J. Phycol. **52,** 356–368 (2016)
© 2016 Phycological Society of America
DOI: 10.1111/jpy.12399

ON THE USE OF HIGH-THROUGHPUT SEQUENCING FOR THE STUDY OF CYANOBACTERIAL DIVERSITY IN ANTARCTIC AQUATIC MATS[1]

Igor Stelmach Pessi, Pedro De Carvalho Maalouf, Haywood Dail Laughinghouse IV

Centre for Protein Engineering, University of Liège, Allée du Six Août 13, B6a, Quartier Agora, Sart-Tilman, 4000 Liège, Belgium

Denis Baurain

Eukaryotic Phylogenomics, Department of Life Sciences/PhytoSYSTEMS, University of Liège, Chemin de la Vallée 4, B22, Quartier Vallée 1, Sart-Tilman, 4000 Liège, Belgium

and Annick Wilmotte[2]

Centre for Protein Engineering, University of Liège, Allée du Six Août 13, B6a, Quartier Agora, Sart-Tilman, 4000 Liège, Belgium

The study of Antarctic cyanobacterial diversity has been mostly limited to morphological identification and traditional molecular techniques. High-throughput sequencing (HTS) allows a much better understanding of microbial distribution in the environment, but its application is hampered by several methodological and analytical challenges. In this work, we explored the use of HTS as a tool for the study of cyanobacterial diversity in Antarctic aquatic mats. Our results highlight the importance of using artificial communities to validate the parameters of the bioinformatics procedure used to analyze natural communities, since pipeline-dependent biases had a strong effect on the observed community structures. Analysis of microbial mats from five Antarctic lakes and an aquatic biofilm from the Sub-Antarctic showed that HTS is a valuable tool for the assessment of cyanobacterial diversity. The majority of the operational taxonomic units retrieved were related to filamentous taxa such as *Leptolyngbya* and *Phormidium*, which are common genera in Antarctic lacustrine microbial mats. However, other phylotypes related to different taxa such as *Geitlerinema*, *Pseudanabaena*, *Synechococcus*, *Chamaesiphon*, *Calothrix*, and *Coleodesmium* were also found. Results revealed a much higher diversity than what had been reported using traditional methods and also highlighted remarkable differences between the cyanobacterial communities of the studied lakes. The aquatic biofilm from the Sub-Antarctic had a distinct cyanobacterial community from the Antarctic lakes, which in turn displayed a salinity-dependent community structure at the phylotype level.

Key index words: **16S rRNA gene; 454 pyrosequencing; Antarctica; biodiversity; bioinformatics; cyanobacteria; high-throughput sequencing; lacustrine environments**

Abbreviations: **DGGE, denaturing gradient gel electrophoresis; HTS, high-throughput sequencing; LCA, lowest common ancestor; OTU, operational taxonomic unit; RDP, Ribosomal Database Project; UPGMA, unweighted pair group method with arithmetic mean**

Cyanobacteria are a major component of Antarctic ecosystems, usually forming complex microbial mat communities in the benthos of lakes, ponds, and streams (Vincent and Quesada 2012). They often dominate total biomass and primary production in lacustrine environments and are also the main source of organic carbon for soil communities (Singh and Elster 2007). Despite their importance as the dominant phototrophs in Antarctic ecosystems, a comprehensive understanding of the ecological and geographical range of Antarctic cyanobacteria, including the environmental processes shaping their distribution and activity, is far from being achieved.

Important studies have been carried out on this subject using morphological identification as well as indirect molecular methods based on 16S rRNA gene sequences retrieved from environmental DNA, such as denaturing gradient gel electrophoresis (DGGE) and clone libraries. It is now known that Antarctic benthic microbial mat communities are dominated by filamentous cyanobacteria, such as *Leptolyngbya*, *Phormidium*, *Microcoleus*, and *Oscillatoria*, which are responsible for the three-dimensional structure of the mats (Komárek 2007, Komárek et al. 2008, Vincent and Quesada 2012). Morpho- and genotypes related to heterocystous cyanobacteria

[1]Received 4 December 2015. Accepted 31 January 2016.
[2]Author for correspondence: e-mail awilmotte@ulg.ac.be.
Editorial Responsibility: A. Post (Associate Editor)

such as *Nostoc*, *Calothrix*, and *Nodularia* also make up a large fraction of the communities, and unicellular cyanobacteria are usually less represented (Sabbe et al. 2004, Jungblut et al. 2005, Taton et al. 2006a, 2008). Earlier studies based on 16S rRNA gene sequences indicated a bioregionalization of cyanobacterial mat communities within Antarctic lacustrine environments and also suggested the existence of endemic Antarctic cyanobacteria (Taton et al. 2003, 2006a,b, Jungblut et al. 2005). Finally, additional molecular studies found that low-temperature cyanobacterial ecotypes are globally distributed throughout the cold biosphere (Jungblut et al. 2010).

Further advances in this area have been hampered due to limitations of morphological identification and traditional molecular tools for the assessment of cyanobacterial diversity. Morphological characters traditionally used in cyanobacterial taxonomy (e.g., cell division pattern, colony formation, length/width measurement of cells, and presence of extracellular envelopes and sheaths) are subjected to plasticity, making the distinction between morphologically similar taxa often difficult or impossible (Wilmotte and Golubic 1991). In vitro biochemical, physiological, and genotypic characterization is also of limited application due to the fact that a part of the cyanobacterial diversity is not easily cultivable (Waterbury 2006). Finally, traditional molecular approaches such as DGGE and cloning of 16S rRNA gene sequences are time consuming, expensive, and only provide information on the most abundant members of the microbial communities (Tringe and Hugenholtz 2008).

High-throughput sequencing (HTS) technologies such as 454 pyrosequencing and Illumina have become valuable tools for the study of complex microbial communities (Shokralla et al. 2012, Lynch and Neufeld 2015). These technologies allow the parallel recovery of genetic information from multiple templates and thus bypass the necessity of vector-based cloning procedures. They also generate several hundreds of thousands to tens of millions of sequences on a single run, therefore providing an in-depth inventory of the microbial diversity on a given environment. Finally, through the use of sample-specific barcode tags, several samples can be multiplexed and sequenced in parallel, thus allowing much more complex experimental designs (Binladen et al. 2007).

HTS has been widely used in amplicon-sequencing studies, where a genetic marker such as the 16S rRNA gene, which has been the golden standard for microbial phylogeny since originally proposed by Woese and Fox (1977), is amplified directly from environmental DNA and sequenced. High-throughput amplicon-sequencing studies have revealed a higher microbial diversity than previously acknowledged, and have allowed the investigation of rare, low-abundance taxa that constitute an important fraction of microbial communities but have been largely neglected (Sogin et al. 2006, Roesch et al. 2007).

One caveat of the HTS approach is that the extreme increase in sequencing depth comes along with a considerable amount of PCR and sequencing artifacts, which contribute significantly to the overestimation of microbial diversity (Kunin et al. 2010). Moreover, each step of PCR-based, community-level assessments can introduce biases to the observed structure of the microbial communities (von Wintzingerode et al. 1997). A correct assessment of microbial diversity using HTS thus relies on robust bioinformatics tools that can efficiently identify and remove PCR and sequencing artifacts. To this day, several studies have tested the effects of different experimental and bioinformatics parameters on the observed structure of microbial communities (Schloss et al. 2009, 2011, Kunin et al. 2010, Bakker et al. 2012, Lee et al. 2012, Pinto and Raskin 2012, Edgar 2013, Gaspar and Thomas 2013). However, given major experimental differences in DNA extraction methods, analyzed regions of the 16S rRNA gene and targeted taxa across studies, the emergence of a universal pipeline that would apply to every amplicon-sequencing study seems unrealistic. More likely, experimental and bioinformatics procedures should be validated on a study-by-study basis to overcome the limitations of the HTS approach and generate ecologically meaningful results. Nevertheless, only few studies have been carried out so far in which HTS coupled with cyanobacteria-specific primers was used for the targeted investigation of cyanobacterial diversity (Dadheech et al. 2013, Hur et al. 2013, Kleinteich et al. 2014, Brandes et al. 2015), and a comprehensive evaluation of different experimental protocols is still lacking.

Here, we assess the use of HTS as a tool for the study of the cyanobacterial diversity in Antarctic lakes. Artificial communities were used to evaluate the performance of different bioinformatics pipelines. The effect of different DNA extraction methods, the use of longer barcoded primers, and the reproducibility of the results were determined using environmental samples. Finally, we present a description of the cyanobacterial diversity in microbial mats from five Antarctic lakes and an aquatic biofilm from the Sub-Antarctic, assessing the strength of HTS in comparison to the well-established DGGE and cloning approaches.

MATERIALS AND METHODS

Reference strains. Twenty-two cyanobacterial strains from the BCCM/ULC Cyanobacteria Collection (http://bccm.belspo.be/about-us/bccm-ulc) were used to assemble the artificial communities (Table 1). These were selected based on a neighbor-joining tree of partial 16S rRNA gene sequences from 108 strains available in the collection (data not shown). A total of 20 polar strains, one from Siberia (S141 *Leptolyngbya* sp. "Doroninskoye") and one from New Zealand (ULC080 *Anabaena* sp. CY-036), were chosen, comprising 11 of the 19

TABLE 1. List of reference strains used in the artificial communities Art1 and Art2.

ID[a]	Strain name	Accession number	Relative abundance (%)	
			Art1	Art2
ULC022	*Phormidesmis priestleyi* ANT.L61.2	AY493582	4.5	8.9
ULC026	*Phormidesmis priestleyi* ANT.L52.6	AY493579	4.5	8.9
ULC049	*Phormidesmis priestleyi* ANT.L66.1	AY493581	4.5	8.9
ULC009	*Plectolyngbya hodgsonii* ANT.LPR2.2	AY493583	4.5	8.9
ULC001	*Leptolyngbya frigida* ANT.L53B.1	AY493608	4.5	8.9
ULC029	*Leptolyngbya frigida* ANT.L52B.3	AY493612	4.5	8.9
ULC043	*Leptolyngbya antarctica* ANT.LWAV6.1	AY493602	4.5	8.9
S141	*Leptolyngbya* sp. "Doroninskoye"	KT753316	4.5	8.9
ULC041	*Leptolyngbya antarctica* ANT.LAC.1	AY493588	4.5	8.9
ULC065	*Cyanobium* sp. "Bylot Island"	KT753317	4.5	1.5
S082	*Cyanobium* sp. "Chester Cone"	KT753318	4.5	1.5
ULC084	*Cyanobium* sp. "Laguna Chica"	KT753319	4.5	1.5
ULC004	*Leptolyngbya* cf. *fragilis* ANT.L52.1	AY493584	4.5	1.5
S111	*Microcoleus vaginatus* JR6	KT753320	4.5	1.5
S120	*Microcoleus favosus* JR20	KT753321	4.5	1.5
ULC117	*Leptolyngbya* sp. ANT07.JR16	KT753322	4.5	1.5
ULC123	*Leptolyngbya* sp. ANT07.JR23	KT753323	4.5	1.5
ULC080	*Anabaena* sp. CY-036	KT753324	4.5	1.5
S133	*Nodularia* sp. "khil 06-sten"	KT753325	4.5	1.5
ULC050	*Nostoc* sp. ANT.L34.1	AY493591	4.5	1.5
ULC069	*Pseudanabaena frigida* O-302	KT753326	4.5	1.5
S075	Oscillatoriaceae Toolik Lake O-103	KT753327	4.5	1.5

[a]A ULCXXX number indicates that the strain is available at the BCCM/ULC Cyanobacteria Collection (http://bccm.belspo.be/about-us/bccm-ulc).

cyanobacterial lineages represented in the phylogenetic tree. Within some clusters, up to three strains were selected, to assess the sensitivity of the method to discriminate between closely related organisms.

The full 16S rRNA gene sequences of 10 strains, previously determined by Taton et al. (2006b), were retrieved from GenBank. For the remaining strains, sequences were obtained using the 16S27F and 23S30R primers as described elsewhere (Taton et al. 2003). Sequencing was carried out bidirectionally on a 3730 DNA Analyser (Applied Biosystems, Foster City, CA, USA) at the GIGA Genomics facility (Liège, Belgium). Sequences of the 22 reference strains were aligned using MUSCLE (Edgar 2004) and trimmed to the expected amplicon product of the V3-V4 variable region – hereafter referred to as "reference sequences." Full 16S rRNA gene sequences obtained in this study were submitted to GenBank under the accession numbers KT753316–KT753327 (Table 1).

Artificial communities and environmental samples. For the construction of artificial communities, genomic DNA was extracted from the reference strains using the DNeasy Plant Mini Kit (Qiagen, Venlo, the Netherlands). DNA extracts were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Carlsbad, CA, USA) and pooled at known abundances (Table 1). Library Art1 ("even") was prepared by combining the DNA of all strains at equal amounts, whereas in library Art2 ("staggered") DNA from nine strains accounted for 80% of the community. The latter was assembled to represent a more realistic scenario of taxa distribution in Antarctic environments, where some species are more abundant than others.

Microbial mats from five Antarctic lakes and an aquatic biofilm from the Sub-Antarctic were also analyzed (Fig. 1). Lake 59b (69° 24′ S; 76° 21′ E), Lake Rauer2 (68° 48′ S; 77° 51′ E), and Highway Lake (68° 30′ S; 78° 14′ E) are located in East Antarctica. Lake 59b is a shallow (0.8 m deep) freshwater lake in Larsemann Hills (Prydz Bay), the second largest ice-free region of East Antarctica (Hodgson et al. 2001a). Lake Rauer2, located in Rauer Islands (south-eastern Prydz Bay, 60 km northeast of Larsemann Hills), is a relatively deep (3 m depth) hypersaline lake (131.5 mS · cm^{-1} conductivity; Hodgson et al. 2001b). Highway Lake is an oligotrophic (3–32 mg · L^{-1} DOC; 0.32–6.74 μg · L^{-1} chlorophyll *a*), slightly brackish lake (4–5‰ salinity) in Vestfold Hills (Laybourn-Parry and Wadham 2014). Forlidas Pond (82° 27′ S; 51° 16′ W) and Lundström Lake (80° 26′ S; 29° 26′ W) are located in the Transantarctic Mountains. Forlidas Pond is a shallow (1.8 m deep), perennially frozen freshwater lake in the north-eastern end of the Dufek Massif, overlying a 20 cm thick layer of hypersaline brine (Hodgson et al. 2010, Fernández-Carazo et al. 2011). Lundström Lake, located in the Shackleton Range (420 km northeast of Forlidas Pond), is a relatively deep (3.1 m depth), perennially frozen freshwater lake (Fernández-Carazo et al. 2011). Finally, one sample from a partially submerged wet biofilm on gravel from Macquarie Island (Sub-Antarctic, 54° 28′ S; 158° 55′ E) was also analyzed.

454 pyrosequencing analysis. DNA was extracted from the mats using the PowerSoil DNA Isolation Kit (MOBIO Laboratories, Carlsbad, CA, USA) according to the manufacturer's instructions with some modifications. Tubes were agitated on a vortex for 20 extra min to ensure a good disintegration of the mats and, if not completely disintegrated, a sterile pestle was used to crush the remaining pieces. For comparison, the Forlidas Pond sample was also processed with the DNA extraction method used in the DGGE analysis (see below). DNA concentration and quality were determined using a NanoVue spectrophotometer (GE Healthcare Life Sciences, Little Chalfont, UK). A blank DNA extraction consisting of sterile Milli-Q water was carried out in parallel.

The cyanobacteria-specific primer set CYA359F and CYA781R(a)/CYA781R(b) (Nübel et al. 1997) was used to amplify the V3-V4 variable region of the 16S rRNA gene from both the artificial and environmental samples. The reverse primers CYA781R(a) and CYA781R(b) differ by only two nucleotides at positions 7 and 23 (5′–3′), but have been shown to preferentially amplify different cyanobacterial lineages (Boutte et al. 2006), and were therefore used separately in independent PCR reactions. Primers were modified

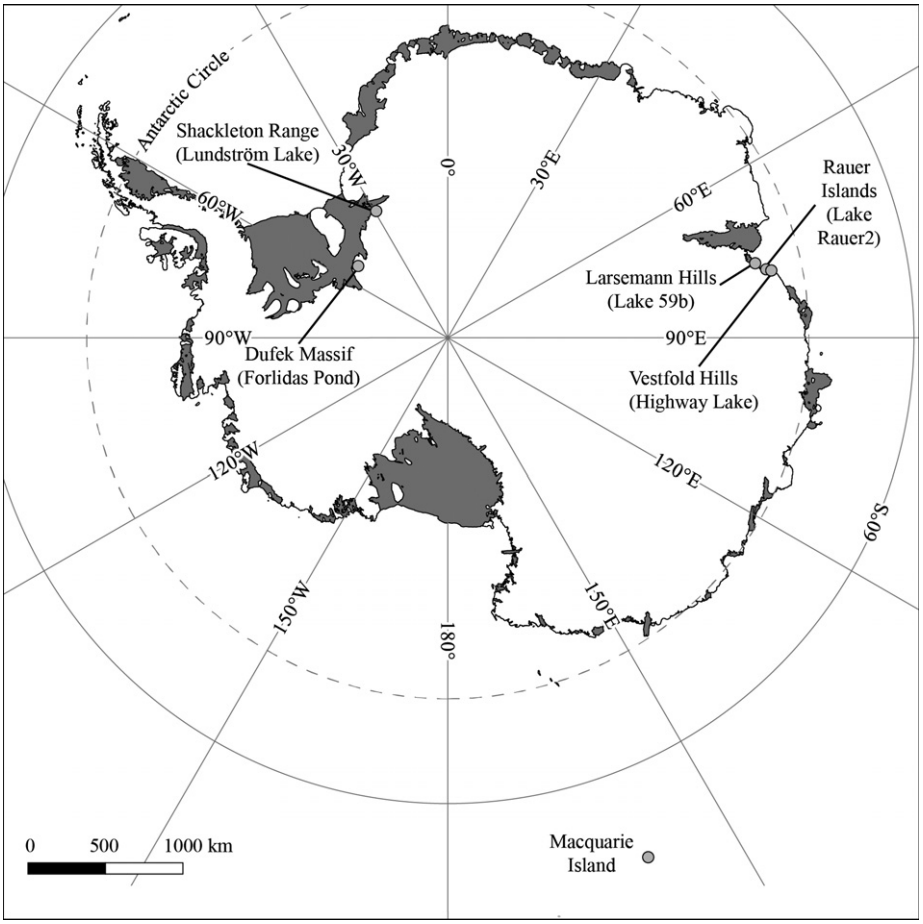

FIG. 1. Map of Antarctica showing the location of the samples analyzed in this study. Map created using the Quantarctica package (Norwegian Polar Institute, Tromsø, Norway).

to include a 10-bp sample-specific barcode tag at the 5′ end, to allow samples to be multiplexed for sequencing. PCR reactions consisted of 1× PCR buffer with 1.5 mM $MgCl_2$, 1 mg · mL^{-1} BSA, 200 μM of each dNTP, 0.2 μM of each primer, 1 U SUPER TAQ *plus* DNA polymerase (HT Biotechnology, Cambridge, UK), and 4 ng · μL^{-1} template DNA in a final volume of 50 μL. Amplification was performed using an initial denaturation step at 94°C for 2 min, followed by 30 cycles of 94°C for 45 s, 60°C (for primer 781Ra) or 57°C (for primer 781Rb) for 45 s and 68°C for 45 s, and a final elongation at 68°C for 5 min. Negative controls (PCR mixes with either no DNA or the blank DNA extracts) were always included during PCR amplifications. To minimize stochastic PCR bias, amplification was carried out as six independent PCR reactions (three for each reverse primer) that were pooled before purification.

The Lake Rauer2 sample was analyzed as two technical replicates to assess the reproducibility of the experimental protocol. Technical replicates were carried out using the same DNA extract but independent PCR reactions and sequencing. Moreover, to assess potential biases arising from direct long primer amplification (Berry et al. 2011), the Lake Rauer2 and Forlidas Pond samples were also analyzed using a modified "two-step" PCR protocol. For this, the target region was first amplified as described above, but using primers without the barcode tag. Then, 10 ng · mL^{-1} of the PCR product were used as templates in an additional five-cycle PCR reaction, which included barcoded primers to extend the barcode tags to the amplicons.

Replicate PCR reactions were pooled and purified using the GeneJet PCR Purification Kit (Thermo Scientific, Waltham, MA, USA). Purified amplicons were quantified using the Quant-iT PicoGreen dsDNA Assay Kit, pooled in equimolar concentrations, and concentrated to 25 μL using the Amicon Ultra-0.5 mL 30K device (EMD Millipore, Billerica, MA, USA). Pooled libraries were sent to Beckman Coulter Genomics (Takeley, UK), where primer dimers were removed using the Agencourt AMPure XP Kit (Beckman Coulter, Brea, CA, USA) and sequencing adapters were ligated to the amplicons. Sequences were obtained using the 454 GS FLX+ Titanium platform (454 Life Sciences, Branford, CT, USA).

DGGE analysis. For DGGE analysis, DNA was extracted from the samples according to Fernández-Carazo et al. (2011). Briefly, cells were disrupted using glass beads (0.17–0.18 mm diameter), SNT solution (500 mM Tris-HCl pH8, 100 mM NaCl, 25% sucrose), 50 mg · mL^{-1} lysozyme, and SII solution (500 mM Tris base, 500 mM EDTA, 1% SDS, 6% phenol). DNA was purified with successive addition of phenol, phenol-chloroform-isoamyl alcohol (25:24:1), and chloroform-isoamyl alcohol (24:1), precipitated with sodium acetate-ethanol and resuspended in TE 1× buffer. DNA was further purified using the Wizard DNA Clean-up System (Promega, Madison, WI, USA).

Sequences of the V3-V4 variable region of the 16S rRNA gene were obtained by semi-nested PCR as in Taton et al. (2003) with some modifications. First, a larger fragment of the 16S rRNA gene was amplified using the primers CYA378F and 23S30R in 25 μL reactions containing 1× PCR buffer with 1.5 mM $MgCl_2$, 1 mg · mL^{-1} BSA, 200 μM of each dNTP, 0.5 μM of each primer, 1 U SUPER TAQ *plus* DNA polymerase, and 4 ng · μL^{-1} DNA. Thermocycling conditions consisted of 94°C for 5 min, followed by 15 cycles of 94°C for

45 s, 54°C for 1 min, and 68°C for 90 s, with a final elongation step at 68°C for 7 min. PCR products were used as templates in a second PCR reaction performed in the same conditions as above, but in a total volume of 50 μL and using the primers CYA378F and CYA781R(a)/CYA781R(b). Reverse primers contained a 38-bp GC clamp at the 5′ end, to ensure the melting stability of the DNA. The second amplification step was carried out using 94°C for 5 min, 25 cycles of 94°C for 1 min, 60°C for 1 min, and 68°C for 1 min, with a final elongation step at 68°C for 20 min.

PCR products were purified using the GeneJet PCR Purification Kit and analyzed by DGGE according to Boutte et al. (2006). To ensure experimental reproducibility, each sample was analyzed twice on two different gels. Bands were excised from the gel using a sterile scalpel and incubated in TE 1× buffer overnight at 4°C. DNA was reamplified using primers CYA378F and CYA781R and sequences were obtained bidirectionally on a 3730 DNA Analyser at the GIGA Genomics facility.

Bioinformatics procedures. Pyrosequencing reads were extracted and assigned to the corresponding sample based on the barcode tag using the "sffinfo" command in mothur with default parameters (Schloss et al. 2009). Quality control of reads, removal of chimeric sequences, and operational taxonomic unit (OTU) clustering were first performed on the data obtained for the artificial communities, using five independent bioinformatics pipelines. Pipelines (I) "shhh.flows (450 flows)," (II) "shhh.flows (360–720 flows)," and (III) "Sliding Window (Q35, 50 bp)" were carried out using mothur following the recommended procedure on http://www.mothur.org/wiki/454_SOP, corresponding to the homonymous pipelines as described in Schloss et al. (2011). Pipelines I and II applied flowgram denoising after trimming flowgrams to 450 and 360–720 flows, respectively, and pipeline III applied an average quality score of 35 over a sliding window of 50 bp. Pipelines (IV) "fastq_maxee" and (V) "fastq_truncqual" were carried out using UPARSE according to Edgar (2013), following the example commands in http://drive5.com/usearch/manual/uparse_cmds.html. In pipelines IV and V, reads were filtered based on a maximum expected errors value of 0.5 or a minimum per-base quality score of 16, respectively. In all pipelines, sequences were clustered into OTUs using a similarity cut-off value of 97.5% according to Taton et al. (2003). One representative sequence, defined as the most abundant unique sequence in the cluster, was determined for each OTU.

The correspondence between the OTUs retrieved at the end of each pipeline and the reference strains used to assemble the artificial communities was determined by mapping the OTU representative sequences to the 22 reference sequences using the "uparse_ref" command in USEARCH with default parameters (Edgar 2010). OTUs were classified as "Perfect" (identical to a reference sequence), "Good" (≥99% similarity), "Noisy" (≥97.5% to <99% similarity), "Other" (<97.5% similarity), and "Chimeric" (composed of two or more parent reference sequences). The relative abundance of each reference strain was inferred considering only "Perfect" and "Good" OTUs, i.e., OTUs with ≥99% similarity to a reference sequence. To avoid biases due to uneven sequencing depths, data sets were subsampled without replacement to 15,820 sequences per sample.

Based on the results obtained for the artificial communities (see below), pipeline (IV) "fastq_maxee" was selected to analyze the data obtained for the environmental samples. However, additional steps were performed after quality control and OTU clustering. First, OTUs were classified using the lowest common ancestor (LCA) algorithm implemented in CREST (Lanzén et al. 2012) based on the Greengenes database (McDonald et al. 2012), which, for cyanobacteria at the suprageneric level, is ultimately based on Hoffmann et al. (2005). Noncyanobacterial OTUs were removed from downstream analyses. OTU representative sequences were submitted to GenBank under the accession numbers listed in Table S1 in the Supporting Information. Further information on the bioinformatics procedures applied in this study and command-line examples are provided in Appendix S1 in the Supporting Information.

For each cyanobacterial OTU, the most closely related isolate and uncultured sequences deposited in the Ribosomal Database Project (RDP) were retrieved using the SeqMatch tool (Cole et al. 2014). These were aligned using MUSCLE along with the sequences obtained by DGGE and the sequences reported by Fernández-Carazo et al. (2011) for the Forlidas Pond and Lundström Lake samples. A neighbor-joining tree using the Jukes-Cantor model and 1,000 bootstraps was then constructed using TREECON (Van de Peer and De Wachter 1994). At this point, a manual curating step based on taxonomic expertise was carried out to identify erroneous OTUs that might have escaped the automated controls. OTUs that had <97% similarity with their closest SeqMatch isolate and uncultured sequences and were sharing some branch length with others but were distinct and isolated were flagged as being potentially artifactual. In this case, their "intermediate" positioning would indicate a possible chimeric origin, which is supported by their isolation from known cultured and uncultured sequences.

Community dissimilarities at the phylotype level were investigated using unweighted pair group method with arithmetic mean (UPGMA) analysis. For this, data sets were subsampled without replacement to 4,500 sequences per sample to avoid biases due to uneven sequencing depths. Community dissimilarities were computed using Bray–Curtis and Unifrac distances. UPGMA analysis was carried out using the "vegan" package in R (Oksanen et al. 2015).

RESULTS AND DISCUSSION

Evaluation of bioinformatics pipelines using artificial communities. Artificial communities Art1 and Art2 were used to test the performance of five bioinformatics pipelines for quality filtering of reads, removal of chimeric sequences, and OTU clustering. Sequence data were processed separately using mothur and UPARSE, and OTUs were mapped back to the strains used to assemble the artificial communities. Our goal was to assess the overall performance of well-established pipelines currently available to the end user. Thorough comparison between the different pipelines and algorithm parameters can be found in the original mothur (Schloss et al. 2009, 2011) and UPARSE (Edgar 2013) publications.

A total of 61,419 reads were obtained for the two artificial communities (34,213 and 27,206 reads for Art1 and Art2, respectively). After applying the bioinformatics pipelines, between 15,820 and 25,105 reads remained, representing a decrease of up to 47% from the original number of reads. Observed relative abundances differed from the theoretical expectations and were similar in communities Art1 and Art2, despite differences in the initial proportions of each template DNA (Fig. 2). Recovered rel-

ative abundances were also consistent across the different pipelines. In general, *Phormidesmis priestleyi* ANT.L52.6 (ULC026), *Leptolyngbya* sp. "Doroninskoye" (S141), *Microcoleus vaginatus* JR6 (S111), and *Leptolyngbya* sp. ANT07.JR23 (ULC123) were three to six times more abundant than expected, whereas *Leptolyngbya frigida* ANT.L53B.1 (ULC001), *L. frigida* ANT.L52B.3 (ULC029), *Leptolyngbya antarctica* ANT.LWAV6.1 (ULC043), *Cyanobium* sp. "Bylot Island" (ULC065), and *Leptolyngbya* cf. *fragilis* ANT.L52.1 (ULC004) were largely underrepresented.

Phylogenetic analysis showed that some of the strains used to assemble the artificial communities have highly similar 16S rRNA gene sequences and, in fact, the 22 reference strains represent 17 OTUs at 97.5% similarity (Fig. 2). This might explain the disparity between the expected and observed relative abundances of some strains, since sequences with ≥97.5% similarity are clustered into the same OTU and thus information on the relative abundance of individual strains is lost. However, even when this phenomenon is accounted for, a discrepancy between expected and observed relative abundances was still evident.

It has been long acknowledged that differential denaturation, binding efficiency, accessibility of the target gene within genomes, and gene copy number can distort the observed structure of microbial communities in multitemplate PCR (Farrelly et al. 1995, Suzuki and Giovanonni 1996, Polz and Cavanaugh 1998, Sipos et al. 2007, Hong et al. 2009). In addition, the fact that the strains used to assemble the artificial communities were available as unialgal but not axenic cultures might have contributed to an underestimation of the relative abundances of some of the strains. Indeed, the presence of DNA from cocultured bacteria in the extracts used would have affected the actual proportion of the cyanobacterial DNA in the pool. Nevertheless, several similar studies dealing with artificial communities were also unable to perfectly recover the community structure information after PCR and sequencing (Lee et al. 2012, Pinto and Raskin 2012, Shakya et al. 2013). This poses a problem when trying to predict the abundance of an organism in the environment based on the number of reads, but it seems to be an intrinsic bias of the amplicon-sequencing approach (Amend et al. 2010, Ushio et al. 2014).

Inferred phylotype richness varied considerably between pipelines, ranging from 20 to 261 OTUs and from 16 to 155 OTUs in communities Art1 and Art2, respectively (Figs. 3 and S1 in the Supporting Information). Phylotype richness obtained with the

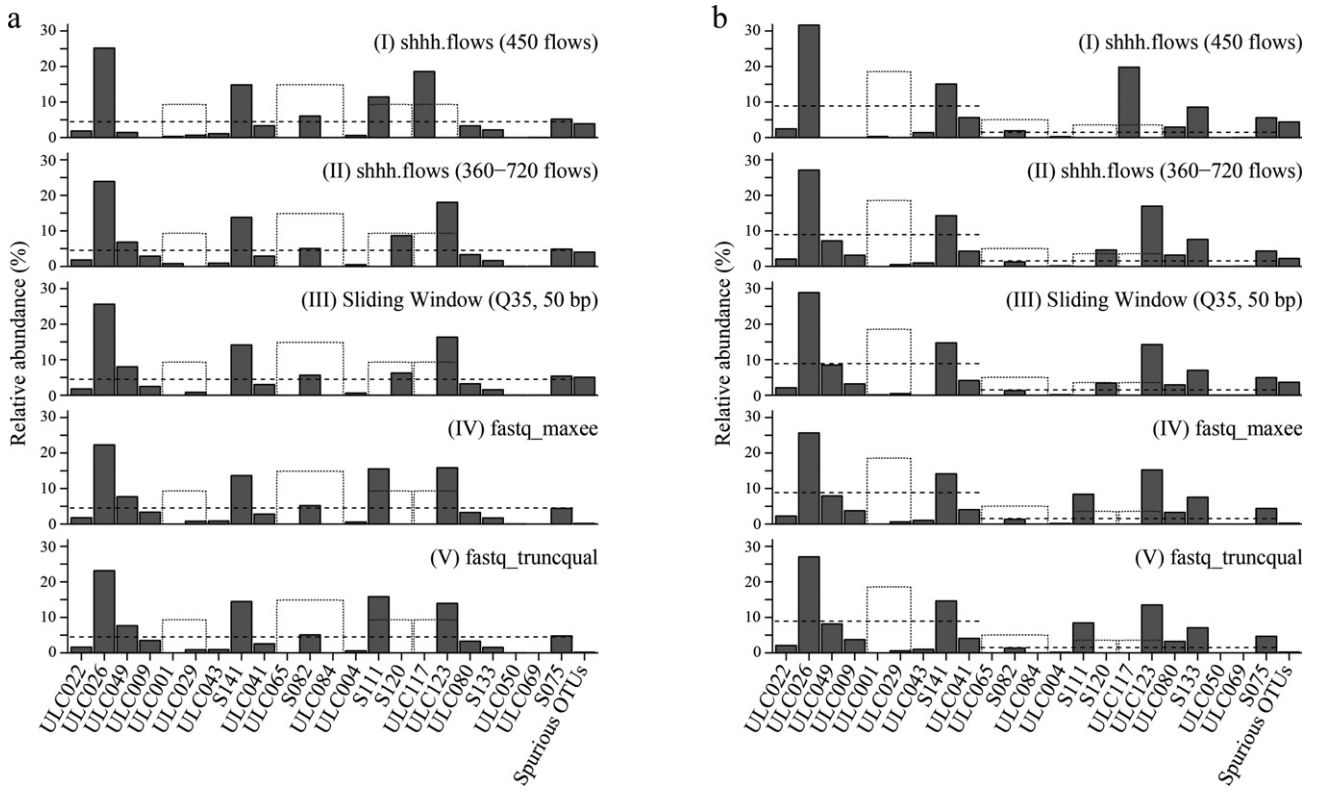

Fig. 2. Recovered relative abundances of reference strains in artificial communities Art1 (a) and Art2 (b) after application of each bioinformatics pipeline. Relative abundances were computed taking into account only OTUs classified as "Perfect" or "Good" (i.e., with ≥99% similarity to a reference sequence). Remaining OTUs ("Noisy," "Other," and "Chimeric") are grouped under "Spurious OTUs." Horizontal dashed lines represent the expected relative abundance of reference strains. Dashed rectangles represent the summed relative abundance of groups of strains with ≥97.5% similar 16S rRNA gene sequences.

mothur-based pipelines (pipelines I-III) was surprisingly high (98–261 OTUs), and rarefaction curves suggested that it would be even higher with increased sequencing depth. OTU richness was much lower (16–21 OTUs) with the UPARSE-based protocols (pipelines IV-V), with rarefaction curves reaching a plateau at around 1,000 sequences. Even at this lower sequencing depth, OTU richness reported by the mothur-based pipelines was 2–3 times higher than expected.

The number of biologically relevant OTUs ("Perfect" and "Good," therefore with ≥99% similarity to a reference sequence) was similar for all pipelines (15–24 OTUs), meaning that all of them were able to identify a consistent number of phylotypes that ultimately correspond to real taxa (Fig. 3). However, the mothur-based protocols reported a high number (82–237 OTUs) of additional spurious phylotypes ("Noisy," "Other," and "Chimeric" OTUs). Although these spurious phylotypes accounted for a small proportion of the data set (0.2%–5.1% of the reads; Fig. 2), they contributed significantly to the overestimation of phylotype richness since their divergent sequences are perceived as new OTUs (Fig. 3). On the other hand, phylotype richness reported by the UPARSE-based protocols was in line with expected results, consisting of 15–16 OTUs with ≥99% similarity to a reference sequence and only 1–5 additional spurious OTUs. Nevertheless, of the 17 expected OTUs, 1–2 OTUs were missing from the UPARSE data sets. No sequence related to *Nostoc* sp. ANT.L34.1 (ULC050) was found in either artificial community, whereas *Pseudanabaena frigida* O-302 (ULC069) was also absent in community Art2. Nevertheless, sequences from these strains were not found in the mothur data sets either, suggesting that their absence was due to PCR and/or sequencing rather than bioinformatics biases.

Bioinformatics pipelines for the analysis of microbial high-throughput amplicon-sequencing data usually consist of three steps: quality filtering of reads, removal of chimeric sequences, and OTU clustering. In mothur, the quality filtering step can be performed using simple quality score thresholds (as in pipeline III) – which is based on the quality score assigned to each base call – or the more computationally demanding flowgram denoising (as in pipelines I and II). Although the per-base error of the 454 pyrosequencing technology is relatively low (Margulies et al. 2005), inconsistencies usually arise from the inaccurate determination of homopolymer lengths. Flowgram denoising algorithms use the flowgram information to assess the confidence of each base call and, therefore, the length of the homopolymer (Quince et al. 2009). Although denoising algorithms represent an improvement over simple quality score-based filtering, they are not able to completely remove sequencing noise and sometimes produce inconsistent results (Bakker et al. 2012, Gaspar and Thomas 2013).

UPARSE, on the other hand, deals with quality filtering of reads in a much simpler way, but nevertheless has been shown to generate highly accurate results, outperforming the methods described above (Edgar 2013). In this case, quality control is simply carried out by truncating sequences when the quality score falls below a determined threshold (as in pipeline V) or based on maximum expected errors, i.e., by using the error probability associated with the quality scores of each base call (as in pipeline IV). The most distinguishing feature of UPARSE is the "UPARSE-OTU" algorithm, which, unlike previous methods, performs chimera filtering and OTU clustering simultaneously. Despite its relatively simple quality filtering approach, community structures obtained in this study using the UPARSE-

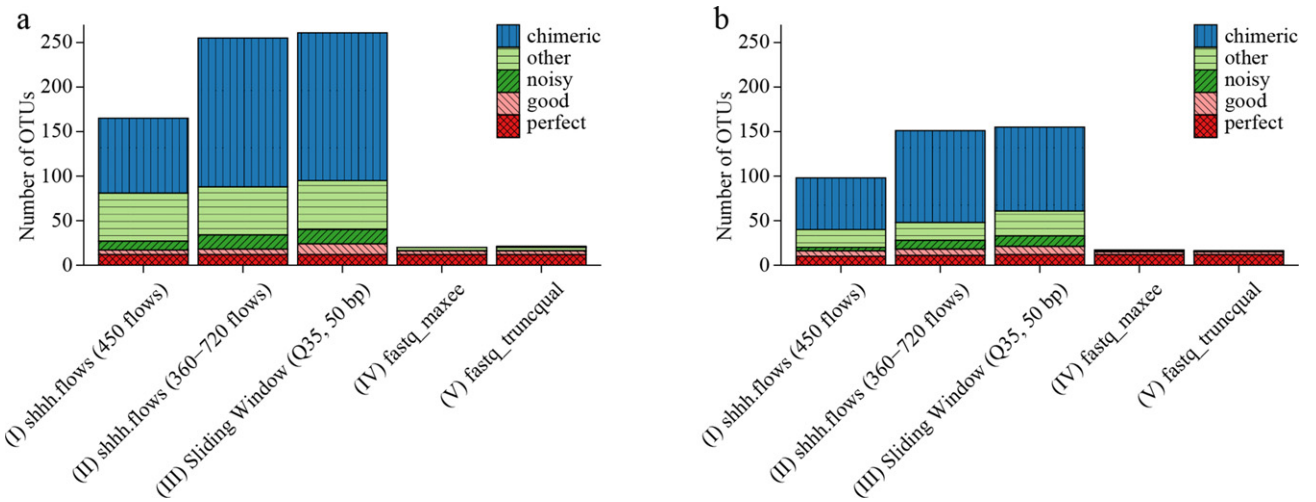

FIG. 3. Classification of OTUs in artificial communities Art1 (a) and Art2 (b) after application of each bioinformatics pipeline. OTUs were classified as "Perfect" (identical to a reference sequence), "Good" (≥99% similarity), "Noisy" (≥97.5% to <99% similarity), "Other" (<97.5% similarity), and "Chimeric" (composed of two or more parent reference sequences).

based pipelines were much more consistent with expected results, mainly due to a drastically improved ability to detect and filter out chimeric sequences (Fig. 3).

Chimeras are usually formed from an incompletely extended amplicon that, in subsequent PCR cycles, anneals to another similar template sequence, thus yielding a hybrid product from multiple parent sequences (Shuldiner et al. 1989). Chimeric sequences are an important source of bias in amplicon-sequencing studies, since they might be indistinguishable from true, novel diversity, especially if formed from closely related sequences (Haas et al. 2011). Many experimental measures can be taken to reduce chimera formation, such as increasing extension time and reducing the number of cycles during PCR, as well as avoiding over shearing of DNA during extraction by using less rigorous cell lysis methods (von Wintzingerode et al. 1997, Ahn et al. 2012). However, none of these approaches has been shown to prevent chimera formation completely, and bioinformatics removal of chimeric sequences still remains a crucial step in amplicon-sequencing studies (Haas et al. 2011).

Description of Antarctic cyanobacterial diversity using HTS. As part of an exploratory effort to validate the use of HTS for the assessment of cyanobacterial diversity, the cyanobacterial community composition in five Antarctic lacustrine microbial mats and one Sub-Antarctic aquatic biofilm was investigated (Fig. 1). All samples were analyzed in parallel using DGGE, and the Forlidas Pond and Lundström Lake samples have also been studied in the past using a polyphasic approach (Fernández-Carazo et al. 2011). Here, we describe the cyanobacterial diversity found in these samples using HTS, assessing the sensitivity of this method for less abundant taxa that are normally missed by traditional techniques. Moreover, samples were used to assess how potential biases introduced at different experimental steps of HTS studies may distort the observed structure of cyanobacterial communities. This included inconsistencies due to the choice of DNA extraction method, stochastic PCR variation, and the use of longer barcoded primers (Appendix S1).

A total of 108,647 16S rRNA gene sequences were obtained for the six samples and submitted to rigorous quality control using the bioinformatics pipeline (IV) "fastq_maxee." Based on the results obtained for the artificial communities, this pipeline showed the highest consistency with expected results among the tested pipelines (see above). After quality control and OTU clustering, 85,290 sequences (78.5%) were retained, of which 83,243 sequences (97.6%) were assigned to cyanobacteria, confirming the specificity of the primer set. From the noncyanobacterial sequences, 498 sequences (0.6%) were assigned to plastid sequences of eukaryotic phototrophs such as chlorophytes, streptophytes, haptophytes, and stramenopiles (heterokonts), and 1,549 sequences (1.8%) were assigned to bacterial phyla such as Acidobateria, Chloroflexi, Planctomycetes, Proteobacteria, TM7, and Verrucomicrobia. Cyanobacterial sequences were clustered into 108 OTUs at 97.5% similarity (Table S1). However, following a manual inspection of OTUs according to SeqMatch and phylogenetic analyses, 37 OTUs were flagged as being potentially artifactual.

Overall, taxonomic, SeqMatch, and phylogenetic analyses of OTUs showed that the retrieved diversity covers a broad spectrum of the overall cyanobacterial diversity (Fig. 4, Table S1 and Fig. S2 in the Supporting Information). Taxonomic assignment of sequences revealed the presence of five cyanobacterial orders (Fig. 4a). Pseudanabaenales was the most OTU-rich order (60 OTUs), followed by Oscillatoriales (10 OTUs), Chroococcales (8 OTUs), Synechococcales and Nostocales (7 OTUs each). No sequences related to Gloeobacterales were found, and 16 OTUs could not be classified at the order level by the LCA classifier algorithm. At the genus level (Fig. 4b), OTUs were classified as *Leptolyngbya* (22 OTUs), *Phormidium*, *Pseudanabaena* (4 OTUs each), *Synechococcus*, *Nostoc* (2 OTUs each), *Planktothrix*, *Cyanobacterium*, *Acaryochloris*, *Arthrospira*, *Calothrix,* and *Rivularia* (1 OTU each), and 68 OTUs remained unclassified at the genus level. With some exceptions, taxonomic classification of OTUs using the LCA classifier algorithm and the Greengenes database was consistent with the taxonomy of the best SeqMatch isolate hit (Table S1).

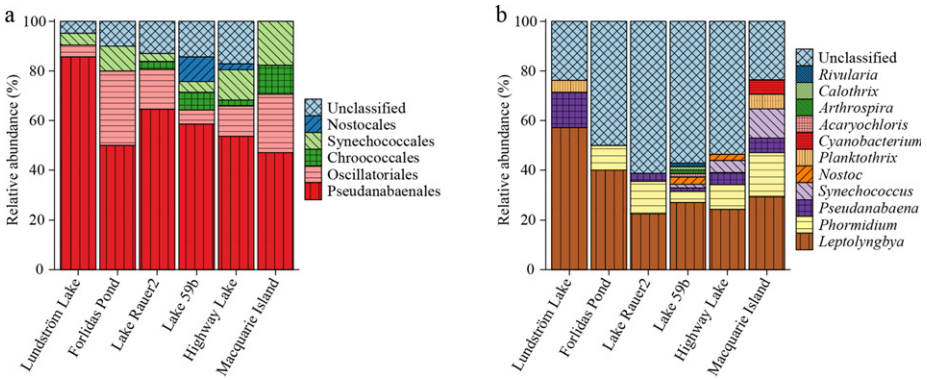

FIG. 4. Taxonomic assignment of cyanobacterial OTUs at the order (a) and genus (b) level. Cyanobacterial suprageneric classification is shown according to Hoffmann et al. (2005). OTUs were determined using a similarity cut-off value of 97.5%. Relative abundances were computed by dividing the number of OTUs assigned to each order/genus by the total number of OTUs in each sample.

Differences in community structures were observed between the samples (Fig. 4). As expected given its geographic origin, the aquatic biofilm community from Macquarie Island (Sub-Antarctic) was very distinct from the communities of the five Antarctic microbial mat samples at the order level, with a higher proportion of Synechococcales and Chroococcales OTUs. Interestingly, OTUs unclassified at the order level were nearly absent in this community, suggesting a cyanobacterial community composed by rather "traditional" taxa. Communities from the two Transantarctic Mountains lakes (Forlidas Pond and Lundström Lake) were distinguished from the remaining samples by an absence of Chroococcales OTUs. Finally, the communities from the three East Antarctic lakes (Lake 59b, Lake Rauer2, and Highway Lake) were relatively similar, although Nostocales OTUs were not found in the Lake Rauer2 sample. Communities were very dissimilar at the genus level, with the Lake 59b sample harboring a much more diverse community than the remaining lakes.

The 10 most abundant OTUs taking all samples together accounted for 84.9% of the reads (Table S1). With the exception of OTU10, they were all assigned to the filamentous form-genera *Leptolyngbya* and *Phormidium*, which are common taxa in Antarctic lacustrine microbial mats (Vincent and Quesada 2012). OTU1 and OTU9 were identical to strains of *L. antarctica* isolated from Ace Lake (Vestfold Hills) and Firelight Lake (Bølingen Islands), respectively (Taton et al. 2006b). OTU3 was 99.2% similar to an unidentified cyanobacterium (LPP-group MBIC 10597) and OTU6 was distantly related (97.8% similarity) to a strain of *Leptolyngbya nostocorum* isolated from a biofilm in the Guadarrama River, Spain (Loza et al. 2013). SeqMatch analysis of other dominant OTUs showed phylotypes highly related (≥99.5% similarity) to strains identified as *Phormidium* cf. *uncinatum* CYN108 (OTU2), *Phormidium lumbricale* UTCC 476 (OTU5), *Synechococcus* sp. PS845 (OTU10), and *Leptolyngbya* sp. LEGE 07074 (OTU14).

Two dominant OTUs had extremely low similarity to their best SeqMatch isolate hits (Table S1). OTU4 had 93.4% similarity to an unidentified lichen photobiont (Miura and Yokota 2006), but was also 99.7% similar to an environmental sequence obtained from a stromatolite in Spain (Santos et al. 2010), suggesting that, rather than a sequencing artifact, this phylotype likely represents a not yet described cyanobacterial taxon. Similarly, OTU7 was 89.6% similar to cf. *Leptolyngbya* sp. Greenland_9 isolated from a hot spring in Greenland (Roeselers et al. 2007). However, in this case, an additional BLAST analysis showed that this phylotype is in fact 99.0% similar to a strain identified as *Oscillatoria* sp. KNUA009, recently isolated from a meltwater stream in King George Island, Antarctic Peninsula (Hong et al. 2013).

A striking observation was the occurrence in the Lake 59b sample (Larsemann Hills) of one OTU affiliated with *Arthrospira* (OTU73, 99.2% similar to *Arthrospira platensis*), a taxon that is currently known to have a tropical and semitropical distribution (Sili et al. 2012). However, it is known that huge amounts of dust are transported between continents every year, serving as a vector for the dispersion of microorganisms around the globe, including to and from Antarctica (Pearce et al. 2009). Most likely, the *Arthrospira*-related OTU found in this study was originated from dead cell material introduced from outside the Antarctic continent rather than from an actual resident population, especially considering that it was reported for only one sample and at very low read abundance (0.1%). Nevertheless, the introduction of alien species in Antarctic environments is a growing concern particularly in light of the expected effects of climate change, which are likely to shift the distributional limits of temperate species toward the poles (Convey et al. 2014). In this sense, due to its unprecedented analytical depth, HTS could be a useful tool for monitoring the development of invasive populations in Antarctic environments at very early stages.

Cyanobacterial community structures differed considerably between the samples at the phylotype level, with 62 OTUs (57.4%) being found in only one sample (Table S1). The single phylotype shared between all the six samples and the most abundant in the whole data set (OTU1) was identical to *L. antarctica* ANT.LAC.1 (strain ULC041 in the BCCM/ULC Cyanobacteria Collection) isolated from Ace Lake (Taton et al. 2006a,b), confirming the ubiquity of this taxon throughout Antarctica (Komárek 2007, Komárek et al. 2008). Five OTUs were common to five samples, and included phylotypes highly related (≥98.1% similarity) to strains identified as *Phormidium* cf. *uncinatum* CYN108 (OTU2), *P. priestleyi* ANT.L52.6 (OTU90), and *Chamaesiphon subglobosus* PCC 7430 (OTU39).

Mirroring the differences observed at the order level (Fig. 4), UPGMA analysis showed that the aquatic biofilm community from Macquarie Island was distinct from the remaining communities also at the phylotype level (Fig. S3 in the Supporting Information). In relation to the Antarctic lacustrine communities, UPGMA analysis clustered the samples irrespective of their geographic location inside the continent, indicating similar cyanobacterial communities in Lake Rauer2 (Rauer Islands), Forlidas Pond (Dufek Massif), and Highway Lake (Vestfold Hill), which might be related to the salinity of these lakes. Lake Rauer2 and Highway Lake are hypersaline and slightly brackish, respectively, and although Forlidas Pond is a freshwater lake, it is evaporating and contains a layer of hypersaline brine at the bottom (Hodgson et al. 2010). In line with this observation, salinity has been shown as an important factor structuring the cyanobacterial com-

munities of Antarctic lakes (Sabbe et al. 2004, Jungblut et al. 2005, Taton et al. 2006a).

Using a polyphasic approach, Fernández-Carazo et al. (2011) reported an impoverished cyanobacterial diversity in microbial mats from Forlidas Pond and Lundström Lake (Transantarctic Mountains, 80–82° S). In our study, these were also the lakes with the lowest cyanobacterial diversity among the studied samples. Phylogenetic analysis (Fig. S2) showed that the widely distributed OTU1 (see above) corresponds to *L. antarctica* TM1FOS73 (strain ULC073) isolated from the bottom brine of Forlidas Pond. Likewise, one phylotype related to *Phormidium murrayi* TM2ULC130 (ULC130) isolated from Forlidas Pond was also recovered (OTU2). Finally, several phylotypes found in the Lundström Lake sample (OTU4, OTU15, and OTU32) and Forlidas Pond (OTU7, OTU8, and OTU16) are related to DGGE and clone sequences previously reported for these lakes, but given the low similarity to sequences from cultured strains, they likely represent novel cyanobacterial taxa not yet in culture.

On the other hand, HTS revealed the presence of several additional phylotypes that had not been previously reported for Forlidas Pond and Lundström Lake (Fig. S2 and Table S1). These were mainly present at low relative abundances (<0.5%) with the exception of OTU6, which corresponded to 23.7% of the reads in the Lundström Lake sample. Novel phylotypes reported for these lakes were highly related (≥98.1% similarity) to different taxa, such as *C. subglobosus* (OTU39), *Oscillatoria* sp. (OTU23), *L. frigida* (OTU43), *Leptolyngbya tenerrima* (OTU115), *P. priestleyi* (OTU90 and OTU132), and *Synechococcus* sp. (OTU52). Finally, three phylotypes previously unreported for these lakes (OTU6, OTU13, and OTU153) had low similarity with sequences from cultured strains, but were highly related (≥98.9%) to sequences retrieved from environmental samples (Table S1).

The improvement represented by HTS in comparison to DGGE becomes even more apparent with the results obtained for the Lake 59b sample (Larsemann Hills). In addition to the three OTUs found in this sample by DGGE, related to *Nostoc edaphicum* (OTU121, 99.2% similarity), *L. nostocorum* (OTU6, 97.8% similarity), and *L. antarctica* (OTU26, 99.5% similarity), 67 additional OTUs were detected by HTS (Fig. S2 and Table S1). Unlike other samples, novel phylotypes were also found at high abundances (up to 13.8% of the reads). These were highly related (≥98.1%) to diverse taxa such as *Synechococcus* sp. (OTU10 and OTU52), *P. murrayi* (OTU2), *P. priestleyi* (OTU21, OTU84, OTU90, OTU132, and OTU135), *P. lumbricale* (OTU5), *Geitlerinema carotinosum* (OTU16), *Coleodesmium* sp. (OTU30), *Calothrix* sp. (OTU35), *Plectolyngbya hodgsonii* (OTU38), and *A. platensis* (OTU73). As for the Forlidas Pond and Lundström Lake samples, several phylotypes had low similarity with sequences from cultured strains but were related to environmental sequences. The same trend was observed for the other studied samples. Overall, 14, 28, and 38 additional OTUs were found by HTS in the Macquarie Island, Lake Rauer2, and Highway Lake samples, respectively, in comparison to the results obtained by DGGE (Fig. S2).

CONCLUSIONS

Altogether, this study shows that HTS is a valuable tool for the assessment of cyanobacterial diversity in Antarctic lakes, being able to detect several low-abundance taxa that are normally missed by traditional methods. Chimeric sequences contributed to a significant overestimation of cyanobacterial diversity in artificial communities and recovered relative abundances did not reflect the original abundance of the taxa in the communities. Nevertheless, recovered community structures were pipeline dependent, with UPARSE being able to generate the most accurate community structure information due to a more efficient detection of chimeric sequences. Still, although the results obtained with UPARSE were more consistent with the expected results, we recognize that this may not hold true for other studies (e.g., targeting different regions of the 16S rRNA gene). More appropriately, these findings show the benefits of using controlled artificial communities to validate the experimental and bioinformatics parameters used to characterize natural communities in HTS studies. These results show that HTS data for cyanobacterial communities should be interpreted with caution, as read abundances may be prone to PCR amplification biases and, therefore, may not reflect the real abundance of taxa in the environment.

Furthermore, our study corroborates the results obtained with traditional techniques but, as expected, revealed a much higher cyanobacterial diversity in Antarctic aquatic samples. In agreement with the current knowledge, samples were dominated by phylotypes related to the filamentous form-genera *Leptolyngbya* and *Phormidium*. However, several additional phylotypes were detected at low abundances, giving insights into a cyanobacterial "rare biosphere." The majority of these rare OTUs had not been previously reported for the studied samples, showing that other filamentous taxa such as *Geitlerinema* and *Pseudanabaena*, unicellular taxa such as *Synechococcus* and *Chamaesiphon*, and heterocystous taxa such as *Calothrix* and *Coleodesmium* may also be important components of Antarctic lacustrine communities. UPGMA analysis evidenced a distinct cyanobacterial community in the aquatic biofilm from Macquarie Island (Sub-Antarctic) and also suggested that lake salinity, rather than geographic location, is an important factor influencing the community structure of Antarctic lakes. These findings are in line with the current knowledge on cyanobac-

terial distribution in Antarctica, but additional research comprising a larger number of samples from different Antarctic provinces is needed to provide a more complete picture.

Another remarkable result was the high incidence of OTUs unclassified at the genus level and with low similarity to sequences from cultured strains available in RDP. A large fraction of this unidentified diversity can be presumed to correspond to real but not yet described cyanobacterial taxa, considering that many were similar to sequences reported by other molecular studies. Nevertheless, the confirmation of the biological authenticity of such unidentified diversity remains a challenge, and, in our understanding, it is not a task that can be resolved by HTS studies alone. As any other indirect molecular approach for the assessment of microbial diversity, HTS results have to be scrutinized from an ecological perspective, and additional work using different techniques has to be carried out to confirm or discard ambiguous findings. In this sense, the isolation and characterization of cyanobacterial strains would be a useful complement. More appropriately, the improved coverage provided by HTS might be particularly important for beta-diversity–driven investigations such as bioregionalization and temporal studies, allowing the investigation of community dissimilarities at a much deeper scale.

This work was supported by the BELSPO projects CCAMBIO (SD/BA/03A), BELDIVA (EA/00/05), and AMBIO (SD/BA/01A), the FRFC project 2.4570.09, and the FRS-FNRS credit CR.CH.10-11-1.5139.11. IS Pessi had a PhD FRIA fellowship from the FRS-FNRS. HD Laughinghouse IV had a mobility postdoctoral fellowship from the FRS-FNRS. A Wilmotte is a research associate of the FRS-FNRS. The authors would like to thank CW Herbold, J Laybourn-Parry, D Velazquez, J Kleinteich, Y Lara, A. Murray, and L Cappelatti for valuable suggestions and discussion. Useful comments from an anonymous reviewer are also very much acknowledged. We are very grateful to Dr D Bergström and the Australian Antarctic Division for providing the sample from Macquarie Island, and to Dr D Hodgson, Prof J Laybourn-Parry and Paul Dyer for the remaining samples. This study is a contribution to the State of the Antarctic Ecosystem (AntEco) research program of the Scientific Committee on Antarctic Research (SCAR). The authors declare no conflict of interest.

Ahn, J.-H., Kim, B.-Y., Song, J. & Weon, H.-Y. 2012. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J. Microbiol.* 50:1071–4.

Amend, A. S., Seifert, K. A. & Bruns, T. D. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol. Ecol.* 19:5555–65.

Bakker, M. G., Tu, Z. J., Bradeen, J. M. & Kinkel, L. L. 2012. Implications of pyrosequencing error correction for biological data interpretation. *PLoS ONE* 7:e44357.

Berry, D., Mahfoudh, K. B., Wagner, M. & Loy, A. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. *Appl. Environ. Microbiol.* 77:7846–9.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2:e197.

Boutte, C., Grubisic, S., Balthasart, P. & Wilmotte, A. 2006. Testing of primers for the study of cyanobacterial molecular diversity by DGGE. *J. Microbiol. Meth.* 65:542–50.

Brandes, M., Albach, D. C., Vogt, J. C., Mayland-Quellhorst, E., Mendieta-Leiva, G., Golubic, S. & Palinska, K. A. 2015. Supratidal extremophiles-cyanobacterial diversity in the rock pools of the Croatian Adria. *Microb. Ecol.* 70:876–88.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R. & Tiedje, J. M. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42:D633–42.

Convey, P., Chown, S. L., Clarke, A., Barnes, D. K. A., Bokhorst, S., Cummings, V., Ducklow, H. W. et al. 2014. The spatial structure of Antarctic biodiversity. *Ecol. Monogr.* 84: 203–44.

Dadheech, P. K., Glöckner, G., Casper, P., Kotut, K., Mazzoni, C. J., Mbedi, S. & Krienitz, L. 2013. Cyanobacterial diversity in the hot spring, pelagic and benthic habitats of a tropical soda lake. *FEMS Microbiol. Ecol.* 85:389–401.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–7.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–1.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10:996–8.

Farrelly, V., Rainey, F. A. & Stackebrandt, E. 1995. Effect of genome size and *rrn* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl. Environ. Microbiol.* 61:2798–801.

Fernández-Carazo, R., Hodgson, D. A., Convey, P. & Wilmotte, A. 2011. Low cyanobacterial diversity in biotopes of the Transantarctic Mountains and Shackleton Range (80-82°S), Antarctica. *FEMS Microbiol. Ecol.* 77:503–17.

Gaspar, J. M. & Thomas, W. K. 2013. Assessing the consequences of denoising marker-based metagenomic data. *PLoS ONE* 8: e60458.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D. et al. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21:494–504.

Hodgson, D. A., Convey, P., Verleyen, E., Vyverman, W., McInnes, S. J., Sands, C. J., Fernández-Carazo, R. et al. 2010. The limnology and biology of the Dufek Massif, Transantarctic Mountains 82° South. *Polar Sci.* 4:197–214.

Hodgson, D. A., Noon, P. E., Vyverman, W., Bryant, C. L., Gore, D. B., Appleby, P., Gilmour, M. et al. 2001a. Were the Larsemann Hills ice-free through the Last Glacial Maximum? *Antarct. Sci.* 13:440–54.

Hodgson, D. A., Vyverman, W. & Sabbe, K. 2001b. Limnology and biology of saline lakes in the Rauer Islands, eastern Antarctica. *Antarct. Sci.* 13:255–70.

Hoffmann, L., Komárek, J. & Kaštovský, J. 2005. System of cyanoprokaryotes (cyanobacteria) – state in 2004. *Algolog. Stud.* 117:95–115.

Hong, S., Bunge, J., Leslin, C., Jeon, S. & Epstein, S. S. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3:1365–73.

Hong, J. W., Kim, S. H., Choi, H. G., Kang, S. H. & Yoon, H. S. 2013. Isolation and description of a globally distributed cryosphere cyanobacterium from Antarctica. *Polar Res.* 32:18529.

Hur, M., Lee, I., Tak, B. M., Lee, H. J., Yu, J. J., Cheon, S. U. & Kim, B. S. 2013. Temporal shifts in cyanobacterial communities at different sites on the Nakdong River in Korea. *Water Res.* 47:6973–82.

Jungblut, A. D., Hawes, I., Mountfort, D., Hitzfeld, B., Dietrich, D. R., Burns, B. P. & Neilan, B. A. 2005. Diversity within cyanobacterial mat communities in variable salinity meltwater ponds of McMurdo Ice Shelf, Antarctica. *Environ. Microbiol.* 7:519–29.

Jungblut, A. D., Lovejoy, C. & Vincent, W. F. 2010. Global distribution of cyanobacterial ecotypes in the cold biosphere. *ISME J.* 4:191–202.

Kleinteich, J., Hildebrand, F., Wood, S. A., Cirés, S., Agha, R., Quesada, A., Pearce, D. A., Convey, P., Küpper, F. C. & Dietrich, D. R. 2014. Diversity of toxin and non-toxin containing cyanobacterial mats of meltwater ponds on the Antarctic Peninsula: a pyrosequencing approach. *Antarct. Sci.* 26:521–32.

Komárek, J. 2007. Phenotype diversity of the cyanobacterial genus *Leptolyngbya* in the maritime Antarctic. *Pol. Polar Res.* 28:211–31.

Komárek, J., Elster, J. & Komárek, O. 2008. Diversity of the cyanobacterial microflora of the northern part of James Ross Island, NW Weddell Sea, Antarctica. *Polar Biol.* 31:853–65.

Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12:118–23.

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L. & Urich, T. 2012. CREST – classification resources for environmental sequence tags. *PLoS ONE* 7:e49334.

Laybourn-Parry, J. & Wadham, J. L. 2014. *Antarctic Lakes*, 1st edn. Oxford University Press, Oxford, UK, 215 pp.

Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R. & Cary, S. C. 2012. Groundtruthing next-gen sequencing for microbial ecology – biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS ONE* 7:e44224.

Loza, V., Berrendero, E., Perona, E. & Mateo, P. 2013. Polyphasic characterization of benthic cyanobacterial diversity from biofilms of the Guadarrama River (Spain): morphological, molecular, and ecological approaches. *J. Phycol.* 49:282–97.

Lynch, M. D. J. & Neufeld, J. D. 2015. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* 13:217–29.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J. et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., Andersen, G. L., Knight, R. & Hugenholtz, P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6:610–8.

Miura, S. I. & Yokota, A. 2006. Isolation and characterization of cyanobacteria from lichen. *J. Gen. Appl. Microbiol.* 52:365–74.

Nübel, U., Garcia-Pichel, F. & Muyzer, G. 1997. PCR primers to amplify 16S rRNA genes from Cyanobacteria. *Appl. Environ. Microbiol.* 63:3327–32.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. & Wagner, H. 2015. Vegan: community ecology package. R package version 2.3-1. Available at http://CRAN.R-project.org/package=vegan.

Pearce, D. A., Bridge, P. D., Hughes, K. A., Sattler, B., Psenner, R. & Russell, N. J. 2009. Microorganisms in the atmosphere over Antarctica. *FEMS Microbiol. Ecol.* 69:143–57.

Pinto, A. J. & Raskin, L. 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE* 7:e43093.

Polz, M. F. & Cavanaugh, C. M. 1998. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* 64:3724–30.

Quince, C., Lanzén, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., Read, L. F. & Sloan, W. T. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6:639–41.

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., Daroub, S. H., Camargo, F. A. O., Farmerie, W. G. & Triplett, E. W. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.* 1:283–90.

Roeselers, G., Norris, T. B., Castenholz, R. W., Rysgaard, S., Glud, R. N., Kühl, M. & Muyzer, G. 2007. Diversity of phototrophic bacteria in microbial mats from Arctic hot springs (Greenland). *Environ. Microbiol.* 9:26–38.

Sabbe, K., Hodgson, D. A., Verleyen, E., Taton, A., Wilmotte, A., Vanhoutte, K. & Vyverman, W. 2004. Salinity, depth and the structure and composition of microbial mats in continental Antarctic lakes. *Freshwater Biol.* 49:296–319.

Santos, F., Peña, A., Nogales, B., Soria-Soria, E., García del Cura, M. Á., González-Martín, J. A. & Antón, J. 2010. Bacterial diversity in dry modern freshwater stromatolites from Ruidera Pools Natural Park, Spain. *Syst. Appl. Microbiol.* 33:209–21.

Schloss, P. D., Gevers, D. & Westcott, S. L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A. et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75:7537–41.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W. & Podar, M. 2013. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. Microbiol.* 15:1882–99.

Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. 2012. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* 21:1794–805.

Shuldiner, A. R., Nirula, A. & Roth, J. 1989. Hybrid DNA artifact from PCR of closely related target sequences. *Nucleic Acids Res.* 17:4409.

Sili, C., Torzillo, G. & Vonshak, A. 2012. *Arthrospira* (Spirulina). *In* Whitton, B. A. [Ed.] *Ecology of Cyanobacteria II*. Springer, Dordrecht, Netherlands, pp. 371–85.

Singh, S. M. & Elster, J. 2007. Cyanobacteria in Antarctic lake environments. *In* Seckbach, J. [Ed] *Algae and Cyanobacteria in Extreme Environments*. Springer, Dordrecht, Netherlands, pp. 303–20.

Sipos, R., Székely, A. J., Palatinszky, M., Révész, S., Márialigeti, K. & Nikolausz, M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol. Ecol.* 60:341–50.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* 103:12115–20.

Suzuki, M. T. & Giovanonni, S. J. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* 62:625–30.

Taton, A., Grubisic, S., Balthasart, P., Hodgson, D. A., Laybourn-Parry, J. & Wilmotte, A. 2006a. Biogeographical distribution and ecological ranges of benthic cyanobacteria in East Antarctic lakes. *FEMS Microbiol. Ecol.* 57:272–89.

Taton, A., Grubisic, S., Brambilla, E., De Wit, R. & Wilmotte, A. 2003. Cyanobacterial diversity in natural and artificial microbial mats of Lake Fryxell (McMurdo Dry Valleys, Antarctica): a morphological and molecular approach. *Appl. Environ. Microbiol.* 69:5157–69.

Taton, A., Grubisic, S., Ertz, D., Hodgson, D. A., Piccardi, R., Biondi, N., Tredici, M. R., Mainini, M., Losi, D., Marinelli, F. & Wilmotte, A. 2006b. Polyphasic study of Antarctic cyanobacterial strains. *J. Phycol.* 42:1257–70.

Taton, A., Hoffmann, L. & Wilmotte, A. 2008. Cyanobacteria in microbial mats of Antarctic lakes (East Antarctica) – a microscopical approach. *Algolog. Stud.* 126:173–208.

Tringe, S. G. & Hugenholtz, P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* 11:442–6.

Ushio, M., Makoto, K., Klaminder, J., Takasu, H. & Nakano, S.-I. 2014. High-throughput sequencing shows inconsistent results with a microscope-based analysis of the soil prokaryotic community. *Soil Biol. Biochem.* 76:53–6.

Van de Peer, Y. & De Wachter, R. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Bioinformatics* 10:569–70.

Vincent, W. F. & Quesada, A. 2012. Cyanobacteria in high latitude lakes, rivers and seas. *In* Whitton, B. A. [Ed.] *Ecology of Cyanobacteria II*. Springer, Dordrecht, Netherlands, pp. 371–85.

Waterbury, J. B. 2006. The Cyanobacteria – isolation, purification and identification. *In* Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E. [Eds.] *The prokaryotes*. Springer, New York, USA, pp. 1053–73.

Wilmotte, A. & Golubic, S. 1991. Morphological and genetic criteria in the taxonomy of Cyanophyta/Cyanobacteria. *Algolog. Stud.* 64:1–24.

von Wintzingerode, F., Göbel, U. B. & Stackebrandt, E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* 21:213–29.

Woese, C. R. & Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74:5088–90.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Figure S1. Rarefaction curves for artificial communities Art1 (a) and Art2 (b) after application of each bioinformatics pipeline.

Figure S2. Neighbor-joining tree using the Jukes-Cantor model and 1,000 bootstraps of OTUs obtained by HTS (in bold), their best SeqMatch isolate and uncultured hits, sequences obtained by DGGE (in blue) and sequences reported by Fernández-Carazo et al. (2011) for the Forlidas Pond and Lundström Lake samples (in red). Bootstraps values >50% are indicated. OTUs marked with an asterisk were flagged as being potentially artifactual.

Figure S3. UPGMA clustering of the studied cyanobacterial communities. Distances correspond to community dissimilarities as computed by the Bray–Curtis index. UPGMA analysis using Unifrac distances yielded the same topology and is therefore not shown.

Table S1. Distribution of cyanobacterial OTUs, taxonomic assignment, and related SeqMatch isolate and uncultured sequences.

Appendix S1. Assessment of potential experimental biases and additional information on the bioinformatics procedures.