# 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, endemism

**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 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 1) different SAR11 ecotypes would be favored within them, and 2) that the distinct characteristics of the Arctic Ocean would favor the existence of endemic SAR11 ecotypes

## 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.

87 **ITS phylogenetic analysis**

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

94

95 **Metagenome-assembled genomes**

96 We conducted metagenomic binning using the Metabat2 pipeline [23] on a 12 sample
97 coassembly [22, 24]. Further SAR11 bin filtering and cleaning was conducted as described in the
98 Supplementary Information. Seven SAR11 MAGs with low contamination values (<4%) and
99 relatively high completeness values estimated using CheckM [25] (36 % to 47%) were selected
100 for further analysis (Table 3) (Genbank accession numbers VBTQ00000000, VBTR00000000,
101 VBTS00000000, VBTT00000000, VBTU00000000, VBTV00000000, VBTW00000000). The
102 distribution of orthologs across SAR11 MAGs and reference genomes was analyzed with
103 ProteinOrtho [26]. A set of 39 single copy ortholog genes were concatenated and used for
104 phylogenetic analysis using MEGA-cc with a JTT substitution model [27], as further described in
105 the Supplementary Information.

106

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

109 To identify Arctic-specific SAR11 genes, we compared all genes within the Arctic MAGs
110 to those found within 41 SAR11 reference genomes using ProteinOrtho [26]. The protein functions
111 of genes only found in Arctic MAGs were retrieved following the IMG annotation [28] or the

112 SwissProt database [29]. We calculated average nucleotide identity (ANI) between Arctic MAGs
113 and reference genomes representing the different subgroups following the method implemented
114 within IMG [28].
115
116 **Fragment recruitment and SAR11 biogeography.**
117 To determine the prevalence of Arctic MAGs across global marine biomes, we performed
118 reciprocal best hit analysis [22, 30] for 163 metagenomic samples from a range of marine biomes.
119 Metagenomic datasets included 129 TARA ocean datasets [31] which were randomly sub-
120 sampled to a size of approximately 1 GB of reads each to facilitate analysis. These were added
121 to the Arctic metagenomes (also randomly subsetted to 1.5 GB of reads) and 22 Antarctic
122 datasets that included 20 marine and 2 ACE lake samples [2]. We examined the distribution of
123 the seven Arctic MAGs and 48 representative SAR11 reference genomes as described in the
124 Supplemental Information. Best hits from the reciprocal blast were filtered to a minimum length of
125 100 bp and a minimum identity of 98% and the number of recorded hits per reference genome
126 and metagenomic sample were used to calculate the number of reads recruited per kilobase
127 genome per gigabase metagenome (RPKG). See Supplemental Table 1 for all metagenomic
128 datasets and reference genomes.
129 We utilized the RPKG matrix for PCoA ordination of its Bray-Curtis dissimilarity. The
130 envdist function as implemented in vegan with 999 permutations was used for *Post-Hoc* tests of
131 environmental variables. [32].
132
133 **Results**
134 **Environmental setting**
135 The stratified upper waters of the Canada Basin in the Western Arctic Ocean sampled
136 during the late summer–autumn of 2015 were typical for the region, with warmer and fresher
137 summer waters above colder slightly saltier winter Pacific-origin water (Table 2). Salinity at the

138 surface ranged from 25.6-27.3 and nitrate concentrations were below the detection limit (0.5 μM).
139 Below the surface water, relative chlorophyll *a* fluorescence at the DCM was calculated to range
140 from 0.23 to 0.33 mg $m^{-3}$. The DCM depth varied from 25-79 m, with the shallowest value at CB11,
141 the most northerly station along the western edge of the Canada Basin (Figure 1A-B). Nitrate
142 concentrations increased with depth to up to 16 mg $m^{-3}$ in the PWW, where chlorophyll *a* was low
143 (0.05 μmol $m^{-3}$).
144
145 **SAR11 ITS sequence diversity**
146 No SAR11 16S rRNA sequences were present in the metagenomes. Within the
147 metagenomic assemblies, we identified 140 high quality SAR11 ITS sequences which were
148 clustered into 111 unique ITS sequence variants (SVs) and combined these SVs with previously
149 published sequences for phylogenetic analysis. Sixty Arctic ITS SVs did not cover the full ITS
150 region and were assigned to phylotypes using their best BLAST hit against full-length SAR11 ITS
151 SVs. In total, 6 distinct phylotypes were evident from ITS SVs (Figure 1C), from all major clades:
152 S1, S2, and S3 (Figure 1C, Supplemental Figure 1A-D).
153 **Clade S1**. Within S1, we identified the S1a and S1b subclades. The Arctic ITS S1a SVs
154 mostly clustered apart from previously published P1a SVs (Supplemental Figure 1A). These P1a-
155 related SVs were common in samples from the PWW and DCM layer, but nearly absent from the
156 surface layer (Supplemental Figure 1A, Figure 2A). Within subclade S1b (Supplemental Figure
157 1B), Arctic Ocean ITS SVs were found in two clusters. One cluster corresponded to a previously
158 described P1b.a group [2], while the other grouped within a previously designated non-
159 monophyletic tropical P1b cluster [2], hereafter termed P1b.b (Supplemental Figure 1B), but we
160 were unable to connect these groupings to the three previously described P1b.1-3 phylotypes [5].
161 P1b.a contained three full-length Arctic SVs and was found nearly exclusively in the PWW
162 (Supplemental Figure 1B, Figure 2A).

163 **Clade S2.** The majority of ITS SVs were assigned to phylotypes within subclade S2
164 (Supplemental Figure 1C). In our phylogenetic analysis, we recovered a monophyletic group
165 within the P2.3 phylotype, with Arctic SVs clustering with two deep water Red Sea SVs, forming
166 a novel cluster P2.3s1 (Figure 1C, Supplemental Figure 1C). Apart from a single DCM sample,
167 which contained a high number of P1a sequences (coverage of ~670, compared to an average
168 coverage of ~40, Fig. 2A), SVs within the P2.3s1 phylotype were the most frequently detected in
169 Arctic waters, specifically in DCM and PWW samples (Supplemental Figure 1C, Figure 2A). The
170 published ITS SVs previously assigned to P2.1 and P2.2 were interspersed with one another, and
171 we therefore refer to this subgroup as P2.1–2.2. P2.1–2.2 was distributed relatively evenly across
172 sampling depths and locations (Supplemental Figure 1C, Figure 2A).
173 **Clade S3.** We recovered two ITS SVs belonging to the brackish Arctic phylotype P3.2
174 (Supplemental Figure 1D). P3.2 was common in the less saline surface water samples and absent
175 below the DCM (Supplemental Figure 1D**,** Figure 2A).
176
177 **Phylotype abundance and biogeography**
178 PCoA ordination of samples based on the relative abundance of phylotypes showed that
179 SAR11 assemblages were structured along the first axis in relation to the water layer sampled
180 (Figure 2B). The second axis had less explanatory power and assemblage structuring was mostly
181 driven by the highly abundant P1a SV in the DCM sample of station CB8. PWW samples
182 contained a high contribution of diverse phylotypes including P1a, P1b.a, P1b, P2.1–P2.2 and
183 P2.3 and P2.3s1, while P3.2 was most frequent in surface layer samples. P2.1–P2.2, was not
184 associated with contributions to a specific depth class of samples likely because multiple poorly
185 resolved ecotypes are contained within this group (Figure 2A, B). Secondary fitting of
186 environmental variables onto the ordination did not yield significant results.
187
188

189 **Characteristics of Arctic Ocean SAR11 MAGs**

190 We binned SAR11 scaffolds from the Arctic Ocean metagenome co-assembly based on
191 tetranucleotide frequency and coverage across samples. After automated binning and manual
192 curation, we selected seven high quality SAR11 MAGs for further analysis (Table 3). Quality was
193 based on high N50 values (19-100Kb), a low number of scaffolds (5-26), relatively high
194 completeness (35-47%), and low contamination (0-4%) values (except for 'completeness'
195 following [33]).

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

206 The remaining six MAGs were members of the S2 clade of SAR11. Four of the MAGs
207 (SAR11–112, SAR11–272, SAR11–410, and SAR11-484) were members of the S2a subclade,
208 while two (SAR11–144 and SAR11–196) were most closely related to two SAGs from the S2b
209 subclade that originated from the ETNP OMZ (Figure 3A). The S2a MAGs exhibited differential
210 distributions in the stratified waters of the Arctic Ocean. Three closely related MAGs (SAR11–
211 112, SAR11–272, SAR11–410) were detected at all depths, while a more distantly related S2a
212 MAG (SAR11–484) was only detected in the surface and DCM samples (Figure 3B). The
213 phylogenetic placement of these MAGs was supported by their maximum ANI values in
214 comparison to the S2a reference genome HIMB058 (>75%). In contrast, MAGs SAR11–144 and

215 SAR11–196 were detected in DCM and PWW samples (Figure 3B) and are the most likely MAGs
216 to represent the novel P2.3s1 phylotype defined in this study. However, as S2b reference
217 genomes lack ITS sequences, we were unable to directly confirm that subclade S2b MAGs belong
218 to the P2.3 phylotype. Due to the absence of any S2b reference genomes with high completeness,
219 we were unable to calculate ANI of SAR11–144 and SAR11–196 with S2b.
220

221 **Comparison of gene content**

222 To investigate the potential of local adaptation through variation in gene content we
223 identified SAR11 genes specific to the Arctic Ocean by comparing the distribution of orthologs
224 between Arctic MAGs and 60 reference genomes representing the known phylogenetic diversity
225 of SAR11. Of the 2 648 orthologs we identified across SAR11 genomes, 233 were only found in
226 Arctic MAGs and four of these were present in more than one MAG. The majority of these
227 orthologs were poorly characterized proteins (58%, 136 orthologs: COG categories S and R). A
228 further 24% (55 orthologs) were involved in metabolism, 3% (7 orthologs) in cellular processes
229 and signaling and 7% (16 orthologs) in information storage and processing (Supplemental Table
230 2).
231

232 **Global biogeography of SAR11 genomes**

233 We investigated the biogeographic distribution of Arctic SAR11 populations using Arctic
234 MAGs and available SAR11 genomes. Arctic MAGs were either bi-polarly distributed or only found
235 in the Arctic Ocean. MAG SAR11–312, belonging to subclade S1a, was detected only in the DCM
236 and PWW layer of the Arctic ocean (Figure 4A). Arctic S2a MAGs fell into two groups. SAR11-
237 410, -112 and -484 were found largely in the surface and DCM layers of the Arctic and Southern
238 Ocean, whereas SAR11-272 was detected only in the DCM and PWW samples of the Arctic
239 Ocean. The Arctic MAGs most closely related to the S2b subclade also showed distributions

240 suggesting Arctic endemism and were only found below the surface layer of the Arctic Ocean
241 (Figure 4A).

242 Other than the Arctic MAGs, most SAR11 genomes recruited poorly across both Arctic
243 and Antarctic metagenomic samples, suggesting they are not significant members of Arctic
244 SAR11 assemblages. However, there were some exceptions: S1a genome HTCC9022, which
245 was among the most widely distributed genomes, was prevalent in the Antarctic and to a lesser
246 degree the Arctic. Two other S1a genomes, HTCC1062 and HTCC1013, were also prevalent in
247 polar samples. Within the S2 clade, only Arctic MAGs recruited in either Arctic or Antarctic
248 samples, indicating that distinct phylotypes inhabit polar niches, in contrast to more temperate
249 generalist S2a genomes such as HIMB058 and AAA795E07 (Figure 4A). Lastly, S3 genome
250 IMCC9063, characterized as a fresher, arctic phylotype [9] was most prevalent in Arctic surface
251 waters, followed by Antarctic samples.

252 To further explore SAR11 biogeography in relation to environmental conditions of ocean
253 biomes, we performed PCoA ordination of the Bray-Curtis distance matrix of the RPKG-values
254 (Figure 4B). Overall, metagenomic samples clustered according to their biome and sampling
255 depth. Along the first axis, samples separated according to a temperature and salinity gradient,
256 with colder, fresher samples such as the Arctic and Antarctic environments investigated here
257 distinguished from warmer, saltier environments. Along the second axis, metagenomic samples
258 were separated according to their depth. Deep samples were characterized by high nitrate
259 concentrations while more shallow samples had higher oxygen and Chl *a* concentrations. Notably,
260 Arctic and Antarctic surface and Arctic DCM samples clustered together (Figure 4B, as polar
261 biome samples in the upper left quadrant), while Arctic PWW samples were distinct in their
262 compositions (Figure 4B, blue crosses). S1a MAG SAR11–312 belonged to an apparently cold-
263 adapted S1a cluster consisting of HTCC1062, HTCC1013, HTCC1040 and HTCC1016 which had
264 a vastly different distribution compared to most other S1a genomes, which were found largely in
265 warmer coastal and trades biomes (e.g. RS39, HIMB5, HIMB083). Arctic S2a MAGs were

266 associated with polar DCM and surface samples. The Arctic MAGs SAR11–144 and 196 were
267 similarly associated with polar samples, and differentially distributed from the most closely related
268 North Pacific genomes A6S6 and B3S13.
269
270 **Discussion**

271 Ribosomal RNA amplicon surveys of the marine environment have documented the
272 occurrence of SAR11 in Arctic marine systems [34–36], but the diversity and biogeography of the
273 different lineages has not been investigated in detail. Here we found evidence for the presence
274 of at least six SAR11 phylotypes in the Arctic. The global success of the SAR11 group across
275 marine environments points to a capacity to adapt to a wide array of environmental conditions.
276 While Arctic and Antarctic Oceans are superficially similar environments with respect to solar
277 radiation, the presence of sea ice and algae blooms following ice melt in spring [17], the Arctic
278 Ocean is surrounded by land and freshwater input from large rivers makes the Arctic Ocean a
279 more estuary-like marine biome compared to the Antarctic. The rivers also bring in terrestrial-
280 derived organic matter which represents a potential substrate and selective force on the bacterial
281 communities [22]. Increasing evidence suggests microbial Arctic endemism in other bacterial
282 groups [22], but this is the first genome-level study to our knowledge within the SAR11 group.

283 The biogeography of ITS phylotypes offers a hint that local selective forces are at work
284 but the relatively coarse nature of the marker masks potential differentiation at the genome level,
285 such that ITS-based phylotypes may not necessarily discern discrete bacterial populations that
286 are locally adapted. Recent advances in sequencing techniques and genome assembly
287 algorithms have resulted in the utilization of MAGs and SAGs to extend the biogeographies of
288 bacteria beyond the phylogenetic marker level and infer patterns of local adaptation at the
289 genome level [7, 23, 37]. Utilizing this approach, gene content of individual SAR11 genomes has
290 been linked to niche partitioning based on nutrients [38], metabolic adaptation to environmental
291 productivity [39] and oxygen minimum zones [7]. However, this approach may have limitations

292 since assembling SAR11 genomes from metagenomes is notoriously difficult due to their high
293 levels of polymorphism [40]. Moreover, recent work suggests that gene content differences alone
294 may not be sufficient to explain SAR11 biogeographic patterns [41]. Nonetheless, our combined
295 MAG-based and marker gene-based results converged, providing support for the existence of
296 novel arctic ecotypes.
297
298 **Structure of Arctic Ocean SAR11 assemblages**
299 Arctic phylotypes belonging to S1a, S2 and S3 showed preferences linked to different
300 water masses [5]. The fresher Arctic surface waters were dominated by a P3.2 phylotype that had
301 been previously associated with Arctic and brackish waters [9, 10]. In contrast, PWW depths were
302 dominated by the cold [2] and deep [5] phylotype P2.3. The majority of all recovered ITS
303 phylotypes from the Arctic Ocean belonged to the P2.3 phylotype, with one highly abundant
304 phylotype (P2.3s1, Figure 1C) that was previously only represented by a few sequences from
305 deep in the Red Sea [5]. The Arctic Ocean may constitute the centre of the P2.3s1 range
306 distribution [42] and its potential for local adaptation as well as its niche requirements needs
307 further exploration. In contrast to P3.2 and P2.3, the P2.1–P2.2 phylotype showed a broader
308 distribution across depths and samples. However, our phylogenetic analyses failed to recover the
309 two distinct previously published phylotypes P2.1 (tropical) and P2.2 (cold) [2].
310 To further investigate the link between SAR11 diversity and the environment we utilized
311 metagenomic binning to assemble seven Arctic Ocean SAR11 MAGs that belonged to or were
312 closely related to subclades S1a and S2a and S2b and investigated their biogeographic
313 distribution and transcriptional activity in the Arctic Ocean. Three of the four S2a MAGs recovered
314 here were widely distributed across depth classes and most stringently corresponded to the
315 distribution of the P2.1–P2.2 phylotype. The two MAGs most closely related to S2b reference
316 genomes were mostly present in DCM and PWW waters, mirroring the distribution of the P2.3
317 and P2.3s1 phylotypes. However, in the absence of direct ITS evidence we were unable to test

318 directly that S2b-related Arctic MAGs correspond to the extremely common P2.3s1 ITS phylotype.
319 Surprisingly, we failed to assemble a S3 MAG, which based on the abundance of corresponding
320 ITS sequences, the high fragment recruitment of the S3 reference genome IMCC9063, should
321 have been recoverable. Moreover, we were unable to detect any long (>10 kb) SAR11 scaffolds
322 mapping to S3. The absence of any other strong S3 signal, apart from the presence of ITS
323 sequences, is intriguing and warrants further investigation.
324
325 **Placing Arctic SAR11 into the global biogeography**
326 To determine the biogeography of Arctic MAGs, in comparison to reference genomes
327 representing SAR11 diversity, we performed fragment recruitment analysis across 163
328 metagenomic datasets from different marine biomes, including the Arctic and two Antarctic
329 habitats: the Southern Ocean and Ace Lake, a saline Antarctic lake. In general, Arctic MAGs
330 showed either Arctic-endemic or bi-polar distributions. Several Arctic S2a MAGs were found only
331 in Arctic and Antarctic biomes, indicating a bi-polar distribution of members of this subclade. One
332 Arctic S1a and S2a MAG, as well as two MAGs which were most closely related to subclade S2b
333 were found only in samples from the Arctic Ocean, providing support for the presence of endemic
334 Arctic SAR11. However, due to the low genome completeness of the MAGs and the lack of closely
335 related reference genomes in the case of the putative S2b genomes it was not feasible to link this
336 potential endemism to gene content [7, 39] or selection acting on genes responsible for the
337 apparent local adaptation to the Arctic, as was recently done for a different and wide-spread
338 SAR11 population [41].
339 It has been discussed whether SAR11 diversity is shaped by neutral evolution [43], or
340 whether the multitude of subclades found within SAR11 represent ecotypes adapted to local
341 environmental conditions (e.g. [44]). Hellweger and colleagues [43] found that distinct populations
342 characterized by up to 0.5% diversity (roughly corresponding, for example, to the rRNA gene
343 diversity between phylotypes P1a.1 and P1a.3 [44]) can arise neutrally. However, a range of other

344 studies [1, 4, 45] have described seasonal and spatial patterning consistent with adaptation to
345 environmental conditions and colonization of environmental niches. Consistent with ecotype
346 adaptations to local conditions, we found that SAR11 assemblages within the metagenomic
347 datasets to be structured by both their biomes and environmental features. This finding, in
348 correspondence with the fact that diversity between most SAR11 subclades and phylotypes is
349 higher than 0.5% [2, 5], indicates that the majority of diversity within the clade is likely to be shaped
350 by adaptation to environmental niches.

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

363 Seasonal variability of SAR11 phylotypes in association with mixing and phytoplankton
364 bloom events has been extensively described at the Bermuda Atlantic Time-series Study (BATS)
365 site in the Sargasso Sea [4, 45] as well as at coastal environments [44]. The absence of seasonal
366 data in our study makes a direct inference of Arctic phylotype seasonality impossible, but Arctic
367 summer/fall patterns can be compared to those described in other marine biomes. In contrast to
368 summer samples from the BATS, phylotypes belonging to S1a and S1b were relatively scarce in
369 the Arctic surface layer [4, 45], though P1a was abundant in most Arctic DCM samples., Brackish

370 phylotype P3.2 was abundant in most Arctic surface water samples, which is an analogous
371 distribution to phylotype P3.1 that blooms in BATS surface waters in the fall [45]. Whether the
372 apparently more cold-tolerating P3.2 phylotype fills the same seasonal niche as P3.1 in tropical
373 and temperate waters remains to be investigated. S2 phylotypes have been found year-round in
374 deeper waters, but bloom specifically in the spring within the upper mesopelagic at BATS [4, 45],
375 where they are likely involved in DOM remineralization following winter deep mixing [4]. In
376 contrast, Arctic S2a MAGs were found to be abundant in the euphotic zones of only Arctic and
377 Antarctic marine biomes, indicating that the MAGs' niche may be spatially, rather than temporally,
378 defined. In agreement with previous work, we found the S2b reference genomes A6S6 and B3S13
379 to be highly abundant in the DCM and mesopelagic layer in the North Pacific but absent in the
380 Arctic Ocean. The absence of any S2b reference genome hits in the Arctic Ocean could point
381 towards Arctic S2b-like MAGs (and the putatively corresponding P2.3s1 phylotype) replacing
382 them as endemic Arctic specialists in the DCM and PWW, but, in the absence of seasonal
383 samples, further work is needed to elucidate the seasonal patterns of Arctic ecotypes.

384 Different degrees of endemism within bacterial communities in the Arctic Ocean have
385 been reported. Patterns vary for different groups of bacteria, but in general Arctic bacterial
386 communities are complex assemblages of bacteria with cosmopolitan, bi-polar, and Arctic-
387 specific distributions, indicating varying degrees of adaptation to local conditions. [6, 16, 17, 46–
388 48]. In the present study, we set out to describe and place Arctic Ocean SAR11 into the global
389 SAR11 biogeography using both marker gene and MAG-based approaches. Assembly of the
390 highly genome-streamlined SAR11 clade, which shows high rates of recombination, from
391 metagenomes is difficult, complicating comparisons with 16S rRNA-based approaches.
392 Nevertheless, we detected a previously nearly undescribed ITS phylotype, P2.3s1, which was the
393 most common phylotype in most Arctic PWW and DCM samples. Moving from phylotypes to
394 MAGs, we detected Arctic SAR11 genomes with both apparently endemic and bi-polar

395 biogeographic distributions. The selective forces shaping these biogeographic distributions as
396 well as the resulting adaptive responses on the genome level merit future investigation.
397

398 **Acknowledgements**

399 Data were collected aboard the CCGS Louis S. St-Laurent in collaboration with
400 researchers from Fisheries and Oceans Canada at the Institute of Ocean Sciences and Woods
401 Hole Oceanographic Institution's Beaufort Gyre Exploration Program and are available at
402 http://www.whoi.edu/beaufortgyre. We would like to thank both the captain and crew of the CCGS
403 Louis S. St-Laurent, the chief scientist, William J. Williams, and the scientific team aboard. The
404 work was conducted in collaboration with the U.S. Department of Energy Joint Genome Institute,
405 a DOE Office of Science User facility, and was supported under Contract No. DE-AC02-
406 05CH11231. Funding Discovery grants (D.W. and C.L.). This study is also a contribution to
407 ArcticNet, a Network of Centers of Excellence (Canada). The Canadian Natural Science and
408 Engineering Research Council (NSERC) Discovery (C.L and D.W) and Northern Supplement (C.L),
409 the Fonds de recherche du Québec Nature et Technologies (FRQNT) supporting Québec-Océan
410 (C.L. D.W) and the Canada Research Chair Program (D.W.) are acknowledged.
411

412 **Conflict of interest statement**

413 The authors declare no conflict of interest.

414 **References**

415 1. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11
416 clade dominates ocean surface bacterioplankton communities. *Nature* 2002; **420**: 806–
417 810.
418 2. Brown M V, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, et al. Global
419 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 II, Cho JC. Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. *J Bacteriol* 2011; **193**: 3379–3380.

11. Bahr M, Hobbie JE, Sogin ML. Bacterial diversity in an arctic lake: A freshwater SAR11 cluster. *Aquat Microb Ecol* 1996; **11**: 271–277.

12. McLaughlin FA, Carmack EC, Macdonald RW, Melling H, Swift JH, Wheeler PA, et al. The joint roles of Pacific and Atlantic-origin waters in the Canada Basin, 1997-1998. *Deep Res Part I Oceanogr Res Pap* 2004; **51**: 107–128.

13. Macdonald RW, McLaughlin FA, Carmack EC. Fresh water and its sources during the

SHEBA drift in the Canada Basin of the Arctic Ocean. *Deep Res Part I Oceanogr Res Pap* 2002; **49**: 1769–1785.

14. Guéguen C, McLaughlin FA, Carmack EC, Itoh M, Narita H, Nishino S. The nature of colored dissolved organic matter in the southern Canada Basin and East Siberian Sea. *Deep Res Part II Top Stud Oceanogr* 2012; **81**–**84**: 102–113.

15. Pedrós-Alió C, Potvin M, Lovejoy C. Diversity of planktonic microorganisms in the Arctic Ocean. *Prog Oceanogr* 2015; **139**: 233–243.

16. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci* 2013; **110**: 11463–11468.

17. Ghiglione J-F, Galand PE, Pommier T, Pedros-Alio C, Maas EW, Bakker K, et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc Natl Acad Sci* 2012; **109**: 176330–17638.

18. Comeau AM, Li WKW, Tremblay JÉ, Carmack EC, Lovejoy C. Arctic ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One* 2011; **6**.

19. Alonso-Sáez L, Sánchez O, Gasol JM, Balagué V, Pedrós-Alio C. Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environ Microbiol* 2008; **10**: 2444–2454.

20. Shimada K, Itoh M, Nishino S, McLaughlin F, Carmack E, Proshutinsky A. Halocline structure in the Canada Basin of the Arctic Ocean. *Geophys Res Lett* 2005; **32**.

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

22. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean

472 Chloroflexi bacteria. *Commun Biol* 2018; **1**.
473 23. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
474 reconstructing single genomes from complex microbial communities. *PeerJ* 2015; **3**.
475 24. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. *Joint Genome Instritute,*
476 *department of energy* . 2014.
477 25. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
478 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
479 *Genome Res* 2015; **25**: 1043–1055.
480 26. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho:
481 Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 2011; **12**.
482 27. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of molecular
483 evolutionary genetics analysis program for automated and iterative data analysis.
484 *Bioinformatics* 2012; **28**: 2685–2686.
485 28. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, et al.
486 The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline
487 (MAP v.4). *Stand Genomic Sci* 2016; **11**.
488 29. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its
489 supplement TrEMBL in 1999. *Nucleic Acids Res* 1999; **27**: 49–54.
490 30. Landry Z, Swa BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. SAR202 genomes from
491 the dark ocean predict pathways for the oxidation of recalcitrant dissolved organic matter.
492 *MBio* 2017; **8**.
493 31. Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, et al. A holistic approach
494 to marine Eco-systems biology. *PLoS Biol* 2011; **9**.
495 32. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. Vegan:
496 Community Ecology Package. *URL https//cran.r-project.org,*
497 *https//github.com/vegandevs/vegan* 2016.

498 33. Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, et al. Microbial life
499 under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-
500 covered Lakes. *Environ Microbiol* 2018; **20**: 2568–2584.
501 34. Bano N, Hollibaugh JT. Phylogenetic composition of bacterioplankton assemblages from
502 the Arctic Ocean. *Appl Environ Microbiol* 2002; **68**: 505–518.
503 35. Kirchman DL, Cottrell MT, Lovejoy C. The structure of bacterial communities in the
504 western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ*
505 *Microbiol* 2010; **12**: 1132–1143.
506 36. Galand PE, Potvin M, Casamayor EO, Lovejoy C. Hydrography shapes bacterial
507 biogeography of the deep Arctic Ocean. *ISME J* 2010; **4**: 564–576.
508 37. Cameron TJ, Temperton B, Swan BK, Landry ZC, Woyke T, Delong EF, et al. Single-cell
509 enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J* 2014; **8**:
510 1440–1451.
511 38. Grote J, Thrash JC, Huggett MJ. Streamlining and Core Genome Conservation among
512 Highly Divergent Members of the SAR11 Clade. 2012; **3**: 1–13.
513 39. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ. The presence of the
514 glycolysis operon in SAR11 genomes is positively correlated with ocean productivity.
515 *Environ Microbiol* 2010; **12**: 490–500.
516 40. Viklund J, Ettema TJG, Andersson SGE. Independent genome reduction and
517 phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* 2012; **29**: 599–
518 615.
519 41. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappe MS, et al. The global
520 biogeography of amino acid variants within a single SAR11 population is governed by
521 natural selection. *bioRxiv* 2017.
522 42. Kirkpatrick M, Barton NH. Evolution of a Species' Range. *Am Nat* 1997; **150**: 1–23.
523 43. Hellweger FL, Van Sebille E, Fredrick ND. Biogeographic patterns in ocean microbes

524 emerge in a neutral agent-based model. *Science (80- )* 2014; **345**: 1246–1349.
525 44. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping:
526 Differentiating between closely related microbial taxa using 16S rRNA gene data.
527 *Methods Ecol Evol* 2013; **4**.
528 45. Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, et al.
529 High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site
530 by phylogenetic placement of pyrosequences. *ISME J* 2013; **7**: 1322–1332.
531 46. Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ. Patterns of diversity in marine
532 phytoplankton. *Science (80- )* 2010; **327**: 1509–1511.
533 47. Thomas MK, Kremer CT, Klausmeier CA, Litchman E. A global pattern of thermal
534 adaptation in marine phytoplankton. *Science (80- )* 2012; **338**: 1085–1088.
535 48. Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, et al. Global
536 marine bacterial diversity peaks at high latitudes in winter. *ISME J* 2013; **7**: 1669–1677.
537 49. Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V., et
538 al. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake
539 Baikal. *Appl Environ Microbiol* 2018; **84**.
540 50. Morris RM, Vergin KL, Cho JC, Rappé MS, Carlson CA, Giovannoni SJ. Temporal and
541 spatial response of bacterioplankton lineages to annual convective overturn at the
542 Bermuda Atlantic Time-series Study site. *Limnol Oceanogr* 2005; **50**: 1687–1696.
543 51. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, et al. Genomic
544 characterization of two novel SAR11 isolates from the Red Sea, including the first strain
545 of the SAR11 Ib clade. *FEMS Microbiol Ecol* 2017; **93**.
546

547 **Tables**

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

550

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

554

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

556

557 **Figure legends**

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

563

564 Figure 2: Distribution of SAR11 phylotypes in the Arctic Ocean. A) ITS region coverage of Arctic
565 SAR11 phylotypes across samples. B) Principal coordinate analysis ordination of Bray-Curtis
566 dissimilarities of Arctic samples based on the coverage of different ITS phylotypes. Scaling 2 is
567 shown. Dark blue diamonds: PWW samples, blue diamonds: SCM samples, light blue samples:
568 surface samples. Asterisks show phylotypes' weighted average frequency.

569

570 Figure 3: Phylogenetic context and distribution of Arctic SAR11 MAGs. A) Maximum likelihood
571 phylogeny based on 39 concatenated orthologous loci. Only bootstrap values higher than 0.6 are
572 shown on the tree. Colored squares and circles at the tips indicate the environment and
573 temperature of origin (if known), matching phylotypes if known are indicated. Colored squares
574 indicate reference genomes with known ITS phylotypes. Grey labeling indicates inferred

575 phylotypes based on phylogenetic placement and distribution of the genomes. B) Coverage of
576 Arctic MAG scaffolds across Western Arctic ocean metagenomes.
577
578 Figure 4: Global biogeography of SAR11 genomes. A) RPKG table of Arctic MAGs and reference
579 genomes by ocean layer and ocean region. Abbreviations for oceans are Arctic (AO), Southern
580 (SO), South Pacific (SP), North Pacific (NP), South Atlantic (SA), North Atlantic (NA), Indian
581 Ocean (IO), Mediterranean (MS) and Red (RS) Seas. For clarity only RPKG values ≥ 0.1 are
582 shown in the figure. We also performed fragment recruitment for five LD12 genomes, but
583 observed no significant recruitment across all metagenomic datasets. B) Principal coordinate
584 analysis ordination of Bray-Curtis dissimilarities of metagenome samples based on RPKG values
585 of arctic MAGs and reference genomes. Scaling 2 is shown. Arrows indicate significant
586 environmental variables after *Post-Hoc* testing. Asterisks show the weighted average of each
587 genomes' frequency.
588
589 Supplementary Information: Supplementary methods description (docx).
590
591 Supplemental Information: Supplemental Figure Legends (docx).
592
593

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 [49]<br><br>-Blooms in summer in surface SS [45] | -Summer SS bloom in surface [45]<br>-Bloom in summer mixed layer of SS (high UV, low nutrients) [4, 50]<br><br>-SS euphotic zone, blooms after mixing [4] | -Cold [2]<br><br>-Blooms December to June in coastal NA [44] |
| | | P1a.2 | N | | | -Temperate [2] |
| | | P1a.3 | Y | | | -Tropical [2]<br>-Tropical, depth generalist [5]<br><br>-Blooms July to November in coastal NA [44] |
| | S1b | P1b.1 | Y | | -Tropical, oxygen generalist [7]<br>-SS spring bloom [45]<br>-Tropical, epi- and bathypelagic [51]<br><br>-Tropical [2] | |
| | | P1b.2 | Y | | | |
| | | P1b.3 | Y | | | -Tropical, upper euphotic layer [5] |
| | S1c | NA2 | Y | | -Deep, tropical [5, 7]<br><br>-Deep SS [45] | |
| S2 | S2a.a | - | Y | -Blooms in mesopelagic NA after deep mixing [4] | -SS spring bloom [45] | -OMZ specialist, tropical [7] |
| | S2a.b | P2.1 | Y | | | -Tropical [2] |
| | | P2.2 | N | | | -Cold [2]<br><br>-Antarctic [8] |
| | S2b | P2.3 | Y | | -Deep SS [45] | -Cold [2]<br>-Deep, tropical [5]<br>-Tropical, OMZ [7]<br>-P2.3s1: Arctic SCM and PWW (this study) |
| S3 | | P3.1 | Y | | | -SS surface in autumn [45] |

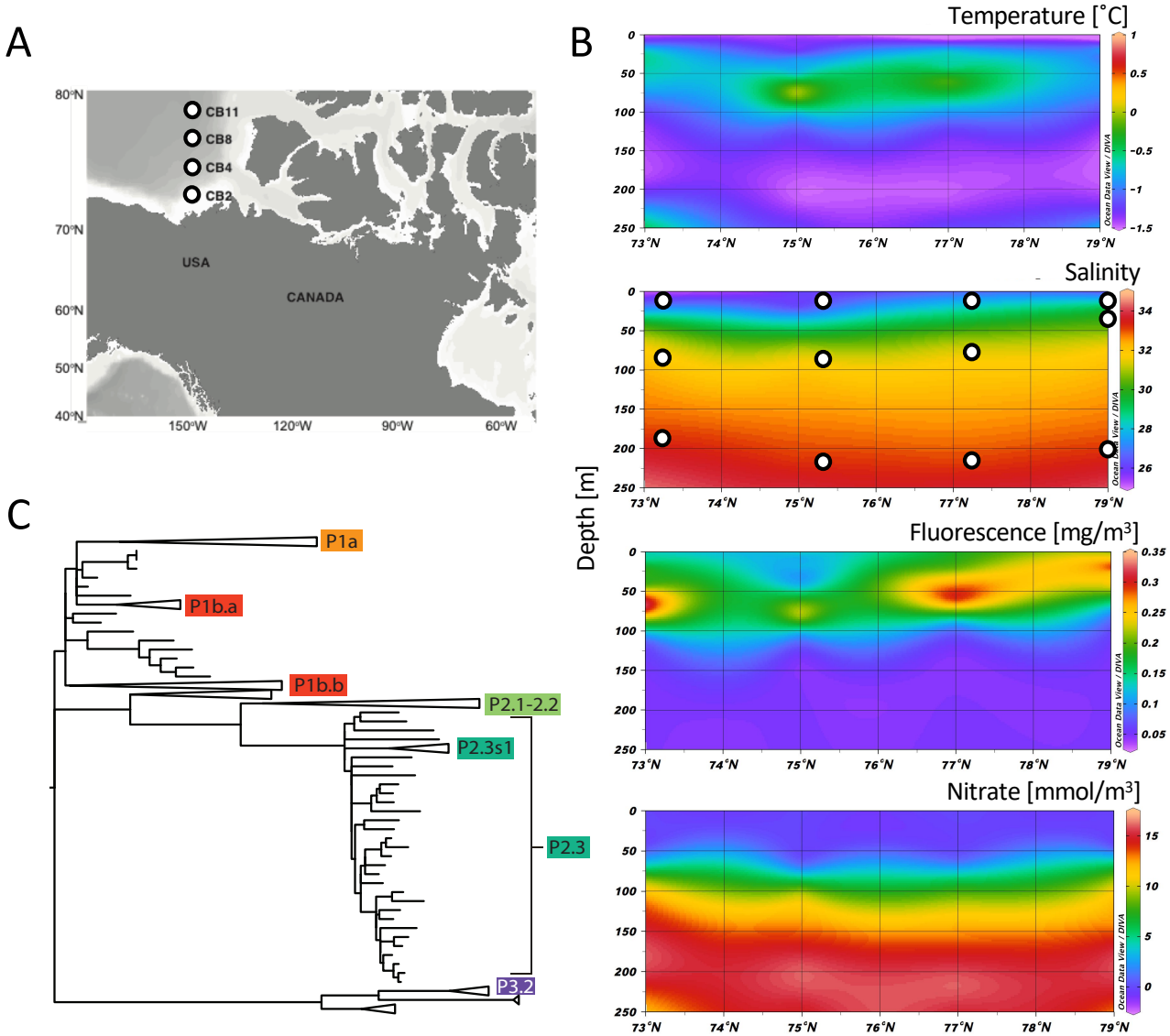
| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | -Coastal, mesohaline, surface [9]<br>-Tropical [2] |
| | | P3.2 | Y | | | -Arctic [10]<br>-Brackish [9]<br>-Temperate [2] |
| | | LD12 | Y | | | -Freshwater [11] |

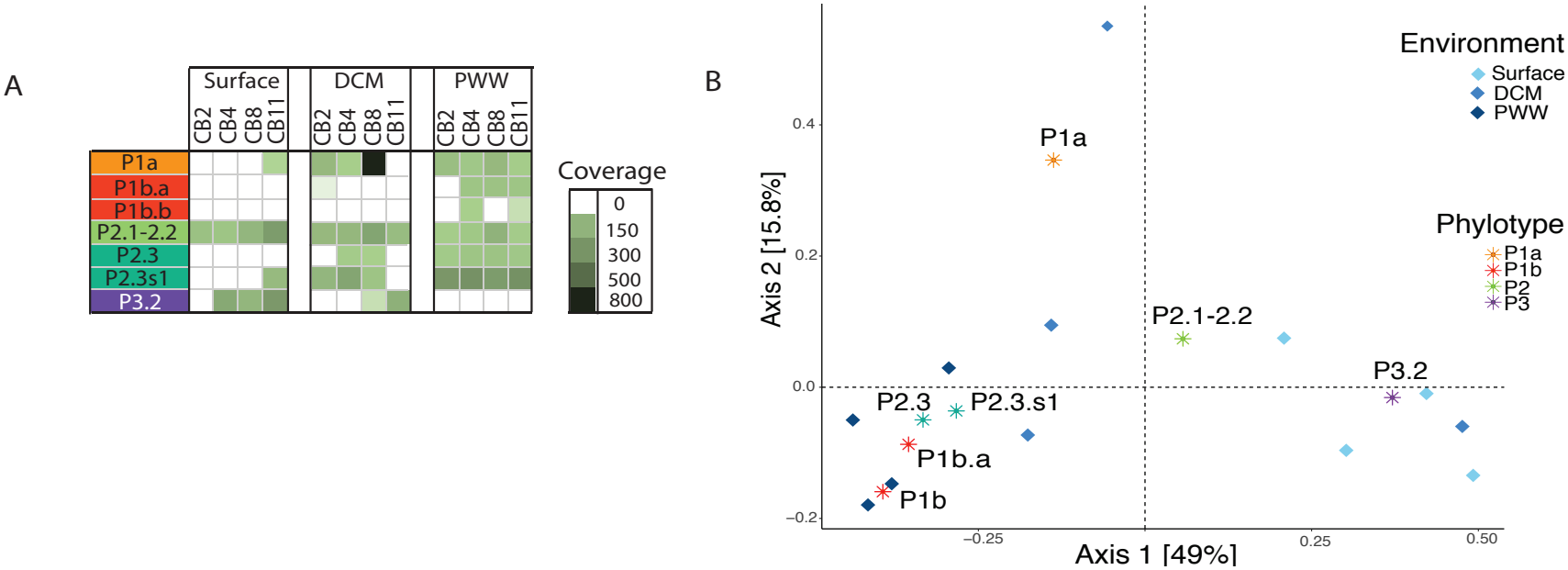
Table 2: Western Arctic Ocean metagenome sampling sites, their environmental feature and environmental parameters including sampling depth (Depth), water temperature (Temp.), Salinity, Chl a Fluorescence, CDOM, Oxygen and Nitrate ($NO_3$) limit of detection 0.02 (nd)

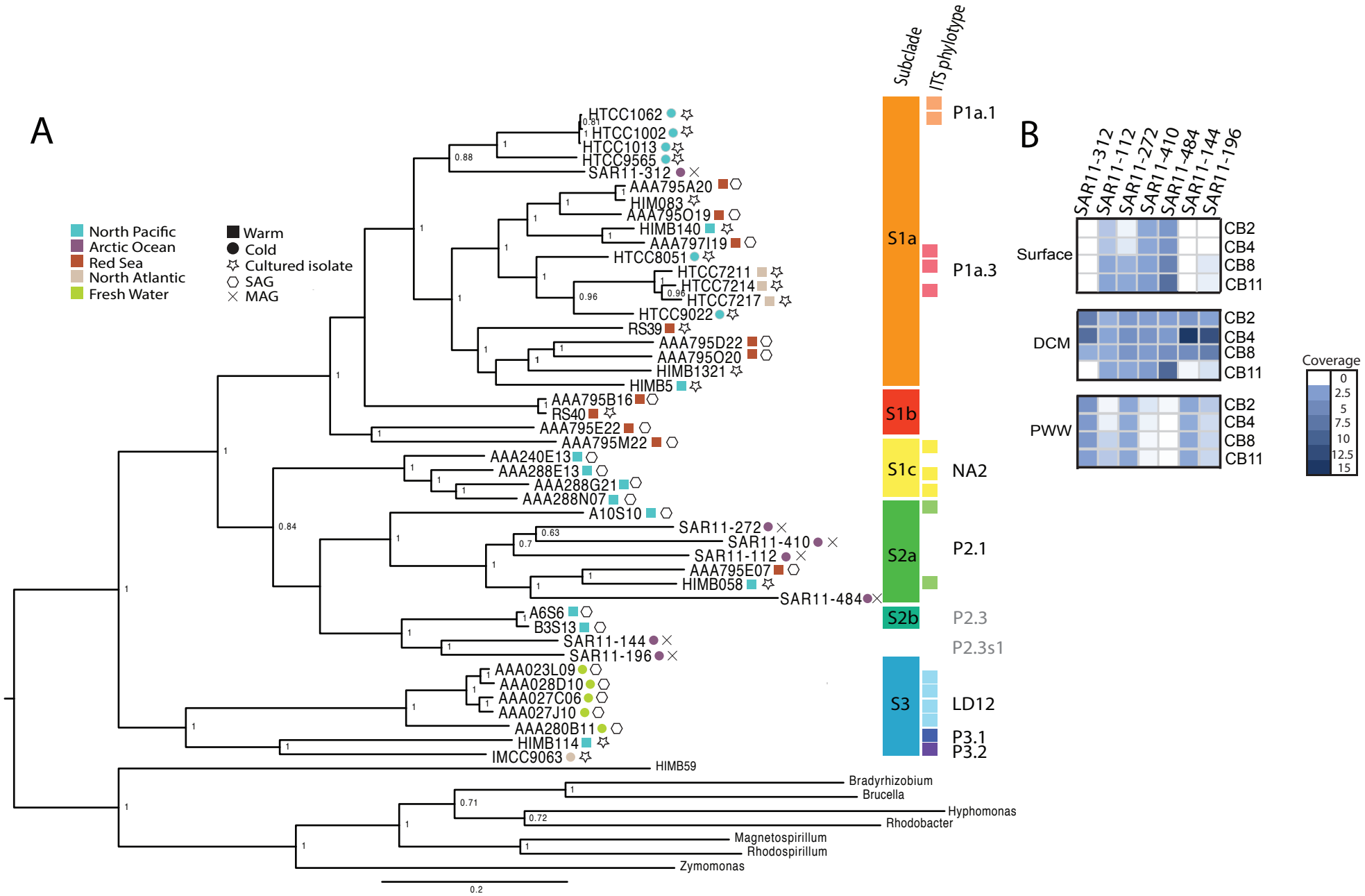
| Station | Lat. | Long. | Feature | Depth [m] | Temp. [˚C] | Salinity | Fluorescence [mg m$^{-3}$] | CDOM [mg m$^{-3}$] | Oxygen [mmol m$^{-3}$] | $NO_3$ [mmol m$^{-3}$] |
|---|---|---|---|---|---|---|---|---|---|---|
| CB2_154 | 73.22 | -150.22 | SUR | 7 | -1.26 | 25.70 | 0.15 | 2.54 | 391.85 | nd |
| 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 |

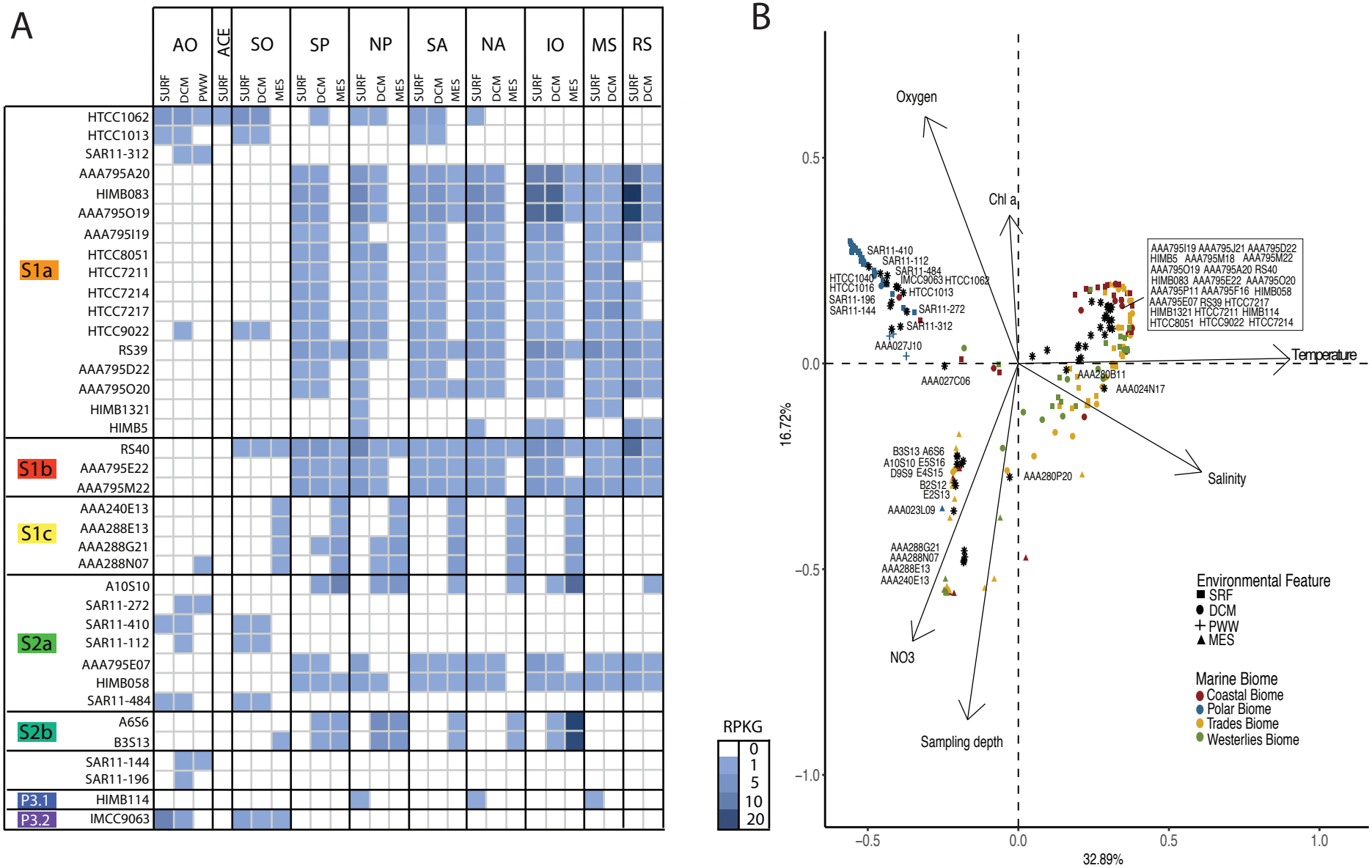
Table 3: Summary statistics of MAGs from the Western Arctic Ocean.

| 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
CB11
CB8
CB4
CB2
USA
CANADA
80°N
70°N
60°N
50°N
40°N
150°W
120°W
90°W
60°W
B
Temperature [˚C]
Salinity
Fluorescence [mg/m³]
Nitrate [mmol/m³]
Depth [m]
73°N
74°N
75°N
76°N
77°N
78°N
79°N
Ocean Data View / DIVA
C
P1a
P1b.a
P1b.b
P2.1-2.2
P2.3s1
P2.3
P3.2


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
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
HIM083
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
S1a
S1b
S1c
S2a
S2b
P3.1
P3.2
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
B
Oxygen
Chl a
Temperature
Salinity
Sampling depth
NO3
16.72%
32.89%
0.5
0.0
−0.5
−1.0
1.0
SAR11-410
SAR11-112
SAR11-484
IMCC9063 HTCC1062
HTCC1040
HTCC1016
HTCC1013
SAR11-196
SAR11-144
SAR11-272
SAR11-312
AAA027J10
AAA027C06
AAA280B11
AAA024N17
AAA280P20
AAA795I19 AAA795J21 AAA795D22
HIMB5 AAA795M18 AAA795M22
AAA795O19 AAA795A20 RS40
HIMB083 AAA795E22 AAA795O20
AAA795P11 AAA795F16 HIMB058
AAA795E07 RS39 HTCC7217
HIMB1321 HTCC7211 HIMB114
HTCC8051 HTCC9022 HTCC7214
B3S13 A6S6
A10S10 E5S16
D9S9 E4S15
B2S12
E2S13
AAA023L09
AAA288G21
AAA288N07
AAA288E13
AAA240E13
Environmental Feature
SRF
DCM
PWW
MES
Marine Biome
Coastal Biome
Polar Biome
Trades Biome
Westerlies Biome
RPKG
0
1
5
10
20