURE 2. Distribution of phylogenetic conflicts in the phylogeny of Nostocales. The tree is a maximum likelihood phylogeny inferred using the concatenated L1648+ng+aa+site-het. dataset with the model LG+PMSF.C60+G4+F for 55 taxa (subset 0). All internodes in the tree have 100% UFboot support. Pie charts show the proportion of amino acid gene trees inferred from the L1648+ng+aa+site-het. dataset that recovered each topological bipartition with strong support, strong conflict, weak support, or weak conflict, or that was not scored due to missing data. We used 95% UFboot as support threshold to assess conflict. Numbers next to pie charts show the bipartitions that we tracked for the analyses of conflict presented in Figure 3. Internodes highlighted in grey, part of the anomaly zone, were detected using the coalescent lengths of the major edge tree in Figure 4 (Table 3).

FIGURE 3. Distribution of conflicts for a) 22 key topological bipartitions across b) 20 concatenated dataset trees and c) 20 ASTRAL trees. Bipartitions 1–15 are concordant, and bipartitions 16–22 are in conflict, with the tree shown in Figure 2. Each colored circle represents the set of taxa highlighted with the same color in Figure 2. The heatmaps show whether each bipartition (columns) was recovered with strong or low support, or in strong or weak conflict, by each species tree (rows). See Table 1 for descriptions of the datasets and models of evolution used to generate these 40 species trees. We used 95% UFboot as support threshold to assess conflict. Internodes highlighted as part of

the anomaly zone were detected using the coalescent lengths of the major edge tree in Figure 4 (Table 3).

FIGURE 4. Anomaly zone linked to four rapid and consecutive speciation events. a) Phylogenetic network inferred with SNaQ based on posterior distributions of amino acid gene trees from 1293 genes selected from the L1648+ng alignments (i.e., excluding loci and taxa with missing data; subset 1; Table 2). Branch lengths represent coalescent time, with shorter internodes indicating higher conflict among gene trees. The internodes linking nodes III–VI are in the anomaly zone (Table 3; Degnan and Rosenberg 2006). Asterisks indicate edges with BootSNaQ > 95%. Hybrid edges are shown with colors. γ values indicate the inferred proportion of genes inherited by a hybrid node from each of its hybrid parents (Solís-Lemus and Ané 2016). b) Marginal prior density of ages for the nodes of interest. c) Posterior density of ages for the nodes of interest. Internodes highlighted as part of the anomaly zone were detected as shown in Table 3.

FIGURE 5. Phylogenetic support for relationships linked to the anomaly zone based on concatenated jackknife sets of increasing number of genes. The jackknifed gene sets were sampled from the amino acid L1648+ng alignments of taxa subset 1 (Table 2). We sampled 50 jackknife replicates for each number of genes. All trees were inferred using the site-heterogeneous model LG+PMSF.C60+G4+F and 1000 UFBoot replicates. Trees were trimmed to include only the lineages whose relationships (a) span the anomaly zone (thick branches and bold taxon names), or d) that were outside the anomaly zone (thick branches and bold taxon names). b) Percentage of anomaly zone subtrees inferred from jackknife sets of increasing number of genes where all internodes where highly supported (UFBoot > 95%). e) Percentage of non-anomaly zone subtrees inferred from jackknife sets of increasing number of genes where all internodes where highly supported (UFBoot > 95%). c) Distribution of distinct, fully supported topologies (i.e., all internodes with UFBoot > 95%) of subtrees in the anomaly zone. f) Distribution of distinct, fully supported topologies (i.e., all internodes with UFBoot > 95%) of

subtrees outside the anomaly zone. Two topologies were considered distinct if the Robinson-Foulds distance was > 0. Numbers above each bar indicate the number of distinct topologies. All topologies referenced in panels c and f are shown in Figure S8.

TABLE 1. Summary of datasets and model types used in this study. The number following the letter L of each loci set refers to the number genes part of each set.

Loci set	**Trimming strategy**	**Data type**	**Aa model type**	**Datasets**
L31	+ng: Remove all sites with gaps.	+aa, +na	+site-het.	L31+ng+aa+site-het. L31+ng+na+site-hom.
L70	+ng: Remove all sites with gaps.	+aa, +na	+site-het.	L70+ng+aa+site-het. L70+ng+na+site-hom.
L746	+ng: Remove all sites with gaps.	+aa, +na	+site-het.	L746+ng+aa+site-het. L746+ng+na+site-hom.
L1082	+ng: Remove all sites with gaps.	+na	+site-hom.	L1082+ng+na+site-hom.
L1233	+ng: Remove all sites with gaps.	+aa	+site-het.	L1233+ng+aa+site-het.
L1648	+ng: Remove all sites with gaps.	+aa, +na	+site-het., +site-hom.	L1648+ng+aa+site-het. L1648+ng+aa+site-hom. L1648+ng+na+site-hom.
L1648	+kcg: Keep constant and parsimony informative sites, trim with smart-gappy[a] algorithm.	+aa, +na	+site-het., +site-hom.	L1648+kcg+aa+site-het. L1648+kcg+aa+site-hom. L1648+kcg+na+site-hom.
L1648	+kcg2: Keep constant and parsimony informative sites, trim sites with > 20% gaps.	+aa, +na	+site-het., +site-hom.	L1648+kcg2+aa+site-het. L1648+kcg2+aa+site-hom. L1648+kcg2+na+site-hom.
L1648	+strict: Trim with trimAl strict algorithm.	+aa, +na	+site-het., +site-hom.	L1648+strict+aa+site-het. L1648+strict+aa+site-hom. L1648+strict+na+site-hom.

Notes: +aa: amino acid, +na: nucleotide, +site-het.: site heterogeneous model, +site-hom.: site homogeneous model. All nucleotide datasets were analyzed with site-homogeneous models.

[a]Dynamic gappyness threshold determination approach implemented in ClipKit (Steenwyk et al. 2020).

TABLE 2. Taxa subset 1 used for phylogenetic network analyses. Representatives from each major clade were selected from the taxa included in subset 0 (Fig. 2). Subset 1 consists of 1293 genes, derived from the L1648+ng+aa alignments and includes only genes and taxa without missing data.

Clade	**Subset 1**
Fischerella	*Fischerella* sp. PCC 9605
Fortiea	*Fortiea contorta* PCC 7126
Nodularia	*Nodularia* sp. NIES-3585
Nostoc I	*Trichormus variabilis* ATCC 29413
Nostoc II	*Nostoc* sp. cyanobiont of *Peltigera aphthosa* JL23
Aphanizomenon	*Anabaena cylindrica* PCC 7122 *Cylindrospermum stagnale* PCC 7417
Rivularia	*Rivularia* sp. PCC 7116
Scytonema	*Scytonema* sp. NIES 2130 (HK-05) v2
Tolypothrix	*Tolypothrix* sp. NIES-4075
Outgroup	*Chroococcidiopsis thermalis* PCC 7203 Cyanobiont of *Peltula cylindrica*

TABLE 3. Delimitation of the anomaly zone. We calculated $a(x)$ using equation 4 from Degnan and Rosenberg (2006) for each internode x in the tree shown in Figure 4. For an internode x that has a descendant internode y, the pair x and y are in the anomaly zone if $y < a(x)$. Roman numerals refer to node labels in Figure 4a.

Internode pair (x, y)	**x length (coal. units[a])**	**y length (coal. units)**	**$a(x)$**	**Anomaly zone? i.e., $y < a(x)$**
I-II, II-III	0.718	1.704	-0.222	No
II-III, III-IV	1.704	0.028	-0.352	No
III-IV, IV-V	0.028	0.163	1.042	Yes
IV-V, V-VI	0.163	0.062	0.314	Yes
V-VI, VI-VII	0.062	0.976	0.552	No
I-VIII, VIII-IX	0.239	0.735	0.028	No

[a]Coal. units: coalescent units, which are equivalent to the number of generations since divergence divided by the effective population size. In the network coalescent model, it is also a direct measure of phylogenetic conflict, such that lower values indicate higher conflict among gene trees evolving under the major tree topology (Solís-Lemus and Ané 2016).

TABLE 4. Results of the goodness-of-fit test for three phylogenetic hypotheses to determine whether the observed proportion of outlier quartets deviates significantly from 0.05, by comparing it to a distribution obtained from simulated gene trees. P-values < 0.05 indicate that there is a significant deviation, and thus the phylogenetic hypothesis is a poor fit to the data. The magnitude of the p-values can be used to compare the fit of different phylogenetic hypotheses relative to each other, where more negative p-values indicate a worst fit.

Phylogentic hypothesis	**Source of conflict**	**p-value**
SNaQ tree with $h = 0$	All conflict is due to ILS	2.15×10^{-61}
SNaQ network with $h = 2$	Conflict is the result of both ILS and reticulations	1.11×10^{-44}
SNaQ tree with $h = 0$ and problematic internodes collapsed	Ancestral panmixia. All conflict is due to ILS	2.15×10^{-66}

Figure 1

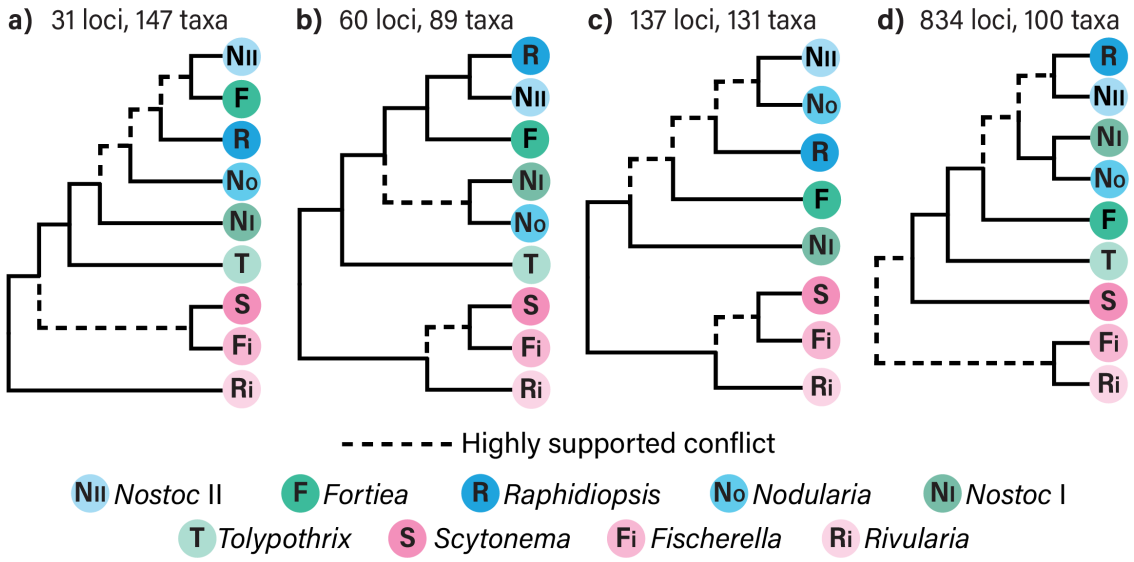

Figure 2

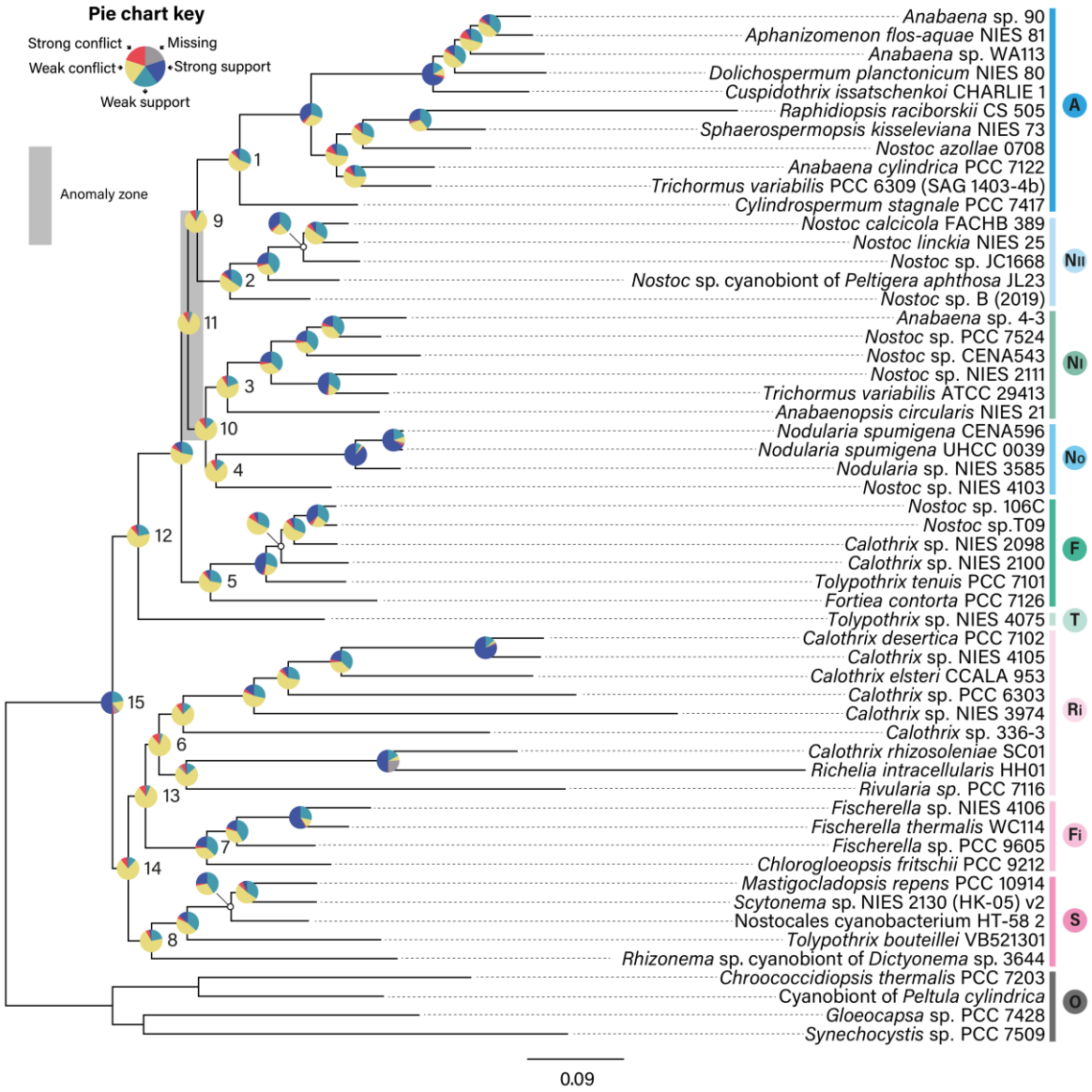

Figure 3

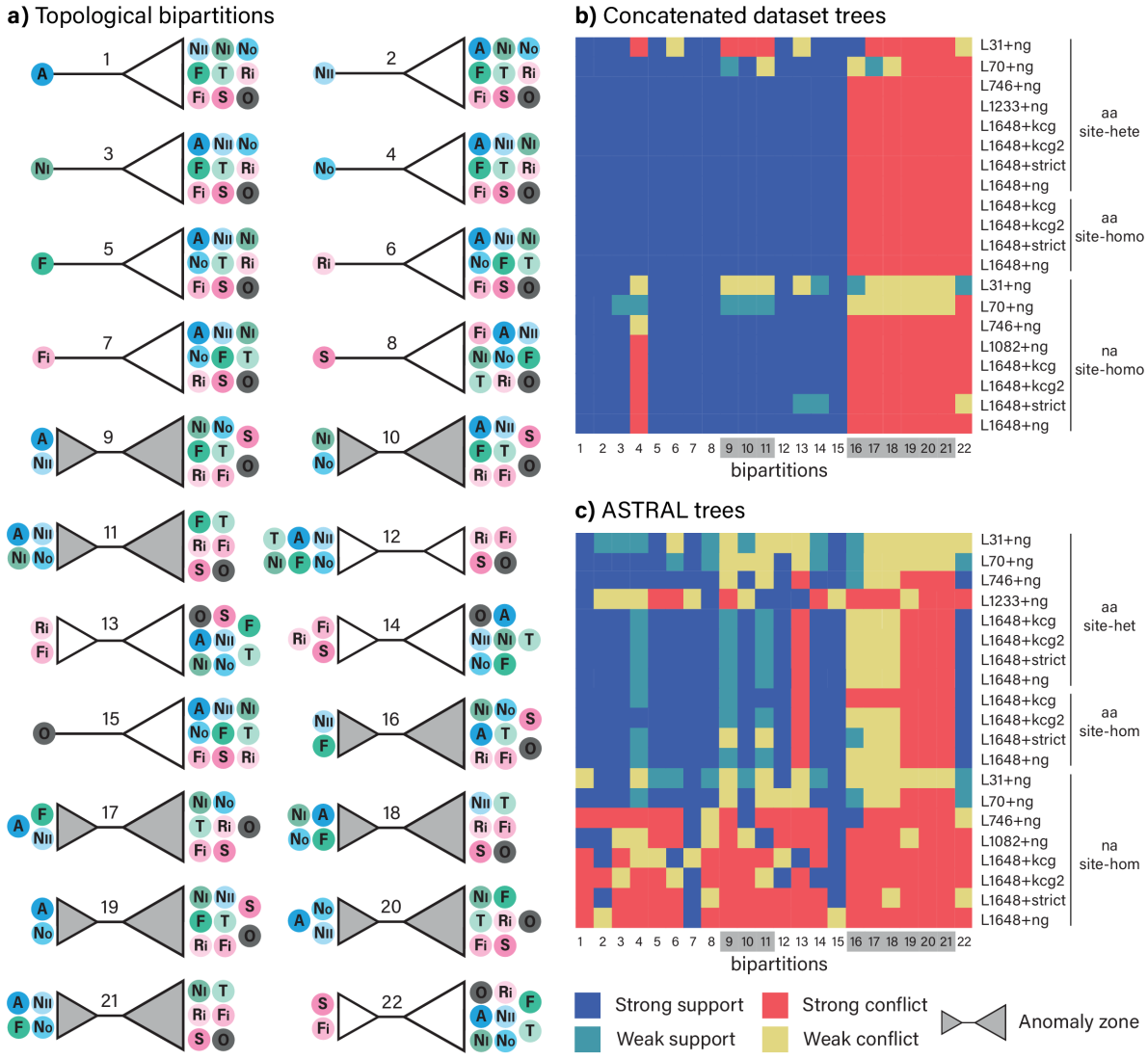

Figure 4

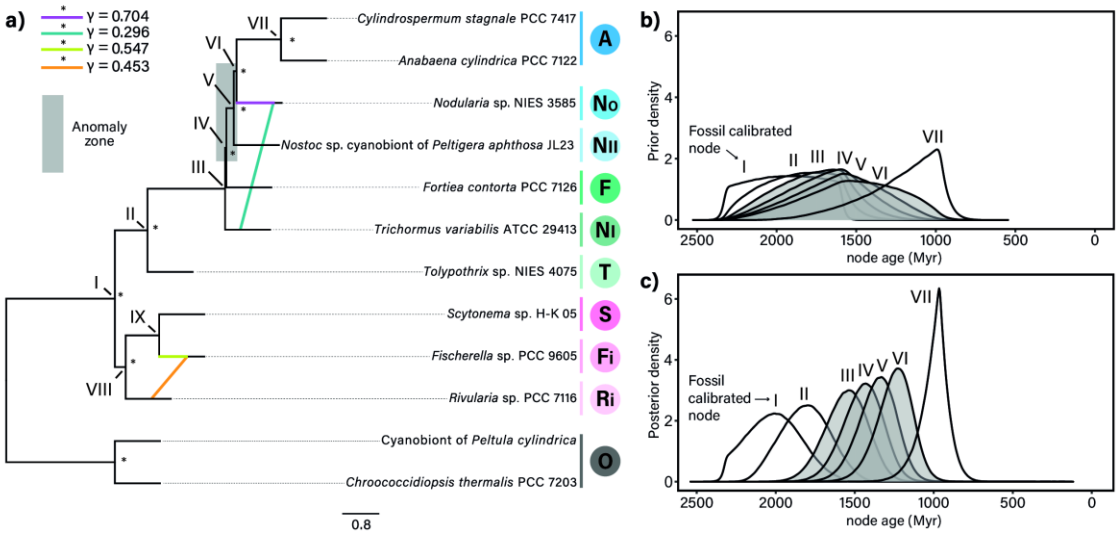

Figure 5