Biogeographic Distribution of Five Antarctic Cyanobacteria Using Large-Scale k-mer Searching with sourmash branchwater

Authors: Jessica Lumian, Dawn Sumner, Christen Grettenberger, Anne D. Jungblut, Luiz Irber, N. Tessa Pierce-Ward, C. Titus Brown (corresponding)

ABSTRACT

Cyanobacteria form diverse communities and are important primary producers in Antarctic freshwater environments, but their geographic distribution patterns in Antarctica and globally are still unresolved. There are however few genomes of cultured cyanobacteria from Antarctica available and therefore metagenome-assembled genomes (MAGs) from Antarctic cyanobacteria microbial mats provide an opportunity to explore distribution of uncultured taxa. These MAGs also allow comparison with metagenomes of cyanobacteria enriched communities from a range of habitats, geographic locations, and climates. However, most MAGs do not contain 16S rRNA gene sequences, making a 16S rRNA gene-based biogeography comparison difficult. An alternative technique is to use large-scale k-mer searching to find genomes of interest in public metagenomes.

This paper presents the results of k-mer based searches for 5 Antarctic cyanobacteria MAGs from Lakes Fryxell and Lake Vanda, assigned the names *Phormidium pseudopriestleyi*, a *Microcoleus,* a *Leptolyngbya*, a *Pseudanabaena*, and a *Neosynechococcus* (Lumian et al., 2021, Lumian et al., 2022, in prep.) in 498,942 unassembled metagenomes from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). The *Microcoleus* MAG was found in a wide variety of environments, *P. pseudopriestleyi* was found in environments with challenging conditions, the *Neosynechococcus* was only found in Antarctica, and the *Leptolyngbya* and *Pseudanabaena* MAGs were found in Antarctic and other cold environments. The findings based on metagenome matches and global comparisons suggest that these Antarctic cyanobacteria have distinct distribution patterns ranging from locally restricted to global distribution across the cold biosphere and other climatic zones.

INTRODUCTION

Cyanobacteria are a diverse group of oxygenic photosynthetic gram-negative bacteria that are prevalent in a wide range of environments. In polar environments, cyanobacteria play an important part in shaping local ecology because of their role as primary producers underlying habitats such as benthic biofilms (Stal, 2007; Quesada & Vincent, 2012; Chrismas et al., 2016). Cyanobacteria that thrive in Antarctica face many challenges including variable light availability, cold temperatures, and freeze-drying conditions. To withstand these conditions, cyanobacteria may have tolerance mechanisms encoded in their genomes (Chrismas et al., 2015, 2016). However, the presence of tolerance genes in their genomes may make it more difficult for polar cyanobacteria to compete with other cyanobacteria in non-polar environments. Consequently, some polar cyanobacteria may only occur in polar environments, while others may also be present in environments that share similar conditions to the stresses they face in Antarctica, such as cold temperatures or light stress (Jungblut et al., 2016; Chrismas, Williamson, et al., 2018b; Lumian et al., 2021).

Currently, polar cyanobacteria are underrepresented in genomic databases, despite the important role they play in primary productivity including in perennially ice-covered lakes in the McMurdo Dry Valleys in Antarctica. Due to a lack of grazers and limited water mixing, vast microbial mats in the Lakes Vanda and Fryxell prosper and sustain complex geochemical gradients (Jungblut et al., 2016; Sumner et al., 2016; Lumian et al., 2021). These geochemical gradients structure competition within communities, which are also dealing with challenging environmental conditions, such as highly seasonal light, nutrient, limitation, and in part of Lake Fryxell, sulfidic water (Lumian et al., 2021; Jungblut et al. 2016; Dillon et al. 2020).

The question of why Antarctic cyanobacteria can survive in challenging conditions and what other environments they grow in can be addressed by biogeography studies (Whitaker et al., 2003;

Martiny et al., 2006; Fierer, 2008; Green et al. 2008). Previous 16S rRNA gene surveys based on amplicon sequencing provided support for the longstanding theory that microbes have unlimited dispersal and community distribution is selected by the environment (Baas-Becking, 1934; Jungblut et al. 2010; Harding et al 2011). However, studies from other environments and climatic zones have shown that 16S rRNA gene and single gene markers might not provide sufficient resolution to resolve genotypes and populations.

Most biogeography studies on polar microbiomes and cyanobacteria to date are based on 16S rRNA gene amplicon sequencing in the context of local environmental conditions of sampling sites or pole-to-pole comparisons using clone library surveys and high throughput sequencing approaches (Taton et al., 2006; Namsaraev et al., 2010; Jungblut et al. 2011; Bahl et al., 2011; Moreira et al., 2013; Harke et al., 2016; Kleinteich et al. 2017; Ribeiro et al., 2018). Although 16S rRNA gene sequences are computationally easier to compare to each other, there are limitations to 16S rRNA gene-based biogeography studies. The 16S rRNA gene is conserved and therefore likely leads to an under estimation of genotype level richness. Furthermore, the short read length of high throughput sequencing only allows the coverage of a few variable regions which further reduces phylogenetic resolution. While recent genomic work has provided advances in biogeography of polar microbes (Chrismas et al. 2015), the 16S rRNA gene sequence may not assemble and bin well from metagenomes, which can prohibit MAGs from being incorporated into 16S rRNA gene-based biogeographical distributions.

An alternative to 16S rRNA gene-based biogeography is to apply comparative genomic approaches, but this is computationally more complicated because of the size and scale of metagenome datasets. One option is to use an alignment-based approach in which the reads are aligned to reference genomes, which has been done for large-scale viral genome discovery with Serratus (Edgar et al., 2022). Another option is to apply large-scale k-mer matching to unassembled metagenomes, which is possible with sourmash branchwater (Irber et al., 2022; Brown & Irber, 2016; Irber, 2020a, 2020b; Brown, 2021).

These techniques open the possibility of using metagenomic data for biogeography studies by searching all publicly available metagenomes on the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (Leinonen et al., 2011) for Antarctic MAGs of interest. In this paper, branchwater was used to search 498,942 unassembled metagenomes from the NCBI SRA for the presence of five Antarctic cyanobacteria MAGs that lack the 16S rRNA gene. These findings provide new insights based on comparative genomic analyses into the distribution patterns of cyanobacteria in cold biospheres.

MATERIALS AND METHODS

Selection of Antarctic Cyanobacteria of Interest

Phormidium pseudopriestleyi is a well characterized cyanobacteria in Lake Fryxell, Antarctica (Lumian et al., 2021). Lake Fryxell is a perennially ice-covered lake located at 77.36° S, 162.6° E in the McMurdo Dry Valleys. The base of the lake is covered with microbial mats, with *P. pseudopriestleyi* dominating the mats at 9.8 m in depth, where light levels are low (1-2 µmol photons m^{-2} s^{-1}) and sulfide is present in the water column (0.091 mg L^{-1}). *P. pseudopriestleyi* performs oxygenic photosynthesis in the presence of hydrogen sulfide, even though sulfide inhibits oxygenic photosynthesis (Sumner et al., 2015; Lumian et al., 2021). Lake conditions and sampling have been described in Jungblut et al. (2016), Dillon et al. (2020), and Lumian et al. (2021).

The *Neosynechococcus*, *Leptolyngbya*, *Microcoleus*, and *Pseudanabaena* MAGs are from microbial mats located in Lake Vanda, McMurdo Dry Valleys. Lake Vanda is also a perennially ice-covered lake and is located at 77.53° S, 161.58° E. Microbial mats in Lake Vanda contain pinnacles that range from millimeters to centimeters tall. Unlike Lake Fryxell, there is no sulfide where we sampled, and it is better illuminated at the sampled location than Lake Fryxell, though samples from the inside of pinnacles receive little light (Sumner et al., 2016). Sampling methods and lake conditions have previously been described in Sumner et al. (2016)

Bioinformatic Processing and Assembly of Antarctic Cyanobacteria Reference MAGs

The methods to obtain the *P. pseudopriestleyi* MAG has been previously described in Lumian et al. (2021). Briefly, the *P. pseudopriestleyi* MAG was obtained from a microbial mat sample sequenced on an Illumina HiSeq 2500 PE250 platform and a laboratory culture was sequenced on an Illumina 2000 PE100 platform. The microbial mat sample was quality filtered, and forward and reverse reads were joined using PEAR v0.9.6 (Zhang et al., 2014). For the isolated strain, trimmomatic v0.36 (Bolger et al., 2014) was used to trim sequencing adapters, and the interleave-reads.py script in khmer v2.1.2 (Crusoe et al., 2015) was used to interleave the reads. Both samples were assembled separately and together as a co-assembly by MEGAHIT v1.1.2 (Li et al., 2015) and mapped with bwa v2.3 (Li, 2013) and samtools v1.9 (Li et al., 2009). A single cyanobacteria bin was obtained using the CONCOCT binning algorithm in anvi'o and identified using CheckM (Eren et al., 2015; Delmont & Eren, 2018; Parks et al., 2015). The *P. pseudopriestleyi* bin was refined with spacegraphcats to extract additional content from the metagenomes with a k-mer size of 21 and a radius of 1 (Brown et al., 2020).

Methods to obtain the *Microcoleus, Pseudanabaena, Leptolyngyba*, and *Neosynechococcus* MAGs from Lake Vanda are described in Lumian et al. 2022 (unpublished). Filtered and quality controlled raw data was retrieved from the NCBI Sequence Read Archive under the accession numbers SRR6448204 - SRR6448219 and SRR 6831528.. MEGAHIT v1.9.6 was used to assemble metagenomes with a minimum contig length of 1500 bp and a paired end setting. Bowtie2 v1.2.2 and samtools v1.7 were used to map reads back to the assembly. A depth file was generated using jgi_summarize_bam_contig_depths from MetaBAT v2.12.1 (Kang et al., 2015), which was also used to generate bins with a minimum contig length of 2500 bp. The completeness and contamination of the bins were calculated with CheckM v1.0.7 (Parks, D.H., et al., 2014). Bins that were contained within the phylum Cyanobacteria in the phylogenetic tree generated by CheckM were retained for further analysis. 139 single copy marker genes (Campbell et al., 2013) were collected using the anvi-run-hmms command in anvi'o v6.2 (Eren et al., 2021) and a phylogenetic tree was constructed using the anvi-gen-phylogenomic-tree command. Genome similarity between bins was computed using the anvi-compute-genome-similarity command. Bins were grouped into taxa if they shared more than 98% average nucleotide similarity and were phylogenetically cohesive. When a taxon was found in multiple metagenomes, the most complete bin with the lowest level of contamination for that taxon was selected for additional analysis. <<Could say something like the bin selected for each taxon will be referred to as the MAG for that taxon -- something to explain why we jump from the phrase "bin" to "MAG">> Taxa were classified using GTDB-tk v.2.1.0 (Chaumeil et al., 2020). MAGs for each taxon are being deposited in the NCBI sequence read archive under and will receive accession numbers.

Sourmash branchwater Software with Large-Scale k-mer Searching for Comparative Metagenomic Analysis

The branchwater software used large-scale k-mer searching to search all metagenomes in the NCBI SRA as of September 2020 for matches with genomes of interest (Brown & Irber, 2016; Pierce et al., 2019). Signature files of the genomes of interest were generated using sourmash v3.5.0 (Brown & Irber, 2016) with k-mer sizes of 21, 31, 51, the scaled parameter set to 1000, and abundance tracking. This generated a unique signature file specific to each of the five Antarctic MAGs. These signature files were searched against signature files previously generated for all 498,942 publicly available unassembled metagenome sets on the SRA as of September, 2020 using exact k-mer matching. Results are organized by containment, which is the proportion of the query MAG k-mers found in the metagenome. The use of k = 31 as a k-mer size enables detection of matches to ~91% ANI at 5% containment and ~97% ANI at 30% containment (Hera et al., 2022). The size of the Antarctic query MAGs ranged from 2.7 Mbp – 6.07 Mbp, so a match with containment value of 5% implies 135,000 –

303,500 matching k-mers with k = 31 and 4,185,000 – 9,408,500 matching base pairs, which indicates significant shared genomic material between MAGs and metagenome matches.

Validation of k-mer results from branchwater was done by mapping the Antarctic MAGs back to the metagenomes from the SRA using minimap2 v2.24 in genome-grist v0.8.4 (Li, 2018; Irber et al., 2022). Environmental metadata for the top hits of all MAGs with hits above 5% were recorded, with the exception of the *Microcoleus*, which had over 1,000 matches above that threshold (Tables 2 – 3). Unassembled metagenomes from geographically distinct environments were assembled with MEGAHIT v1.9.6, mapped with bowtie2 v1.2.2 and samtools v1.7, and binned with MetaBAT v.2.12.1. However, none of the assemblies were high enough quality to yield bins (Table 4). The code from this project is available at: https://github.com/dib-lab/2022-pipeline-antarctic-biogeography.

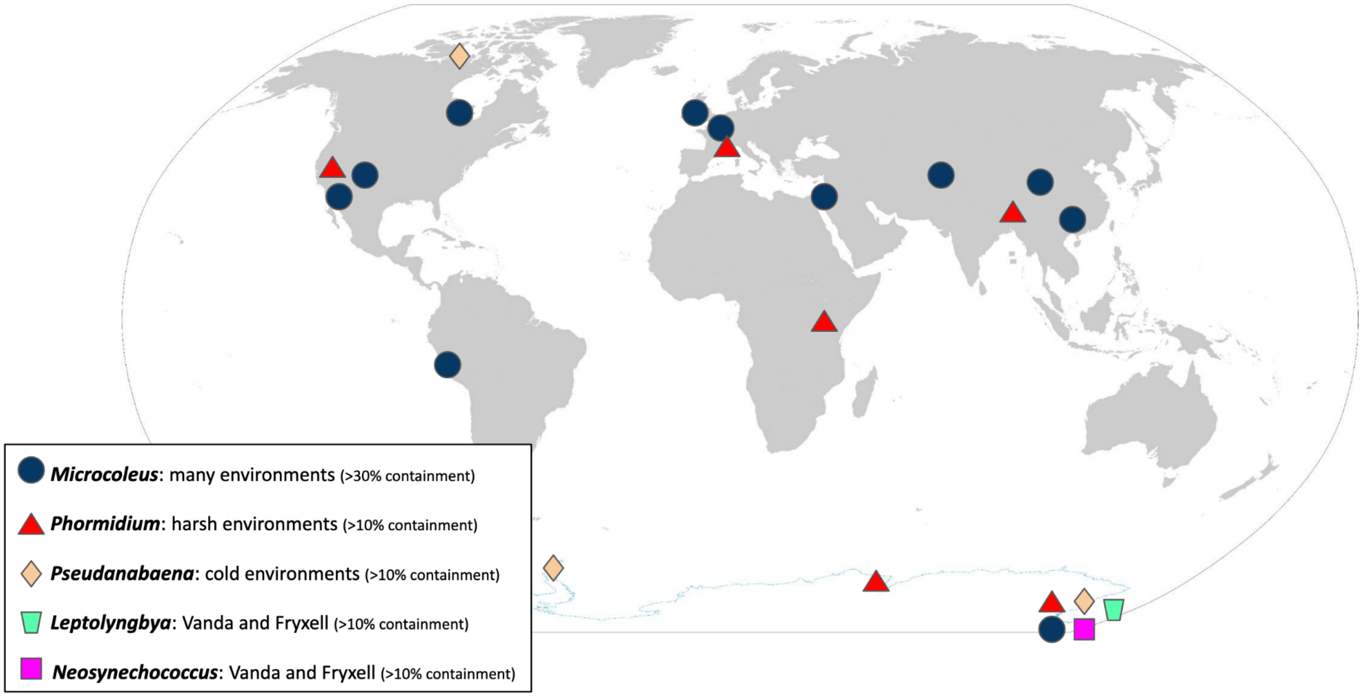
RESULTS

The five polar cyanobacteria MAGs used as search queries were found in a variety of non-polar metagenomic data sets in a range of environmental conditions (Table 1, Figure 1). The metagenome data sets with >5% containment of the MAGs described in Tables 1 – 3. Information about additional environments where the *Microcoleus* MAG was found with over 20% containment is displayed in Table 3. Validation mapping numbers are available in Table 5. The SRA accession numbers of additional hits are available in Supplementary Tables S1 – S5.

The purpose of branchwater was to find shared genomic data between Antarctic MAGs and SRA metagenomes from different habitats, geographic location, and climate zones. Matches of our selected Antarctic cyanobacteria MAGs in these metagenomes may indicate the occurrence of Antarctic cyanobacteria or closely related taxa in environments across the globe. A k-mer size of 31 with at least 5% containment indicates a ~91% ANI between matched sequences (Hera et al., 2022). At 30%

containment, this value increases to ~97% ANI (Hera et al., 2022). Thus, a high containment value indicates the presence in the metagenome of genomic DNA similar to the MAG, and supports the presence of a related organism in the sampling location of that metagenome. Low k-mer containment values may represent smaller regions of shared genomic material or the presence of a related species, but cannot definitively support the presence of the same species in that environment. Containment, particularly at low values, can be affected by factors such as plasmids, metagenome sequencing depth, or small portions of shared contamination between the MAG and metagenome.

Figure 1 Global Distribution of Antarctic Cyanobacteria


● ***Microcoleus***: many environments (>30% containment)
▲ ***Phormidium***: harsh environments (>10% containment)
◇ ***Pseudanabaena***: cold environments (>10% containment)
⏢ ***Leptolyngbya***: Vanda and Fryxell (>10% containment)
■ ***Neosynechococcus***: Vanda and Fryxell (>10% containment)

Global distribution of Antarctic MAGs above 30% for the *Microcoleus* MAG and above 10% for the remaining MAGs. The symbols over Lake Vanda and Fryxell are not positioned for scale to avoid crowding.

Table 1 Summary of branchwater hits

	Microcoleus	***Phormidium pseudopriestleyi***	***Pseudanabaena***	***Neosynechococcus***	***Leptolyngbya***
# Hits >75% Containment	6	30	6	3	5

# Hits >50% Containment	12	33	6	5	5
# Hits >25% Containment	119	38	10	6	6
# Hits >5% Containment	1,121	131	24	16	22
Total Hits	6,184	2,739	3,769	2,796	2,999
# Geographically Distinct Locations >25% Containment	27	3	3	1	1

Table 2 Metagenomes from the NCBI SRA with the Highest Containment Values

MAG	Containment (%)	Location	Accession Number	Latitude and Longitude	BioSample Metadata#
Microcoleus	99.18*	Mat lift-off from Lake Fryxell, Antarctica	SRR5468150	77.605 S, 163.1630 E	Isolation Source: Ice surface mat Collection Date: 2014-12-05
	65.02*	Polar Desert Sand Communities, Antarctica	SRR6266358	78.0741 S, 163.8918 E	Isolation Source: Antarctic Sand Collection Date 2010-01-15
	57.50*	Moab Green Butte, Utah, USA	SRR5855414	38.42 N, 109.41 W	Isolation Source: biocrust samples from Green Butte Site near Canyonlands National Park along an apparent maturity gradient of Cyanobacteria-dominated biocrusts Collection Date: 2014-09 Sample Collection: soil coring with Petri dishes
	41.65*	Ningxia, China	SRR2952554	Not provided	Isolation Source: algae crusts Collection Date: 2013-04-15 Geographic Location: China: Ningxia
	41.10*	Sonoran Desert, Colorado Plateau, USA	SRR5247052	38.42 N, 109.4099 W	Isolation Source: Colorado Plateau and Sonoran Desert Plateau and Sonoran Desert
	40.61*	Pig Farm, UK	ERR3588763	55.95 N, -3.188 W	Environmental Context: Pig farm soil Collection Date: 2017-01-11 Project Name: The dynamics of antimicrobial resistance gene prevalence on a commercial pig farm: implications for policy

	39.54*	Glacier Snow, China	SRR5891573	38.2186 N, 81.120 E	Environmental Context: Glacier snow from glacier Collection Date: 2013-09 Depth: 0.01 m Elevation: 5800 m
	38.36*	Mine Tailing Pool Sediment near Shaoyang, China	ERR1333181	27.745 N, 111.46 E	Environmental Context: Mine tailing pool, sediment Collection Date: 2014-12-27 Depth: 0.1 m (?) Elevation: 286 m (?)
	37.04*	Wastewater in Milwaukee, Wisconsin, USA	SRR5459769	43.023 N, 87.895 W	Environmental Context: Wastewater communities Collection Date: 2014-07-17
	36.30*	Puca Glacier, Peru	SRR6048908	13.773 S, 71.071 W	Environmental Context: Early successional soil, N + P addition Collection Date: 2012
	35.71*	Negev Desert, Israel	SRR12473531	30.785 N, 34.767 E	Environmental Context: Temperate “Mediterranean” desert biome Collection Date: 2017-05-10 Depth: 0.2 cm Elevation: 0 m
	33.58*	Southwest Germany	ERR3192241	Not Provided	BioProject Information: Short read whole genome sequencing of 276 wild *Arabidopsis thaliana* rosettes from southwest Germany
Phormidium pseudopriestleyi	98.49*	Microbial mat in Lake Fryxell, Antarctica	SRR7769747	77.6167 S, 163.1833 E	Collection Date: 2012-11 Isolation source: benthic surface of ice-covered lake
	55.80*	Ace Lake, Antarctica	SRR7528444	68.473 S, 78.188 E	Collection Date: 2014-02-15 Isolation Source: saline lake
	23.54*	Rauer Islands, Antarctica	SRR5216658	68.556 S, 78.191 E	Collection Date: 2015-01-11 Isolation Source: saline lake
	20.63*	Les Salins du Lion Bird Reserve, France	SRR7428116	43.453 N, 5.230 E	Environmental Context: Microbial mat from brackish lagoon in a natural zone of ecological interest Collection Date: 2011-09 Depth: 0.2 cm Elevation: 0 m
	19.04*	Big Soda Lake, Nevada	SRR12522841	39.523 N, 118.870 W	Collection Date: 2016-06-22
	18.25*	Étang de Berre Lagoon, France	SRR7428132	43.485 N, 5.188 E	Environmental Context: Microbial mat from hydrocarbon retention basic Collection Date: 2012-04 Depth: 0.2 cm Elevation: 2 m

	11.99*	Sewage in Nairobi, Kenya	ERR3503286	1.19 S, 36.47 E	Environmental Context: Stream collection for survey of infectious diseases and antimicrobial resistance Collection Date: 2014-07-28
	10.37	Wetland soil in Yanghu, China	SRR9691033	29.19 N, 90.59 E	Collection Date: 2013-07-12
	8.98*	Salar del Huasco salt flat, Chile	SRR10186387	20.264 S, 68.875 W	Collection Date: 2018-06-02
	8.48*	Simulated Metagenome	ERR738546	Not Applicable	Simulated metagenome based on real Illumina HiSeq 2000 data to benchmark metagenome analysis tools
	7.61*	Human Gut	SRR6262267	40.431 N, 79.959 W	Sample Context: Fecal sample from infant Collection Date: 2015
Pseudanabaena	99.49*	Mat lift-off from Lake Fryxell, Antarctica	SRR5468149	78.072 S, 163.719 E	Collection Date: 2009-01-15 Isolation Source: Antarctic Sand
	37.45	Dry Valley Sand Communities, Antarctica	SRR6266338	81.017 N, 81.583 W	Collection Date: 2015-07-18 Organism: Glacier metagenome
	33.54*	Nunavut, Canada	SRR5829599	77.605 S, 163.1630 E	Isolation Source: Ice surface mat Collection Date: 2014-12-05
	18.31*	Deception Island, Antarctica (Whaler's Bay Sediment)	ERR4192538	62.97 S, 60.55 W	Collection Date: 2017-12 Broad Scale Environmental Context: Lake Local Scale Environmental Context: Lake Environmental Medium: sediment
	7.45*	Microbial mat in Lake Fryxell, Antarctica	SRR7769784	77.6167 S, 163.1833 E	Collection Date: 2012-11 Isolation Source: Benthic surface of ice-covered lake Host: Lake pH: 7.47
Neosynechococcus	97.82*	Mat lift-off from Lake Fryxell, Antarctica	SRR5208701	77.605 S, 163.163 E	Collection Date: 2014-12-05 Isolation source: Ice surface mat
Leptolyngbya	98.72*	Mat lift-off from Lake Fryxell, Antarctica	SRR5468150	77.605 S, 163.163 E	Collection Date: 2014-12-05 Isolation Source: Ice surface mat
	8.40	Spitsbergen, Svalbard, Norway, Arctic	SRR6683740	78.58 N, 12.05 E	Collection Date: 2016-07-01 Isolation Source: Soil and sediment

An asterisk (*) denotes where multiple samples from the same location above 5% containment were identified but are not shown in this table, which only shows distinct geographical hits. Only the sample with the highest containment is shown. For an extended full list of hits, see Supplementary Tables 1.1 – 1.5.

Note: Question marks are used where units are assumed but not given in SRA metadata. #Metadata provided as available from SRA, which differed for many datasets.

Table 3 Additional Metagenomes from Unique Locations >20% Containment for *Microcoleus* MAG

MAG	Containment (%)	Location	Latitude and Longitude	Accession Number
Microcoleus	32.14	Qing River, China	40.029 N, 116.368 E	SRR10571243
	31.77	Fecal Metagenome of Great Black-Headed Gulls around Qinghai Lake, China	36.78 N, 100.00 E	SRR10492798
	31.35	Agave Microbial Communities from Guanajuato, Mexico	21.766 N, 100.163 W	SRR4142282
	31.14	Miers Valley, Antarctica	78.160 S, 164.100 E	SRR3471615
	31.06	Glacial meltwater from Laohugou glacier, China	39.50 N, 96.52 E	SRR9965273
	28.87	Polar Desert Sand Communities, Antarctica	78.024 S, 163.917 E	SRR6266336
	28.77	Ace Lake, Antarctica	68.473 S, 78.188 E	SRR7528444
	28.41	Particulate Aerosol Particulate Matter, Beijing, China	40.01 N, 116.33 E	SRR10613504
	28.40	Soil crust, Chicken Creek, Germany	51.36 N, 14.15 E	SRR8357461
	28.17	Wetland soil, Lanzhou, China	36.09 N, 103.71 E	SRR9691044
	27.50	Terrestrial metagenome from Alberta, Canada	50.34 N, 113.77 W	SRR5678923
	27.15	Cave, Twin Sisters, Idaho	42.02 N, 113.72 W	SRR7774479
	25.88	Soil communities, Rifle, Colorado	39.53 N, 107.78 W	SRR3969602
	25.27	Semi-synthetic marine metagenomes from University of Algarve, Faro, Portugal	NA	ERR1992808
	25.25	Saskatchewan, Canada	50.28 N, 107.80 W	SRR7013884
	24.64	Wheat and chickpea soil microbiome, Australia	34.538 S, 138.690 E	ERR3029103
	23.58	Soil communities, Uluru, Australia	25.350 S, 131.052	ERR671932
	22.19	Rhizosphere soil, Mafikeng, South Africa	25.79 S, 25.61 E	SRR11128415
	20.79	Microbial mat, Eel River, California, USA	39.840 N, 123.710 W	SRR244337
	20.33	Biofilm in Wai-iti River, New Zealand	NA	SRR9948934

	20.29	Soil rhizosphere, Durango, Mexico	19.321 N, 99,194 W	SRR11092592

The *Microcoleus* MAG was the most widely distributed MAG with 27 globally distinct locations above 25% containment (Tables 1 – 3). The *Microcoleus* and *P. pseudopriestleyi* MAGs were present in the most time series and subsamples from the same environmental location, which resulted in 1,121 hits above 5% for the *Microcoleus* MAG and 131 hits for *P. pseudopriestleyi* MAG (Table 1). The *Pseudanabaena* and *P. pseudopriestleyi* MAGs were found in three distinct locations above 25% containment while the *Neosynechococcus* and *Leptolyngbya* MAGs were only found in one location above 25% containment (Table 1).

The *Microcoleus* MAG was found in diverse environments with conditions ranging from hot to cold climates and including both arid and wet locations (Tables 2 – 3). Some environments are cold year-round such as Puca Glacier in Peru (36.30% containment), glacier snow in China (39.54% containment), and the ice-covered Lake Vanda, while others are temperate, like Wisconsin, USA (37.04% containment), or Southwest Germany (33.58% containment). *P. pseudopriestleyi* MAG was found in three Antarctic metagenome data sets: Lake Fryxell mat samples (98.49% containment), Ace Lake (55.8% containment) and the Rauer Islands (23.54% containment). The highest 30 hits for the *P. pseudopriestleyi* MAG, including the three samples used to create the MAG, were from Lake Fryxell. This search revealed that *P. pseudopriestleyi* is likely present in other depths of Lake Fryxell than 9.8 m despite not being prevalent at those depths based on 16S sequencing (Jungblut et al., 2010). Besides Antarctica, the *P. pseudopriestleyi* MAG was found in a bird reserve next to a lagoon in France called Les Salins du Lion (20.63% containment) as well as a hydrocarbon polluted saline lagoon called Étang de Berre (18.25% containment) which were part of a study on the effects of hydrocarbon pollution on microbial communities (Aubé et al., 2016). The *P. pseudopriestleyi* MAG was also found in the Salar del Huasco salt flat in Chile (8.98% containment), antimicrobial treated sewage collected in Nairobi, Kenya

(11.99% containment) and an infant gut fecal sample (7.61% containment). All these environments represent extreme conditions for cyanobacteria.

Although the *Microcoleus*, *Pseudanabaena*, *Neosynechococcus,* and *Leptolyngbya* MAGs were obtained from microbial mat pinnacles in Lake Vanda, they were all present in high containment (>97%) in mat lift-off samples from Lake Fryxell where the *P. pseudopriestleyi* MAG was not detected. The *Pseudanabaena* MAG was also found in a dry sand community in the McMurdo Dry Valleys (37.45% containment), where lakes Vanda and Fryxell are located, as well as Whaler's Bay on Deception Island in Antarctic (18.31 % containment) and the Canadian High Arctic such as Nunavut, Canada (33.54 % containment), which is cold but geographically distant from the Antarctic.

Table 4 Quality Metrics of Metagenome Assemblies and Mapping Statistics

MAG	SRA Accession Number and Location	Number of Contigs (over 500 bp)	Total Length of Contigs (bp, over 500 bp)	Largest Contig (bp)	N50
Microcoleus	SRR5855414 Moab Green Butte, Utah, USA	779, 743	497,286,103	4,248	612
	SRR2952554 Ningxia, China	296,291	187,141,964	7,947	604
	SRR5247052 Sonoran Desert, Colorado Plateau, USA	781,528	498,867,695	3,982	613
	ERR3588763 Pig Farm, UK	402,432	243,270,443	3,395	581
	SRR5891573 Glacier Snow, China	979,446	619,490,337	3,614	606
	ERR1333181 Antimony Polluted Sediment, China	415,411	260,914,760	5,701	607
	SRR5459769 Wastewater in Milwaukee, Wisconsin USA	1,259,786	1,086,536,868	57,883	819
	SRR12473531 Negev Desert, Israel	297,517	187,591,328	4,696	600
	ERR192241 Southwest Germany	108,078	63,950,717	1,868	567
Phormidium pseudopriestleyi	SRR7528444 Ace Lake, Antarctica	218,966	195,047,488	62,020	822
	SRR5216658 Rauer Islands, Antarctica	250,605	212,798,142	64,599	794
	SRR7428116 Bird Reserve, France	462,930	371,142,339	36,453	146,664
Pseudanabaena	SRR6266338	651,239	424,874,538	9,934	624

	Dry Valley, Antarctica				
	SRR5829599 Canada	704,963	564,862,171	27,496	751
Neosynechococcus	SRR5468153 Lake Fryxell, Antarctica	1,119,433	887,670,078	21,258	760
Leptolyngbya	SRR5468153 Lake Fryxell, Antarctica	1,119,433	887,670,078	21,258	760

Table 5 Mapping Validation of MAGs in SRA Metagenomes

MAG	SRA Accession Number and Location	K-mer Containment (%)	Effective Coverage	Percentage of MAG Detected in Metagenome (%)	Number of Mapped Reads from MAG
Microcoleus	SRR5468150 Mat lift-off from Lake Fryxell, Antarctica	99.18*	125.84	99.35	5,864,248
	SRR6266358 Polar Desert Sand Communities, Antarctica	65.02*	93.34	88.34	3,832,909
	SRR5855414 Moab Green Butte, Utah, USA	57.50*	407.19	86.11	15,915,624
	SRR2952554 Ningxia, China	41.65*	18.83	73.53	899,792
	SRR5247052 Sonoran Desert, Colorado Plateau, USA	41.10*	180.87	73.08	10,101,904
	ERR3588763 Pig Farm, UK	40.61*	9.38	76.14	329,215
	SRR5891573 Glacier Snow, China	39.54*	14.36	75.66	482,590
	ERR1333181 Mine Tailing Pool Sediment near Shaoyang, China	38.36*	28.59	73.24	1,120,980
	SRR5459769 Wastewater in Milwaukee, Wisconsin, USA	37.04*	13.67	76.29	636,988
	SRR6048908 Puca Glacier, Peru	36.30*	7.76	73.49	280,909

	SRR12473531 Negev Desert, Israel	35.71*	18.06	74.46	639,468
	ERR3192241 Southwest Germany	33.58*	8.80	69.98	288,838
Phormidium pseudopriestleyi	SRR7769747 Microbial mat in Lake Fryxell	98.49*	22.61	98.93	602,728
	SRR7528444 Ace Lake, Antarctica	55.80*	2.21	61.50	103,088
	SRR5216658 Rauer Islands, Antarctica	23.54*	1.59	27.23	21,770
	SRR7428116 Les Salins du Lion Bird Reserve, France	20.63*	11.68	60.33	471,918
	SRR12522841 Big Soda Lake, Nevada	19.04*	10.67	67.52	340,050
	SRR7428132 Étang de Berre Lagoon, France	18.25*	2.75	47.58	90,501
	ERR3503286 Sewage in Nairobi, Kenya	11.99*	2.12	40.00	38,112
	SRR9691033 Wetland soil in Yanghu, China	10.37	1.69	30.33	24,130
	SRR10186387 Salar del Huasco salt flat, Chile	8.98*	3.64	24.24	50,222
	ERR738546 Simulated Metagenome	8.48*	1.45	19.54	20,152
	SRR6262267 Human Gut	7.61*	2.06	25.27	25,271
Pseudanabaena	SRR5468149 Mat lift-off from Lake Fryxell, Antarctica	99.49*	142.44	99.89	2,748,583
	SRR6266338 Dry Valley Sand Communities, Antarctica	37.45	2.65	45.28	28,689
	SRR5829599 Nunavut, Canada	33.54*	5.91	78.56	97,210

	ERR4192538 Deception Island, Antarctica (Whaler's Bay Sediment)	18.31*	2.48	46.67	24,011
	SRR7769784 Microbial mat in Lake Fryxell	7.45*	1.13	9.00	1,405
Neosynechococcus	SRR5208701 Mat lift-off from Lake Fryxell, Antarctica	97.82*	4.023	90.79	133,786
Leptolyngbya	SRR5468150 Mat lift-off from Lake Fryxell, Antarctica	98.72*	62.71	99.35	2,759,236
	SRR6683740 Spitsbergen, Svalbard, Norway	8.40	1.71	16.67	17,394

Metagenomes representing geographically distinct locations were selected for further analysis to compare genomic data from different environments to the Antarctic MAGs. These data sets were run through an assembly and binning pipeline to obtain bins that could be compared to the Antarctic MAGs. However, metagenome assemblies were poor quality with the majority of the N50s under 1,000 base pairs, which is the minimum contig length required to bin with MetaBAT. Thus, bins were not generated, and it would not have been possible to identify the presence of the MAGs in these metagenomes without using an assembly-independent technique. Validation of the branchwater results was done by mapping the MAGs to metagenomes (Table 5). The percentage of the MAG detected in metagenome and average MAG coverage confirm the results of branchwater independent of k-mer comparisons, with all but one sample exhibiting higher mapping-based detection in the metagenome than k-mer containment.

DISCUSSION

Environmental Diversity of Microcoleus

The presence of the *Microcoleus* MAG in diverse environments indicates that it can survive in a range of different ecological conditions and climatic zones. The findings agree with previous biogeographic assessments of cultured cyanobacteria belonging to the species *Microcoleus vaginatus* and the *Microcoleus* spp. based on the 16S rRNA gene (Dvořák et al., 2012; Strunecký et al., 2013). In order to survive cold temperatures in Lake Vanda, the *Microcoleus* must deal with cellular membranes becoming brittle and slowed metabolism. However, some environments where the *Microcoleus* was found are only cold for part of the year (Moab Green Butte Desert; Ningxia, China; Southwest Germany; Milwaukee, Wisconsin; and the UK) while other environments are cold year-round (Puca Glacier, Peru, and glacial snow in China). In contrast to cold conditions, hot temperatures can cause proteins to denature and prolonged exposure to sunlight can cause high light and UV stress. These conditions occur in the Moab Green Butte Desert, the Sonoran Desert, and the Negev Desert. Furthermore, the Moab Desert and Sonoran Desert experience extreme temperature changes between morning and night (Turnage & Hinckley, 1938; Balling et al., 1998; McCann et al., 2018), forcing the *Microcoleus* to adapt to both conditions on a 24-hour cycle.

In addition to temperature range, the *Microcoleus* MAG was found in metagenomes environments with different levels of water availability and habitat types. Locations entailed arid desert soil crusts (Moab and Negev Deserts), mine tailings (Shaoyang, China; the United Kingdom; Milwaukee), epiphyte on plant microbiomes (Southwest, Germany), and freshwater environments (Qing River, China and Ace Lake, Antarctica). The *Microcoleus* MAG was also found in data from both high and low elevation environments (5800 m elevation in glacial snow in China and 0 m elevation in the Negev Desert). Interestingly, in Southwest Germany the MAG was found in metagenomic data of wild Arabidopsis plants. Overall, the variety of conditions where the *Microcoleus* MAG was found indicates that it may live in an impressive range of environments ranging from moderate climates to extreme heat or cold.

Environmental Diversity of Phormidium pseudopriestleyi

P. pseudopriestleyi is a sulfide-tolerant cyanobacteria found in a low light environment in Lake Fryxell, Antarctica. Our study identified the *P. pseudopriestleyi* MAG in metagenomes from additional locations in Antarctica such as the saline Ace Lake (Vestfold Hills) and lakes on the Rauer Islands, which agrees with previous 16S rRNA gene sequencing where the species was documented from Salt Pond and Fresh Pond on McMurdo Ice Shelf (Jungblut et al., 2005; Lumian et al., 2021) as well as Ace lake (Taton et al., 2006). Interestingly, *P. pseudopriestleyi* or a close relative is present also at low abundance in a pond at Les Salins du Lion, a bird reserve (20.63% containment, 95% cANI), and Étang de Berre, a hydrocarbon polluted saline lagoon (18.25% containment, 94% cANI), both in southern France (Aubé et al., 2016). Four environmental conditions can be compared in these locations: irradiance, salinity, temperature, and sulfide concentrations. The irradiance at Les Salins du Lion pond and Étang de Berre lagoon was not measured when environmental sampling occurred, but the elevation of the lagoon was recorded to be at 0 m, indicating that irradiance is higher at the surface of the pond than the low irradiance at the depth of sampling in Lake Fryxell (1-2 µmol/photon m^{-2} s^{-1}) (Sumner et al., 2015). Furthermore, Salt Pond and Fresh Pond have high illumination levels in the summer (Roos & Vincent, 1998; Jungblut et al., 2005), indicating that *P. pseudopriestleyi* may have the capability to overcome high irradiation and UV levels for prolonged periods. Les Salins du Lion (14 g L^{-1} NaCl) and Étang de Berre (20 g L^{-1} NaCl) have a lower salinity than Lake Fryxell (70.13 g L^{-1} NaCl) and Salt Pond (~990 g L^{-1} NaCl), which is hypersaline (Jungblut et al., 2005; Aubé et al., 2016; Lumian et al., 2021). Previous work has showed that *P. pseudopriestleyi* increases the thickness of its extracellular polymeric substance layer in response to saline stress (Agrawal & Singh, 1999). Sulfide is also present in Les Salins du Lion, with a concentration of ~0.24 g L^{-1} at the time of sampling (Aubé et al., 2016), which was the highest value at any location or time sampled included in the study. This demonstrates a much higher sulfide tolerance than what was previously recorded in the Lake Fryxell sampling site, which was 9.8×10^{-5} g L^{-1} (Lumian et al., 2021).

In addition to Les Salins du Lion and Étang de Berre, the *P. pseudopriestleyi* MAG was found in globally distributed challenging environments such as a salt flat in Chile, antimicrobial treated sewage in Kenya, and infant gut. The fact that *P. pseudopriestleyi* thrives in environments with harsh conditions suggests that it has capabilities to overcome diverse environmental stresses. In Lake Fryxell, *P. pseudopriestleyi* dominates microbial mats at 9.8 m depth in low light and sulfidic conditions but it is less abundant at shallower depths, even though there is more light availability and no sulfide (Jungblut et al., 2016; Dillon et al., 2020). Thus, *P. pseudopriestleyi* may grow slowly and find ecological success in environments that are too harsh for faster growing cyanobacteria, which is consistent with the slow growth rate of *P. pseudopriestleyi* seen in unpublished laboratory observations. The other environments where genomes similar to *P. pseudopriestleyi* were found may provide challenges that prohibit many other cyanobacteria from growing (polar environments, alkaline lake in Big Soda lake, antimicrobial treated sewage in Nairobi, Kenya), allowing *P. pseudopriestleyi* to survive in a nonpolar environment.

Environmental Diversity of Pseudanabaena, Neosynechococcus, and Leptolyngbya

The top matches for the *Pseudanabaena, Neosynechococcus*, and *Leptolyngbya* MAGs showed that they were also present in Lake Fryxell and that the *Pseudanabaena* MAG was in sediment in the McMurdo Dry Valleys. The *Neosynechococcus* MAG was only present in the McMurdo Dry Valleys, however the presence of the *Leptolyngbya* and *Pseudanabaena* MAGs in geographically distant locations in the Arctic (Norway and Canada respectively) suggests that the cyanobacteria forming these MAGs have a global distribution in cold environments and might have undergone long range dispersal. The mechanism of long-range distribution could be wind; atmospheric studies show bacteria from the Saharan desert are transported by wind throughout the Atlantic (Griffin et al., 2002; Gorbushina et al., 2007; A. D. Jungblut et al., 2010). A similar process is expected to allow Antarctic cyanobacteria to cross large distances and populate diverse geographic regions. However, the lack of non-polar locations suggests that they are not as successful at integrating into non-polar environments. Thus, these cyanobacteria may be specific to polar environments even though they may be transported globally, which agrees with 16S rRNA gene analysis that proposed the presence of cosmopolitan cold ecotypes (Jungblut et al., 2010). The genus *Neosynechococcus* was described by Dvořák et al., (2014) based on cyanobacteria isolated from peat bog in Slovakia and was described based on 16S rRNA gene, 16S- 23S ITS, and *rbcl* loci, however in the approached presented here it could only be identified in Antarctica.

Implications for Biogeographic Distributions

The perceived distributions of organisms in biogeography studies are affected by sampling and publishing biases. Sampling in remote locations is logistically difficult and is often centered around established sampling locations which may be near research stations and infrastructure. This results in many studies and publications from established sampling locations and a deeper understanding of local ecology and geochemical processes in these environments. Biogeography studies, however, benefit from widespread sampling in many locations. Conducting widespread ecological sampling is expensive and can be impractical, so it is advantageous to search existing datasets for as much information as possible. Using branchwater to search public metagenomes makes the most out of data from remote areas by revealing previously unknown locations of organisms of interest. Furthermore, results from this analysis included remote areas, including various sites in Antarctica, which may not have otherwise been identified as locations of the query MAGs.

Despite being affected by sampling bias like all biogeography studies, the results showed that the *Microcoleus* MAG was globally distributed over a wide variety of environments, the *P. pseudopriestleyi* MAG was found in predominantly in harsh environments, the *Neosynechococcus* and *Leptolyngbya* MAGs were in the Antarctic, and the *Pseudanabaena* MAG was in geographically separated polar environments. The numerous sites containing the *Microcoleus* MAG imply that this species has the genetic capacity to adapt to many types of environments. It may also have a faster growth rate than an extreme conditions specialist, like *P. pseudopriestleyi* MAG, which would allow it to compete in a variety of ecological communities, some of which experience stressful conditions. Previous work has shown *Microcoleus* to be a cosmopolitan genus (Garcia-Pichel et al., 1996, 2001).

Although the *Microcoleus* MAG is by far the most globally diverse cyanobacterial genome in this study, there is variety in the distributions of the other four MAGs. The prevalence of the *P. pseudopriestleyi* MAG in harsh environments indicates that it finds ecological success in stressful environments, and it is likely outperformed by other organisms in moderate environments. The *Pseudanabaena*, *Neosynechococcus*, and *Leptolyngbya* MAGs were only found in polar environments, indicating they may be outcompeted in moderate environments. Diving deeper into the metabolic potential of each organism and interactions between metagenome community members may offer insights as to how and why some organisms are prevalent in a multitude of environments while others are prevalent in only certain conditions.

CONCLUSIONS

This paper presents the first biogeography study using a large-scale k-mer-based approach and characterizes the global distribution of five distinct Antarctic cyanobacteria based on public data. We show that the *Microcoleus* MAG has cosmopolitan distribution and presence in a variety of environments, whereas the *P. pseudopriestleyi* MAG is also globally distributed but mostly present in harsh environments. *Leptolyngbya,* and *Pseudanabaena* MAGs were only found in polar environments from Arctic to Antarctica suggesting the existence of cosmopolitan cold ecotypes. The *Neosynechococcus* MAG was only detected in Antarctica and provides support for more restricted distribution patterns and potential endemicity. Further in situ transcriptomic studies of these MAGs may reveal adaptation mechanisms including why the *Microcoleus* is so pervasive compared to the other cyanobacteria in this study.

Branchwater can search ~500,000 metagenomes with a query genome in under 24 hours on commodity hardware (Irber et al. unpublished). The ability to quickly find genomes similar to query MAGs in publicly available unassembled metagenomic data sets has important implications for biogeography studies, which have been predominantly based on 16S rRNA gene sequencing due to the prevalence of data and ease of comparison. Branchwater greatly increases the amount of data that can be used for biogeography studies. This technique is especially helpful for organisms that are in remote locations and underrepresented in genomic data, such as polar cyanobacteria, by providing a much larger number of known environments than would be possible with targeted field studies. Additionally, branchwater can be used to identify accessible sampling locations of organisms from remote environments, such as the *Microcoleus* being identified in the Moab Green Butte Desert in Colorado, USA at 41.10% containment. As more metagenome datasets are made publicly available on the NCBI SRA, more information about the distribution of cryosphere cyanobacteria can be attained. The results further demonstrate the potential of metagenomics and k-mer based MAG approaches in investigating biogeography and ecology of cyanobacteria and environmental microbiology in the Polar Regions.

REFERENCES

Agrawal, S. C., & Singh, V. (1999). Viability of dried vegetative trichomes, formation of akinetes and heterocysts and akinete germination in some blue-green algae under water stress. *Folia Microbiologica*, *44*(4), 411–418. https://doi.org/10.1007/BF02903715

Aubé, J., Senin, P., Pringault, O., Bonin, P., Deflandre, B., Bouchez, O., Bru, N., Biritxinaga-Etchart, E., Klopp, C., Guyoneaud, R., & Goñi-Urriza, M. (2016). The impact of long-term hydrocarbon exposure on the structure, activity, and biogeochemical functioning of microbial mats. *Marine Pollution Bulletin*, *111*(1), 115–125. https://doi.org/10.1016/j.marpolbul.2016.07.023

Bahl, J., Lau, M. C. Y., Smith, G. J. D., Vijaykrishna, D., Cary, S. C., Lacap, D. C., Lee, C. K., Papke, R. T., Warren-Rhodes, K. A., Wong, F. K. Y., McKay, C. P., & Pointing, S. B. (2011). Ancient origins determine global biogeography of hot and cold desert cyanobacteria. *Nature Communications*, *2*(1), 163. https://doi.org/10.1038/ncomms1167

Baas-Becking, L. G. M. (1934). *Geobiologie; of inleiding tot de milieukunde*. WP Van Stockum & Zoon NV.

Balling, R. C., Klopatek, J. M., Hildebrandt, M. L., Moritz, C. K., & Watts, C. J. (1998). Impacts of Land Degradation on Historical Temperature Records from the Sonoran Desert. *Climatic Change*, *40*(3), 669–681. https://doi.org/10.1023/A:1005370115396

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brown, C. T. (2021, June 8). Searching all public metagenomes with sourmash. *Living in an Ivory Basement*. http://ivory.idyll.org/blog/2021-MAGsearch.html

Brown, C. T., & Irber, L. (2016). sourmash: A library for MinHash sketching of DNA. *Journal of Open Source Software*, *1*(5), 27. https://doi.org/10.21105/joss.00027

Brown, C. T., Moritz, D., O'Brien, M. P., Reidl, F., Reiter, T., & Sullivan, B. D. (2020). Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. *Genome Biology*, *21*(1), 164. https://doi.org/10.1186/s13059-020-02066-4

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., Söll, D., & Podar, M. (2013). UGA is an Additional Glycine Codon in Uncultured SR1 Bacteria from the Human Microbiota.

Proceedings of the National Academy of Science, 110, 5540–5545. https://doi.org/10.1073/pnas.1303090110

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chrismas, N. A. M., Anesio, A. M., & Sánchez-Baracaldo, P. (2015). Multiple adaptations to polar and alpine environments within cyanobacteria: A phylogenomic and Bayesian approach. *Frontiers in Microbiology*, *6*, 1070. https://doi.org/10.3389/fmicb.2015.01070

Chrismas, N. A. M., Barker, G., Anesio, A. M., & Sánchez-Baracaldo, P. (2016). Genomic mechanisms for cold tolerance and production of exopolysaccharides in the Arctic cyanobacterium Phormidesmis priestleyi BC1401. *BMC Genomics*, *17*. https://doi.org/10.1186/s12864-016-2846-4

Chrismas, N. A. M., Williamson, C. J., Yallop, M. L., Anesio, A. M., & Sánchez‐Baracaldo, P. (2018). Photoecology of the Antarctic cyanobacterium Leptolyngbya sp. BC1307 brought to light through community analysis, comparative genomics and in vitro photophysiology. *Molecular Ecology*, *27*(24), 5279–5293. https://doi.org/10.1111/mec.14953

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., Fenton, J., Fenzl, T., Fish, J., Garcia-Gutierrez, L., Garland, P., Gluck, J., González, I., Guermond, S., Guo, J., … Brown, C. T. (2015). The khmer software package: Enabling efficient nucleotide sequence analysis. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.6924.1

Delmont, T. O., & Eren, A. M. (2018). Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ*, *6*, e4320. https://doi.org/10.7717/peerj.4320

Dillon, M. L., Hawes, I., Jungblut, A. D., Mackey, T. J., Eisen, J. A., Doran, P. T., & Sumner, D. Y. (2020). Energetic and Environmental Constraints on the Community Structure of Benthic Microbial Mats in Lake Fryxell, Antarctica. *FEMS Microbiology Ecology*, *96*(2). https://doi.org/10.1093/femsec/fiz207

Dvořák, P., Hašler, P., & Poulíčková, A. (2012). Phylogeography of the Microcoleus vaginatus (Cyanobacteria) from Three Continents – A Spatial and Temporal Characterization. PLOS ONE, 7(6), e40153. https://doi.org/10.1371/journal.pone.0040153

Edgar, R. C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J. F., de la Peña, M., Korobeynikov, A., Chikhi, R., & Babaian, A. (2022). Petabase-scale sequence alignment catalyses viral discovery. *Nature*, *602*(7895), 142–147. https://doi.org/10.1038/s41586-021-04332-2

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ö. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., … Willis, A. D. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. Nature Microbiology, 6(1), Art. 1. https://doi.org/10.1038/s41564-020-00834-3

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ*, *3*, e1319. https://doi.org/10.7717/peerj.1319

Fierer, N. (2008). Microbial Biogeography: Patterns in Microbial Diversity across Space and Time. In Accessing Uncultivated Microorganisms (pp. 95–115). John Wiley & Sons, Ltd. https://doi.org/10.1128/9781555815509.ch6

Garcia-Pichel, F., López-Cortés, A., & Nübel, U. (2001). Phylogenetic and Morphological Diversity of Cyanobacteria in Soil Desert Crusts from the Colorado Plateau. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.67.4.1902-1910.2001

Garcia-Pichel, F., Prufert-Bebout, L., & Muyzer, G. (1996). Phenotypic and phylogenetic analyses show Microcoleus chthonoplastes to be a cosmopolitan cyanobacterium. *Applied and Environmental Microbiology*. https://journals.asm.org/doi/abs/10.1128/aem.62.9.3284-3291.1996

Gorbushina, A. A., Kort, R., Schulte, A., Lazarus, D., Schnetger, B., Brumsack, H.-J., Broughton, W. J., & Favet, J. (2007). Life in Darwin's dust: Intercontinental transport and survival of microbes in the nineteenth century. *Environmental Microbiology*, *9*(12), 2911–2922. https://doi.org/10.1111/j.1462-2920.2007.01461.x

Green, J. L., Bohannan, B. J. M., & Whitaker, R. J. (2008). Microbial Biogeography: From Taxonomy to Traits. Science, 320(5879), 1039–1043. https://doi.org/10.1126/science.1153475

Grettenberger, C. L., Sumner, D. Y., Wall, K., Brown, C. T., Eisen, J. A., Mackey, T. J., Hawes, I., Jospin, G., & Jungblut, A. D. (2020). A phylogenetically novel cyanobacterium most closely related to Gloeobacter. *The ISME Journal*, *14*(8), 2142–2152. https://doi.org/10.1038/s41396-020-0668-5

Griffin, D. W., Kellogg, C. A., Garrison, V. H., & Shinn, E. A. (2002). The Global Transport of Dust: An Intercontinental river of dust, microorganisms and toxic chemicals flows through the Earth’s atmosphere. *American Scientist*, *90*(3), 228–235.

Harding, T. Jungblut, A. D., Lovejoy, C., & Vincent, W. F. 2011. “Microbes in High Arctic Snow and Implications for the Cold Biosphere.” Applied and Environmental Microbiology 77 (10): 3234–43. https://doi.org/10.1128/AEM.02611-10.

Harke, M. J., Steffen, M. M., Gobler, C. J., Otten, T. G., Wilhelm, S. W., Wood, S. A., & Paerl, H. W. (2016). A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, Microcystis spp. *Harmful Algae*, *54*, 4–20. https://doi.org/10.1016/j.hal.2015.12.007

Hera, M. R., Pierce-Ward, N. T., & Koslicki, D. (2022). Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances (p. 2022.01.11.475870). bioRxiv. https://doi.org/10.1101/2022.01.11.475870

Irber, L. C. (2020a). Decentralizing Indices for Genomic Data [Ph.D., University of California, Davis]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/docview/2503641751/abstract/7B8543548D284D81PQ/1

Irber, L. C. (2020b, July 24). MinHashing all the things: A quick analysis of MAG search results. *Gabbleblotchits*. https://blog.luizirber.org/2020/07/24/mag-results/

Irber, L. C., Brooks, P. T., Reiter, T. E., Pierce-Ward, N. T., Hera, M. R., Koslicki, D., & Brown, C. T. (2022). Lightweight compositional analysis of metagenomes with FracMinHash and minimum metagenome covers. *BioRxiv*, 2022.01.11.475838. https://doi.org/10.1101/2022.01.11.475838

Irber, L., Pierce-Ward, N. T., & Brown, C. T. (2022). Sourmash Branchwater Enables Lightweight Petabyte-Scale Sequence Search. In Manubot. Manubot. https://dib-lab.github.io/2022-paper-branchwater-software/

Jungblut, A. D., Hawes, I., Mackey, T. J., Krusor, M., Doran, P. T., Sumner, D. Y., Eisen, J. A., Hillman, C., & Goroncy, A. K. (2016). Microbial Mat Communities along an Oxygen Gradient in a Perennially Ice-Covered Antarctic Lake. *Applied and Environmental Microbiology*, *82*(2), 620–630. https://doi.org/10.1128/AEM.02699-15

Jungblut, A. D., Lovejoy, C., & Vincent, W. F. (2010). Global distribution of cyanobacterial ecotypes in the cold biosphere. *The ISME Journal*, *4*(2), 191–202. https://doi.org/10.1038/ismej.2009.113

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. *Environmental Microbiology*, *7*(4), 519–529. https://doi.org/10.1111/j.1462-2920.2005.00717.x

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *3*, e1165. https://doi.org/10.7717/peerj.1165

Kleinteich, J., Hildebrand, F., Bahram, M., Voigt, A. Y., Wood, S. A., Jungblut, A. D., Küpper, F. C., Quesada, A., Camacho, A., Pearce, D. A., Convey, P., Vincent, W. F., Zarfl, C., Bork, P., & Dietrich, D. R. (2017). Pole-to-Pole Connections: Similarities between Arctic and Antarctic Microbiomes and Their Vulnerability to Environmental Change. https://doi.org/10.3389/fevo.2017.00137

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(suppl_1), D19–D21. https://doi.org/10.1093/nar/gkq1019

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lumian, J. E., Jungblut, A. D., Dillion, M. L., Hawes, I., Doran, P. T., Mackey, T. J., Dick, G. J., Grettenberger, C. L., & Sumner, D. Y. (2021). Metabolic Capacity of the Antarctic Cyanobacterium Phormidium pseudopriestleyi That Sustains Oxygenic Photosynthesis in the Presence of Hydrogen Sulfide. *Genes*, *12*(3), 426. https://doi.org/10.3390/genes12030426

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. Nature Reviews Microbiology, 4(2), Art. 2. https://doi.org/10.1038/nrmicro1341

McCann, R. B., Lynch, J., & Adams, J. (2018). Mitigating Projected Impacts of Climate Change and Building Resiliency Through Permaculture. In *Addressing Climate Change at the Community Level in the United States*. Routledge.

Moreira, C., Vasconcelos, V., & Antunes, A. (2013). Phylogeny and Biogeography of Cyanobacteria and Their Produced Toxins. *Marine Drugs*, *11*(11), 4350–4369. https://doi.org/10.3390/md11114350

Namsaraev, Z., Mano, M.-J., Fernandez, R., & Wilmotte, A. (2010). Biogeography of terrestrial cyanobacteria from Antarctic ice-free areas. *Annals of Glaciology*, *51*(56), 171–177. https://doi.org/10.3189/172756411795931930

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., & Brown, C. T. (2019). *Large-scale sequence comparisons with* sourmash (8:1006). F1000Research. https://doi.org/10.12688/f1000research.19675.1

Pierce-Ward, N. T. (2022, February 7). *Personal Communication, in prep.* [Personal communication].

Quesada, A., & Vincent, W. F. (2012). Cyanobacteria in the Cryosphere: Snow, Ice and Extreme Cold. In B. A. Whitton (Ed.), *Ecology of Cyanobacteria II: Their Diversity in Space and Time* (pp. 387–399). Springer Netherlands. https://doi.org/10.1007/978-94-007-3855-3_14

Ribeiro, K. F., Duarte, L., & Crossetti, L. O. (2018). Everything is not everywhere: A tale on the biogeography of cyanobacteria. *Hydrobiologia*, *820*(1), 23–48. https://doi.org/10.1007/s10750-018-3669-x

Roos, J. C., & Vincent, W. F. (1998). Temperature Dependence of Uv Radiation Effects on Antarctic Cyanobacteria. *Journal of Phycology*, *34*(1), 118–125. https://doi.org/10.1046/j.1529-8817.1998.340118.x

Stal, L. J. (2007). Cyanobacteria. In J. Seckbach (Ed.), *Algae and Cyanobacteria in Extreme Environments* (pp. 659–680). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6112-7_36

Strunecký, O., Komárek, J., Johansen, J., Lukešová, A., & Elster, J. (2013). Molecular and morphological criteria for revision of the genus Microcoleus (Oscillatoriales, Cyanobacteria). Journal of Phycology, 49(6), 1167–1180. https://doi.org/10.1111/jpy.12128

Sumner, D. Y., Hawes, I., Mackey, T. J., Jungblut, A. D., & Doran, P. T. (2015). Antarctic microbial mats: A modern analog for Archean lacustrine oxygen oases. *Geology*, *43*(10), 887–890. https://doi.org/10.1130/G36966.1

Sumner, D. Y., Jungblut, A. D., Hawes, I., Andersen, D. T., Mackey, T. J., & Wall, K. (2016). Growth of elaborate microbial pinnacles in Lake Vanda, Antarctica. *Geobiology*, *14*(6), 556–574. https://doi.org/10.1111/gbi.12188

Taton, A., Grubisic, S., Balthasart, P., Hodgson, D. A., Laybourn-Parry, J., & Wilmotte, A. (2006). Biogeographical distribution and ecological ranges of benthic cyanobacteria in East Antarctic lakes. FEMS Microbiology Ecology, 57(2), 272–289. https://doi.org/10.1111/j.1574-6941.2006.00110.x

Turnage, W. V., & Hinckley, A. L. (1938). Freezing Weather in Relation to Plant Distribution in the Sonoran Desert. *Ecological Monographs*, *8*(4), 529–550. https://doi.org/10.2307/1943083

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. Science, 301(5635), 976–978. https://doi.org/10.1126/science.1086909

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593