Metagenomic assembly of new (sub)polar Cyanobacteria and their associated microbiome from non-axenic cultures

Luc Cornet,1,2 Amandine R. Bertrand,1,3 Marc Hanikenne,3 Emmanuelle J. Javaux,2 Annick Wilmotte4 and Denis Baurain1,*

Abstract
Cyanobacteria form one of the most diversified phyla of Bacteria. They are important ecologically as primary producers, for Earth evolution and biotechnological applications. Yet, Cyanobacteria are notably difficult to purify and grow axenically, and most strains in culture collections contain heterotrophic bacteria that were probably associated with Cyanobacteria in the environment. Obtaining cyanobacterial DNA without contaminant sequences is thus a challenging and time-consuming task. Here, we describe a metagenomic pipeline that enables the easy recovery of genomes from non-axenic cultures. We tested this pipeline on 17 cyanobacterial cultures from the BCCM/ULC public collection and generated novel genome sequences for 12 polar or subpolar strains and three temperate ones, including three early-branching organisms that will be useful for phylogenomics. In parallel, we assembled 31 co-cultivated bacteria (12 nearly complete) from the same cultures and showed that they mostly belong to Bacteroidetes and Proteobacteria, some of them being very closely related in spite of geographically distant sampling sites.

DATA SUMMARY
1. Cornet L and Baurain D, National Center for Biotechnology Information (NCBI) BioProject, accession PRJNA436342 (2018).
2. Cornet L and Baurain D, National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), accessions SAMN08623419 to SAMN08623435 (2018).
3. Cornet L and Baurain D, National Center for Biotechnology Information (NCBI) DDBJ/ENA/GenBank, accessions QBSL00000000 to QBLZ00000000 and QBMA00000000 to QBMS00000000 (2018).
4. Cornet L and Baurain D, National Center for Biotechnology Information (NCBI) BioSample, accessions SAMN08895976 to SAMN0889600 (2018).

INTRODUCTION
Cyanobacteria, also called blue-green algae, are an intensively studied group of prokaryotes. This focus is notably due to their ecological importance, as they colonize a very diverse range of ecosystems and are a major component of the phytoplankton [1, 2]. They are also of primary interest in terms of evolution and palaeobiogeology, cyanobacteria having been present on Earth since the Proterozoic [3–5]. Emergence of oxygenic photosynthesis in this phylum, which led to the Great Oxygenation Event (GOE) around 2.4 billion years ago, had a critical impact on early Earth and evolution by increasing the level of free oxygen and subsequently creating new ecological niches [6–8]. Moreover, Cyanobacteria played a role in another major biological event, the spread of photosynthesis to eukaryotic lineages through an initial endosymbiosis termed ‘primary’, followed by several higher-order endosymbioses [9]. Finally, Cyanobacteria produce a large number of bioactive compounds (e.g. alkaloids, non-ribosomal peptides, polyketides), which make them promising for both biotechnological and biomedical applications [10–12]. The generation of an axenic cyanobacterial culture is notoriously difficult [1], especially for polar strains [13], and hence the need for tedious purification protocols [14]. In consequence, all cyanobacterial culture collections include...
many non-axenic cultures (e.g. American Type Culture Collection, ATCC; Czech Collection of Algae and Cyanobacteria, CCALA; University of Toronto Culture Collection of Algae and Cyanobacteria, UTCC; Culture Collection of Algae at the University of Texas, UTEX), with the notable exception of the Pasteur Culture Collection of Cyanobacteria, PCC. The difficulty of reaching axenicity results from bacterial communities living in close relationship with *Cyanobacteria* in nature. This microbiome has been described both from environmental samples [15–19] and from non-axenic cultures [20–22]. Moreover, *Bacterial* *Cyanobacteria* associations appear to be stable in culture, as no significant differences could be found between bacterial communities accompanying *Cyanobacteria* in fresh samples and collection cultures [21]. Complex trophic interactions between *Cyanobacteria* and other bacterial phyla feeding on their sheaths, such as *Proteobacteria* and *Bacteroidetes*, have been described [23], as well as specific interactions, such as adhesion to heterocysts [20]. The presence of these bacterial communities consequently limits the use of non-axenic cyanobacterial cultures for genomic applications, because fragments of their genomes can eventually become part of published cyanobacterial genomes. Hence, we have recently shown that a large proportion (52%) of publicly available genomes of *Cyanobacteria* are contaminated by such foreign sequences [24]. In 5% of the surveyed genomes, these non-cyanobacterial contaminants even reach up to 41.5% of the genome sequences deposited in the databases.

Owing to their clear scientific interest, obtaining authentic genome sequences of *Cyanobacteria* is an important issue. During the last decade, the rise of metagenomics has allowed an ever-better separation of the different components of a mixture of organisms, based on various properties of the metagenomic contigs, e.g. sequencing coverage and oligonucleotide signatures [25]. In this work, we use a straightforward pipeline that enables the efficient isolation of cyanobacterial genomes from non-axenic cultures. Easy to set up, this pipeline is composed of state-of-the-art metagenomic tools, metaSPAdes [26], MetaBAT [27], CheckM [28], followed by DIAMOND BLASTX analyses [29] and SSPACE [30] scaffolding. This pipeline allowed us to assemble 15 novel cyanobacterial genomes (12 high-quality, two medium-quality and one low-quality) from 17 polar, subpolar and temperate cultures of the BCCM/ULC public culture collection hosted by the University of Liège (Belgium), of which three appear to belong to early-branching strains in the cyanobacterial tree of life. In the process, we also characterized 31 different co-cultivated bacteria out of the 17 cyanobacterial cultures. Those contaminants’ organisms mostly belong to *Proteobacteria* and *Bacteroidetes*, and some of them are very closely related to each other. Finally, we investigated why small subunit (SSU) rRNA (16S) genes are often lost during metagenomic binning and developed a new metric to compare genome bins with different levels of completeness.

**IMPACT STATEMENT**

Complete genomes of cold-adapted *Cyanobacteria* are underrepresented in databases, due to the difficulty of growing them axenically. In this work, we report the genome sequencing of 12 (sub)polar and three temperate *Cyanobacteria*, along with 21 *Proteobacteria* and five *Bacteroidetes* recovered from their microbiome. Following the use of a state-of-the-art metagenomic pipeline, 12 of our new cyanobacterial genome assemblies are of high quality, which indicates that even non-axenic cultures can yield complete genomes suitable for phylogenomics and comparative genomics. Beyond this main theme, we also address two methodological issues in self-standing Supplemental Appendices. Firstly, we investigated the fate of small subunit rRNA (16S) genes during metagenomic binning and observe that multi-copy rRNA operons are lost because of their higher sequencing coverage and divergent tetranucleotide frequencies. Secondly, we devised a measure of genomic identity to compare metagenomic bins of different completeness, which allowed us to show that *Cyanobacteria*-associated bacteria can be closely related in spite of considerable geographical distance between collection points.

**METHODS**

**Cyanobacterial cultures and DNA extraction**

The 17 cyanobacterial cultures were selected in order to sequence new genomes of interesting Arctic and Antarctic organisms, the biodiversity of which is still not well known. All the strains used in this study were indeed collected from (sub)polar regions, with the exception of three Belgian strains, ULC335, added to the sequencing batch to obtain the first genome of the genus *Snowella*, and ULC186 and ULC187, both related to the (sub)polar strains but of temperate origin. All the *Cyanobacteria* from the present study are from freshwater. The cultures (deposited in the BCCM/ULC collection during the period 2011–2014; Table 1) were incubated at 15 °C in BG11 or BG110 medium and exposed to a constant white fluorescent light source (about 40 μmol photons m⁻² s⁻¹) for 4 weeks. DNA was extracted using the GenElute Bacterial Genomic DNA kit (Sigma-Aldrich) following the recommendations of the manufacturer. After control of the integrity of the genomic DNA by electrophoresis and quantification of the dsDNA concentration using the Quan-iT Picogreen dsDNA Assay kit (Thermo Fisher Scientific), a minimum of 1 μg of dsDNA was sent to the sequencing platform.

**Metagenome sequencing and assembly**

The 17 cyanobacterial cultures were sequenced (PE 2 × 250 nt) on the Illumina MiSeq sequencing platform (GIGA Genomics, University of Liège). Nextera XT libraries had a fragment size estimated at 800–900 nt. Raw
sequencing reads were trimmed using Trimmomatic v0.35 [31]. Sequencing adapters were removed with the option illuminaclip NexteraPE-PE.fa : 2 : 30 : 20. Trimming values were selected to maximize genome bin sizes (in terms of bp), after preliminary testing. Trailing/leading values were set at 20, the sliding window at 10 : 20, the crop value at 145 and the minimal length at 80. Trimmed paired-end reads were assembled with metaSPAdes v3.10.1 [26] using default settings. Trimmed paired-end reads were then re-mapped on the metaSPAdes assemblies with BamM v1.7.3 (http://ecogenomics.github.io/BamM/), yielding BAM files suitable for the metagenomic analyses. Genome bins were determined with MetaBAT v0.30.1 [27], trying each built-in parameter set in turn (i.e. verysensitive, sensitive, specific, veryspecific and superspecific). CheckM v1.0.7 [28] was then used with the option lineage_wf to select the best MetaBAT parameter set for each metaSPAdes assembly. In practice, we first tried to select the MetaBAT parameter set that was the most suitable for the largest genome bin of a given metagenome (in terms of total assembly length),

Table 1. Details of the ULC strains

| Assembly | Strain | Name | Type | Prior affiliation | Morphology | Sheath | Deposit date | Habitat | Culture medium | Temperature (°C) |
|----------|--------|------|------|------------------|------------|--------|--------------|---------|----------------|------------------|
| QBLS000000000 | ULC187 | Pseudanabaena sp. FW039 | Non-axenic | Clade F | Filamentous | No | 2012 | Belgium, lake Ri Jaune | BG11 | RT |
| QBM000000000 | ULC066 | Pseudanabaena frigida O-155 | Non-axenic | Clade F | Filamentous | No | 2011 | Canadian Arctic, Bylot Island | BG11 | 12 |
| QBMK000000000 | ULC068 | Pseudanabaena sp. O-202 | Non-axenic | Clade F | Filamentous | No | 2011 | Canadian Subarctic, Québec, Kuujjuaq | BG11 | 12 |
| QBM000000000 | ULC065 | Cyanobium sp. O-154 | Non-axenic | Clade C1 | Unicellular | No | 2011 | Canadian Arctic, Bylot Island | BG11 | 12 |
| QBMG000000000 | ULC082 | Cyanobium sp. Chester Cone | Non-axenic | Clade C1 | Unicellular | No | 2011 | Antarctic, Livingston Island | BG11 | 12 |
| QBMF000000000 | ULC084 | Cyanobium sp. Laguna Chica | Non-axenic | Clade C1 | Unicellular | No | 2011 | Antarctic, Livingston Island | BG11 | 12 |
| QBMH000000000 | ULC077 | Leptolyngbya sp. O-157 | Non-axenic | Clade C3 | Filamentous | No | 2011 | Canadian Arctic, Bylot Island | BG11 | 12 |
| QBMQ000000000 | ULC007 | Phormidium priestleyi ANT.LHS52.4 | Axenic | Clade C3 | Filamentous | No | 2011 | Antarctic, Larsemann Hills | BG11 | 18 |
| NA | ULC165 | Leptolyngbya sp. OTC1/1 | Non-axenic | Clade C3 | Filamentous | Yes | 2012 | Antarctica, Sor Rondane Mountains | BG11 | 12 |
| QBM000000000 | ULC129 | Leptolyngbya fovealarum TM2FOS129 | Non-axenic | Clade C3 | Filamentous | No | 2011 | Antarctic, Transantarctic Mountains | BG11 | 12 |
| QBM000000000 | ULC027 | Phormidium priestleyi ANT. PROGRESS2.5 | Non-axenic | Clade C3 | Filamentous | No | 2011 | Antarctic, Transantarctic Mountains | BG11 | 18 |
| QBLT000000000 | ULC186 | Leptolyngbya sp. FW074 | Non-axenic | Clade C3 | Filamentous | No | 2012 | Belgium, Renipont lake | BG11 | RT |
| QBMN000000000 | ULC041 | Leptolyngbya antarctica ANT.ACE.1 | Non-axenic | Clade C3 | Filamentous | No | 2011 | Antarctic, Vestfold Hills | BG11 | 12 |
| QBMJ000000000 | ULC073 | Leptolyngbya glacialis TM1FOS73 | Non-axenic | Clade C3 | Filamentous | Yes | 2011 | Antarctic, Transantarctic Mountains | BG11 | 18 |
| QBS000000000 | ULC335 | Snowella sp. FW024 | Non-axenic | Clade B2 | Unicellular | Yes | 2014 | Belgium, lake Falemprise | BG11 | RT |
| NA | ULC146 | Nostoc sp. ANT. UTS.183 | Non-axenic | Clade B1 | Filamentous heterocystous | Yes | 2012 | Antarctica, Sor Rondane Mountains | BG110 | 18 |
| NA | ULC179 | Nostoc sp. OTCControl | Non-axenic | Clade B2 | Filamentous heterocystous | Yes | 2012 | Antarctica, Sor Rondane Mountains | BG110 | 12 |
considering CheckM output statistics in the following order: (1) contamination, (2) strain heterogeneity and (3) completeness. When multiple parameter sets were equally optimal for the largest bin, we turned to the next-largest bin(s) for parameter selection. The non-assignment of a given contig to multiple bins was checked using the unique option of CheckM, while binning accuracy was assessed using the merge and tree_qa options after generating a marker set for Bacteria. The automatic taxonomic classification of CheckM was then extracted to determine the nature of each bin, either cyanobacterial or foreign. The strain names of cyanobacterial bins were attributed based on phenotypic observations during cultivation. Bins classified as root (i.e. unclassified) by CheckM were discarded from phylogenomic analyses. Contaminants (with respect to the taxon determined by CheckM) in each genome bin were further characterized using DIAMOND BLASTX v0.8.22 [29] and the companion parser developed in our article regarding the contamination of public cyanobacterial genomes [24]. To this end, we split the genome bins into non-overlapping pseudo-reads of 250 nt (with a custom Perl script), so as to increase the sensitivity of the analyses. We then used DIAMOND BLASTX to blast these pseudo-reads against a curated database derived from the release 30 of Ensembl Bacteria that we developed for our genome contamination analyses [24]. In parallel, contigs within each genome bin were scaffolded with SSPACE v3.0 [30] using default settings, except that contigs were first extended using paired-end reads (-x 1) and that the minimum of read pairs required to compute a scaffold was set to 3 (-k 3). The fragmentation of the scaffolded genome bins was then analysed with QUAST v2.3 [32] using default settings, whereas their sequencing coverage was determined with BBMap v37.24 (http://bbmap.sourceforge.net/). Finally, protein sequences were predicted for all genome bins with Prodigal v2.6.2 [33] using the ab initio mode. In Appendix S1, we provide the stepwise tutorial describing the set up and use of the metagenomic pipeline.

Phylogenetic analyses

The complete proteomes of 64 cyanobacterial strains chosen to represent the diversity of the whole phylum were downloaded from the NCBI portal [34]. Details and download links for the selected proteomes are available in Tables 2 and S1 (available in the online version of this article), respectively. Orthology inference was performed with USEARCH v8.1 (64 bits) [35] and OrthoFinder v1.1.2, using the standard inflation parameter of 1.5 [36]. Out of 37 261 orthologous groups (OGs), 675 were selected with classify-ali.pl (part of the Bio-MUST-Core software package; [45]). Visual inspection of the tree inferred from this smaller dataset led us to further discard four fast-evolving organisms, yielding a total of 196 representative organisms. Both the large (3474 organisms) and the small (196 organisms) datasets were used in subsequent analyses. Hence, the 53 alignments (both large and small versions) were enriched (using again ‘42’) with sequences from the foreign (i.e. non-cyanobacterial) bins assembled from our 17 cyanobacterial cultures (31 bins in total, excluding unclassified CheckM bins). To control the origins of the enriching sequences, taxonomic filters of ‘42’ were enabled, so as to require all new sequences to belong to the taxon determined by CheckM during its analysis of each whole bin. After this step, four incomplete genome bins (ULC066-bin3, ULC073-bin4, ULC082-bin4, ULC146-bin6) were discarded due to their low prevalence in the alignments (<10 %). Enriched alignments were then processed as above with either ali2phylip.pl (large dataset) or BMGE (small dataset). The two resulting supermatrices assembled with SCAFoS contained 3501 organisms × 6613 unambiguously aligned amino-acid positions (6.0 % missing character states) and 223 organisms × 7060 unambiguously aligned amino-acid positions (7.8 % missing character states), respectively. Finally, two different trees were inferred using either RAxML (large dataset) or PhyloBayes (small dataset). The trees were rooted on the branch leading to Archaea.
Table 2. Details regarding reference proteomes
All details were extracted from the NCBI metadata.

| Assembly          | Bioproject     | Taxid       | Name                                                   |
|-------------------|----------------|-------------|--------------------------------------------------------|
| GCA_000484535.1   | PRJNA162637    | 1183438     | Gloeobacter kilaeuensis JS1                           |
| GCF_000011385.1   | PRJNA58011     | 251221      | Gloeobacter violaceus PCC 7421                        |
| GCF_000013205.1   | PRJNA224116    | 321327      | Synechococcus sp. JA-3-3Ab                            |
| GCF_0003322275.1  | PRJNA224116    | 195250      | Synechococcus sp. PCC 7336                            |
| GCF_000317065.1   | PRJNA224116    | 82654       | Pseudanabaena sp. PCC 7367                           |
| GCF_000332215.1   | PRJNA224116    | 927668      | Pseudanabaena biceps PCC 7429                        |
| GCF_000317085.1   | PRJNA224116    | 1173263     | Synechococcus sp. PCC 7502                            |
| GCF_000332175.1   | PRJNA224116    | 118173      | Pseudanabaena sp. PCC 6802                            |
| GCF_000018105.1   | PRJNA224116    | 329726      | Acaryochloris marina MBIC11017                        |
| GCA_000022045.1   | PRJNA224116    | 395961      | Cyanothece sp. PCC 7425                              |
| GCF_000505665.1   | PRJNA224116    | 1394889     | Thermosynechococcus sp. NK55a                         |
| GCF_000316885.1   | PRJNA224116    | 195253      | Synechococcus sp. PCC 6312                            |
| GCF_000775285.1   | PRJNA224116    | 1497020     | Neosynechococcus sphyngicola sy1                      |
| GCF_000309945.1   | PRJNA224116    | 864702      | Oscillatoriales cyanobacterium JSC-12                 |
| GCF_001895925.1   | PRJNA224116    | 1920490     | Phormidemis priestleyi ULC007                        |
| GCF_001650195.1   | PRJNA224116    | 1850361     | Phormidemis priestleyi BC1401                        |
| GCF_000353285.1   | PRJNA224116    | 272134      | Leptolyngbya irionana PCC 6306                        |
| GCF_000333415.1   | PRJNA224116    | 1487953     | Leptolyngbya sp. ISC-1                               |
| GCF_000332095.2   | PRJNA224116    | 1173264     | Leptolyngbya sp. PCC 6406                            |
| GCF_000763385.1   | PRJNA224116    | 1229172     | Leptolyngbya sp. KIOST-1                             |
| GCF_000309385.1   | PRJNA224116    | 118166      | Nodosilleana nodolosa PCC 7104                        |
| GCF_000155595.1   | PRJNA224116    | 91464       | Synechococcus sp. PCC 7335                           |
| GCF_000482245.1   | PRJNA224116    | 1385935     | Leptolyngbya sp. Heron Island J                       |
| GCF_000316115.1   | PRJNA224116    | 102129      | Leptolyngbya sp. PCC 7375                            |
| GCF_000464785.1   | PRJNA224116    | 1255374     | Planctothrix rubescens NIVA-CYA 407                   |
| GCF_000175415.3   | PRJNA224116    | 634502      | Arthropira platensis str. Paraca                      |
| GCF_000478195.2   | PRJNA224116    | 1348334     | Lyngbya aestuarii BL J                               |
| GCF_000832155.1   | PRJNA224116    | 402777      | Kamptonea formosum PCC 6407                          |
| GCF_000317475.1   | PRJNA224116    | 179408      | Oscillatoria nigro-viridis PCC 7112                   |
| GCF_000317105.1   | PRJNA224116    | 56110       | Oscillatoria acuminatea PCC 6304                      |
| GCF_000317515.1   | PRJNA224116    | 1173027     | Microcoleus sp. PCC 7113                             |
| GCF_000021825.1   | PRJNA224116    | 65393       | Cyanothecace sp. PCC 7424                            |
| GCA_000307995.2   | PRJEA88171     | 1160280     | Microcystis aeruginosa PCC 9432                       |
| GCF_000021805.1   | PRJNA224116    | 41431       | Cyanothecace sp. PCC 8801                            |
| GCF_000737945.1   | PRJNA256120    | 1527444     | Candida tarda Atelecyanobacterium thalassa isolate SIO64986 |
| GCF_000284135.1   | PRJNA224116    | 1080228     | Synechocystis sp. PCC 6803 substr. GT-1              |
| GCF_000715475.1   | PRJNA224116    | 490193      | Synechococcus sp. NKBG042902                         |
| GCF_000317655.1   | PRJNA39697     | 292563      | Cyanobacterium staniier PCC 7202                     |
| GCF_000332055.1   | PRJNA224116    | 102125      | Xenococcus sp. PCC 7305                              |
| GCF_000317575.1   | PRJNA224116    | 111780      | Stanieria cyanosaqua PCC 7437                        |
| GCF_000380225.1   | PRJNA224116    | 1128427     | filamentous cyanobacterium ESFC-1                    |
| GCF_000317615.1   | PRJNA224116    | 13035       | Dactylococcopsis salina PCC 8305                     |
| GCF_000317495.1   | PRJNA224116    | 1173022     | Crinalium epipsammum PCC 9333                        |
| GCF_000317355.1   | PRJNA224116    | 1173026     | Gloeocapsa sp. PCC 7428                              |
| GCF_000317125.1   | PRJNA224116    | 251229      | Chroococcidiopsis thermals PCC 7203                  |
| GCF_000582685.1   | PRJNA224116    | 1469607     | [Scytonema hofmannii] UTEX 2349                      |
| GCF_00079435.1    | PRJNA224116    | 1532906     | Anaphysomonas flas-aquae 2012/KM1/D3                 |
| GCF_000196515.1   | PRJNA224116    | 551115      | 'Nostoc azollae' 0708                               |
| GCF_000316645.1   | PRJNA224116    | 28072       | Nostoc sp. PCC 7524                                 |
| GCF_000204075.1   | PRJNA10642     | 240292      | Anabaena variabilis ATCC 29413                       |
Table 2. cont.

| Assembly | Bioproject | Taxid    | Name                      |
|----------|------------|----------|---------------------------|
| GCA_000340565.3 | PRJNA185469 | 313624   | Nodularia spumigena CCY9414 |
| GCF_000020025.1 | PRJNA224116 | 63737    | Nostoc punctiforme PCC 73102 |
| GCF_000332295.1 | PRJNA224116 | 64347    | Forteia contorta PCC 7126 |
| GCF_000346485.2 | PRJNA224116 | 128403   | Syctonema holmianum PCC 7110 |
| GCF_000734895.2 | PRJNA224116 | 1337936  | Calothrix sp. 336/3 |
| GCF_000332255.1 | PRJNA224116 | 1173021  | cyanobacterium PCC 7702 |
| GCF_000317225.1 | PRJNA224116 | 98439    | Fischerella thermals PCC 7521 |
| GCF_000012525.1 | PRJNA224116 | 1160     | Synechococcus elongatus PCC 7942 |
| GCF_000586015.1 | PRJNA224116 | 1451353  | Candidatus Synechococcus spongiarum SH4 |
| GCF_000155635.1 | PRJNA224116 | 180281   | Cyanobium sp. PCC 7001 |
| GCA_000015705.1 | PRJNA13496   | 59922    | Prochlorococcus marinus str. MIT 9303 |
| GCF_000011485.1 | PRJNA224116 | 74547    | Prochlorococcus marinus str. MIT 9313 |
| GCF_000153805.1 | PRJNA224116 | 313625   | Synechococcus sp. BL107 |

All phylogenetic trees were formatted using the script format-tree.pl (part of Bio-MUST-Core), FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) and further arranged in Inkscape v0.92 [46].

**SSU rRNA (16S) analyses**

SSU rRNA (16S) genes were predicted using RNAmmer v1.2 [47] in all genome bins for the selected MetaBAT parameter set. Beyond regular bins, we also investigated an additional bin (called nobin) for each metagenome, which contained all the scaffolds rejected by MetaBAT during the binning process. Predicted rRNA sequences were taxonomically classified by SINA v1.2.11 [48], using release 128 of the SILVA database composed of 1,922,213 SSU rRNA reference sequences [49].

**RESULTS**

**Metagenome sequencing and assembly**

We obtained a total of 55 different genome bins from the separate sequencing and metagenomic assembly of the 17 cyanobacterial cultures (Table 3). Among those, we identified 15 bins as cyanobacterial (ULC007-bin1, ULC027-bin1, ULC041-bin1, ULC065-bin1, ULC066-bin1, ULC068-bin1, ULC073-bin1, ULC077-bin1, ULC082-bin1, ULC084-bin3, ULC129-bin1, ULC165-bin4, ULC186-bin1, ULC187-bin1, ULC335-bin1), based on CheckM classification [28], except for ULC165-bin4, which was classified after DIAMOND BLASTX results. For the two Nostocales strains (ULC146 and ULC179), we failed to recover any cyanobacterial bin (but see below for the analysis of the other bins). For 12 metagenomes, the cyanobacterial bin corresponded to the largest predicted bin, in terms of both total length and sequencing coverage (Table 3; see also Appendix S2). For two cultures, however, cyanobacterial bins were the smallest predicted (ULC084-bin3 and ULC165-bin4). Genome completeness, evaluated with CheckM, was ≥90 % [median=97.74 %, interquartile range (IQR)=4.04 %] for all cyanobacterial bins but lower for ULC165-bin4 (24.14 %). As expected, completeness correlated positively with the sequencing coverage of the bins in the metagenomic assemblies, but this correlation was barely significant (Pearson’s r=0.52, P=0.05). The contamination level was evaluated to be <1.63 % (median=0.47 %, IQR=0.83 %) with CheckM and <2.62 % (median=1.26 %, IQR=0.40 %) with our DIAMOND BLASTX parser [24]. As our libraries were only composed of paired-ends (and not of mate pairs), the number of scaffolds obtained after metaspades assembly and SSPACE scaffolding was ≥60 for all cyanobacterial genome bins (median=238, IQR=292) (Tables 3 and S2).

Altogether, we identified 40 bins that were not of cyanobacterial origin out of our 17 cyanobacterial cultures. Among these foreign genome bins, we classified 21 as Proteobacteria and five as Bacteroidetes, and thus 26 bins contained organisms belonging to two bacterial phyla known to participate in the cyanobacterial microbiome [21, 50]. The remaining 14 bins could only be classified as Bacteria (five) or were left unclassified (nine) by CheckM. While unclassified bins were discarded from subsequent analyses, bins identified at the Bacteria level were retained. Genome completeness of these 31 bacterial bins was very heterogeneous (median=71.96 %, IQR=51.84 %). As for cyanobacterial bins, but more significantly, completeness correlated positively with sequencing coverage, lowly covered bins being the less complete (Pearson’s r=0.52, P=0.007). Nevertheless, we managed to recover 13 nearly complete foreign bins (completeness ≥90 %). According to CheckM, the contamination level (foreign sequences not belonging to the taxonomic label of the bin under study) of the 26 classified non-cyanobacterial bins was always <9.28 % (median=0.8 %, IQR=1.13 %), except for ULC179-bin1 (60.19 %). The contamination level of the bins classified as Bacteria was not recorded, because such a high taxonomic rank made its evaluation meaningless. As for cyanobacterial bins, the number of scaffolds of the 31 bacterial bins remained quite high (>53, median=232, IQR=205). In spite of three cases of possible complementar-
Table 3. Assembly statistics, taxonomy, completeness, contamination and coverage of genome bins

The taxonomic label (CM taxon), genome completeness (CM compl.) and contamination level (CM contam.) were computed with CheckM. Sequencing coverage (med) was computed with BBMap, while bin length was extracted from QUAST output. Length (%) represents the proportion of assembled data in a bin with respect to the total amount of data of the corresponding metagenome. In the Nature column, cyanobacterial bins are denoted by C, microbiome bins by M, unclassified bins by U and nobins by No. Genome bins used in phylogenetic inference are marked by an asterisk (*) and discarded bins by a dash (—). NA, not applicable.

| Strain  | MetaBAT setting | Bin | CM taxon   | Nature | No. of scaffolds | Length (%) | Coverage (med) | CM compl. | CM contam. |
|---------|----------------|-----|------------|--------|-----------------|------------|----------------|-----------|------------|
| ULC335  | Veryspecific   | 1   | Cyanobacteria* | C      | 238             | 20.84      | 10.90          | 98.91     | 0.51       |
|         |                | 2   | Flavobacteriaceae* | M      | 67              | 13.73      | 11.12          | 99.29     | 0.12       |
|         |                | 3   | Bacteroidetes*  | M      | 576             | 12.83      | 4.46           | 65.45     | 0.49       |
|         |                | 4   | Alphaproteobacteria* | M      | 271             | 4.79       | 4.13           | 32.28     | 0          |
|         |                | 0   | Nobin | No      | 23 056          | 47.81      | 1.88           | NA        | NA         |
| ULC007  | Superspecific  | 1   | Cyanobacteria* | C      | 84              | 91.14      | 26.62          | 98.11     | 0          |
|         |                | 2   | Unclassified | U      | 12              | 4.95       | 72.12          | 0         | 0          |
|         |                | 0   | Nobin | No      | 358             | 3.91       | 1.48           | NA        | NA         |
| ULC027  | Verysensitive  | 1   | Cyanobacteria* | C      | 439             | 21.40      | 6.27           | 90.43     | 0.27       |
|         |                | 2   | Alphaproteobacteria* | M      | 190             | 16.16      | 7.71           | 95.02     | 1.16       |
|         |                | 3   | Sphingomonadales* | M      | 293             | 12.03      | 6.18           | 60.21     | 2.35       |
|         |                | 4   | Unclassified | U      | 164             | 4.16       | 5.09           | 4.17      | 0          |
|         |                | 0   | Nobin | No      | 24 364          | 46.24      | 1.89           | NA        | NA         |
| ULC041  | Verysensitive  | 1   | Cyanobacteria* | C      | 287             | 84.76      | 31.38          | 96.2      | 1.63       |
|         |                | 2   | Unclassified | U      | 24              | 9.36       | 44.33          | 0         | 0          |
|         |                | 0   | Nobin | No      | 441             | 5.88       | 3.97           | NA        | NA         |
| ULC065  | Veryspecific   | 1   | Cyanobacteria* | C      | 95              | 22.36      | 38.37          | 99.09     | 0.27       |
|         |                | 2   | Xanthomonadaeae* | M      | 332             | 19.33      | 6.19           | 83.73     | 1.23       |
|         |                | 0   | Nobin | No      | 20 555          | 58.31      | 1.73           | NA        | NA         |
| ULC066  | Superspecific  | 1   | Cyanobacteria* | C      | 67              | 28.81      | 21.86          | 98.82     | 0.47       |
|         |                | 2   | Bacteroidetes*  | M      | 401             | 13.94      | 4.93           | 76.91     | 1.23       |
|         |                | 3   | Betaproteobacteria* | M      | 152             | 2.86       | 3.48           | 15.86     | 0          |
|         |                | 0   | Nobin | No      | 24 558          | 54.38      | 1.69           | NA        | NA         |
| ULC068  | Superspecific  | 1   | Cyanobacteria* | C      | 60              | 57.04      | 29.34          | 97.09     | 0.71       |
|         |                | 2   | Unclassified | U      | 3               | 2.56       | 22.60          | 0         | 0          |
|         |                | 0   | Nobin | No      | 10 385          | 40.41      | 1.42           | NA        | NA         |
| ULC073  | Verysensitive  | 1   | Cyanobacteria* | C      | 476             | 22.70      | 10.74          | 92.03     | 1.42       |
|         |                | 2   | Betaproteobacteria* | M      | 65              | 16.26      | 7.99           | 97.92     | 0.67       |
|         |                | 3   | Sphingomonadales* | M      | 603             | 15.78      | 4.94           | 70.57     | 5.3        |
|         |                | 4   | Bacteria*—     | M      | 156             | 2.79       | 4.39           | 10.71     | 0          |
|         |                | 5   | Unclassified | U      | 26              | 1.40       | 15.02          | 0         | 0          |
|         |                | 6   | Unclassified | U      | 29              | 1.38       | 6.45           | 0         | 0          |
|         |                | 0   | Nobin | No      | 16 790          | 39.68      | 1.94           | NA        | NA         |
| ULC077  | Veryspecific   | 1   | Cyanobacteria* | C      | 407             | 47.37      | 15.08          | 97.64     | 0.47       |
|         |                | 0   | Nobin | No      | 14 903          | 52.63      | 1.83           | NA        | NA         |
| ULC082  | Veryspecific   | 1   | Cyanobacteria* | C      | 124             | 11.49      | 19.85          | 97.74     | 0.27       |
|         |                | 2   | Bacteria*      | M      | 529             | 9.77       | 4.50           | 62.77     | 7.54       |
|         |                | 3   | Bacteria*      | M      | 542             | 8.16       | 3.88           | 46.21     | 9.28       |
|         |                | 4   | Bacteria*—     | M      | 120             | 1.72       | 4.73           | 11.13     | 0          |
|         |                | 5   | Unclassified | U      | 74              | 1.67       | 4.57           | 0         | 0          |
|         |                | 0   | Nobin | No      | 30 077          | 67.18      | 2.15           | NA        | NA         |
| ULC084  | Superspecific  | 1   | Betaproteobacteria* | M      | 232             | 23.15      | 5.67           | 93.61     | 1.73       |
|         |                | 2   | Alphaproteobacteria* | M      | 222             | 22.39      | 6.65           | 92.46     | 1.38       |
|         |                | 3   | Cyanobacteria* | C      | 116             | 21.88      | 20.78          | 98.55     | 0          |
|         |                | 0   | Nobin | No      | 10 835          | 32.58      | 1.59           | NA        | NA         |
| ULC129  | Verysensitive  | 1   | Cyanobacteria* | C      | 299             | 38.35      | 18.46          | 98.64     | 0.77       |
|         |                | 0   | Nobin | No      | 21 968          | 61.65      | 1.62           | NA        | NA         |
| ULC146  | Superspecific  | 1   | Burkholderiales* | M      | 177             | 16.18      | 10.96          | 96.57     | 0.93       |
first involving unclassified bins, the corresponding bins were not merged because CheckM phylogenetic placement was never congruent. Details about genome bins are available in Table S2. We released scaffolded assemblies and protein predictions for all the bins having a completeness $\geq 90\%$, whether classified as cyanobacterial (14) or probable microbiome organisms (13).

**Cyanobacterial phylogenomics**

A phylogenomic analysis based on 675 genes and 64 reference *Cyanobacteria* showed that three cyanobacterial bins (i.e. excluding ULC335) were situated in the basal part of the cyanobacterial tree, here defined as clades G, F and E [51] (Fig. 1). Clade C, mainly composed of *Leptolyngbya* species and picoplanktonic *Cyanobacteria*, contains 11 cyanobacterial bins (Fig. 1). Statistical support [Bayesian posterior probability (PP)] was maximal except for three nodes. In the following, we refer to the cyanobacterial clades using the nomenclature defined by Shih *et al.* [52], since theirs was the first to fully sample the cyanobacterial morphological diversity (i.e. Sections I–V from [1]). Three ULC strains (*Pseudanabaena* sp. ULC187, *Pseudanabaena frigida* ULC066 and *Leptolyngbya* sp. ULC068) are located at a very basal (i.e. 'early-branching') position in clade F, and form a cluster with the reference strain *Pseudanabaena biceps* PCC 7429. Three other strains, identified as *Cyanobium* sp. (ULC065, ULC082 and ULC084), emerge together from the picocyanobacteria clade C1. Although their C1 membership is indisputable, the exact branching point within clade C1 is not resolved (PP=0.51). The six *Leptolyngbya* strains (*Leptolyngbya* sp. ULC077/ULC165/ULC186, *L. antarctica* ULC041, *L. glacialis* ULC073 and *L. foveolarum* ULC129) and the two *Phormidium*/*Phormidium priestleyi* strains (ULC007 and ULC027) are located in clade C3, mainly composed of reference *Leptolyngbya* strains. While two strains (*Leptolyngbya* sp. ULC077 and ULC165) each form an additional single branch within clade C3, five other strains emerge as two new sub-groups: *Leptolyngbya foveolarum* ULC129 and *Phormidium priestleyi* ULC027 on the one hand (yet weakly supported; PP=0.51), and *Leptolyngbya* sp. ULC186, *Leptolyngbya antarctica* ULC041 and *Leptolyngbya glacialis* ULC073 on the other. As expected, our new assembly of *Phormidmsis/Phormidium priestleyi* strains (ULC007) is extremely close to the first release of the same genome (*Phormidmsis priestleyi* ULC007 GCF_001895925.1), which we used as positive control for our pipeline [53]. Finally, *Snowella* sp. ULC335 is part of clade B2, composed of various cyanobacterial genera from the orders *Pleurocapsales* and *Chroococcales* [54]. This strain branches with *Synechocystis* sp. PCC 6803, which is among the most comprehensively studied *Cyanobacteria*, again with maximal support.

### Microbiome phylogenomics

To identify the organisms in the putative microbiome bins recovered from the 17 cultures, we built two phylogenomic

---

**Table 3. cont.**

| Strain | MetaBAT setting | Bin | CM taxon | Nature | No. of scaffolds | Length (%) | Coverage (med) | CM compl. | CM contam. |
|--------|-----------------|-----|----------|--------|-----------------|------------|--------------|-----------|------------|
| ULC165 | Verrysensitive | 1   | *Xanthomonadaceae* | M      | 53              | 15.37      | 24.76        | 99.54     | 0.8        |
| ULC179 | Superspecific   | 1   | *Alphaproteobacteria* | M      | 247             | 18.89      | 16.30        | 98.54     | 0.19       |
| ULC186 | Verrysensitive | 1   | *Cyanobacteria* | C      | 412             | 67.38      | 21.10        | 98.73     | 0.22       |
| ULC187 | Verryspecific  | 1   | *Cyanobacteria* | C      | 62              | 62.18      | 33.11        | 99.29     | 0.47       |

---
trees with different taxon samplings of reference prokaryotes from a concatenation of 53 ribosomal proteins (see Materials and Methods). Fig. 2 shows the small tree (193 Bacteria and 30 Archaea), surrounded by zooms in specific regions of the large tree (3374 Bacteria and 127 Archaea; Fig. S1). Only 27 out of 31 non-cyanobacterial bins could be included in the tree, four bins (marked by a dash in Table 3) being too incomplete to be positioned robustly (see Materials and Methods). The resolution of the small tree was quite good, with 78 % of the nodes having PP/C2 ≥0.90 and no node having a PP<0.50. This analysis showed that all 27 analysed microbiome bins fall either in Bacteroidetes (five bins) or in Proteobacteria (14 bins in Alphaproteobacteria, five bins in Betaproteobacteria and three bins in Gammaproteobacteria) (Fig. 2), the tree allowing us to precisely determine the CheckM ‘bacterial’ affiliation of ULC082-bin3 to Gammaproteobacteria. In all cases, microbiome bins were sisters to one or more of the representative organisms with PP≥0.99, except for ULC179-bin3 (PP=0.63). Insets A–C of Fig. 2 demonstrate that the five Bacteroidetes bins correspond to different organisms, despite the fact that they appear closely clustered in the small tree. However, the picture is different for the bins falling in Proteobacteria (insets E–H). Whereas they are globally scattered across the phylum, there exist five cases (involving 11 bins) for which two or three bins from different cyanobacterial cultures appear extremely close in the large tree: ULC073-bin2/ULC084-bin1/ULC146-bin4 (D), ULC146-bin1/ULC165-bin3 (D), ULC065-bin2/ULC165-bin1 (E), ULC027-bin3/ULC146-bin3 (G) and ULC084-bin2/ULC165-bin2 (H). Taking this into account, the 27 microbiome bins only create 21 terminal branches in the large tree, five of them (representing six strains) clustering with a reference strain of Brevundimonas subvibrioides (H).

SSU rRNA (16S) analyses

In an attempt to refine the taxonomic analysis of all our genome bins, we predicted their SSU rRNA (16S) with RNAmmer [47]. Hence, we managed to predict 38 sequences (Table 4). Unfortunately, the vast majority (33) of the rRNA genes were predicted from unbinned metagenomic contigs (nobins; see Materials and Methods). When the taxon corresponding to the rRNA was straightforward to match with the taxon of one of the bins from the same cyanobacterial culture (based on congruent CheckM and SINA classifications), we manually affiliated the rRNA gene to that bin. This was possible for 20 predicted rRNA genes, but 13 sequences could not be reliably affiliated to any genome bin (empty cells in Table 4). According to SINA [48], only 10 of the predicted SSU
rRNA genes were of cyanobacterial origin, whereas eight sequences were left unclassified. The 20 remaining sequences were of either Proteobacteria or Bacteroidetes origin, thereby confirming the results of our phylogenomic analysis of microbiome bins based on rRNA proteins. Two best hits were encountered more than once by SINA: Blastomonas sp. AAP25 (from a Czech freshwater lake) in ULC073-bin6 and ULC146-bin3, and 'Uncultured bacterium' clone B3NR69D12 (from a drinking water biofilm) in ULC073-bin2 and ULC084-bin1.

**DISCUSSION**

According to the standards developed by the Genomic Standards Consortium for the minimum information about metagenomes of bacteria and archaea [25], the vast majority (14) of the cyanobacterial bins are of medium-quality, as their genome completeness is ≥90% and their contamination level ≤5% (both with CheckM and with DIAMOND BLASTX). Yet, they are still composed of a large number of scaffolds (≥60), due to the use of short insert DNA libraries for sequencing (Tables 3 and S2). In contrast, the only low-quality cyanobacterial assembly obtained here (ULC165-bin4) shows a completeness of 24.14%, in agreement with the lowest coverage obtained over all four ULC165 bins (3.90%). The situation is worse with the two Nostocales cultures (ULC146 and ULC179), for which we could not isolate any cyanobacterial bin. This lack of cyanobacterial contigs can be explained by the fact that these three strains (ULC146, ULC165 and ULC179) produce a thick polysaccharidic sheath that hinders DNA extraction [1]. Such a thick sheath is thought to protect the organisms from the harsh conditions of their hostile environment (Sør Rondane Mountains in Antarctica in all three cases). The use of a DNA extraction protocol more adapted to these organisms with a thick sheath (e.g. [55]) might have given different results and should be considered for future applications. Regardless, the recovery of only one cyanobacterium per sample provides molecular evidence for the integrity of the cultures in the BCCM/ULC collection.

When MetaBAT partitioned the metagenomic contigs, it produced nine small bins that were left unclassified by CheckM. In two cases, unclassified bins were identified as complementary (of CheckM marker genes) to another bin.
from the same metagenome (ULC027-bin3/ULC027-bin4; ULC146-bin3/ULC146-bin7; see above). Despite similar values in GC content and sequencing coverage, we did not merge these bins, thereby following the recommendations in the CheckM manual, because we had no indication about the phylogenetic affiliation of the unclassified bins. Because they only represented a very small fraction of the metagenomes, we discarded these bins from our phylogenetic analyses. Puzzlingly, such a bin was also recovered from strain ULC007, for which no foreign bin was expected due to its axenicity. While the sequencing coverage of the unclassified bin (ULC007-bin2) was more than twice that of the main bin (ULC007-bin1), tetranucleotide frequencies (TNFs) were undistinguishable between the two bins (Figs S1 and S3). This suggests that the corresponding contigs originate from the same organism but that the small bin contains contigs encoded in multiple copies in the genome. We attempted to characterize some unclassified bins from a functional point a view using Prodigal [33] and Blast2GO [56]. Unfortunately, the results were largely inconclusive.

Table 4. SSU rRNA (16S) gene prediction, taxonomy and coverage

The last-common ancestor (LCA) classification and top hits were retrieved from SINA analyses. The bins with SSU rRNA (16S) genes directly predicted from the genome bins (without manual assignment) are indicated by an asterisk (*). Coverage values were computed with BBMap. NA, not applicable.

| Strain | SSUref_128 taxon | SSUref_128 top hit | Bin affiliation | Coverage |
|--------|------------------|--------------------|-----------------|----------|
| ULC335 | Snowella         | Snowella litoralis | bin0 bin1      | 37.00    |
| ULC335 | Brevundimonas    | Uncultured Brevundimonas sp. | bin0 | 22.23 |
| ULC335 | Flavobacterium   | Uncultured bacterium clone N4_091 | bin0 | 58.44 |
| ULC335 | Unclassified     | NA                 | bin2 | 10.25 |
| ULC335 | Hydrogenophaga   | Hydrogenophaga pallonii | bin0 | 9.64  |
| ULC335 | Rhodobacteriaeae | Uncultured bacterium clone ZWB3-3 | bin0 | 7.79  |
| ULC007 | Leptolyngbya     | Phormiøøesmis priestleyi ANT.LG2.4 16S | bin0 | 85.23 |
| ULC027 | Unclassified     | NA                 | bin2 bin2* | 54.08 |
| ULC041 | Leptolyngbya     | Leptolyngbya antarctica ANT.LAV6.1 | bin0 | 97.23 |
| ULC065 | Arenimonas       | Uncultured bacterium clone a33 | bin2 | 40.66 |
| ULC065 | Synechococcus    | Cyanobium sp. JJ17-5 | bin0 | 165.13 |
| ULC066 | Limnobacter      | Uncultured bacterium clone S25 | bin0 | 14.15 |
| ULC066 | Unclassified     | NA                 | bin0 | 21.51 |
| ULC066 | Familyl          | Pseudanabaena biceps PCC 7429 | bin0 | 50.06 |
| ULC068 | Familyl          | Pseudanabaena sp. Sa012 | bin0 | 68.53 |
| ULC073 | Sphingomonadaceae| Blastomonas sp. AAP25 | bin6 | 31.87 |
| ULC073 | Leptolyngbya     | Leptolyngbya antarctica ANT.LAV6.1 | bin0 | 33.60 |
| ULC073 | Limnobacter      | Uncultured bacterium clone B3NR69D12 | bin0 | 19.58 |
| ULC077 | Unclassified     | NA                 | bin0 | 52.80 |
| ULC082 | Hydrogenophaga   | Uncultured Comamonadaceae bacterium | bin0 | 18.85 |
| ULC082 | Brevundimonas    | Uncultured alphaproteobacterium clone KWK6S.50 | bin0 | 25.08 |
| ULC082 | Unclassified     | NA                 | bin0 | 32.77 |
| ULC082 | Pseudomonas      | Pseudomonas sp. WCS374 | bin0 | 32.71 |
| ULC082 | Synechococcus    | Synechococcus sp. MW97C4 | bin0 | 93.93 |
| ULC084 | Brevundimonas    | Uncultured alphaproteobacterium | bin0 | 31.39 |
| ULC084 | Synechococcus    | Uncultured bacterium clone MS81 | bin0 | 87.30 |
| ULC084 | Limnobacter      | Uncultured bacterium clone B3NR69D12 | bin0 | 16.99 |
| ULC129 | Phormidium       | Uncultured bacterium clone GBII-52 | bin0 | 52.71 |
| ULC146 | Sphingomonadaceae| Blastomonas sp. AAP25 | bin3 | 81.39 |
| ULC146 | Flavobacterium   | Flavobacterium sp. Leaf359 | bin0 | 25.91 |
| ULC146 | Hydrogenophaga   | Hydrogenophaga sp. Root209 | bin1 | 61.13 |
| ULC165 | Unclassified     | NA                 | bin0 | 85.20 |
| ULC165 | Unclassified     | NA                 | bin0 | 98.62 |
| ULC179 | Devosia          | Devosia psychrophila strain Cr7-05 | bin0 | 97.88 |
| ULC179 | Unclassified     | NA                 | bin0 | 16.43 |
| ULC179 | Polymorphobacter | Uncultured Sphingomonadaceae bacterium | bin3 | 91.23 |
| ULC186 | Familyl          | Leptolyngbya sp. 08B32S02 | bin0 | 116.04 |
| ULC187 | Familyl          | Pseudanabaena sp. Sa010 | bin0 | 81.01 |
and we could not ascertain whether these bins (containing some transferases, e.g. acyltransferases, transferring one-carbon groups, transferring nitrogenous groups) correspond to aberrant chromosomal regions (e.g. laterally transferred segments, repetitive elements) or to plasmids (data not shown).

Even if our assemblies are globally of medium quality, they often lack SSU rRNA (16S) genes. Hence, out of 38 predicted rRNA genes, as few as five were predicted from genome bins (all of which are foreign bins), leaving 50 bins without any rRNA gene. Apparently, rRNA genes are rejected by MetaBAT, because we could only predict them from unbinned contigs (nobins) in all remaining cases (33). Importantly, this outcome was independent of the parameter set used for MetaBAT (data not shown). We nonetheless elected to favour this software because its binning performance in terms of completeness is better than that of other recent tools, such as CONCOCT [57], GroopM [58], MaxBin [59] and Canopy [60] (see figure 3 of Kang et al. [27]). Whenever SINA [48] successfully classified a predicted SSU rRNA (16S) gene, we did our best to manually affiliate it to the corresponding genome bin (Table 4). Consequently, 10 of our 15 cyanobacterial bins turned into high-quality genomes [25]. In this respect, it is worth mentioning that, among the 651 cyanobacterial genome assemblies available on the NCBI as of December 2017, only 458 have an SSU rRNA (16S) gene, based on RNAmmer [47] predictions (data not shown). According to our analyses, the frequent loss of rRNA genes is caused by the presence of multiple copies of the rRNA operon in many bacterial genomes [61], resulting in short rRNA-bearing contigs due to incomplete assembly of repeated regions. Because these contigs are dominated by the rRNA operon, they feature both a higher sequencing coverage and divergent TNFs, two properties that interfere with the binning process carried out by MetaBAT and other metagenomic software (Appendix S2). Yet, an improved sequencing depth might have positively impacted the results of our study. Even if sequencing coverage (ranging between 6.27 and 38.37) was sufficient to ensure reliable binning of the cyanobacterial contigs, deeper coverage would have resulted in more complete bins, whether cyanobacterial or corresponding to the microbiome bacteria. More data could also have improved assembly contiguity (in terms of scaffold size), which in turn might have helped with the binning of rRNA genes. This is particularly important because SSU rRNA (16S) is still the standard for microbial taxonomy [49]. Another way to improve the assembly quality is to use third-generation sequencing (TGS), such as Pacific Bioscience (PacBio) or Oxford Nanopore Technology (ONT). These approaches use long reads of 10 kb (instead of 250 nt with Illumina), which has been shown to increase the contiguity of assemblies, especially in bacteria [62, 63]. Regarding the exploitation of non-axenic cultures, it has been recently shown that plasmid binning from PacBio data could avoid the production of small unclassified bins by considering features others than TNF and coverage alone [64].

Our phylogenomic tree of Cyanobacteria is based on the largest supermatrix (in terms of conserved positions) to date (64 non-contaminated and complete reference strains; >170,000 unambiguously aligned amino-acid positions). It is congruent with other recent cyanobacterial phylogenies [52, 65]. We chose to root the tree on the Gloeobacter species (clade G), following the practice of many recent cyanobacterial phylogenies (e.g. [8, 40, 52, 65–68]). Nevertheless, it is worth mentioning that the basal position of Gloeobacter has been criticized [69] and that an alternative rooting has been recently proposed [70]. Interestingly, three of the cyanobacterial bins corresponding to polar or subpolar strains are clearly located in the basal part of the tree. The BCCM/ULC collection has a focus on (sub)polar cyanobacterial strains that may present interesting features to survive freeze/thaw cycles, seasonally contrasted light intensities, high UV radiation, desiccation and other stresses. Cyanobacterial diversity from such environments is presently underrepresented in comparison to that of marine Cyanobacteria. This is notably due to the difficulty of cultivating these organisms from ‘cold regions’, such as polar or alpine Cyanobacteria [13]. Hence, increasing the sampling of (cyano)bacteria from these environments may lead to a better understanding of their functional adaptation to environmental pressures, which is especially important in the context of climate change [13]. Moreover, the three ‘early-branching’ Pseudanabaena strains (ULC066, ULC068 and ULC187 in clade F) should prove useful to improve the resolution of the phylogeny of Cyanobacteria in further studies by increasing their taxon sampling. Two of these strains were isolated from Canadian samples and ULC066 even originates from the Arctic (Table 1).

When the sequencing coverage was sufficient, we also assembled the foreign (i.e. non-cyanobacterial) bins. According to Bowers et al. [25], 13 of these bins are of medium quality (completeness ≥90 %) and 18 bins are of low quality (completeness <90 %) (Table 3). All are either of Proteobacteria or Bacteroidetes origin, as assessed by both CheckM and phylogenomic inference. All the Cyanobacteria of the present study are freshwater organisms. Consequently, the cyanobacterial microbiome from other environments might be completely different. From our phylogenomic analysis, it appears that the 27 analysed bins represent 21 different terminal branches in the tree (Fig. 2). As 11 were indistinguishable (or very closely related) in spite of the use of 53 ribosomal proteins, we investigated whether they represented genuinely different samplings of highly similar associated organisms or were the result of cross-contamination during Cyanobacteria isolation/cultivation or DNA processing (Appendix S3). Altogether, genome-wide similarity measurements suggest that cross-contamination may not be involved, even if sampling sites were occasionally very distant (i.e. Arctic and Antarctic samples). Inset H of Fig. 2 shows a group of six foreign bins clustered around a reference strain of Bre-vundimonas subvibrioides. As this alphaproteobacterium
frequently appears as a last common ancestor taxon in SINA classifications of SSU rRNA (16S) sequences (Table 4), this indicates that Brevundimonas (or related taxa) is regularly present in ULC cultures and probably naturally associated with Cyanobacteria. More generally, the classification of all identifiable foreign bins as either Proteobacteria or Bacteroidetes suggests that the associated organisms come from the original environment and accompanied the Cyanobacteria through the isolation steps. Indeed, these two phyla are known to co-evolve with Cyanobacteria through complex trophic relations [21, 50].

We probably identified only these two phyla in our foreign bins because they are the most abundant [21], whereas other associated bacterial phyla (Actinobacteria, Gemmatimonadetes, Planctomycetes, Verrucomicrobia) have been described in the cyanobacterial microbiome [15–17, 21]. This result is completely in line with our recent analysis of the level of contamination in publicly available cyanobacterial genomes, in which foreign sequences were also mainly classified as Proteobacteria and Bacteroidetes [24]. In other words, the difficulty with purifying non-axenic cyanobacterial cultures, possibly combined with the accidental transfer of associated bacteria during the isolation process (or any subsequent step), is probably the main cause for genome contamination. This certainly highlights the importance of careful bioinformatic protocols for genome data processing. In this respect, we compared our new assembly of ULC007 to the previous release of the same strain, based on a HiSeq run in addition to the MiSeq run used here [53]. Interestingly, all CheckM values (completeness, contamination, strain heterogeneity) for ULC007-bin1 were slightly better than those obtained for our previously published assembly (completeness 98.11 vs 95.99, contamination 0 vs 1.18, strain heterogeneity 0 vs 100). As the latter had used more primary data and benefited from a thorough curation by hand, this indicates that the fully automated metagenomic pipeline of the present study is also applicable for axenic strains.

## Conclusion

In this work, we showed that a quite straightforward metagenomic protocol allows us to take advantage of non-axenic cyanobacterial cultures. Our pipeline yields medium-quality genomes with a high level of completeness (high sensitivity) for a very low level of contaminant sequences (high specificity), which could be very useful for phylogenomic analyses. In contrast, it has the disadvantage of regularly discarding multi-copy SSU rRNA (16S) genes during the binning of metagenomic contigs. We have shown that this loss is due to their higher sequencing coverage and divergent TNFs, which are especially detrimental for short contigs. The metagenomic pipeline reported here has nevertheless the advantage of facilitating the assembly of cyanobacterial genomes, as long as enough genomic DNA can be extracted from the strains. Our results further indicate that the microbiome of different cultures can sometimes contain associated bacteria that are very closely related, even when sampling sites are very distant. Finally, we have released 14 novel cyanobacterial assemblies, including 11 (sub)polar strains, and 13 assemblies of organisms belonging to their microbiome.

## Funding information

This work was supported by operating funds from FRS-FNRS (National Fund for Scientific Research of Belgium) (AW), the European Research Council Stg ELITE FP7/308074 (EJJ), the BELSPO project CCAMBO (SD/BA/03A) (AW), and the BELSPO Interuniversity Attraction Pole Planet TOPERS (EJJ and LC). LC was and ARB is a FRIA fellow of the FRS-FNRS, and LC is now an iAP Planet Topers PhD scholar. MH and AW are Stewards of the FRS-FNRS. Computational resources were provided by the Fédération Wallonie-Bruxelles (Tier-1; funded by Walloon Region, grant no. 1117545), the Consortium des Équipements de Calcul Intensif (CECI; 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).

## Acknowledgements

We thank Dr. Yannick Lara (University of Liège) for the isolation of strains ULC186, ULC187, ULC335, his help with cultivation and DNA extraction, and insightful comments on preliminary versions of the manuscript. Professor Warwick Vincent (Laval University) is acknowledged for the deposit of strains ULC065, ULC066, ULC68, ULC077, ULC082 and ULC084. We are also grateful to the reviewers for their constructive criticisms.

## Conflicts of interest

The authors declare that there are no conflicts of interest.

## References

1. Rippka R, Stanier RY, Deruelles J, Herdman M, Waterbury JB. Generic assignments, strain histories and properties of pure cultures of Cyanobacteria. Microbiology 1979;111:1–61.
2. Whitton BA, Potts M. Introduction to the Cyanobacteria. In: Whitton BA (editor). Ecology of Cyanobacteria II. Netherlands: Springer; 2012. pp. 1–13.
3. Hofmann HJ, Microflora P, Islands B. Canada: significance and systematics. J Paleontol 1976;50:1040–1073.
4. Golubic S, Hofmann HJ. Comparison of holocene and middle-pre cambrian Entophysalidaceae (Cyanophyta) in stromatolitic algal mats: cell division and degradation. J Paleontol 1976;50:1074–1082.
5. Knoll AH, Golubic S. Proterozoic and living cyanobacteria. In: Schidlowski PDM, Golubic PDS, Kimberley PDMM, Trudinger DPA (editors). Early Organic Evolution. Berlin Heidelberg: Springer; 1992. pp. 450–462.
6. Knoll AH. The geological consequences of evolution. Geobiology 2003;1:3–14.
7. Kopp RE, Kirschvink JL, Hilburn IA, Nash CZ. The paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. Proc Natl Acad Sci USA 2005;102:11131–11136.
8. Ochoa de Alda JA, Esteban R, Diago ML, Houmard J. The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun 2014;5:4937.
9. Archibald JM. The puzzle of plastid evolution. Curr Biol 2009;19: R81–R88.
10. Namikoshi M, Rinheart KL. Bioactive compounds produced by cyanobacteria. J Ind Microbiol Biotechnol 1996;17:373–384.
11. Singh S, Kate BN, Banerjee UC. Bioactive compounds from cyanobacteria and microalgae: an overview. Crit Rev Biotechnol 2005;25:73–95.
12. Sharma NK, Stal LJ. The economics of cyanobacteria-based biofuel production: challenges and opportunities. In: Sharma NK, Rai AK, Stal LJ (editors). Cyanobacteria. New Jersey, USA: John Wiley & Sons, Ltd; 2014. pp. 167–180.
56. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005;21:3674–3676.

57. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;11:1144–1146.

58. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2014;2:e603.

59. Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2014;2:26.

60. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;32:822–828.

61. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* 2000;66:1328–1333.

62. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* 2015;23:110–120.

63. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VG, McHardy AC et al. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* 2016;6:25373.

64. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol* 2018;36:61–69.

65. Ponce-Toledo R, Deschamps P, López-García P, Zivanovic Y, Benzerara K et al. An early-branching freshwater cyanobacterium at the origin of plastids. *Curr Biol* 2017;27:386–391.

66. Blank CE. Origin and early evolution of photosynthetic eukaryotes in freshwater environments: reinterpreting protorozoic paleobiology and biogeochemical processes in light of trait evolution. *J Phycol* 2013;49:1040–1055.

67. Uyeda JC, Harmon LJ, Blank CE. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. *PLoS One* 2016;11:e0162539.

68. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc Natl Acad Sci USA* 2017;114:E7737–E7745.

69. Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci USA* 2012;109:17513–17518.

70. Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol* 2017;1:193.

---

**Five reasons to publish your next article with a Microbiology Society journal**

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as ‘excellent’ or ‘very good’.
5. Your article will be published on an interactive journal platform with advanced metrics.

Find out more and submit your article at microbiologyresearch.org.