







Red Sea SAR11 and *Prochlorococcus* Single-Cell Genomes Reflect Globally Distributed Pangenomes

 Luke R. Thompson,^{a,b,c} Mohamed F. Haroon,^a
 Ahmed A. Shibl,^{a,*} Matt J. Cahill,^a
 David K. Ngugi,^a Gareth J. Williams,^d James T. Morton,^{e,f} Rob Knight,^{e,f,g}
 Kelly D. Goodwin,^c Ulrich Stingl^{a,h}

^aRed Sea Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

^bDepartment of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, Mississippi, USA

^cOcean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, La Jolla, California, USA

^dSchool of Ocean Sciences, Bangor University, Anglesey, United Kingdom

^eDepartment of Pediatrics, University of California San Diego, La Jolla, California, USA

^fDepartment of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

^gCenter for Microbiome Innovation, University of California, San Diego, California, USA

^hDepartment of Microbiology & Cell Science, Fort Lauderdale Research and Education Center, UF/Institute of Food and Agricultural Sciences, University of Florida, Davie, Florida, USA

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

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

Citation Thompson LR, Haroon MF, Shibl AA, Cahill MJ, Ngugi DK, Williams GJ, Morton JT, Knight R, Goodwin KD, Stingl U. 2019. Red Sea SAR11 and *Prochlorococcus* single-cell genomes reflect globally distributed pangenomes. *Appl Environ Microbiol* 85:e00369-19. <https://doi.org/10.1128/AEM.00369-19>.

Editor Shuang-Jiang Liu, Chinese Academy of Sciences

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

Address correspondence to Luke R. Thompson, lukethompson@gmail.com, or Ulrich Stingl, ulistingl@gmail.com.

* Present address: Ahmed A. Shibl, Marine Microbial Ecology Lab, Biology Program, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates.

Received 13 February 2019

Accepted 19 April 2019

Accepted manuscript posted online 26 April 2019

Published 17 June 2019

everywhere, but the environment selects" also applies to gene ortholog groups. This widely dispersed functional diversity may give oceanic microbial communities the functional capacity to respond rapidly to changing conditions.

KEYWORDS *Pelagibacter*, metagenomics, metatranscriptomics, population genomics, single-cell genomics

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

Examples of adaptive gene presence/absence patterns are seen in the most numerous groups of bacteria in the oligotrophic tropical and subtropical surface ocean, the photoautotrophic picocyanobacteria *Prochlorococcus* and *Synechococcus* and the chemoheterotrophic *Alphaproteobacteria* SAR11 clade ("*Candidatus Pelagibacter ubique*"). Genomes of these genera are smaller than their relatives in less nutrient-poor environments (6, 8), suggestive of genome streamlining to conserve resources used for genome replication (3). Consistent with genome streamlining, the genes maintained in *Prochlorococcus* and SAR11 genomes are correlated with physical features in parts of the water column in which they are found, for example, genes for acquisition of nitrite and nitrate in genomes found where those compounds are available (3, 8). Examples revealed through comparative community genomics include an enrichment of phosphorus acquisition gene ortholog groups in the Atlantic relative to the Pacific Ocean (11, 13) and an enrichment in osmolyte oxidation gene ortholog groups in the Mediterranean and Red Seas relative to the Atlantic and Pacific Oceans (12).

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

Evidence of genomic adaptation to high light and high salinity in the Red Sea has been revealed through comparative metagenomics, showing increased relative abundance of known gene ortholog groups in *Prochlorococcus* and SAR11 (12). Relative to *Prochlorococcus* populations in the North Pacific, Sargasso Sea, and Mediterranean Sea, the Red Sea *Prochlorococcus* population had increased frequencies of high light stress and DNA repair gene ortholog groups (12), the latter likely an adaptation to UV-induced DNA damage. Relative to SAR11 populations in those same seas, the Red Sea SAR11 population had increased frequencies of gene ortholog groups for osmolyte degradation (12); osmolytes are important molecules for surviving high salinity in many organisms. Across 45 metagenomes along latitudinal and depth gradients from the surface to 500 m in the Red Sea, temperature explained more variation in gene ortholog groups than any other environmental parameter, and the relative abundance

of gene ortholog groups linked to high irradiance, high salinity, and low nutrients was correlated with those parameters (16).

The above-mentioned patterns observed in comparative metagenomics studies were all based on relative abundance of known gene ortholog groups, dependent on a reference genome database with no representatives from the Red Sea. Therefore, the question remains if there are gene functions in the *Prochlorococcus* and SAR11 populations in the Red Sea not found in any other *Prochlorococcus* and SAR11 populations in the ocean. Because of its relative geographic isolation, we might expect the Red Sea to be genetically isolated, with endemic genomic adaptations to its unique combination of high solar irradiance, high temperature, high salinity, and low nutrient levels. Newly identified gene ortholog groups could be informative for understanding microbial adaptation and mechanisms of stress tolerance and have potential biotechnological applications.

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

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

RESULTS AND DISCUSSION

Single-cell genome properties and taxonomic classification. Following collection of surface seawater from the east-central Red Sea, flow sorting, and amplification, we sequenced and assembled 21 SAR11 and 5 *Prochlorococcus* single-cell amplified genomes (SAGs). These SAGs represent reference genomes in an ocean region with sparse coverage: only one cultured *Prochlorococcus* genome (20) and two cultured SAR11 genomes (21) are currently available from the Red Sea. The SAR11 SAGs also represent genomes from clades without other sequenced representatives: two SAGs from subclade Ib and three SAGs from subclade IId (Fig. 1).

To account for and remove any possible contaminating DNA sequences, assembled contigs were retained only if they were part of an SAR11 or *Prochlorococcus* Metawatt bin or if they had a top-10 BLASTN hit to a *Prochlorococcus* or SAR11 genome (see

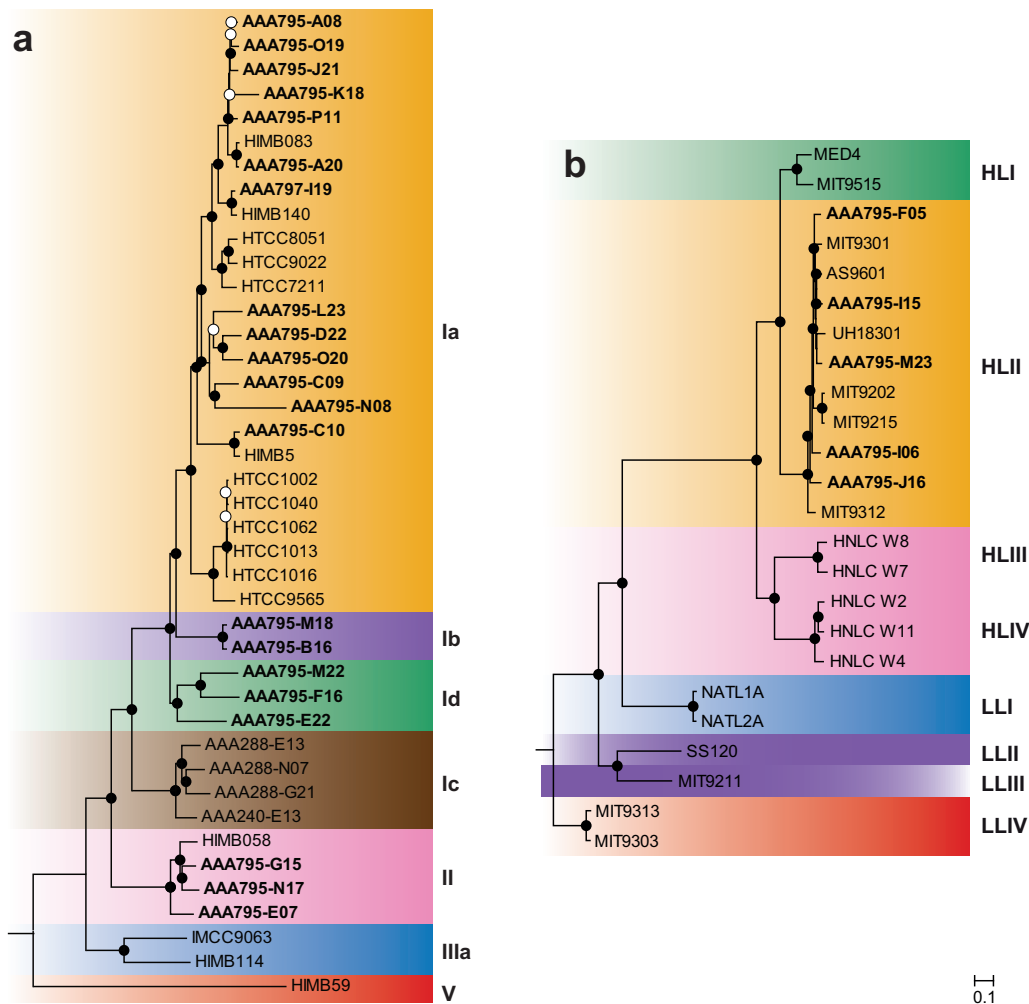


FIG 1 Maximum-likelihood proteomic trees for single-cell genomes from this study (boldface), plus a representative set of cultured genomes. Trees were built from concatenated alignments of 89 SAR11 (a) and 96 *Prochlorococcus* (b) single-copy orthologous genes. Bootstrap values are indicated at the nodes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

Materials and Methods). In Metawatt, assignment to bins is based on tetranucleotide frequency, and the average taxonomy of the bin is determined by BLAST of 500-bp fragments of all the contigs against a prokaryotic database (22). A contig matching the tetranucleotide frequency of an SAR11 or *Prochlorococcus* bin could be retained even if it contained contradictory or missing taxonomic information. However, to check if our secondary, BLASTN-based assignment process could be biased against short contigs, which might lack a neighboring anchor gene, we analyzed the distribution of contig lengths between retained and removed contigs for each SAG. We found that in most cases (20 of 26 SAGs) the median sizes of retained and removed contigs were not different (see Fig. S2 in the supplemental material); in 6 SAGs the retained contigs were larger than the removed contigs ($P < 0.05$ by two-tailed Mann-Whitney U test).

Genome size and completeness was greater for *Prochlorococcus* SAGs than SAR11 SAGs. Sizes of *Prochlorococcus* SAGs ranged from 1.28 to 1.46 Mbp in 85 to 221 contigs, containing 1,428 to 1,710 genes; SAR11 SAGs ranged from 0.29 to 1.14 Mbp in 55 to 157 contigs, containing 342 to 1,199 genes (Table 1). Completeness was calculated by two methods: fraction of single-copy core genes observed and CheckM completeness score. Genome redundancy was calculated by CheckM. Completeness of *Prochlorococcus* SAGs ranged from 85.9 to 90.3% core completeness and 70.7 to 78.7% CheckM

TABLE 1 Genomic features of *Prochlorococcus* and SAR11 single-cell genomes^a

Genus	SAG reference no.	Clade	No. of contigs	Assembled size (bp)	No. of genes	No. of single-copy core genes	Completeness (%)		Redundancy (CheckM, %)	G+C (%)
							Core	CheckM		
<i>Prochlorococcus</i>	SCGC AAA795-F05	HLII	136	1,418,374	1,632	1,033	90.2	78.6	0.27	31.4
<i>Prochlorococcus</i>	SCGC AAA795-I06	HLII	120	1,388,767	1,604	981	85.9	77.5	0.10	31.1
<i>Prochlorococcus</i>	SCGC AAA795-I15	HLII	221	1,282,941	1,428	989	86.6	70.7	0.97	31.3
<i>Prochlorococcus</i>	SCGC AAA795-J16	HLII	85	1,463,721	1,691	1,033	90.3	78.7	0.52	31.0
<i>Prochlorococcus</i>	SCGC AAA795-M23	HLII	93	1,443,989	1,710	1,012	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	1,199	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	1,082	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	1,017	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	1,072	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	1,074	526	81.0	69.0	0.04	29.0
SAR11	SCGC AAA795-P11	Ia	127	977,727	1,021	485	74.7	52.4	1.32	29.2
SAR11	SCGC AAA797-I19	Ia	77	1,016,895	1,071	468	72.1	66.4	0.59	29.2

^aSingle cells were isolated from a surface sample from the eastern Red Sea (19.75°N, 40.05°E). *Prochlorococcus* clades are ecotypes, and SAR11 clades are subclades. Completeness is reported as the fraction of 1,144 *Prochlorococcus* or 649 SAR11 single-copy core OGs found in each SAG; completeness is also reported as the percentage 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.

completeness; SAR11 SAGs ranged from 20.3 to 90.0% core completeness and 19.1 to 76.7% CheckM completeness (Table 1). Genome redundancy of *Prochlorococcus* SAGs ranged from 0.1 to 1.0% and of SAR11 SAGs ranged from 0.0 to 1.4% (Table 1). Plotting the number of single-copy core genes as a function of total contig size (Fig. S3) showed a strong correlation between total contig size and number of single-copy core genes; this analysis illustrates the greater completeness of the *Prochlorococcus* SAGs relative to that of the SAR11 SAGs.

Taxonomic assignment of SAGs to clades was done by comparing SAGs against reference genomes using several methods. Phylogenetic analysis was done on concatenated proteins (89 SAR11 and 96 *Prochlorococcus* shared single-copy orthologous genes) using the maximum likelihood method (see Materials and Methods). Nucleotide composition (G+C content and *k*-mer composition) was calculated and compared to that of reference genomes. Ordination using principal component analysis (PCA) of *k*-mer composition and OG composition (presence/absence of each OG in each genome) was used to visualize SAGs in relation to known clades of SAR11 and *Prochlorococcus*.

Phylogenetic analysis of concatenated proteins (Fig. 1) revealed that *Prochlorococcus* SAGs were all ecotype HLII (5/5). Surveys of the Red Sea using 16S-23S rRNA internal transcribed spacer (ITS) amplicon sequencing (23), *rpoC1* gene amplicon sequencing (24), and metagenomic sequencing (12) have each shown that HLII is the dominant *Prochlorococcus* ecotype in the surface of the Red Sea. This pattern is consistent with temperature-driven ecotype distribution patterns of *Prochlorococcus*, where ecotype HLII is predominant in warm/tropical surface waters (and has a higher thermal tolerance in culture) and ecotype HLI is predominant in cool/subtropical surface waters (25). SAR11 SAGs were predominantly subclade Ia (13/21), with the remainder being subclades Ib (2/21), Id (3/21), and II (3/21). Placement of the SAR11 SAGs in these respective

clades is supported by a previous phylogenetic analysis of 16S rRNA gene sequences that included these SAGs (10). Surveys using amplicon sequencing of the 16S rRNA gene (26) and metagenomic sequencing (12) have both shown that SAR11 subclade Ia dominates the Red Sea surface. Subclade distributions in the 16S survey (26) approximately matched the distribution of the SAG subclades here, suggesting that the SAGs approximate the natural SAR11 population.

DNA G+C content of the *Prochlorococcus* SAGs ranged from 31.0 to 31.4% (Table 1), which is typical of genomes of *Prochlorococcus* ecotype HLII (27). G+C content of the SAR11 SAGs was lower, ranging from 27.8 to 30.5% (Table 1). We have previously shown, using the SAR11 SAGs and other SAR11 genomes, that the ratio of nonsynonymous to synonymous nucleotide mutations and other genomic evidence in SAR11 genomes is consistent with selection for low nitrogen driving the low G+C content in marine SAR11 (10).

Ordination by PCA of genome properties provided visualization and in some cases improved resolution of genome taxonomy relative to that of tree-based methods. For nucleotide composition analysis, six-nucleotide words (6-mers) were chosen to balance computational time and information content. The distribution of all 4,096 possible 6-mers across the genomes was subject to dimensionality reduction using PCA and plotted as the first two principal components (PCs). The first PC explains 27% and 67% of the variation, respectively, for the SAR11 genomes (Fig. 2a) and the *Prochlorococcus* genomes (Fig. 2b). The PCA plots show wider spread in the SAR11 genomes than in the *Prochlorococcus* genomes; both cluster by clade, but the *Prochlorococcus* genomes are more tightly clustered, with three main clusters (Fig. 2b): HLI nested within HLII and near HLIII/IV (lower left), LLII (middle left) was next closest, followed by LLII and LLIII (upper left), and LLIV was distant from the others and more dispersed (lower right).

Ordination by PCA of OG composition was done by random subsampling of OG counts down to 800 gene counts per SAR11 genome and 1,400 gene counts per *Prochlorococcus* genome (see Materials and Methods). This had the effect of dropping 9 SAR11 SAGs, but it allowed the genomes to have even depth of coverage for PCA calculation. PCA ordination revealed patterns of OG composition of SAR11 genomes (Fig. 2c) and *Prochlorococcus* genomes (Fig. 2d). PC1 and PC2 each explained 6 to 9% of the variation for both sets of genomes. For SAR11, ordination of OG composition clustered by clade approximately as well as 6-mer composition. For *Prochlorococcus*, PCA of OG composition provided good separation of the low-light ecotypes (LLI, LLII, LLIII, and LLIV), whereas the high-light ecotypes HLI and HLII formed a single cluster with HLIII/IV nearby.

Of particular interest to investigations of the low-light-adapted *Prochlorococcus* ecotypes, we note that OG composition clearly distinguished between genomes of ecotypes LLII and LLIII. It has previously been observed that phylogenetic analysis (ITS region) (28, 29) does not resolve ecotypes LLII and LLIII (identified as high B/A II and III in reference 30). Similarly, our analysis of 6-mer composition also could not resolve these two low-light ecotypes. Our method of OG ordination, however, did distinguish these ecotypes. Thus, OG distributions can be a helpful tool to assign genomes to ecotypes that are indistinguishable by other taxonomic or phylogenetic methods. The rich genotypic information provided by OG distribution patterns, combined with an ordination method like PCA, could be applied to other microbial groups for taxonomic classification of closely related genomes.

Gene clustering and identification of Red Sea-associated ortholog groups. The SAGs described here come from an undersampled region of the ocean (the Red Sea) and in part from undersampled clades of marine bacteria (SAR11 subclades Ib, Id, and II) and therefore provide the opportunity to identify OGs specific for these clades or possibly endemic to this ocean region. To investigate these patterns, we combined the Red Sea SAGs with available cultured genomes and SAGs (separately for *Prochlorococcus* and SAR11), clustered genes into OGs using a Markov clustering algorithm

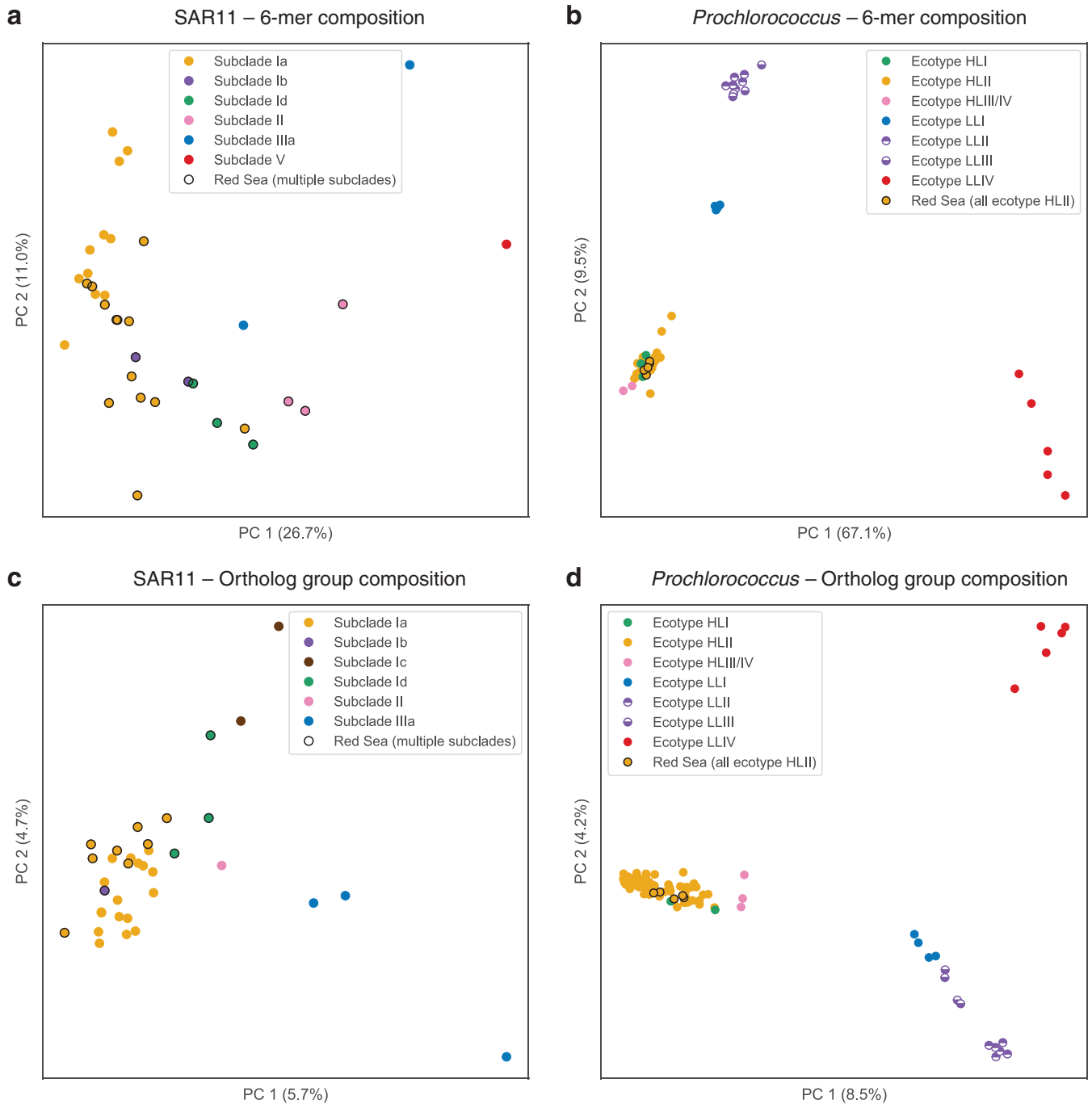


FIG 2 PCA ordination of SAGs and genomes based on hexanucleotide (6-mer) composition (a and b) and OG composition (c and d). Genomes are colored by clade; single-cell genomes from the Red Sea (this study) are circled in black. OG counts, prior to PCA ordination, were randomly subsampled to 800 (SAR11) or 1,400 (*Prochlorococcus*). While both nucleotide composition and OG composition cluster genomes into discrete groups by clade, OG composition differentiates clades more clearly, as exemplified by the separation of *Prochlorococcus* clades LLII and LLIII (d).

(OrthoMCL; see Materials and Methods), and identified OGs found only in the Red Sea SAGs (RS-OGs) and OGs found only in certain clades.

We identified 878 SAR11 RS-OGs and 96 *Prochlorococcus* RS-OGs, that is, OGs not found (in this analysis) in genomes from outside the Red Sea (File S1). These totals represent 16.7% of all (19.1% of noncore) SAR11 OGs and 0.9% of all (1.0% of noncore) *Prochlorococcus* OGs. Many of the RS-OGs were found only in a single clade: 96 in *Prochlorococcus* ecotype HLII, 484 in SAR11 subclade Ia, 101 in SAR11 subclade Ib, 101

in SAR11 subclade Id, and 132 in SAR11 subclade II. The numerous clade-specific OGs present targets for understanding ecotype-specific physiology.

The first pattern of note was that there were more RS-OGs in the SAR11 SAGs than in the *Prochlorococcus* SAGs. This reflects the large contribution of our SAR11 SAGs to the sequenced SAR11 pangenome: the number of SAR11 Red Sea SAGs (21) was nearly as high as the number of SAR11 reference genomes (26). In contrast, the number of *Prochlorococcus* Red Sea SAGs (5) was only 3% of the number of *Prochlorococcus* reference genomes (140). Emphasizing the effect of the genome reference database on estimates of OG endemicity, after new *Prochlorococcus* genomes (9, 28) were added to the clustering, the number of RS-OGs dropped from 1,192 to 96 (Fig. S4). Another explanation for the greater number of new SAR11 OGs is that the SAR11 SAGs span previously unsampled or undersampled clades: 334 of the 878 Red Sea-associated SAR11 OGs were found in only one of subclade Ib, Id, or II. Furthermore, SAR11 is a broader phylogenetic group, based on 16S rRNA diversity, than *Prochlorococcus* (31); therefore, its pangenome may be expected to be larger. In summary, we suspect that the large number of new SAR11 OGs (878) in general more likely reflects the current dearth of sequence data for SAR11 rather than a significant degree of endemism due to isolation and/or selection.

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

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

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

Among SAR11 SAGs, there were 21 nonhypothetical RS-OGs found in two or more SAGs and another 134 found in only one SAG (Table S2). These OGs show links to high light adaptation, motility, and nitrogen and phosphorus assimilation. One OG (pelag14710, found in one SAG) encodes a photolyase enzyme that repairs DNA damage caused by exposure to UV light. Pyrophosphatase (pelag15064; found in one SAG) is involved in the hydrolysis of inorganic pyrophosphate into two orthophosphates and may have a role in phosphorus utilization. Allantoinase (pelag15247) and urease accessory protein UreF (pelag14490) are each found in one SAR11 SAG. These enzymes involved in phosphorus and nitrogen metabolism may provide an adaptive advantage in the Red Sea, which exhibits colimitation to both elements and may be relatively more nitrogen limited (12, 15). Several of the SAR11 RS-OGs encode enzymes with biotechnological relevance. DNA polymerase I (pelag12679, pelag14776, and pelag14807) from this higher temperature environment could have heat-resistant

properties, for example, marginal thermostability conferred by amino acid substitutions (32).

After the major analyses had been completed for this study, two SAR11 genomes (21) and one *Prochlorococcus* genome (20) derived from cultivated strains were sequenced, and four *Prochlorococcus* genomes were assembled from metagenomes (33). Of the SAR11 genomes, one was assigned to subclade Ia and the other to subclade Ib (21). Of note, the subclade Ia genome (RS39) contained several OGs also found among the Red Sea-associated SAR11 OGs: 3-oxoacyl-acyl-carrier-protein synthase, ABC branched amino acid transporter, arylsulfotransferase, formate dehydrogenases, glycosyl transferases, methyltransferases, sialic acid synthase, sucrose synthase, sulfotransferases, and a type II restriction-modification system. Several of these functions may play roles in one-carbon and sugar metabolism by SAR11 in the Red Sea (21). The *Prochlorococcus* genome was assigned to the HLII ecotype and notably contained a pathway for biosynthesis of the osmolyte (compatible solute) glucosylglycerol (20). This pathway represents a possible adaptation to the higher salinity of the Red Sea. However, the three genes in this pathway were not found among the Red-Sea-associated *Prochlorococcus* OGs, and they were not found elsewhere among the retained or removed contigs from the Red Sea SAGs (BLASTN).

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

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

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

Distribution of ortholog groups across the global ocean. The analysis to this point has focused on the distribution of OGs among cultured and single-cell genomes and their expression in the Red Sea water column. A set of OGs has been found that is exclusive to Red Sea genomes (to date), and a majority of them are expressed in the water column. However, we cannot rule out the possibility that these OGs appear endemic only because more genomes are not available from around the World Ocean. If we extended our search to global marine metagenomes, instead of just genomes, would we in fact find these putative endemic OGs in other seas?

To investigate the possibility that, contrary to our original hypothesis, there may be few OGs truly endemic to the Red Sea microbial community, we analyzed metagenomes collected from across the global ocean by the *Tara* Oceans expedition. We searched for SAR11 and *Prochlorococcus* OGs in 139 prokaryote-fraction metagenomes from the *Tara* Oceans expedition (34), which come from several depths in the water column: surface, mixed layer, deep chlorophyll maximum, and mesopelagic zone. We

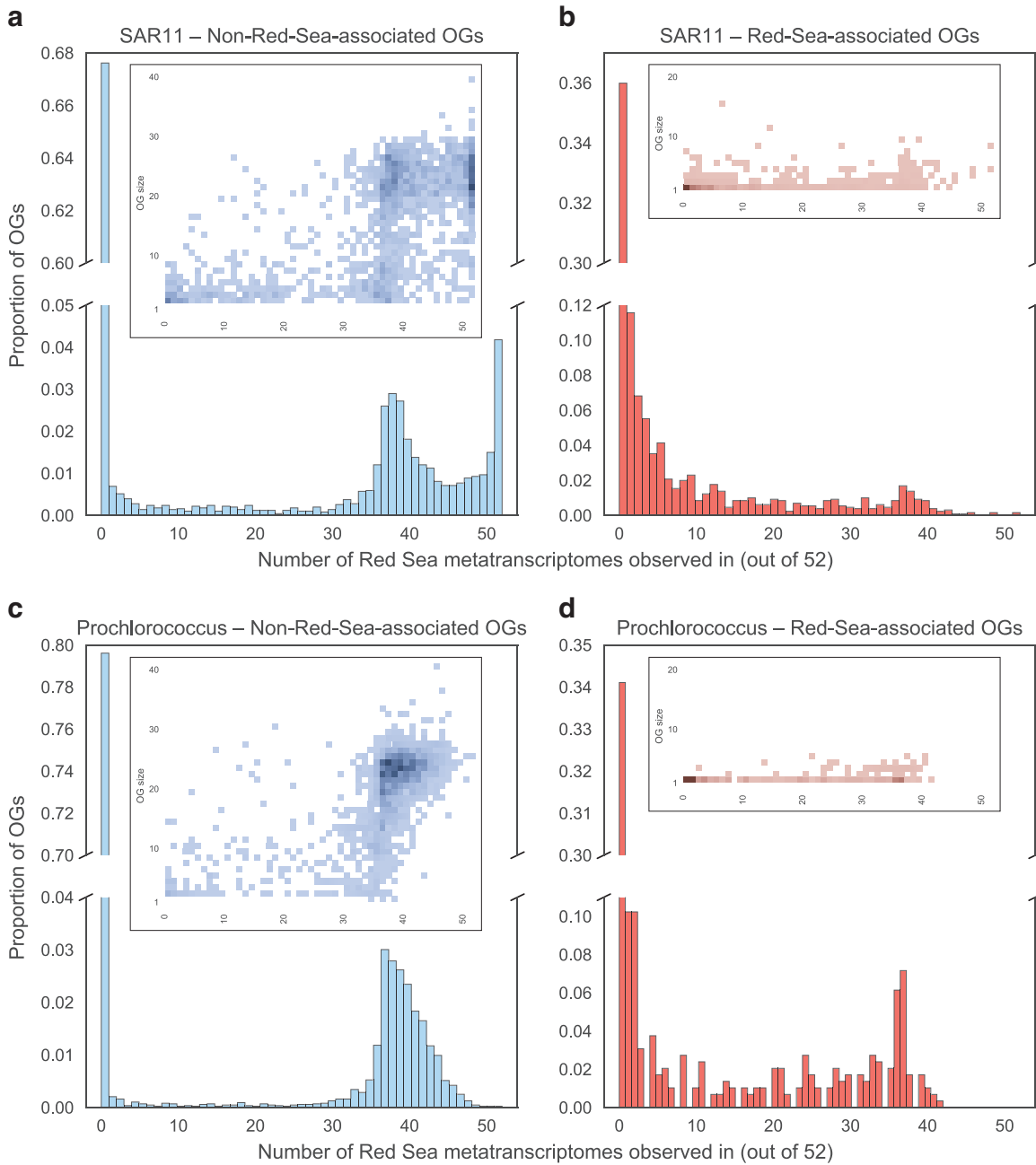


FIG 3 Expression of SAG OGs in Red Sea metatranscriptomes. The 52 metatranscriptomes span a broad range of the water column at a station in the central Red Sea: four depths and 13 time points over a 48-h period (every 4 h). Histograms show the number of metatranscriptomes found in SAR11 non-RS-OGs (a), SAR11 RS-OGs (b), *Prochlorococcus* non-RS-OGs (c), and *Prochlorococcus* RS-OGs (d). Heatmaps (inset) show the density of OGs based on OG size (number of total copies across the SAGs) and the number of metatranscriptomes an OG is found in. RS-OGs were more likely than other OGs to be expressed in one or more samples, and non-RS-OGs that were expressed were more likely to be expressed in a large number of samples.

queried the data set to determine what fraction of all OGs and what fraction of RS-OGs could be found outside the Red Sea. If RS-OGs represent gene content endemic to the Red Sea, we would expect to find them absent from metagenomes from other regions. Our approach was complementary to a recent study that analyzed the global metapangenome of *Prochlorococcus* in the *Tara* metagenomes, showing the distributions of gene clusters (OGs) with strain-level resolution across the *Tara* samples (35). In the work here, we employed rarefaction and ordination techniques, with a particular focus on RS-OGs.

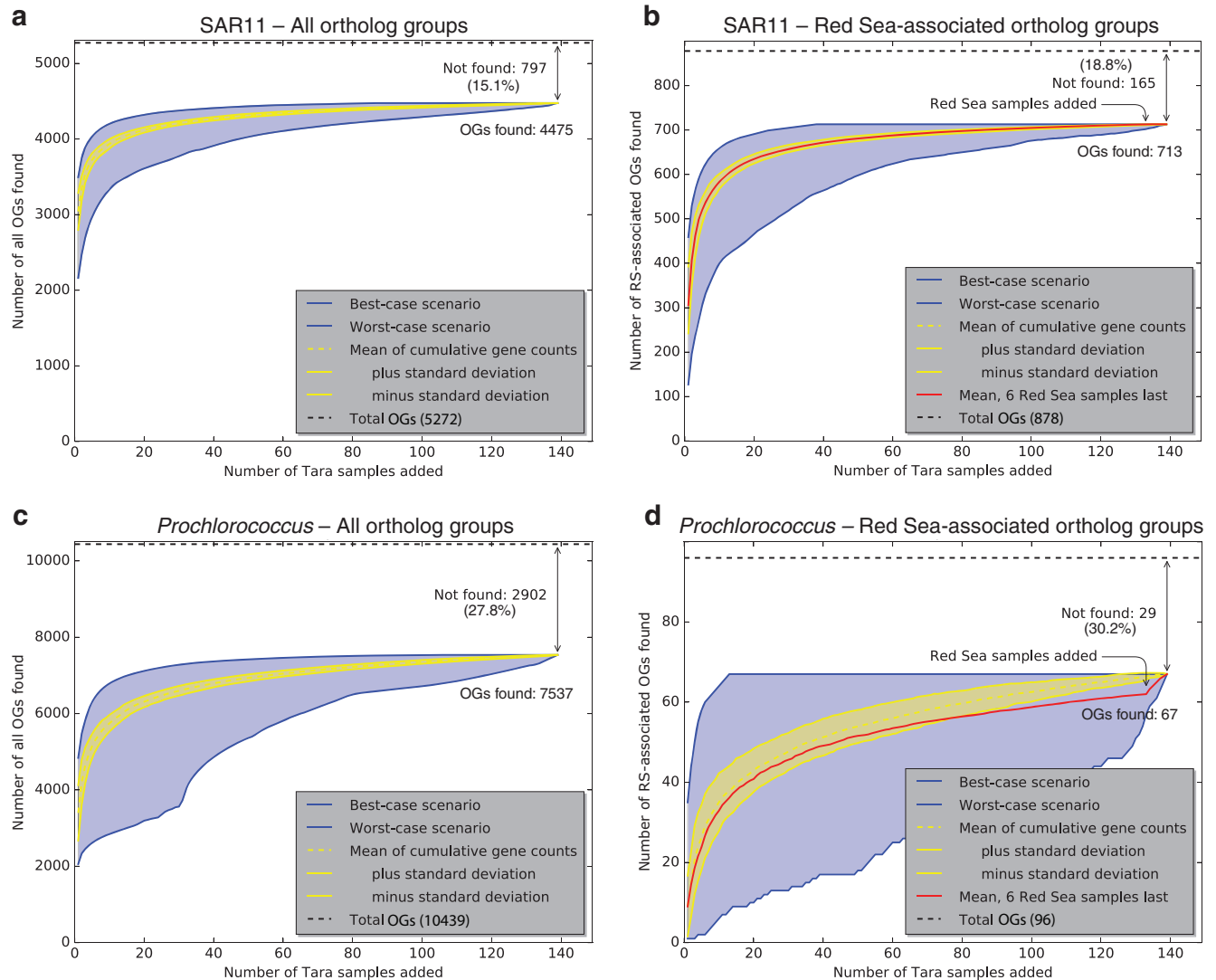


FIG 4 Rarefaction analysis showing the proportion of all OGs (a and c) and RS-OGs of SAR11 (b and d) and *Prochlorococcus* observed in *Tara* Oceans metagenome samples. Curves show the cumulative number of OGs observed in *Tara* Oceans samples (E value of $<1e-5$) as more samples are added. Yellow lines show the averages \pm standard deviations from 1,000 permutations of randomly added samples. Blue lines show the best-case scenario (each sample added is that with the most number of new OGs observed) and worst-case scenario (each sample added is that with the fewest number of new OGs observed). Red lines show the means from 1,000 permutations of randomly added samples but with Red Sea samples (031_SRF_0.22-1.6, 032_DCM_0.22-1.6, 032_SRF_0.22-1.6, 033_SRF_0.22-1.6, 034_DCM_0.22-1.6, and 034_SRF_0.22-1.6) added last. As more *Tara* metagenome samples are added to the analysis, the number of new OGs identified approaches a plateau where new samples do not reveal many new OGs. The same is true with RS-OGs, even when samples from the Red Sea are added last, with the exception of 5 *Prochlorococcus* OGs (proch20367, proch20368, proch20390, proch20423, and proch20438).

The presence or absence of SAR11 and *Prochlorococcus* orthologs in *Tara* Oceans prokaryote-fraction metagenomes (Files S7 and S8) was plotted as rarefaction curves (Fig. 4). *Tara* Oceans metagenomes were added randomly one by one, and the fraction of SAR11 and *Prochlorococcus* OGs found was tallied and plotted. The rarefaction curves show the averages \pm standard deviations from 1,000 permutations. They also show the best-case (and worst-case) scenarios, that is, the fraction of OGs found if each new metagenome adds the most (or fewest) new OGs. Between 70 and 85% of OGs could be found in one or more *Tara* Oceans metagenomes (Fig. 4), and in the best-case scenarios it took at most ten metagenomes to find 90% of these OGs (Table S3). The percentage of OGs not found (15 to 30%) was independent of whether they were Red Sea associated or not. This result, combined with the rarefaction analysis, suggests these OGs are unlikely to be found in the *Tara* samples with deeper sequencing. It is possible that some OGs are rare and/or divergent enough to be undetectable with the current methodological approach.

Across the 139 *Tara* Oceans prokaryote-fraction metagenomes, we found 84.9% (4,475/5,272) of all SAR11 OGs in one or more metagenomes (leaving 15.1% not found) (Fig. 4a) and 72.2% (7,537/10,439) of all *Prochlorococcus* OGs in one or more metagenomes (leaving 27.8% not found) (Fig. 4c). In the best-case scenarios, it took only 5 metagenomes to find 90% of the found SAR11 OGs and 50 metagenomes to find 99%; it took only 10 metagenomes to find 90% of the found *Prochlorococcus* OGs and 60 metagenomes to find 99% (Table S3). The fractions of OGs found were similar for RS-OGs, where 81.2% (713/878) of SAR11 OGs were found (leaving 18.8% not found) (Fig. 4b) and 69.8% (67/96) of *Prochlorococcus* OGs were found (leaving 30.2% not found) (Fig. 4d). That is, RS-OGs were about as likely to be found across the World Ocean as non-RS-OGs. For both SAR11 (Fig. S5a) and *Prochlorococcus* (Fig. S5b), considering the number of *Tara* metagenomes in which each OG was found, RS-OGs were less likely to be found in a large fraction of metagenomes than all OGs. This is not surprising: the set of non-RS-OGs contains all of the core OGs, which would be expected to be found in most if not all samples.

To evaluate whether *Tara* Red Sea metagenomes contained any RS-OGs not already found in the non-Red Sea metagenomes, we tested scenarios where the Red Sea metagenomes were added last in the rarefaction analysis. There was no change in the mean curve of cumulative SAR11 OGs found when the six *Tara* Red Sea metagenomes were added last (Fig. 4b): all of the SAR11 RS-OGs could be found without examining the Red Sea metagenomes. In contrast, there were five *Prochlorococcus* RS-OGs that were added to the cumulative total when the *Tara* Red Sea metagenomes were added last (Fig. 4d). These five OGs, all with unknown function, represent a small fraction of the total *Prochlorococcus* pangenome (10,439 OGs total). Given the available genomes, this study may have uncovered a small set of OGs (Table S2) that possibly reflect gene content endemic to or generally associated with Red Sea environmental conditions, and this marks an area for further research. In light of this metagenomic analysis, however, it appears that the putative RS-OGs provide a relatively minor contribution to the whole and that these new SAR11 and *Prochlorococcus* genomes from the Red Sea generally reflect global pangenomes.

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

We used the tables of OG counts in the 63 *Tara* surface (SRF) prokaryote-fraction metagenomes to do PCA ordination on the *Tara* metagenomes (Fig. 5; top OGs driving separation among the metagenomes are provided in Table S4). SAR11 OG composition (Fig. 5a) was not obviously structured by temperature differences in the temperate and tropical ranges, although Red Sea samples clustered together, and polar samples were separate from the others. *Prochlorococcus* OG composition (Fig. 5b), however, was structured by temperature differences in the temperate and tropical ranges. The four Red Sea samples were split, with two samples clustering with the warm samples and two samples with the cooler samples. These Red Sea samples are positioned where they would be expected based on temperature: the two southern samples (latitudes 18.4°N and 22.0°N) were warmer (temperature, 27.6°C and 27.3°C) and clustered with other warm/tropical samples (Fig. 5b, left side of PC1); the two northern samples (latitudes 23.36°N and 27.16°N) were cooler (temperature, 25.8°C and 25.1°C) and clustered closer to the cool/temperate samples (Fig. 5b, right side of PC1). Note these temperatures are lower than average for Red Sea surface waters because the *Tara* Red Sea samples were collected in winter (January); in contrast, the Red Sea samples in the World Ocean Database (described above) were collected in spring (April). Given that temperature tolerances generally lack known genetic markers (36), these data suggest an area for future investigation.

In summary, the analysis of *Prochlorococcus* and SAR11 OGs in *Tara* Oceans metagenomes shows that (i) most Red Sea-associated OGs are actually widely distributed across the World Ocean, not endemic to the Red Sea, and (ii) OG distribution patterns

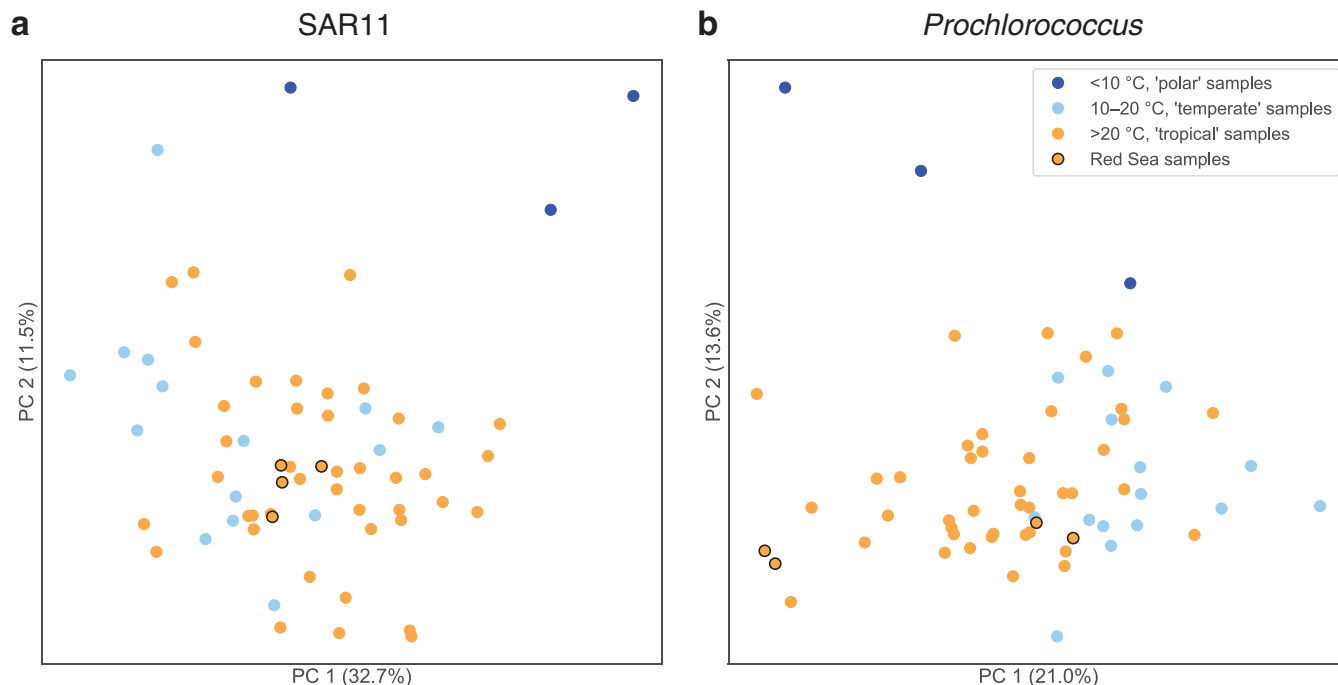


FIG 5 Principal component analysis of *Tara* Oceans surface samples by the abundance of SAR11 (a) and *Prochlorococcus* (b) OGs. The ordination shows the similarity of *Tara* Oceans samples to each other along the first two principal components. Samples are colored by *Tara* temperature categories: polar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

as a whole, taking relative abundance into account, place the Red Sea on a continuum with other seas, with patterns explained by environmental factors, including temperature. Supporting this idea, differences in the relative abundance of OGs, with physicochemical properties covarying with OG functions, have been observed among the North Pacific, Sargasso Sea, Mediterranean Sea, and Red Sea in previous comparative metagenomics studies (11, 12). Despite the Red Sea existing at the periphery of multiple physicochemical parameters in the World Ocean, its distinctiveness may best be revealed by the relative abundance of OGs rather than in the wholesale presence or absence of OGs. In addition to this general pattern, this effort also identified a small set of putative and nonhypothetical proteins that warrant further ecological and biotechnological study.

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

How marine microbes are able to respond to a changing ocean will be critical to understanding the future biosphere of planet Earth. At the population and community levels, the cosmopolitan distribution of genetic functions may confer an advantage, enabling marine microbial populations and communities, as a whole, to rapidly respond and adapt to changing ocean conditions. Here, we generally considered the Baas Becking hypothesis (“everything is everywhere, but the environment selects”) from the perspective of gene ortholog groups (“every OG is everywhere, but the environment

selects"). The overall data analysis lends support to the Baas Beeking hypothesis as applied to OGs. We described a small set of OGs that may be related to Red Sea environmental conditions and that mark areas for further investigation. However, the overall analysis was not consistent with endemism as a primary feature. Instead, we found Red Sea OGs to be nearly as prevalent across global ocean metagenomes as in Red Sea metagenomes. This view was supported by analysis of OG relative abundance rather than absolute presence/absence of OGs. Perhaps OGs are present but undetectable in a region, and they become detectable after OG frequencies increase in response to environmental conditions (via the growth of cells containing those OGs). Therefore, genomic adaptations in a given ocean region may not simply reflect the presence of OGs unique to a region but rather the relative abundance of generally cosmopolitan OGs.

MATERIALS AND METHODS

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

Seawater samples for metatranscriptomics were taken 3 to 5 March 2013 from an open-ocean site in the Red Sea (Kebrit Deep; 24.7244°N, 36.2785°E [referred to as "Station 3" in the Sequence Read Archive]). To obtain broad coverage of the water column by both time of day and water depth, one sample per depth was collected every 4 h over a 48-h period at four depths: surface (10 m), below the mixed layer (40 m; bottom of mixed layer was 35 m), chlorophyll maximum (75 m), and oxygen minimum zone (420 m). For each time point and depth, 1 liter of seawater was filtered using a peristaltic pump with two in-line filters in series: a 1.6- μm GF/A prefilter (Whatman) and then a 0.22- μm Sterivex filter (Millipore). RNAlater (Qiagen) was added immediately to fill the dead space of the Sterivex filter, which was then flash frozen in liquid nitrogen and stored at -80°C .

Nucleic acid extraction and amplification. Single bacterioplankton cells in the preserved samples were flow sorted, whole-genome amplified (MDA, or multiple displacement amplification), and PCR screened at the Bigelow Laboratory Single Cell Genomics Center (SCGC; Boothbay Harbor, ME, USA), by following previously described protocols (37), with SYTO-13 nucleic acid stain used to stain cells for flow sorting. SAG identification was carried out with SCGC protocol S-102 for bacteria using 16S rRNA primers 27F and 907R (38, 39). Totals of 21 and 5 cells were identified from 16S PCR screening and subjected to a second round of MDA before sequencing.

The RNA extraction protocol for metatranscriptomics was adapted from references 40–42. After expelling RNAlater from the Sterivex filter, 2 ml lysozyme solution (1 mg/ml in lysis buffer, containing 40 mM EDTA, 50 mM Tris, pH 8.3, 0.73 M sucrose) was added and then filter incubated at 37°C with rotation for 45 min. Proteinase K solution (50 μl at 20 mg/ml; Qiagen/5PRIME) and SDS solution (100 μl at 20%) were added, and then the filter was incubated at 55°C with rotation for 2 h. Lysate was expelled to a separate tube; meanwhile, 1 ml lysis buffer was added to the filter to wash at 55°C for 15 min. The two lysates were pooled, to which was added 1.5 ml absolute ethanol. RNA was then extracted from this solution using the RNeasy Protect bacterial midi kit (Qiagen). RNA was eluted with two volumes of RNase-free water. RNA sample was concentrated using a speed vacuum, from 250 μl to 60 μl . To this volume we added DNase (1 μl Ambion TURBO DNA free, 6 μl 10 \times buffer, 60 μl RNA) and incubated at 37°C for 30 min. This solution was purified using the RNeasy MinElute cleanup kit (Qiagen) and eluted with RNase-free water. The final yield was 1 to 2 ng total RNA. Total RNA was amplified using the C&E Version ExpressArt bacterial mRNA amplification nano kit, which preferentially amplifies mRNA [independent of poly(A) tail] and selects against rRNAs. A single round of amplification was performed on 2 to 4 ng of total RNA, which yielded about 10 μg final amplified RNA.

Nucleic acid sequencing. For single-cell genome sequencing, genomic library preparation with Illumina TruSeq and sequencing with Illumina GAIIx and Illumina HiSeq 2000 was done at the KAUST Bioscience Core Laboratory, generating paired 105-bp reads. The assembled contigs (assembly methods are described below) are available from NCBI with accession numbers [PRJEB9287](https://www.ncbi.nlm.nih.gov/assembly/PRJEB9287) (BioProject) and [SAMEA3368552](https://www.ncbi.nlm.nih.gov/assembly/SAMEA3368552) to [SAMEA3368577](https://www.ncbi.nlm.nih.gov/assembly/SAMEA3368577) (BioSample) and from Integrated Microbial Genomes (43) with accession numbers 2630968236, 2630968238 to 2630968254, 2630968277 to 2630968281, and 2630968285 to 2630968287.

For metatranscriptomics, sequence data were processed as described in reference 20. Amplified RNA was used to construct sequencing libraries using the TruSeq stranded RNA LT sample preparation kit (Illumina) according to the manufacturer's protocol. Libraries were paired-end sequenced with the Illumina HiSeq 2000 platform (2×100 bp). A link to the raw FASTQ sequences is provided at the end of Materials and Methods. Low-quality reads and sequencing adapters were removed using Trimmomatic v0.32 (44). Sequence reads shorter than 50 bp were discarded. Bowtie 2 v2.2.4 (45) was used to identify and remove PhiX contamination sequences. The remaining sequences were error corrected using the BayesHammer algorithm (46) implemented in SPAdes v3.5.0 (47), followed by removal of putative rRNA gene transcripts with SortMeRNA v2.0 (48).

Genome assembly and annotation. *De novo* assemblies were generated using CLC Genomics Workbench 4.9. The genomes were assembled independently, and unless otherwise specified, the

following applies to all of the SAGs. The reads were first imported and quality trimmed with a limit of 0.01. They were then assembled using CLC's *de novo* assembler with a word size (*k*-mer) of 64 and with the minimum/maximum of the insert size set to 100/1,000 bp. Only those contigs greater than 200 bp in length were included in downstream analyses. The reads were mapped to the consensus sequence of the assembled contigs using CLC's default parameters but with the length fraction set to 1.0 and the similarity set to 0.95.

Assembled SAG contigs were ordered and oriented relative to SAR11 HTCC1062 (NC_007205.1) or *Prochlorococcus* MIT 9202 (NZ_DS999537) using ABACAS 1.3.1 (49). The ordered sequences were then imported into GAP4 (50), and additional joins were made between overlapping contigs if conserved synteny supported the arrangement. To identify and remove possible contaminating sequences from the assemblies, each contig was retained only if it met one or both of the following criteria: (i) the contig was binned into a bin annotated as SAR11 or *Prochlorococcus* using Metawatt 3.5 (22), using the medium bin level, with a minimum bin size of 50 kbp and minimum contig size of 500 bp; (ii) the contig had a top-10 BLASTN hit against the GenBank nucleotide database, with an E value of $<1e-5$ to SAR11 or *Prochlorococcus*.

Prediction of gene open reading frames (ORFs) and functional annotation of SAGs were performed by the RAST web service (51) with FIGfam, release 59.

OG clustering. Predicted proteins from SAGs were clustered with proteins from published cultured and SAG genomes (see File S1 in the supplemental material) into OGs using OrthoMCL 2.0 (52). OrthoMCL configuration settings were a percentMatchCutoff of 50 and evaluateExponentCutoff of -5 . This yielded 5,272 SAR11 OGs and 10,439 *Prochlorococcus* OGs. After OrthoMCL clustering, OGs were assigned as core and noncore based on copy number in the non-Red Sea, cultured (non-SAG) genomes: core OGs are those found at least once in each of the non-Red Sea, cultured genomes, and noncore OGs are those not found in at least one of the non-Red Sea, cultured genomes. Among SAR11, there were 683 core OGs and 4,589 noncore OGs. Among *Prochlorococcus*, there were 1,152 core OGs and 9,287 noncore OGs. Protein sequence identifiers and FASTA sequences for each OG have been archived at <https://zenodo.org> (<https://doi.org/10.5281/zenodo.2634561>).

Estimation of genome completeness. Completeness of SAGs was assessed using two methods. First, completeness was assessed using single-copy core OGs, i.e., those OGs found once and only once in each complete genome based on the OrthoMCL clusters (analyzed separately for SAR11 and *Prochlorococcus*). Completeness was calculated as the number of core orthologs present in each SAG out of 649 SAR11 or 1,144 *Prochlorococcus* single-copy core OGs. Second, genome completeness of the SAGs was assessed using CheckM 1.0.13 (53) using lineage-specific workflow (lineage_wf) with database file checkm_data_2015_01_16.tar.gz downloaded from https://data.ace.uq.edu.au/public/CheckM_databases; CheckM was also used to estimate genome redundancy (called contamination in CheckM). For comparison, CheckM completeness and redundancy were calculated for the reference genomes used in this study (Table S1).

Genome taxonomy and phylogenetics. A total of 89 SAR11 and 96 *Prochlorococcus* shared single-copy orthologous genes were identified using the GET_HOMOLOGUES software (54). Amino acid sequences translated from gene sequences were aligned using the MAFFT software (55). These alignments were concatenated, sites with gaps were deleted, and the concatenated data were partitioned using the PartitionFinder software (56) to account for variations of evolutionary processes among gene families. With the Bayesian information criterion (BIC) statistic, a 16-partition framework was chosen to optimally describe the variability, in which the LG rate matrix with gamma distribution of rate variation (LG+G) was selected for 15 partitions and the VT rate matrix with gamma distribution of rate variation (VT+G) was selected for the remaining partition. This partition model was used in the maximum-likelihood phylogenomic construction using the RAxML software (57).

Ordination of SAGs and genomes using *k*-mer composition and ortholog composition. SAGs and reference genomes (Table S1) were analyzed using PCA of nucleotide composition and OG composition. Nucleotide composition of the SAGs and reference genomes (SAR11 and *Prochlorococcus* scaffolds of >200 kbp from Integrated Microbial Genomes; <https://img.jgi.doe.gov>) was determined as 6-nucleotide words or *k*-mers (6-mers). *k*-mer frequencies were calculated using Jellyfish 2.2.5; the main command used was jellyfish count $-m$ 6 $-t$ 8 $-s$ 1 M. This resulted in a table of 6-mer frequencies in the SAGs and genomes, one table each for SAR11 and *Prochlorococcus*. OG composition was derived from tables of OrthoMCL clusters, which, as the SAGs had variable levels of completeness and gene counts (Table 1), were randomly subsampled so that all genomes had the same number of gene counts in the tables; five replicate subsamples produced very similar results (only one SAR11 subsample and one *Prochlorococcus* subsample are shown). The number of OG counts subsampled was chosen to balance the number of OG counts with the number of genomes retained (less complete SAGs were excluded): the OG composition tables (with counts of 5,272 unique SAR11 OGs and 10,439 unique *Prochlorococcus* OGs) were subsampled down to 800 gene counts per SAR11 SAG (keeping 12 of 21 SAGs) and 1,400 gene counts per *Prochlorococcus* genome (keeping 5 of 5 SAGs). Prior to PCA, a pseudocount of 1 was added to *k*-mer and OG count tables to account for zero values; *k*-mer counts were then converted to relative abundances for each genome (unnecessary for OG counts because of the subsampling procedure); *k*-mer relative abundances were then standardized to *z* scores (not done for OG counts because this reduced the resolving power of PCA). PCA was then performed using the Scikit-Learn function `sklearn.decomposition.PCA` (58).

Mapping of metatranscriptomic reads to OGs. The quality-filtered mRNA reads from the 52 samples were mapped against the SAGs using Bowtie 2 (45) with default settings. Each read mapping above the threshold was assigned to exactly one gene in a SAG contig. The resultant read counts were normalized based on the FPKM (fragments per kilobase of gene per million mapped reads) metric.

Per-sample FPKM counts for each gene were then summed by OGs, resulting in per-sample FPKM counts for each OG. For downstream analysis, counts were converted to a simple presence/absence measure: if any gene belonging to the OG had one or more mapped transcript, that OG was marked as present in that sample.

Detection and rarefaction analysis of OGs in Tara Oceans metagenomes. A set of 139 prokaryote-enriched Tara Oceans metagenomic gene files (34) was downloaded from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena>; accession numbers ERZ096909 to ERZ097150). Each file contains nucleotide sequences for genes predicted on Tara Oceans metagenomic contigs that were assembled from shotgun sequencing reads from individual Tara Oceans samples. The prokaryote fraction was 0.22 to 1.6 μm for stations 004 to 052 and 0.22 to 3 μm for stations 056 to 152; the environmental features of the samples were indicated as SRF (surface), MIX (mixed layer), DCM (deep chlorophyll maximum), and MES (mesopelagic zone). The metagenomic gene sequences were queried against a database of translated proteins from the SAGs and genomes with DIAMOND 0.8.26 (59) using the program blastx with parameters $-p\ 40 -k\ 25 -e\ 1e-3$. The top hit (SAG or genome protein sequence) for each Tara gene sequence (E value of $<1e-5$) was retained. E value cutoffs of $1e-10$ and $1e-15$ were also tested, which showed the same trends as an E value of $<1e-5$ but with fewer total OGs identified. Counts of the number of times each protein was a top hit were then summed across each OG. This resulted in a table of OGs by samples where each OG was either present (at least one constituent protein was a top hit at least once) or absent in each sample. These presence/absence tables (one for SAR11, one for *Prochlorococcus*) were used to generate rarefaction curves: samples were added one-by-one randomly (1,000 permutations), and the cumulative number of OGs found was recorded.

Ordination of Tara Oceans metagenomes by OG composition. OG counts (total, not presence/absence) in Tara Oceans surface (SRF) sample metagenomes were used for ordination using PCA. Prior to PCA, a pseudocount of 1 was added to OG count tables to account for zero values; counts were then converted to relative abundances for each metagenome; OGs with an average relative abundance across all metagenomes of less than 0.0001 (0.01%) were removed; relative abundances were then standardized to z scores. PCA was then performed using the Scikit-Learn function `sklearn.decomposition.PCA` (58).

World ocean temperature and salinity data. Surface temperature and salinity data (WOD13_ALL_SUR_OBS) from the World Ocean Database 2013 (<https://www.nodc.noaa.gov/OC5/WOD13/>) were downloaded from the Research Data Archive at the National Center for Atmospheric Research (<https://rda.ucar.edu/datasets/ds285.0/>).

Data availability. Assembled SAG contigs have been deposited in NCBI GenBank with BioProject number PRJEB9287. Raw metatranscriptomic FASTQ sequences have been deposited in the NCBI Sequence Read Archive with BioProject number PRJNA289956.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/AEM.00369-19>.

SUPPLEMENTAL FILE 1, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.04 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.6 MB.

SUPPLEMENTAL FILE 5, XLSX file, 0.7 MB.

SUPPLEMENTAL FILE 6, XLSX file, 4.1 MB.

SUPPLEMENTAL FILE 7, XLSX file, 2.4 MB.

SUPPLEMENTAL FILE 8, XLSX file, 3.7 MB.

SUPPLEMENTAL FILE 9, PDF file, 2.6 MB.

ACKNOWLEDGMENTS

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

This paper was prepared by L.R.T. under award NA06OAR4320264 06111039 to the Northern Gulf Institute by NOAA's Office of Oceanic and Atmospheric Research, U.S. Department of Commerce. This work was supported by the NOAA Atlantic Oceanographic and Meteorological Laboratory 'Omics Program.

REFERENCES

1. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. 2008. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A* 105:3805–3810. <https://doi.org/10.1073/pnas.0708897105>.
2. Waldbauer JR, Rodrigue S, Coleman ML, Chisholm SW. 2012. Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PLoS One* 7:e43432. <https://doi.org/10.1371/journal.pone.0043432>.

3. Coleman ML, Chisholm SW. 2007. Code and context: Prochlorococcus as a model for cross-scale biology. *Trends Microbiol* 15:398–407. <https://doi.org/10.1016/j.tim.2007.07.001>.
4. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. *Nature* 551:45–50. <https://doi.org/10.1038/nature24287>.
5. Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304. <https://doi.org/10.1038/35012500>.
6. Rocop G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–1047. <https://doi.org/10.1038/nature01947>.
7. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. 2009. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* 4:e6864. <https://doi.org/10.1371/journal.pone.0006864>.
8. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* 3:e00252-12. <https://doi.org/10.1128/mBio.00252-12>.
9. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344:416–420. <https://doi.org/10.1126/science.1248575>.
10. Luo H, Thompson LR, Stingl U, Hughes AL. 2015. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol* 32:2738–2748. <https://doi.org/10.1093/molbev/msv149>.
11. Coleman ML, Chisholm SW. 2010. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci U S A* 107:18634–18639. <https://doi.org/10.1073/pnas.1009480107>.
12. Thompson LR, Field C, Romanuk T, Ngugi D, Siam R, El Dorry H, Stingl U. 2013. Patterns of ecological specialization among microbial populations in the Red Sea and diverse oligotrophic marine environments. *Ecol Evol* 3:1780–1797. <https://doi.org/10.1002/ece3.593>.
13. Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH, Ackerman M, Moore LR, Meisel JD, Sher D, Thompson LR, Campbell L, Martiny AC, Chisholm SW. 2015. Physiology and evolution of nitrate acquisition in Prochlorococcus. *ISME J* 9:1195–1207. <https://doi.org/10.1038/ismej.2014.211>.
14. Edwards FJ. 1987. Climate and oceanography, p 45–68. *In* Edwards AJ, Head SM (ed), *Key environments: Red Sea*. Pergamon, Oxford, United Kingdom.
15. Post AF. 2005. Nutrient limitation of marine cyanobacteria, p 87–107. *In* Huisman J, Matthijs HCP, Visser PM (ed), *Harmful cyanobacteria*. Springer, New York, NY.
16. Thompson LR, Williams GJ, Haroon MF, Shibl A, Larsen P, Shorenstein J, Knight R, Stingl U. 2017. Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *ISME J* 11:138–151. <https://doi.org/10.1038/ismej.2016.99>.
17. Baas Becking LGM. 1934. *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.
18. Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. 2013. Evidence for a persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci U S A* 110:4651–4655. <https://doi.org/10.1073/pnas.1217767110>.
19. Gonnella G, Böhnke S, Indenbirken D, Garbe-Schönberg D, Seifert R, Mertens C, Kurtz S, Perner M. 2016. Endemic hydrothermal vent species identified in the open ocean seed bank. *Nat Microbiol* 1:16086. <https://doi.org/10.1038/nmicrobiol.2016.86>.
20. Shibl AA, Ngugi DK, Talarmin A, Thompson LR, Blom J, Stingl U. 2018. The genome of a novel isolate of Prochlorococcus from the Red Sea contains transcribed genes for compatible solute biosynthesis. *FEMS Microbiol Ecol* 94:fy182. <https://doi.org/10.1093/femsec/fy182>.
21. Jimenez-Infante F, Ngugi DK, Vinu M, Blom J, Alam I, Bajic VB, Stingl U. 2017. Genomic characterization of two novel SAR11 isolates from the Red Sea, including the first strain of the SAR11 lb clade. *FEMS Microbiol Ecol* 93:fx083.
22. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. 2012. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* 3:410. <https://doi.org/10.3389/fmicb.2012.00410>.
23. Shibl AA, Thompson LR, Ngugi DK, Stingl U. 2014. Distribution and diversity of Prochlorococcus ecotypes in the Red Sea. *FEMS Microbiol Lett* 356:118–126. <https://doi.org/10.1111/1574-6968.12490>.
24. Shibl AA, Haroon MF, Ngugi DK, Thompson LR, Stingl U. 24 June 2016. Distribution of Prochlorococcus ecotypes in the Red Sea basin based on analyses of rpoC1 sequences. *Front Mar Sci*. <https://doi.org/10.3389/fmars.2016.00104>.
25. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. 2006. Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science* 311:1737–1740. <https://doi.org/10.1126/science.1118052>.
26. Ngugi DK, Stingl U. 2012. 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* 7:e50274. <https://doi.org/10.1371/journal.pone.0050274>.
27. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferreira S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW. 2007. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet* 3:e231. <https://doi.org/10.1371/journal.pgen.0030231>.
28. Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L, Roache-Johnson KH, Ding H, Giovannoni SJ, Rocop G, Moore LR, Chisholm SW. 2014. Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Sci Data* 1:140034. <https://doi.org/10.1038/sdata.2014.34>.
29. Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. *Nat Rev Microbiol* 13:13–27. <https://doi.org/10.1038/nrmicro3378>.
30. Rocop G, Distel DL, Waterbury JB, Chisholm SW. 2002. Resolution of Prochlorococcus and Synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* 68:1180–1191. <https://doi.org/10.1128/AEM.68.3.1180-1191.2002>.
31. Ngugi DK, Antunes A, Brune A, Stingl U. 2012. Biogeography of pelagic bacterioplankton across an antagonistic temperature-salinity gradient in the Red Sea. *Mol Ecol* 21:388–405. <https://doi.org/10.1111/j.1365-294X.2011.05378.x>.
32. Somero GN, Lockwood BL, Tomanek L. 2016. *Biochemical adaptation: response to environmental challenges, from life's origins to the anthropocene*. Sinauer Associates, Sunderland, MA.
33. Haroon MF, Thompson LR, Parks DH, Hugenholtz P, Stingl U. 2016. A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci Data* 3:160050. <https://doi.org/10.1038/sdata.2016.50>.
34. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P, Boss E, Bowler C, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D. 2015. Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348:1261359–1261359. <https://doi.org/10.1126/science.1261359>.
35. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* 6:e4320. <https://doi.org/10.7717/peerj.4320>.
36. Hickey DA, Singer GA. 2004. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol* 5:117. <https://doi.org/10.1186/gb-2004-5-10-117>.
37. Stepanauskas R, Sieracki ME. 2007. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci U S A* 104:9052–9057. <https://doi.org/10.1073/pnas.0700496104>.
38. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin M, Pace NR. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* 82:6955–6959. <https://doi.org/10.1073/pnas.82.20.6955>.
39. Page KA, Cannon SA, Giovannoni SJ. 2004. Representative freshwater bacterioplankton isolated from Crater Lake, Oregon. *Appl Environ Microbiol* 70:6542–6550. <https://doi.org/10.1128/AEM.70.11.6542-6550.2004>.
40. Massana R, Murray AE, Preston CM, Delong EF. 1997. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* 63:50–56.
41. Bèjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T,

- Eisen JA, Fraser CM, DeLong EF. 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415:630–633. <https://doi.org/10.1038/415630a>.
42. Stewart FJ, Dalsgaard T, Young CR, Thamdrup B, Revsbech NP, Ulloa O, Canfield DE, DeLong EF. 2012. Experimental incubations elicit profound changes in community transcription in OMZ bacterioplankton. *PLoS One* 7:e37118. <https://doi.org/10.1371/journal.pone.0037118>.
 43. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-M, Chu K, Kyrpides NC. 2009. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics (Oxford, England)* 25:2271–2278. <https://doi.org/10.1093/bioinformatics/btp393>.
 44. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
 45. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.
 46. Nikolenko SI, Korobeynikov AI, Alekseyev MA. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 14(Suppl 1):S7. <https://doi.org/10.1186/1471-2164-14-S1-S7>.
 47. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshtkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>.
 48. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211–3217. <https://doi.org/10.1093/bioinformatics/bts611>.
 49. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics (Oxford, England)* 25:1968–1969. <https://doi.org/10.1093/bioinformatics/btp347>.
 50. Bonfield JK, Smith KF, Staden R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res* 23:4992–4999. <https://doi.org/10.1093/nar/23.24.4992>.
 51. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. <https://doi.org/10.1186/1471-2164-9-75>.
 52. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189. <https://doi.org/10.1101/gr.1224503>.
 53. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>.
 54. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl Environ Microbiol* 79:7696. <https://doi.org/10.1128/AEM.02411-13>.
 55. Katoh K, Kuma K-I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518. <https://doi.org/10.1093/nar/gki198>.
 56. Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* 29:1695–1701. <https://doi.org/10.1093/molbev/mss020>.
 57. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
 58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830.
 59. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60. <https://doi.org/10.1038/nmeth.3176>.