Diversity, biogeography, and evidence for endemism 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, metatranscriptomics, 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, metagenome-assembled genomes (MAGs) and metatranscriptomic datasets. 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.

48 **Introduction**

49 The SAR11 (Pelagibacterales) group accounts for roughly 30% of bacteria in the ocean
50 surface and 25% of mesopelagic bacteria [1–3]. High phylogenetic diversity and divergence into
51 ecological lineages (i.e. ecotypes) of SAR11 tends to mirror distinct conditions in the oceanic
52 environment [2, 4–6]. SAR11 are classified into clades and subclades based on 16S rRNA genes
53 and further classified into phylotypes based on rRNA internal transcribed spacers (ITS) (Table 1).
54 Three major SAR11 clades (S1, S2 and S3) are currently recognized, with several subclades
55 defined within S1 (S1a, S1b, S1c). Within this diversity approximately 12 phylotypes have been
56 described. To date, many of the phylotypes are thought to have restricted distributions. For
57 example, there are three phylotypes in subclade S1a: P1a.1 is associated with cold environments
58 [7], while P1a.2 and P1a.3 are associated with temperate and tropical environments, respectively
59 [5, 7]. Subclade S1b is generally associated with tropical environments [5, 7, 8], while subclade
60 S1c is generally associated with deep marine samples [5]. Recently, the clade S2 was divided
61 into subclades S2a and S2b. The S2a subclade is further divided into an oxygen minimum zone
62 subclade (S2a.a) and a tropical subclade (S2a.b) [8]. The S2 subclade phylotypes have been
63 further divided by environment. P2.1, is associated with tropical environments [7], P2.2 is reported
64 from cold, especially Antarctic, waters [7, 9], and the P2.3 phylotype is more ambiguous;
65 associated with cold waters and tropical deep-sea environments [5, 7]. To date, three phylotypes
66 have been distinguished for clade S3; P3.1 is associated with coastal, mesohaline surface [10]
67 and tropical samples [7], P3.2 is found in brackish [10], temperate [7], as well as Arctic samples
68 [11] and the third S3 phylotype, LD12, is found in freshwater [12] (Table 1).

69 The Arctic Ocean is small, geographically isolated, and more influenced by freshwater and
70 ice compared to other oceans [13–15]. These distinct characteristics would favour the evolution
71 of locally adapted microbial assemblages, however evidence for this is difficult to document.
72 Biogeographic patterns of taxa reported from the Arctic vary, with reports of cosmopolitan, bi-
73 polar and endemic species, suggesting a potential for different levels of specialization to local

conditions [6, 16–18]. Although SAR11 are reported widely and account for 25 to 30% of Arctic Ocean bacterial assemblages [19, 20], comparatively little information exists on Arctic SAR11 diversity [7]. The extensive knowledge of SAR11 diversity and ecology in other oceans makes it an attractive clade to investigate ecotypes adapted to specific 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 [21], These three water masses were previously found to have distinct microbial communities [22] and we hypothesised that 1) different SAR11 ecotypes would be favored within the Arctic water masses, and 2) 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 September 2015, during the Joint Ocean Ice Study cruise in the Canada Basin (Table 2). Sample collection and preservation, DNA extraction, and metagenomic data generation are described previously [23], details are provided in the Supplementary Information.

ITS phylogenetic analysis

SAR11 ITS sequences were retrieved from twelve assembled metagenomes, clustered and filtered (Supplementary information). We combined the Arctic ITS sequences with reference

100 sequences from SAR11 genomes and ITS sequences from two previous biogeographic studies
101 [2, 5] and assigned Arctic ITS sequences to phylotypes. We determined phylotype distribution
102 across the samples with PCoA ordination of the Bray-Curtis dissimilarity as described in the
103 Supplementary Information.
104
105 **Metagenome-assembled genomes**
106 We conducted metagenomic binning using the Metabat2 pipeline [24] on a 12 sample
107 coassembly [23, 25]. Further SAR11 bin filtering and cleaning was conducted as described in the
108 Supplementary Information. Seven SAR11 MAGs with low contamination values (<4%) and
109 relatively high completeness values estimated using CheckM [26] (36 % to 47%) were selected
110 for further analysis (Table 3). The distribution of orthologs across SAR11 MAGs and reference
111 genomes was analyzed with ProteinOrtho [27]. A set of 39 single copy ortholog genes were
112 concatenated and used for phylogenetic analysis using MEGA-cc with a JTT substitution model
113 [28], as further described in the Supplementary Information.
114
115 **Arctic SAR11 transcriptome**
116 RNA samples from surface water, DCM, and PWW at station CB4 (75.24˚ N), were
117 preserved in RNAlater and stored at -80 C. Detailed sample processing is described in the
118 Supplementary Information. Metatranscriptomic data was generated by paired-end Illumina
119 sequencing at the Department of Energy Joint Genome Institute as in [23]. Unassembled
120 metatranscriptomic data was recruited against the protein-coding gene sequences from SAR11
121 MAGs and the S3 reference genome IMCC9063 using bbmap with a 98% cutoff [25]. Recruited
122 reads per kilobase metagenome (RPKB) were calculated as: ((#Number of mapped reads of gene
123 × Total number of mapped reads) / #Total reads in metatranscriptome) × 1000. To obtain MAG-
124 level expression we utilized the average number of mapped reads across all genes.

125 **Comparative genome content, average nucleotide identity and signatures of selection of**
126 **Arctic MAGs**

127 To identify Arctic-specific SAR11 genes, we compared all genes within the Arctic MAGs
128 to those found within 41 SAR11 reference genomes using ProteinOrtho [27]. The protein functions
129 of genes only found in Arctic MAGs were retrieved following the IMG annotation [29] or the
130 SwissProt database [30]. We calculated average nucleotide identity (ANI) between Arctic MAGs
131 and reference genomes representing the different subgroups following the method implemented
132 within IMG [29].

133 We compared the genome content of MAG SAR11-484 to two closely related S2a
134 reference genomes, HIMB058 and AAA795E07. Orthologous genes were identified across the
135 three genomes using ProteinOrtho [27]. We calculated dN/dS between these shared genes using
136 the Young and Nielson method in yn00 as implemented in paml [31] as described in the
137 Supplemental Information, where values greater than one are indicative of diversifying selection.
138

139 **Fragment recruitment and SAR11 biogeography.**

140 To determine the prevalence of Arctic MAGs across global marine biomes, we performed
141 reciprocal best hit analysis [23, 32] for 168 metagenomic samples from a range of marine biomes.
142 Metagenomic datasets included 134 TARA ocean datasets [33] which were randomly sub-
143 sampled to a size of approximately 1 GB of reads each to facilitate analysis. These were added
144 to the Arctic metagenomes and 22 Antarctic datasets that included 20 marine and 2 ACE lake
145 samples [2]. We examined the distribution of the seven Arctic MAGs, 48 representative SAR11
146 reference genomes, and 38 TARA ocean MAGs previously described as SAR11 [34] as described
147 in the Supplemental Information. Best hits from the reciprocal blast were filtered to a minimum
148 length of 100 bp and a minimum identity of 98% and the number of recorded hits per reference
149 genome and metagenomic sample were used to calculate the number of reads recruited per

150 megabase genome per gigabase metagenome (RPMG). See Supplemental Table 1 for all
151 metagenomic datasets and reference genomes.
152 We utilized the RPMG matrix for PCoA ordination of the Bray-Curtis dissimilarity of the
153 RPMG matrix. The envdist function as implemented in vegan with 999 permutations was used for
154 *Post-Hoc* tests of environmental variables. [35].
155
156 **Results**
157 **Environmental setting**
158 The stratified upper waters of the Canada Basin in the Western Arctic Ocean sampled
159 during the late summer–autumn of 2015 were typical for the region, with warmer and fresher
160 summer waters above colder slightly saltier winter Pacific-origin water (Table 2). Salinity at the
161 surface ranged from 25.6-27.3 and nitrate concentrations were below the detection limit (0.5 μM).
162 Below the surface water, relative chlorophyll *a* fluorescence at the DCM was calculated to range
163 from 0.23 to 0.33 mg m^{-3}. The DCM depth varied from 25-79 m, with the shallowest value at CB11,
164 the most northerly station along the western edge of the Canada Basin (Figure 1A-B). Nitrate
165 concentrations increased with depth to up to 16 mg m^{-3} in the PWW, where chlorophyll *a* was low
166 (0.05 μmol m^{-3}).
167
168 **SAR11 ITS sequence diversity**
169 No SAR11 16S rRNA sequences were present in the metagenomes. Within the
170 metagenomic assemblies, we identified 140 high quality SAR11 ITS sequences belonging to 111
171 unique ITS sequence variants and combined these with previously published sequences for
172 phylogenetic analysis. Sixty Arctic ITS sequences did not cover the full ITS region and were
173 assigned to phylotypes using their best BLAST hit against full-length SAR11 ITS sequences. In

174 total, 6 distinct phylotypes were evident from ITS sequences (Figure 1C), from all major clades
175 S1, S2, and S3 (Figure 1C, Figure 2A-D).

176 **Clade S1**. Within S1, we identified the S1a and S1b subclades. The Arctic ITS S1a
177 sequences mostly clustered apart from previously published P1a sequences (Figure 2A). These
178 P1a-related sequences were common in samples from the PWW and DCM layer, but nearly
179 absent from the surface layer (Figure 2A, Figure 3A). Within subclade S1b (Figure 2B), Arctic
180 Ocean ITS sequences were found in two clusters. One cluster corresponded to a previously
181 described P1b.a group [7], while the other grouped within a previously designated non-
182 monophyletic tropical P1b cluster [2], hereafter termed P1b.b (Figure 2B), but we were unable to
183 connect these groupings to the three previously described P1b.1-3 phylotypes [5]. P1b.a
184 contained three full-length Arctic sequences and was found nearly exclusively in the PWW (Figure
185 2B, Figure 3A).

186 **Clade S2.** The majority of ITS sequences were assigned to phylotypes within subclade
187 S2 (Figure 2C). In our phylogenetic analysis, we recovered a monophyletic group within the P2.3
188 phylotype, with Arctic sequences clustering with two deep water Red Sea sequences, forming a
189 novel cluster P2.3s1 (Figure 2C). Apart from a single DCM sample, which contained a high
190 number of P1a sequences (coverage of ~670, compared to an average coverage of ~40, Fig. 3A),
191 sequence types within the P2.3s1 phylotype were the most frequently detected in Arctic waters,
192 specifically in DCM and PWW samples (Figure 2C, Figure 3A). The published ITS sequences
193 previously assigned to P2.1 and P2.2 were interspersed with one another, and we therefore refer
194 to this subgroup as P2.1–2.2. P2.1–2.2 was distributed relatively evenly across sampling depths
195 and locations (Figure 2C, Figure 3A).

196 **Clade S3.** We recovered two ITS sequence type variants belonging to the brackish Arctic
197 phylotype P3.2 (Figure 2D). P3.2 was common in the less saline surface water samples and
198 absent below the DCM (Figure 2D**,** Figure 3A).

199

200 **Phylotype abundance and biogeography**

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

210

211 **Characteristics of Arctic Ocean SAR11 MAGs**

212 We binned SAR11 scaffolds from the Arctic Ocean metagenome co-assembly based on
213 tetranucleotide frequency and coverage across samples. After automated binning and manual
214 curation, we selected seven high quality SAR11 MAGs for further analysis (Table 3). Quality was
215 based on high N50 values (19-100Kb), a low number of scaffolds (5-26), relatively high
216 completeness (35-47%), and low contamination (0-4%) values.

217 SAR11 MAGs did not contain rRNA genes or ITS sequences, and thus their placement
218 into 16S rRNA or ITS-based phylogenies was not possible. Instead we investigated their
219 phylogeny using a concatenated gene tree of Arctic MAGs, reference genomes, and SAGs from
220 the Red Sea and the Eastern Tropical North Pacific oxygen minimum zone (ETNP OMZ) [8]
221 (Figure 4A). The Arctic MAG SAR11–312 belonged to subclade S1a and was prevalent at and
222 below the DCM layer, but not at the surface (Figure 4B). This MAG was more closely related to
223 phylotype P1a.1 than to P1a.3; a placement supported by its maximum ANI value in comparison
224 to a P1a.1 reference genome (80%). Its phylotype could not be determined, as the ITS regions
225 for most of the reference genomes were missing.

226 The remaining six MAGs were members of the S2 clade of SAR11. Four of the MAGs
227 (SAR11–112, SAR11–272, SAR11–410, and SAR11-484) were members of the S2a subclade,
228 while two (SAR11–144 and SAR11–196) were most closely related to two SAGs from the S2b
229 subclade that originated from the ETNP OMZ (Figure 4A). The S2a MAGs exhibited differential
230 distributions in the stratified waters of the Arctic Ocean. Three closely related MAGs (SAR11–
231 112, SAR11–272, SAR11–410) were detected at all depths, while a more distantly related S2a
232 MAG (SAR11–484) was only detected in the surface and DCM samples (Figure 4B). The
233 phylogenetic placement of these MAGs was supported by their maximum ANI values in
234 comparison to the S2a reference genome HIMB058 (>75%). In contrast, MAGs SAR11–144 and
235 SAR11–196 were detected in DCM and PWW samples (Figure 4B), and are the most likely MAGs
236 to represent the novel P2.3s1 phylotype defined in this study. Due to the absence of any S2b
237 reference genomes with high completeness, we were unable to calculate ANI of SAR11–144 and
238 SAR11–196 with S2b.
239
240 **Arctic SAR11 transcriptomes**
241 To assess transcriptional activity of the Arctic SAR11 assemblages, we recruited
242 metatranscriptome data from surface, DCM, and PWW depths at one station (CB4, 75.24˚ N)
243 against the MAGs. RPKB values for all genes within the MAGs roughly followed their coverage in
244 the metagenomes (Figure 4C). S1a MAG SAR11–312 was transcriptionally active in DCM and
245 PWW waters. Transcripts from S2a MAGs SAR11–112, SAR11–272 and SAR11–410 were found
246 at all depths, however they were mostly associated with the DCM layer, and, in the case of MAG
247 SAR11–272, the PWW layer. In contrast, the more distantly related S2a MAG SAR11–484 was
248 found to be active in both DCM and surface waters. S2b-related MAGs SAR11–144 and SAR11–
249 196 were only active in DCM and PWW samples. As we were unable to assemble a S3 genome
250 from the Arctic metagenomes, we utilized the S3-P3.2 genome IMCC9063 to test its
251 transcriptional activity in the Arctic. IMCC9063 transcript abundance was high in the surface layer.

252 Transcripts from most SAR11 genes were not detected or were very rare in the
253 metatranscriptomes. We detected only 13 genes with RPKB values higher than one, belonging to
254 MAGs SAR11–144, SAR11–196 and SAR11–312. All but one of these highly expressed genes
255 were found in the DCM, reflecting the relatively high activities of these MAGs there. Of the 13
256 genes, six are of unknown function. The remaining seven highly expressed genes are involved in
257 nucleotide transport and metabolism (two genes), amino acid transport and metabolism (two
258 genes), co-enzyme transport and metabolism (one gene), cell membrane biogenesis (one gene)
259 and cell cycle control (one gene) (Supplemental Table 2).
260
261 **Comparison of gene content**
262 To investigate the potential of local adaptation through variation in gene content we
263 identified SAR11 genes specific to the Arctic Ocean by comparing the distribution of orthologs
264 between Arctic MAGs and 60 reference genomes representing the known phylogenetic diversity
265 of SAR11. Of the 2 648 orthologs we identified across SAR11 genomes, 233 were only found in
266 Arctic MAGs and four of these were present in more than one MAG. The majority of these
267 orthologs were poorly characterized proteins (58%, 136 orthologs: COG categories S and R). A
268 further 24% (55 orthologs) were involved in metabolism, 3% (7 orthologs) in cellular processes
269 and signaling and 7% (16 orthologs) in information storage and processing (Supp. Table 3).
270 To determine patterns of adaptation, we investigated the relationship between Arctic S2a
271 MAG SAR11–484 and its two most closely related reference genomes: Pacific HIMB058 and Red
272 Sea SAG AAA795E07 (Figure 4D). SAR11–484 shared more orthologs with HIMB058 than with
273 the Red Sea SAG (though this could reflect the higher completeness of the HIMB058 genome).
274 Of the 1 170 orthologs identified in these three genomes, we identified 221 complete genes which
275 were common to all three genomes (Figure 4D) and tested these genes for diversifying selection.
276 None of the genes showed dN/dS values > 1 in any comparison between the three genomes.

277 Thus, none of the orthologs investigated here showed signatures that may indicate diversifying
278 selection and adaptation to local conditions.
279
280 **Global biogeography of SAR11 genomes**
281 We investigated the biogeographic distribution of Arctic SAR11 populations using Arctic
282 MAGs and available SAR11 genomes. In addition to being present in Arctic DCM and PWW
283 samples, MAG SAR11–312, belonging to subclade S1a, was also detected in the mesopelagic
284 zone of colder oceans (Figure 5A). Arctic S2a MAGs were present in metagenomic datasets that
285 were mostly from polar regions, with the exception of SAR11–112 and SAR11–272, which were
286 present in Mediterranean samples. The Arctic MAGs most closely related to the S2b subclade
287 were even more narrowly distributed. Outside of Arctic samples, SAR11–144 was only detected
288 in South Atlantic DCM layer and North Atlantic mesopelagic samples. In contrast, MAG SAR11–
289 196 was only found in Arctic DCM and PWW samples (Figure 5A).
290 Other than the Arctic MAGs, most SAR11 genomes recruited poorly across both Arctic
291 and Antarctic metagenomic samples, suggesting they are not significant members of Arctic
292 SAR11 assemblages. However, there were some exceptions: S1a genome HTCC1062, which
293 was the most widely distributed genome, was prevalent in the Antarctic and to a lesser degree
294 the Arctic. The generalist S2a genome HIMB058, which is prevalent in most marine biomes, was
295 present in Arctic and Antarctic samples (Figure 5A). Lastly, S3 genome IMCC9063, characterized
296 as a fresher, arctic phylotype [10], was found across surface samples of several marine biomes,
297 but was most prevalent in Arctic surface waters, followed by Antarctic samples.
298 To further explore SAR11 biogeography in relation to environmental conditions of ocean
299 biomes, we performed PCoA ordination of the Bray-Curtis distance matrix of the RPMG-values
300 (Figure 5B). Overall, metagenomic samples clustered according to their biome and sampling
301 depth. Along the first axis, samples separated according to a temperature and salinity gradient,
302 with colder, fresher samples such as the Arctic and Antarctic environments investigated here

303 distinguished from warmer, saltier environments. Along the second axis, metagenomic samples
304 were separated according to their depth. Deep samples were characterized by high nitrate
305 concentrations while more shallow samples had higher oxygen and Chl *a* concentrations. Notably,
306 Arctic and Antarctic surface and DCM samples clustered together (Figure 5B, as polar biome
307 samples in lower left quadrant) while Arctic PWW samples were clearly distinct in their
308 compositions (Figure 5B, blue circles). S1a MAG SAR11–312 had a vastly different distribution
309 compared to most other S1a genomes, which were found largely in warmer coastal and trades
310 biomes (e.g. RS39, HIMB5, HIMB083). Arctic S2a MAGs were associated with polar DCM and
311 surface samples. MAGs SAR11–144 and 196 were differentially distributed from all other
312 genomes investigated here, including the most closely related North Pacific A6S6 and B3S13.
313 These results were robust to expanding the analysis to MAGs assembled from the initial TARA
314 Oceans project, which did not include Arctic samples (Supp. Figure 1).

315

316 **Discussion**

317 Ribosomal RNA amplicon surveys of the marine environment have documented the
318 occurrence of SAR11 in Arctic marine systems [36–38], but the diversity and biogeography of the
319 different evolutionary lineages has not been investigated in detail. Here we found evidence for
320 the presence of at least six SAR11 groups in the Arctic. The global success of the SAR11 group
321 across marine environments points at a capacity to adapt to a wide array of environmental
322 conditions. While Arctic and Antarctic Oceans are superficially similar environments with respect
323 to solar radiation, the presence of sea ice and algae blooms following ice melt in spring [18], the
324 Arctic Ocean is surrounded by land and freshwater input from large rivers makes the Arctic Ocean
325 a more estuary-like marine biome compared to the Antarctic. The rivers also bring in terrestrial-
326 derived organic matter which represents a potential substrate and selective force on the bacterial
327 communities [23]. Increasing evidence suggests microbial Arctic endemism in other bacterial
328 groups [23], but this is the first genome-level study to our knowledge within the SAR11 group.

329 The biogeography of ITS phylotypes offers a hint that local selective forces are at work
330 but the relatively coarse nature of the marker masks potential differentiation at the genome level,
331 such that ITS-based phylotypes may not necessarily discern discrete bacterial populations that
332 are locally adapted. Recent advances in sequencing techniques and genome assembly
333 algorithms have resulted in the utilization of MAGs and SAGs to extend the biogeographies of
334 bacteria beyond the phylogenetic marker level and infer patterns of local adaptation at the
335 genome level [8, 24, 39]. Utilizing this approach, gene content of individual SAR11 genomes has
336 been linked to niche partitioning based on nutrients [40], metabolic adaptation to environmental
337 productivity [41] and oxygen minimum zones [8]. However, this approach may have limitations
338 since assembling SAR11 genomes from metagenomes is notoriously difficult due to their high
339 levels of polymorphism [42]. Moreover, recent work suggests that gene content differences alone
340 may not be sufficient to explain SAR11 biogeographic patterns [43]. Nonetheless, our combined
341 MAG-based and marker gene-based results converged, providing evidence of novel arctic
342 ecotypes.
343
344 **Structure of Arctic Ocean SAR11 assemblages**
345 Arctic phylotypes belonging to S1a, S2 and S3 showed preferences linked to different
346 water masses [5]. The fresher Arctic surface waters were dominated by a P3.2 phylotype that had
347 been previously associated with Arctic and brackish waters [10, 11]. In contrast, PWW depths
348 were dominated by the cold [2] and deep [5] phylotype P2.3. The majority of all recovered ITS
349 phylotypes from the Arctic Ocean belonged to the P2.3 phylotype, with one highly abundant
350 phylotype (P2.3s1, Figure 2C) that was previously only represented by a few sequences from
351 deep in the Red Sea [5]. The Arctic Ocean may constitute the centre of the P2.3s1 range
352 distribution [44] and its potential for local adaptation as well as its niche requirements needs
353 further exploration. In contrast to P3.2 and P2.3, the P2.1–P2.2 phylotype showed a broader

354 distribution across depths and samples. However, our phylogenetic analyses failed to recover the
355 two distinct previously published phylotypes P2.1 (tropical) and P2.2 (cold) [2].
356 To further investigate the link between SAR11 diversity and the environment we utilized
357 metagenomic binning to assemble seven Arctic Ocean SAR11 MAGs that belonged to or were
358 closely related to subclades S1a and S2a and S2b (Table 1), and investigated their biogeographic
359 distribution and transcriptional activity in the Arctic Ocean. Three of the four S2a MAGs recovered
360 here were widely distributed across depth classes and most stringently corresponded to the
361 distribution of the P2.1–P2.2 phylotype. The two MAGs most closely related to S2b reference
362 genomes were mostly present in DCM and PWW waters, mirroring the distribution of the P2.3
363 and P2.3s1 phylotypes. However, in the absence of direct ITS evidence we were unable to test
364 directly that S2b-related Arctic MAGs correspond to the extremely common P2.3s1 ITS phylotype.
365 Surprisingly, we failed to assemble a S3 MAG, which based on the abundance of corresponding
366 ITS sequences, the high fragment recruitment of the S3 reference genome IMCC9063, and the
367 high transcriptome signal of this genome in surface waters, should have been recoverable.
368 Moreover, we were unable to detect any long (>10 kb) SAR11 scaffolds mapping to S3. The
369 absence of any other strong S3 signal, apart from the presence of ITS sequences, is intriguing
370 and warrants further investigation.
371
372 **Placing Arctic SAR11 into the global biogeography**
373 To determine the biogeography of Arctic MAGs, in comparison to reference genomes
374 representing SAR11 diversity, we performed fragment recruitment analysis across 168
375 metagenomic datasets from different marine biomes, including the Arctic and two Antarctic
376 habitats: the Southern Ocean and Ace Lake, a saline Antarctic lake. While Arctic MAGs belonging
377 to clade S1 and subclade S2a were also detected in other metagenomic samples, one MAG,
378 which was most closely related to subclade S2b, SAR11–196, was found only in samples from
379 the Arctic Ocean, providing evidence in support of endemism. However, due to the low genome

380 completeness of this MAG and the lack of closely related reference genomes it was not feasible
381 to link this apparent endemism to gene content [8, 41] or selection acting on genes responsible
382 for the apparent local adaptation to the Arctic, as was recently done for a different and wide-
383 spread SAR11 population [43]. Arctic S2a MAGs were sparsely distributed across global
384 metagenomes but were consistently found in Arctic, Antarctic and South Atlantic biomes,
385 indicating a bi-polar distribution of this subclade. However, two of the three MAGs showed
386 recruitment to the warm and salty Mediterranean Sea in both the surface and the DCM layer.
387 Other reference genomes within the subclade were isolated from warm and oxygen-poor
388 environments (e.g. A6S6 and A10S10 [8]), indicating that cross-binning across closely related
389 genomes during the fragment recruitment is theoretically possible. Lastly, in contrast to the
390 majority of other S1a reference genomes investigated, the Arctic S1a MAG SAR11–312 was
391 found across the mesopelagic layer of most Oceans.

392 More broadly speaking, SAR11 assemblages were structured by their biomes and
393 environmental features. The majority of SAR11 reference genomes (including most S1a and S1b
394 genomes), as well as MAGs assembled from the global TARA ocean metagenomes [33], are
395 associated with warmer, more shallow environments from the Trades, Coastal and Westerlies
396 Biomes. A second group of reference genomes (mostly S1c) were most common in deeper water
397 with lower oxygen and higher nitrate concentrations, as previously reported for S1c genotypes
398 found in nitrate replete oceanic OMZs [8]. Polar biome samples formed a distinct third
399 environmental group. Within the polar samples, PWW samples clustered away from all other
400 environments, providing support for the hypothesis that distinct SAR11 assemblages in the
401 samples are selected by environmental conditions, specifically with respect to the Arctic S2b
402 MAGs, which were most frequent here. In contrast to PWW samples, other Arctic samples
403 clustered closely with samples from the Southern Ocean, pointing towards the bi-polar distribution
404 for, among others, the Arctic S2a MAGs that were the most frequent in these samples.

405 Different degrees of endemism within bacterial communities in the Arctic Ocean have
406 been reported. Patterns vary for different groups of bacteria, but in general Arctic bacterial
407 communities are complex assemblages of bacteria with cosmopolitan, bi-polar, and Arctic-
408 specific distributions, indicating varying degrees of adaptation to local conditions. [6, 17, 18, 45–
409 47]. In the present study, we set out to describe and place Arctic Ocean SAR11 into the global
410 SAR11 biogeography using both marker gene and MAG-based approaches. Assembly of the
411 highly genome-streamlined SAR11 clade, which shows high rates of recombination, from
412 metagenomes is difficult, complicating comparisons with 16S rRNA-based approaches.
413 Nevertheless, we detected a previously nearly undescribed ITS phylotype, P2.3s1, which was the
414 most common phylotype in most Arctic PWW and DCM samples. Moving from phylotypes to
415 MAGs, we detected Arctic SAR11 genomes with both apparently endemic and bi-polar
416 biogeographic distributions. The selective forces shaping these biogeographic distributions as
417 well as the resulting adaptive responses on the genome level merit future investigation.
418

419 **Acknowldgements**

420 Data were collected aboard the CCGS Louis S. St-Laurent in collaboration with
421 researchers from Fisheries and Oceans Canada at the Institute of Ocean Sciences and Woods
422 Hole Oceanographic Institution's Beaufort Gyre Exploration Program and are available at
423 http://www.whoi.edu/beaufortgyre. We would like to thank both the captain and crew of the CCGS
424 Louis S. St-Laurent, the chief scientist, William J. Williams, and the scientific team aboard. The
425 work was conducted in collaboration with the U.S. Department of Energy Joint Genome Institute,
426 a DOE Office of Science User facility, and was supported under Contract No. DE-AC02-
427 05CH11231. Funding from the Canadian Natural Science and Engineering Research Council
428 (NSERC) Discovery grants (D.W. and C.L.) and the Canada Research Chair Program (D.W.) are
429 acknowledged.
430

431 **References**

432 1. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11
433 clade dominates ocean surface bacterioplankton communities. *Nature* 2002.
434 2. Brown M V, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, et al. Global
435 biogeography of SAR11 marine bacteria. *Mol Syst Biol* 2012; **8**: 1–13.
436 3. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev*
437 *Mar Sci* 2017.
438 4. Carlson C a, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. Seasonal
439 dynamics of SAR11 populations in the euphotic and mesopelagic zones of the
440 northwestern Sargasso Sea. *ISME J* 2009; **3**: 283–95.
441 5. Ngugi DK, Stingl U. Combined Analyses of the ITS Loci and the Corresponding 16S
442 rRNA Genes Reveal High Micro- and Macrodiversity of SAR11 Populations in the Red
443 Sea. *PLoS One* 2012.
444 6. Pommier T, Pinhassi J, Hagström Å. Biogeographic analysis of ribosomal RNA clusters
445 from marine bacterioplankton. *Aquat Microb Ecol* 2005.
446 7. Brown M V., Lauro FM, Demaere MZ, Muir L, Wilkins D, Thomas T, et al. Global
447 biogeography of SAR11 marine bacteria. *Mol Syst Biol* 2012.
448 8. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, et al. SAR11
449 bacteria linked to ocean anoxia and nitrogen loss. *Nature* 2016.
450 9. García-Martínez J, Rodríguez-Valera F. Microdiversity of uncultured marine prokaryotes:
451 The SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* 2000.
452 10. Herlemann DPR, Woelk J, Labrenz M, Jürgens K. Diversity and abundance of
453 'Pelagibacterales' (SAR11) in the Baltic Sea salinity gradient. *Syst Appl Microbiol* 2014;
454 **37**: 601–604.
455 11. Oh HM, Kang I, Lee K, Jang Y, Lim S II, Cho JC. Complete genome sequence of strain
456 IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. *J Bacteriol* .

2011.

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

13. 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.

14. 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.

15. 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.

16. Pedrós-Alió C, Potvin M, Lovejoy C. Diversity of planktonic microorganisms in the Arctic Ocean. *Prog Oceanogr* 2015.

17. 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.

18. 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.

19. 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.

20. 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.

21. Shimada K, Itoh M, Nishino S, McLaughlin F, Carmack E, Proshutinsky A. Halocline

483 structure in the Canada Basin of the Arctic Ocean. *Geophys Res Lett* 2005.
484 22. Monier A, Comte J, Babin M, Forest A, Matsuoka A, Lovejoy C. Oceanographic structure
485 drives the assembly processes of microbial eukaryotic communities. *ISME J* 2015.
486 23. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic
487 evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean
488 Chloroflexi bacteria. *Commun Biol* 2018.
489 24. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
490 reconstructing single genomes from complex microbial communities. *PeerJ* 2015.
491 25. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. *Joint Genome Instritute,*
492 *department of energy* . 2014.
493 26. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
494 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
495 *Genome Res* 2015.
496 27. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho:
497 Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 2011.
498 28. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of molecular
499 evolutionary genetics analysis program for automated and iterative data analysis.
500 *Bioinformatics* 2012.
501 29. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, et al.
502 The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline
503 (MAP v.4). *Stand Genomic Sci* 2016.
504 30. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its
505 supplement TrEMBL in 1999. *Nucleic Acids Res* . 1999.
506 31. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007.
507 32. Landry Z, Swa BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. SAR202 genomes from
508 the dark ocean predict pathways for the oxidation of recalcitrant dissolved organic matter.

509 *MBio* 2017.
510 33. Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, et al. A holistic approach
511 to marine Eco-systems biology. *PLoS Biol* 2011.
512 34. Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Lucker S, et al. Nitrogen-Fixing
513 Populations Of Planctomycetes And Proteobacteria Are Abundant In The Surface Ocean.
514 *doi.org* 2017.
515 35. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, et al. Vegan:
516 Community Ecology Package. *URL https//cran.r-project.org,*
517 *https//github.com/vegandevs/vegan* 2016.
518 36. Bano N, Hollibaugh JT. Phylogenetic composition of bacterioplankton assemblages from
519 the Arctic Ocean. *Appl Environ Microbiol* 2002.
520 37. Kirchman DL, Cottrell MT, Lovejoy C. The structure of bacterial communities in the
521 western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ*
522 *Microbiol* 2010.
523 38. Galand PE, Potvin M, Casamayor EO, Lovejoy C. Hydrography shapes bacterial
524 biogeography of the deep Arctic Ocean. *ISME J* 2010; **4**: 564–576.
525 39. Cameron TJ, Temperton B, Swan BK, Landry ZC, Woyke T, Delong EF, et al. Single-cell
526 enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J* 2014.
527 40. Grote J, Thrash JC, Huggett MJ. Streamlining and Core Genome Conservation among
528 Highly Divergent Members of the SAR11 Clade. 2012; **3**: 1–13.
529 41. Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ. The presence of the
530 glycolysis operon in SAR11 genomes is positively correlated with ocean productivity.
531 *Environ Microbiol* 2010; **12**: 490–500.
532 42. Viklund J, Ettema TJG, Andersson SGE. Independent genome reduction and
533 phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* 2012.
534 43. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappe MS, et al. The global

535 biogeography of amino acid variants within a single SAR11 population is governed by
536 natural selection. *bioRxiv* 2017.
537 44. Kirkpatrick M, Barton NH. Evolution of a Species' Range. *Am Nat* 1997.
538 45. Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ. Patterns of diversity in marine
539 phytoplankton. *Science (80-)* 2010.
540 46. Thomas MK, Kremer CT, Klausmeier CA, Litchman E. A global pattern of thermal
541 adaptation in marine phytoplankton. *Science (80-)* 2012.
542 47. Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, et al. Global
543 marine bacterial diversity peaks at high latitudes in winter. *ISME J* 2013.
544 48. Cabello-Yeves PJ, Zemskay TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov V V., et
545 al. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake
546 Baikal. *Appl Environ Microbiol* 2018.
547 49. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, et al. Genomic
548 characterization of two novel SAR11 isolates from the Red Sea, including the first strain
549 of the SAR11 Ib clade. *FEMS Microbiol Ecol* 2017.
550

551

552 **Tables**

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

555

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

559

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

561

562 **Figure legends**

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

568

569 Figure 2: ITS phylogeny of SAR11 subclades. The longest sequence type per Arctic cluster was
570 used in the phylogeny and only sequence types containing complete ITS regions are included.
571 Arctic samples are indicated by coloured squares; Surface (green), DCM (orange) and PWW
572 (blue), Subclades from Table 1: A) S1a, B) S1b , C) S2 and D) S3.

573

574 Figure 3: Distribution of SAR11 phylotypes in the Arctic Ocean. A) ITS region coverage of Arctic
575 SAR11 phylotypes across samples. B) Principal coordinate analysis ordination of Bray-Curtis
576 dissimilarities of Arctic samples based on the coverage of different ITS phylotypes. Scaling 2 is

577 shown. Dark blue diamonds: PWW samples, blue diamonds: SCM samples, light blue samples:
578 surface samples. Asterisks show phylotypes' weighted average frequency.
579
580 Figure 4: Phylogenetic context, distribution, transcriptional activity and gene orthology of Arctic
581 SAR11 MAGs. A) Maximum likelihood phylogeny based on 39 concatenated orthologous loci.
582 Only bootstrap values higher than 0.6 are shown on the tree. Colored squares and circles at the
583 tips indicate the environment and temperature of origin (if known), matching phylotypes if known
584 are indicated. B) Coverage of Arctic MAG scaffolds across Western Arctic ocean metagenomes.
585 C) RPKB values of the transcriptomes of station CB4 of Arctic MAGs. D) Venn diagram of gene
586 orthology between the Arctic S2a MAG SAR11-484 and the reference genomes AAA795E07 and
587 HIMB058.
588
589 Figure 5: Global biogeography of SAR11 genomes. A) RPMG table of Arctic MAGs and reference
590 genomes by ocean layer and ocean region. Abbreviations for oceans are Arctic (AO), Southern
591 (SO), South Pacific (SP), North Pacific (NP), South Atlantic (SA), North Atlantic (NA), Indian
592 Ocean (IO), Mediterranean (MS) and Red (RS) Seas. For clarity only RPMG values > 3 are shown
593 in the figure. B) Principal coordinate analysis ordination of Bray-Curtis dissimilarities of
594 metagenome samples based on RPMG values of arctic MAGs and reference genomes. Scaling
595 2 is shown. Arrows indicate significant environmental variables after *Post-Hoc* testing. Asterisks
596 show the weighted average of each genomes' frequency.
597
598 Supplemental information
599
600 Supplemental Table 1: List of all genomes and metagenomic samples used in this study.
601

602 Supplemental Table 2: List of transcribed genes of Arctic SAR11 MAGs, as well as their COGs

603

604 Supplemental Table 3: List of genes unique to arctic SAR11 MAGs, as well as their COGs.

605 Supplemental Figure 1: Principal coordinate analysis ordination of Bray-Curtis dissimilarities of
606 metagenome samples based on RPMG values of arctic MAGs and reference genomes, including
607 TARA ocean-derived SAR11 MAGs. Scaling 2 is shown. Arrows indicate significant environmental
608 variables after Post-Hoc testing. Asterisks show the weighted average of each genomes'
609 frequency.

610

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

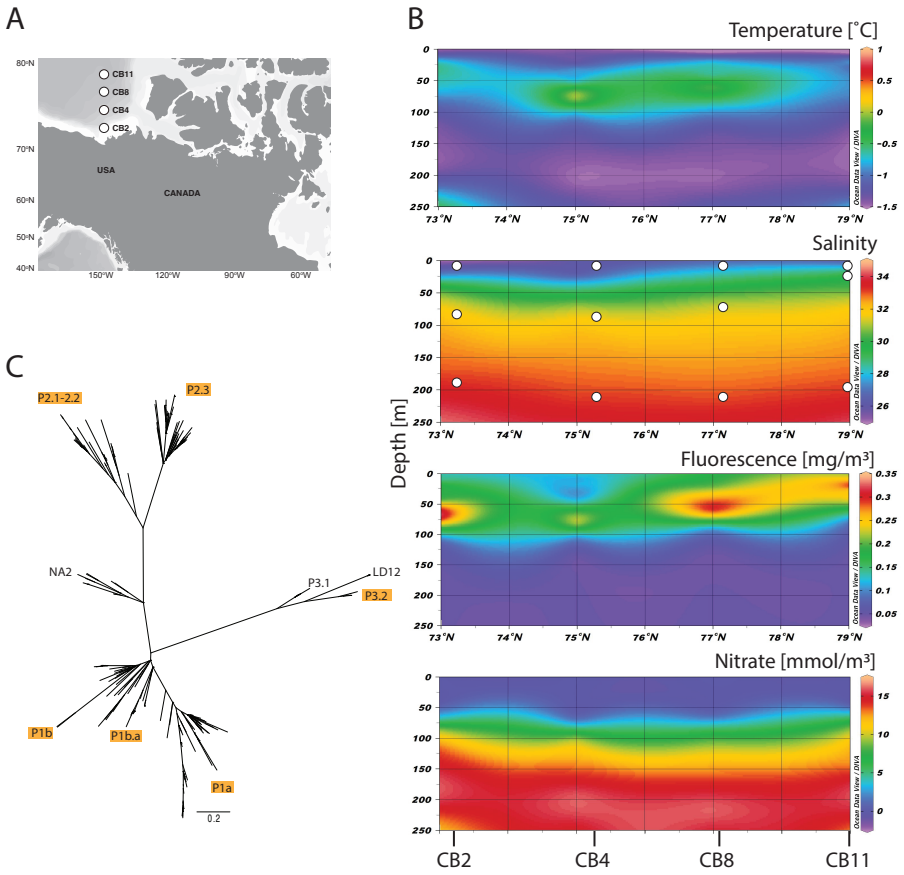
Clade (16S rRNA)	Subclade (16S rRNA)	Phylotype (ITS)	Reference genome available?	Environmental description
S1	S1a	P1a.1	Y	-Cold [2] -S1: Freshwater Lake Baikal [48]
		P1a.2	N	-Temperate [2]
		P1a.3	Y	-Tropical [2] -Tropical, depth generalist [5]
	S1b	P1b.1	Y	-P1b: Tropical [2] -P1b: Tropical, epi- and bathypelagic [49] -S1b: Tropical, oxygen generalist [8] -P1b.3: Tropical, upper euphotic layer [5]
		P1b.2	Y	
		P1b.3	Y	
	S1c	NA2	Y	-Deep, tropical [5, 8]
S2	S2a.a	-	Y	-OMZ specialist, tropical [8]
	S2a.b	P2.1	Y	-P2.1: Tropical [2]
		P2.2	N	-P2.2: Cold [2] -P2.2: Antarctic [9]
	S2b	P2.3	Y	-P2.3: Cold [2] -P2.3: Deep, tropical [5] -S2b: Tropical, OMZ [8] -P2.3s1: Arctic SCM and PWW (this study)
S3		P3.1	Y	-Coastal, mesohaline, surface [10] -Tropical [2]
		P3.2	Y	-Brakish [10] -Temperate [2] -Arctic [11]
		LD12	Y	-Freshwater [12]

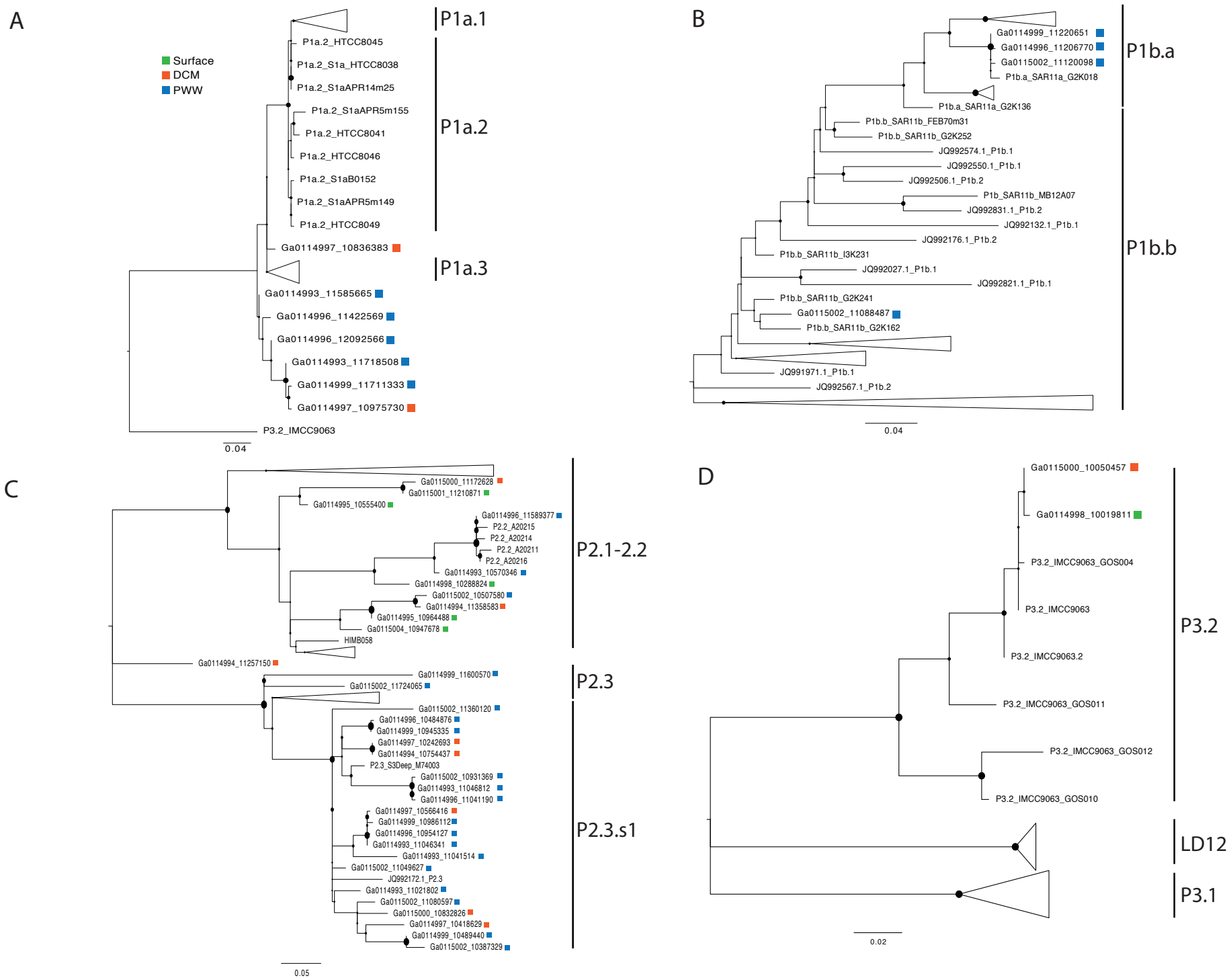
Table 2: Western Arctic Ocean metagenome sampling sites, their environmental feature and environmental parameters including sampling depth (Depth), water temperature (Temp.), Salinity, Chl a Fluorescence, CDOM, Oxygen and Nitrate (NO_3).

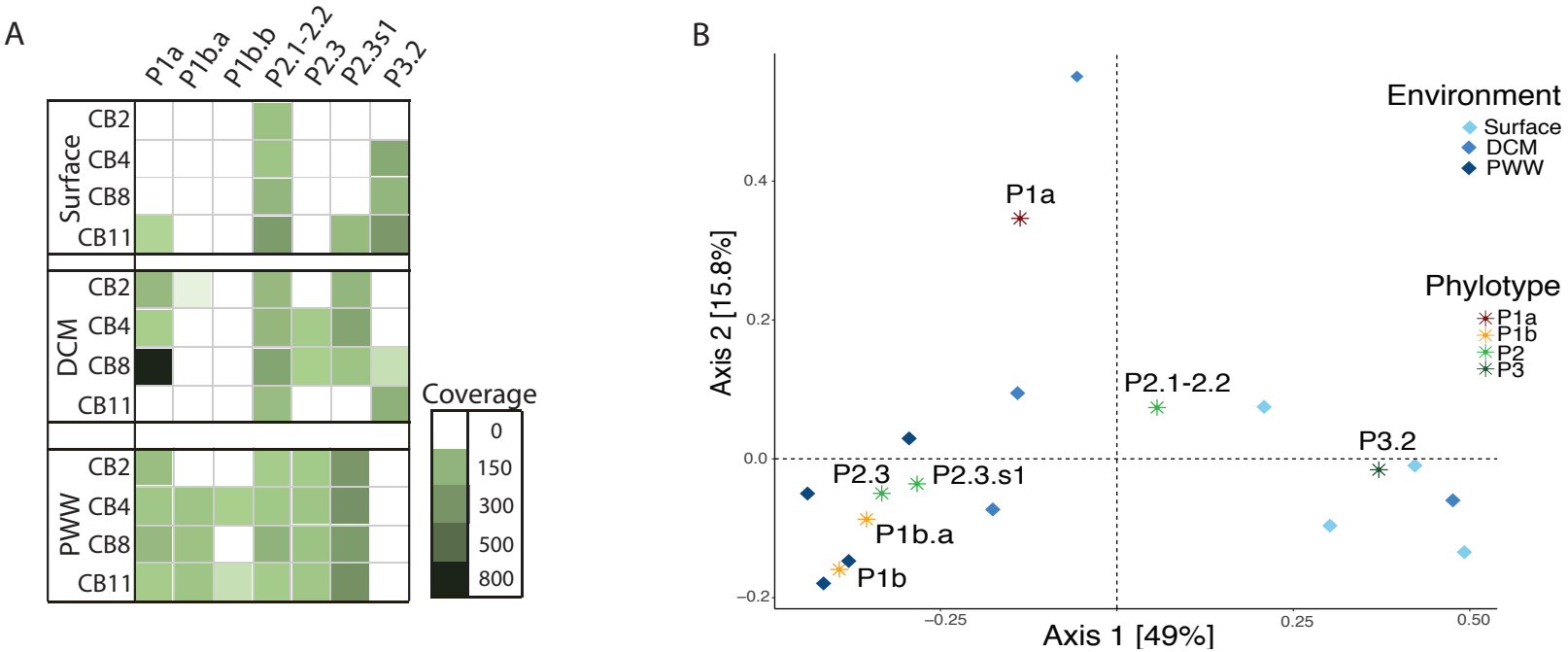
Station	Lat.	Long.	Feature	Depth [m]	Temp. [˚C]	Salinity	Fluorescence [mg m^{-3}]	CDOM [mg m^{-3}]	Oxygen [mmol m^{-3}]	NO_3 [mmol m^{-3}]
CB2_154	73.22	-150.22	SUR	7	-1.26	25.70	0.15	2.54	391.85	0
CB2_152	73.22	-150.22	DCM	67	-0.86	31.48	0.33	3.9	351.56	4.45
CB2_150	73.22	-150.22	PWW	177	-1.45	33.18	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	0
CB4_136	75.26	-150.07	DCM	79	-0.021	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	395.598	0
CB8_128	77.1	-150.23	DCM	58	-0.15	30.97	0.3	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	0
CB11_88	79.25	-150.06	DCM	25	-1.04	29.69	0.25	3.22	401.18	0
CB11_86	79.25	-150.06	PWW	190	-1.46	33.15	0.05	4.35	281.89	15.82

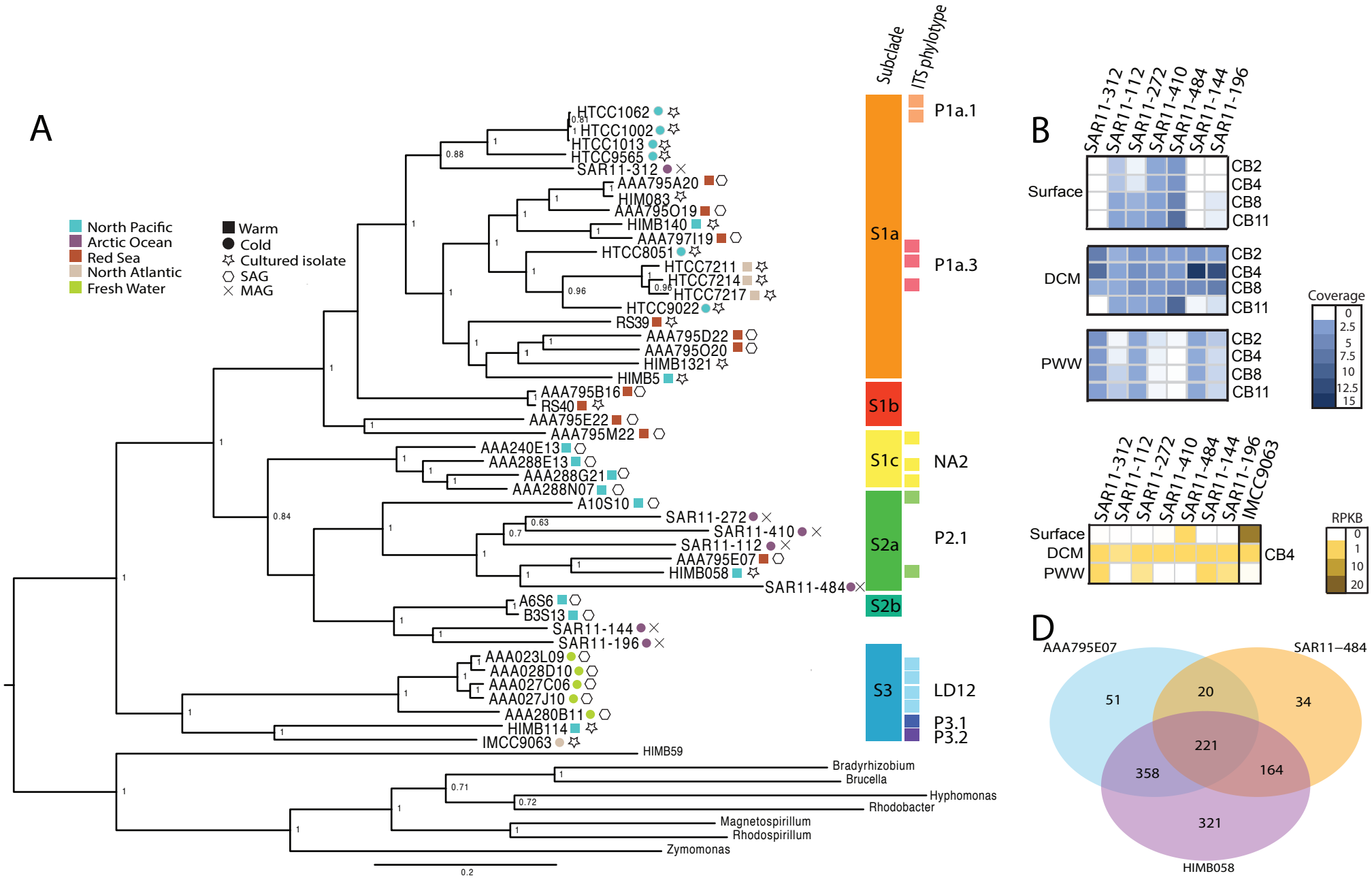
Table 3: Summary statistics of MAGs from the Western Arctic Ocean.

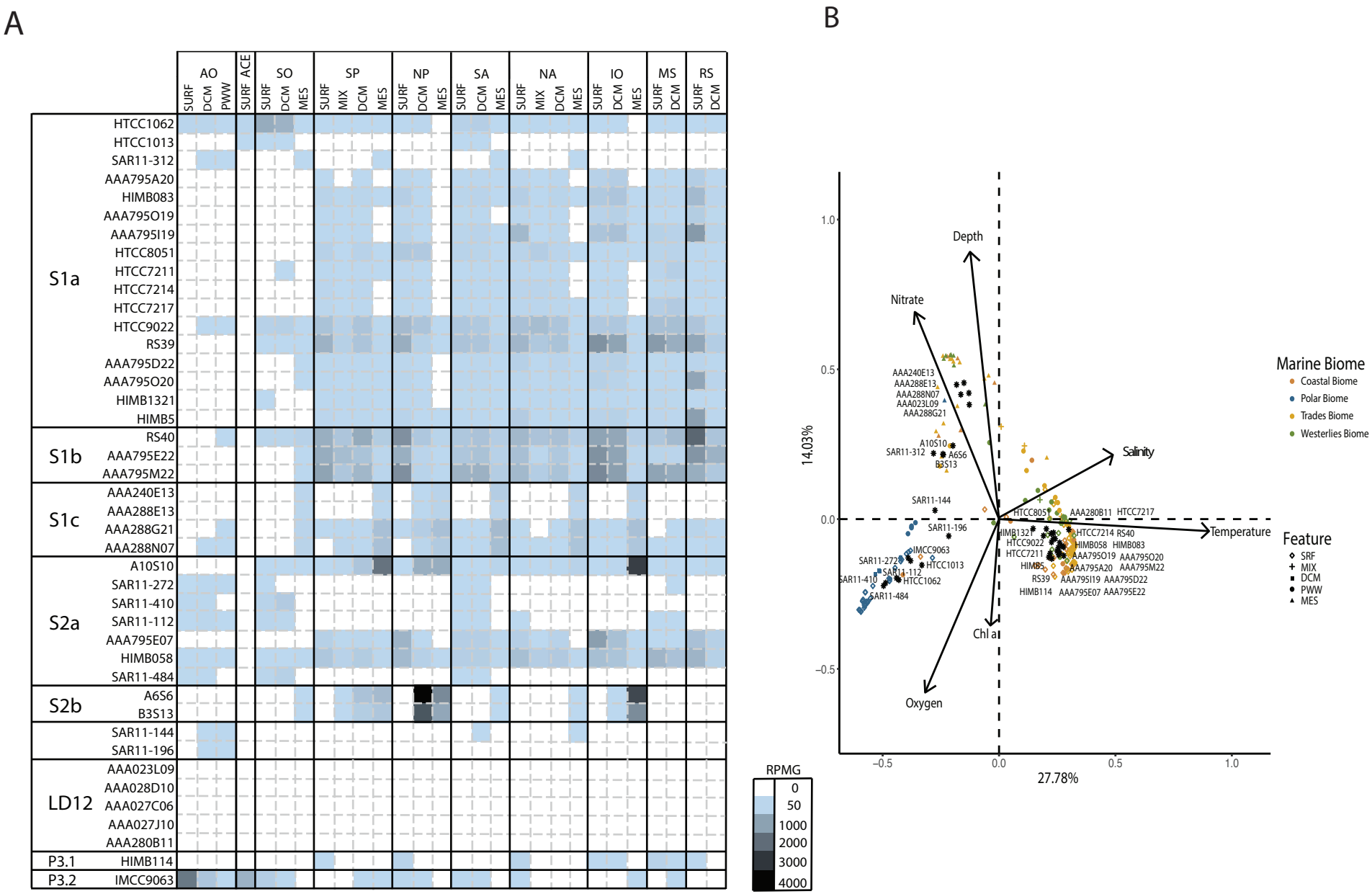
Subgroup	MAG	Size (mb)	Cov (x)	GC (%)	Completeness (%)	Contamination (%)	N50 (kb)	# of scaffolds	# of genes	# of unique genes
S1a	312	0.42	27.83	27.3	47.17	1.89	21.23	20	462	33
S2a	112	0.42	8.66	28.31	38.68	1.89	18.76	20	462	49
S2a	272	0.3	17.96	29.52	35.85	0	98.98	5	320	10
S2a	410	0.42	16.77	30.68	24.53	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
CB2
CB4
CB8
CB11
C
P2.1-2.2
P2.3
NA2
P3.1
LD12
P3.2
P1b
P1b.a
P1a
0.2


A
Surface
DCM
PWW
P1a.1
P1a.2
P1a.3
P1a.2_HTCC8045
P1a.2_S1a_HTCC8038
P1a.2_S1aAPR14m25
P1a.2_S1aAPR5m155
P1a.2_HTCC8041
P1a.2_HTCC8046
P1a.2_S1aB0152
P1a.2_S1aAPR5m149
P1a.2_HTCC8049
Ga0114997_10836383
Ga0114993_11585665
Ga0114996_11422569
Ga0114996_12092566
Ga0114993_11718508
Ga0114999_11711333
Ga0114997_10975730
P3.2_IMCC9063
0.04
B
P1b.a
P1b.b
Ga0114999_11220651
Ga0114996_11206770
Ga0115002_11120098
P1b.a_SAR11a_G2K018
P1b.a_SAR11a_G2K136
P1b.b_SAR11b_FEB70m31
P1b.b_SAR11b_G2K252
JQ992574.1_P1b.1
JQ992550.1_P1b.1
JQ992506.1_P1b.2
P1b_SAR11b_MB12A07
JQ992831.1_P1b.2
JQ992132.1_P1b.1
JQ992176.1_P1b.2
P1b.b_SAR11b_I3K231
JQ992027.1_P1b.1
JQ992821.1_P1b.1
P1b.b_SAR11b_G2K241
Ga0115002_11088487
P1b.b_SAR11b_G2K162
JQ991971.1_P1b.1
JQ992567.1_P1b.2
0.04
C
P2.1-2.2
P2.3
P2.3.s1
Ga0115000_11172628
Ga0115001_11210871
Ga0114995_10555400
Ga0114996_11589377
P2.2_A20215
P2.2_A20214
P2.2_A20211
P2.2_A20216
Ga0114993_10570346
Ga0114998_10288824
Ga0115002_10507580
Ga0114994_11358583
Ga0114995_10964488
Ga0115004_10947678
HIMB058
Ga0114994_11257150
Ga0114999_11600570
Ga0115002_11724065
Ga0115002_11360120
Ga0114996_10484876
Ga0114999_10945335
Ga0114997_10242693
Ga0114994_10754437
P2.3_S3Deep_M74003
Ga0115002_10931369
Ga0114993_11046812
Ga0114996_11041190
Ga0114997_10566416
Ga0114999_10986112
Ga0114996_10954127
Ga0114993_11046341
Ga0114993_11041514
Ga0115002_11049627
JQ992172.1_P2.3
Ga0114993_11021802
Ga0115002_11080597
Ga0115000_10832826
Ga0114997_10418629
Ga0114999_10489440
Ga0115002_10387329
0.05
D
P3.2
LD12
P3.1
Ga0115000_10050457
Ga0114998_10019811
P3.2_IMCC9063_GOS004
P3.2_IMCC9063
P3.2_IMCC9063.2
P3.2_IMCC9063_GOS011
P3.2_IMCC9063_GOS012
P3.2_IMCC9063_GOS010
0.02


A
P1a
P1b.a
P1b.b
P2.1-2.2
P2.3
P2.3s1
P3.2
Surface
DCM
PWW
CB2
CB4
CB8
CB11
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
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
IMCC9063
CB4
RPKB
0
1
10
20
D
AAA795E07
SAR11–484
HIMB058
51
20
34
221
358
164
321


A
AO
ACE
SO
SP
NP
SA
NA
IO
MS
RS
SURF
DCM
PWW
MIX
MES
S1a
S1b
S1c
S2a
S2b
LD12
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
AAA023L09
AAA028D10
AAA027C06
AAA027J10
AAA280B11
HIMB114
IMCC9063
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
Marine Biome
Coastal Biome
Polar Biome
Trades Biome
Westerlies Biome
Feature
SRF
MIX
DCM
PWW
MES

Supplemental information

Supplemental methods

Sampling and DNA extractions

Samples from 4 stations (CB2, CB4, CB8, and CB11) were analysed, with a total of 12 water samples collected from 10-LNiskin bottles mounted on a rosette system. Samples were from the surface layer (Surface: 5 to 7 m), the DCM (25 to 79 m) and the PWW layers (177 to 213 m). The rosette was equipped with profilers for conductivity, temperature and depth (Sea-Bird Electronics Inc., California, USA) coupled with sensors for dissolved oxygen (DO), coloured dissolved organic matter (CDOM) and chlorophyll *a* (Chl *a*) fluorescence. Sample depth was determined during the downward casts, to ensure that target water masses were sampled, with the DCM identified from the Chl *a* fluorescence and PWW centered on a salinity of 31. Nutrient samples and Chl *a* were collected directly from the Niskin bottles and analysed on board as previously described [1]. Consistency of the targeted and actual sampling depths were verified by bottle salinities analyzed onboard using a Guideline Portasal and IAPSO standard seawater [2], while dissolved oxygen (DO) was determined by Winkler-based titration (Kit B, Scripps Institute of Oceanography, USA). Water for planktonic nucleic acid (DNA) analysis was sequentially filtered through a 3-µm polycarbonate filter and 0.22 µm pore size Sterivex® filters (Millipore, Billerica, MA, USA). Filters were preserved in RNAlater for 20 to 40 min and then kept at -80˚C until laboratory processing. DNA was extracted from the Sterivex® filters (3 to 0.22 µm size fraction) following the protocol described previously and stored at -80˚C [3]. Physical oceanographic profiles were visualised using Ocean Data View [4].

Metagenomic sequencing and assembly

The 12 Arctic Ocean samples were sequenced at the Department of Energy Joint Genome Institute (Walnut Creek, CA, USA) using an Illumina HiSeq 2500-1TB platform as described in

Colatriano et al. [3]. Raw reads were assembled using MEGAHIT [5] and gene annotation was performed with the Integrated Microbial Genomes (IMG) database tool (version 4.11.0) [6].

Internal transcribed spacer (ITS) phylogenetic analysis

We used the complete ITS region from a randomly chosen SAR11 reference genome (HIMB114) to extract ITS regions from each of the twelve single sample metagenome assemblies using BLASTn [7]. All hits were filtered to a maximum e-value of 10e-35 and a minimum scaffold size of 350 bp. The retrieved sequences were aligned using the MUSCLE algorithm as implemented in MEGA v.7 [8] and poorly aligned sequences removed after visual inspection. Remaining sequences were clustered using CD-hit [9] at an identity of 100%. Arctic ITS sequences not covering both of the ITS-region tRNAs were removed from the alignment. To assign Arctic sequences to specific phylotypes, we analysed the Arctic ITS sequences with reference sequences from published genomes and ITS sequences used in two previous biogeographic studies [10, 11]. All sequences were realigned using MUSCLE as implemented in MEGA v.7 [8], and the resulting alignment curated in Mesquite [12]. Maximum likelihood trees using GTR+Gamma models of nucleotide substitution were calculated in MEGA v.7 with 100 bootstraps [8]. Subsequently, we assigned phylotype membership of the Arctic ITS sequences (and their CD-hit clusters) based on tree topology. ITS CD-hit clusters representing short ITS fragments were blasted against a custom ITS database of full-length Arctic ITS samples using BLASTn and assigned by their best hit [7].

We performed PCoA ordination of the Bray-Curtis dissimilarity of the coverage of each Arctic ITS phylotype within each sample, correcting for negative eigenvalues using the Caillez method [13]. Ordination was performed using the R packages phyloseq, ade4 and vegan [14–16] and visualized using the package ggplot2 [17]. *Post-hoc* tests were performed using the envdist function as implemented in vegan [16] with 999 permutations. All computational parallelization was performed with parallel [18].

Metagenomic binning

We conducted metagenomic binning using the Metabat2 pipeline [19] on a coassembly of all twelve samples [3]. Fastq reads of all Arctic samples were mapped to scaffolds larger than 10 kb using bbmap [20] with a minimum id of 100 to achieve a high degree of between-bin differentiation. 44 bins belonging to the SAR11 cluster were identified using the checkM lineage_wf function [21]. These bins were cleaned with the refineM filter_bins and taxon_filter options using a taxonomic reference genome database with added SAR11 reference genomes. Scaffolds with aberrant coverage across individual samples were removed after visual inspection.

Concatenated gene phylogeny

Single-copy orthologous genes were identified in seven Arctic SAR11 MAGs, in 41 SAR11 reference genomes representing the clade's phylogenetic diversity, as well as in eight outgroup genomes using ProteinOrtho [22]. We removed all loci present in less than five of the seven MAGs, resulting in 39 orthologs. We aligned their amino acid sequences using the MUSCLE algorithm as implemented in MEGA v.7 [8]. Aligned orthologs were concatenated in Mesquite [12] and positions were masked if they had ZORRO-calculated confidence scores <0.5 [23], resulting in a concatenated alignment of 12 516 amino acids. A maximum likelihood phylogeny was constructed using MEGA-cc with a JTT substitution model, a gamma-distribution with an invariant sites model of rate variation and nearest-neighbour interchange heuristic search method, and bootstrapped 100 times [24].

Arctic SAR11 transcriptome

RNA samples were collected during a cruise in September 2017 and stored as described previously for DNA samples. To extract the RNA, RNAlater was expelled and the filters inundated with 1.5mL Lysis buffer from the mirVana RNA isolation kit (Invitrogen, California, USA), followed

by two minutes of vortexing. 150 µl of miRNA homogenate were added to the filter, followed by vortexing and ten minutes of incubation on ice. The lysate was then removed from the filter and RNA extracted using the phenol-chloroform extraction protocol as described in the mirVana RNA isolation standard protocol (Invitrogen, CA, USA).

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

We compared the genome content of the Arctic S2a MAG SAR11-484 to two closely related S2a reference genomes, HIMB058 and AAA795E07. Orthologous genes were identified across the three genomes using ProteinOrtho [22]. We investigated the selective pressures acting on these shared genes by performing codon-based nucleotide alignments with MUSCLE as implemented in MEGA v.7 [8]. Pairwise dN/dS ratios of each gene were calculated using the Young and Nielson method in yn00 as implemented in paml [25].

Fragment recruitment and SAR11 biogeography.

Seven Arctic MAGs, 48 representative SAR11 reference genomes, and 38 TARA ocean MAGs previously described as SAR11 [26] were cleaned of rRNA sequences, blasted against the metagenomic datasets using blastall (e-value = 0.00001), and all hits recorded. Subsequently, all recruited reads were extracted from the metagenomic datasets and used in a reciprocal best-hit blast against a database of SAR11 MAGs and reference genomes. 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 used to calculate the number of reads recruited per megabase genome per gigabase metagenome (RPMG).

PCoA ordination of the Bray-Curtis dissimilarity of the RPMG matrix was conducted using the R packages; vegan, ade4 and phyloseq [14–16]. Negative eigenvalues were corrected using

the Caillez method [13]. The depth, temperature, salinity, nitrate and Chl *a* concentrations for each sample were projected onto the ordination.

1. Onda DFL, Medrinal E, Comeau AM, Thaler M, Babin M, Lovejoy C. Seasonal and Interannual Changes in Ciliate and Dinoflagellate Species Assemblages in the Arctic Ocean (Amundsen Gulf, Beaufort Sea, Canada). *Front Mar Sci* 2017.
2. 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.
3. 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 Chloroflexi bacteria. *Commun Biol* 2018.
4. Schlitzer R. Ocean Data View.
5. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015.
6. Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, et al. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand Genomic Sci* 2016.
7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990.
8. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* 2016.

9. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012.

10. 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.

11. 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.

12. Maddison W, Maddison D. Mequite: a modular system for evolutionary analysis. Version 3.04. *http://mesquiteproject.org* 2015.

13. Legendre P, Andersson MJ. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecol Monogr* 1999.

14. McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 2013.

15. Dray S, Dufour A-B. The **ade4** Package: Implementing the Duality Diagram for Ecologists. *J Stat Softw* 2007.

16. 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.

17. Wickham H. Elegant Graphics for Data Analysis. *Media* 2015.

18. Tang D, Gupta G. An efficient parallel dynamic programming algorithm. *Comput Math with Appl* 1995.

19. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately

reconstructing single genomes from complex microbial communities. *PeerJ* 2015.

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

21. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015.

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

23. Wu M, Chatterji S, Eisen JA. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 2012.

24. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 2012.

25. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007.

26. Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Lucker S, et al. Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria Are Abundant In The Surface Ocean. *doi.org* 2017.