Molecular Phylogenetics and Evolution 162 (2021) 107100

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

# Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: A case study with Cyanobacteria and cyanolichens

Luc Cornet [a,1], Nicolas Magain [b,c,1], Denis Baurain [a,*], François Lutzoni [b]

[a] InBioS – PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
[b] Department of Biology, Duke University, Durham, NC, USA
[c] Evolution and Conservation Biology, InBioS, University of Liège, Liège, Belgium

ARTICLE INFO

*Keywords:*
Cyanobacteria
Lichen
Horizontal gene transfers
Phylogenomics

ABSTRACT

Understanding the evolutionary history of symbiotic Cyanobacteria at a fine scale is essential to unveil patterns of associations with their hosts and factors driving their spatiotemporal interactions. As for bacteria in general, Horizontal Gene Transfers (HGT) are expected to be rampant throughout their evolution, which justified the use of single-locus phylogenies in macroevolutionary studies of these photoautotrophic bacteria. Genomic approaches have greatly increased the amount of molecular data available, but the selection of orthologous, congruent genes that are more likely to reflect bacterial macroevolutionary histories remains problematic. In this study, we developed a synteny-based approach and searched for Collinear Orthologous Regions (COR), under the assumption that genes that are present in the same order and orientation across a wide monophyletic clade are less likely to have undergone HGT. We searched sixteen reference Nostocales genomes and identified 99 genes, part of 28 COR comprising three to eight genes each. We then developed a bioinformatic pipeline, designed to minimize inter-genome contamination and processed twelve *Nostoc*-associated lichen metagenomes. This reduced our original dataset to 90 genes representing 25 COR, which were used to infer phylogenetic relationships within Nostocales and among lichenized Cyanobacteria. This dataset was narrowed down further to 71 genes representing 22 COR by selecting only genes part of one (largest) operon per COR. We found a relatively high level of congruence among trees derived from the 90-gene dataset, but congruence was only slightly higher among genes within a COR compared to genes across COR. However, topological congruence was significantly higher among the 71 genes part of one operon per COR. Nostocales phylogenies resulting from concatenation and species tree approaches based on the 90- and 71-gene datasets were highly congruent, but the most highly supported result was obtained when using synteny, collinearity, and operon information (i.e., 71-gene dataset) as gene selection criteria, which outperformed larger datasets with more genes.

## 1. Introduction

Symbiotic interactions with other species play a key role in the evolution and ecology of most eukaryotes (e.g., Thompson, 1999; Guimarães et al., 2011). Lichens are symbiotic associations between two to three main partners: a fungus (mycobiont) and one or two photosynthetic partners (photobionts), which can be green algae and/or Cyanobacteria (Nash, 2008). These interactions are considered mutualistic, as both main partners are assumed to benefit from the association (Lutzoni and Miadlikowska, 2009; however see Hyvärinen et al., 2002). Only about 120 species of photobionts are described in comparison to ca. 15,000 lichen-forming fungal species (Honegger, 2012). The identity of the photobionts beyond the genus or family level has been generally overlooked (Friedl and Büdel, 2008). However, recent molecular systematic studies on specific lichen groups showed that photobiont biodiversity is much higher than previously expected, and that the number of photobiont species may be close to the number of mycobiont species for some lichen genera (Kroken and Taylor, 2000; Skaloud and Peksa, 2010; Magain et al., 2017a; Muggia et al., 2020).

Most phylogenetic studies of lichenized Cyanobacteria (cyanobionts)

* Corresponding author.
*E-mail address:* denis.baurain@uliege.be (D. Baurain).
[1] Contributed equally.

https://doi.org/10.1016/j.ympev.2021.107100
Received 28 August 2020; Received in revised form 22 January 2021; Accepted 1 February 2021
Available online 14 February 2021

are based on a single locus, such as *rbcLX* (Otálora et al., 2010; O'Brien et al., 2013; Magain et al., 2017a, 2018; Pardo-De la Hoz et al., 2018), SSU rRNA (16S) (Elvebakk et al., 2008; Lohtander et al., 2003; Magain and Sérusiaux, 2014), tRNA (UAA) (Paulsrud et al., 1998; Fedrowitz et al., 2011; Jüriado et al., 2019), or sometimes on a combination of these loci, or comparison of single-locus analyses using these loci (O'Brien et al., 2005; Myllys et al., 2007; Kaasalainen et al., 2015). Because single-locus phylogenies often lack resolution and statistical support, systematic studies of mycobionts and photobionts have been unbalanced. On the one hand, fungal species are usually characterized using multilocus datasets and several species delimitation methods. On the other hand, lichenized Cyanobacteria are defined as operational taxonomic units (OTUs) characterized by monophyly (e.g., phylogroups), often relying on poorly resolved phylogenetic trees, or based on sequence similarity alone (e.g.. Dal Forno et al., 2020; Magain et al., 2017a; O'Brien et al., 2013). This difference in our ability to recognize species among the main interacting partners inside lichen thalli leads to limitations in our understanding of eco-evolutionary patterns of associations, and their spatiotemporal dynamics. Moreover, relying on a single locus may not depict an accurate representation of the evolutionary history of these organisms (e.g., O'Brien et al., 2005; Kaasalainen et al., 2015). Phylogenies based on commonly used single markers are showing their limits as we continue to add more new taxa to continuously larger phylogenetic trees. Therefore, new approaches are urgently needed for the next generation of studies on symbiont interactions in lichen symbioses. The relatively small genome sizes of lichen partners (Armaleo and May, 2009; McDonald et al., 2013; Armaleo et al., 2019) allow the efficient sequencing of whole lichen metagenomes that are, in principle, well suited for multilocus studies. However, whereas genome-based phylogenetic studies have resolved problematic relationships in many groups of organisms (Delsuc et al., 2005; Xi et al., 2013; Zhang et al., 2020), such genome-wide datasets can include foreign DNA because of Horizontal Gene Transfer (HGT) and/or environmental contamination (Cornet et al., 2018a; Philippe et al., 2011, 2017).

HGT is common in prokaryotes (Doolittle, 1999) and causes phylogenetic incongruence (Hilario and Gogarten, 1993). It has been argued that the amount of genetic material transmitted horizontally can be so prevalent that the evolution of prokaryotic genomes might be better represented by networks than by phylogenetic trees (Boekels-Gogarten et al., 2009; Gribaldo and Brochier, 2009; but see also Abby et al., 2012). As a result, concerns were raised about the use of gene concatenation in phylogenetic studies of prokaryotes (Gribaldo and Brochier, 2009). Cyanobacteria genomes show occurrences of HGT events (Khan et al., 2016; Manen and Falquet, 2002; Popa et al., 2017; Shi and Falkowski, 2008; Tooming-Klunderud et al., 2013; Zhaxybayeva et al., 2006), especially in diazotrophic taxa (Shi and Falkowski, 2008). The second potential source of foreign DNA in alignments used for phylogenetic studies is contamination (Laurin-Lemay et al., 2012; Philippe et al., 2011, 2017). This phenomenon, where orthologous sequences from a distinct organism are incorrectly selected, has been reported for Cyanobacteria (Cornet et al., 2018a), notably because of complex trophic relationships between Cyanobacteria and other bacterial phyla (Lee et al., 2014; Stuart et al., 2016). In recent years, several methods (e.g., Snir and Rao, 2012; Mirarab and Warnow, 2015) have been developed to account for HGT and other sources of incongruence, such as incomplete lineage sorting, *after* assembling the dataset. However, minimizing incongruence *before* assembling the dataset, by adequately selecting loci, can only reinforce these methods, and is their logical complement.

For lichens, the long-standing paradigm of a single photobiont genotype in each individual bi-membered thallus was rejected by recent molecular studies. For example, the presence of multiple green algal photobionts inside individual thalli was shown for some lichens (Casano et al., 2011; Dal Grande et al., 2017; Onuț-Brännström et al., 2018). Consequently, metagenomics, where the DNA of all organisms present in a sample is sequenced, appears to be an ideal tool to study a morphologically defined, easy to collect symbiotic unit such as the lichen thallus. However, the presence of multiple fungi (lichen mycobiota) and prokaryotes (lichen microbiota) inside lichen thalli (Arnold et al., 2009; U'Ren et al., 2010; Hodkinson et al., 2012; Aschenbrenner et al., 2016), sometimes in addition to multiple photobionts, makes contamination a more acute issue for metagenomic data obtained from individual lichen thalli.

In this context, it is crucial to develop reliable bioinformatic pipelines to carefully handle data gathering and processing for phylogenetic studies. Here we present an approach to perform phylogenetic inference on lichenized Cyanobacteria from metagenomic data. One focus of this study was on twelve metagenomes from lichen-forming fungi (Peltigerales, Lecanoromycetes, Ascomycota) associated with Cyanobacteria of the genus *Nostoc* (Nostocales). We first used a pipeline to process and assemble those 12 metagenomes. We then successfully assigned metagenomic contigs to genomic bins corresponding to each of the organisms present, so as to avoid contamination that could result from having a mixture of sequences from different organisms co-living in lichen thalli.

We assume that HGT events are less likely to have occurred in genomic regions where the order (synteny) and orientation (collinearity) of genes are conserved across a wide range of organisms, because it would imply that the horizontally-transferred gene has not only been inserted next to its ortholog but also that the ortholog has been deleted (see Rolland et al., 2009, for a similar rationale). A recent study, using a similar approach, identified 125 Collinear Orthologous Regions (COR) spanning two classes of fungi (Lecanoromycetes and Eurotiomycetes) and demonstrated the potential of such regions to resolve relationships in fungal species complexes (Magain et al., 2017b).

While tools are available to generate syntenic datasets for phylogenetic analyses (e.g., GET_HOMOLOGUES, Contreras-Moreira and Vinuesa, 2013; SynChro, Drillon et al., 2014; HomBlocks, Bi et al., 2018), no software seems available to take advantage of synteny within metagenomes. Here we developed a bioinformatic pipeline to ensure the orthology, synteny and collinearity of metagenomic sequences added to a syntenic dataset (Fig. 1). It includes a new program (hereafter called MESYRES; MEtagenomic SYntenic REgionS), which starts from the output of an orthogroup (OG) inference program, such as OrthoFinder (Emms and Kelly, 2015), which is a basic step needed for any phylogenomic analysis (Tekaia, 2016). MESYRES is designed to create both the initial dataset of COR from reference genomes and to enrich such a syntenic dataset by searching multiple sets of metagenomic contigs corresponding to distinct samples (here, individual lichen thalli). We used the resulting COR generated with MESYRES, when applied on cyanobacterial genomic and cyanolichen metagenomic data, to infer the evolutionary history of lichenized Cyanobacteria in the phylogenetic context of the order Nostocales. We analyzed the selected genes and COR individually (including a set of analyses restricted to genes part of the largest operon within each COR), as well as part of concatenated datasets, for phylogenetic resolution, support and congruence, and compared them to results obtained with genes that have been widely used in phylogenetic studies of Nostocales (such as SSU rRNA [16S] and *rbcLX*).

## 2. Materials and methods

### 2.1. Phylogenetic analyses of reference genomes

For this study, we selected 16 reference Nostocales strains (Table 1) with a maximal number of four scaffolds in their genome assembly and a low level of contamination (<1.08% as estimated by DIAMOND blastx [Buchfink et al., 2015]) according to the method developed in Cornet et al. (2018a). We chose OrthoFinder v1.1.2 (Emms and Kelly, 2015) using the standard inflation parameter (1.5) and USEARCH v8.1 (64 bits) (Edgar, 2010) to define the orthologous groups (OGs) from these 16 strains. Of the 27,061 OGs, 1160 genes were selected with classify-ali.pl (part of the Bio-MUST-Core software package; D. Baurain; https://metac

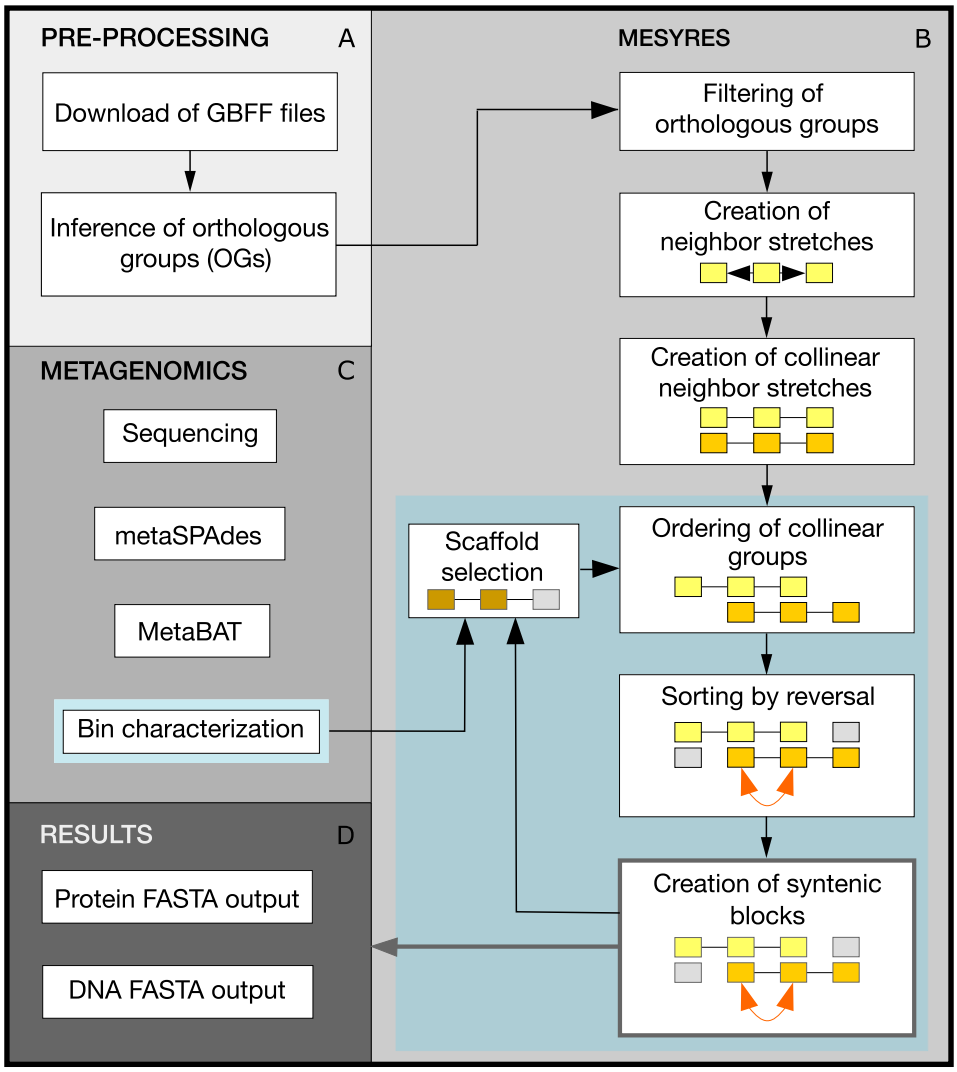


**Fig. 1.** MESYRES pipeline. A. The pipeline begins with the pre-processing step where orthologous groups are inferred. B. The six main boxes from the MESYRES core represent the six main steps of the algorithm as described in the Materials and methods section. These six steps were first run on the 16 reference genomes, which resulted in the detection of 99 syntenic genes part of 28 COR. The three last steps of MESYRES are iterative, and the sorting by reversal was performed in two directions. C. Metagenomic pipeline leading to the characterization of 12 main cyanobacterial bins. The 99 syntenic genes were then enriched with orthologous sequences from 12 cyanobacterial metagenomic bins (light blue parts of the flow chart). The scaffold selection, based on the frequency of presence of each metagenomic scaffold over a whole collinear group, was carried out before running again the three last steps of MESYRES. D. MESYRES produces both DNA and protein FASTA files, for each of the orthologous groups retained by the pipeline. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pan.org/release/Bio-MUST-Core) by enforcing in each OG the presence of the 16 reference species in ≤1.1 gene copy per organism on average. OGs were then aligned with MAFFT v7.273 (Katoh and Standley, 2013) and conserved sites were selected with BMGE v1.12 (Criscuolo and Gribaldo, 2010) using moderately severe settings (entropy cut-off 0.5, gap cut-off 0.2). A supermatrix of 16 organisms × 21,281 unambiguously aligned amino-acid positions (AA) was produced with SCaFoS v1.30k (Roure et al., 2007). Finally, a phylogenomic tree was inferred with RAxML v8.2.9 under the PROTGAMMALGF model (Stamatakis, 2006).

The 27,061 OGs were also used to produce 100 datasets with the program jack-ali-dir.pl (also part of the Bio-MUST-Core software package) by enforcing the supermatrix size to ca. 21,200 AA. For each jackknife replicate, the genes were aligned with MAFFT, conserved sites were selected with BMGE and a supermatrix was assembled with SCaFoS. Then, 100 phylogenomic trees were inferred using RAxML, as described above.

### 2.2. Taxon sampling and molecular data acquisition

Total genomic DNA was extracted from six lichen thalli (NMS1, NMS2, NMS4, NMS7, NMS8 and NMS9, Table 2) following a protocol modified from Zolan and Pukkila (1986), using 2% sodium dodecyl sulphate (SDS) as the extraction buffer. DNA libraries (DNA-Seq, 300 bp insert) and sequencing of the metagenomes (Illumina HiSeq 4000 150-bp PE) was performed by the Sequencing and Genomic Technologies Shared Resource, Duke Center for Genomic and Computational Biology, Duke University (Durham, North Carolina, USA). Genomic DNA was also extracted from a culture of *Nostoc* (NOS) isolated from cephalodia of *Peltigera aphthosa*, using a CTAB-based extraction protocol (Cubero et al., 1999) and sequenced on the Illumina MiSeq sequencing platform (GIGA Genomics, University of Liège). Nextera XT libraries had an insert size estimated at 800–900 bp. Metagenomes from four additional *Peltigera* thalli (JL23, JL31, JL33, JL34, Logsdon et al., unpublished) and from *Leptogium austroamericanum* (McDonald et al., 2013) were added to our dataset, resulting in a total of twelve metagenomes (11 lichen metagenomes and one *Nostoc* genome, Table 2).

It is not possible to determine with certainty the identity of *Nostoc* symbionts before sequencing. However, we expected the sequenced Cyanobacteria to represent a broad phylogenetic diversity of symbionts within *Nostoc* clade II (sensu Magain et al., 2017a). This is because metagenomes were selected to cover the fungal phylogenetic diversity of the genus *Peltigera,* i.e., representing seven of eight sections of this genus (sensu Miadlikowska and Lutzoni, 2000), as well as the sister genus *Solorina* (Peltigeraceae, Peltigerales), and *Leptogium austroamericanum*, which belongs to the family Collemataceae (Peltigerales).

**Table 1**
Reference genomes used to generate the 99-gene 28-COR dataset.

| Strain | GCF Ref. No. | Reference | GenBank No. | JGI project ID |
|---|---|---|---|---|
| *Anabaena* sp. PCC 7108 | GCF_000332135 | Shih et al., 2013 | — | 1077465 |
| *Anabaena* sp. WA102 | GCF_001277295 | Brown et al., 2016 | CP011456 | Go0118624 |
| *Calothrix* sp. 336/3 | GCF_000734895 | Isojärvi et al., 2015 | CP011382 | Ga0111347 |
| *Calothrix* sp. PCC 6303 | GCF_000317435 | Shih et al., 2013 | CP003610 | 1077462 |
| *Calothrix* sp. PCC 7507 | GCF_000316575 | Shih et al., 2013 | CP003943 | 1077456 |
| *Cylindrospermum stagnale* PCC 7417 | GCF_000317535 | Shih et al., 2013 | CP003642 | 1077801 |
| *Fischerella* sp. NIES-3754 | GCF_001548455 | Hirose et al., 2016a | AP017305 | Ga0123612 |
| *Fortiea contorta* PCC 7126 (*Microchaete* sp. PCC 7126) | GCF_000332295 | Shih et al., 2013, Hauer et al., 2014 | — | 1077460 |
| *Mastigocladopsis repens* PCC 10914 | GCF_000315565 | Shih et al., 2013 | — | 1079022 |
| *Nodularia spumigena* CCY9414 | GCA_000340565 | Voß et al., 2013 | CP007203 | Ga0067561 |
| *Nostoc azollae* 0708 | GCF_000196515 | Ran et al., 2010 | CP002059 | 4084143 |
| *Nostoc piscinale* CENA21 | GCF_001298445 | Leão et al., 2016 | CP012036 | Ga0100955 |
| *Nostoc* sp. NIES-3756 | GCF_001548375 | Hirose et al., 2016b | AP017295 | Ga0123611 |
| *Nostoc* sp. PCC 7107 | GCF_000316625 | Shih et al., 2013 | CP003548 | 1077464 |
| *Nostoc* sp. PCC 7524 | GCF_000316645 | Shih et al., 2013 | CP003552 | 1077798 |
| *Rivularia* sp. PCC 7116 | GCF_000316665 | Shih et al., 2013 | CP003549 | 1077466 |

**Table 2**
Mycobiont and photobiont sequenced, as well as voucher information, for 11 lichen metagenomes used in this study and one *Nostoc* strain (see asterisk) isolated in culture from *Peltigera aphthosa* (ID: NOS). For fungal species associated with two photobionts (one green alga [*Coccomyxa*] and one cyanobacterium [*Nostoc*], i.e., trimembered lichens), both photobionts are listed.

| ID | Mycobiont | Photobiont | Voucher information |
|---|---|---|---|
| JL23 | *Peltigera aphthosa* | *Coccomyxa* sp. and *Nostoc* sp. | Canada, Alberta, Miadlikowska & Lutzoni s. n., DUKE |
| JL31 | *Peltigera extenuata* | *Nostoc* sp. | Canada, Alberta, Miadlikowska & Lutzoni s. n., DUKE |
| JL33 | *Peltigera malacea* | *Nostoc* sp. | Canada, Alberta, Miadlikowska & Lutzoni s. n., DUKE |
| JL34 | *Peltigera evansiana* | *Nostoc* sp. | USA, Michigan, Miadlikowska & Lutzoni s. n., DUKE |
| LPT | *Leptogium austroamericanum* | *Nostoc* sp. | USA, North Carolina, T. McDonald s.n., DUKE |
| NMS1 | *Peltigera hydrothyria* | *Nostoc* sp. | Canada, Nova Scotia, F. Anderson 16529, NSPM |
| NMS2 | *Peltigera hydrothyria* | *Nostoc* sp. | Canada, Nova Scotia, F. Anderson 16530, NSPM |
| NMS4 | *Peltigera venosa* | *Coccomyxa* sp. and *Nostoc* sp. | Finland, A. Simon 80, LG. |
| NMS7 | *Solorina crocea* | *Coccomyxa* sp. and *Nostoc* sp. | Finland, A. Simon 79, LG. |
| NMS8 | *Peltigera dolichorhiza* | *Nostoc* sp. | Panama, N. Magain P6117, DUKE |
| NMS9 | *Peltigera phyllidiosa* | *Nostoc* sp. | USA, North Carolina, N. Magain P3103, DUKE |
| NOS | *Peltigera aphthosa* | *Coccomyxa* sp. and *Nostoc* sp.* | Sweden, R. Darnajoux s.n., DUKE |

### 2.3. Genome assembly

#### 2.3.1. Metagenome assembly and genome binning

Reads from the 12 samples (Table 2) were trimmed with Trimmomatic v0.35 (Bolger et al., 2014). Sequencing adapters were removed using the option illuminaclip TruSeq3-PE.fa:2:30:20. The leading/trailing values were set at 20, the sliding window at 10:20, the crop value at 145 and the minimal length at 100 (with the exception of the *Nostoc* sp. in culture (NOS) for which we used a minimal length of 80). Trimmed paired-end reads were assembled with the metagenomic version of SPAdes, metaSPAdes v3.10.1 (Nurk et al., 2017) and then mapped back on the assemblies using BamM v1.7.3 (http://ecogenomics.github.io/BamM/). Genome bins were determined with MetaBAT v0.30.1, using the option resulting in the best specificity (i.e., super-specific) (Kang et al., 2015).

#### 2.3.2. Taxonomic analysis of genome bins

Cyanobacterial bins were identified through DIAMOND blastx analyses against a curated database of 272 complete proteomes including representative genomes from the three domains of Life (M. Van Vlierberghe, University of Liège, available at https://doi.org/10.6084/m9.figshare.13550267.v2). To increase DIAMOND blastx sensitivity, genome bins were split into non-overlapping pseudo-reads of 250 nt in length with split-ali.pl (https://doi.org/10.6084/m9.figshare.5006039.v3) as in Cornet et al. (2018a). A Last Common Ancestor (LCA) algorithm post-processor developed for DIAMOND blastx (Cornet et al., 2018a) was then used to taxonomically classify the pseudo-reads from each genome bin. Overall, for the twelve metagenomes, we identified 249 bins containing ≥0.1% of cyanobacterial sequences (1 for NOS, 6 for LPT, 8 for NMS1, 13 for NMS8, 14 for NMS9, 16 for JL23, 17 for NMS4, 17 for NMS7, 24 for NMS2, 29 for JL33, 49 for JL34, 55 for JL31).

### 2.4. Synteny analyses

#### 2.4.1. MESYRES generated syntenic and collinear dataset

We developed the Perl program MESYRES (MEtagenomic SYntenic REgionS) to recover syntenic and collinear regions among a collection of orthologous groups (OGs), i.e., determined by orthogroup inference programs (Tekaia, 2016). These OGs may (or may not) be part of metagenome bins. MESYRES is preceded by a pre-processor (Fig. 1), GBK-reader, which has a dual purpose: (1) to create specially-formatted FASTA files from GenBank GBFF files for the orthogroup inference program, and (2) to collect gene metadata (such as gene ID, position, strand, product) and reformat the definition lines of the output of the orthogroup inference program for MESYRES.

MESYRES is a six-step greedy algorithm (see Fig. 1). The first step is the filtering of OGs, produced with 16 reference genomes (see phylogenetic analyses of reference species). First, the OGs found in <16 genomes are discarded from the analysis. To explain the next step of the filtration process, we introduce the concept of neighborhood. Two proteins are neighbors if their genes are next to each other on the chromosome. To pass this first step, an OG (composed of orthologous proteins) needs to have one or more neighbor OGs. It means that a user specified fraction (by default 100%) of its component proteins must have one or more neighbor proteins among the component proteins of at least one other OG. This first step generated a dataset composed of 378 neighbor genes.

The second step is the creation of neighbor stretches. From the remaining OGs, we collect the protein identification numbers (IDs).

Then, we search for the longest possible stretches of neighbors. In practice, we start with a triplet of neighbor proteins that serves as a seed for extension and we search for different stretches of neighbors. We then iteratively merge them into longer stretches if they share common neighbors at their extremities. During this step, paralogous proteins that can be present in OGs are discarded. Indeed, paralogous sequences often affect specific genomes and have thus little chance to have neighbors among the remaining OGs. These sequences without neighbors are thus discarded from the analysis. The duplicated sequences that co-localize (e.g., in tandem) on a genome are conserved in the analysis. In practice, only one of such duplicate sequences can be used in the subsequent steps of MESYRES but they are both conserved in its final output.

The third step is the creation of collinear neighbor stretches. We define here the concept of shared groups: two neighbor stretches from two different genomes are shared if they have at least one OG in common. Through the different genomes, in a greedy way, we merge the different stretches if they share common OGs, so as to create shared groups.

The fourth step is the ordering of the shared groups. The objective of this step is to create an ordered suite of the OGs composing the shared group. To that end, we select the longest neighbor stretch of the shared group and order the OGs according to the gene order of this "master" stretch, based on the gene order in the corresponding genome. If necessary, we continue the ordering, by changing the "master" stretch, until every OG in the shared group is ordered.

The fifth step is the sorting by reversal. For each genome involved in the ordered shared group, we compute the rearrangement events required and additionally count them by OG. This means that, at the end of this step, the ordered suite of OGs is considered as a consensus chain that summarizes all the events of all genomes. The possible events are of three different types: (1) an inversion, when a given genome has two (or more) neighbor genes that do not co-localize in successive OGs; (2) a missing gene, when a given genome has no protein in an OG; and (3) an out-of-order event, which is linked to the orthogroup inference process. This occurs when a gene is present in a given genome but has no neighbor in the shared group.

The sixth step is the creation of syntenic blocks (Syntenic Orthologous Region = SOR). We define two different thresholds, one for the missing/out-of-order events and one for the inversion events. The OG consensus chain is read and OGs added to a syntenic block until reaching one of the thresholds. When any threshold is reached, the current block is closed, the thresholds are reset and a new syntenic block is opened on the next OG. In practice, inversion events in a genome circumscribe windows delimited by distal OGs involved in the inversion. All inversion windows (through all genomes) must be closed to add an OG to a syntenic group.

To increase the number of syntenic OGs composing the SOR, we implement different iterative processes. First, the two last steps are performed in two directions (from left-to-right and right-to-left) and the direction yielding the largest number of syntenic OGs is selected. In the same way, all the steps from the fourth step (i.e., ordering of collinear groups) are iterative. The first iteration starts with the longest neighbor stretch, while the next ones each select a shorter stretch as their master stretch. We perform a maximum of three iterations, and the iteration yielding the largest number of syntenic OGs is retained. Both MESYRES and its GBK-reader are available at https://bitbucket.org/phylogeno/MESYRES.git. By using 0 inversion and 0 missing/out-of-order thresholds, we found 99 syntenic OGs corresponding to 28 SOR (Supplementary Table S1). The directionality of the genes composing the 28 SOR was verified for all reference genomes using the graphical overview provided by the NCBI portal to GenBank (nt) (Table 3; see https://www.ncbi.nlm.nih.gov/nuccore/NZ_KB235930.1?report=graph for an example). Because the relative orientation (collinearity) of genes was conserved across reference genomes, all 28 SOR are considered to be Collinear Orthologous Regions (COR). The operons of the 28 COR were identified with Operon-mapper (Taboada et al., 2018) using *Cylindrospermum stagnale* PCC 7417 as a reference (Table 3).

### 2.4.2. Enrichment of the COR dataset

This step corresponds to the addition of metagenomic bins on the three last steps of MESYRES (Fig. 1B, C). The 99 syntenic and collinear OGs, corresponding to 28 COR, are enriched with orthologous sequences from metagenomic bins in order to be reused in MESYRES. We used a pre-release version of our software Forty-Two (Simion et al., 2017; Irisarri et al., 2017) (now available at https://metacpan.org/release/Bio-MUST-Apps-FortyTwo) to enrich these 99 OGs with orthologous sequences from our twelve main cyanobacterial bins (MCBs, see below Metagenome completeness). We then re-introduced these OGs at step four of MESYRES (ordering of collinear groups, Fig. 1B) and computed the synteny taking into account metagenomic sequences. Unfortunately, due to insufficient scaffolding of the metagenomes, the number of syntenic OGs dropped to 43 (data not shown). Consequently, we decided to disable the sorting by reversal step and instead select metagenomic sequences based on scaffold completeness in collinear groups. Therefore, before step 4 of MESYRES, we computed the frequency of presence of each metagenomic scaffold over a whole collinear group (i.e., before the creation of COR). We used this metric as a threshold, in the scaffold selection step (see Fig. 1), for either retaining or rejecting metagenomic sequences, with all sequences of a scaffold being rejected in case of insufficient presence. Supplementary Table S1 shows the effect of different thresholds (from 0 to 80%) on COR completeness. As all MCBs showed a small amount of contaminants, an advantage of this scaffold selection step is to allow the filtering of metagenomic bins contaminations, while assuming that duplicated sequences are contaminants. Indeed, after the raw enrichment by Forty-Two (T0 of Supplementary Table S1), all the metagenomes showed $\geq$1 sequences for some OGs (2.35%). By increasing our threshold, we reduced the amount of foreign sequences by OG (0.93% at T20, 0.001% at T40, 0% at T60 and T80), so as to extract more single-copy proteins from our metagenomes. Unfortunately, this procedure had the disadvantage to also increase the quantity of missing sequences in the OGs (0.75% at T0, 6.82% at T20, 10.86% at T40, 20.96% at T60, 23.65% at T80). That is why we decided to use a moderate threshold of 40% for this study. Finally, the 99 syntenic OGs composed of aligned protein sequences were back-translated into aligned DNA sequences using a pre-release version of our software Leel (Rodríguez et al., 2017), which is now distributed as part of the Forty-Two software suite (see above). Both protein and nucleotide multiple sequence alignment files are available at https://doi.org/10.6084/m9.figshare.12093882.v2.

### 2.4.3. Metagenome completeness

We identified, through DIAMOND blastx analysis (see 2.3.2 Taxonomic analysis of genome bins), a main cyanobacterial bin (MCB) for each metagenome (at the exception of NMS4, which had two such bins) composed of $\geq$60% of cyanobacterial sequences (bin 4 for NMS1, bin 6 for NMS2, bins 6 and 15 for NMS4, bin 6 for NMS7, bin 5 for NMS8, bin 6 for NMS9, bin 7 for JL31, bin 9 for JL33, bin 6 for JL34, bin 6 for JL23, bin 3 for LPT, bin 1 for NOS) (Supplementary Table S1). To evaluate their completeness, we analyzed the 249 bins containing cyanobacterial sequences with CheckM v1.0.7, a program that uses lineage-specific microbial marker genes (Parks et al., 2015). MCBs appeared to have a high completeness level: $\geq$98.89% for the majority of the bins (96.67% for NOS, which had only one predicted bin). Yet, CheckM failed to evaluate the completeness of two MCBs (bin 37 for JL31, bin 6 for JL34). Interestingly, we identified that for the two largest cyanobacterial bins of NMS4 (bin 6 and bin 15), CheckM returned low and "complementary", levels of completeness (86.15% for bin 6 and 15.52% for bin 15).

Then we used the 99 syntenic OGs of the *Nostoc* group to evaluate the completeness of our MCBs (Supplementary Table S1). We used Forty-Two to enrich the OGs with orthologous sequences from the 249 bins containing cyanobacterial sequences. It appears from this analysis that the two largest cyanobacterial bins of NMS4 are indeed complementary,

as suggested by CheckM values. Hence, when a protein was missing in some OG for the large bin 6, it was found in the small bin 15 and vice versa (15 cases). Moreover, CheckM indicated that these two bins are both Nostocales. Consequently, we decided to merge these two bins into a single MCB (bin6-15), which resulted in a final completeness of 96.04%. We did not identify such complementarity patterns for other bins, showing that the highly specific option used for MetaBAT (see 2.3.1 Metagenome assembly and genome binning) was well suited to our dataset.

### 2.5. Phylogenetic analyses based on 28 taxa

#### 2.5.1. Dataset assembly and phylogenetic inference

At the T40 threshold, nine OGs were missing from all twelve lichenized Cyanobacteria (Supplementary Table S1, Table 3). Therefore, we assembled DNA and protein alignments for the 90 remaining OGs, for the 16 reference taxa and the twelve lichenized Cyanobacteria (28 taxa in total, Tables 1 and 2).

Sequences were aligned using MAFFT v7.305b (Katoh and Standley, 2013) with default parameters. Alignments were then carefully checked manually and slightly adjusted. Ambiguous regions were manually identified and excluded from the datasets. All single-gene DNA alignments were partitioned following their codon positions, whereas the best models of evolution for protein alignments were determined using ProtTest v3.4.2 (Darriba et al., 2011, Supplementary Table S2).

Both DNA and protein single-gene phylogenies were generated using RAxML v8.2.9 (Stamatakis, 2006). Optimal tree and bootstrap searches were conducted with the rapid hill-climbing algorithm for 1000 replicates under the GTR + $\Gamma_4$ model (Rodriguez et al., 1990). Because the degree of variation and phylogenetic resolution inside the clade of interest (lichenized Cyanobacteria) was more adequate in DNA sequences, subsequent analyses were performed on the DNA alignments only.

We assembled single genes into concatenated datasets using SCaFoS (Roure et al., 2007) where each concatenated dataset represented one COR, for a total of 90 genes concatenated into 25 COR. Fifteen COR consisted of three genes, eight COR consisted of four genes, one COR consisted of five genes, and one COR consisted of eight genes (Table 3). For all these COR datasets, the optimal partitioning scheme was determined using PartitionFinder2 v2.1.1 (Lanfear et al., 2017), as implemented on the CIPRES portal (Miller et al., 2010) with the tested subsets corresponding to each codon position of each gene, using the greedy algorithm (Lanfear et al., 2012) and the Bayesian Information Criterion (BIC). COR phylogenies were generated using RAxML v8.2.10 (Stamatakis et al., 2008) as implemented on the CIPRES portal, using the same parameters as for the single-gene analyses.

A concatenated dataset containing all 90 genes was also assembled. The best partitioning scheme was determined using PartitionFinder2 with the same parameters as for the COR analyses, and with the optimal partitioning schemes of each COR as the subsets to test. The concatenated 90-gene phylogeny was generated using RAxML on the CIPRES portal, with the same parameters as described above. A subset of this dataset containing the 71 genes belonging to the same operon within each COR (see Table 3) was assembled, and a concatenated 71-gene phylogeny was generated using RAxML on the CIPRES portal, with the same parameters as described above.

Finally, alignments of two commonly-used loci, SSU rRNA (16S) and *rbcLX*, consisting of the genes *rbcL*, *rbcX*, and a spacer, were retrieved from the reference genomes and genome bins, and assembled for the 28 taxa. For *rbcLX*, the spacer was excluded from phylogenetic analyses and a partition was assembled following the codon positions of *rbcL* and *rbcX*. Phylogenies were generated for these two genes using RAxML with the same parameters as above.

#### 2.5.2. Test of congruence between genes

We measured the congruence between phylogenetic trees using the Robinson-Foulds distances (Robinson and Foulds, 1981), as implemented in PAUP v4.0a (Swofford, 2003). Distances were computed on the best trees resulting from the ML searches, and on the same trees, after nodes with bootstrap values <70% were collapsed using TreeGraph v2.14 (Stöver and Müller, 2010). We compared the congruence of the 90 single-gene trees among each other and against the topologies from the 90- and 71-gene concatenated datasets, and of the 25- and 22-COR trees among each other and against the topologies from the 90- and 71-gene concatenated datasets. We also compared the SSU rRNA (16S) and *rbcLX* topologies against the topologies from the 90- and 71-gene concatenated datasets. A majority-rule consensus tree was generated using PAUP for the COR topologies. COR 3–634 tree was excluded from the analysis because it only includes 23 of the 28 taxa, and the consensus tree was thus generated using the 24 remaining COR trees.

#### 2.5.3. Species tree estimation

We estimated a species tree using MP-EST v1.6 (Liu et al., 2010), with five independent runs, and using the 90 single-genes ML best trees from the RAxML analyses on DNA datasets as input. We also estimated a species tree using the same approach with the 25 COR ML best trees as input. We repeated these analyses for the 71 genes belonging to the same operon within each COR (see Table 3), and their 22 COR containing only genes from the same operon within each COR.

## 3. Results and discussion

### 3.1. Data acquisition and dataset assembly

From an initial set of 16 publicly available genomes of reference cyanobacterial strains classified in the Nostocales (Table 1), and by imposing very strict thresholds (see Materials and methods), we obtained 99 syntenic genes belonging to 28 COR. In a second step, we enriched these COR with sequences from cyanobacterial bins derived from 11 lichen metagenomes and from sequences from one genome of a *Nostoc* strain in culture (Table 2). To minimize intergenome contamination and maximize COR completeness, we tested increasing threshold values for MESYRES (see Materials and methods, Supplementary Table S1). At the selected (T40) threshold, nine of the syntenic genes were absent from all lichenized Cyanobacteria (Table 3). Further analyses were thus conducted on a dataset composed of the remaining 90 syntenic and collinear genes representing 25 COR (Table 3). Our selection of COR imposed the gene orientation to be conserved across genomes, but not within a COR. For 21 of our 25 COR, all genes were in the same

**Table 3**
Operon composition and orientation of transcription for 99 genes part of 28 COR across 16 reference genomes. Horizontal lines delimit the 28 COR and show the gene composition of each COR. The numbers in bold in the COR group column represent genes in different operons within each COR (i.e., COR groups with bold numbers include more than one operon). The numbering system of the operon column correspond to the Operon-mapper numbering for *Cylindrospermum stagnale* and the number in parentheses correspond to the number of genes in the operon. The orientation of genes within genomes is based on the genome of *Cylindrospermum stagnale,* however, the orientation of genes within a COR is conserved across all genomes included in this study. The four genes in opposite orientation inside a COR (bidirectional COR) are highlighted in orange. The nine genes with asterisks are part of the 99-gene dataset, but not part of the 90-gene dataset. Genes highlighted in green are part of the 71-gene dataset, i.e., including only genes belonging to the same operon within a COR. (For interpretation of the references to color in this table legend, the reader is referred to the web version of this article.)

(*continued on next page*)

**Table 3** (*continued*)

| Orthologous group | Gene | COR group | Operon | Orientation |
|---|---|---|---|---|
| OG0001687 | hypothetical protein | 2-116-1 | 3890(4) | >>> |
| OG0001688 | nitrogenase-stabilizing/protective protein nifW | 2-116-1 | 3890(4) | >>> |
| OG0001689 | protein hesA | 2-116-1 | 3890(4) | >>> |
| OG0001763 | peptide chain release factor 1 | 26-4970-1 | 2191(30) | <<< |
| OG0001764 | 50S ribosomal protein L31 | 26-4970-1 | 2191(30) | <<< |
| OG0001765 | 30S ribosomal protein S9 | 26-4970-1 | 2191(30) | <<< |
| OG0001766 | 50S ribosomal protein L13 | 26-4970-1 | 2191(30) | <<< |
| OG0001768 | DNA-directed RNA polymerase subunit alpha | 26-4970-2 | 2191(30) | <<< |
| OG0001769 | 30S ribosomal protein S11 | 26-4970-2 | 2191(30) | <<< |
| OG0001770 | 30S ribosomal protein S13 | 26-4970-2 | 2191(30) | <<< |
| OG0001772 | adenylate kinase | 26-4970-3 | 2191(30) | <<< |
| OG0001773 | preprotein translocase subunit SecY | 26-4970-3 | 2191(30) | <<< |
| OG0001774 | 50S ribosomal protein L15 | 26-4970-3 | 2191(30) | <<< |
| OG0001775 | 30S ribosomal protein S5 | 26-4970-3 | 2191(30) | <<< |
| OG0001776 | 50S ribosomal protein L18 | 26-4970-3 | 2191(30) | <<< |
| OG0001777 | 50S ribosomal protein L6 | 26-4970-3 | 2191(30) | <<< |
| OG0001778 | 30S ribosomal protein S8 | 26-4970-3 | 2191(30) | <<< |
| OG0001779 | 50S ribosomal protein L5 | 26-4970-3 | 2191(30) | <<< |
| OG0001783 | 30S ribosomal protein S3 | 26-4970-4 | 2191(30) | <<< |
| OG0001784 | 50S ribosomal protein L22 | 26-4970-4 | 2191(30) | <<< |
| OG0001785 | 30S ribosomal protein S19 | 26-4970-4 | 2191(30) | <<< |
| OG0001786 | LSU ribosomal protein L2p (L8e) | 26-4970-4 | 2191(30) | <<< |
| OG0000640 | cysteine desulfurase | 3-170-1 | 4294(4) | >>> |
| OG0000641 | Fe-S cluster assembly protein SufD | 3-170-1 | 4294(4) | >>> |
| OG0000642 | ABC transporter ATP-binding protein | 3-170-1 | 4294(4) | >>> |
| OG0001511 | [protein-PII] uridylyltransferase | 3-3418-1 | 892(4) | >>> |
| OG0001512 | YggS family pyridoxal phosphate enzyme | 3-3418-1 | 892(4) | >>> |
| OG0001513 | cell division protein SepF | 3-3418-1 | 892(4) | >>> |
| OG0001575 | uracil phosphoribosyltransferase | **3-3594-1** | 2094(1) | <<< |
| OG0001576 | membrane protein | **3-3594-1** | 2093(1) | <<< |
| OG0001577 | Cell division protein YlmG/Ycf19 (putative) | **3-3594-1** | 2092(1) | <<< |
| OG0001629 | NAD(+) kinase | 3-3706-1 | 2970(4) | >>> |
| OG0001630 | NADH-quinone oxidoreductase subunit K | 3-3706-1 | 2970(4) | >>> |
| OG0001631 | ubiquinone oxidoreductase subunit J | 3-3706-1 | 2970(4) | >>> |
| OG0000745 | cytochrome b6 | 3-410-1 | 4327(3) | <<< |
| OG0000746 | cytochrome C oxidase subunit I | 3-410-1 | 4327(3) | <<< |
| OG0000747 | cytochrome c oxidase subunit II | 3-410-1 | 4327(3) | <<< |
| OG0001735 | hypothetical protein | **3-4474-1** | 4337(1) | <<< |
| OG0001736 | crossover junction endodeoxyribonuclease RuvA | 3-4474-1 | 4338(2) | <<< |
| OG0001737 | chemotaxis protein CheY | 3-4474-1 | 4338(2) | <<< |
| OG0000757 | response regulator | 3-490-1 | 3975(7) | <<< |
| OG0000758 | chemotaxis protein | 3-490-1 | 3975(7) | <<< |
| OG0000759 | methyl-accepting chemotaxis sensory transducer | 3-490-1 | 3975(7) | <<< |
| OG0000823 | potassium transporter | 3-634-1 | 2060(2) | <<< |
| OG0000824 | Na+ -ATPase subunit J | 3-634-1 | 2060(2) | <<< |
| OG0000825 | delta(24)-sterol C-methyltransferase | **3-634-1** | 2061(3) | >>> |
| OG0001645 | rod shape-determining protein MreC | 4-3834-1 | 3485(3) | >>> |
| OG0001646 | rod shape-determining protein | 4-3834-1 | 3485(3) | >>> |
| OG0001647 | single-strand-binding protein | **4-3834-1** | 3484(1) | <<< |
| OG0001648 | SIMPL domain-containing protein | **4-3834-1** | 3483(1) | >>> |
| OG0001650 | solanesyl diphosphate | **4-3898-1** | 3479(1) | <<< |
| OG0001651 | hypothetical protein | **4-3898-1** | 3478(1) | >>> |
| OG0001652 | ATPase involved in DNA repair | 4-3898-1 | 3477(2) | <<< |
| OG0001653 | PatU | 4-3898-1 | 3477(2) | <<< |

(*continued on next page*)

**Table 3** (*continued*)

| | | | | |
|---|---|---|---|---|
| OG0000828 | aldehyde oxygenase (deformylating) | **4-730-1** | 502(1) | <<< |
| OG0000829 | long-chain acyl-[acyl-carrier-protein] | **4-730-1** | 501(1) | <<< |
| OG0000830 | acetyl-CoA carboxylase carboxyltransferase | **4-730-1** | 500(1) | <<< |
| OG0000831 | short-chain dehydrogenase | **4-730-1** | 499(2) | <<< |
| OG0001114 | DUF2499 domain-containing protein | 5-1756-1 | 2556(2) | >>> |
| OG0001115 | ABC-type dipeptide/oligopeptide/nickel transport | 5-1756-1 | 2556(2) | >>> |
| OG0001116 | 11-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide isomerase | **5-1756-1** | 2557(1) | >>> |
| OG0001172 | fimbrial protein | 5-2026-1 | 3096(3) | <<< |
| OG0001173 | Tfp pilus assembly protein PilO | 5-2026-1 | 3096(3) | <<< |
| OG0001174 | type II secretory pathway, component HofQ | **5-2026-1** | 3095(1) | <<< |
| OG0001705 | hypothetical protein | 5-4409-1 | 3919(3) | <<< |
| OG0001706 | hypothetical protein | 5-4409-1 | 3919(3) | <<< |
| OG0001707 | hypothetical protein | 5-4409-1 | 3919(3) | <<< |
| OG0002002 | serine/threonine protein kinase | **5-5994-1** | 163(1) | >>> |
| OG0002003 | 23S rRNA (uracil-5-)-methyltransferase RumA | **5-5994-1** | 162(1) | >>> |
| OG0002004 | allophycocyanin alpha-B subunit apoprotein | **5-5994-1** | 161(1) | <<< |
| OG0000889 | 50S ribosomal protein L10 | **5-970-1** | 3684(2) | <<< |
| OG0000890 | 50S ribosomal protein L1 | 5-970-1 | 3686(4) | <<< |
| OG0000891 | 50S ribosomal protein L11 | 5-970-1 | 3686(4) | <<< |
| OG0000892 | transcription termination/antitermination | 5-970-1 | 3686(4) | <<< |
| OG0000893 | acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferasepreprotein translocase su | 5-970-1 | 3686(4) | <<< |
| OG0001256 | ATP synthase subunit C | 6-2491-1 | 379(7) | <<< |
| OG0001257 | F0F1 ATP synthase subunit B' | 6-2491-1 | 379(7) | <<< |
| OG0001258 | ATP synthase F0F1 subunit B | 6-2491-1 | 379(7) | <<< |
| OG0001259 | ATP synthase F0F1 subunit delta | 6-2491-1 | 379(7) | <<< |
| OG0001792 | lipid-A-disaccharide synthase | 6-5067-1 | 2206(7) | <<< |
| OG0001793 | acyl-[acyl-carrier-protein]--UDP-N- | 6-5067-1 | 2206(7) | <<< |
| OG0001794 | beta-hydroxyacyl-ACP dehydratase | 6-5067-1 | 2206(7) | <<< |
| OG0001795 | UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine | 6-5067-1 | 2206(7) | <<< |
| OG0001671 | carbon dioxide-concentrating protein CcmK | 7-4106-1 | 2801(4) | <<< |
| OG0001672 | carbon dioxide concentrating mechanism protein | 7-4106-1 | 2801(4) | <<< |
| OG0001673 | carbon dioxide concentrating mechanism protein | 7-4106-1 | 2801(4) | <<< |
| OG0001674 | carbonic anhydrase | 7-4106-1 | 2801(4) | <<< |
| OG0001935 | ATP-dependent protease ATP-binding subunit ClpX | 8-5734-1 | 185(4) | >>> |
| OG0001936 | ATP-dependent Clp protease proteolytic subunit | 8-5734-1 | 185(4) | >>> |
| OG0001937 | trigger factor | 8-5734-1 | 185(4) | >>> |
| OG0001388* | TIGR00303 family protein | **2-250-1** | 2633(1) | <<< |
| OG0001389* | membrane protein | 2-250-1 | 2634(2) | <<< |
| OG0001390* | CarD family transcriptional regulator | 2-250-1 | 2634(2) | <<< |
| OG0001020* | hypothetical protein | **2-3002-1** | 3046(1) | >>> |
| OG0001021* | hypothetical protein | **2-3002-1** | 3045(1) | >>> |
| OG0001022* | FliA/WhiG subfamily RNA polymerase sigma-28 | **2-3002-1** | 3044(1) | >>> |
| OG0001419* | ribonuclease P | **2-3002-2** | 2672(2) | <<< |
| OG0001420* | preprotein translocase YidC | 2-3002-2 | 2671(2) | <<< |
| OG0001421* | RNA-binding protein | 2-3002-2 | 2671(2) | <<< |

orientation within each COR across all reference genomes (unidirectional COR). In four COR, one gene was in an opposite orientation compared to the other genes in the same COR (bidirectional COR). In all four cases the gene oriented in opposite orientation belonged to a different operon (Table 3). This opposite orientation was conserved in all reference genomes. Therefore, for all 25 COR, the relative orientation of genes within any given COR is conserved in all genomes. However, the conservation is only observed within a COR, the orientation of the COR compared to the rest of the genome being variable among reference genomes (Table 3). This suggests that parts of genomes have undergone inversion events during their evolution. However, it is very interesting that 21 of 25 COR have all their genes in the same orientation, given that no filters on conservation of the orientation of genes was enforced in MESYRES. This result also demonstrates the efficiency of our methodology, as collinearity is maintained (i.e., gene orientation is conserved across all analyzed genomes) for all 25 COR. Among the 25 COR, 15

were composed of genes from the same operon (Table 3). The 10 remaining COR were composed of genes from two to three different operons (five and two cases, respectively) or only composed of "solitary" genes, each of which being controlled by a distinct promotor (3 cases).

### 3.2. Phylogenies based on syntenic versus non-syntenic datasets

#### 3.2.1. Comparison of the phylogenies based on jackknifing of 1160 orthologous genes versus the 99 COR genes datasets for the 16 reference species

To assess the resolving power of the 99-gene syntenic dataset (21,280 amino-acid [AA] positions), we compared the topologies and statistical support of trees inferred using our 99-gene DNA and protein alignments, derived from 16 reference taxa, to a consensus protein tree from 100 jackknife replicates of ca. 21,200 AA each, sampled from a 1160 orthologous genes dataset. The latter were first filtered based on the number of average gene copies per organism, to avoid artifacts due to the inclusion of multigenic families. The trees were generated with a standard phylogenomic approach composed of gene alignment, conserved site selection, supermatrix creation and phylogenomic inference (Fig. 2A).

Our 99-gene DNA tree has the exact same topology as the 1160-gene protein tree based on jackknifed datasets (Fig. 3A). The 99-gene protein tree has the same topology as both gene trees except for two conflicts outside of the ingroup. The 99-gene DNA tree consistently has the highest support values, except for the node supporting the sister relationship of the *Fortiea* and *Anabaena* groups (70/80; Fig. 3A).

Therefore, the 99-gene DNA tree seems to be very powerful at resolving relationships within Nostocales, compared to AA based trees. The average of bootstrap (BS) values supporting branches of the 99-gene

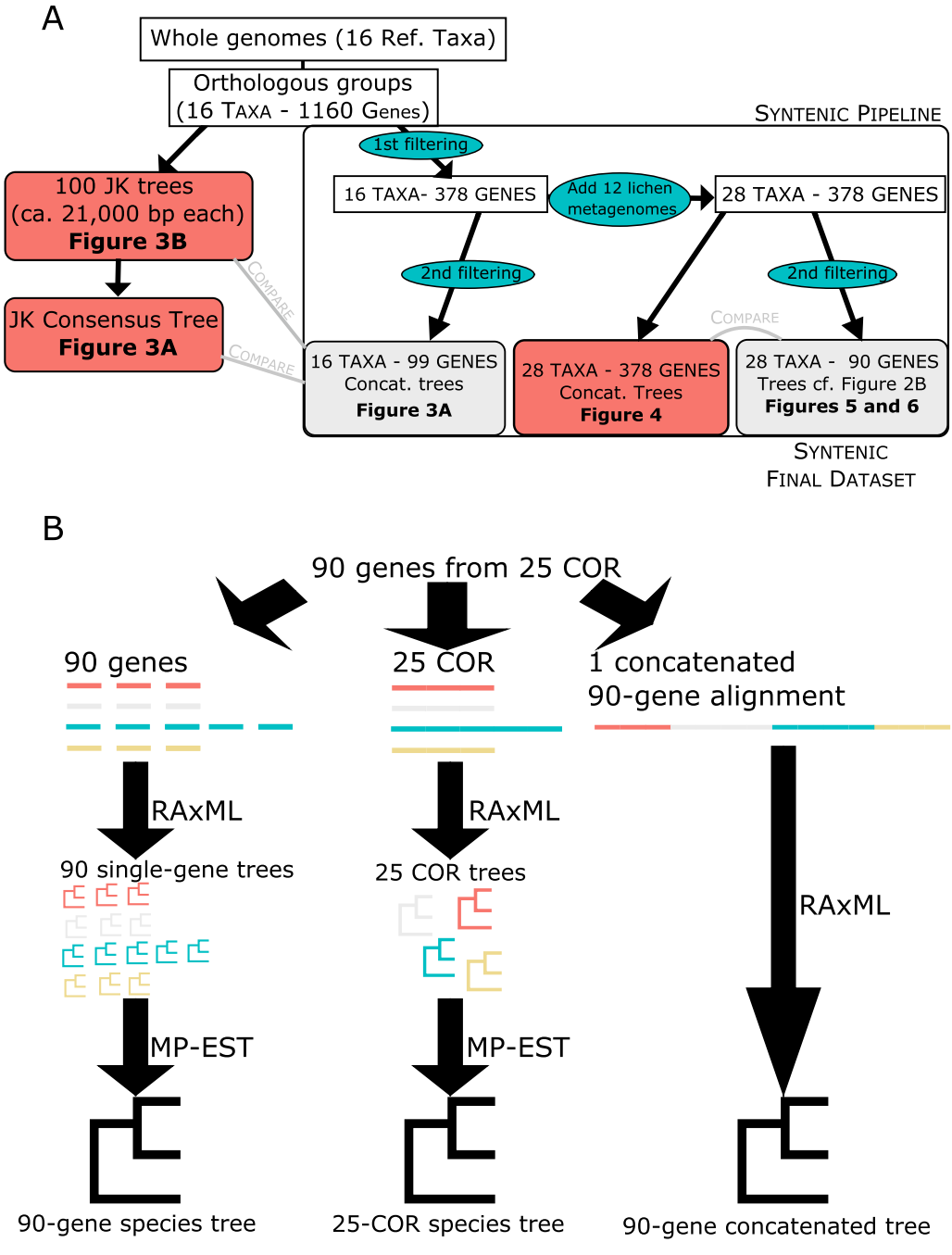


**Fig. 2.** Overview of phylogenetic analyses. A. Flowchart showing the datasets and trees part of this study. Gray boxes represent the syntenic and collinear datasets, whereas red boxes show datasets used for comparisons, and blue ovals represent various filtering steps or the addition of genomes. JK = jackknife. B. Description of the three sets of phylogenetic analyses performed on the final dataset of 90 genes belonging to 25 COR. The same set of analyses was done for the 71-gene 22-COR dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

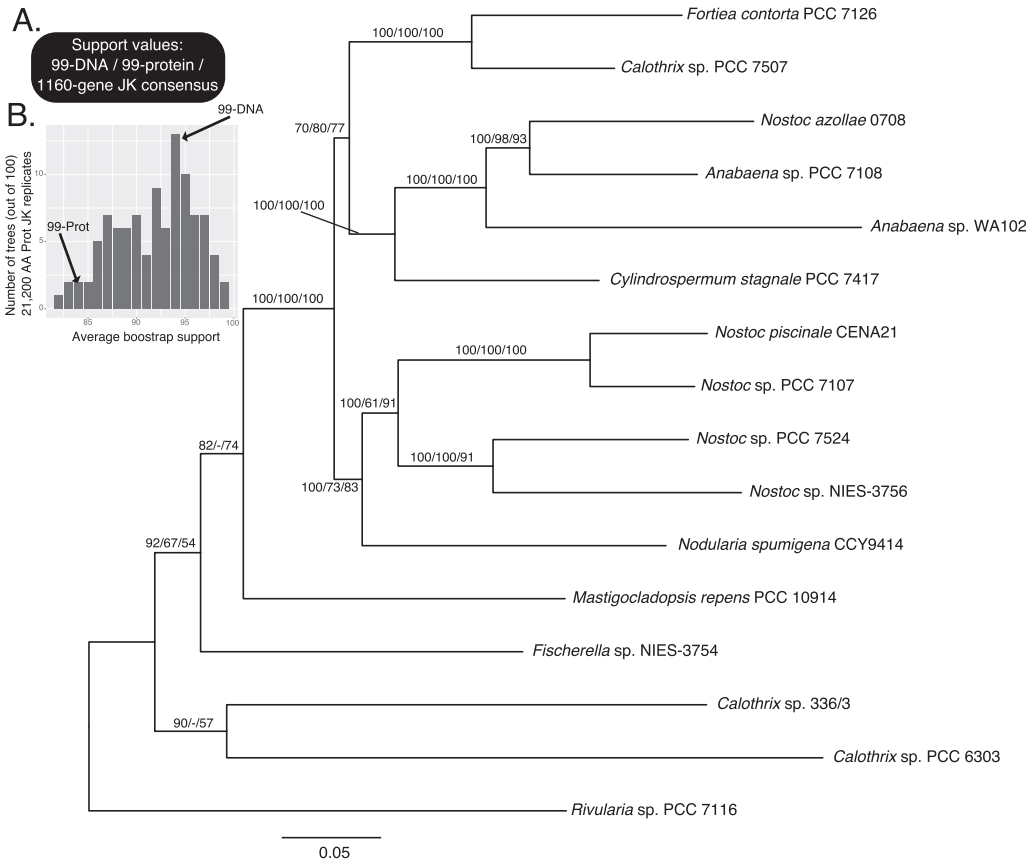


**Fig. 3.** Comparison of phylogenies derived from the 99-gene syntenic and collinear dataset with trees inferred with jackknifed datasets derived from the pool of 1160 orthologous genes for 16 reference cyanobacterial species (Fig. 2A). A. Best 99-gene DNA ML tree showing relationships among the 16 reference taxa. Support values above the branches represent bootstrap (BS) values of the best ML 99-gene DNA tree, BS values of the best 99-gene protein tree, and Jackknife (JK) protein tree values, respectively. A dash means that this relationship was not present in the tree. Five species (*Rivularia-Mastigocladopsis* grade) formed the outgroup. The scale corresponds to nucleotide substitutions per site. B. Distribution of the average bootstrap support values for the 100 jackknife replicates used to generate the jackknife protein consensus tree. Bars correspond to support classes, e.g., averages between 99 and 100. Arrows point to the category to which belong the 99-gene best DNA tree (shown in panel A) and the 99-gene best ML protein tree.

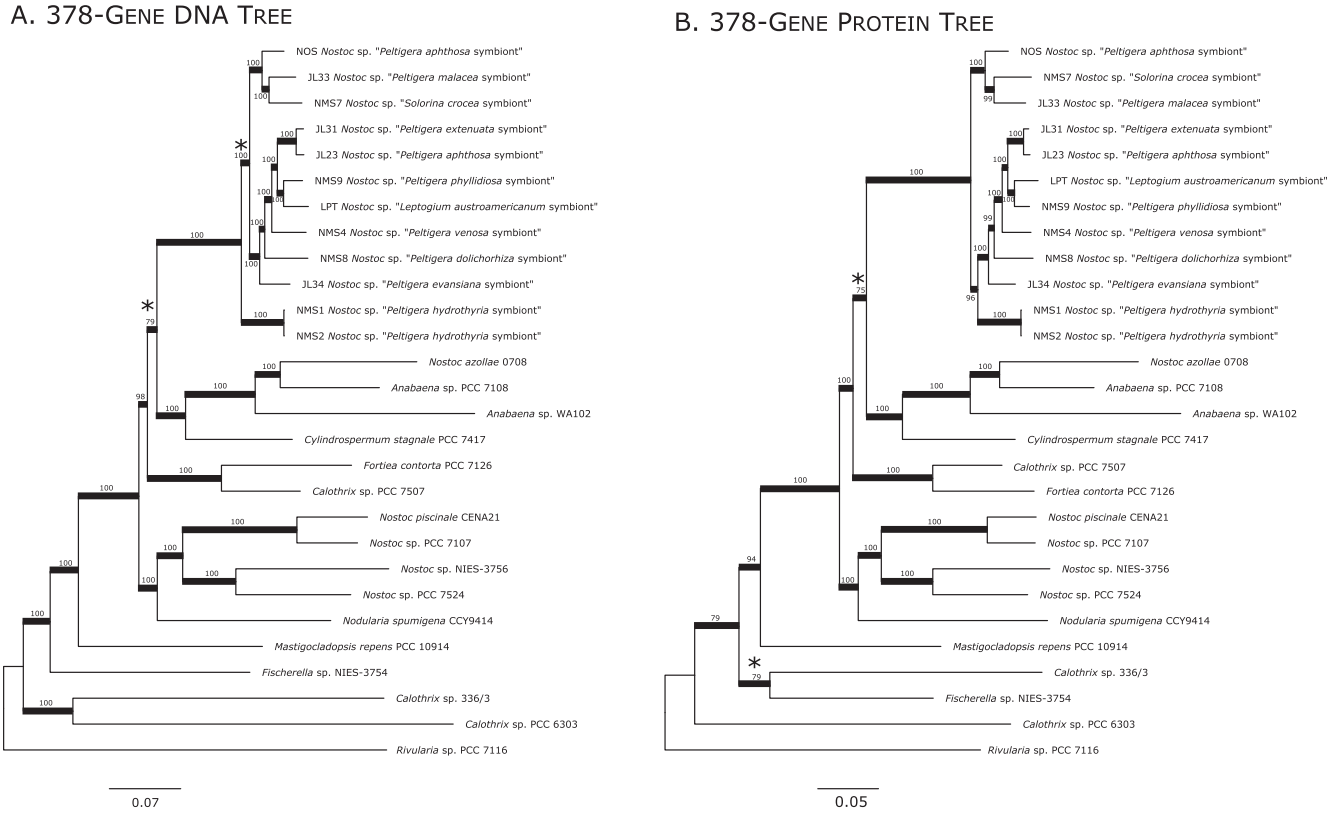


**Fig. 4.** Best trees resulting from the ML analyses of the 378-gene DNA dataset (A) and 378-gene protein dataset (B) before synteny filtering. Thick branches have bootstrap support (BS) $\geq$ 70% (values shown above internodes). Asterisks highlight topological conflicts with the 90-gene concatenated tree (Fig. 5A). The scales in (A) and (B) correspond to nucleotide and amino-acid substitutions per site, respectively.

nucleotide tree (94.9%) is slightly higher than the average of the bootstrap values derived from the 100 individual bootstrap protein trees inferred from the datasets generated by jackknife resampling (92.9%) (Fig. 3B). Support values of the 99-gene protein tree are lower (average 84.8%). The lengths of the jackknifed datasets sampled from the 1160 orthologous genes dataset are similar to our 99-gene alignment. However, many of the individual optimal phylogenetic trees derived from the 100 jackknifed datasets have highly supported conflicts when compared to the 99-gene trees and the consensus tree from 100 jackknife replicates. In contrast, neither the 99-gene DNA tree (0 conflict) nor the 99-gene protein tree (2 conflicts with BS = 61% and BS = 63%) have highly supported conflicts against the topology of the consensus protein tree from 100 jackknife replicates.

### 3.2.2. *Comparisons of phylogenies based on 378 orthologous genes with trees based on 90 or 71 COR genes*

To further demonstrate the use of the syntenic pipeline, we compared a DNA tree and a protein tree generated from our 90-gene syntenic dataset to DNA and protein trees generated from a 378-gene dataset for 28 taxa (Fig. 2A, 4, 5). The latter dataset results from an earlier stage of MESYRES, i.e., after the 1160 orthologous genes selection, but before syntenic filtering (see Fig. 2A). These 378 genes are genes that satisfy the concept of neighborhood, i.e., all proteins of an OG must have at least one neighbor protein among the 378-gene dataset (see the first step of MESYRES in Materials and methods).

Both the DNA and protein 378-gene trees recover the *Anabaena* clade as sister to the lichenized clade (i.e., *Nostoc* symbiont found in *Peltigera, Leptogium,* and *Solorina* lichens; Fig. 4) with high support, whereas the 90-gene synteny tree recovered the *Fortiea* clade as sister to the lichenized clade with a higher support value (Fig. 5A). The 378-gene DNA tree recovers a clade composed of NMS1 and NMS2 as the first diverging event within the lichenized clade with high support (Fig. 4A), whereas all other analyses recover a clade composed of NOS, NMS7 and JL33 as the result of the first divergence event within that clade (Fig. 4B, 5A–B, 6B). The 378-gene protein tree recovers a sister relationship between *Fischerella* sp. and *Calothrix* sp. 336/3 with high support (Fig. 4B), whereas the other trees recover *Calothrix* sp. 336/3 forming a monophyletic group with *Calothrix* sp. PCC 6303 (Fig. 4A, 5A–B). The strongly supported conflicts between the 378-gene DNA and protein trees (Fig. 4) highlight one of the caveats of phylogenomic studies based on the concatenation of a large number of orthologous genes, which is that it can lead to well-supported but wrong relationships (Kubatko and Degnan, 2007). Recently published studies based on different datasets have recovered different relationships among major clades within Nostocales (Warshan et al., 2018; Nelson et al., 2019; Bell-Doyon et al., 2020), emphasizing the need for adequate loci and method selection.

The three analyses relying on the 90-gene dataset (Fig. 5A–B, 6B) resulted in highly congruent topologies and delimited five main groups in the tree, corresponding to the same groups as in the 16-taxon analyses (Fig. 3A), with the addition of the lichenized clade. The first group is the outgroup, which contains five taxa, *Rivularia* sp., *Fischerella* sp., *Mastigocladopsis repens*, *Calothrix* sp. 336/3 and *Calothrix* sp. PCC 6303. However, the relationships of these five taxa are not congruent among these three analyses. Whereas they are identical among the 90-gene concatenated DNA tree (with 28 taxa, Fig. 5A), the 99-gene concatenated DNA tree (Fig. 3A), and the 1160-gene protein tree using

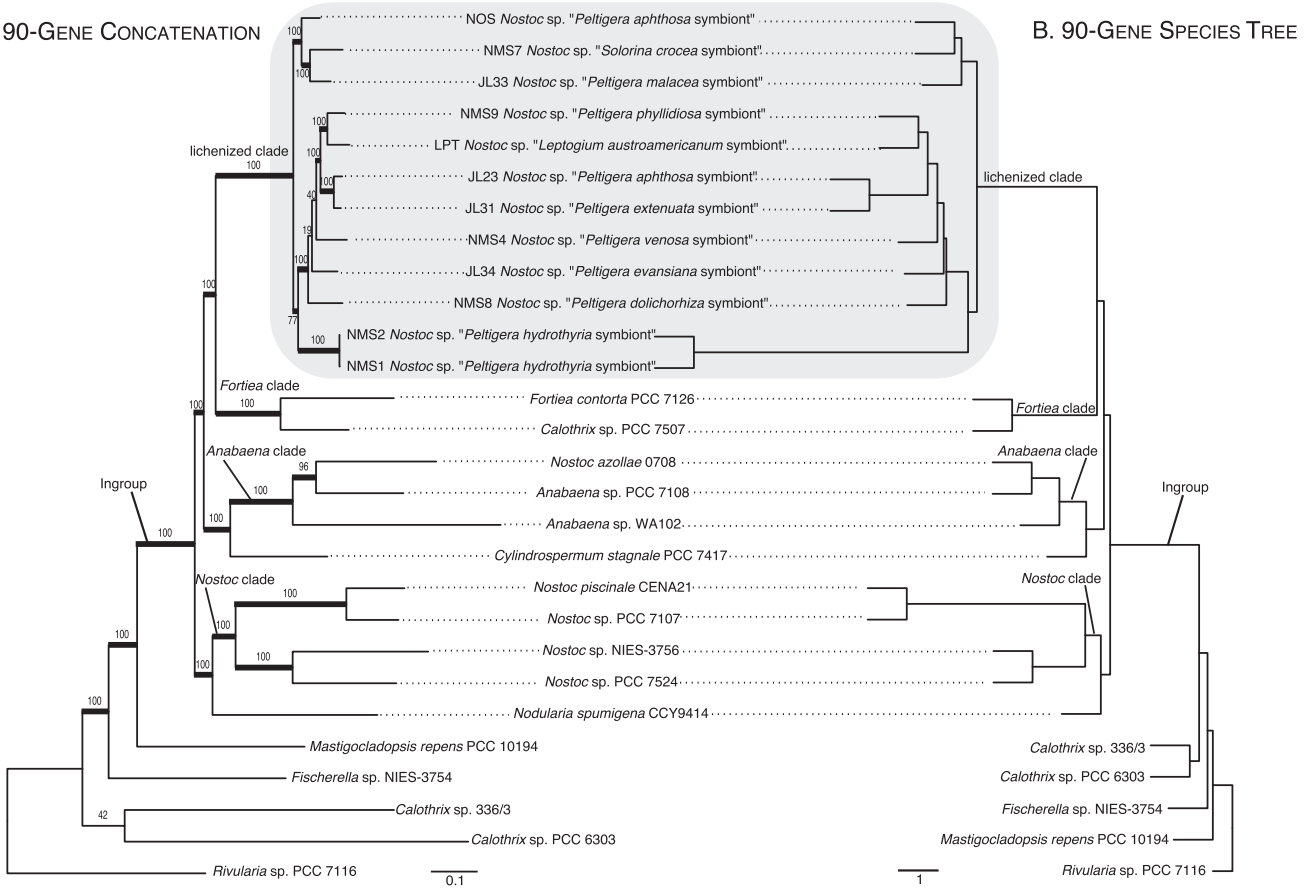


**Fig. 5.** Comparison of tree topologies resulting from the concatenated 90-gene analysis (A) and the 90-gene species tree analysis (B). The trees were rooted using *Rivularia* sp. following Shih et al. (2013). Thick branches on the concatenated tree have bootstrap support (BS) ≥ 70% (values shown above internodes). The gray box highlights the lichenized cyanobacterial clade, which corresponds to *Nostoc* clade II (O'Brien et al., 2005; Otálora et al., 2010; Magain et al., 2017a). The scales in (A) and (B) correspond to nucleotide substitutions per site and coalescence units, respectively.

jackknifed datasets (on 16 taxa), differences are present in the species trees (Fig. 5B and 6B). Of the 90 single-gene trees and the 25 COR trees, 49 and 21 trees, respectively, also group these five taxa together (Supplementary Table S3).

All three analyses (Fig. 5A–B, 6B) show *Nostoc piscinale*, *Nostoc* sp. PCC 7107, *Nostoc* sp. NIES-3756 and *Nostoc* sp. PCC 7524 forming a monophyletic group (*Nostoc* clade I, according to O'Brien et al. [2005] and Magain et al. [2017a]). Of the 90 single-gene trees and 25 COR trees, 38 and 17 trees, respectively, group these four taxa together (Supplementary Table S3). All three analyses also placed *Nodularia spumigena* sister to this group (Fig. 5A–B, 6B), which is also the case for 13 single-gene trees and 10 COR trees. The *Anabaena* clade, which includes *Nostoc azollae* 0708, *Anabaena* sp. PCC 7108 and *Anabaena* sp. WA-102, was recovered by all three analyses. A total of 57 single-gene trees and 23 COR trees also support this relationship (Supplementary Table S3). All three analyses (Fig. 5A–B, 6B), as well as 27 single-gene trees and 16 COR trees, show *Cylindrospermum stagnale* sharing a most recent common ancestor with this clade. *Fortiea contorta* and *Calothrix* sp. PCC 7507 form the "*Fortiea* clade", which is resolved as sister to the group formed by the 12 lichenized *Nostoc* (lichenized clade) in the 90-gene concatenated and species trees (Fig. 5). The lichenized clade was recovered by 84 single-gene trees and in all 25 COR trees (Supplementary Table S3). These two clades share a most recent common ancestor with the *Anabaena* clade (Fig. 5). In the 25-COR species tree, those three clades form a polytomy (Fig. 6).

Overall, the 90-gene concatenated and species trees are slightly more congruent with each other, and more resolved, than the 25-COR species tree. These three topologies are similar to the topology retrieved by Shih et al. (2013), which was based on an analysis of 31 conserved protein-coding genes. The taxonomic span of our trees corresponds to their clade B1.

When restricting our dataset to only genes belonging to the same operon within a COR (71-gene dataset, Table 3), the concatenated analysis recovered the same topology as in the 90-gene concatenated analysis (Fig. 5A, 7A), except for the relative positions of JL34 and NMS8 within the lichenized clade (relationships not supported in the 90-gene concatenated analysis). Furthermore, three internodes that were not well supported (BS $<$ 70%) in the 90-gene analysis are well supported in the 71-gene analysis (BS $\geq$ 70%). These relationships are the monophyly of *Calothrix* sp. PCC 6303 and *Calothrix* sp. 336/3; and two relationships within the lichenized clade (Fig. 7A, bootstrap values in bold). All internodes received BS $\geq$ 70% in the 71-gene concatenated analysis. The 71-gene (not shown) and 22-COR species trees are also highly congruent with the concatenated analyses, but with small differences, such as the position of *Fischerella* sp. in the 22-COR species tree (Fig. 7B). The relative positions of NMS4, NMS8 and JL34 within the lichenized clade are probably the most unstable relationships within these 28-taxa trees. In the 71-gene concatenated analysis, JL34 splits first, followed by NMS8 and NMS4. The three nodes supporting these splits are highly supported (BS $\geq$ 70%) in the 71-gene concatenated analysis (Fig. 7A). The same relationships are found in the 22-COR and 25-COR species trees (Fig. 6B, 7B). In the 90-gene concatenated analysis and the 90-gene species tree (Fig. 5), NMS8 splits first (but without support in the concatenated analysis). Including more genes might be better for species tree methods, whereas phylogenetic analyses of concatenated datasets seem to benefit from a more rigorous selection of

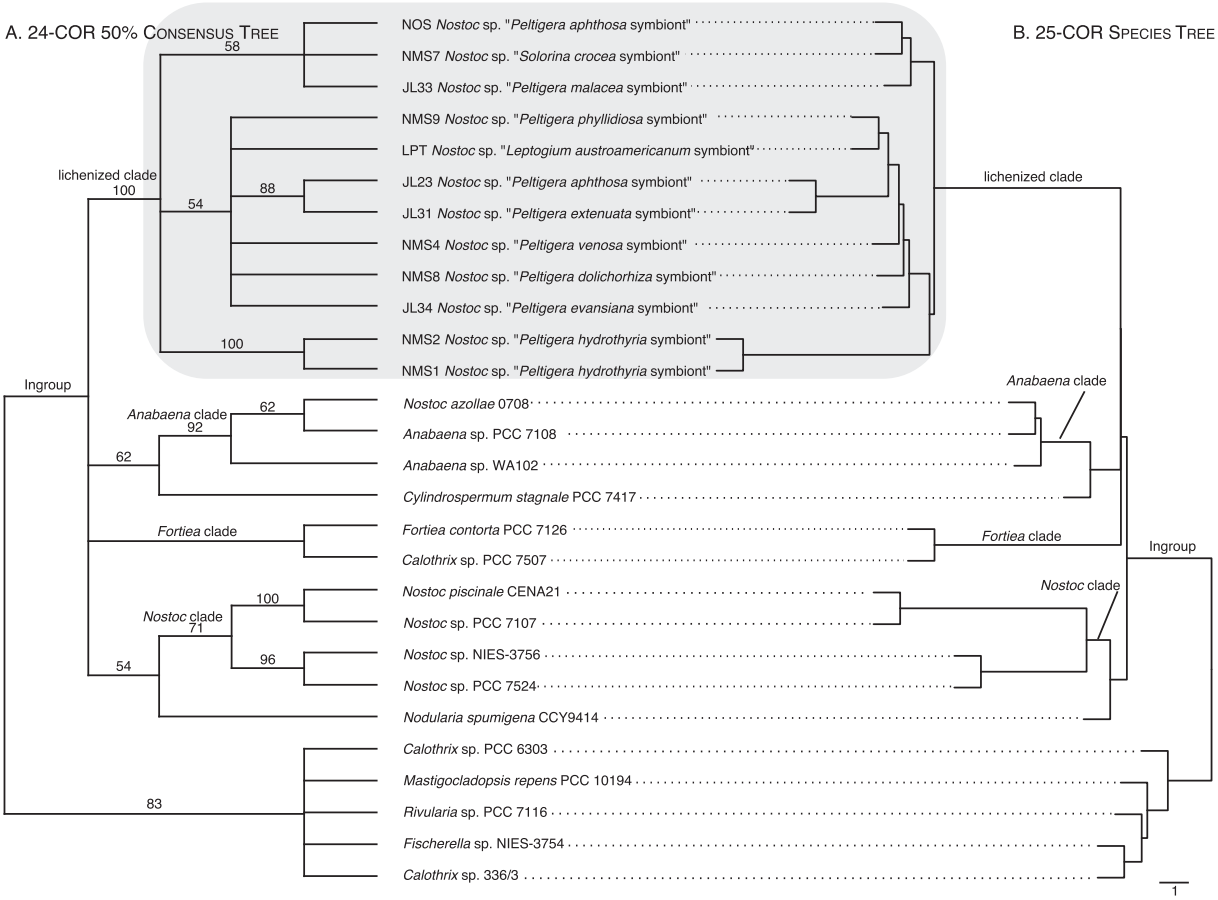


**Fig. 6.** Comparison of tree topologies resulting from the 50% majority-rule consensus tree based on 24 COR trees without missing data (A) and the 25-COR species tree analysis (B). Values above branches on the consensus tree represent the proportion of trees (%) where these specific topological bipartitions are present. The gray box indicates the clade of lichenized Cyanobacteria, which corresponds to *Nostoc* clade II (O'Brien et al., 2005; Otálora et al., 2010; Magain et al., 2017a). Scale in (B) corresponds to coalescence units.

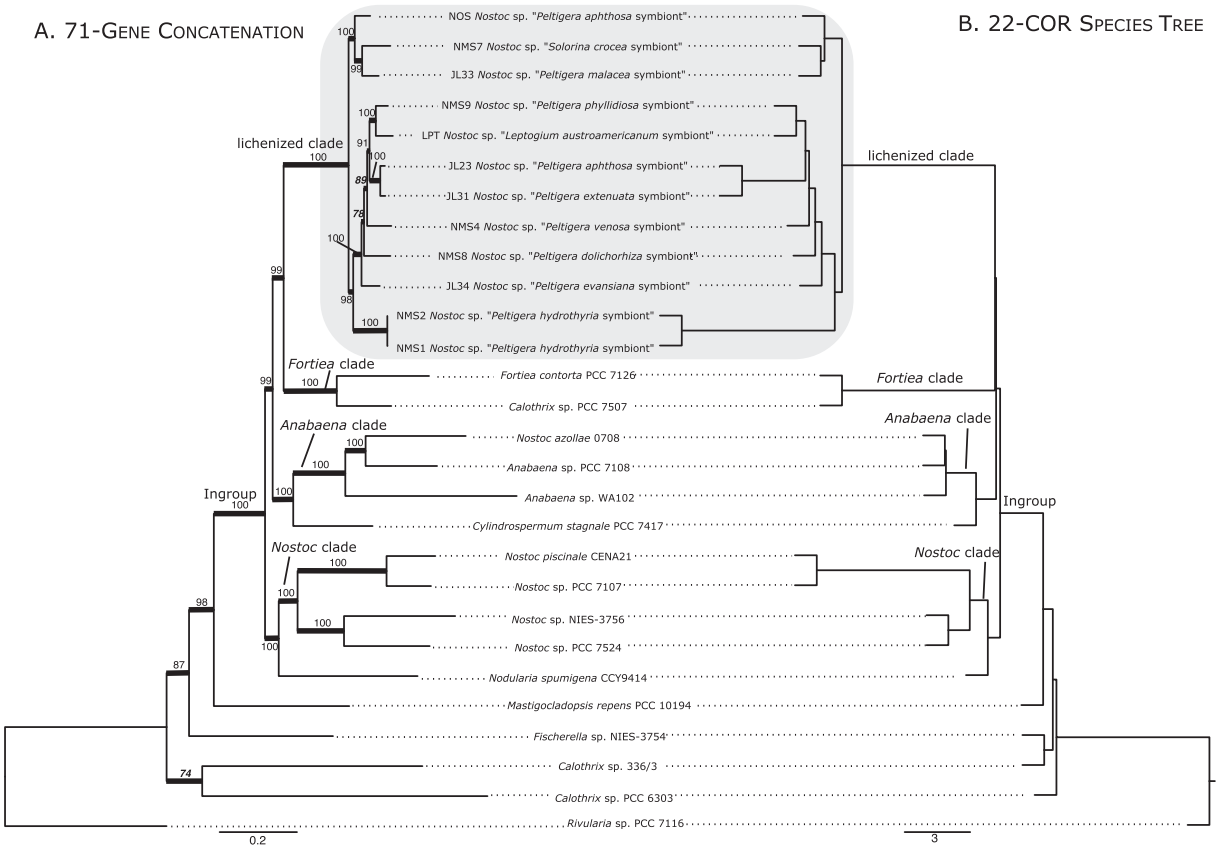


**Fig. 7.** Comparison of tree topologies resulting from the concatenated 71-gene analysis (A) and the 22-COR species tree analysis (B) after removal of genes that do not belong to the same operon within a COR (Table 3). The trees were rooted using *Rivularia* sp. following Shih et al. (2013). Thick branches on the concatenated tree have bootstrap support (BS) ≥ 70% (values shown above or below internodes). Nodes that were not supported with BS ≥ 70% in the 90-gene concatenation (Fig. 5A) but supported on the 71-gene concatenated tree are indicated with bold italic bootstrap values. The gray box highlights the lichenized cyanobacterial clade, which corresponds to *Nostoc* clade II (O'Brien et al., 2005; Otálora et al., 2010; Magain et al., 2017a). The scales in (A) and (B) correspond to nucleotide substitutions per site and coalescence units, respectively.

genes using criteria such as synteny, collinearity, and operon delimitations.

### 3.2.3. Phylogeny of the lichenized Nostoc clade

As most phylogenetic studies on lichenized Cyanobacteria have relied on single-locus analyses of *rbcLX* or SSU rRNA (16S), it is important to assess the congruence and resolution power of these markers compared to the 90-gene and 71-gene trees derived from concatenated datasets. Both *rbcLX* and 16S topologies retrieved the lichenized clade (*Nostoc* clade II) as a strongly supported monophyletic group (Fig. 8). However, discordances with the concatenated topology are present. For example, NMS7 *Nostoc* sp. is sister to JL33 *Nostoc* sp. in the 71-gene and 90-gene topologies as well as in the 22-COR species tree (Fig. 5A, 7), whereas NMS7 *Nostoc* sp. is sister to NOS *Nostoc* sp. in the 90-gene species tree and the *rbcLX* tree, but the latter is not well supported (Fig. 5B, 8A). In the 16S tree, no incongruent relationship is highly supported (i.e., with BS ≥ 70%) compared to the topology from the 71-gene concatenated dataset, except for the position of *Nodularia spumigena*. This is mostly because the 16S is too conserved to resolve with high confidence these relatively close phylogenetic relationships.

Contrary to the 16S tree, but in agreement with the *rbcLX* tree, 41 of the 90 single-gene trees, 18 of the 25 COR trees (Supplementary Table S3), all phylogenies resulting from the three analyses based on the 90-gene dataset (Fig. 5, 6B) and all trees derived from the 71-gene dataset (Fig. 7) separate NOS, NMS7 and JL33 (i.e., *Nostoc* clade II, subclade 2 in Fig. 8A) from a monophyletic group containing all other lichenized *Nostoc* (i.e., subclade 3 in Fig. 8A). The NMS1-NMS2 clade, which represents symbionts of the aquatic lichen *Peltigera hydrothyria* (Miadlikowska et al., 2014), is sister to the remaining seven *Nostoc* of subclade 3. All 24 COR trees group NMS1 and NMS2 together, and 13 COR trees group the seven remaining lichenized *Nostoc* together (Fig. 6A). Within the latter seven-taxon monophyletic group, four taxa form a clade that is subdivided into two groups of two taxa (NMS9 with LPT, and JL23 with JL31, Fig. 5, 6B, 7, 8A). All 90-gene analyses and 71-gene analyses converged on this result, which is further strongly supported by the bootstrap analysis of the 90-gene concatenated dataset (Fig. 5A). In contrast, the relationships among the three remaining taxa, NMS4, JL34 and NMS8, which are identical in the 90-gene concatenation and species trees (Fig. 5), but different from relationships based on the 71-gene dataset, are not well supported by any analysis. The only analysis with strong support for the relationships of these three taxa is based on the 71-gene concatenated dataset (Fig. 7A), which is the same topology revealed by the 22-COR species tree analysis (Fig. 7B).

The sister relationship of symbionts from *Peltigera hydrothyria* (NMS1 and NMS2) to a clade containing the other taxa from subclade 3 (Fig. 5–7) is revealed here for the first time. Deep nodes of subclade 3 appeared as a polytomy in previous studies (e.g., Miadlikowska et al., 2014; Magain et al., 2017a). Our *rbcLX* tree (Fig. 8A) also shows support for this relationship. This increased resolution based on *rbcLX* alone likely results from the use of the entire sequence of *rbcL* obtained from genomic data combined with the use of far less taxa. Previous studies (e.g., O'Brien et al., 2013; Miadlikowska et al., 2014; Magain et al., 2017a) only used a fragment of this gene and included far more taxa. Support for the monophyly of NMS9, LPT, JL23 and JL31 in the 90- and 71-gene analyses (Fig. 5, 6B, 7) also appears to be an improvement in phylogenetic resolution within *Nostoc* clade II subclade 3, compared to previous

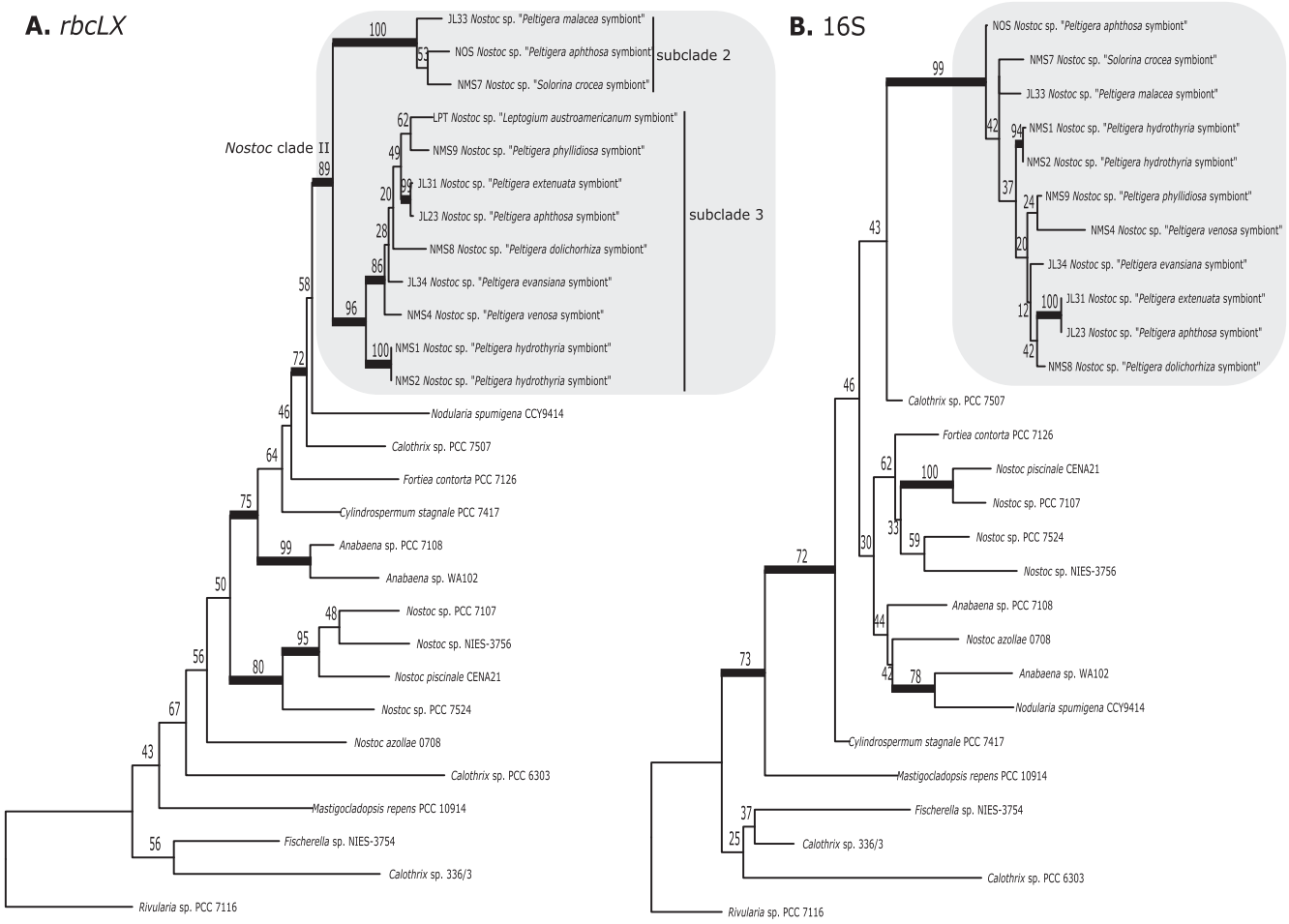


**Fig. 8.** Best ML trees inferred from our 28-taxon datasets using *rbcLX* (A) and SSU rRNA (16S) (B). Thick branches have bootstrap support (BS) $\geq$ 70% (values shown above or below internodes). Names of clades in the *rbcLX* tree follow Otálora et al. (2010). The gray boxes delineate the lichenized Cyanobacteria clade. The rooting of these trees follows Shih et al. (2013). The scales correspond to nucleotide substitutions per site.

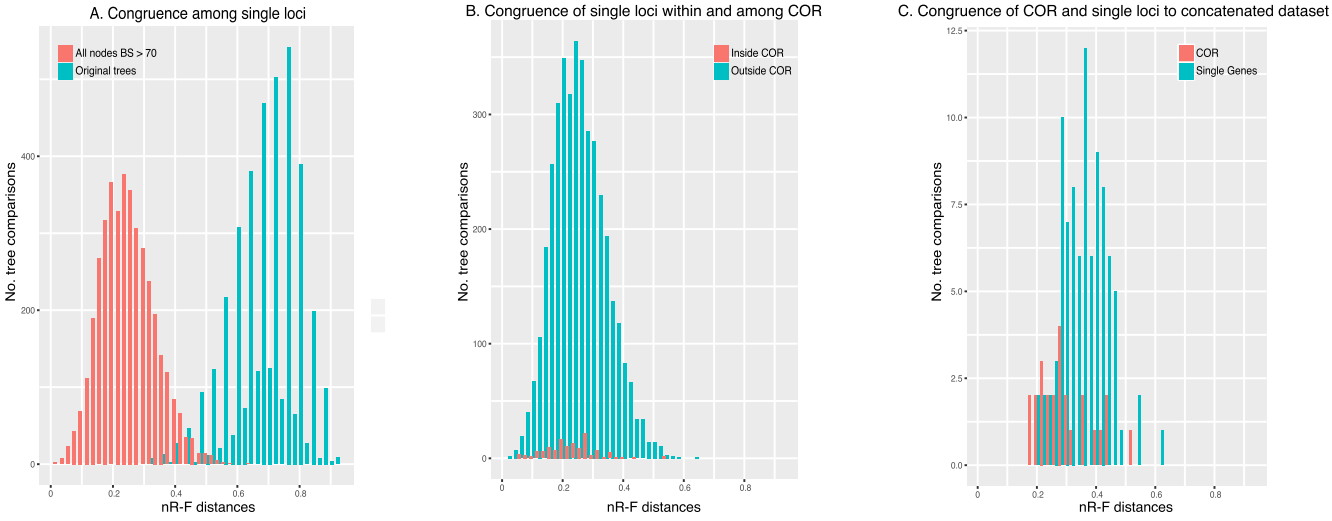


**Fig. 9.** Distribution of pairwise congruence among single-locus and COR trees using normalized Robinson-Foulds (nR-F) distances. (A) Congruence among 90 single-gene trees using the best ML trees (blue) and trees where all nodes with bootstrap support (BS) < 70% were collapsed (red). (B) Congruence among single-gene tree topologies among COR (blue) and genes within the same COR (red) using trees where all nodes with BS < 70% were collapsed. (C) Comparison of the 90 single-gene trees (blue) and the 25 COR trees (red) to the topology from the 90-gene concatenated dataset where all nodes with BS < 70% were collapsed. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

studies (e.g., Magain et al., 2017a).

Overall, the topology within the lichenized clade of our *rbcLX* tree (Fig. 8A) matches the 90-gene concatenated topology and the topologies based on the 71-gene dataset (Fig. 5A, 7). The few topological differences are never well-supported by the phylogenetic analysis of the *rbcLX* dataset. Therefore, it seems that *rbcLX* is a good marker to infer relationships among lichenized *Nostoc* (i.e., within *Nostoc* clade II), but not at a broader phylogenetic level inside Nostocales (see above). Moreover, relying on a single marker does not allow for sufficient phylogenetic resolution on datasets with high numbers of taxa (see e.g., Magain et al. 2017a).

### 3.3. Resolution and congruence among genes and collinear orthologous regions (COR)

#### 3.3.1. Congruence among the 90 single-gene topologies

We used the Robinson-Foulds (R-F) distance as a measure of incongruence (Robinson and Foulds, 1981). After normalization (nR-F), a value of 0 indicates that two tree topologies are identical, whereas a value of 1 means that the trees are totally different, i.e., do not share a single topological bipartition. The nR-F pairwise distances between our 90 single-gene DNA trees ranged from nR-F = 0.32 to 0.92, with an average of 0.69 (Fig. 9A). Similarly, their incongruence with the topology from the 90-gene concatenated dataset ranged from 0.28 to 0.88, with an average of 0.55 (Supplementary Table S4). A large proportion of this incongruence can be explained by unsupported relationships in our trees, due to insufficient phylogenetic information contained in single-gene datasets. To account for this stochastic error, we also measured incongruence of trees after collapsing all nodes that have bootstrap values <70%, the widely-used threshold for a node to be considered highly supported (Hillis and Bull, 1993). When comparing tree topologies that only contain strongly supported relationships, the nR-F pairwise distances among single-gene topologies ranged from nR-F = 0.02 to 0.64, with an average of 0.252 (Fig. 9A). The incongruence with the 90-gene concatenated tree ranged from nR-F = 0.20 to 0.62, with an average of 0.36 (Supplementary Table S4).

#### 3.3.2. Congruence among genes within versus among collinear orthologous regions

We tested the hypothesis that genes inside a COR would be more congruent among each other than to other genes of our 90-gene dataset. On average, nR-F distances for single-gene pairwise comparisons inside COR are indeed slightly lower than comparisons between genes from different COR: 0.23 versus 0.25 for trees where all nodes with BS < 70% were collapsed (Fig. 9B). These differences are statistically significant (*t*-test $P = 0.0044$). However, there are exceptions, and these are especially easy to detect when the gene topologies are very different from the 90-gene concatenated tree (Supplementary Table S4).

For example, in COR 6–2491, genes 6–2491-1256, 6–2491-1257 and 6–2491-1258 (ATP synthase subunit C, F0F1 ATP synthase subunit B' and ATP synthase F0F1 subunit B, respectively, Supplementary Table S1) show strong support to delimit a clade consisting of *Mastigocladopsis repens*, *Rivularia* sp., *Anabaena* sp. PCC 7108 (*Anabaena* clade), *Nostoc* sp. PCC 7524, *Nostoc* sp. NIES-3756 (*Nostoc* clade I), *Cylindrospermum stagnale* and *Nodularia spumigena* sister to the remaining 21 taxa. Otherwise, expected relationships are maintained within these two monophyletic groups. This pattern could be explained by undetected hidden paralogy: a gene duplication event, or a HGT next to its ortholog, followed by differential loss of one copy in the group of seven taxa, and the other copy in the group of 21; even if other reasons such as model inadequacy cannot be ruled out. Interestingly, gene 6–2491-1259 (ATP synthase F0F1 subunit delta), which is from the same COR and same operon (Table 3) in all 28 genomes, does not show such incongruence and is more congruent to the 90-gene concatenated tree, with nR-F = 0.30 compared to 0.42–0.46 for the three other genes (with unsupported nodes collapsed) (Supplementary Table S4). Moreover, the three trees with this unique "21vs7" topology are more similar among themselves (nR-F = 0.48–0.60) than to gene 1259 (nR-F = 0.68–0.76). All four genes within this COR and operon have the same orientation (Table 3).

This COR is among the four COR that are the most incongruent with the 90-gene concatenated-dataset tree topology (Supplementary Table S4). Each of these COR, with nR-F values ranging from 0.42 to 0.52, includes genes part of a single operon except for the COR with the highest nR-F value (0.52; COR 3–634), which includes three genes part of two different operons (Table 3). There are thus cases where unique, well-supported, relationships are found in several or all genes within a COR. In these cases, genes from the same COR share a more similar evolutionary history compared to genes from other COR. However, when looking at the entire dataset, congruence is only slightly higher among individual genes within a COR than genes from different COR (average nR-F $0.23 \pm 0.085$ vs $0.252 \pm 0.087$, unpaired two-tailed *t*-test $P = 0.0044$; Fig. 9B).

Whether all genes have the same orientation or not within a COR does not seem to explain the congruence among COR trees consistently. One of the four COR where one gene is in the opposite direction compared to the other genes (belonging to another operon) of the COR (COR 3–634) has the worst congruence score of all 25 COR (nR-F = 0.52; Table 3, Supplementary Table S4). However, the three other COR that include one gene in the opposite direction are among the most congruent, including the single most congruent COR (4–3898 nR-F = 0.18; 5–5994, nR-F = 0.20; 4–3834, nR-F = 0.26, when nodes with BS < 70% are collapsed).

When comparing congruence between individual genes part of the same operon within a COR (n = 97, average nR-F = $0.21 \pm 0.07$, nodes with BS < 70% collapsed) with congruence of genes part of different operons within the same COR (n = 34, average nR-F = $0.29 \pm 0.099$), we found a very significant difference (two-tailed *t*-test *P*-value < 0.0001). Consequently, we also found a higher congruence between genes part of a COR consisting only of genes from a single operon (average nR-F = $0.21 \pm 0.076$, n = 85) compared to congruence of genes within COR consisting of at least two operons (average nR-F = $0.27 \pm 0.09$, n = 46, two-tailed *t*-test *P*-value < 0.0001). We found no significant difference in congruence between genes part of the same operon regardless of whether the entire COR is part of the same operon (average nR-F = $0.21 \pm 0.076$, n = 85), or whether only part of the COR is in the same operon (average nR-F = $0.23 \pm 0.04$, n = 12, two-tailed *t*-test $P = 0.37$).

#### 3.3.3. Congruence among entire COR versus single genes

Concatenating genes from a COR is expected to increase resolving power compared to each gene being analyzed separately. Indeed, COR topologies are more congruent to the 90-gene concatenated topology (average nR-F = 0.296 when nodes with BS < 70% are collapsed) than single genes (average nR-F = 0.360) (Fig. 9C, Supplementary Table S4), and this difference is statistically significant (*t*-test $P = 0.008$). In general, when comparing seven major relationships (i.e., the monophyly of the ingroup, the *Anabaena*, *Nostoc* and lichenized clades, the relationships of *Nodularia spumigena*, the position of *Cylindrospermum stagnale*, and the segregation of taxa from the lichenized clade subclade 2 versus subclade 3, Supplementary Table S3), 17 of 25 COR trees include at least five of these relationships, and eight COR trees show all seven. In contrast, only five of the 90 single-gene trees show all seven relationships, and 23 show five of these relationships. Thirty-two single-gene trees show two or less of these relationships, whereas a single COR tree (COR 6–2491, discussed above) only shows two of these relationships.

### 3.4. Exploring new markers for phylogenetic studies of Nostocales

When comparing trees with collapsed internodes (BS < 70%), the widely used markers 16S and *rbcLX* both have a nR-F value of 0.36 compared to the 90-gene concatenated topology, which is very similar to the mean of our 90-gene distribution, i.e., 0.358. Forty single-gene topologies from the 90 single-gene matrices have better congruence than

16S and *rbcLX*, 38 are worse, and eight have the same nR-F distance of 0.36 (Supplementary Table S4). Among the 25 COR trees, 18 are more congruent to the 90-gene concatenated topology than 16S and *rbcLX*, two are equally congruent, and five are less congruent. Even though the congruence of *rbcLX* and 16S is in the range of our 90 genes, this result suggests that using a COR, or even a well-chosen single gene, would give better results to reconstruct phylogenetic relationships within Nostocales than using *rbcLX* or the SSU rRNA (16S).

Among the 90 single-gene trees, the nine most congruent trees have nR-F values to the 90-gene concatenated topology comprised between 0.20 and 0.26, whereas the ten next most congruent genes all have a nR-F value of 0.28 (Supplementary Table S4). Interestingly, each of the nine most congruent genes belongs to a different COR. Among them, the topology of 5–5994-2003 retrieves all seven reference relationships (see above, Supplementary Table S3) and has 16 nodes with BS $\geq$ 70%. Topologies of genes 4–3898-1652 and 5–970-892 show all reference relationships, except for the position of *Nodularia spumigena*, and have 15 and 12 nodes with BS $\geq$ 70%, respectively, whereas 3–3706-1629 and 6–5067-1795 show all reference relationships, except for the position of *Cylindrospermum stagnale*, and have 11 and 13 nodes with BS $\geq$ 70%, respectively.

Other single-gene trees displaying all seven reference relationships (Supplementary Table S3) but with higher nR-F distances (Supplementary Table S4) include 3–4474-1737 (chemotaxis protein CheY, nR-F distance to the 90-gene concatenated topology = 0.28, 13 nodes BS $\geq$ 70%), 6–5067-1792 (lipid-A-disaccharide synthase, nR-F distance = 0.30, 11 nodes BS $\geq$ 70%), 5–2026-1172 (fimbrial protein, nR-F distance = 0.32, 14 nodes BS $\geq$ 70%), 3–490-759 (methyl-accepting chemotaxis sensory transducer, nR-F distance = 0.34, 22 nodes BS $\geq$ 70%). All these genes seem very promising and could be considered for single- or multi-gene based studies of Nostocales.

When looking at COR trees where all nodes with BS < 70% were collapsed, eleven out of 25 trees have nR-F values $\leq$ 0.26 (compared to nine out of 90 single-gene trees) (Supplementary Table S4). These eleven most congruent COR trees are 4–3898 (20 nodes BS $\geq$ 70%), 6–5067 (15 nodes BS $\geq$ 70%), 5–5994 (20 nodes BS $\geq$ 70%), 26–4970-3 (20 nodes BS $\geq$ 70%), 3–490 (21 nodes BS $\geq$ 70%), 3–3418 (15 nodes BS $\geq$ 70%), 3–4474 (19 nodes BS $\geq$ 70%), 4–730 (20 nodes BS $\geq$ 70%), 5–970 (18 nodes BS $\geq$ 70%), 4–3834 (11 nodes BS $\geq$ 70%) and 7–4106 (13 nodes BS $\geq$ 70%). Only five of these 11 COR include genes that are part of a single operon. The remaining six COR have genes part of two to four operons. The seven reference relationships are present in seven of the trees from these 11 regions (Supplementary Table S3). These COR would be good candidates for studies of cyanobacterial evolution, especially COR 26–4970-3 and COR 6–5067-1, provided that the amplification of such long fragments is possible. For example, COR 26–4970-3 is composed of eight genes (six of which are ribosomal proteins), which is the COR with the highest number of genes considered in this study (Supplementary Table S4).

## 4. Conclusion

Phylogenies resulting from the concatenated dataset and species trees based on the 90 genes belonging to 25 COR, and the 71 genes distributed across 22 COR, are highly congruent, highly supported, and overall, match the topologies reported in other studies (e.g., Shih et al., 2013). They are also congruent with a phylogenomic consensus tree based on 100 jackknifed datasets. Moreover, several single-gene trees and most COR trees recover the overall topology for our 28-taxa dataset.

Contrary to what we expected, congruence is only slightly (but significantly) better within a COR than among other genes of our 90-gene dataset. Yet, using entire syntenic regions rather than single genes improve the resolution and congruence of the trees. Most COR trees have a topology where the main clades and most relationships of the 90-gene concatenated topology are recovered, often with high support. However, restricting the analysis to genes part of the same operon within each COR (71-gene in 22 COR for this study) generated trees that were nearly perfectly congruent when based on a concatenated dataset versus generated with a species tree analysis, with the maximum amount of bootstrap support for the ML analysis on the concatenated dataset.

In spite of the power of phylogenomic approaches, sequencing of whole genomes or even targeting dozens of genes is not always possible, especially for studies based on a high number of specimens. Two alternative strategies can help solving this problem: (1) Build a reference phylogeny for exemplar taxa for which genomic or metagenomic data is available and use it as a backbone constraint tree for a larger phylogenetic analysis including many specimens for which one locus (used for the phylogenomic analysis) that is particularly efficient at resolving relationships with high phylogenetic confidence has been sequenced (Cornet et al., 2018b). (2) Choose a few genes part of a single COR or two COR that have been shown congruent in closely related groups and sequence them across a large number of specimens. Our study presents a number of potential markers with high congruence, resolution, and phylogenetic confidence that could be used in single- or multi-gene studies within Nostocales.

Our study also demonstrated that widely used loci, such as SSU rRNA (16S) and *rbcLX*, are not among the best single loci available to resolve phylogenetic relationships within Nostocales. Better loci should be considered for future studies of lichenized Cyanobacteria, either as replacement, or for comparison, to these widely used loci. The synteny-collinearity-operon approach proposed here also shows promising opportunities to select multiple genes to resolve phylogenetic relationships among other groups of Cyanobacteria and prokaryotes where HGT is likely to be problematic. Finally, the metagenomic pipeline presented here, including MESYRES, allows the direct use of metagenomic data and the obtention of a set of trustful orthologous genes, which saves considerable amounts of time, as it does not require the isolation of *Nostoc* in pure cultures.

## 5. Availability of data and material

All data used in this study are available at the National Center for Biotechnology Information (NCBI) under the BioProject PRJNA616181 http://www.ncbi.nlm.nih.gov/bioproject/616181. Raw reads are accessible on the Sequence Read Archive (SRA) under the accessions SAMN14483408 to SAMN14483418. Assemblies are available on GenBank under the accessions SAMN14531965 to SAMN14531976.

MESYRES is available in BitBucket at https://bitbucket.org/phylogeno/MESYRES.git.

All datasets used in this study, including individual genes and supermatrices are available at https://doi.org/10.6084/m9.figshare.12093882.v2.

## CRediT authorship contribution statement

**Luc Cornet:** Conceptualization, Data Curation, Formal Analysis, Software, Writing - original draft, Writing - review & editing. **Nicolas Magain:** Conceptualization, Data Curation, Formal Analysis, Visualization, Writing - original draft, Writing - review & editing. **Denis Baurain:** Conceptualization, Formal Analysis, Software, Writing - original draft, Writing - review & editing. **François Lutzoni:** Conceptualization, Writing - original draft, Writing - review & editing.

## Acknowledgments

LC was a FRIA fellow of the FRS-FNRS and then an IAP Planet Topers PhD scholar. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI; funded by FRS-FNRS, grant no. 2.5020.11), and through two grants to DB (University of Liège "Crédit de démarrage 2012" SFRD-12/04; FRS-FNRS "Crédit de recherche 2014" CDR J.0080.15). We thank Mark Miller and the CIPRES portal team for providing us with these clusters and assistance for the

analyses. NM and FL were supported by grants from the National Science Foundation (DEB-1025930 and DEB-1556995 to Jolanta Miadlikowska and FL). The authors thank John Logdson, Cindy Toll and Elizabeth Savelkoul for the sequencing of four metagenomes, as well as Romain Darnajoux for providing the culture of one *Nostoc* for which the genome was sequenced, Mick van Vlierberghe for providing his dataset "life of Mick" and Jolanta Miadlikowska, Antoine Simon and Frances Anderson for providing *Peltigera* and *Solorina* specimens that were used to sequence metagenomes for this study. Rosa Gago is acknowledged for the improvement of Fig. 1. Finally, we are grateful to the anonymous reviewers for their insightful comments.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107100.

## References

Abby, S.S., Tannier, E., Gouy, M., Daubin, V., 2012. Lateral gene transfer as a support for the tree of life. Proc. Natl. Acad. Sci. 109, 4962–4967.
Arnold, A.E., Miadlikowska, J., Higgins, K.L., Sarvate, S.D., Gugger, P., Way, A., Hofstetter, V., Kauff, F., Lutzoni, F., 2009. A phylogenetic estimation of trophic transition networks for ascomycetous fungi: Are lichens cradles of symbiotrophic fungal diversification? Syst. Biol. 58, 283–297.
Armaleo, D., May, S., 2009. Sizing the fungal and algal genomes of the lichen *Cladonia grayi* through quantitative PCR. Symbiosis 49, 43–51.
Armaleo, D., Müller, O., Lutzoni, F., Andrésson, Ó.S., Blanc, G., Bode, H.B., Collart, F.R., DalGrande, F., Dietrich, F., Grigoriev, I.V., Joneson, S., Kuo, A., Larsen, P.E., Logsdon, J.M., Lopez, D., Martin, F., May, S.P., McDonald, T.R., Merchant, S.S., Miao, V., Morin, E., Oono, R., Pellegrini, M., Rubinstein, N., Sanchez-Puerta, M.V., Savelkoul, E., Schmitt, I., Slot, J.C., Soanes, D., Szoevenyi, P., Talbot, N.J., Veneault-Fourrey, C., Xavier, B.B., 2019. The lichen symbiosis re-viewed through the genomes of Cladonia grayi and its algal partner Asterochloris glomerata. BMC Genomics 20, 605.
Aschenbrenner, I.A., Cernava, T., Berg, G., Grube, M., 2016. Understanding microbial multi-species symbioses. Front. Microbiol. 7, 180.
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.
Bi, G., Mao, Y., Xing, Q., Cao, M., 2018. HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. Genomics 110, 18–22.
Boekels-Gogarten, M., Gogarten, J.P., Olendzenski, L. (Eds.), 2009. Horizontal gene transfer: genomes in flux. Humana Press, New York, p. 500.
Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.
Brown, N.M., Mueller, R.S., Shepardson, J.W., Landry, Z.C., Morré, J.T., Maier, C.S., Hardy, J., Dreher, T.W., 2016. Structural and functional analysis of the finished genome of the recently isolated toxic *Anabaena* sp. WA102. BMC Genomics 17 (1), 457.
Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60.
Casano, L.M., del Campo, E.M., García-Breijo, F.J., Reig-Armiñana, J., Gasulla, F., Del Hoyo, A., Guéra, A., Barreno, E., 2011. Two *Trebouxia* algae with different physiological performances are ever-present in lichen thalli of *Ramalina farinacea*. Coexistence versus Competition? Environ. Microbiol. 13, 806–818.
Contreras-Moreira, B., Vinuesa, P., 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl. Environ. Microbiol. 79, 7696–7701.
Cornet, L., Meunier, L., Vlierberghe, M.V., Léonard, R.R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E.J., Philippe, H., et al., 2018a. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. PLoS ONE 13, e0200323.
Cornet, L., Wilmotte, A., Javaux, E.J., Baurain, D., 2018b. A constrained SSU-rRNA phylogeny reveals the unsequenced diversity of photosynthetic Cyanobacteria (Oxyphotobacteria). BMC Res. Notes 11, 435.
Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10, 210.
Cubero, O.F., Crespo, A.N.A., Fatehi, J., Bridge, P.D., 1999. DNA extraction and PCR amplification method suitable for fresh, herbarium-stored, lichenized, and other fungi. Plant Syst. Evol. 216 (3), 243–249.
Dal Forno, M., Lawrey, J.D., Sikaroodi, M., Gillevet, P.M., Schuettpelz, E., Lücking, R., 2020. Extensive photobiont sharing in a rapidly radiating cyanolichen clade. Mol. Ecol. https://doi.org/10.1111/mec.15700.
Dal Grande, F., Rolshausen, G., Divakar, P.K., Crespo, A., Otte, J., Schleuning, M., Schmitt, I., 2017. Environment and host identity structure communities of green algal symbionts in lichens. New Phytol. 217 (1), 277–289.
Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27 (8), 1164–1165.
Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.
Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284 (5423), 2124–2128.
Drillon, G., Carbone, A., Fischer, G., 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS ONE 9, e92621.
Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.
Elvebakk, A., Papaefthimiou, D., Robertsen, E.H., Liaimer, A., 2008. Phylogenetic patterns among *Nostoc* cyanobionts within bi- and tripartite lichens of the genus *Pannaria*. J. Phycol. 44 (4), 1049–1059.
Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157.
Fedrowitz, K., Kaasalainen, U., Rikkinen, J., 2011. Genotype variability of *Nostoc* symbionts associated with three epiphytic *Nephroma* species in a boreal forest landscape. The Bryologist 114 (1), 220–230.
Friedl, T., Büdel, B., 2008. Photobionts. In: Nash, T.H. (Ed.), Lichen Biology. Cambridge University Press, Cambridge, UK, pp. 9–26.
Gribaldo, S., Brochier, C., 2009. Phylogeny of prokaryotes: does it exist and why should we care? Res. Microbiol. 160 (7), 513–521.
Guimarães Jr, P.R., Jordano, P., Thompson, J.N., 2011. Evolution and coevolution in mutualistic networks. Ecol. Lett. 14 (9), 877–885.
Hauer, T., Bohunická, M., Johansen, J.R., Mareš, J., Berrendero-Gomez, E., 2014. Reassessment of the cyanobacterial family Microchaetaceae and establishment of new families Tolypothrichaceae and Godleyaceae. J. Phycol. 50 (6), 1089–1100.
Hilario, E., Gogarten, J.P., 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. Biosystems 31 (2–3), 111–119.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42 (2), 182–192.
Hirose, Y., Fujisawa, T., Ohtsubo, Y., Katayama, M., Misawa, N., Wakazuki, S., Shimura, Y., Nakamura, Y., Kawachi, M., Yoshikawa, H., et al., 2016a. Complete genome sequence of cyanobacterium *Fischerella* sp. NIES-3754, providing thermoresistant optogenetic tools. J. Biotechnol. 220, 45–46.
Hirose, Y., Fujisawa, T., Ohtsubo, Y., Katayama, M., Misawa, N., Wakazuki, S., Shimura, Y., Nakamura, Y., Kawachi, M., Yoshikawa, H., et al., 2016b. Complete genome sequence of cyanobacterium *Nostoc* sp. NIES-3756, a potentially useful strain for phytochrome-based bioengineering. J. Biotechnol. 218, 51–52.
Hodkinson, B.P., Gottel, N.R., Schadt, C.W., Lutzoni, F., 2012. Photoautotrophic symbiont and geography are major factors affecting highly structured and diverse bacterial communities in the lichen microbiome. Environ. Microbiol. 14 (1), 147–161.
Honegger, R., 2012. The symbiotic phenotype of lichen-forming ascomycetes and their endo- and epi-bionts. Fungal associations, second ed. Springer, Berlin Heidelberg, pp. 165–188.
Hyvärinen, M., Härdling, R., Tuomi, J., 2002. Cyanobacterial lichen symbiosis: The fungal partner as an optimal harvester. Oikos 98, 498–504.
Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., Philippe, H., 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol. 1, 1370–1378.
Isojärvi, J., Shunmugam, S., Sivonen, K., Allahverdiyeva, Y., Aro, E.M., Battchikova, N., 2015. Draft genome sequence of *Calothrix* strain 336/3, a novel H2-producing cyanobacterium isolated from a Finnish lake. Genome Announcements 3 (1), e01474–e1514.
Jüriado, I., Kaasalainen, U., Jylhä, M., Rikkinen, J., 2019. Relationships between mycobiont identity, photobiont specificity and ecological preferences in the lichen genus *Peltigera* (Ascomycota) in Estonia (northeastern Europe). Fungal Ecol. 39, 45–54.
Kaasalainen, U., Olsson, S., Rikkinen, J., 2015. Evolution of the tRNALeu (UAA) intron and congruence of genetic markers in lichen-symbiotic *Nostoc*. PLoS ONE 10 (6), e0131223.
Kang, D.D., Froula, J., Egan, R., Wang, Z., 2015. MetaBAT, an efficient tool for accurately reconstring single genomes from complex microbial communities. PeerJ 3, e1165.
Katoh, K., Standley, D.M., 2013. MAFFT Multiple sequence alignment software Version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
Khan, M.A., Mahmudi, O., Ullah, I., Arvestad, L., Lagergren, J., 2016. Probabilistic inference of lateral gene transfer events. BMC Bioinf. 2016 (17), 431.
Kroken, S., Taylor, J.W., 2000. Phylogenetic species, reproductive mode, and specificity of the green alga *Trebouxia* forming lichens with the fungal genus *Letharia*. The Bryologist 103 (4), 645–660.
Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56 (1), 17–24.
Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B., 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34, 772–773.
Laurin-Lemay, S., Brinkmann, H., Philippe, H., 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. 22, R593–R594.
Leão, T., Guimarães, P.I., de Melo, A.G.C., Ramos, R.T.J., Leão, P.N., Silva, A., Fiore, M. F., Schneider, M.P.C., 2016. Draft genome sequence of the N2-fixing cyanobacterium *Nostoc piscinale* CENA21, isolated from the Brazilian Amazon floodplain. Genome Announcements 4 (2), e00189–e216.

Lee, J.Z., Burow, L.C., Woebken, D., Everroad, R.C., Kubo, M.D., Spormann, A.M., et al., 2014. Fermentation couples Chloroflexi and sulfate-reducing bacteria to Cyanobacteria in hypersaline microbial mats. Front. Microbiol. 2014, 5–61.
Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10 (1), 302.
Lohtander, K., Oksanen, I., Rikkinen, J., 2003. Genetic diversity of green algal and cyanobacterial photobionts in *Nephroma* (Peltigerales). The Lichenologist 35 (4), 325–339.
Lutzoni, F., Miadlikowska, J., 2009. Lichens. Curr. Biol. 19 (13), R502–R503.
Magain, N., Miadlikowska, J., Goffinet, B., Sérusiaux, E., Lutzoni, F., 2017a. Macroevolution of specificity in cyanolichens of the genus *Peltigera* section *Polydactylon* (Lecanoromycetes, Ascomycota). Syst. Biol. 66 (1), 74–99.
Magain, N., Miadlikowska, J., Mueller, O., Gajdeczka, M., Truong, C., Salamov, A.A., Dubchak, I., Grigoriev, I.V., Goffinet, B., Sérusiaux, E., Lutzoni, F., 2017b. Conserved genomic collinearity as a source of broadly applicable, fast evolving, markers to resolve species complexes: a case study using the lichen-forming genus *Peltigera* section *Polydactylon*. Mol. Phylogenet. Evol. 117, 10–29.
Magain, N., Sérusiaux, E., 2014. Do photobiont switch and cephalodia emancipation act as evolutionary drivers in the lichen symbiosis? A case study in the Pannariaceae (Peltigerales). PLoS ONE 9 (2), e89876.
Magain, N., Truong, C., Goward, T., Niu, D., Goffinet, B., Sérusiaux, E., Lutzoni, F., Miadlikowska, J., 2018. Species delimitation at a global scale reveals high species richness with complex biogeography and patterns of symbiont association in *Peltigera* section *Peltigera* (lichenized Ascomycota: Lecanoromycetes). Taxon 67 (5), 836–870.
Manen, J.F., Falquet, J., 2002. The cpcB-cpcA locus as a tool for the genetic characterization of the genus *Arthrospira* (Cyanobacteria): evidence for horizontal transfer. Int. J. Syst. Evol. Microbiol. 52 (3), 861–867.
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 (1), 225.
Miadlikowska, J., Lutzoni, F., 2000. Phylogenetic revision of the genus *Peltigera* (lichen-forming Ascomycota) based on morphological, chemical, and large subunit nuclear ribosomal DNA data. Int. J. Plant Sci. 161 (6), 925–958.
Miadlikowska, J., Richardson, D., Magain, N., Ball, B., Anderson, F., Cameron, R., Lendemer, J., Truong, C., Lutzoni, F., 2014. Phylogenetic placement, species delimitation, and cyanobiont identity of endangered aquatic *Peltigera* species (lichen-forming Ascomycota, Lecanoromycetes). Am. J. Bot. 101 (7), 1141–1156.
Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Gateway Computing Environments Workshop (GCE), 2010, pp. 1–8. IEEE.
Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31 (12), i44–i52.
Muggia, L., Nelsen, M.P., Kirika, P.M., Barreno, E., Beck, A., Lindgren, H., Lumbsch, H.T., Leavitt, S.D., Trebouxia working group, 2020. Formally described species woefully underrepresent phylogenetic diversity in the common lichen photobion genus Trebouxia (Trebouxiophyceae, Chlorophyta): an impetus for developing an integrated taxonomy. Mol. Phylogenet. Evol. 149, 106821.
Myllys, L., Stenroos, S., Thell, A., Kuusinen, M., 2007. High cyanobiont selectivity of epiphytic lichens in old growth boreal forest of Finland. New Phytol. 173 (3), 621–629.
Nash, T.H., 2008. Lichen Biology, Second Edition. Cambridge University Press, New York.
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 (7), 1959–1964.
Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834.
O'Brien, H.E., Miadlikowska, J., Lutzoni, F., 2005. Assessing host specialization in symbiotic cyanobacteria associated with four closely related species of the lichen fungus *Peltigera*. Eur. J. Phycol. 40 (4), 363–378.
O'Brien, H.E., Miadlikowska, J., Lutzoni, F., 2013. Assessing population structure and host specialization in lichenized cyanobacteria. New Phytol. 198 (2), 557–566.
Onuț-Brännström, I., Benjamin, M., Scofield, D.G., Heiðmarsson, S., Andersson, M.G., Lindström, E.S., Johannesson, H., 2018. Sharing of photobionts in sympatric populations of *Thamnolia* and *Cetraria* lichens: evidence from high-throughput sequencing. Sci. Rep. 8 (1), 4406.
Otálora, M.A., Martínez, I., O'Brien, H., Molina, M.C., Aragón, G., Lutzoni, F., 2010. Multiple origins of high reciprocal symbiotic specificity at an intercontinental spatial scale among gelatinous lichens (Collemataceae, Lecanoromycetes). Mol. Phylogenet. Evol. 56 (3), 1089–1095.
Pardo-De la Hoz, C., Magain, N., Lutzoni, F., Goward, T., Restrepo, S., Miadlikowska, J., 2018. Contrasting symbiotic patterns in two closely related lineages of trimembered lichens of the genus *Peltigera*. Front. Microbiol. 9, 2770.
Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.
Paulsrud, P., Rikkinen, J., Lindblad, P., 1998. Cyanobiont specificity in some *Nostoc*-containing lichens and in a *Peltigera aphthosa* photosymbiodeme. New Phytologist 139 (3), 517–524.
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, e1000602.
Philippe, H., de Vienne, D.M., Ranwez, V., Roure, B., Baurain, D., Delsuc, F., 2017. Pitfalls in supermatrix phylogenomics. Eur. J. Taxon. 283, 1–25.
Popa, O., Landan, G., Dagan, T., 2017. Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J. 11 (2), 543.
Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J.A.A., Ininbergs, K., Zheng, W.-W., Lapidus, A., Lowry, S., Haselkorn, R., Bergman, B., 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS ONE 5, e11486.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53 (1–2), 131–147.
Rodriguez, F.J.L.O.J., Oliver, J.L., Marin, A., Medina, J.R., 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142 (4), 485–501.
Rodríguez, A., Burgon, J.D., Lyra, M., Irisarri, I., Baurain, D., Blaustein, L., Göçmen, B., Künzel, S., Mable, B.K., Nolte, A.W., Veith, M., 2017. Inferring the shallow phylogeny of true salamanders (Salamandra) by multiple phylogenomic approaches. Mol. Phylogenet. Evol. 115, 16–26.
Rolland, T., Neuvéglise, C., Sacerdot, C., Dujon, B., 2009. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. PLoS ONE 4 (8), e6515.
Roure, B., Rodriguez-Ezpeleta, N., Philippe, H., 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7, S2.
Shi, T., Falkowski, P.G., 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. PNAS 105, 2510–2515.
Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., de Marsac, N.T., Rippka, R., Herdman, M., 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. PNAS 110, 1053–1058.
Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapebie, P., 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. 27, 958–967.
Skaloud, P., Peksa, O., 2010. Evolutionary inferences based on ITS rDNA and actin sequences reveal extensive diversity of the common lichen alga *Asterochloris* (Trebouxiophyceae, Chlorophyta). Mol. Phylogenet. Evol. 54 (1), 36–46.
Snir, S., Rao, S., 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. Mol. Phylogenet. Evol. 62 (1), 1–8.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.
Stöver, B.C., Müller, K.F., 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinf. 11 (1), 7.
Stuart, R.K., Mayali, X., Lee, J.Z., Craig Everroad, R., Hwang, M., Bebout, B.M., Weber, P. K., Pett-Ridge, J., Thelen, M.P., 2016. Cyanobacterial reuse of extracellular organic carbon in microbial mats. ISME J. 10, 1240–1251.
Swofford, D.L., 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0 b10.
Taboada, B., Estrada, K., Ciria, R., Merino, E., 2018. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. Bioinformatics 34, 4118–4120.
Tekaia, F., 2016. Inferring orthologs: open questions and perspectives. Genomics Insights 9, 17–28.
Thompson, J.N., 1999. The evolution of species interactions. Science 284 (5423), 2116–2118.
Tooming-Klunderud, A., Sogge, H., Rounge, T.B., Nederbragt, A.J., Lagesen, K., Glöckner, G., et al., 2013. From green to red: horizontal gene transfer of the phycoerythrin gene cluster between planktothrix strains. Appl. Environ. Microbiol. 79, 6803–6812.
U'Ren, J.M., Lutzoni, F., Miadlikowska, J., Arnold, A.E., 2010. Community analysis reveals close affinities between endophytic and endolichenic fungi in mosses and lichens. Microb. Ecol. 60 (2), 340–353.
Voß, B., Bolhuis, H., Fewer, D.P., Kopf, M., Möke, F., Haas, F., El-Shehawy, R., Hayes, P., Bergman, B., Sivonen, K., Dittmann, E., 2013. Insights into the physiology and ecology of the brackish-water-adapted cyanobacterium *Nodularia spumigena* CCY9414 based on a genome-transcriptome analysis. PLoS ONE 8 (3), e60224.
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 (5), 1160–1175.
Xi, Z., Rest, J.S., Davis, C.C., 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. PLoS ONE 8 (11), e80870.
Zhang, J., Fu, X.X., Li, R.Q., Zhao, X., Liu, Y., Li, M.H., Zwaenepoel, A., Ma, H., Goffinet, B., Guan, Y.L., Xue, J.Y., 2020. The hornwort genome and early land plant evolution. Nat. Plants 6 (2), 107–118.
Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., Papke, R.T., 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16, 1099–1108.
Zolan, M.E., Pukkila, P.J., 1986. Inheritance of DNA methylation in *Coprinus cinereus*. Mol. Cell. Biol. 6 (1), 195–200.