



PRIFYSGOL  
**BANGOR**  
UNIVERSITY

## Red Sea SAR11 and Prochlorococcus Single-cell Genomes Reflect Globally Distributed Pangenomes

Thompson, Luke R.; Haroon, Mohamed F.; Shibl, Ahmed A.; Cahill, Matt J.; Ngugi, David K.; Williams, Gareth J.; Morton, James T.; Knight, Rob; Goodwin, Kelly D.; Stingl, Ulrich

### Applied and Environmental Microbiology

DOI:

[10.1128/AEM.00369-19](https://doi.org/10.1128/AEM.00369-19)

Published: 01/07/2019

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

*Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):*

Thompson, L. R., Haroon, M. F., Shibl, A. A., Cahill, M. J., Ngugi, D. K., Williams, G. J., Morton, J. T., Knight, R., Goodwin, K. D., & Stingl, U. (2019). Red Sea SAR11 and Prochlorococcus Single-cell Genomes Reflect Globally Distributed Pangenomes. *Applied and Environmental Microbiology*, 85(13), [e00369-19]. <https://doi.org/10.1128/AEM.00369-19>

#### Hawliau Cyffredinol / General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

#### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

1 **Red Sea SAR11 and *Prochlorococcus* Single-cell Genomes Reflect**  
2 **Globally Distributed Pangenomes**

3 Luke R. Thompson<sup>1,2,3</sup>, Mohamed F. Haroon<sup>1</sup>, Ahmed A. Shibl<sup>1</sup>, Matt J. Cahill<sup>1</sup>, David K.  
4 Ngugi<sup>1</sup>, Gareth J. Williams<sup>4</sup>, James T. Morton<sup>5,6</sup>, Rob Knight<sup>5,6,7</sup>, Kelly D. Goodwin<sup>3</sup>, Ulrich  
5 Stingl<sup>1,8</sup>

6 1. Red Sea Research Center, King Abdullah University of Science and Technology, Thuwal,  
7 Saudi Arabia

8 2. Department of Biological Sciences and Northern Gulf Institute, University of Southern  
9 Mississippi, Hattiesburg, MS, United States of America

10 3. Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological  
11 Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest  
12 Fisheries Science Center, La Jolla, CA, United States of America

13 4. School of Ocean Sciences, Bangor University, Anglesey, United Kingdom

14 5. Department of Pediatrics, University of California San Diego, La Jolla, CA, United States  
15 of America

16 6. Department of Computer Science and Engineering, University of California San Diego, La  
17 Jolla, CA, United States of America

18 7. Center for Microbiome Innovation, University of California, San Diego, CA, United States  
19 of America

20 8. Department of Microbiology & Cell Science, Fort Lauderdale Research and Education  
21 Center, UF/Institute of Food and Agricultural Sciences, University of FL, Davie, Florida,  
22 United States of America

23 Correspondence: lukethompson@gmail.com, ulistingl@gmail.com

24 Keywords: metagenomics, *Pelagibacter*, population genomics, SAG, single-cell genomics

25 Short title: Single-cell Genomes of Red Sea Phytoplankton

## 26 **Abstract**

27 Evidence suggests many marine bacteria are cosmopolitan, with widespread but sparse strains  
28 poised to seed abundant populations upon conducive growth conditions. However, studies  
29 supporting this “microbial seed bank” hypothesis have analyzed taxonomic marker genes rather  
30 than whole genomes/metagenomes, leaving open the possibility that disparate ocean regions  
31 harbor endemic gene content. The Red Sea is isolated geographically from the rest of the ocean  
32 and has a combination of high irradiance, high temperature, and high salinity that is unique  
33 among the ocean; we therefore asked whether it harbors endemic gene content. We sequenced  
34 and assembled single-cell genomes of 21 SAR11 (subclades Ia, Ib, Id, II) and 5 *Prochlorococcus*  
35 (ecotype HLII) cells from the Red Sea and combined them with globally-sourced reference  
36 genomes to cluster genes into ortholog groups (OGs). Ordination of OG composition could  
37 distinguish clades, including phylogenetically cryptic *Prochlorococcus* ecotypes LLII and LLIII.  
38 Compared with reference genomes, 1% of *Prochlorococcus* and 17% of SAR11 OGs were  
39 unique to the Red Sea genomes (RS-OGs). Most (83%) RS-OGs had no annotated function, but  
40 65% of RS-OGs were expressed in diel Red Sea metatranscriptomes, suggesting they could be  
41 functional. Searching *Tara* Oceans metagenomes, RS-OGs were as likely to be found as non-RS-  
42 OGs; nevertheless, Red Sea and other warm samples could be distinguished from cooler samples  
43 using the relative abundances of OGs. The results suggest that the prevalence of OGs in these  
44 surface ocean bacteria is largely cosmopolitan, with differences in population metagenomes  
45 manifested by differences in relative abundance rather than complete presence–absence of OGs.

## 46 **Importance**

47 Studies have shown that as we sequence seawater from a selected environment deeper and  
48 deeper, we approach finding every bacterial taxon known for the ocean as a whole. However,  
49 such studies have focused on taxonomic marker genes rather than on whole genomes, raising the  
50 possibility that the lack of endemism results from the method of investigation. We took a  
51 geographically isolated water body, the Red Sea, and sequenced single cells from it. We  
52 compared those single-cell genomes to available genomes from around the ocean, and to ocean-  
53 spanning metagenomes. We showed that gene ortholog groups found in Red Sea genomes but  
54 not in other genomes are nevertheless common across global ocean metagenomes. These results

55 suggest that Baas Becking's hypothesis "everything is everywhere, but the environment selects"  
56 also applies to gene ortholog groups. This widely dispersed functional diversity may give  
57 oceanic microbial communities the functional capacity to respond rapidly to changing  
58 conditions.

## 59 **Introduction**

60 Marine bacteria thrive throughout the surface ocean despite low nutrients, high irradiation, and  
61 other physicochemical stressors. Adaptations enabling survival can be at the level of  
62 transcriptional, translational, and other methods of cellular regulation that occur at time-scales of  
63 minutes to hours (1, 2). Alternatively, microbial genomes can evolve new functions on the scale  
64 of thousands to millions of generations (3, 4). Evolution via horizontal gene transfer enables the  
65 introduction of entirely new functionality (gene gain) as well as genome streamlining (gene loss)  
66 for more efficient resource (e.g., nitrogen, phosphorus) allocation (5). Therefore, it is expected  
67 that the genomes of marine bacteria will display differences in gene content correlated with the  
68 physicochemical environment in which they live. Indeed, both individual genomes (cultured and  
69 single-cell genomes) (6–10) and community genomes (metagenomes) (11, 12) show that bacteria  
70 in the oligotrophic (nutrient-poor) surface ocean carry streamlined genomes finely tuned to their  
71 environments.

72 Examples of adaptive gene presence–absence patterns are seen in the most numerous groups of  
73 bacteria in the oligotrophic tropical and sub-tropical surface ocean, the photoautotrophic  
74 picocyanobacteria *Prochlorococcus* and *Synechococcus* and the chemoheterotrophic  
75 Alphaproteobacteria SAR11 clade (*Candidatus Pelagibacter ubique*). Genomes of these genera  
76 are smaller than their relatives in less nutrient-poor environments (6, 8), suggestive of genome  
77 streamlining to conserve resources used for genome replication (3). Consistent with genome  
78 streamlining, the genes maintained in *Prochlorococcus* and SAR11 genomes are correlated with  
79 physical features in parts of the water column in which they are found, for example, genes for  
80 acquisition of nitrite and nitrate in genomes found where those compounds are available (3, 8).  
81 Examples revealed through comparative community genomics include an enrichment of  
82 phosphorus acquisition gene ortholog groups in the Atlantic relative to the Pacific Ocean (11, 13)

83 and an enrichment in osmolyte oxidation gene ortholog groups in the Mediterranean and Red  
84 Seas relative to the Atlantic and Pacific Oceans (12).

85 The Red Sea is an attractive environment for the study of genomic adaptations. Geographically,  
86 the Red Sea is largely isolated from the rest of the World Ocean, with only a small sill (the Bab  
87 el Mandeb) connecting it to the Indian Ocean (14). Among surface waters catalogued in the  
88 World Ocean Database, the Red Sea lies at the high end of the global temperature distribution  
89 and is higher than any other sea in the global salinity distribution (Fig. S1). The Red Sea,  
90 straddling the Tropic of Cancer, experiences year-round high irradiance, and cloud cover across  
91 North Africa and the Arabian Peninsula is among the lowest on the planet (NASA Aqua satellite  
92 MODIS sensor). The Red Sea is also oligotrophic, with production thought to be limited by  
93 nitrogen (15).

94 Evidence of genomic adaptation to high light and high salinity in the Red Sea has been revealed  
95 through comparative metagenomics, showing increased relative abundance of known gene  
96 ortholog groups in *Prochlorococcus* and SAR11 (12). Relative to the North Pacific, Sargasso  
97 Sea, and Mediterranean Sea, the Red Sea *Prochlorococcus* population had increased frequencies  
98 of high-light stress and DNA repair gene ortholog groups (12), the latter likely an adaptation to  
99 UV-induced DNA damage. Relative to these same seas, the SAR11 population had increased  
100 frequencies of gene ortholog groups for osmolyte degradation (12); osmolytes are important  
101 molecules for surviving high salinity in many organisms. Across 45 metagenomes along  
102 latitudinal and depth gradients from the surface to 500 m in the Red Sea, temperature explained  
103 more variation in gene ortholog groups than any other environmental parameter, and the relative  
104 abundance of gene ortholog groups linked to high irradiance, high salinity, and low nutrients  
105 were correlated with those parameters (16).

106 The above-mentioned patterns observed in comparative metagenomics studies were all based on  
107 relative abundance of known gene ortholog groups, dependent on a reference genome database  
108 with no representatives from the Red Sea. Therefore, the question remains if there are gene  
109 functions in the *Prochlorococcus* and SAR11 populations in the Red Sea not found in any other  
110 *Prochlorococcus* and SAR11 populations in the ocean. Because of its relative geographic  
111 isolation, we might expect the Red Sea to be genetically isolated, with endemic genomic  
112 adaptations to its unique combination of high solar irradiance, high temperature, high salinity,

113 and low nutrient levels. Newly identified gene ortholog groups could be informative for  
114 understanding microbial adaptation and mechanisms of stress tolerance, and have potential  
115 biotechnological applications.

116 The question of whether there are genetic functions found in only one sea of the global ocean  
117 speaks to theoretical questions of microbial biogeography as well. A prevailing idea in microbial  
118 ecology is that most microbial species are found at a given site provided the conditions are  
119 conducive for their growth. This is known as the Baas Becking hypothesis: “Everything is  
120 everywhere, but the environment selects” (17). Among microbial taxa found in seawater, there is  
121 growing evidence for a cosmopolitan distribution of these taxa throughout the global ocean (18,  
122 19). Support for the “microbial seed bank” hypothesis has come from deep sequencing of ocean  
123 samples, revealing for example that nearly all 16S rRNA operational taxonomic units (OTUs)  
124 from a deep-sea hydrothermal vent can be found in the open ocean (19), and that we approach  
125 identifying all OTUs in the ocean as sequencing effort increases for a single marine sample (18).  
126 Despite this evidence supporting a cosmopolitan distribution of OTUs throughout the ocean,  
127 these amplicon sequences (16S rRNA OTUs) are only taxonomic proxies and do not represent  
128 the extensive gene-level diversity in microbial genomes. Even if such marker gene sequences are  
129 omnipresent across the ocean, genome evolution and diversification, e.g., via horizontal gene  
130 transfer, could be occurring that generates gene-level adaptations that are endemic to particular  
131 locations. Are microbial gene ortholog groups, defined at the level of genus (SAR11 or  
132 *Prochlorococcus*), as widely distributed as microbial 16S rRNA gene sequences?

133 Here, to investigate microbial genomic diversity in SAR11 and *Prochlorococcus*, including  
134 possible endemic adaptation in Red Sea populations, we have sequenced single-cell amplified  
135 genomes (SAGs) from the Red Sea and compared their gene ortholog group (OG) content to  
136 genomes and metagenomes from around the World Ocean. We have quantified expression of  
137 OGs in metatranscriptomes from the Red Sea collected over two consecutive 24-hour day–night  
138 cycles. This effort has resulted in 21 SAR11 SAGs, including the first genomes from subclades  
139 Ib and Id, and 5 *Prochlorococcus* SAGs. Using these Red Sea SAGs and the OGs they contain as  
140 queries for genomic and metagenomic analyses, we have analyzed globally-sourced genomes  
141 and metagenomes to investigate the extent to which OGs from surface-ocean *Prochlorococcus*  
142 and SAR11 are distributed across the World Ocean.

## 143 **Materials and Methods**

### 144 **Sample collection**

145 A single seawater sample (100 mL) was collected in a polycarbonate bottle from the surface  
146 (depth of 0 m) of an open-ocean site in the east-central Red Sea (19.75 °N, 40.05 °E), near the  
147 Farasan Banks region, on June 15, 2010. The sample was preserved with dimethyl sulfoxide (5%  
148 final concentration), flash frozen in liquid nitrogen, and stored at –80 °C.

149 Seawater samples for metatranscriptomics were taken March 3–5, 2013, from an open-ocean site  
150 in the Red Sea (Kebrit Deep, 24.7244 °N, 36.2785 °E). To obtain broad coverage of the water  
151 column by both time of day and water depth, one sample per depth was collected every 4 h over  
152 a 48-h period at four depths: surface (10 m), below the mixed layer (40 m; bottom of mixed layer  
153 was 35 m), chlorophyll maximum (75 m), and oxygen minimum zone (420 m). For each  
154 timepoint and depth, 1 L seawater was filtered using a peristaltic pump with two in-line filters in  
155 series: a 1.6- $\mu$ m GF/A pre-filter (Whatman), then a 0.22- $\mu$ m Sterivex filter (Millipore).  
156 RNAlater (QIAGEN) was added immediately to fill the dead space of the Sterivex filter, which  
157 was then flash frozen in liquid nitrogen and stored at –80 °C.

### 158 **Nucleic acid extraction and amplification**

159 Single bacterioplankton cells in the preserved samples were flow-sorted, whole-genome  
160 amplified (MDA, multiple displacement amplification), and PCR-screened at the Bigelow  
161 Laboratory Single Cell Genomics Center (SCGC, Boothbay Harbor, ME, USA), following  
162 previously described protocols (20), with SYTO-13 nucleic acid stain used to stain cells for  
163 flow-sorting. SAG identification was carried out with SCGC protocol S-102 for bacteria using  
164 16S rRNA primers 27F and 907R (21, 22). A total of 21 and 5 cells were identified from 16S  
165 PCR screening and subjected to a second round of MDA before sequencing. The 16S rRNA gene  
166 sequences are available from the European Nucleotide Archive with accession numbers  
167 LN850141–LN850161.

168 The RNA extraction protocol for metatranscriptomics was adapted from (23–25). After expelling  
169 RNAlater from the Sterivex filter, 2 mL lysozyme solution (1 mg/mL in lysis buffer: 40 mM  
170 EDTA, 50 mM Tris pH 8.3, 0.73 M sucrose) was added, then filter incubated at 37 °C with



171 rotation for 45 min. Proteinase K solution (50  $\mu$ L at 20 mg/mL, QIAGEN/5PRIME) and SDS  
172 solution (100  $\mu$ L at 20%) were added, then filter incubated at 55 °C with rotation for 2 h. Lysate  
173 was expelled to a separate tube; meanwhile, 1 mL lysis buffer was added to the filter to wash at  
174 55 °C for 15 min. The two lysates were pooled, to which was added 1.5 mL absolute ethanol.  
175 RNA was then extracted from this solution using the RNeasy Protect Bacteria Midi Kit  
176 (QIAGEN). RNA was eluted with two volumes of RNase-free water. RNA sample was  
177 concentrated using a speed vacuum, from 250  $\mu$ L to 60  $\mu$ L. To this volume we added DNase (1  
178  $\mu$ L Ambion TURBO DNA-free, 6  $\mu$ L 10x buffer, 60  $\mu$ L RNA) and incubated at 37 °C for 30  
179 min. This solution was purified using the RNeasy MinElute Cleanup Kit (QIAGEN) and eluted  
180 with RNase-free water. Final yield was 1–2 ng total RNA. Total RNA was amplified using the  
181 C&E Version ExpressArt Bacterial mRNA Amplification Nano Kit, which preferentially  
182 amplifies mRNA (independent of poly-A tail) and selects against rRNAs. A single round of  
183 amplification was performed on 2–4 ng of total RNA which yielded about 10  $\mu$ g final amplified  
184 RNA.

### 185 **Nucleic acid sequencing**

186 For single-cell genome sequencing, genomic library preparation with Illumina TruSeq and  
187 sequencing with Illumina GAIIX and Illumina HiSeq 2000 was done at the KAUST Bioscience  
188 Core Laboratory, generating paired 105-bp reads. The assembled contigs (assembly methods  
189 below) are available from NCBI with accession numbers PRJEB9287 (BioProject) and  
190 SAMEA3368552–SAMEA3368577 (BioSample), and can also be visualized in Integrated  
191 Microbial Genomes system (26) under accession numbers 2630968236, 2630968238–  
192 2630968254, 2630968277–2630968281, and 2630968285–2630968287.

193 For metatranscriptomics, sequence data were processed as described in (27). Amplified RNA  
194 was used to construct sequencing libraries using the TruSeq Stranded RNA LT Sample Prep Kit  
195 (Illumina) according to the manufacturer's protocol. Libraries were paired-end sequenced with  
196 the Illumina HiSeq 2000 platform ( $2 \times 100$  bp). Raw RNA sequences have been deposited in  
197 NCBI GenBank with Bioproject number PRJNA289956. Low-quality reads and sequencing  
198 adapters were removed using Trimmomatic v0.32 (28). Sequence reads shorter than 50 bp were  
199 discarded. Bowtie 2 v2.2.4 (29) was used to identify and remove PhiX contamination sequences.  
200 The remaining sequences were error-corrected using the BayesHammer algorithm (30)



201 implemented in the SPAdes v3.5.0 (31), followed by removal of putative ribosomal RNA  
202 (rRNA) gene transcripts with SortMeRNA v2.0 (32).

### 203 **Genome assembly and annotation**

204 De novo assemblies were generated using CLC Genomics Workbench 4.9. The genomes were  
205 assembled independently and, unless otherwise specified, the following applies to all of the  
206 SAGs. The reads were first imported and quality trimmed with a limit of 0.01. They were then  
207 assembled using CLC's *de novo* assembler with a word size (*k*-mer) of 64 and with the min/max  
208 of the insert size set to 100/1000 bp. Only those contigs greater than 200 bp in length were  
209 included in downstream analyses. The reads were mapped to the consensus sequence of the  
210 assembled contigs using CLC's default parameters but with the length fraction set to 1.0 and the  
211 similarity set to 0.95.

212 Assembled SAG contigs were ordered and oriented relative to SAR11 HTCC1062  
213 (NC\_007205.1) or *Prochlorococcus* MIT 9202 (NZ\_DS999537) using ABACAS 1.3.1 (33). The  
214 ordered sequences were then imported into GAP4 (34) and additional joins were made between  
215 overlapping contigs if conserved synteny supported the arrangement. To identify and remove  
216 possible contaminating sequences from the assemblies, each contig was retained only if it met  
217 one or both of the following criteria: (i) the contig was binned into a bin annotated as SAR11 or  
218 *Prochlorococcus* using Metawatt 3.5 (35), using the "medium" bin level, with a minimum bin  
219 size of 50 kbp and minimum contig size of 500 bp; (ii) the contig had a top-10 BLASTN hit  
220 against GenBank nt, with E-value  $<1e-5$ , to SAR11 or *Prochlorococcus*.

221 Prediction of gene open reading frames (ORFs) and functional annotation of SAGs was  
222 performed by the RAST web service (36) with FIGfam Release 59.

### 223 **Ortholog group clustering**

224 Predicted proteins from SAGs were clustered with proteins from published cultured and SAG  
225 genomes (supplemental file 1) into ortholog groups (OGs) using OrthoMCL 2.0 (37). OrthoMCL  
226 configuration settings were as follows: percentMatchCutoff=50, evaluateExponentCutoff=-5. This  
227 yielded 5272 SAR11 OGs and 10439 *Prochlorococcus* OGs. After OrthoMCL clustering, OGs  
228 were assigned as core and non-core based on copy number in the non-Red Sea, cultured (non-

229 SAG) genomes: core OGs are those found at least once in each of the non-Red Sea, cultured  
230 genomes, and non-core OGs are those not found in at least one of the non-Red Sea, cultured  
231 genomes. Among SAR11, there were 683 core OGs and 4589 non-core OGs. Among  
232 *Prochlorococcus*, there were 1152 core OGs and 9287 non-core OGs. Protein sequence  
233 identifiers and FASTA sequences for each OG have been archived at <https://zenodo.org> with  
234 DOI 10.5281/zenodo.2634561.

### 235 **Estimation of genome completeness**

236 Completeness of SAGs was assessed using two methods. First, completeness was assessed using  
237 single-copy ‘core’ OGs, i.e., those OGs found once and only once in each complete genome  
238 based on the OrthoMCL clusters (analyzed separately for SAR11 and *Prochlorococcus*).  
239 Completeness was calculated as the number of core orthologs present in each SAG out of 649  
240 SAR11 or 1144 *Prochlorococcus* single-copy core OGs. Second, genome completeness of the  
241 SAGs was assessed using CheckM 1.0.13 (38) using the lineage-specific workflow (lineage\_wf)  
242 with database file `checkm_data_2015_01_16.tar.gz` downloaded from  
243 [https://data.ace.uq.edu.au/public/CheckM\\_databases](https://data.ace.uq.edu.au/public/CheckM_databases); CheckM was also used to estimate genome  
244 redundancy (called “contamination” in CheckM). For comparison, CheckM completeness and  
245 redundancy were calculated for the reference genomes used in this study (Table S1).

### 246 **Genome taxonomy and phylogenetics**

247 A total of 89 SAR11 and 96 *Prochlorococcus* shared single-copy orthologous genes were  
248 identified using the GET\_HOMOLOGUES software (39). Amino acid sequences translated from  
249 gene sequences were aligned using the MAFFT software (40). These alignments were  
250 concatenated, sites with gaps were deleted, and the concatenated data were partitioned using the  
251 PartitionFinder software (41) to account for variations of evolutionary processes among gene  
252 families. With the Bayesian information criterion (BIC) statistic, a 16-partition framework was  
253 chosen to optimally describe the variability, in which the LG rate matrix with Gamma  
254 distribution of rate variation (LG+G) was selected for 15 partitions and the VT rate matrix with  
255 Gamma distribution of rate variation (VT+G) was selected for the remaining partition. This  
256 partition model was used in the maximum-likelihood phylogenomic construction using the  
257 RAxML software (42).

258 **Ordination of SAGs and genomes using *k*-mer composition and ortholog**  
259 **composition**

260 SAGs and reference genomes (Table S1) were analyzed using principal components analysis  
261 (PCA) of nucleotide composition and OG composition. Nucleotide composition of the SAGs and  
262 reference genomes (SAR11 and *Prochlorococcus* scaffolds >200 kbp from Integrated Microbial  
263 Genomes, <https://img.jgi.doe.gov>) was determined as 6-nucleotide words or *k*-mers (6-mers). *k*-  
264 mer frequencies were calculated using Jellyfish 2.2.5; the main command used was jellyfish  
265 count -m 6 -t 8 -s 1M. This resulted in a table of 6-mer frequencies in the SAGs and genomes,  
266 one table each for SAR11 and *Prochlorococcus*. OG composition was derived from tables of  
267 OrthoMCL clusters, which—as the SAGs had variable levels of completeness and gene counts  
268 (Table 1)—were subsampled so that all genomes had the same number of gene counts in the  
269 table. The number of OG counts subsampled was chosen to balance the number of OG counts  
270 with the number of genomes retained (less complete SAGs were excluded): the OG composition  
271 tables (with counts of 5272 unique SAR11 OGs and 10439 unique *Prochlorococcus* OGs) were  
272 subsampled down to 800 gene counts per SAR11 SAG (keeping 12 of 21 SAGs) and 1400 gene  
273 counts per *Prochlorococcus* genome (keeping 5 of 5 SAGs). Prior to PCA, a pseudo-count of 1  
274 was added to *k*-mer and OG count tables to account for zero values; *k*-mer counts were then  
275 converted to relative abundances for each genome (unnecessary for OG counts because of the  
276 subsampling procedure); *k*-mer relative abundances were then standardized to *z*-scores (not done  
277 for OG counts because this reduced the resolving power of PCA). PCA was then performed  
278 using the Scikit-Learn function `sklearn.decomposition.PCA` (43).

279 **Mapping of metatranscriptomic reads to OGs**

280 The quality-filtered mRNA reads from the 52 samples were mapped against the SAGs using  
281 Bowtie 2 (29) with default settings. Each read mapping above the threshold was assigned to  
282 exactly one gene in a SAG contig. The resultant read counts were normalized based on the  
283 FPKM metric (fragments per kilobase of gene per million mapped reads). Per-sample FPKM  
284 counts for each gene were then summed by OGs, resulting in per-sample FPKM counts for each  
285 OG. For downstream analysis, counts were converted to a simple presence–absence measure: if

286 any gene belonging to the OG had one or more mapped transcript, that OG was marked as  
287 present in that sample.

### 288 **Detection and rarefaction analysis of OGs in *Tara* Oceans metagenomes**

289 A set of 139 prokaryote-enriched *Tara* Oceans metagenomic gene files (44) was downloaded  
290 from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>, ERZ096909–ERZ097150).  
291 Each file contains nucleotide sequences for genes predicted on *Tara* Oceans metagenomic  
292 contigs that were assembled from shotgun sequencing reads from individual *Tara* Oceans  
293 samples. The prokaryote fraction was 0.22–1.6  $\mu\text{m}$  for stations 004–052 and 0.22–3  $\mu\text{m}$  for  
294 stations 056–152; the environmental features of the samples were indicated as “SRF” (surface),  
295 “MIX” (mixed layer), “DCM” (deep chlorophyll maximum), and “MES” (mesopelagic zone).  
296 The metagenomic gene sequences were queried against a database of translated proteins from the  
297 SAGs and genomes with DIAMOND 0.8.26 (45) using the program blastx with parameters  $-p$  40  
298  $-k$  25  $-e$   $1e-3$ . The top hit (SAG or genome protein sequence) for each *Tara* gene sequence (E-  
299 value  $< 1e-5$ ) was retained. E-value cutoffs of  $1e-10$  and  $1e-15$  were also tested, which showed  
300 the same trends as E-value  $< 1e-5$  but with fewer total OGs identified. Counts of the number of  
301 times each protein was a top hit were then summed across each OG. This resulted in a table of  
302 OGs by samples where each OG was either present (at least one constituent protein was a top hit  
303 at least once) or absent in each sample. These presence–absence tables (one for SAR11, one for  
304 *Prochlorococcus*) were used to generate rarefaction curves: samples were added one-by-one  
305 randomly (1000 permutations), and the cumulative number of OGs found was recorded.

### 306 **Ordination of *Tara* Oceans metagenomes by OG composition**

307 OG counts (total, not presence–absence) in *Tara* Oceans surface (SRF) sample metagenomes  
308 were used for ordination using PCA. Prior to PCA, a pseudo-count of 1 was added to OG count  
309 tables to account for zero values; counts were then converted to relative abundances for each  
310 metagenome; OGs with an average relative abundance across all metagenomes less than 0.0001  
311 (0.01%) were removed; relative abundances were then standardized to z-scores. PCA was then  
312 performed using the Scikit-Learn function `sklearn.decomposition.PCA` (43).

## 313 **World Ocean temperature and salinity data**

314 Surface temperature and salinity data (WOD13\_ALL\_SUR\_OBS) from the World Ocean  
315 Database 2013 (<https://www.nodc.noaa.gov/OC5/WOD13/>) were downloaded from the Research  
316 Data Archive at the National Center for Atmospheric Research  
317 (<https://rda.ucar.edu/datasets/ds285.0/>).

## 318 **Results and Discussion**

### 319 **Single-cell genome properties and taxonomic classification**

320 Following collection of surface seawater from the east-central Red Sea, flow sorting, and  
321 amplification, we sequenced and assembled 21 SAR11 and 5 *Prochlorococcus* single-cell  
322 amplified genomes (SAGs). These SAGs represent reference genomes in an ocean region with  
323 sparse coverage: only one cultured *Prochlorococcus* genome (27) and two cultured SAR11  
324 genomes (46) are currently available from the Red Sea. The SAR11 SAGs also represent  
325 genomes from clades without other sequenced representatives: two SAGs from subclade Ib and  
326 three SAGs from subclade IId (Fig. 1).

327 To account for and remove any possible contaminating DNA sequences, assembled contigs were  
328 retained only if they were part of a SAR11 or *Prochlorococcus* Metawatt bin or if they had a top-  
329 10 BLASTN hit to a *Prochlorococcus* or SAR11 genome (methods). In Metawatt, assignment to  
330 bins is based on tetranucleotide frequency, and the average taxonomy of the bin is determined by  
331 BLAST of 500-bp fragments of all the contigs against a prokaryotic database (35). A contig  
332 matching the tetranucleotide frequency of a SAR11 or *Prochlorococcus* bin could be retained  
333 even if it contained contradictory or missing taxonomic information. However, to check if our  
334 secondary, BLASTN-based assignment process could be biased against short contigs, which  
335 might lack a neighboring anchor gene, we analyzed the distribution of contig lengths between  
336 retained and removed contigs for each SAG. We found that in most cases (20 of 26 SAGs) the  
337 median sizes of retained and removed contigs were not different (Fig. S2); in 6 SAGs the  
338 retained contigs were larger than the removed contigs (Mann–Whitney U,  $p < 0.05$ , two-tailed).

339 Genome size and completeness was greater for *Prochlorococcus* SAGs than SAR11 SAGs. Size  
340 of *Prochlorococcus* SAGs ranged from from 1.28–1.46 Mbp in 85–221 contigs, containing

341 1428–1710 genes; SAR11 SAGs ranged from 0.29–1.14 Mbp in 55–157 contigs, containing  
342 342–1199 genes (Table 1). Completeness was calculated by two methods: fraction of single-copy  
343 core genes observed and CheckM completeness score; genome redundancy was calculated by  
344 CheckM. Completeness of *Prochlorococcus* SAGs ranged from 85.9–90.3% core completeness  
345 and 70.7–78.7% CheckM completeness; SAR11 SAGs ranged from 20.3–90.0% core  
346 completeness and 19.1–76.7% CheckM completeness (Table 1). Genome redundancy of  
347 *Prochlorococcus* SAGs ranged from 0.1–1.0%, and of SAR11 SAGs ranged from 0.0–1.4%  
348 (Table 1). Plotting the number of single-copy core genes as a function of total contig size (Fig.  
349 S3) showed a strong correlation between total contig size and number of single-copy core genes;  
350 this analysis illustrates the greater completeness of the *Prochlorococcus* SAGs relative to the  
351 SAR11 SAGs.

352 Taxonomic assignment of SAGs to clades was done by comparing SAGs against reference  
353 genomes using several methods. Phylogenetic analysis was done on concatenated proteins (89  
354 SAR11 and 96 *Prochlorococcus* shared single-copy orthologous genes) using the maximum  
355 likelihood method (methods). Nucleotide composition (G+C content and *k*-mer composition)  
356 was calculated and compared to reference genomes. Ordination using principal components  
357 analysis (PCA) of *k*-mer composition and OG composition (presence–absence of each OG in  
358 each genome) was used to visualize SAGs in relation to known clades of SAR11 and  
359 *Prochlorococcus*.

360 Phylogenetic analysis of concatenated proteins (Fig. 1) revealed that *Prochlorococcus* SAGs  
361 were all ecotype HLII (5/5). Surveys of the Red Sea using 16S–23S rRNA internal transcribed  
362 spacer (ITS) amplicon sequencing (47), *rpoC1* gene amplicon sequencing (48), and  
363 metagenomic sequencing (12) have each shown that HLII is the dominant *Prochlorococcus*  
364 ecotype in the surface Red Sea. This pattern is consistent with temperature-driven ecotype  
365 distribution patterns of *Prochlorococcus*, where ecotype HLII is predominant in warm/tropical  
366 surface waters (and has a higher thermal tolerance in culture) and ecotype HLI is predominant in  
367 cool/subtropical surface waters (49). SAR11 SAGs were predominantly subclade Ia (13/21), with  
368 the remainder subclades Ib (2/21), Id (3/21), and II (3/21). Placement of the SAR11 SAGs in  
369 these respective clades is supported by a previous phylogenetic analysis of 16S rRNA gene  
370 sequences that included these SAGs (10). Surveys using amplicon sequencing of the 16S rRNA

371 gene (50) and metagenomic sequencing (12) have both shown that SAR11 subclade Ia dominates  
372 the surface Red Sea. Subclade distributions in the 16S survey (50) approximately matched the  
373 distribution of the SAG subclades here, suggesting that the SAGs may approximate the natural  
374 SAR11 population.

375 DNA G+C content of the *Prochlorococcus* SAGs ranged from 31.0–31.4% (Table 1), which is  
376 typical of genomes of *Prochlorococcus* ecotype HLII (51). G+C content of the SAR11 SAGs  
377 was lower, ranging from 27.8–30.5% (Table 1). We have previously shown, using the SAR11  
378 SAGs and other SAR11 genomes, that the ratio of nonsynonymous to synonymous nucleotide  
379 mutations and other genomic evidence in SAR11 genomes is consistent with selection for low  
380 nitrogen driving the low G+C content in marine SAR11 (10).

381 Ordination by PCA of genome properties provided visualization and in some cases improved  
382 resolution of genome taxonomy relative to tree-based methods. For nucleotide composition  
383 analysis, six-nucleotide words (6-mers) were chosen to balance computational time and  
384 information content. The distribution of all 4096 possible 6-mers across the genomes was subject  
385 to dimensionality reduction using PCA and plotted as the first two principal components (PCs).  
386 The first PC explains 27% and 67% of the variation, respectively, for the SAR11 genomes (Fig.  
387 2a) and the *Prochlorococcus* genomes (Fig. 2b). The PCA plots show wider spread in the SAR11  
388 genomes than in the *Prochlorococcus* genomes; both cluster by clade, but the *Prochlorococcus*  
389 genomes are more tightly clustered, with three main clusters (Fig. 2b): HLI nested within HLII  
390 and near HLIII/IV (lower-left), then LLI (middle-left) next-closest followed by LLII and LLIII  
391 (upper-left), and then LLIV distant from the others and more disperse (lower-right).

392 Ordination by PCA of OG composition was done following subsampling of OG counts down to  
393 800 gene counts per SAR11 genome and 1400 gene counts per *Prochlorococcus* genome  
394 (methods). This had the effect of dropping 9 SAR11 SAGs, but it allowed the genomes to have  
395 even depth of coverage for PCA calculation. PCA ordination revealed patterns of OG  
396 composition of SAR11 genomes (Fig. 2c) and *Prochlorococcus* genomes (Fig. 2d). PC1 and PC2  
397 each explained 6–9% of the variation for both sets of genomes. For SAR11, ordination of OG  
398 composition clustered by clade approximately as well as 6-mer composition. For  
399 *Prochlorococcus*, PCA of OG composition provided good separation of the low-light ecotypes



400 (LLI, LLII, LLIII, and LLIV), whereas the high-light ecotypes HLI and HLII formed a single  
401 cluster with HLIII/IV nearby.

402 Of particular interest to investigations of the low-light adapted *Prochlorococcus* ecotypes, we  
403 note that OG composition clearly distinguished between genomes of ecotypes LLII and LLIII. It  
404 has previously been observed that phylogenetic analysis (ITS region) (52, 53) does not resolve  
405 ecotypes LLII and LLIII (identified as high B/A II and III by (54)). Similarly, our analysis of 6-  
406 mer composition also could not resolve these two low-light ecotypes. Our method of “OG  
407 ordination”, however, did distinguish these ecotypes. Thus OG distributions can be a helpful tool  
408 to assign genomes to ecotypes that are indistinguishable by other taxonomic or phylogenetic  
409 methods. The rich genotypic information provided by OG distribution patterns, combined with  
410 an ordination method like PCA, could be applied to other microbial groups for taxonomic  
411 classification of closely related genomes.

#### 412 **Gene clustering and identification of Red-Sea-associated ortholog groups**

413 The SAGs described here come from an undersampled region of the ocean (the Red Sea) and in  
414 part from undersampled clades of marine bacteria (SAR11 subclades Ib, Id, and II), and therefore  
415 provide the opportunity to identify OGs specific for these clades or possibly endemic to this  
416 ocean region. To investigate these patterns, we combined the Red Sea SAGs with available  
417 cultured genomes and SAGs (separately for *Prochlorococcus* and SAR11), clustered genes into  
418 OGs using a Markov clustering algorithm (OrthoMCL, methods), and identified those OGs  
419 found only in the Red Sea SAGs and/or only in certain clades.

420 We identified 878 SAR11 OGs and 96 *Prochlorococcus* Red-Sea-associated OGs (RS-OGs), that  
421 is, OGs not found (in this analysis) in genomes from other parts of the ocean (supplemental file  
422 1). These totals represent 16.7% of all (19.1% of non-core) SAR11 OGs and 0.9% of all (1.0% of  
423 non-core) *Prochlorococcus* OGs. Many of the RS-OGs were found only in a single clade: 96 in  
424 *Prochlorococcus* ecotype HLII, 484 in SAR11 subclade Ia, 101 in SAR11 subclade Ib, 101 in  
425 SAR11 subclade Id, and 132 in SAR11 subclade II. The numerous clade-specific OGs present  
426 targets for understanding ecotype-specific physiology.

427 The first pattern of note was that there were more RS-OGs in the SAR11 SAGs than in the  
428 *Prochlorococcus* SAGs. This reflects the large contribution of our SAR11 SAGs to the

429 sequenced SAR11 pangenome: the number of SAR11 Red Sea SAGs (=21) was nearly as many  
430 as the number of SAR11 reference genomes (=26). In contrast, the number of *Prochlorococcus*  
431 Red Sea SAGs (=5) was only 3% of the number of *Prochlorococcus* reference genomes (=140).  
432 Emphasizing the effect of the genome reference database on estimates of OG endemicity, after  
433 new *Prochlorococcus* genomes (9, 52) were added to the clustering, the number of RS-OGs  
434 dropped from 1192 to 96 (Fig. S4). Another explanation for the greater number of new SAR11  
435 OGs is that the SAR11 SAGs span previously unsampled or undersampled clades: 334 of the 878  
436 Red-Sea-associated SAR11 OGs were found in only one of subclade Ib, Id, or II. Furthermore,  
437 SAR11 is a broader phylogenetic group, based on 16S rRNA diversity, than *Prochlorococcus*  
438 (55), and therefore its pangenome may be expected to be larger. In summary, we suspect that the  
439 large number of new SAR11 OGs (=878), in general, more likely reflects the current dearth of  
440 sequence data for SAR11 rather than a significant degree of endemism due to isolation and/or  
441 selection.

442 The second pattern we examined was inspired by our question about possible endemic gene  
443 content in the Red Sea: based on the geographic isolation of the Red Sea and its unique  
444 combination of physicochemical conditions (simultaneously high irradiance, high salinity, high  
445 temperature, and low nutrients), do genomes isolated from the Red Sea exhibit endemic OG  
446 content encoding adaptive functions for this environment? The answer that emerged to this  
447 question is that there were some indications of possible endemic adaptations to the Red Sea;  
448 however, there were no new pathways identifiable, most of the OGs with annotated functions  
449 were found in only one or two SAGs, and the majority of OGs encoded hypothetical proteins  
450 with no assigned function.

451 The majority of RS-OGs were hypothetical proteins: 82% (723 of 878) for SAR11 and 91% (87  
452 of 96) for *Prochlorococcus*. It was difficult to infer possible adaptive functions for OGs with no  
453 predicted functions; however, these OGs may be referenced later when new approaches for  
454 annotating conserved hypotheticals are developed. The remaining non-hypothetical OGs (155  
455 SAR11, 9 *Prochlorococcus*), i.e., those with predicted functions, are listed in Table S2. While  
456 we could not detect a widespread signature of adaptation to the Red Sea environment—i.e., RS-  
457 OGs with annotated functions represented across multiple SAGs—below we highlight a few

458 sparsely represented RS-OGs that may have adaptive functionality in the Red Sea environment,  
459 some with possible biotechnological potential.

460 Among *Prochlorococcus* SAGs, none of the 9 non-hypothetical RS-OGs (Table S2) were found  
461 in more than one SAG. One OG (proch20425) found in SCGC AAA795-M23 encodes UvrABC  
462 system protein B, responsible for repair of DNA damage. We could posit that this enzyme is  
463 found preferentially in the Red Sea because of the year-round high irradiance, which increases  
464 the rate of DNA damage in cells.

465 Among SAR11 SAGs, there were 21 non-hypothetical RS-OGs found in two or more SAGs and  
466 another 134 found in only one SAG (Table S2). These OGs show links to high light adaptation,  
467 motility, and nitrogen and phosphorus assimilation. One OG (pelag14710, found in one SAG)  
468 encodes a photolyase enzyme that repairs damaged DNA caused by exposure to ultraviolet light.  
469 Pyrophosphatase (pelag15064, found in one SAG) is involved in the hydrolysis of inorganic  
470 pyrophosphate into two orthophosphates and may have a role in phosphorus utilization.  
471 Allantoinase (pelag15247) and urease accessory protein UreF (pelag14490) are each found in  
472 one SAR11 SAG. These enzymes involved in phosphorus and nitrogen metabolism may provide  
473 an adaptive advantage in the Red Sea, which exhibits co-limitation to both elements and may be  
474 relatively more nitrogen-limited (12, 15). Several of the SAR11 RS-OGs encode enzymes with  
475 biotechnological relevance. DNA polymerase I (pelag12679, pelag14776, pelag14807) from this  
476 higher temperature environment could have heat-resistant properties, for example, marginal  
477 thermostability conferred by amino acid substitutions (56).

478 After the major analyses had been completed for this study, two SAR11 genomes (46) and one  
479 *Prochlorococcus* genome (27) derived from cultivated strains were sequenced, and four  
480 *Prochlorococcus* genomes were assembled from metagenomes (57). Of the SAR11 genomes,  
481 one was assigned to subclade Ia and the other to subclade Ib (46). Of note, the subclade Ia  
482 genome (RS39) contained several OGs also found among the Red-Sea-associated SAR11 OGs:  
483 3-oxoacyl-acyl-carrier-protein synthase, ABC branched amino acid transporter,  
484 arylsulfotransferase, formate dehydrogenases, glycosyl transferases, methyltransferases, sialic  
485 acid synthase, sucrose synthase, sulfotransferases, and a type II restriction–modification system.  
486 Several of these functions may play roles in one-carbon and sugar metabolism by SAR11 in the  
487 Red Sea (46). The *Prochlorococcus* genome was assigned to the HLII ecotype and notably

488 contained a pathway for biosynthesis of the osmolyte (compatible solute) glucosylglycerol (27).  
489 This pathway represents a possible adaptation to the higher salinity of the Red Sea. However, the  
490 three genes in this pathway were not found among the Red-Sea-associated *Prochlorococcus*  
491 OGs, nor were they found elsewhere among the retained or removed contigs from the Red Sea  
492 SAGs (BLASTN).

### 493 **Expression of ortholog groups in the Red Sea water column**

494 To further test the idea that there could be OGs of ecological importance endemic to the Red Sea,  
495 we analyzed metatranscriptomes from the Red Sea. Any OGs with functional roles would be  
496 expected to be expressed in the Red Sea water column. We collected seawater and filtered the  
497 prokaryotic fraction from a station in the central Red Sea over a broad temporal and depth range:  
498 samples were collected at four depths and 13 timepoints over a 48-hour period. We extracted and  
499 sequenced RNA from these samples and mapped the reads to the Red Sea SAGs.

500 We found that around two-thirds of RS-OGs were expressed in one or more sample: 64% for  
501 SAR11 (Fig. 3b), 66% for *Prochlorococcus* (Fig. 3d). This was more than the fraction of non-  
502 RS-OGs expressed: 32% for SAR11 (Fig. 3a), 20% for *Prochlorococcus* (Fig. 3c). We were  
503 curious if the high fraction of non-RS-OGs that were unexpressed was due to many of these OG  
504 being singletons (OGs having only one member). To the contrary, heatmaps of OG size  
505 vs. number of metatranscriptomes in which the OG was found (Fig. 3, inset) do not show a high  
506 density of singleton OGs having no expression in non-RS-OGs, and rather the trend toward  
507 singletons is more common in RS-OGs.

508 Of OGs expressed in at least one sample, non-RS-OGs (Fig. 3a,c) tended to be expressed in more  
509 samples than RS-OGs (Fig. 3b,d). This is consistent with many of the non-RS-OGs being core  
510 genes, many of which are housekeeping genes that are often constitutively expressed. Overall,  
511 the expression patterns indicate that the majority of RS-OGs are transcribed to messenger RNA,  
512 consistent with the synthesis of functional gene products.

### 513 **Distribution of ortholog groups across the global ocean**

514 The analysis to this point has focused on the distribution of OGs among cultured and single-cell  
515 genomes and their expression in the Red Sea water column. A set of OGs has been found that is

516 exclusive to Red Sea genomes (to date), and a majority of them are expressed in the water  
517 column. However, we cannot rule out the possibility that these OGs appear endemic only  
518 because more genomes are not available from around the World Ocean. If we extended our  
519 search to global marine metagenomes, instead of just genomes, would we in fact find these  
520 putative endemic OGs in other seas?

521 To investigate the possibility that, contrary to our original hypothesis, there may be few truly  
522 endemic OGs in the Red Sea microbial community, we analyzed metagenomes collected from  
523 across the global ocean by the *Tara* Oceans expedition. We searched for SAR11 and  
524 *Prochlorococcus* OGs in 139 prokaryote-fraction metagenomes from the *Tara* Oceans expedition  
525 (44), which come from several depths in the water column: surface, mixed layer, deep  
526 chlorophyll maximum, and mesopelagic zone. We queried the dataset to determine what fraction  
527 of all OGs and what fraction of RS-OGs could be found outside the Red Sea. If RS-OGs  
528 represent endemic gene content of the Red Sea, we would expect to find them absent from  
529 metagenomes from other regions. Our approach was complementary to a recent study that  
530 analyzed the global metapangenome of *Prochlorococcus* in the *Tara* metagenomes, showing the  
531 distributions of gene clusters (OGs) with strain-level resolution across the *Tara* samples (58). In  
532 the work here, we employed rarefaction and ordination techniques, with a particular focus on  
533 RS-OGs.

534 The presence or absence of SAR11 and *Prochlorococcus* orthologs in *Tara* Oceans prokaryote-  
535 fraction metagenomes (supplemental files 7 and 8) was plotted as rarefaction curves (Fig. 4).  
536 *Tara* Oceans metagenomes were added randomly one by one, and the fraction of SAR11 and  
537 *Prochlorococcus* OGs found was tallied and plotted. The rarefaction curves show the average  $\pm$   
538 standard deviation of 1000 permutations. They also show the best-case (and worst-case)  
539 scenarios, that is, the fraction of OGs found if each new metagenome adds the most (or fewest)  
540 new OGs. Between 70–85% of OGs could be found in one or more *Tara* Oceans metagenome  
541 (Fig. 4), and in the best-case scenarios it took at most ten metagenomes to find 90% of these OGs  
542 (Table S3). The percentage of OGs not found (15–30%) was independent of whether they were  
543 ‘Red-Sea-associated’ or not. This result combined with the rarefaction analysis suggests these  
544 OGs would be unlikely to be found in the *Tara* samples with deeper sequencing. It is possible

545 that some OGs may be rare and/or divergent enough to be undetectable with the current  
546 methodological approach.

547 Across the 139 *Tara* Oceans prokaryote-fraction metagenomes, we found 84.9% (4475/5272) of  
548 all SAR11 OGs in one or more metagenomes (leaving 15.1% not found; Fig. 4a) and 72.2%  
549 (7537/10439) of all *Prochlorococcus* OGs in one or more metagenome (leaving 27.8% not  
550 found; Fig. 4c). In the best-case scenarios, it took only 5 metagenomes to find 90% of the  
551 ‘found’ SAR11 OGs and 50 metagenomes to find 99%; it took only 10 metagenomes to find  
552 90% of the ‘found’ *Prochlorococcus* OGs and 60 metagenomes to find 99% (Table S3). The  
553 fractions of OGs found were similar for RS-OGs, where 81.2% (713/878) of SAR11 OGs were  
554 found (leaving 18.8% not found; Fig. 4b) and 69.8% (67/96) of *Prochlorococcus* OGs were  
555 found (leaving 30.2% not found; Fig. 4d). That is, RS-OGs were about as likely to be found  
556 across the World Ocean as non-RS-OGs. For both SAR11 (Fig. S5a) and *Prochlorococcus* (Fig.  
557 S5b), considering the number of *Tara* metagenomes in which each OG was found, RS-OGs were  
558 less likely to be found in a large fraction of metagenomes, relative to all OGs. This is not  
559 surprising: the set of non-RS-OGs contains all of the core OGs, which would be expected to be  
560 found in most if not all samples.

561 To evaluate whether *Tara* Red Sea metagenomes contained any RS-OGs not already found in the  
562 non-Red Sea metagenomes, we tested scenarios where the Red Sea metagenomes were added  
563 last in the rarefaction analysis. There was no change in the mean curve of cumulative SAR11  
564 OGs found when the six *Tara* Red Sea metagenomes were added last (Fig. 4b): all of the SAR11  
565 RS-OGs could be found without examining the Red Sea metagenomes. In contrast, there were  
566 five *Prochlorococcus* RS-OGs that were added to the cumulative total when the *Tara* Red Sea  
567 metagenomes were added last (Fig. 4d). These five OGs, all with unknown function, represent a  
568 small fraction of the total *Prochlorococcus* pangenome (10439 OGs total). Given the available  
569 genomes, this study may have uncovered a small set of OGs (Table S2) that possibly reflect gene  
570 content endemic to or generally associated with Red Sea environmental conditions, and this  
571 marks an area for further research. In light of this metagenomic analysis, however, it appears that  
572 the putative RS-OGs provide a relatively minor contribution to the whole and that these new  
573 SAR11 and *Prochlorococcus* genomes from the Red Sea generally reflect global pangenomes.

574 Finally, we were curious if OG composition as a whole could show the Red Sea metagenomes to  
575 be different from the other metagenomes, despite the lack of evidence of endemic OGs. More  
576 generally, could the relative abundance of OGs across *Tara* be used to distinguish populations of  
577 *Prochlorococcus* and SAR11?

578 We used the tables of OG counts in the 63 *Tara* surface (SRF) prokaryote-fraction metagenomes  
579 to do PCA ordination on the *Tara* metagenomes (Fig. 5; top OGs driving separation among the  
580 metagenomes provided in Table S4). SAR11 OG composition (Fig. 5a) was not obviously  
581 structured by temperature differences in the temperate and tropical ranges, though Red Sea  
582 samples clustered together, and polar samples were separate from the others. *Prochlorococcus*  
583 OG composition (Fig. 5b), however, was structured by temperature differences in the temperate  
584 and tropical ranges. The four Red Sea samples were split, with two samples clustering with the  
585 warm samples and two samples with the cooler samples. These Red Sea samples are positioned  
586 where they would be expected based on temperature: the two southern samples (latitude: 18.4  
587 °N, 22.0 °N) were warmer (temperature: 27.6 °C, 27.3 °C) and clustered with other  
588 warm/tropical samples (left side of PC1 in Fig. 5b); the two northern samples (latitude: 23.36 °N,  
589 27.16 °N) were cooler (temperature: 25.8 °C, 25.1 °C) and clustered closer to the cool/temperate  
590 samples (right side of PC1 in Fig. 5b). Note these temperatures are lower than average Red Sea  
591 surface waters because the *Tara* Red Sea samples were collected in winter (January); by contrast,  
592 the Red Sea samples in the World Ocean Database (see above) were collected in spring (April).  
593 Given that temperature tolerances generally lack known genetic markers (59), these data suggest  
594 an area for future investigation.

595 In summary, the analysis of *Prochlorococcus* and SAR11 OGs in *Tara* Oceans metagenomes  
596 shows that (i) most “Red-Sea-associated” OGs are actually widely distributed across the World  
597 Ocean, not endemic to the Red Sea; and (ii) OG distribution patterns as a whole, taking relative  
598 abundance into account, place the Red Sea on a continuum with other seas, with patterns  
599 explained by environmental factors including temperature. Supporting this idea, differences in  
600 the relative abundance of OGs—with physicochemical properties covarying with OG  
601 functions—have been observed among the North Pacific, Sargasso Sea, Mediterranean Sea, and  
602 Red Sea in previous comparative metagenomics studies (11, 12). Despite the Red Sea existing at  
603 the periphery of multiple physicochemical parameters in the World Ocean, its distinctiveness



604 may best be revealed by the relative abundance of OGs rather than in the wholesale presence or  
605 absence of OGs. In addition to this general pattern, this effort also identified a small set of  
606 putative and non-hypothetical proteins that warrant further ecological and biotechnological  
607 study.

## 608 **Conclusions and future directions**

609 Here we analyzed SAR11 and *Prochlorococcus* SAGs from an undersampled ocean region, the  
610 Red Sea. This single-cell sequencing effort included SAR11 SAGs from undersampled clades  
611 and provided the first genomes from SAR11 subclades 1b and 1d. Our analysis of these genomes  
612 provided significant contributions to the reference databases of these organisms, adding 878 new  
613 ortholog groups to the SAR11 pangenome and 96 new ortholog groups to the *Prochlorococcus*  
614 pangenome. We described a new method called “OG ordination” that uses PCA of ortholog  
615 group composition to resolve phylogenetic differences in closely related genomes and used it to  
616 distinguish *Prochlorococcus* ecotypes LLII and LLIII in our samples.

617 How marine microbes are able to respond to a changing ocean will be critical to understanding  
618 the future biosphere of planet Earth. At the population and community levels, the cosmopolitan  
619 distribution of genetic functions may confer an advantage, enabling marine microbial  
620 populations and communities, as a whole, to rapidly respond and adapt to changing ocean  
621 conditions. Here we generally considered the Baas Becking hypothesis (“Everything is  
622 everywhere, but the environment selects”) from the perspective of gene ortholog groups (“Every  
623 OG is everywhere, but the environment selects”). The overall data analysis lends support to the  
624 Baas Becking hypothesis as applied to OGs. We described a small set of OGs that may be related  
625 to Red Sea environmental conditions and that mark areas for further investigation. However, the  
626 overall analysis was not consistent with endemism as a primary feature. Instead, we found Red  
627 Sea OGs to be nearly as prevalent across global ocean metagenomes as in Red Sea  
628 metagenomes. This view was supported by analysis of OG relative abundance rather than  
629 absolute presence–absence of OGs. Perhaps OGs may be present but undetectable in a region,  
630 and they become detectable after OG frequencies increase in response to environmental  
631 conditions (via the growth of cells containing those OGs). Therefore, genomic adaptations in a  
632 given ocean region may not simply reflect the presence of OGs unique to a region, but rather the  
633 relative abundance of generally cosmopolitan OGs.

## 634 Acknowledgements

635 We thank Haiwei Luo for assistance building genome trees, Mamoon Rashid for consultation  
636 about decontamination methods, Qiyun Zhu for assistance with genome analysis, and Ramunas  
637 Stepanauskas and Nicole Poulton for assistance with the single-cell genomics protocol.

## 638 Figure Legends

639 **Figure 1.** Maximum-likelihood proteomic trees for single-cell genomes from this study (bold),  
640 plus a representative set of cultured genomes. Trees were built from concatenated alignments of  
641 (a) 89 SAR11 and (b) 96 *Prochlorococcus* single-copy orthologous genes. Bootstrap values are  
642 indicated at the nodes (solid circles  $\geq 80\%$  and open circles  $\geq 50\%$ ). Scale bar equals 0.1 change  
643 per site. The Red Sea SAR11 SAGs cluster with subclades Ia, Ib, Id, and II. The Red Sea  
644 *Prochlorococcus* SAGs all cluster with ecotype HLII.

645 **Figure 2.** PCA ordination of SAGs and genomes based on (a, b) hexanucleotide (6-mer)  
646 composition and (c, d) ortholog group (OG) composition. Genomes are colored by clade; single-  
647 cell genomes from the Red Sea (this study) are circled in black. OG counts, prior to PCA  
648 ordination, were subsampled to 800 (SAR11) or 1400 (*Prochlorococcus*). While both nucleotide  
649 composition and OG composition cluster genomes into discrete groups by clade, OG  
650 composition differentiate clades more clearly, as exemplified by the separation of  
651 *Prochlorococcus* clades LLII and LLIII (panel d).

652 **Figure 3.** Expression of SAG ortholog groups (OGs) in Red Sea metatranscriptomes. The 52  
653 metatranscriptomes span a broad range of the water column at a station in the central Red Sea:  
654 four depths and 13 timepoints over a 48-hour period (every 4 hours). Histograms show the  
655 number of metatranscriptomes found in (a) SAR11 non-RS-OGs, (b) SAR11 RS-OGs, (c)  
656 *Prochlorococcus* non-RS-OGs, and (d) *Prochlorococcus* RS-OGs. Heatmaps (inset) show the  
657 density of OGs based on OG size (number of total copies across the SAGs) and the number of  
658 metatranscriptomes an OG is found in. RS-OGs were more likely than other OGs to be expressed  
659 in one or more samples, and non-RS-OGs that were expressed were more likely to be expressed  
660 in a high number of samples.

661 **Figure 4.** Rarefaction analysis showing the proportion of (a, c) all OGs and (b, d) RS-OGs of  
662 SAR11 and *Prochlorococcus* observed in *Tara* Oceans metagenome samples. Curves show the  
663 cumulative number of OGs observed in *Tara* Oceans samples ( $e$ -value  $< 1e-5$ ) as more samples  
664 are added. Yellow lines show the average  $\pm$  standard deviation of 1000 permutations of  
665 randomly added samples. Blue lines show the “best-case scenario” (each sample added is that  
666 with the most number of new OGs observed) and “worst-case scenario” (each sample added is  
667 that with the fewest number of new OGs observed). Red lines show the mean of 1000  
668 permutations of randomly added samples but with Red Sea samples (031\_SRF\_0.22-1.6,  
669 032\_DCM\_0.22-1.6, 032\_SRF\_0.22-1.6, 033\_SRF\_0.22-1.6, 034\_DCM\_0.22-1.6,  
670 034\_SRF\_0.22-1.6) added last. As more *Tara* metagenome samples are added to the analysis, the  
671 number of new OGs identified approaches a plateau where new samples do not reveal many new  
672 OGs. The same is true with RS-OGs, even when samples from the Red Sea are added last, with  
673 the exception of 5 *Prochlorococcus* OGs (proch20367, proch20368, proch20390, proch20423,  
674 and proch20438).

675 **Figure 5.** Principal components analysis of *Tara* Oceans surface samples by the abundance of  
676 (a) SAR11 and (b) *Prochlorococcus* OGs. The ordination shows the similarity of *Tara* Oceans  
677 samples to each other along the first two principal components. Samples are colored by *Tara*  
678 temperature categories: ‘polar’ samples ( $<10$  °C) are dark blue, ‘temperate’ samples (10–20 °C)  
679 are light blue, ‘tropical’ samples ( $>20$  °C) are orange, and Red Sea ‘tropical’ samples are orange  
680 with black edges. Red Sea samples and *Tara* samples generally show more separation based on  
681 temperature when ordinated by *Prochlorococcus* OG composition than by SAR11 OG  
682 composition.

## 683 References

- 684 1. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF.  
685 2008. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci USA  
686 105:3805–3810.
- 687 2. Waldbauer JR, Rodrigue S, Coleman ML, Chisholm SW. 2012. Transcriptome and proteome  
688 dynamics of a light–dark synchronized bacterial cell cycle. PLoS ONE 7:e43432.

- 689 3. Coleman ML, Chisholm SW. 2007. Code and context: Prochlorococcus as a model for cross-  
690 scale biology. *Trends Microbiol* 15:398–407.
- 691 4. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of  
692 molecular evolution over 60,000 generations. *Nature* 551:45–50.
- 693 5. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of  
694 bacterial innovation. *Nature* 405:299–304.
- 695 6. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman  
696 M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL,  
697 Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. 2003.  
698 Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation.  
699 *Nature* 424:1042–1047.
- 700 7. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. 2009. Whole  
701 genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* 4:e6864.
- 702 8. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012.  
703 Streamlining and core genome conservation among highly divergent members of the SAR11  
704 clade. *mBio* 3:e00252–12.
- 705 9. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen  
706 P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. 2014. Single-cell  
707 genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*  
708 344:416–420.
- 709 10. Luo H, Thompson LR, Stingl U, Hughes AL. 2015. Selection Maintains Low Genomic GC  
710 Content in Marine SAR11 Lineages. *Mol Biol Evol* 32:2738–2748.
- 711 11. Coleman ML, Chisholm SW. 2010. Ecosystem-specific selection pressures revealed through  
712 comparative population genomics. *Proc Natl Acad Sci USA* 107:18634–18639.
- 713 12. Thompson LR, Field C, Romanuk T, Ngugi D, Siam R, El Dorry H, Stingl U. 2013. Patterns  
714 of ecological specialization among microbial populations in the Red Sea and diverse oligotrophic  
715 marine environments. *Ecol Evol* 3:1780–1797.

- 716 13. Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH,  
717 Ackerman M, Moore LR, Meisel JD, Sher D, Thompson LR, Campbell L, Martiny AC,  
718 Chisholm SW. 2015. Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J*  
719 9:1195–1207.
- 720 14. Edwards FJ. 1987. Climate and oceanography, pp. 45–68. *In* Edwards, AJ, Head, SM (eds.),  
721 Key environments: Red sea. Pergamon, Oxford.
- 722 15. Post AF. 2005. Nutrient limitation of marine cyanobacteria, pp. 87–107. *In* Huisman, J,  
723 Matthijs, HCP, Visser, PM (eds.), Harmful cyanobacteria. Springer.
- 724 16. Thompson LR, Williams GJ, Haroon MF, Shibl A, Larsen P, Shorenstein J, Knight R, Stingl  
725 U. 2016. Metagenomic covariation along densely sampled environmental gradients in the Red  
726 Sea. *ISME J* 11:138–151.
- 727 17. Baas Becking LGM. 1934. Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum  
728 & Zoon, The Hague, Netherlands.
- 729 18. Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. 2013. Evidence for a  
730 persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci USA*.
- 731 19. Gonnella G, Böhnke S, Indenbirken D, Garbe-Schönberg D, Seifert R, Mertens C, Kurtz S,  
732 Perner M. 2016. Endemic hydrothermal vent species identified in the open ocean seed bank. *Nat*  
733 *Microbiol* 1:16086.
- 734 20. Stepanauskas R, Sieracki ME. 2007. Matching phylogeny and metabolism in the uncultured  
735 marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* 104:9052–9057.
- 736 21. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin M, Pace NR. 1985. Rapid determination of 16S  
737 ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* 82:6955–6959.
- 738 22. Page KA, Cannon SA, Giovannoni SJ. 2004. Representative Freshwater Bacterioplankton  
739 Isolated from Crater Lake, Oregon. *Appl Environ Microbiol* 70:6542–6550.

- 740 23. Massana R, Murray AE, Preston CM, Delong EF. 1997. Vertical distribution and  
741 phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl*  
742 *Environ Microbiol* 63:50–56.
- 743 24. Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T, Eisen JA, Fraser  
744 CM, DeLong EF. 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs.  
745 *Nature* 415:630–633.
- 746 25. Stewart FJ, Dalsgaard T, Young CR, Thamdrup B, Revsbech NP, Ulloa O, Canfield DE,  
747 DeLong EF. 2012. Experimental incubations elicit profound changes in community transcription  
748 in OMZ bacterioplankton. *PLoS ONE* 7:e37118.
- 749 26. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC. 2009. IMG  
750 ER: a system for microbial genome annotation expert review and curation. *Bioinformatics*  
751 (Oxford, England) 25:2271–2278.
- 752 27. Shibl AA, Ngugi DK, Talarmin A, Thompson LR, Blom J, Stingl U. 2018. The genome of a  
753 novel isolate of *Prochlorococcus* from the Red Sea contains transcribed genes for compatible  
754 solute biosynthesis. *FEMS Microbiology Ecology*.
- 755 28. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina  
756 sequence data. *Bioinformatics* (Oxford, England) 30:2114–2120.
- 757 29. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient  
758 alignment of short DNA sequences to the human genome. *CORD Conference Proceedings*  
759 10:R25–R25.
- 760 30. Nikolenko SI, Korobeynikov AI, Alekseyev MA. 2013. BayesHammer: Bayesian clustering  
761 for error correction in single-cell sequencing. *BMC Genomics* 14 Suppl 1:S7.
- 762 31. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,  
763 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev  
764 MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to  
765 Single-Cell Sequencing. *J Comput Biol* 19:455–477.

- 766 32. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal  
767 RNAs in metatranscriptomic data. *Bioinformatics (Oxford, England)* 28:3211–3217.
- 768 33. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based  
769 automatic contiguation of assembled sequences. *Bioinformatics (Oxford, England)* 25:1968–  
770 1969.
- 771 34. Bonfield JK, Smith KF, Staden R. 1995. A new DNA sequence assembly program. *Nucleic  
772 Acids Res* 23:4992–4999.
- 773 35. Strous M, Kraft B, Bisdorf R. 2012. The binning of metagenomic contigs for microbial  
774 physiology of mixed cultures. *Front Microbiol.*
- 775 36. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S,  
776 Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK,  
777 Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V,  
778 Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology.  
779 *BMC Genomics* 9:75.
- 780 37. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for  
781 eukaryotic genomes. *Genome Res* 13:2178–2189.
- 782 38. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing  
783 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.  
784 *Genome Res* 25:1043–1055.
- 785 39. Contreras-Moreira B, Vinuesa P. 2013. GET\_HOMOLOGUES, a versatile software package  
786 for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol* 79:7696.
- 787 40. Katoh K. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment.  
788 *Nucleic Acids Res* 33:511–518.
- 789 41. Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of  
790 partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* 29:1695–  
791 1701.



- 792 42. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of  
793 large phylogenies. *Bioinformatics* (Oxford, England).
- 794 43. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,  
795 Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot  
796 M, Duchesnay É. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–  
797 2830.
- 798 44. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B,  
799 Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen  
800 S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G,  
801 Poulain J, Poulos BT, Royo-Llonch M, Sarmiento H, Vieira-Silva S, Dimier C, Picheral M,  
802 Searson S, Kandels-Lewis S, Bowler C, Vargas C de, Gorsky G, Grimsley N, Hingamp P,  
803 Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemann L, Sullivan MB,  
804 Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton.  
805 Structure and function of the global ocean microbiome. *Science* 348:1261359–1261359.
- 806 45. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND.  
807 *Nat Meth* 12:59–60.
- 808 46. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, Stingl U. 2017. Genomic  
809 characterization of two novel SAR11 isolates from the Red Sea, including the first strain of the  
810 SAR11 Ib clade. *FEMS Microbiol Ecol* 93.
- 811 47. Shibl AA, Thompson LR, Ngugi DK, Stingl U. 2014. Distribution and diversity of  
812 *Prochlorococcus* ecotypes in the Red Sea. *FEMS Microbiol Lett* 356:118–126.
- 813 48. Shibl AA, Haroon MF, Ngugi DK, Thompson LR, Stingl U. 2016. Distribution of  
814 *Prochlorococcus* Ecotypes in the Red Sea Basin Based on Analyses of *rpoC1* Sequences. *Front*  
815 *Mar Sci* 3.
- 816 49. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. 2006. Niche  
817 partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients.  
818 *Science* 311:1737–1740.

- 819 50. Ngugi DK, Stingl U. 2012. Combined analyses of the ITS loci and the corresponding 16S  
820 rRNA genes reveal high micro- and macrodiversity of SAR11 populations in the Red Sea. *PLoS*  
821 *ONE* 7:e50274.
- 822 51. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A,  
823 Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. 2007. Patterns and  
824 implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231.
- 825 52. Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L, Roache-  
826 Johnson KH, Ding H, Giovannoni SJ, Rocap G, Moore LR, Chisholm SW. 2014. Genomes of  
827 diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data* 1.
- 828 53. Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. *Prochlorococcus*: the structure and  
829 function of collective diversity. *Nat Rev Microbiol* 13:13–27.
- 830 54. Rocap G, Distel DL, Waterbury JB, Chisholm SW. 2002. Resolution of *Prochlorococcus* and  
831 *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer  
832 sequences. *Appl Environ Microbiol* 68:1180–1191.
- 833 55. Ngugi DK, Antunes A, Brune A, Stingl U. 2012. Biogeography of pelagic bacterioplankton  
834 across an antagonistic temperature-salinity gradient in the Red Sea. *Mol Ecol* 21:388–405.
- 835 56. Somero GN, Lockwood BL, Tomanek L. 2016. *Biochemical Adaptation: Response to*  
836 *Environmental Challenges, from Life’s Origins to the Anthropocene*.
- 837 57. Haroon MF, Thompson LR, Parks DH, Hugenholtz P, Stingl U. 2016. A catalogue of 136  
838 microbial draft genomes from Red Sea metagenomes. *Scientific Data* 3:160050.
- 839 58. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the *Prochlorococcus*  
840 metapangenome. *PeerJ* 6:e4320.
- 841 59. Hickey DA, Singer GA. 2004. Genomic and proteomic adaptations to growth at high  
842 temperature. *Genome Biol* 5:117.

**Table 1.** Genomic features of *Prochlorococcus* and SAR11 single-cell genomes. Single cells were isolated from a surface sample from the Eastern Red Sea (19.75 °N, 40.05 °E). *Prochlorococcus* clades are ecotypes; SAR11 clades are subclades. Completeness is reported as the fraction of 1144 *Prochlorococcus* or 649 SAR11 single-copy core OGs found in each SAG; completeness is also reported as the percent of bacterial single-copy core OGs present as determined by CheckM. Redundancy of bacterial single-copy core OGs is defined as the “contamination” parameter from the CheckM software.

Genus	SAG ref. no.	Clade	Contigs	Assembled size (bp)	Genes	Single-copy core genes	Completeness (core, %)	Completeness (CheckM, %)	Redundancy (CheckM, %)	G+C (%)
<i>Prochlorococcus</i>	SCGC AAA795-F05	HLII	136	1,418,374	1632	1033	90.2	78.6	0.27	31.4
<i>Prochlorococcus</i>	SCGC AAA795-I06	HLII	120	1,388,767	1604	981	85.9	77.5	0.10	31.1
<i>Prochlorococcus</i>	SCGC AAA795-I15	HLII	221	1,282,941	1428	989	86.6	70.7	0.97	31.3
<i>Prochlorococcus</i>	SCGC AAA795-J16	HLII	85	1,463,721	1691	1033	90.3	78.7	0.52	31.0
<i>Prochlorococcus</i>	SCGC AAA795-M23	HLII	93	1,443,989	1710	1012	88.7	74.6	0.34	31.2
SAR11	SCGC AAA795-A08	Ia	61	374,567	384	158	24.3	24.5	0.00	28.3
SAR11	SCGC AAA795-A20	Ia	63	1,140,609	1199	584	90.0	76.7	0.00	29.1
SAR11	SCGC AAA795-B16	Ib	95	551,717	600	331	51.0	34.7	0.06	29.4
SAR11	SCGC AAA795-C09	Ia	82	667,038	734	390	60.1	44.6	0.88	28.4
SAR11	SCGC AAA795-C10	Ia	55	477,445	503	213	32.8	34.9	0.23	29.3
SAR11	SCGC AAA795-D22	Ia	68	1,010,421	1082	555	85.5	69.9	0.60	28.8
SAR11	SCGC AAA795-E07	II	101	681,366	737	418	64.4	56.9	1.37	29.7
SAR11	SCGC AAA795-E22	Ib	63	801,227	820	417	64.3	47.6	0.34	29.0
SAR11	SCGC AAA795-F16	Ib	74	945,491	1017	509	78.4	65.9	0.00	29.1
SAR11	SCGC AAA795-G15	II	62	294,337	342	132	20.3	19.1	0.46	30.5
SAR11	SCGC AAA795-J21	Ia	77	872,902	954	404	62.2	51.5	0.70	29.1
SAR11	SCGC AAA795-K18	Ia	114	731,292	782	314	48.4	48.7	0.70	29.9
SAR11	SCGC AAA795-L23	Ia	150	834,822	910	489	75.3	54.4	0.60	27.8
SAR11	SCGC AAA795-M18	Ib	61	1,050,527	1072	456	70.3	58.9	1.41	29.2
SAR11	SCGC AAA795-M22	Ib	80	860,157	921	515	79.4	64.2	0.13	29.4
SAR11	SCGC AAA795-N08	Ia	157	575,315	622	272	41.9	33.3	0.55	29.1
SAR11	SCGC AAA795-N17	II	94	611,592	620	361	55.6	38.0	0.42	29.5
SAR11	SCGC AAA795-O19	Ia	62	804,609	862	379	58.4	54.2	0.04	29.1
SAR11	SCGC AAA795-O20	Ia	62	1,009,143	1074	526	81.0	69.0	0.04	29.0
SAR11	SCGC AAA795-P11	Ia	127	977,727	1021	485	74.7	52.4	1.32	29.2
SAR11	SCGC AAA797-I19	Ia	77	1,016,895	1071	468	72.1	66.4	0.59	29.2











