Diversity and biogeography of SAR11 bacteria from the Arctic Ocean

Susanne Kraemer[1], Arthi Ramachandran[1], David Colatriano[1], Connie Lovejoy[2], and David A. Walsh[1*]

[1] Department of Biology, Concordia University, 7141 Sherbrooke St. West, Montreal, Quebec H4B 1R6, Canada

[2] Département de biologie, Institut de Biologie Intégrative et des Systèmes (IBIS) and Québec-Océan, Université Laval, Québec, G1K 7P4, Canada

*Corresponding author: Phone: (514) 848-2424 (ext. 3477)

E-mail: david.walsh@concordia.ca

Key words: ITS phylogeny, metagenome assembled genomes, ecotype, evolution

Abstract

The Arctic Ocean is relatively isolated from other oceans and consists of strongly stratified water masses with distinct histories, nutrient, temperature and salinity characteristics, therefore providing an optimal environment to investigate local adaptation. The globally distributed SAR11 bacterial group consists of multiple ecotypes that are associated with particular marine environments, yet relatively little is known about Arctic SAR11 diversity. Here, we examined SAR11 diversity using ITS analysis and metagenome-assembled genomes (MAGs). Arctic SAR11 assemblages were comprised of the S1a, S1b, S2, and S3 clades, and structured by water mass and depth. The fresher surface layer was dominated by an ecotype (S3-derived P3.2) previously associated with Arctic and brackish water. In contrast, deeper waters of Pacific origin were dominated by the P2.3 ecotype of the S2 clade, within which we identified a novel subdivision (P2.3s1) that was rare outside the Arctic Ocean. Arctic S2-derived SAR11 MAGs were restricted to high latitudes and included MAGs related to the recently defined S2b subclade, a finding consistent with bi-polar ecotypes and showing the potential for Arctic endemism. These results place the stratified Arctic Ocean into the SAR11 global biogeography and have identified SAR11 lineages for future investigation of adaptive evolution in the Arctic Ocean.

Introduction

The SAR11 (Pelagibacterales) group accounts for roughly 30% of bacteria in the ocean surface and 25% of mesopelagic bacteria [1–3]. High phylogenetic diversity and divergence into ecological lineages (i.e. ecotypes) of SAR11 tends to mirror distinct conditions in the oceanic environment [2, 4–6]. SAR11 are classified into clades and subclades based on 16S rRNA genes and further classified into phylotypes based on rRNA internal transcribed spacers (ITS) (Table 1). Three major SAR11 clades (S1, S2 and S3) are currently recognized, with several subclades defined within S1 (S1a, S1b, S1c). Within this diversity approximately 12 phylotypes have been described [2, 5]. To date, many of the phylotypes are thought to have restricted distributions. For example, there are three phylotypes in subclade S1a: P1a.1 is associated with cold environments [2], while P1a.2 and P1a.3 are associated with temperate and tropical environments, respectively [2, 5]. Subclade S1b is generally associated with tropical environments [2, 5, 7], while subclade S1c is associated with deep marine samples [5]. Recently, the clade S2 was divided into subclades S2a and S2b. The S2a subclade is further divided into an oxygen minimum zone subclade (S2a.a) and a tropical subclade (S2a.b) [7]. The S2 subclade phylotypes have been further divided by environment. P2.1, is associated with tropical environments [2], P2.2 is reported from cold, especially Antarctic, waters [2, 8], and the P2.3 phylotype is more ambiguous; associated with both cold waters and tropical deep-sea environments [2, 5]. To date, three phylotypes have been distinguished for clade S3; P3.1 is associated with coastal, mesohaline surface [9] and tropical samples [2], P3.2 is found in brackish [9], temperate [2], as well as Arctic samples [10] and the third S3 phylotype, LD12, is found in freshwater [11] (Table 1).

The Arctic Ocean is small, geographically isolated, and more influenced by freshwater and ice compared to other oceans [12–14]. These distinct characteristics would favour the evolution of locally adapted microbial assemblages, however evidence for this is difficult to document. Biogeographic patterns of taxa reported from the Arctic vary, with reports of

cosmopolitan, bi-polar and endemic species, suggesting a potential for different levels of specialization to local conditions [6, 15–17]. Although SAR11 are reported widely and account for 25 to 30% of Arctic Ocean bacterial assemblages [18, 19], comparatively little information exists on Arctic SAR11 diversity [2]. The extensive knowledge of SAR11 diversity and ecology in other oceans makes it an attractive clade to investigate the potential for ecotypes adapted to Arctic Ocean conditions. The objective of this study was to place SAR11 from the Arctic Ocean within a global context using ITS phylogenetic analysis and comparative genomics using metagenome-assembled genomes (MAGs). We then examined the distribution of SAR11 along a latitudinal transect of the stratified waters of the Canada Basin in the western Arctic Ocean. We targeted three water masses in the upper 200 m, the surface layer, the deep chlorophyll maximum (DCM), which corresponds to a halocline formed by Pacific Summer Water, and the Pacific Winter Water (PWW) layer [20]. These three water masses were previously found to have distinct microbial communities [21] and we hypothesised that different SAR11 ecotypes would be favored within them.

Methods

Sampling and metagenomic data generation

Samples were collected aboard the Canadian Coast Guard Icebreaker CCGS Louis S. St-Laurent from the Western Arctic Ocean from latitudes 73˚ to 79˚ N in October 2015, during the Joint Ocean Ice Study cruise in the Canada Basin (Table 2). Sample collection and preservation, DNA extraction, and metagenomic data generation were as described previously [22], and further details given in the Supplementary Information. The metagenomic data is deposited in the Integrated Microbial Genomes database at the Joint Genome Institute at https://img.jgi.doe.gov, GOLD project ID Ga0133547.

88 **ITS phylogenetic analysis**

89 SAR11 ITS sequences were retrieved from twelve assembled metagenomes, clustered
90 and filtered (Supplementary information). We combined the Arctic ITS sequences with reference
91 sequences from SAR11 genomes and ITS sequences from two previous biogeographic studies
92 [2, 5] and assigned Arctic ITS sequences to phylotypes. We determined phylotype distribution
93 across the samples with PCoA ordination of the Bray-Curtis dissimilarity as described in the
94 Supplementary Information.

95

96 **Metagenome-assembled genomes**

97 We conducted metagenomic binning using the Metabat2 pipeline [23] on a 12 sample
98 coassembly [22, 24]. Further SAR11 bin filtering and cleaning was conducted as described in
99 the Supplementary Information. Seven SAR11 MAGs with low contamination values (<4%) and
100 relatively high completeness values estimated using CheckM [25] (36 % to 47%) were selected
101 for further analysis (Table 3) (Genbank accession numbers XXX). The distribution of orthologs
102 across SAR11 MAGs and reference genomes was analyzed with ProteinOrtho [26]. A set of 39
103 single copy ortholog genes were concatenated and used for phylogenetic analysis using MEGA-
104 cc with a JTT substitution model [27], as further described in the Supplementary Information.

105

106 **Comparative genome content, average nucleotide identity and signatures of selection of**
107 **Arctic MAGs**

108 To identify Arctic-specific SAR11 genes, we compared all genes within the Arctic MAGs
109 to those found within 41 SAR11 reference genomes using ProteinOrtho [26]. The protein
110 functions of genes only found in Arctic MAGs were retrieved following the IMG annotation [28]
111 or the SwissProt database [29]. We calculated average nucleotide identity (ANI) between Arctic

MAGs and reference genomes representing the different subgroups following the method implemented within IMG [28].

Fragment recruitment and SAR11 biogeography.

To determine the prevalence of Arctic MAGs across global marine biomes, we performed reciprocal best hit analysis [22, 30] for 168 metagenomic samples from a range of marine biomes. Metagenomic datasets included 134 TARA ocean datasets [31] which were randomly sub-sampled to a size of approximately 1 GB of reads each to facilitate analysis. These were added to the Arctic metagenomes and 22 Antarctic datasets that included 20 marine and 2 ACE lake samples [2]. We examined the distribution of the seven Arctic MAGs, 48 representative SAR11 reference genomes, and 38 TARA ocean MAGs previously described as SAR11 [32] as described in the Supplemental Information. Best hits from the reciprocal blast were filtered to a minimum length of 100 bp and a minimum identity of 98% and the number of recorded hits per reference genome and metagenomic sample were used to calculate the number of reads recruited per megabase genome per gigabase metagenome (RPMG). See Supplemental Table 1 for all metagenomic datasets and reference genomes.

We utilized the RPMG matrix for PCoA ordination of the Bray-Curtis dissimilarity of the RPMG matrix. The envdist function as implemented in vegan with 999 permutations was used for *Post-Hoc* tests of environmental variables. [33].

Results

Environmental setting

The stratified upper waters of the Canada Basin in the Western Arctic Ocean sampled during the late summer–autumn of 2015 were typical for the region, with warmer and fresher summer waters above colder slightly saltier winter Pacific-origin water (Table 2). Salinity at the surface ranged from 25.6-27.3 and nitrate concentrations were below the detection limit (0.5

138 μM). Below the surface water, relative chlorophyll *a* fluorescence at the DCM was calculated to
139 range from 0.23 to 0.33 mg m^{-3}. The DCM depth varied from 25-79 m, with the shallowest value
140 at CB11, the most northerly station along the western edge of the Canada Basin (Figure 1A-B).
141 Nitrate concentrations increased with depth to up to 16 mg m^{-3} in the PWW, where chlorophyll *a*
142 was low (0.05 μmol m^{-3}).
143
144 **SAR11 ITS sequence diversity**
145 No SAR11 16S rRNA sequences were present in the metagenomes. Within the
146 metagenomic assemblies, we identified 140 high quality SAR11 ITS sequences which were
147 clustered into 111 unique ITS sequence variants (SVs) and combined these SVs with previously
148 published sequences for phylogenetic analysis. Sixty Arctic ITS SVs did not cover the full ITS
149 region and were assigned to phylotypes using their best BLAST hit against full-length SAR11
150 ITS SVs. In total, 6 distinct phylotypes were evident from ITS SVs (Figure 1C), from all major
151 clades S1, S2, and S3 (Figure 1C, Supplemental Figure 1A-D).
152 **Clade S1**. Within S1, we identified the S1a and S1b subclades. The Arctic ITS S1a SVs
153 mostly clustered apart from previously published P1a SVs (Supplemental Figure 1A). These
154 P1a-related SVs were common in samples from the PWW and DCM layer, but nearly absent
155 from the surface layer (Supplemental Figure 1A, Figure 2A). Within subclade S1b
156 (Supplemental Figure 1B), Arctic Ocean ITS SVs were found in two clusters. One cluster
157 corresponded to a previously described P1b.a group [2], while the other grouped within a
158 previously designated non-monophyletic tropical P1b cluster [2], hereafter termed P1b.b
159 (Supplemental Figure 1B), but we were unable to connect these groupings to the three
160 previously described P1b.1-3 phylotypes [5]. P1b.a contained three full-length Arctic SVs and
161 was found nearly exclusively in the PWW (Supplemental Figure 1B, Figure 2A).
162 **Clade S2.** The majority of ITS SVs were assigned to phylotypes within subclade S2
163 (Supplemental Figure 1C). In our phylogenetic analysis, we recovered a monophyletic group

within the P2.3 phylotype, with Arctic SVs clustering with two deep water Red Sea SVs, forming a novel cluster P2.3s1 (Figure 1C, Supplemental Figure 1C). Apart from a single DCM sample, which contained a high number of P1a sequences (coverage of ~670, compared to an average coverage of ~40, Fig. 2A), SVs within the P2.3s1 phylotype were the most frequently detected in Arctic waters, specifically in DCM and PWW samples (Supplemental Figure 1C, Figure 2A). The published ITS SVs previously assigned to P2.1 and P2.2 were interspersed with one another, and we therefore refer to this subgroup as P2.1–2.2. P2.1–2.2 was distributed relatively evenly across sampling depths and locations (Supplemental Figure 1C, Figure 2A).

Clade S3. We recovered two ITS SVs belonging to the brackish Arctic phylotype P3.2 (Supplemental Figure 1D). P3.2 was common in the less saline surface water samples and absent below the DCM (Supplemental Figure 1D**,** Figure 2A).

Phylotype abundance and biogeography

PCoA ordination of samples based on the relative abundance of phylotypes showed that SAR11 assemblages were structured along the first axis in relation to the water layer sampled (Figure 2B). The second axis had less explanatory power and assemblage structuring was mostly driven by the highly abundant P1a SV in the DCM sample of station CB8. PWW samples contained a high contribution of diverse phylotypes including P1a, P1b.a, P1b, P2.1–P2.2 and P2.3 and P2.3s1, while P3.2 was most frequent in surface layer samples. P2.1–P2.2, was not associated with contributions to a specific depth class of samples likely because multiple poorly resolved ecotypes are contained within this group (Figure 2A, B). Secondary fitting of environmental variables onto the ordination did not yield significant results.

Characteristics of Arctic Ocean SAR11 MAGs

We binned SAR11 scaffolds from the Arctic Ocean metagenome co-assembly based on tetranucleotide frequency and coverage across samples. After automated binning and manual

curation, we selected seven high quality SAR11 MAGs for further analysis (Table 3). Quality was based on high N50 values (19-100Kb), a low number of scaffolds (5-26), relatively high completeness (35-47%), and low contamination (0-4%) values (except for 'completeness' following [34]).

While SAR11 ITS sequences were abundant in the metagenomic dataset, none were present in the SAR11 MAGs. Moreover, SAR11 MAGs did not contain rRNA genes and thus their placement into 16S rRNA or ITS-based phylogenies was not possible. Instead we investigated their phylogeny using a concatenated gene tree of Arctic MAGs, reference genomes, and SAGs from the Red Sea and the Eastern Tropical North Pacific oxygen minimum zone (ETNP OMZ) [7] (Figure 3A). The Arctic MAG SAR11–312 belonged to subclade S1a and was prevalent at and below the DCM layer, but not at the surface (Figure 3B). This MAG was more closely related to phylotype P1a.1 than to P1a.3; a placement supported by its maximum ANI value in comparison to a P1a.1 reference genome (80%). Its phylotype could not be determined, as no closely related reference genome contained ITS sequences.

The remaining six MAGs were members of the S2 clade of SAR11. Four of the MAGs (SAR11–112, SAR11–272, SAR11–410, and SAR11-484) were members of the S2a subclade, while two (SAR11–144 and SAR11–196) were most closely related to two SAGs from the S2b subclade that originated from the ETNP OMZ (Figure 3A). The S2a MAGs exhibited differential distributions in the stratified waters of the Arctic Ocean. Three closely related MAGs (SAR11–112, SAR11–272, SAR11–410) were detected at all depths, while a more distantly related S2a MAG (SAR11–484) was only detected in the surface and DCM samples (Figure 3B). The phylogenetic placement of these MAGs was supported by their maximum ANI values in comparison to the S2a reference genome HIMB058 (>75%). In contrast, MAGs SAR11–144 and SAR11–196 were detected in DCM and PWW samples (Figure 3B), and are the most likely MAGs to represent the novel P2.3s1 phylotype defined in this study. However, as S2b reference genomes lack ITS sequences, we were unable to directly confirm that subclade S2b MAGs

belong to the P2.3 phylotype. Due to the absence of any S2b reference genomes with high completeness, we were unable to calculate ANI of SAR11–144 and SAR11–196 with S2b.

Comparison of gene content

To investigate the potential of local adaptation through variation in gene content we identified SAR11 genes specific to the Arctic Ocean by comparing the distribution of orthologs between Arctic MAGs and 60 reference genomes representing the known phylogenetic diversity of SAR11. Of the 2 648 orthologs we identified across SAR11 genomes, 233 were only found in Arctic MAGs and four of these were present in more than one MAG. The majority of these orthologs were poorly characterized proteins (58%, 136 orthologs: COG categories S and R). A further 24% (55 orthologs) were involved in metabolism, 3% (7 orthologs) in cellular processes and signaling and 7% (16 orthologs) in information storage and processing (Supp. Table 3).

Global biogeography of SAR11 genomes

We investigated the biogeographic distribution of Arctic SAR11 populations using Arctic MAGs and available SAR11 genomes. In addition to being present in Arctic DCM and PWW samples, MAG SAR11–312, belonging to subclade S1a, was also detected in the mesopelagic zone of colder oceans (Figure 4A). Arctic S2a MAGs were present in metagenomic datasets that were mostly from polar regions, with the exception of SAR11–112 and SAR11–272, which were present in Mediterranean samples. The Arctic MAGs most closely related to the S2b subclade were even more narrowly distributed. Outside of Arctic samples, SAR11–144 was only detected in South Atlantic DCM layer and North Atlantic mesopelagic samples. In contrast, MAG SAR11–196 was only found in Arctic DCM and PWW samples (Figure 4A).

Other than the Arctic MAGs, most SAR11 genomes recruited poorly across both Arctic and Antarctic metagenomic samples, suggesting they are not significant members of Arctic SAR11 assemblages. However, there were some exceptions: S1a genome HTCC1062, which

242 was the most widely distributed genome, was prevalent in the Antarctic and to a lesser degree
243 the Arctic. The generalist S2a genome HIMB058, which is prevalent in most marine biomes,
244 was present in Arctic and Antarctic samples (Figure 4A). Lastly, S3 genome IMCC9063,
245 characterized as a fresher, arctic phylotype [9], was found across surface samples of several
246 marine biomes, but was most prevalent in Arctic surface waters, followed by Antarctic samples.
247 To further explore SAR11 biogeography in relation to environmental conditions of ocean
248 biomes, we performed PCoA ordination of the Bray-Curtis distance matrix of the RPMG-values
249 (Figure 4B). Overall, metagenomic samples clustered according to their biome and sampling
250 depth. Along the first axis, samples separated according to a temperature and salinity gradient,
251 with colder, fresher samples such as the Arctic and Antarctic environments investigated here
252 distinguished from warmer, saltier environments. Along the second axis, metagenomic samples
253 were separated according to their depth. Deep samples were characterized by high nitrate
254 concentrations while more shallow samples had higher oxygen and Chl *a* concentrations.
255 Notably, Arctic and Antarctic surface and Arctic PWW samples clustered together (Figure 4B, as
256 polar biome samples in lower left quadrant) while Arctic DCM samples were clearly distinct in
257 their compositions (Figure 4B, blue circles). S1a MAG SAR11–312 had a vastly different
258 distribution compared to most other S1a genomes, which were found largely in warmer coastal
259 and trades biomes (e.g. RS39, HIMB5, HIMB083). Arctic S2a MAGs were associated with Arctic
260 PWW and polar surface samples. MAGs SAR11–144 and 196 were differentially distributed
261 from all other genomes investigated here, including the most closely related North Pacific A6S6
262 and B3S13. These results were robust to expanding the analysis to MAGs assembled from the
263 initial TARA Oceans project, which did not include Arctic samples (Supp. Figure 2).
264
265 **Discussion**
266 Ribosomal RNA amplicon surveys of the marine environment have documented the
267 occurrence of SAR11 in Arctic marine systems [35–37], but the diversity and biogeography of

the different lineages has not been investigated in detail. Here we found evidence for the presence of at least six SAR11 phylotypes in the Arctic. The global success of the SAR11 group across marine environments points to a capacity to adapt to a wide array of environmental conditions. While Arctic and Antarctic Oceans are superficially similar environments with respect to solar radiation, the presence of sea ice and algae blooms following ice melt in spring [17], the Arctic Ocean is surrounded by land and freshwater input from large rivers makes the Arctic Ocean a more estuary-like marine biome compared to the Antarctic. The rivers also bring in terrestrial-derived organic matter which represents a potential substrate and selective force on the bacterial communities [22]. Increasing evidence suggests microbial Arctic endemism in other bacterial groups [22], but this is the first genome-level study to our knowledge within the SAR11 group.

The biogeography of ITS phylotypes offers a hint that local selective forces are at work but the relatively coarse nature of the marker masks potential differentiation at the genome level, such that ITS-based phylotypes may not necessarily discern discrete bacterial populations that are locally adapted. Recent advances in sequencing techniques and genome assembly algorithms have resulted in the utilization of MAGs and SAGs to extend the biogeographies of bacteria beyond the phylogenetic marker level and infer patterns of local adaptation at the genome level [7, 23, 38]. Utilizing this approach, gene content of individual SAR11 genomes has been linked to niche partitioning based on nutrients [39], metabolic adaptation to environmental productivity [40] and oxygen minimum zones [7]. However, this approach may have limitations since assembling SAR11 genomes from metagenomes is notoriously difficult due to their high levels of polymorphism [41]. Moreover, recent work suggests that gene content differences alone may not be sufficient to explain SAR11 biogeographic patterns [42]. Nonetheless, our combined MAG-based and marker gene-based results converged, providing support for the existence of novel arctic ecotypes.

294 **Structure of Arctic Ocean SAR11 assemblages**

295 Arctic phylotypes belonging to S1a, S2 and S3 showed preferences linked to different
296 water masses [5]. The fresher Arctic surface waters were dominated by a P3.2 phylotype that
297 had been previously associated with Arctic and brackish waters [9, 10]. In contrast, PWW
298 depths were dominated by the cold [2] and deep [5] phylotype P2.3. The majority of all
299 recovered ITS phylotypes from the Arctic Ocean belonged to the P2.3 phylotype, with one highly
300 abundant phylotype (P2.3s1, Figure 1C) that was previously only represented by a few
301 sequences from deep in the Red Sea [5]. The Arctic Ocean may constitute the centre of the
302 P2.3s1 range distribution [43] and its potential for local adaptation as well as its niche
303 requirements needs further exploration. In contrast to P3.2 and P2.3, the P2.1–P2.2 phylotype
304 showed a broader distribution across depths and samples. However, our phylogenetic analyses
305 failed to recover the two distinct previously published phylotypes P2.1 (tropical) and P2.2 (cold)
306 [2].

307 To further investigate the link between SAR11 diversity and the environment we utilized
308 metagenomic binning to assemble seven Arctic Ocean SAR11 MAGs that belonged to or were
309 closely related to subclades S1a and S2a and S2b, and investigated their biogeographic
310 distribution and transcriptional activity in the Arctic Ocean. Three of the four S2a MAGs
311 recovered here were widely distributed across depth classes and most stringently corresponded
312 to the distribution of the P2.1–P2.2 phylotype. The two MAGs most closely related to S2b
313 reference genomes were mostly present in DCM and PWW waters, mirroring the distribution of
314 the P2.3 and P2.3s1 phylotypes. However, in the absence of direct ITS evidence we were
315 unable to test directly that S2b-related Arctic MAGs correspond to the extremely common
316 P2.3s1 ITS phylotype. Surprisingly, we failed to assemble a S3 MAG, which based on the
317 abundance of corresponding ITS sequences, the high fragment recruitment of the S3 reference
318 genome IMCC9063, should have been recoverable. Moreover, we were unable to detect any

long (>10 kb) SAR11 scaffolds mapping to S3. The absence of any other strong S3 signal, apart from the presence of ITS sequences, is intriguing and warrants further investigation.

Placing Arctic SAR11 into the global biogeography

To determine the biogeography of Arctic MAGs, in comparison to reference genomes representing SAR11 diversity, we performed fragment recruitment analysis across 168 metagenomic datasets from different marine biomes, including the Arctic and two Antarctic habitats: the Southern Ocean and Ace Lake, a saline Antarctic lake. While Arctic MAGs belonging to clade S1 and subclade S2a were also detected in other metagenomic samples, one MAG, which was most closely related to subclade S2b, SAR11–196, was found only in samples from the Arctic Ocean, providing support for the presence of endemic arctic SAR11. However, due to the low genome completeness of this MAG and the lack of closely related reference genomes it was not feasible to link this potential endemism to gene content [7, 40] or selection acting on genes responsible for the apparent local adaptation to the Arctic, as was recently done for a different and wide-spread SAR11 population [42].

Arctic S2a MAGs were sparsely distributed across global metagenomes but were consistently found in Arctic, Antarctic and South Atlantic biomes, indicating a bi-polar distribution of this subclade. However, two of the three MAGs showed recruitment to the warm and salty Mediterranean Sea in both the surface and the DCM layer. Other reference genomes within the subclade were isolated from warm and oxygen-poor environments (e.g. A6S6 and A10S10 [7]), indicating that cross-binning across closely related genomes during the fragment recruitment is theoretically possible. Lastly, in contrast to the majority of other S1a reference genomes investigated, the Arctic S1a MAG SAR11–312 was found across the mesopelagic layer of most Oceans.

It has been discussed whether SAR11 diversity is shaped by neutral evolution [44], or whether the multitude of subclades found within SAR11 represent ecotypes adapted to local

345 environmental conditions (e.g. [45]). Hellweger and colleagues [44] found that distinct
346 populations characterized by up to 0.5% diversity (roughly corresponding, for example, to the
347 rRNA gene diversity between phylotypes P1a.1 and P1a.3 [45]) can arise neutrally. However, a
348 range of other studies [1, 4, 46] have described seasonal and spatial patterning consistent with
349 adaptation to environmental conditions and colonization of environmental niches. Consistent
350 with ecotype adaptations to local conditions, we found that SAR11 assemblages within the
351 metagenomic datasets to be structured by both their biomes and environmental features. This
352 finding, in correspondence with the fact that diversity between most SAR11 subclades and
353 phylotypes is higher than 0.5% [2, 5], indicates that the majority of diversity within the clade is
354 likely to be shaped by adaptation to environmental niches.

355 The majority of SAR11 reference genomes (including most S1a and S1b genomes), as
356 well as MAGs assembled from the global TARA ocean metagenomes [31], are associated with
357 metagenomic datasets from warmer, more shallow environments from the Trades, Coastal and
358 Westerlies Biomes. A second group of reference genomes (mostly S1c) were most common in
359 deeper water metagenomic datasets with lower oxygen and higher nitrate concentrations, as
360 previously reported for S1c genotypes found in nitrate replete oceanic OMZs [7]. Polar biome
361 metagenomic datasets formed a distinct third environmental group. Within the polar datasets,
362 those from the DCM clustered away from all other environments, providing support for the
363 hypothesis that distinct SAR11 assemblages in the samples are selected by environmental
364 conditions, specifically with respect to the Arctic S2a MAGs, which were most frequent here. In
365 contrast to DCM metagenomic datasets, other Arctic datasets clustered closely with those from
366 the Southern Ocean, pointing towards the bi-polar distribution for several of the Arctic S2 MAGs
367 found in these datasets.

368 Seasonal variability of SAR11 phylotypes in association with mixing and phytoplankton
369 bloom events has been extensively described at the Bermuda Atlantic Time-series Study
370 (BATS) site in the Sargasso Sea [4, 46] as well as at coastal environments [45]. The absence of

371 seasonal data in our study makes a direct inference of Arctic phylotype seasonality impossible,
372 but Arctic summer/fall patterns can be compared to those described in other marine biomes. In
373 contrast to summer samples from the BATS, phylotypes belonging to S1a and S1b were
374 relatively scarce in the Arctic surface layer [4, 46], though P1a was abundant in most Arctic
375 DCM samples., Brackish phylotype P3.2 was abundant in most Arctic surface water samples,
376 which is an analogous distribution to phylotype P3.1 that blooms in BATS surface waters in the
377 fall [46]. Whether the apparently more cold-tolerating P3.2 phylotype fills the same seasonal
378 niche as P3.1 in tropical and temperate waters remains to be investigated. S2 phylotypes have
379 been found year-round in deeper waters, but bloom specifically in the spring within the upper
380 mesopelagic at BATS [4, 46], where they are likely involved in DOM remineralization following
381 winter deep mixing [4]. In accordance with these findings, Arctic S2a MAGs were found to be
382 abundant in the euphotic zones of several marine biomes, indicating that the MAGs' niche may
383 be temporally, rather than spatially, defined. In agreement with previous work, we found the S2b
384 reference genomes A6S6 and B3S13 to be highly abundant in the DCM and mesopelagic layer
385 in the North Pacific, but absent in the Arctic Ocean. The absence of any S2b reference genome
386 hits in the Arctic Ocean could point towards Arctic S2b-like MAGs (and the putatively
387 corresponding P2.3s1 phylotype) replacing them as endemic Arctic specialists in the DCM and
388 PWW, but, in the absence of seasonal samples, further work is needed to elucidate the
389 seasonal patterns of Arctic ecotypes.

390 Different degrees of endemism within bacterial communities in the Arctic Ocean have
391 been reported. Patterns vary for different groups of bacteria, but in general Arctic bacterial
392 communities are complex assemblages of bacteria with cosmopolitan, bi-polar, and Arctic-
393 specific distributions, indicating varying degrees of adaptation to local conditions. [6, 16, 17, 47–
394 49]. In the present study, we set out to describe and place Arctic Ocean SAR11 into the global
395 SAR11 biogeography using both marker gene and MAG-based approaches. Assembly of the
396 highly genome-streamlined SAR11 clade, which shows high rates of recombination, from

metagenomes is difficult, complicating comparisons with 16S rRNA-based approaches. Nevertheless, we detected a previously nearly undescribed ITS phylotype, P2.3s1, which was the most common phylotype in most Arctic PWW and DCM samples. Moving from phylotypes to MAGs, we detected Arctic SAR11 genomes with restricted biogeographic distributions indicating the potential for bi-polar and endemic ecotypes. The selective forces shaping these biogeographic distributions as well as the resulting adaptive responses on the genome level merit future investigation.

Acknowledgements

Data were collected aboard the CCGS Louis S. St-Laurent in collaboration with researchers from Fisheries and Oceans Canada at the Institute of Ocean Sciences and Woods Hole Oceanographic Institution's Beaufort Gyre Exploration Program and are available at http://www.whoi.edu/beaufortgyre. We would like to thank both the captain and crew of the CCGS Louis S. St-Laurent, the chief scientist, William J. Williams, and the scientific team aboard. The work was conducted in collaboration with the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User facility, and was supported under Contract No. DE-AC02-05CH11231. Funding Discovery grants (D.W. and C.L.). This study is also a contribution to ArcticNet, a Network of Centers of Excellence (Canada). The Canadian Natural Science and Engineering Research Council (NSERC) Discovery (C.L and D.W) and Northern Supplement (C.L), the Fonds de recherche du Québec Nature et Technologies (FRQNT) supporting Québec-Océan (C.L. D.W) and the Canada Research Chair Program (D.W.) are acknowledged.

Conflict of interest statement

The authors declare no conflict of interest.

References

1. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* 2002; **420**: 806–810.

2. Brown M V, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, et al. Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* 2012; **8**: 1–13.

3. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev Mar Sci* 2017; **9**: 231–255.

4. Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* 2009; **3**: 283–95.

5. Ngugi DK, Stingl U. Combined Analyses of the ITS Loci and the Corresponding 16S rRNA Genes Reveal High Micro- and Macrodiversity of SAR11 Populations in the Red Sea. *PLoS One* 2012; **8**.

6. Pommier T, Pinhassi J, Hagström Å. Biogeographic analysis of ribosomal RNA clusters from marine bacterioplankton. *Aquat Microb Ecol* 2005; **41**: 79–89.

7. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* 2016; **536**: 179–183.

8. García-Martínez J, Rodríguez-Valera F. Microdiversity of uncultured marine prokaryotes: The SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* 2000; **9**: 935–948.

9. Herlemann DPR, Woelk J, Labrenz M, Jürgens K. Diversity and abundance of 'Pelagibacterales' (SAR11) in the Baltic Sea salinity gradient. *Syst Appl Microbiol* 2014; **37**: 601–604.

10. Oh HM, Kang I, Lee K, Jang Y, Lim S Il, Cho JC. Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. *J Bacteriol* 2011; **193**: 3379–3380.

448 11. Bahr M, Hobbie JE, Sogin ML. Bacterial diversity in an arctic lake: A freshwater SAR11
449 cluster. *Aquat Microb Ecol* 1996; **11**: 271–277.
450 12. McLaughlin FA, Carmack EC, Macdonald RW, Melling H, Swift JH, Wheeler PA, et al.
451 The joint roles of Pacific and Atlantic-origin waters in the Canada Basin, 1997-1998.
452 *Deep Res Part I Oceanogr Res Pap* 2004; **51**: 107–128.
453 13. Macdonald RW, McLaughlin FA, Carmack EC. Fresh water and its sources during the
454 SHEBA drift in the Canada Basin of the Arctic Ocean. *Deep Res Part I Oceanogr Res*
455 *Pap* 2002; **49**: 1769–1785.
456 14. Guéguen C, McLaughlin FA, Carmack EC, Itoh M, Narita H, Nishino S. The nature of
457 colored dissolved organic matter in the southern Canada Basin and East Siberian Sea.
458 *Deep Res Part II Top Stud Oceanogr* 2012; **81**–**84**: 102–113.
459 15. Pedrós-Alió C, Potvin M, Lovejoy C. Diversity of planktonic microorganisms in the Arctic
460 Ocean. *Prog Oceanogr* 2015; **139**: 233–243.
461 16. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, et al.
462 Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the
463 surface ocean. *Proc Natl Acad Sci* 2013; **110**: 11463–11468.
464 17. Ghiglione J-F, Galand PE, Pommier T, Pedros-Alio C, Maas EW, Bakker K, et al. Pole-to-
465 pole biogeography of surface and deep marine bacterial communities. *Proc Natl Acad Sci*
466 2012; **109**: 176330–17638.
467 18. Comeau AM, Li WKW, Tremblay JÉ, Carmack EC, Lovejoy C. Arctic ocean microbial
468 community structure before and after the 2007 record sea ice minimum. *PLoS One* 2011;
469 **6**.
470 19. Alonso-Sáez L, Sánchez O, Gasol JM, Balagué V, Pedrós-Alio C. Winter-to-summer
471 changes in the composition and single-cell activity of near-surface Arctic prokaryotes.
472 *Environ Microbiol* 2008; **10**: 2444–2454.
473 20. Shimada K, Itoh M, Nishino S, McLaughlin F, Carmack E, Proshutinsky A. Halocline

474 structure in the Canada Basin of the Arctic Ocean. *Geophys Res Lett* 2005; **32**.

475 21. Monier A, Comte J, Babin M, Forest A, Matsuoka A, Lovejoy C. Oceanographic structure
476 drives the assembly processes of microbial eukaryotic communities. *ISME J* 2015; **9**:
477 990–1002.

478 22. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic
479 evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean
480 Chloroflexi bacteria. *Commun Biol* 2018; **1**.

481 23. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
482 reconstructing single genomes from complex microbial communities. *PeerJ* 2015; **3**.

483 24. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. *Joint Genome Instritute,*
484 *department of energy* . 2014.

485 25. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
486 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
487 *Genome Res* 2015; **25**: 1043–1055.

488 26. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho:
489 Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 2011; **12**.

490 27. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of molecular
491 evolutionary genetics analysis program for automated and iterative data analysis.
492 *Bioinformatics* 2012; **28**: 2685–2686.

493 28. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, et al.
494 The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline
495 (MAP v.4). *Stand Genomic Sci* 2016; **11**.

496 29. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its
497 supplement TrEMBL in 1999. *Nucleic Acids Res* 1999; **27**: 49–54.

498 30. Landry Z, Swa BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. SAR202 genomes from
499 the dark ocean predict pathways for the oxidation of recalcitrant dissolved organic matter.

MBio 2017; **8**.

31. Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, et al. A holistic approach to marine Eco-systems biology. *PLoS Biol* 2011; **9**.

32. Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Lucker S, et al. Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria Are Abundant In Surface Ocean Metagenomes. *Nat Rev Microbiol* 2018; **3**: 804–813.

33. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. Vegan: Community Ecology Package. *URL https//cran.r-project.org, https//github.com/vegandevs/vegan* 2016.

34. Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, et al. Microbial life under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environ Microbiol* 2018; **20**: 2568–2584.

35. Bano N, Hollibaugh JT. Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Appl Environ Microbiol* 2002; **68**: 505–518.

36. Kirchman DL, Cottrell MT, Lovejoy C. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* 2010; **12**: 1132–1143.

37. Galand PE, Potvin M, Casamayor EO, Lovejoy C. Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *ISME J* 2010; **4**: 564–576.

38. Cameron TJ, Temperton B, Swan BK, Landry ZC, Woyke T, Delong EF, et al. Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J* 2014; **8**: 1440–1451.

39. Grote J, Thrash JC, Huggett MJ. Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. 2012; **3**: 1–13.

40. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ. The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity.

Environ Microbiol 2010; **12**: 490–500.

41. Viklund J, Ettema TJG, Andersson SGE. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* 2012; **29**: 599–615.

42. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappe MS, et al. The global biogeography of amino acid variants within a single SAR11 population is governed by natural selection. *bioRxiv* 2017.

43. Kirkpatrick M, Barton NH. Evolution of a Species' Range. *Am Nat* 1997; **150**: 1–23.

44. Hellweger FL, Van Sebille E, Fredrick ND. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science* 2014; **345**: 1246–1349.

45. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 2013; **4**.

46. Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, et al. High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* 2013; **7**: 1322–1332.

47. Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ. Patterns of diversity in marine phytoplankton. *Science (80-)* 2010; **327**: 1509–1511.

48. Thomas MK, Kremer CT, Klausmeier CA, Litchman E. A global pattern of thermal adaptation in marine phytoplankton. *Science (80-)* 2012; **338**: 1085–1088.

49. Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, et al. Global marine bacterial diversity peaks at high latitudes in winter. *ISME J* 2013; **7**: 1669–1677.

50. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* 2018; **5**.

51. Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V., et al. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake

Baikal. *Appl Environ Microbiol* 2018; **84**.

52. Morris RM, Vergin KL, Cho JC, Rappé MS, Carlson CA, Giovannoni SJ. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnol Oceanogr* 2005; **50**: 1687–1696.

53. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, et al. Genomic characterization of two novel SAR11 isolates from the Red Sea, including the first strain of the SAR11 Ib clade. *FEMS Microbiol Ecol* 2017; **93**.

Tables

Table 1: SAR11 clades, subclades and phylotypes, as well as their environmental associations as described in the literature.

Table 2: Western Arctic Ocean metagenome sampling sites, their environmental feature and environmental parameters including sampling depth (Depth), water temperature (Temp.), Salinity, Chl a Fluorescence, CDOM, Oxygen and Nitrate (NO_3).

Table 3: Summary statistics of MAGs from the Western Arctic Ocean.

Figure legends

Figure 1: Study metadata and overview of SAR11 ITS phylogeny. A) Sampling locations of the Western Arctic Ocean metagenomes. B) Environmental profiles of sampling locations showing temperature (˚C), salinity, chlorophyll *a* fluorescence (mg m^{-3}) and nitrate (mmol m^{-3}). C) Maximum likelihood ITS phylogeny including Arctic and reference sequences. Only subclades which contain Arctic ITS sequence types are labeled.

577 Figure 2: Distribution of SAR11 phylotypes in the Arctic Ocean. A) ITS region coverage of Arctic
578 SAR11 phylotypes across samples. B) Principal coordinate analysis ordination of Bray-Curtis
579 dissimilarities of Arctic samples based on the coverage of different ITS phylotypes. Scaling 2 is
580 shown. Dark blue diamonds: PWW samples, blue diamonds: SCM samples, light blue samples:
581 surface samples. Asterisks show phylotypes' weighted average frequency.
582
583 Figure 3: Phylogenetic context and distribution of Arctic SAR11 MAGs. A) Maximum likelihood
584 phylogeny based on 39 concatenated orthologous loci. Only bootstrap values higher than 0.6
585 are shown on the tree. Colored squares and circles at the tips indicate the environment and
586 temperature of origin (if known), matching phylotypes if known are indicated. Colored squares
587 indicate reference genomes with known ITS phylotypes. Grey labeling indicates inferred
588 phylotypes based on phylogenetic placement and distribution of the genomes. B) Coverage of
589 Arctic MAG scaffolds across Western Arctic ocean metagenomes.
590
591 Figure 4: Global biogeography of SAR11 genomes. A) RPMG table of Arctic MAGs and
592 reference genomes by ocean layer and ocean region. Abbreviations for oceans are Arctic (AO),
593 Southern (SO), South Pacific (SP), North Pacific (NP), South Atlantic (SA), North Atlantic (NA),
594 Indian Ocean (IO), Mediterranean (MS) and Red (RS) Seas. For clarity only RPMG values > 3
595 are shown in the figure. We also performed fragment recruitment for five LD12 genomes, but
596 observed no recruitment across all metagenomic datasets. B) Principal coordinate analysis
597 ordination of Bray-Curtis dissimilarities of metagenome samples based on RPMG values of
598 arctic MAGs and reference genomes. Scaling 2 is shown. Arrows indicate significant
599 environmental variables after *Post-Hoc* testing. Asterisks show the weighted average of each
600 genomes' frequency.
601

602 Supplementary Information: Supplementary methods description (docx).

603

604 Supplemental Figure 1: ITS phylogeny of SAR11 subclades. The longest sequence type per
605 Arctic cluster was used in the phylogeny and only sequence types containing complete ITS
606 regions are included. Arctic samples are indicated by coloured squares; Surface (green), DCM
607 (orange) and PWW (blue), Subclades from Table 1: A) S1a, B) S1b, C) S2 and D) S3 (pdf).

608

609 Supplemental Figure 2: Principal coordinate analysis ordination of Bray-Curtis dissimilarities of
610 metagenome samples based on RPMG values of arctic MAGs and reference genomes,
611 including SAR11 MAGs from the TARA ocean circumnavigation expedition [50]. Scaling 2 is
612 shown. Arrows indicate significant environmental variables after *Post-Hoc* testing. Asterisks
613 show the weighted average of each genomes' frequency (pdf).

614

Table 1: SAR11 clades, subclades and phylotypes, as well as their environmental associations as described in the literature (SS: Sargasso Sea).

Clade (16S rRNA)	Subclade (16S rRNA)	Phylotype (ITS)	Reference genome available?	Environmental description		
				Clade	Subclade	Phylotype
S1	S1a	P1a.1	Y	-Freshwater Lake Baikal [51] -Blooms in summer in surface SS [46]	-Summer SS bloom in surface [46] -Bloom in summer mixed layer of SS (high UV, low nutrients) [4, 52] -SS euphotic zone, blooms after mixing [4]	-Cold [2] -Blooms December to June in coastal NA [45]
		P1a.2	N			-Temperate [2]
		P1a.3	Y			-Tropical [2] -Tropical, depth generalist [5] -Blooms July to November in coastal NA [45]
	S1b	P1b.1	Y		-Tropical, oxygen generalist [7] -SS spring bloom [46] -Tropical, epi- and bathypelagic [53] -Tropical [2]	
		P1b.2	Y			
		P1b.3	Y			-Tropical, upper euphotic layer [5]
	S1c	NA2	Y		-Deep, tropical [5, 7] -Deep SS [46]	
S2	S2a.a	-	Y	-Blooms in mesopelagic NA after deep mixing [4]	-SS spring bloom [46]	-OMZ specialist, tropical [7]
	S2a.b	P2.1	Y			-Tropical [2]
		P2.2	N			-Cold [2] -Antarctic [8]
	S2b	P2.3	Y		-Deep SS [46]	-Cold [2] -Deep, tropical [5] -Tropical, OMZ [7] -P2.3s1: Arctic SCM and PWW (this study)
S3		P3.1	Y			-SS surface in autumn [46] -Coastal, mesohaline,

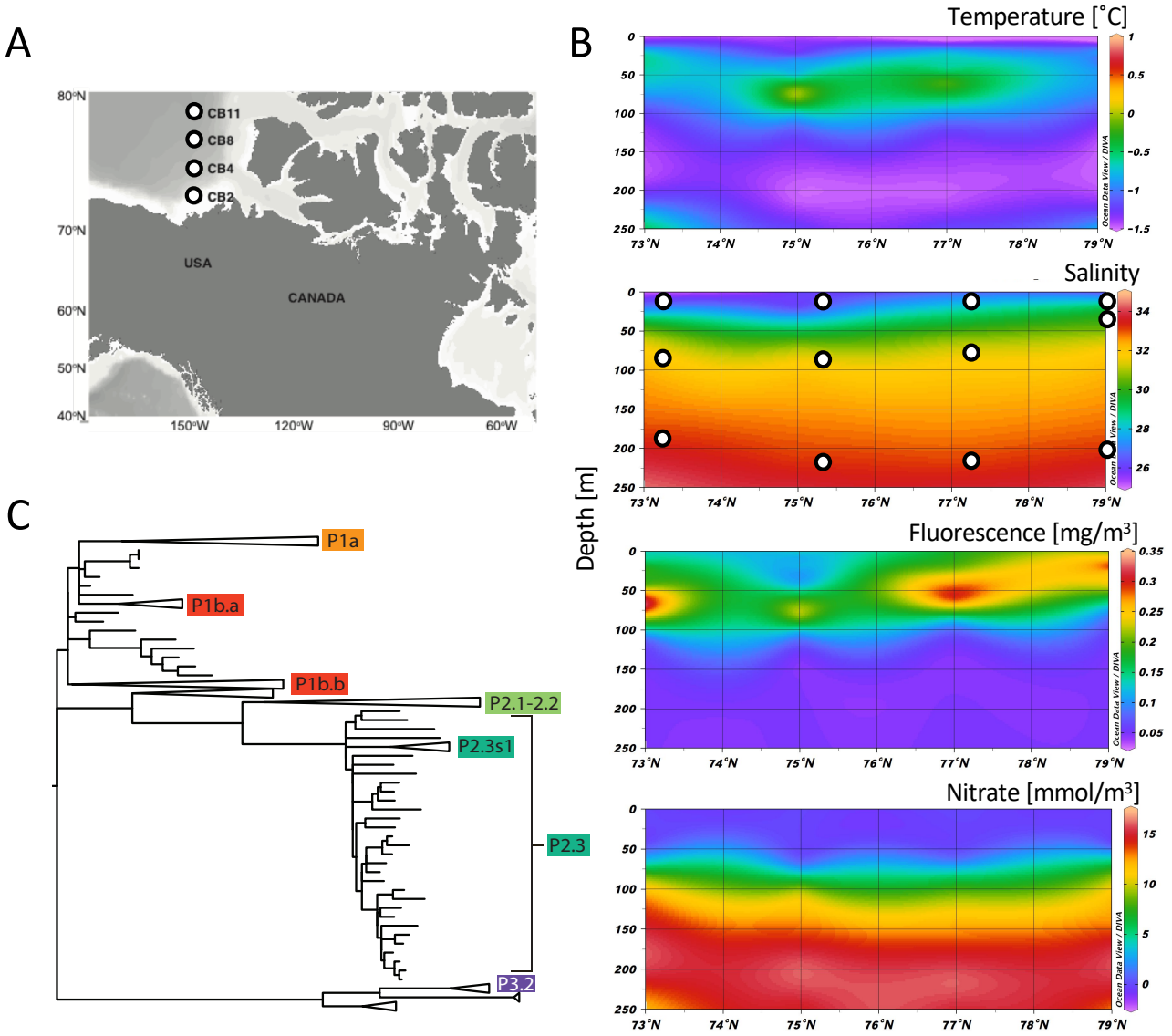
						surface [9] -Tropical [2]
		P3.2	Y			-Arctic [10] -Brackish [9] -Temperate [2]
		LD12	Y			-Freshwater [11]

Table 2: Western Arctic Ocean metagenome sampling sites, their environmental feature and environmental parameters including sampling depth (Depth), water temperature (Temp.), Salinity, Chl a Fluorescence, CDOM, Oxygen and Nitrate (NO_3) limit of detection 0.02 (nd)

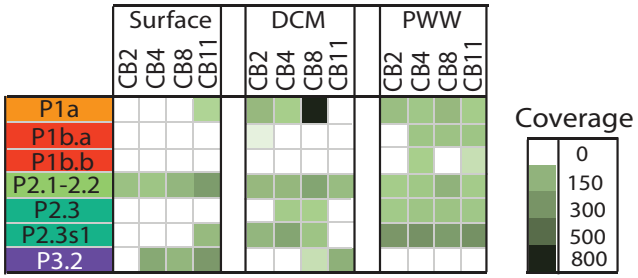
Station	Lat.	Long.	Feature	Depth [m]	Temp. [˚C]	Salinity	Fluorescence [mg m^{-3}]	CDOM [mg m^{-3}]	Oxygen [mmol m^{-3}]	NO_3 [mmol m^{-3}]
CB2_154	73.22	-150.22	SUR	7	-1.26	25.70	0.15	2.54	391.85	nd
CB2_152	73.22	-150.22	DCM	67	-0.86	31.48	0.33	3.9	351.56	4.45
CB2_150	73.22	-150.22	PWW	177	-1.45	33.18	0.05	4.21	281.85	15.98
CB4_138	75.26	-150.07	SUR	5	-1.39	26.14	0.12	2.58	394.08	nd
CB4_136	75.26	-150.07	DCM	79	-0.02	31.16	0.23	3.9	329.14	4.65
CB4_134	75.26	-150.07	PWW	208	-1.47	33.14	0.05	4.36	282.39	16.13
CB8_130	77.1	-150.23	SUR	5	-1.46	27.2	0.19	2.78	395.60	nd
CB8_128	77.1	-150.23	DCM	58	-0.15	30.97	0.30	3.89	383.72	0.35
CB8_126	77.1	-150.23	PWW	213	-1.45	33.14	0.05	4.42	281.09	16.24
CB11_90	79.25	-150.06	SUR	5	-1.48	27.31	0.24	2.67	394.88	nd
CB11_88	79.25	-150.06	DCM	25	-1.04	29.69	0.25	3.22	401.18	nd
CB11_86	79.25	-150.06	PWW	190	-1.46	33.15	0.05	4.35	281.89	15.82

Table 3: Summary statistics of MAGs from the Western Arctic Ocean.

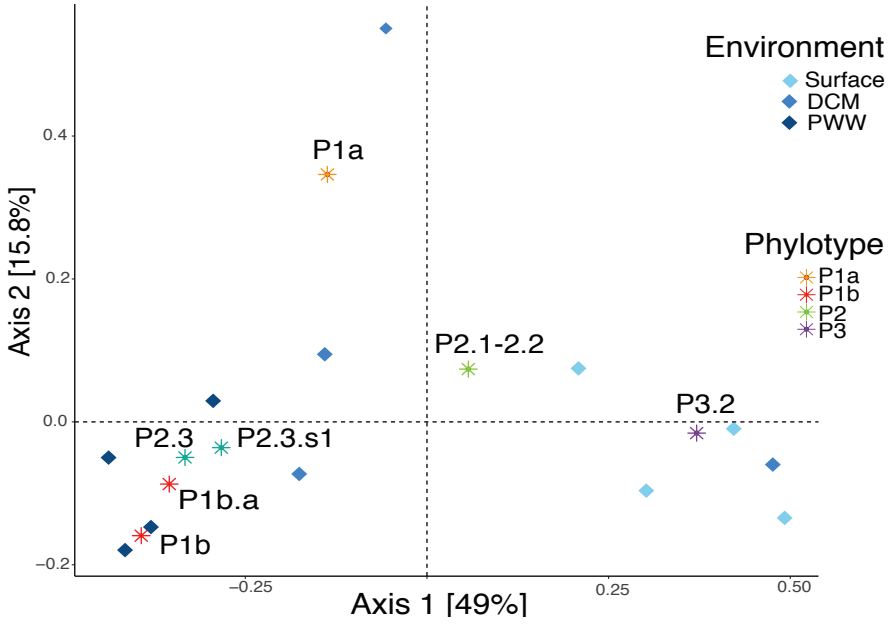
Subclade	MAG	Size (mb)	Cov (x)	GC (%)	Completeness (%)	Contamination (%)	N50 (kb)	# of scaffolds	# of genes	# of unique genes
S1a	312	0.42	27.83	27.3	47.17	1.89	21.23	20	462	33
S2a	112	0.41	8.66	28.31	38.68	1.89	18.76	20	462	49
S2a	272	0.3	17.96	29.52	35.85	0	98.98	5	320	10
S2a	410	0.42	16.77	30.68	24.53	1.89	18.96	26	462	36
S2a	484	0.44	38.18	28.54	35.09	0	38.17	14	481	35
S2	144	0.63	27.78	29.21	41.51	3.77	30.34	23	672	47
S2	196	0.39	23.43	28.5	46.99	0	93.24	8	430	29

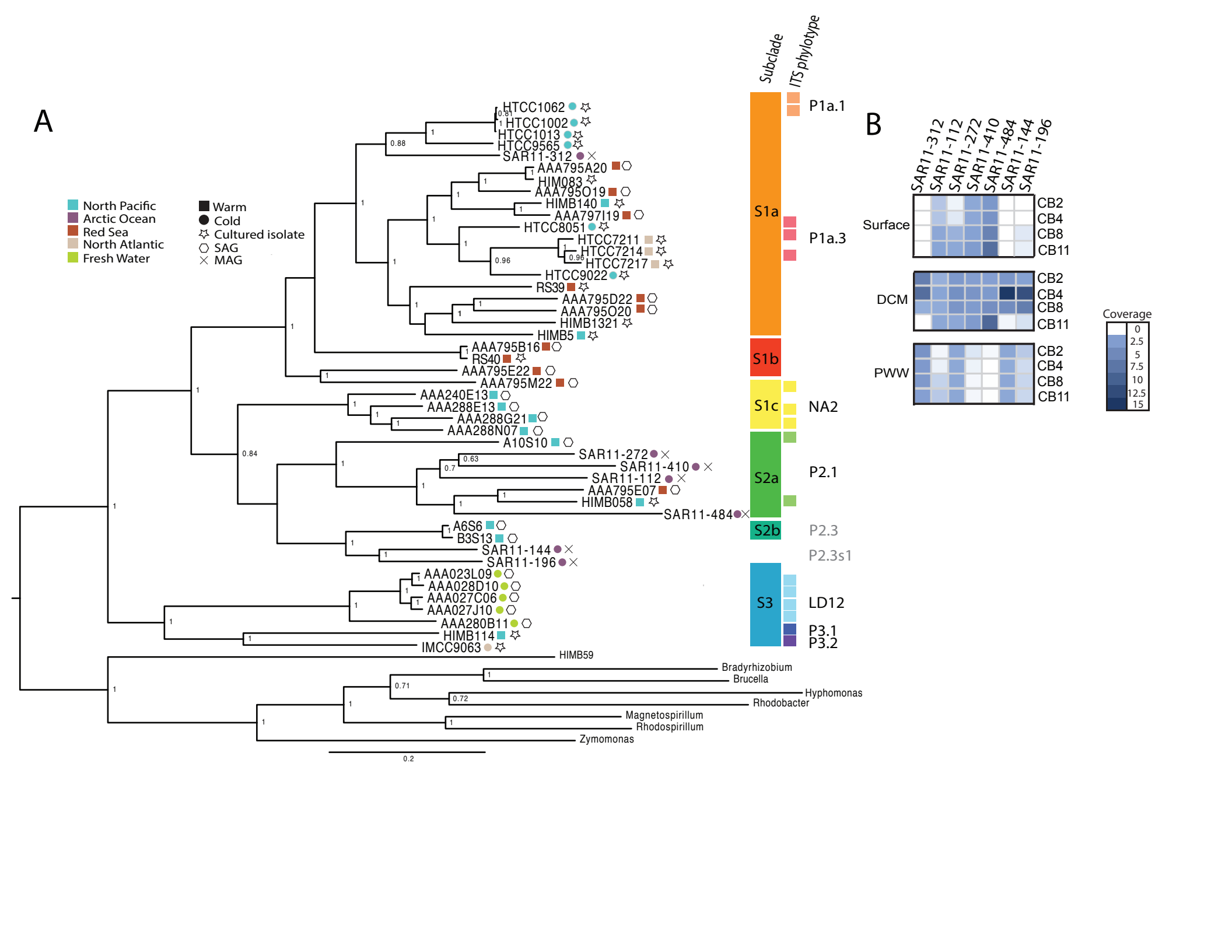


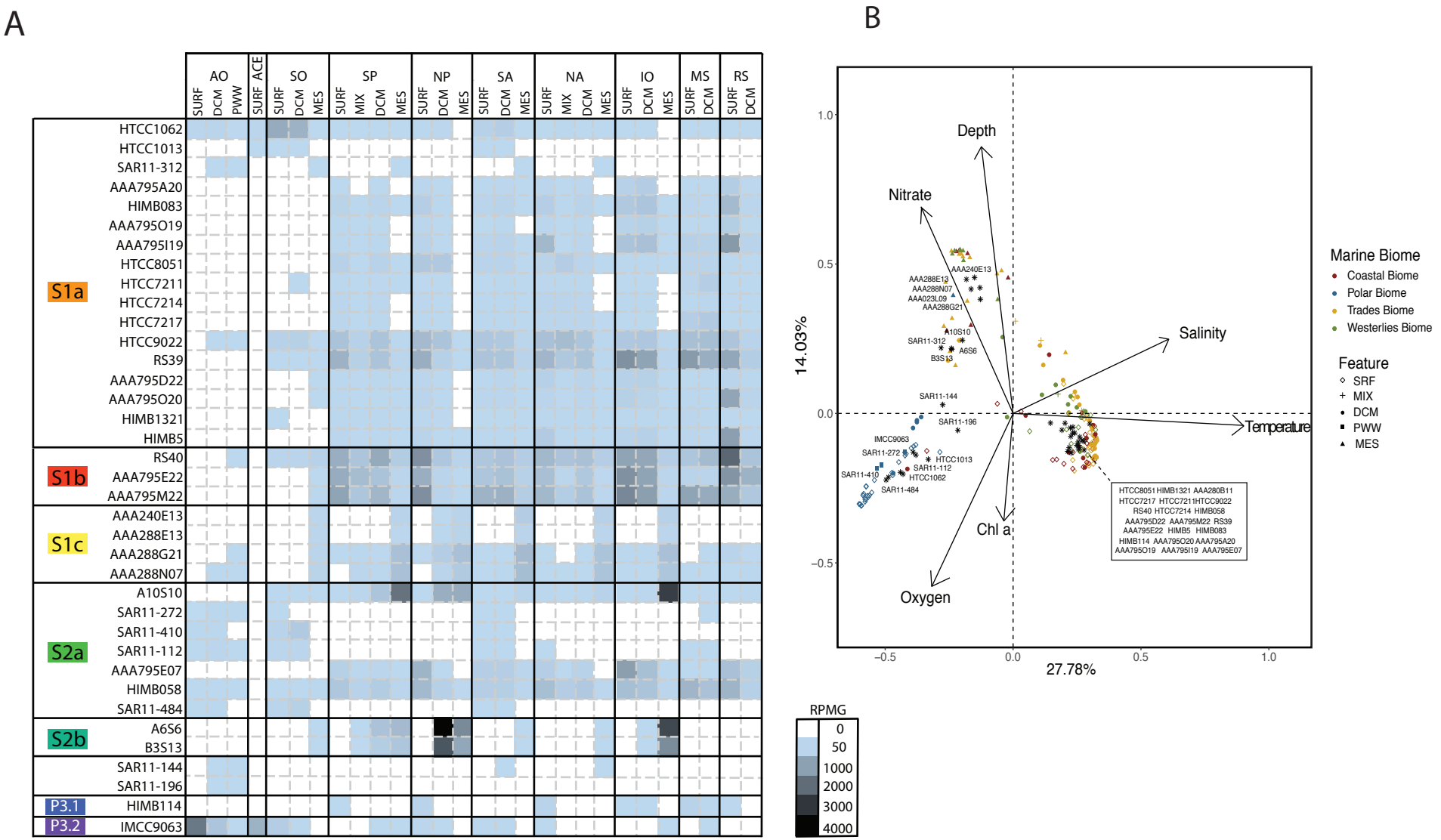
A


Surface
DCM
PWW
CB2
CB4
CB8
CB11
P1a
P1b.a
P1b.b
P2.1-2.2
P2.3
P2.3s1
P3.2
Coverage
0
150
300
500
800

B


Environment
Surface
DCM
PWW
Phylotype
P1a
P1b
P2
P3
P1a
P2.1-2.2
P3.2
P2.3
P2.3.s1
P1b.a
P1b
Axis 2 [15.8%]
Axis 1 [49%]
0.4
0.2
0.0
−0.2
−0.25
0.25
0.50


A
North Pacific
Arctic Ocean
Red Sea
North Atlantic
Fresh Water
Warm
Cold
Cultured isolate
SAG
MAG
Subclade
ITS phylotype
HTCC1062
HTCC1002
HTCC1013
HTCC9565
SAR11-312
AAA795A20
HIMB083
AAA795O19
HIMB140
AAA797I19
HTCC8051
HTCC7211
HTCC7214
HTCC7217
HTCC9022
RS39
AAA795D22
AAA795O20
HIMB1321
HIMB5
AAA795B16
RS40
AAA795E22
AAA795M22
AAA240E13
AAA288E13
AAA288G21
AAA288N07
A10S10
SAR11-272
SAR11-410
SAR11-112
AAA795E07
HIMB058
SAR11-484
A6S6
B3S13
SAR11-144
SAR11-196
AAA023L09
AAA028D10
AAA027C06
AAA027J10
AAA280B11
HIMB114
IMCC9063
HIMB59
Bradyrhizobium
Brucella
Hyphomonas
Rhodobacter
Magnetospirillum
Rhodospirillum
Zymomonas
0.2
S1a
S1b
S1c
S2a
S2b
S3
P1a.1
P1a.3
NA2
P2.1
P2.3
P2.3s1
LD12
P3.1
P3.2
B
SAR11-312
SAR11-112
SAR11-272
SAR11-410
SAR11-484
SAR11-144
SAR11-196
Surface
DCM
PWW
CB2
CB4
CB8
CB11
Coverage
0
2.5
5
7.5
10
12.5
15


A
AO
ACE
SO
SP
NP
SA
NA
IO
MS
RS
SURF
DCM
PWW
MES
MIX
HTCC1062
HTCC1013
SAR11-312
AAA795A20
HIMB083
AAA795O19
AAA795I19
HTCC8051
HTCC7211
HTCC7214
HTCC7217
HTCC9022
RS39
AAA795D22
AAA795O20
HIMB1321
HIMB5
RS40
AAA795E22
AAA795M22
AAA240E13
AAA288E13
AAA288G21
AAA288N07
A10S10
SAR11-272
SAR11-410
SAR11-112
AAA795E07
HIMB058
SAR11-484
A6S6
B3S13
SAR11-144
SAR11-196
HIMB114
IMCC9063
S1a
S1b
S1c
S2a
S2b
P3.1
P3.2
RPMG
0
50
1000
2000
3000
4000
B
Depth
Nitrate
Salinity
Temperature
Chl a
Oxygen
14.03%
27.78%
1.0
0.5
0.0
−0.5
HTCC8051 HIMB1321 AAA280B11
HTCC7217 HTCC7211HTCC9022
RS40 HTCC7214 HIMB058
AAA795D22 AAA795M22 RS39
AAA795E22 HIMB5 HIMB083
HIMB114 AAA795O20 AAA795A20
AAA795O19 AAA795I19 AAA795E07
Marine Biome
Coastal Biome
Polar Biome
Trades Biome
Westerlies Biome
Feature
SRF
MIX
DCM
PWW
MES