Systematic and Applied Microbiology 44 (2021) 126185

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: http://www.elsevier.com/locate/syapm

Description of *Candidatus* Mesopelagibacter carboxydoxydans and *Candidatus* Anoxipelagibacter denitrificans: Nitrate-reducing SAR11 genera that dominate mesopelagic and anoxic marine zones

Carlos A. Ruiz-Perez [a], Anthony D. Bertagnolli [a], Despina Tsementzi [b], Tanja Woyke [c], Frank J. Stewart [a,d,e], Konstantinos T. Konstantinidis [a,b,f,*]

[a] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[c] DOE Joint Genome Institute, One Cyclotron Road, Mail Stop 91R0183, Berkeley, CA 94720, USA
[d] Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, GA 30332, USA
[e] Department of Microbiology & Immunology, Montana State University, Bozeman, MT 59717, USA
[f] Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 10 August 2020
Received in revised form
25 November 2020
Accepted 9 December 2020

Keywords:
SAR11
OMZ
Metagenome-assembled genomes
Denitrification
Uncultivated
Genome-based classification

ABSTRACT

The diverse and ubiquitous members of the SAR11 lineage (*Alphaproteobacteria*) represent up to 30–40% of the surface and mesopelagic oceanic microbial communities. However, the molecular and ecological mechanisms that differentiate closely related, yet distinct, SAR11 members that often co-occur under similar environmental conditions remain speculative. Recently, two mesopelagic and oxygen minimum zone (OMZ)-associated subclades of SAR11 (Ic and IIa.A) were described using single-cell amplified genomes (SAGs) linked to nitrate reduction in OMZs. In this current study, the collection of genomes belonging to these two subclades was expanded with thirteen new metagenome-assembled genomes (MAGs), thus providing a more detailed phylogenetic and functional characterization of these subclades. Gene content-based predictions of metabolic functions revealed similarities in central carbon metabolism between subclades Ic and IIa.A and surface SAR11 clades, with small variations in central pathways. These variations included more versatile sulfur assimilation pathways, as well as a previously predicted capacity for nitrate reduction that conferred unique versatility on mesopelagic-adapted clades compared to their surface counterparts. Finally, consistent with previously reported abundances of carbon monoxide (CO) in surface and mesopelagic waters, subclades Ia (surface) and Ic (mesopelagic) have the genetic potential to oxidize carbon monoxide (CO), presumably taking advantage of this abundant compound as an electron donor. Based on genomic analyses, environmental distribution and metabolic reconstruction, we propose two new SAR11 genera, *Ca.* Mesopelagibacter carboxydoxydans (subclade Ic) and *Ca.* Anoxipelagibacter denitrificans (subclade IIa.A), which represent members of the mesopelagic and OMZ-adapted SAR11 clades.

Introduction

The SAR11 clade is a monophyletic group of heterotrophic *Alphaproteobacteria* that are observed in high abundances in nearly all marine biomes [38]. Depending on the oceanic region, physical conditions and season, these cells can constitute between 30 and 50% of the total bacterioplankton [74]. Their heterotrophic metabolism coupled with their abundance throughout the oceans makes them important for nutrient cycling, especially through the oxidation of dissolved organic matter [38,109]. The specific genomic and physiological characteristics that enable SAR11 dominance have been the subject of intense research for nearly 20 years [38]. These characteristics are thought to include small cell sizes, streamlined genomes with adaptations to take advantage of an ocean's dissolved organic matter (DOM), and relatively abundant transport systems to uptake nutrients under oligotrophic conditions [38,40,43]. Much of our knowledge regarding the SAR11 clade is based on isolates recovered from aerobic coastal or oligotrophic environments [39,102]. However, heterogeneity in physical and

* Corresponding author at: School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA.
E-mail address: kostas@ce.gatech.edu (K.T. Konstantinidis).

https://doi.org/10.1016/j.syapm.2021.126185
0723-2020/

chemical conditions throughout the ocean's water column is often extreme. Clear patterns of niche differentiation within the SAR11 clade along these gradients have been observed using 16S rRNA gene datasets [10], as well as genomes from isolates, single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) [44], suggesting a differential taxonomic distribution with (different) depth or environmental conditions. However, the extent to which these changes are coupled to discrete shifts in functional gene content remains less clear.

To date, nine monophyletic subclades (based on 16S rRNA gene and single-copy marker protein phylogenies) of the SAR11 clade have been recovered from marine and freshwater environments [38]. Of these, the most widely studied is subclade Ia, represented by *Candidatus* Pelagibacter ubique. First isolated in 2002 by Rappé et al. [88], *Ca.* Pelagibacter ubique has been used to study the metabolic and genomic adaptations of SAR11 members [40,96,110] and correlate them with geographic abundances estimated using 16S rRNA amplicons and microscopy abundance estimations [74]. More generally, subclades Ia, Ib, IIa and V dominate ocean surface layers with strong seasonal variability in relative abundances, while subclades Ic and IIb dominate the oceanic mesopelagic zones (200–1000 m) [38,46]. On the other hand, subclade IIIa dominates brackish, surface waters [46] and subclade IIIb dominates freshwater systems [111]. In several cases, clade abundance also correlated with environmental conditions, including DOM and phosphate concentrations [16,38,93]. An interesting pattern observed for mesopelagic subclades Ic, IIb, and the more recently proposed IIa.A (sublineage of clade IIa) is their predominance in open ocean OMZs, such as those found in the Eastern Tropical North (ETNP) and South Pacific (ETSP) [108,112]. In these OMZs, the dissolved oxygen is typically below the detection limit and therefore they are known as anoxic marine zones (AMZs) [64]. In OMZs, the increased concentrations of nitrate and nitrite make them important nutrients for microbial metabolism [113]. Indeed, SAGs coupled with metagenomics and metatranscriptomics have revealed the presence of nitrate reductase genes (*narGHJI*) in multiple SAR11 SAGs from anoxic layers of the ETNP [112]. These features had not been observed previously in SAR11 subclades associated with aerobic or surface waters. These results implicated OMZ-specific members of SAR11 as drivers of nitrate reduction to nitrite, thus, indicating an important microbial pathway for nitrogen transformation from these systems and suggesting that subtle variations in genome content could drive niche differentiation and genome adaptation. Additional analysis of the OMZ SAR11 community has revealed that the genome-aggregate average amino acid identity values (AAI) were >65% within clades vs. <60% between clades [112], which following previous findings on subclade Ic [108], suggested that the mesopelagic SAR11 subclades belonged to genera distinct from the *Pelagibacter* genus (subclade Ia) dominating surface waters.

Currently, genome sequences from isolates belonging to epipelagic (0–200 m) marine (Ia, Ib, IIIa and V) and freshwater (IIIb) subclades are publicly available [45,50,81,88,101,111]. However, there are no isolates or complete genomes from subclades dominating the oceanic mesopelagic or anoxic waters (Ic, IIa.A and IIb). The only genomic information available for mesopelagic subclades Ic and IIa.A comes from a limited number of SAGs [108,112]. Thus, significant knowledge gaps exist regarding the genomic complexity and physiology of SAR11 cells below the photic zone. Comparisons between coastal and surface genomes have shown high conservation of metabolic features. These include the requirement of reduced sulfur and inorganic phosphate for growth, the utilization of a range of organic compounds for respiration and carbon incorporation, and the presence of genes for the oxidation of diverse C1 compounds for energy production but an inability to incorporate C1 compounds into biomass, such as carbon fixation [38,102,110]. However, notable exceptions to metabolic conservation exist, including the presence of nitrate-reducing pathways in mesopelagic subclades and other small variations in gene content that differ between subclades [50,108]. Therefore, a detailed description of the poorly studied mesopelagic subclades is important to further understand their ecological breadth and functional differentiation within the SAR11 clade. In this study, subclade IIa.A was described as a novel SAR11 genus based on its almost exclusive presence in OMZs and its ability to respire nitrate, and subclade Ic was described as a second SAR11 genus that is abundant in the oxic mesopelagic zone of the water column (200–1000 m), as well as in OMZs. The descriptions are based on previously studied SAGs and newly recovered MAGs from ETNP metagenomic samples.

Materials and methods

Sampling and sequencing

Seawater samples from the ETNP OMZ were collected at stations 2 (18° 54.053 N, 108° 48.159 N), 6 (18° 54.0 N, 104° 54.0 W) and 10 (20° 37.8 N, 107° 51.8 W) from six depths: upper oxycline (30 m), lower oxycline (85 m), secondary chlorophyll maximum (100 m), secondary nitrite maximum OMZ (125 m), OMZ core (300 m), and sub-OMZ (800 m) from June 13–28, 2013, during the Oxygen Minimum Zone Microbial Biogeochemistry Expedition (OMZoMBiE) cruise (R/V Horizon) [112]. Additional samples were collected from May 19 – June 2, 2014, at stations 6 (oxycline – 60 m; lower oxycline – 68 and 80 m; OMZ – 100, 120, 150, 200, and 400 m; Sub-OMZ – 800 and 1000 m) and stations F14, F10, and F12 (oxycline – 40, 48, 70, and 75 m; OMZ – 90, 95, 140, and 150 m; sub-OMZ – 2600 m). Microbial biomass in the 0.2–1.6 μm biomass size fraction was collected by filtration and used for DNA extraction as described in [112]. All environmental measurements for these samples – including measurements of temperature and salinity, as well as concentrations of dissolved oxygen, nitrate and nitrite – were previously reported [35]. In total, DNA extracts representing 34 samples were sequenced at the Joint Genome Institute (JGI) using an Illumina HiSeq 2500 sequencer with a 2 × 150 paired-end library. Two ETNP samples (station F10 – 150 m and station 6–400 m) were also sequenced using PacBio long reads at the JGI, processed, and quality checked using JGI's pipeline. Accession numbers and metadata associated with all datasets used in this study are detailed in Table S1.

Catalyzed reporter deposition fluorescence in situ hybridization (CARD-FISH)

CARD-FISH samples were collected and preserved according to previous sampling protocols [42]. Briefly, seawater was collected in 1 L acid-washed polycarbonate bottles, and formaldehyde was added (1% final concentration). Fixed seawater was then incubated at 4 °C for 24 h. Sample volumes of 10 mL were then vacuum-filtered through 0.2 μm-pore size 25 mm-diameter hydrophilic polycarbonate membrane filters (Millipore, GTTP02500). Filters were subsequently washed with DI-water, placed in glass Petri dishes for 2 h to dry, and finally stored at -20 °C until further processing. For visualization and cell size estimation, cells were labeled using CARD-FISH, as previously described [84,97,111]. A 5' HRP-labeled probe specific for subclade IIa.A was designed with the PROBE_DESIGN tool in ARB [67] based on the sequences of the 16S rRNA phylogenetic tree shown in Fig. 1, which included representatives of subclade IIa.A and other SAR11 subclades. Probe specificity was further tested using SILVA's TestProbe [87] and hybridization conditions were determined using MathFISH [121]. In this regard, the 16S rRNA-based abundance of 97% for the populations match-

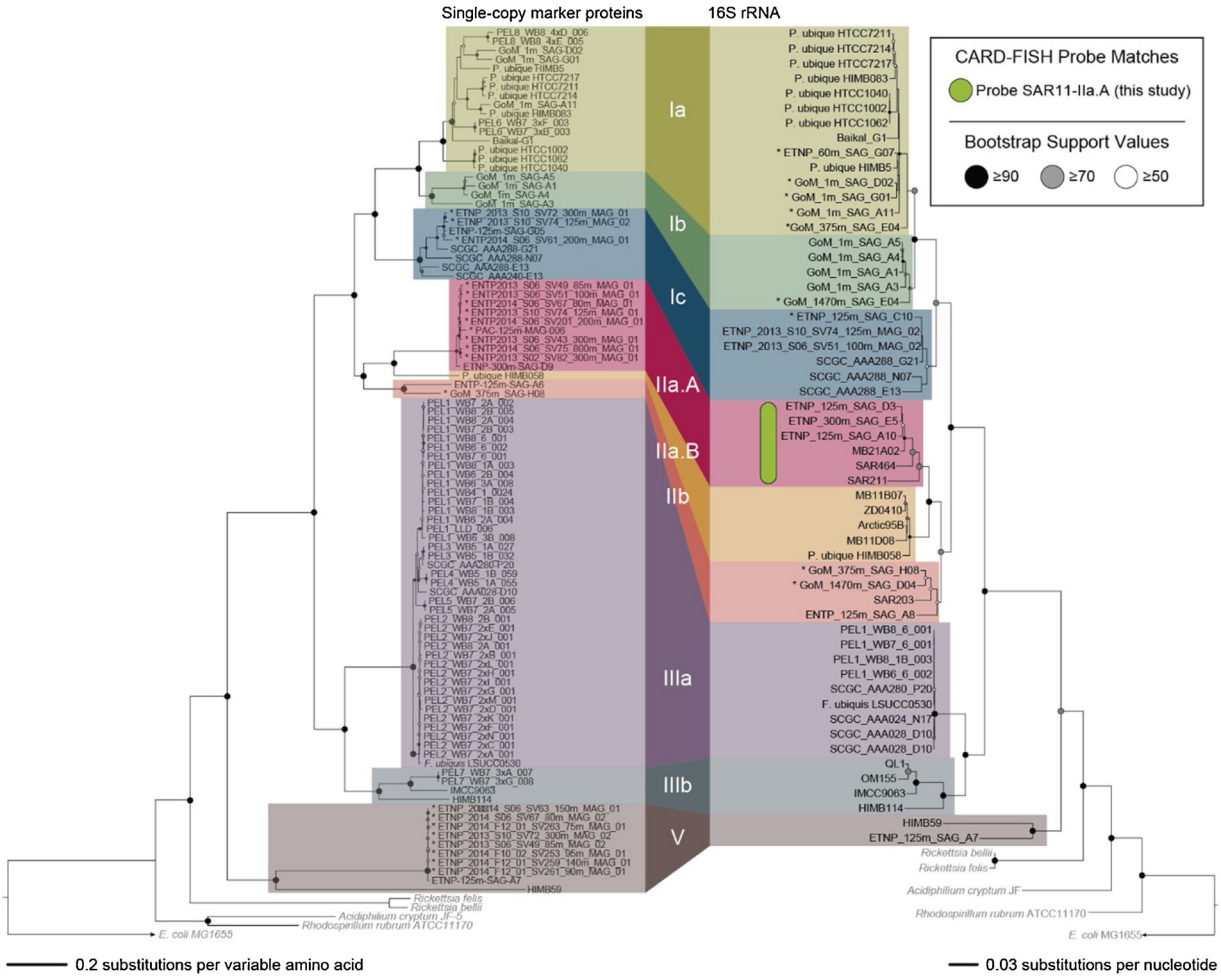

Fig. 1. Phylogeny of SAR11 subclades. Phylogenetic reconstruction of SAR11 genomes based on single-copy marker proteins ($n = 71$) (left) and 16S rRNA gene sequences (right). The protein marker tree was built using 98 complete (or nearly complete) SAR11 genomes capturing all known subclades (to date). The 16S rRNA gene tree was constructed using a representative collection of full-length sequences (>1 kbp) from clones and genomes encompassing the same subclades as in the protein marker tree. The CARD-FISH probe is specificity indicated in the 16S rRNA gene tree by the green bar next to the subclade it targets (IIa.A). Genomes and 16S rRNA gene sequences recovered in this study are marked with asterisks. SCGC – Single Cell Genomics Center, SAG – Single Amplified Genome, MAG – Metagenome Assembled Genome.

ing the two most prevalent mismatches with 88% identity were below 1% in all ETNP OMZ samples, which contrasted with approximately 23% for SAR11 populations. The sequence of the probe and the subclades it targets together with recommendations for the use of competitor probes are shown in Figure S7. The probe (5' – CAGAAAGTTGCCTTCGCT – 3') was synthesized by Integrated DNA Technologies (Iowa, USA). Finally, hybridization conditions were similar to those described in [111], but with a formamide concentration of 21% and incubation at 46 °C for 3 h. DAPI staining was performed using a 0.2 $\mu g\,mL^{-1}$ DAPI solution on filters and then washed in MilliQ H_2O followed by 100% ethanol. Finally, the filters were air-dried (in darkness) and viewed with a Zeiss Axio Observer D1 inverted phase-contrast fluorescence microscope using DAPI (Zeiss filter set 49) and GFP (filter set 41012, Chroma, USA) filters.

Recovery of single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs)

We attempted to compile a comprehensive collection of publicly available SAR11 reference genomic sequences representing all SAR11 subclades from previously published studies that included isolates ($n = 12$), MAGs ($n = 47$), and SAGs ($n = 27$). Additionally, 13 publicly available SAGs obtained from ETNP (60, 125 and 300 m) and GoM (1, 375 and 1470 m) samples [112] identified from the JGI genome database were included, and their taxonomy was confirmed using MiGA [91]. Finally, 23 SAR11 MAGs were obtained from the present study as described below. Accession numbers and metadata associated with reference isolates, SAGs, and MAGs are detailed in Table S2.

A supervised, targeted assembly and binning approach was performed to recover new MAGs belonging to mesopelagic and OMZ-dominating subclades Ic and IIa.A. Briefly, publicly available datasets from the ETNP, ETSP, and GoM were downloaded and quality checked using FaQCs [66] with a PHRED score cutoff of 25 and minimum fragment length of 50 bp (Table S1). Individual assemblies of all ETSP and GoM metagenomic datasets were performed using SPADES v.3.1.0 [4] with the – *meta* flag for metagenomes. In addition, ETNP datasets were co-assembled in combination with PacBio long reads from two ETNP datasets (150 and 400 m, Table S1) using the – *pacbio* flag. For each assembly, contigs longer than 3 kbp were binned using MaxBin v2.2.7 [119], MetaBAT v2.12.1 [51], and CONCOCT v1.1.0 [1] followed by MAG refinement using DASTool [98]. The resulting MAGs were visually inspected using Anvi'o v.6.1 [32]. Completeness and redundancy estimates for each MAG were determined using single-copy gene (SCG) datasets with universally present marker genes in CheckM v1.0.13 [82] and two additional commonly used SCG datasets included in MiGA (Table S2) [27,65,91]. A MAG was excluded from all further analyses if

its redundancy estimate was above 10% in any of the three SCG datasets used.

Selection of genomospecies representatives

Sequence relatedness in the form of genome-aggregate average nucleotide identity (gANI) and AAI between MAGs, SAGs, and reference SAR11 genomes (Tables S2, S5, and S6) was calculated using FastANI v1.2 [49] and the *aai.rb* script from the enveomics collection [90] with default parameters, respectively. Genomes with completeness above 50% in at least two out of three SCG datasets used to estimate completeness were clustered into *genomospecies* defined as genomes with ANI ≥95% [111]. For this, clusters of genomes with ANI values above 95% between themselves were identified using the Markov Clustering algorithm (MCL) [114] (Tables S2 and S5). A genome representative was selected for each genomospecies; when available, an isolate genome was selected as the representative, otherwise, the MAG or SAG with the best available genome quality was selected as the representative.

Single-copy marker protein and 16S rRNA gene phylogenetic reconstruction

The evolutionary relationships between SAR11 MAGs/SAGs and reference genomes (Table S2) were assessed using a set of 71 universally present single protein marker genes derived from previous studies [32,65,125]. Briefly, proteins predicted from each MAG were searched against the set of proteins encoded by these marker genes using HMMER v3.3 [28]. Only genomes with at least 30 marker proteins were included in the analysis in order to build robust phylogenies ($n = 98$). Furthermore, marker proteins found in less than 20 genomes were also excluded from the tree computation, which resulted in a set of 71 usable marker proteins. Subsequently, the proteins identified in the SAR11 genomes were aligned using Mafft v7.310 (accurate option L-INS-i) [53,54] and the resulting alignments were trimmed using Trimal v1.2 [14]. The most appropriate evolutionary model for each single copy marker, as estimated using Prottest v3.4.2 [22], was used in RAxML v8.2.12 [100] to perform a bootstrapped (*–autoMRE*) maximum likelihood phylogenetic reconstruction of each marker. These individual marker gene phylogenies were then combined into a species tree using ASTRAL v5.6.3 [123]. The branch distances of the resulting tree were re-calculated as substitutions per site in IQ-TREE v.1.6.12 [79] using the topology resulting from ASTRAL as a constraint for the computation, as previously reported [125]. For comparison, a 16S rRNA gene-based phylogeny was calculated with sequences identified in the newly recovered genomes and previously reported reference sequences representing all known SAR11 clades (Table S3) using Infernal v1.1.2 [75] with default parameters. All 16S rRNA gene sequences were aligned with SINA v.1.2.11 [85] using the SILVA v132 database as the reference [87]. The maximum likelihood phylogeny was built using MEGA X [62] with a General Time Reversible model [77].

Global distribution of SAR11 clades

The abundance of reference and recovered SAR11 genomes in available metagenomic datasets (Table S1) was estimated by competitively mapping metagenomic reads to a custom database built with Magic-BLAST v1.4.0 [8]. The database included 2970 closed genomes from the NCBI RefSeq database (2805 bacterial and 165 archaeal genomes) and the collection of 122 SAR11 genomes. In addition to the datasets used for MAG recovery, other available OMZ and open ocean datasets were also used for read recruitment and abundance estimations. These included datasets from the ETSP [70,110, 200 and 1000 m; [36]], ETNP [3085, 100, 125 and 300 m; [41]], Gulf of Mexico [GoM; 1, 25, 73, 150, 300, 600, 1000, 1470 and 2107 m; [112]], Hawaii Ocean Time-series (HOT) station ALOHA [25,75, 125 and 500 m; [25]], Bermuda Atlantic Time Series [BATS; 20, 50 and 100 m; [20]], Sea of Marmara [1000 m; [86]], the Mediterranean Sea [50 m; [37]], the Puerto Rico Trench [6000 m; [30]] and the TARA Oceans dataset comprising metagenomic samples from the Pacific, Atlantic, Southern and Indian Oceans, as well as from the Red and Mediterranean Seas at different depth levels [52,104]. Accession numbers and the metadata associated with these additional datasets used in this study are shown in Table S1. Reads mapping with high identity (≥95%) over at least 80% of the read length to SAR11 genomes were filtered using *MagicBlast_Tab_Filter.py* (**Star Methods**) and identified as SAR11 reads. This subset of reads was re-mapped to the set of 33 genomospecies identified previously for the calculation of the relative abundance of each genomospecies representative as follows. The per-base sequencing depth of all contigs of a genome was estimated using an in-house script, *MagicBlast_SeqDepth.py*, followed by the estimation of the central 80% of the truncated average sequencing depth (TAD80) of each genome using *TAD_Calculator.py* (**Star Methods**). This TAD estimation removes the top and bottom 10% of positions, as ranked by per-base sequencing coverage, in order to remove highly (e.g. conserved genes such as the 16S rRNA gene) or poorly (e.g. variable genes within the population) covered regions and multi-copy genes (outliers). The adjusted sequencing depth was then used to calculate the relative abundance of each genome over the total bacterial fraction in each metagenome by dividing the TAD80 by the number of genome equivalents calculated using MicrobeCensus v.1.1.0 [76]. A genome was assumed to be present in a dataset when its TAD was greater than 0, which translated to at least 10% sequencing breadth coverage across the genome (i.e. at least 10% of the genome was covered). This threshold was previously reported to be robust for the determination of the presence/absence of a bacterial genome in a metagenome [19]. High-sensitivity recruitment plots of metagenomic reads against genomospecies representatives were constructed using Blastn [12] with the '*-task blastn*' flag, which allows for more distant mismatches. The results were visualized using the RecPlot4 tool (**Star Methods**).

Functional gene annotation and metabolic reconstruction

Proteins from each genome were predicted using Prodigal v2.6.3 [48] and annotated iteratively using several approaches. First, all proteins were searched against the recently developed KOfam HMM database using KOfamscan [3]. High-quality matches were retained, and their assigned KO numbers and annotations were parsed and stored. Proteins with low confidence matches were recovered from the initial protein dataset and searched against the Swissprot database using Sword v1.0.3 [115]. High-quality hits (amino-acid identity ≥40% along 60% of the query protein) were retained, and the process was repeated successively with searches against the Trembl and NCBI non-redundant (NR) Rel. 96 databases [80,107]. Protein annotations and KO numbers of each high-quality match were extracted from their respective databases using an *in-house* script, *SQLite3_Search_ID.py*, which uses a manually created SQL database to store information of every protein in the Swissprot, Trembl and NCBI NR Rel. 96 databases. In addition, proteorhodopsin protein searches were performed against the curated rhodopsin database MICrhoDE [7]. A pangenome approach using all genomes was performed in order to summarize the gene content and reconstruct the metabolic potential of SAR11 genomes belonging to specific clades. For this, all proteins from all SAR11 genomes were clustered using CD-HIT with identity thresholds ≥50% along ≥60% of the shortest sequence, thus, aiming to represent an orthologous protein family group. The annotation of the representative

sequence from each orthologous cluster (longest sequence) was inherited from the previous annotation step and, in cases where there was no annotation available for the representative, a consensus annotation was manually assigned based on the annotation of other members of the orthologous cluster. In cases where no annotation was found, the sequence representative was annotated as hypothetical. Additional annotations for cluster representatives were obtained using KEGG KAAS [73] and MAPLE v2.3.1 [2]. Pathway reconstructions were performed using KEGG Mapper and an *in-house* script. Finally, considering that the completeness levels of SAGs in all SAR11 clades ranged between approximately 3.7 and 99%, a gene was defined as "core" when it was present in at least 70% of genomes with a completeness level ≥50%. This excluded highly incomplete SAGs and avoided the underestimation of the presence of a conserved gene in each clade. The pangenome size of the SAR11 group, on the other hand, was estimated considering genes present in at least one genome, regardless of its completeness level, and this metric was also used to calculate the pangenome size per subclade of interest [106].

Results and discussion

Sequencing, SAG and MAG recovery

Metadata associated with all metagenomic datasets used in this study are shown in Table S1, whereas general assembly and sequence statistics for each SAG, MAG and representative genomes are presented in Table S2. The targeted assembly and binning approach allowed the recovery of five MAGs belonging to subclade Ic, nine belonging to subclade IIa.A, and nine belonging to subclade V. In addition, 13 newly reported SAGs were analyzed. Together, the MAGs and SAGs represented 36 novel SAR11 genomes from different clades. Of these genomes, 21 MAGs and three SAGs were above medium quality (≥50% complete, ≤10% redundant), as defined in [9]. However, depending on the SCG dataset used in the analysis, the completeness estimate varied (Table S2). Using the average genome size of *Ca.* Pelagibacter ubique isolates as a reference (1,364,468 bp), the recovered subclade IIa.A MAGs represented approximately 67% (57–70%) of the reference. On the other hand, the average genome size of the four single-cell amplified genomes previously reported for subclade Ic members (1,019,735 bp) [108] indicated that the MAGs represented approximately 86% (57–108%) of the reported genome size of subclade Ic members. In both cases, the MAGs recovered in this study represented a genome size mean improvement of more than 50% compared with previously reported SAGs from both clades [112]. Finally, MAGs recovered from subclade V were between 54 and 94% complete compared with the only reported isolate genome from this subclade (HIMB59). The inclusion of additional genomes from subclade V was important in order to shed light on the controversial position of this subclade within the SAR11 phylogeny [43,117].

Genomic sequence diversity among SAR11 MAGs

Phylogenetic placement of the newly recovered and reference SAR11 genomes using the 16S rRNA gene confirmed the previously established clustering of SAGs and reference genomes into eight different subclades [Ia, Ib, Ic, IIa.A, IIa.B, IIb, III(a,b) and V]. Specifically, the 16S rRNA gene phylogeny showed that two MAGs and one SAG were placed in subclade Ic together with three previously reported SAGs [108], while three SAGs were placed in subclade IIa.A (Fig. 1). As previously reported for the SAR11 clades [44,112], our phylogenetic placement reconstructions indicated that both subclades Ic and IIa.A might represent distinct genera within the SAR11 group based on their 16S rRNA gene identities against other SAR11 clades or between themselves, which ranged, on average, between 93.1 and 91.4% (excluding subclade V; Table S4). These values were lower than thresholds commonly used to delineate species and genera [55,120]. The phylogenetic placement based on single-copy marker genes confirmed the 16S rRNA gene-based results above and showed the monophyletic nature of subclades Ic and IIa.A (Fig. 1).

In addition to phylogenetic evidence, ANI measurements and whole-genome content analysis provided additional resolution of genomic variation within and between subclades (Figure S1). As we have previously reported for subclades Ic and IIa.A [111], ANI genomic comparisons including the newly described genomes reported here, confirmed that each clade represented a distinct lineage (not the same species) with inter-cluster ANI values below 80%. Consistent with this previous report, AAI comparisons, which can resolve distantly related genomes [61,89], showed that subclades Ic and IIa.A likely represent different genera from the surface *Ca.* Pelagibacter ubique with inter-clade AAI values of approximately 60% (Fig. 2, Table S5). Moreover, the AAI value between subclades Ic and IIa.A was approximately 65%, which was very close to the 60% AAI value that commonly represents sequence relatedness between closely related (yet distinct) genera [60]. These genus-level AAI values were also consistent with previous reports on SAR11 genome comparisons that were, however, dominated by subclade Ia and Ib genomes [44]. In addition, these values are mostly driven by AAI comparisons between MAGs, given that no SAGs could be included in the calculation due to high levels of incompleteness (too few matches for robust AAI calculation). Nonetheless, the fact that AAI estimates were close to the genus-delimiting value and their monophyly was supported by both 16S rRNA and SCG phylogenies, suggested that subclades IIa.A and Ic did indeed represent distinct genera.

The ANI and AAI comparisons also revealed that the surface subclade Ia had large intra-clade diversity, as revealed by the lower AAI values among clade members (Tables S5 and S6). To test if this was also the case for subclades Ic and IIa.A, a finer resolution approach was used to study the intra-clade diversity. For this, all SAR11 genomes were clustered into "genomospecies", which was defined based on a 95% ANI threshold (<95% between genomospecies). This level of ANI corresponds well to the species level [89] and also follows the previously reported scheme for SAR11 genome clustering of freshwater subclades [111]. In total, 33 genomospecies were recovered from among the complete SAR11 genome dataset. Of these, not surprisingly, subclade Ia had the largest number of genomospecies ($n = 11$), while subclades Ic and IIa.A had four and one, respectively (Table S2). The lower number of genomospecies recovered in this study, compared with the 62 recovered in our previous analysis focused on freshwater clades [111], reflected better genome representatives for subclades Ic and IIa.A, and the exclusion of highly incomplete genomes (mostly SAGs) from the clustering analysis. On the other hand, the relatively large number of genomospecies recovered for subclade Ia, as well as the large number of subclades and genomospecies previously reported [44], suggested that subclade Ia had a larger intra-clade diversity compared to Ic and IIa.A. However, it should be mentioned that the subclade Ia genomes analyzed here were recovered from different oceanic locations, seasons and depths, while the genomes of subclades Ic and IIa.A were recovered from a single oceanic location. In fact, subclade Ic showed a similar diversity pattern as subclade Ia when SAGs previously recovered from Station ALOHA at a depth of 770 m in the Pacific Ocean [108] were included in the analysis (Table S5). This suggested that subclade Ic might exhibit intra-clade diversity in a similar way to that of *Ca.* P. ubique (subclade Ia) if assessed using a sample set spanning a comparable environmental range.

To explore further and provide additional evidence of the high intra-population diversity of subclades Ic and IIa.A (i.e. the pres-

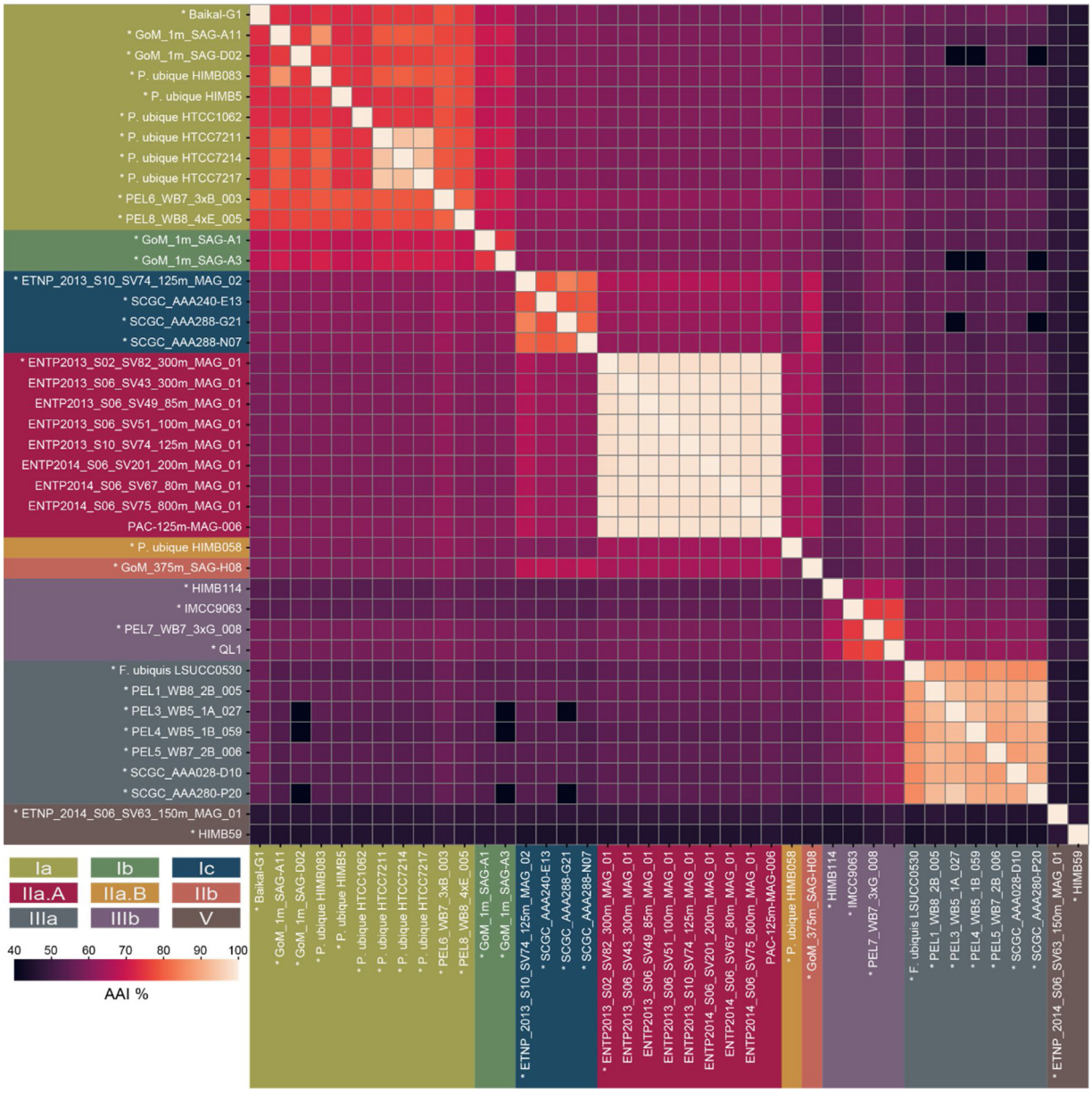

Fig. 2. Genome-aggregate average amino acid identity (AAI) comparisons of representative SAR11 genomes. AAI comparisons of genomospecies representatives (marked with asterisks) are shown. In addition, non-genomospecies representative genomes were included for subclade IIa.A because it represented a single genomospecies. The complete AAI matrix that includes a comparison with non-genomospecies representatives can be found in Table S6. Note that the inter-clade AAI distances are close to, or lower than, 60%, suggesting that they represent different genera by frequently used genomic standards.

ence of multiple closely related species in each subclade, as well as sequence variation within the same species), high sensitivity metagenome read mapping was performed for one genomospecies representative per subclade. The only genomospecies recovered from the ETNP metagenome datasets belonging to subclade Ic was selected as the representative, together with the only genomospecies recovered from subclade IIa.A. Each representative was used as a reference for read mapping, where, in each analysis, the mapped reads were taken only from the metagenome dataset in which the genomospecies representative was most abundant. This method allowed the observation of a "sequence gap area" in the identity distribution of mapped reads, indicative of sequence-discrete populations, in other words, populations with clear genomic differentiation from other closely-related and possibly co-occurring populations [58]. Typically, bacterial populations exhibit this sequence gap at 95% ANIr (average nucleotide identity of all reads mapping to a genome) [69], but for species with higher intra-population diversity, this threshold may be lower, as is the case for the marine cyanobacterium *Prochlorococcus* [17,58].

The distribution of nucleotide identities for the reads mapped along the selected genomospecies for subclades Ic and IIa.A indicated the sequence gap area had an approximately 90% sequence identity (Figure S2). This confirmed that both subclades had high intra-population diversity, similar to that of the surface-dwelling subclade represented by *Ca.* P. ubique. In addition to the drop in coverage at ANIr values close to 90%, there was an increase in coverage at ANIr values between 70–80%, thus, illustrating the co-presence of almost equally abundant, related populations in the same metagenome (different species of the same clade or even different clades). These results indicated that despite the high core genome conservation among SAR11 members, observed even in divergent freshwater clades [43], there was high intra-clade diversity in the SAR11 clade. As in *Prochlorococcus*, this high diversity of populations (or genomospecies) could be structured around a

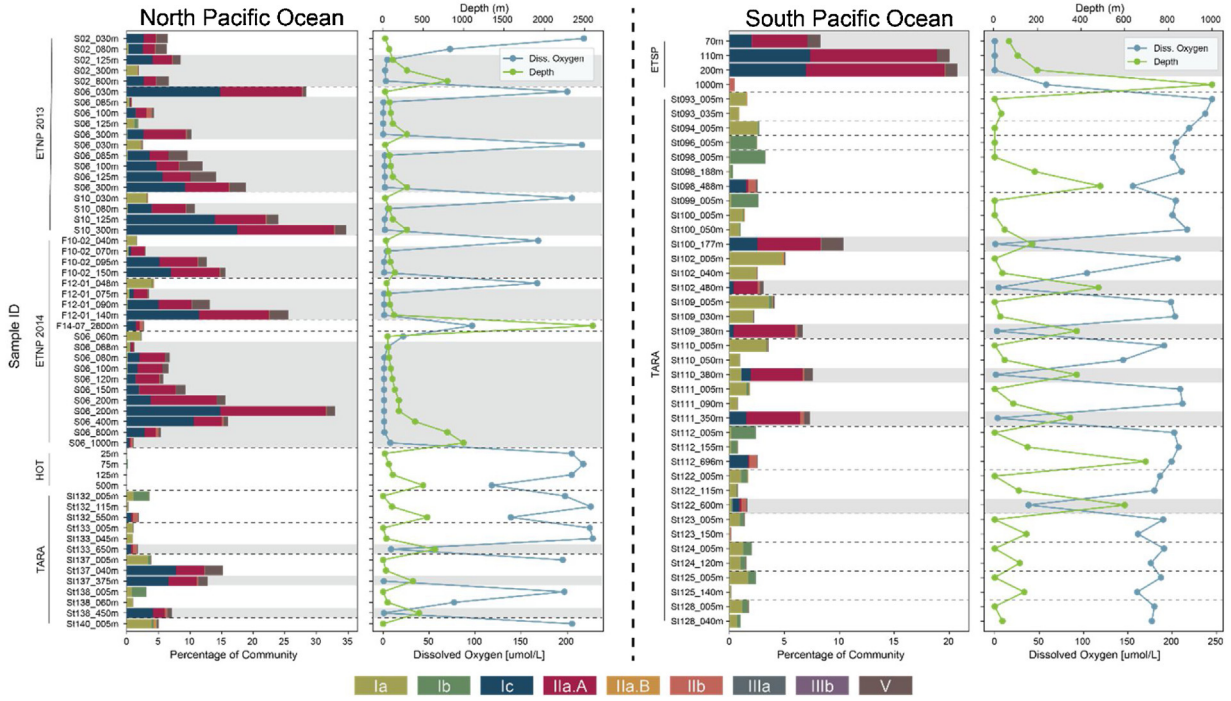

Fig. 3. Relative abundance of SAR11 subclades throughout the water column in various North and South Pacific Ocean regions with the associated metadata. The barplots show the estimated abundances of SAR11 subclades estimated using metagenome read mapping against all genomospecies representatives and normalized by genome equivalents. Data are organized based on the geographic location of the oceanic region and the depth of the metagenomic sample. Line plots to the right of each dataset show the corresponding depths (green) and dissolved oxygen (blue) concentrations associated with each metagenome sample. Different colors indicate the composite relative abundance per subclade. Dotted lines within each plot separate different stations and/or locations. Notice that subclades Ic and IIa.A dominate most OMZs and AMZs (shaded regions), while the more oxygenated waters are dominated by subclades Ia and Ib.

'genetic backbone' shared by all, with small genetic variations characterizing each population leading to a slightly increased fitness compared to other populations [6].

Worldwide ecological distribution

To estimate the biogeographical distribution of subclades Ic and IIa.A, and compare it with that of surface-dwelling clades, the relative abundance of each of the 33 genomospecies was estimated using competitive Magic-BLAST searches (Tables S7 and S8). On average, SAR11 members represented approximately 2.6% (ranging from <1% to ~8.2%) of the total microbial community in highly oxygenated waters around the world (>10 μmol L^{-1}), regardless of depth, with subclades Ia and Ib being the most abundant in these zones (~1.1%; Figure S3). Moreover, in surface oxygenated waters from the Pacific, Atlantic, Indian and Southern Oceans, and the Red and Mediterranean Seas, subclades IIa.A and Ic were barely detectable. However, in OMZs, SAR11 members increased in abundance and accounted, on average, for 11.6% of the total microbial population (1–34.7%). Of these, subclades Ic, IIa.A and V were the most abundant.

The metagenome-based abundances obtained for SAR11 populations were low when compared to previous estimates using (CARD-)FISH, which have been shown to range between approximately 10–35% in surface waters and approximately 15% in mesopelagic waters [46,95]. This discrepancy may be due, at least in part, to the different datasets used and the high conservation of the 16S rRNA gene across subclades together with other experimental factors and noise, such as DNA extraction efficiency. For instance, SAR11 genomes showed relatively low abundances or were not present in 454 datasets due to the low number of reads. Indeed, the relative abundance of SAR11 increased in the Illumina datasets and comprised, on average, 8% of the total (vs. 4% in 454 datasets). To test further the effect of genome-based abundance estimations, these values were compared with those obtained using metagenome-derived 16S rRNA gene-based estimations. Figure S4 shows that SAR11 OTUs made up approximately 22% of the bacterial community, which was much closer to FISH estimates. Moreover, there was a significant linear correlation between the metagenome- and rRNA gene-based estimates (Pearson correlation: 0.46, p-value <0.01) revealing that, while the rRNA gene-based abundances tended to be higher, the overall population abundance trends were similar. Therefore, it also appears that the higher conservation of the 16S rRNA gene relative to the whole genome could account, at least in part, for the differences observed; that is, the 16S rRNA gene (and hence FISH) covered the whole SAR11 clade, which is as diverse as some bacterial orders, if not phyla, whereas the genome would only cover the subclades that had sequenced representatives (which was not all subclades of SAR11). In addition, even though the metagenome-derived abundances were generally lower, this approach allowed differentiation between different SAR11 subclades, which was not possible using 16S rRNA gene sequences. Hence, we preferably report on the metagenome-based abundances throughout this current study.

In cases where a representative genomospecies of a given clade was detected in a sample, all genomospecies of the same clade -when available- were also detected in the same sample (Table S7). This pattern has also been previously observed for the surface clades [24], highlighting the large intra-clade diversity (multiple coexisting closely related species) present in the same samples throughout different oceans. The relative abundances of SAR11 members in OMZ samples (mostly from the Pacific Ocean; Fig. 3) showed that subclade IIa.A was restricted to waters with oxygen levels below detection limits and was absent (or in low abundances; <0.2%) in other mesopelagic, but oxygenated, waters. In contrast, subclade Ic, a known bathytype [108], dominated most deep, oxygenated waters (100–6000 m), while also being abundant in OMZ samples together with IIa.A members (Figure S3). Regardless of the

oxygen conditions in mesopelagic waters, subclade IIb was consistently present together with subclades Ic and IIa.A, although typically at lower abundances. This co-occurrence of clades has been previously linked to seasonality, depth and water mixing events in the Bermuda Atlantic Time-series (BATS) [116], suggesting these factors can influence SAR11 abundances in other regions. Moreover, Vergin et al. [116] also suggested that niche sub-division is also the result of "evolutionary innovations" that, associated with competition, leads to niche expansion and overlap between niches. In fact, an earlier description of subclade Ic genomes [108] demonstrated large similarities in the metabolic potential with surface subclade Ia, suggesting that subtle differences in metabolic content allow these organisms to co-occur in deeper waters.

Central metabolism in mesopelagic and surface SAR11 clades

The metabolic reconstructions were performed by including all members of the subclade, regardless of their genomospecies assignment, in order to provide for a robust assessment of the gene content differences (and similarities) between the different subclades. Genome representatives from subclades Ia, Ic and IIa.A encoded an average of 1037 (114–1547), 851 (266–1593) and 767 (221–1102) genes, respectively. The large variation in the number of genes per genome was mainly driven by the general low completeness of SAGs in each clade (average 37%), with approximately 78% of the SAGs (18 out of 23) encoding less than 800 genes. The incomplete nature of these genomes made it challenging to estimate the complete metabolic potential of each clade using a single representative from each clade (or even from each genomospecies). To circumvent this issue, metabolic reconstructions and comparisons were performed between genomes using the pangenomes calculated for each clade. For this purpose, proteins from all SAR11 genomes, excluding those with high contamination (>10%, Tables S1 and S5), were grouped into orthologous clusters (OCs), resulting in a total of 13,266 clusters, of which, approximately 50% (or 6674) were present only in one genome. Surface subclade Ia had a total of 3628 OCs present in at least one genome, of which 852 (12.8%) were part of the core genome (defined as being present in more than 70% of the genomes), and 1624 (24.3%) were genome-specific genes. These estimates resembled those obtained for subclade Ic (2460 total, 958 or 38.9% genome-specific) and subclade IIa.A (1204 total, 187 or 15.5% genome-specific). Therefore, it appears that the relatively large number of OCs observed within each clade was not solely attributable to the effects of genome incompleteness but rather to a relatively large and open species pangenome, which was consistent with previous estimates [43] (Figure S5A). Indeed, the analysis using only SAR11 genomes with more than 50% completeness showed that the pangenome of each clade modeled with a power-law function (KN^{γ}) had a γ parameter between 0 and 1 (Figure S5A), indicating an open pangenome [124] despite the relatively small SAR11 genome size. The core genome of all SAR11 clades combined and modeled using an exponential decay function $Ke^{-(N/\tau)} + \Omega$ [124] was estimated to plateau at approximately 280 genes (Figure S5A) or approximately 133 when including all genomes regardless of completeness (Figure S5C), which were lower values compared to approximately 598 genes predicted by a previous study with a much smaller number of genomes [43], but consistent with the higher diversity of genomes used here. Note that depending on the amino-acid identity threshold used for clustering, more than one protein cluster may have the same annotation, suggesting that the proteins in such clusters might be performing the same biological function. Indeed, such occurrences were often observed in our dataset (Table S5). When these proteins were merged and the core set was re-calculated, the number of core genes was estimated to be 505 (Figure S5B), which was much closer to previous estimates [43]. This result stresses the importance of using a function-based pangenome analysis approach in addition to sequence similarity-based clustering.

Despite the open pangenome and the low number of core genes, SAR11 members have relatively conserved metabolic capabilities [109], even when including more divergent mesopelagic and OMZ-clades (Fig. 4). For instance, subclade Ic and IIa.A members are capable of synthesizing most amino acids from precursors or by interconverting between amino acids, as previously reported for subclade Ia members [40]. The only exceptions to this were the additional capacity of some Ia and Ic members to synthesize glycine from glycolate, and the apparent auxotrophies for tryptophan, tyrosine and phenylalanine in subclade IIa.A caused by the lack of two steps in the shikimate pathway (Tables S9 and S10). Concomitantly, the presence of general L-amino acid and branched amino acid transporters was observed (with some missing subunits; Fig. 4, Tables S9 and S10), which might serve to compensate for the auxotrophies observed [109]. Additionally, and consistent with previous metabolic analyses of subclade Ia members [40,96], no proteins were recovered from the glucose phosphoenolpyruvate (PEP)-dependent phosphotransferase system (PTS) or the dedicated glucose transport system AglEFGK. However, multiple sugar transport systems were recovered that were present in *Ca.* P. ubique HTCC1062 (subclade Ia), which was previously reported to use glucose as a sole carbon source [96]. These transport systems were not universally present in subclade Ic genomes (present in approximately 43% Ic genomes; Table S9) and were absent in subclade IIa.A genomes, suggesting a similar phenomenon as in subclade Ia, in which certain strains can import and use glucose while others cannot. Likewise, although glucose utilization has been reported for some subclade Ia strains [96], the lack of genes encoding for key steps in the glycolysis pathway indicates that this pathway is probably not used by members of the surface clades [40,43,96]. The same enzymatic steps missing from the glycolysis pathway reconstruction were also observed (i.e. the absence of hexokinase [EC2.7.1.1], phosphofructokinase [EC2.7.1.11] and pyruvate kinase [EC2.7.1.40] in the genomes of subclades Ic and IIa.A). Of these, the hexokinase activity could be replaced by the repressor ORF kinases (ROKs), which have promiscuous kinase activities toward several sugars, including glucose [70,71].

ROKs have also been suggested to be involved in the Entner–Doudoroff (ED) pathway as an alternative for glucose utilization in subclade Ia members [96]. In fact, genes encoding proteins involved in the ED pathway were present in approximately 43–63% of subclade Ia and Ic genomes, indicating that some strains could complete all reactions in this pathway (Table S9). Of note, genes encoding enzymes for the last two steps of conversion, from 2-keto-3-deoxy-6-phosphogluconate to 2,4-keto-3-deoxy-6-phosphogluconate and then to 3-phosphoglycerate and pyruvate, appeared to be present in only a small fraction of subclade Ic members (14 and 29%, respectively). Likewise, subclade IIa.A members completely lacked the genes for the enzymes required to perform ED (Fig. 4). Finally, the surface subclades (Ia and Ib) also encoded the capacity for linking the ED pathway to the pentose phosphate shunt via a 6-phosphogluconate dehydrogenase that catalyzed the conversion between 6-phospho-D-gluconate and D-ribulose-5-phosphate (Fig. 4). In this regard, although no clade encoded the oxidative portion of the pentose phosphate pathway, all clades encoded the non-oxidative phase that generates precursors for nucleotides (ribose-5-phosphate and 5-phospho-D-ribose-α-1-pyrophosphate, PRPP) and amino acids (erythrose 4-phosphate). These results, in combination with the lack of glucose transport systems, suggested that glucose utilization (by means of glycolysis or ED) likely does not occur in subclade IIa.A members and, therefore, the origin of carbon for gluconeogenesis (for anabolic reactions) and the TCA cycle must come from other sources.

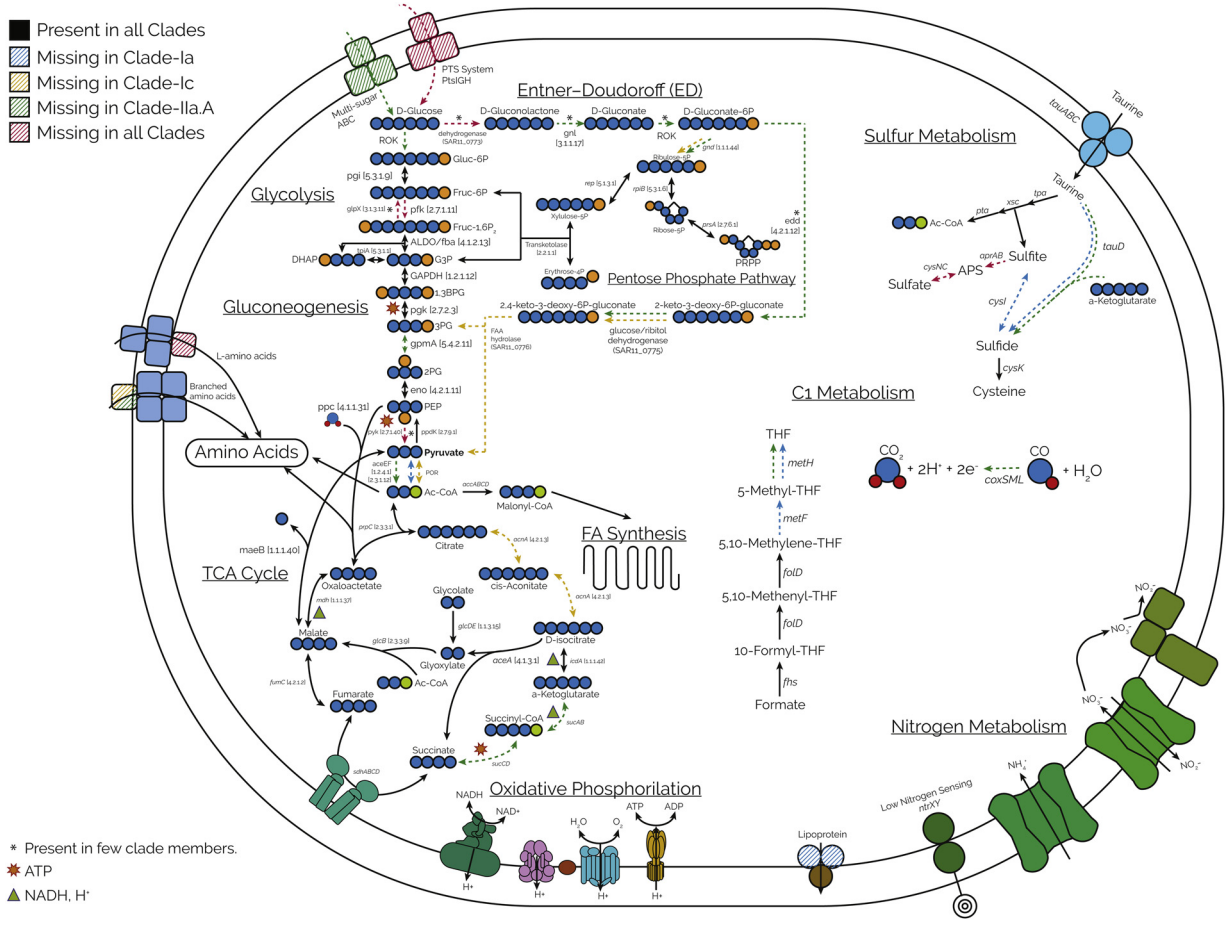

Fig. 4. Metabolic reconstruction of subclade Ia, Ic and IIa.A genomes. Metabolic predictions were based on subclade-specific pangenome protein annotations. Solid black lines indicate genes present in all subclades. Dashed lines indicate missing genes of a pathway (see legend for color designation). Most differences can be observed in the glycolysis, TCA cycle and sulfur metabolism pathways. Membrane associated proteins (transporters and complexes) are also shown. Ac-CoA: Acetyl-CoA, FA: Fatty Acid, APS: Adenylyl Sulfate, PRPP: 5-Phospho-alpha-D-ribose-1-diphosphate.

One such potential source, pyruvate, has been identified as the central carbon compound in SAR11 members [109]. The first step for pyruvate to be incorporated into the tricarboxylic acid (TCA) cycle is its transformation into acetyl-CoA. In aerobic microorganisms, this transformation can be mediated by the pyruvate dehydrogenase complex (PDC) in an irreversible fashion [23,83]. All components of the PDC were found in genomes from subclades Ia and Ic but not in the genomes of subclade IIa.A. In the latter, this transformation was predicted to be mediated by a pyruvate oxidoreductase, an enzyme widespread in anaerobic microorganisms [11], which can catalyze the conversion of pyruvate to acetyl-CoA and CO_2, and has also been reported to have pyruvate synthase capacity [34]. This capacity of subclade IIa.A directly linked the TCA cycle to glycolysis/gluconeogenesis, while subclades Ia and Ic required additional steps (Fig. 4). In the case of the TCA cycle, most steps were encoded in SAR11 genomes, as previously reported for the surface clades [40], with some exceptions for mesopelagic and OMZ-associated clades. For instance, subclade Ic lacked the gene encoding aconitate hydratase (AcnA [EC:4.2.1.3]) that catalyzes the conversion of citrate into D-isocitrate (Fig. 4). It was previously reported that genes for a complete TCA cycle were present in subclade Ic SAGs [108] but, in the current study, it was only possible to find the gene for these enzymes in one SAG out of seven genomes available (2/14 if the excluded incomplete genomes were considered; Table S9). Therefore, it can be expected that this enzyme would not be widespread in subclade Ic genomes and occurs only in a few of its representatives. On the other hand, subclade IIa.A lacked genes for both components of the 2-oxoglutarate dehydrogenase (SucAB) that catalyzes the conversion of α-ketoglutarate into succinyl-CoA and both components of the succinyl-CoA synthase (SucCD) that further catalyzes the conversion into succinate. These missing enzymatic functions of the TCA cycle appeared to be circumvented by alternative routes for the synthesis of the required precursors. In the case of subclade IIa.A, succinate production can be bypassed through the activity of glyoxylate cycle enzymes (the glyoxylate bypass), as commonly observed in freshwater and marine microbes [45,109]. The case of Ic members is different given that most members cannot produce D-isocitrate as a precursor for the glyoxylate bypass. They instead use glycolate from environmental sources [94] as a precursor for glycine [15] or oxidize it to glyoxylate that can then be incorporated into the glyoxylate bypass. Considering the number of metabolic enzymes present in subclade Ic and IIa.A members that have pyruvate as an intermediate and final product, and the previously recognized importance of pyruvate in the general metabolism of surface SAR11 members [109], pyruvate also appeared to be a central carbon molecule in the mesopelagic SAR11 clades. Consistent with the previous characterization of mesopelagic subclade genomes, the presence of *bd*-type oxidases was observed in three SAGS from subclade IIa.A [112]. However, these oxidases were not present in any other MAG or SAG from either subclade Ic or IIa.A, suggesting these might not be widespread or they were not recovered by the binning method used.

Metabolic versatility in mesopelagic and OMZ clades

Nitrogen metabolism

Nitrogen is an important limiting nutrient in marine environments [31] and its importance together with sulfur has been demonstrated in OMZs [63]. Indeed, *Ca.* P. ubique genomes encode multiple transport functions for nitrogen compounds such as ammonia, urea, basic amino acids, spermidine and putrescine [40]. Metabolic reconstruction identified the genes encoding the high-affinity two-component nitrogen sensor NtrXY in subclades Ia, IIa.A and Ic in more than 50% of the genomes. This, in conjunction with the presence of the *amtB* gene encoding an ammonium transmembrane transporter, indicated that mesopelagic and OMZ-associated SAR11 clades also used reduced inorganic nitrogen for growth, as does *Ca.* P. ubique [88]. In addition, the ammonia-producing xanthine metabolic pathway previously reported in subclade Ic SAGs [108] was also recovered in Ic MAGs from this study. Congruent with previous SAG analyses and biochemical characterization [112], genes encoding dissimilatory nitrate reductases (alpha, beta, gamma subunits) were identified in 100% of subclade IIa.A genomes, but in only 28% of subclade Ic genomes (encoding two or three of the subunits; Table S9). NarH, the beta subunit of the complex, requires a Fe-S cluster for its proper conformation [92]. The formation of Fe-S clusters requires iron and sulfur. Accordingly, all genes encoding the iron transport (AfuABC) and cysteine desulfurase (SufS) systems were found in subclade IIa.A members. However, a lower fraction of subclade Ic genomes (29%) carried genes for AfuABC but encoded several subunits of iron complex-binding proteins (Table S9). These results indicated that nitrate reduction was present in subclade IIa.A but it was probably not widespread in subclade Ic members. Finally, oxygen depletion to non-detectable levels in OMZs, which was not observed in the surface and mesopelagic samples, likely selected for the retention of nitrate reduction in OMZ-associated subclade IIa.A members, but not necessarily in subclade Ic from oxygenated, deeper water layers.

Sulfur metabolism

While sulfate reduction appears to be important in OMZs [13], the sulfur requirements for mesopelagic and OMZ-associated SAR11 members are still not well understood. In general, no sulfate assimilation pathway was found in SAR11 members (including subclades Ic and IIa.A), as previously reported, indicating an inability to uptake, reduce and incorporate sulfate (assimilatory sulfate reduction), or to grow solely on sulfate [15,110]. However, in *Ca.* P. ubique, reduced inorganic sulfur in the form of dimethylsulfoniopropionate (DMSP) or methionine can be used for growth and biosynthesis [110]. DMSP can be metabolized to methionine using the demethylation pathway or to dimethyl sulfide (DMS) using the cleavage pathway mediated by DMSP lyases [103]. In this current study, no DMSP lyase (encoded by *dddP* or *dddK*) was recovered in subclade Ic or IIa.A, indicating that mesopelagic and OMZ-associated clades probably could not transform DMSP into DMS. Instead, all associated gene homologs required for the demethylation of DMSP were identified in all clades (Table S9) [99,103], suggesting DMSP use was similar between clades. The presence of a complete DMSP demethylation pathway and an incomplete assimilatory sulfate reduction pathway suggested that SAR11 cells preferentially metabolized DMSP, which is an abundant *in situ* sulfur source [47,122], while undergoing adaptive gene losses for unnecessary pathways consistent with previous results [109].

Interestingly, the presence of adenylylsulfate reductase genes (*aprAB*) did not appear to be related to assimilatory sulfate reduction given the absence of most genes in this pathway. However, *aprAB* was found to be abundant in deep water samples compared to surface waters and it was suggested that these genes may be involved in taurine metabolism [108,118]. Indeed, in this current study, all the genes encoding for the taurine import system (TauABC) were found in the genomes of subclade Ia, Ic and IIa.A members. The proposed fate of the imported taurine shared by members of subclades Ia, Ic and IIa.A would be its conversion to acetyl-CoA with the subsequent detoxification of sulfite by oxidation into sulfate mediated by adenylylsulfate reductase (AprABM), which is a source for the possible production of sulfate in the oceans [118]. All genes encoding this taurine metabolic pathway were found in more than 70% of Ia and IIa.A genomes, but in only approximately 50% of subclade Ic genomes (Table S9), suggesting the possible existence of other mechanisms for metabolizing taurine. Furthermore, an alternative pathway to metabolize taurine was found encoded almost exclusively in subclade IIa.A genomes (90%). The presence of a gene encoding a sulfite reductase (CysI) that catalyzes the formation of sulfide, subsequently used in the formation of cysteine and acetate via a cysteine synthase (CysK), could provide subclade IIa.A members with an alternative taurine metabolism [78]. However, the *cysI* gene was found to be present in only approximately 30% of Ic genomes, suggesting it was not widespread in this clade. Finally, the potential ability to metabolize taurine directly to sulfide via taurine deoxygenase (TauD) was encoded exclusively in subclade Ic genomes and not in subclade IIa.A genomes. This apparent absence of the *tauD* gene in subclade IIa.A genomes was consistent with the occurrence of this clade in low oxygen waters, since the dioxygenase requires oxygen to function [29]. The fact that different alternatives for the use of taurine were found in subclade Ic genomes suggested that members of this clade had more versatility for sulfur metabolism than surface- and other OMZ-associated clades, which most likely expands their ability to compete in OMZs and oxygen-rich deeper waters. Although the sulfur metabolism alternatives reported here for novel SAR11 clades remain to be biochemically confirmed, the gene content patterns observed provide evidence for higher versatility in pathways related to sulfur cycling (Fig. 4), confirming the importance of these processes in OMZs [13,18].

Carbon monoxide oxidation

The ocean has long been recognized as a source of carbon monoxide (CO) with concentrations above atmospheric levels, suggesting it is a global source of CO [105]. CO can be metabolized aerobically and anaerobically by bacteria as a carbon or electron source [26,68]. Therefore, finding evidence of SAR11 CO metabolism would suggest that these microorganisms also take advantage of this abundant carbon source, which would further expand the metabolic capabilities of this versatile bacterial group. In addition to the genetic differences in nitrogen and sulfur metabolism found in subclades Ic and IIa.A, an additional potential to metabolize CO was found in subclades Ia and Ic. A carbon monoxide dehydrogenase operon (*coxSML*) was identified in 57% and 50% of subclade Ic and Ia genomes, respectively (Fig. 4, Table S9). Genes encoding small, medium and large subunits were present, and a highly similar large subunit was found among subclade Ic SAGs and MAGs, indicating that this finding was not an assembly artifact. The large subunit CoxL is commonly used to assess the phylogenetic affiliation of CO-oxidizers. Here, all CoxL protein sequences from SAR11 isolates, SAGs and MAGs formed a monophyletic subcluster deep within all form II Cox proteins, which was perhaps indicative of a common ancestry that has not subsequently experienced frequent horizontal gene transfer (Figure S6). The *cox* genes identified are also ancestral to the carbon monoxide/acetyl-CoA pathway, one of the most widely distributed and ancient carbon fixation pathways (Wood-Ljungdahl pathway) [33]. However, CO can also be used as an electron donor and carbon source by other taxa, referred to as carboxydotrophs, or exclusively used as a source of electrons by even more taxa, referred to as carboxydovores [57].

The latter display high affinities for CO but are inhibited by elevated concentrations of CO. These organisms appear to use CO in a 'mixotrophic' fashion, whereby the energy gained through CO oxidation is used to supplement other metabolisms. The gene arrangement reported here (*coxSLM*) is the same as that reported for cultivated marine carboxydovores [72]. Furthermore, CoxL from the subclade Ia isolate HIMB1321 was consistently the closest phylogenetic (cultivated) relative to the CoxL recovered from subclade Ic and formed a monophyletic sub-cluster deep within form II of CoxL, indicating a common ancestry (Figure S6). Considering the highly streamlined nature of SAR11 genomes [40,43], the presence of the *cox* genes highlights the importance of carbon monoxide oxidation in the surface and deep ocean where CO is produced and is abundant [21]. Oxygen-tolerant *cox* genes have recently been identified in other OMZ-associated taxa such as *Marinimicrobia*-SAR406 [5]. While it is not yet possible to conclude without experimental data that subclade Ic members can use CO *in situ* as a reductant, it is tempting to speculate that CO is used to sustain metabolic activities under anoxic conditions. It is also notable that the carbon monoxide dehydrogenase operon is widespread in surface water SAR11 members [44], indicating that this might be a more general metabolic property of the SAR11 clade. Facultative anaerobic taxa that utilize nitrate as an electron acceptor to oxidize CO have previously been reported [56]. Therefore, the activity and regulation of CO utilization enzymes, especially as it concerns nitrate reduction in OMZs, merits future investigation.

Finally, while the genomes used in the metabolic reconstruction analyses were at least 50% complete, it is likely that the fraction of genes (or pathways) present in a clade was in reality higher than those reported above, and the values reported in this current study would be a lower bound estimate. Furthermore, the pangenome approach used in this study provided confidence that the predicted pathways were present in more than one genome, effectively reducing the possibility of a false positive signal resulting from errors in the assembly or binning steps.

Cell morphology of subclade IIa.A members

CARD-FISH [84] was used instead of FISH to assess the morphology of SAR11 members of subclade IIa.A because of their small size and low ribosome content. Members of this clade were small, curved rods of approximately 1 μm × 0.2 μm (Figure S7), which was consistent with the morphology reported for other SAR11 members in general, including those for freshwater clades [45,88]. Figure S7 also shows the target for the probe designed here, together with its nucleotide bases that discriminate subclade IIa.A from other subclades.

Concluding remarks: mesopelagic SAR11 members as models to study microdiversity, niche partitioning and global geographical distribution

In this study, MAGs were recovered from SAR11 subclades Ic and IIa.A, lineages that have been detected primarily in mesopelagic and low-oxygen marine waters, respectively. This dataset increased the number of available SAR11 genomes and contributed toward a more comprehensive description, genome comparison, characterization and classification of SAR11 bacteria. Notably, whole-genome comparisons using AAI, ANI and marker gene phylogenies (Figs. 1 and 2, and S1) congruently showed that subclade Ic and IIa.A members each belonged to different genera compared to the surface *Ca.* Pelagibacter subclade Ia and Ib members. Furthermore, gene-content comparisons and metabolic reconstructions justified the taxonomic description of subclades Ic and IIa.A as distinct genera. Considering that subclades Ic and IIa.A co-occur in the same samples, there must be functional (e.g. different metabolic pathways) and/or ecological (e.g. phage predation, preference for oxygen and/or nitrate) differentiation that prevents the members of the two clades from out-competing one another. Indeed, it was found that there were several potentially important metabolic differences between the two clades, most notably the lack of enzymes to carry out the Entner–Doudoroff (ED) pathway in subclade IIa.A and the presence of carbon monoxide oxidation enzymes in subclade Ic (Fig. 4). These differences suggested that subclade IIa.A members were probably not capable of importing and metabolizing glucose and, therefore, must rely on other molecules (probably pyruvate) as the main carbon source, while subclade Ic members were more versatile in the range of nutrients they could exploit in both low-oxygen and oxygen-rich mesopelagic waters.

Finally, the characterization of the two clades presumably involved in carbon and nitrogen cycling in low-oxygen OMZ waters opens the door for future studies on microdiversity, evolution and niche partitioning in these key habitats for oceans and climate change. For instance, the genome and gene sequences recovered here provided the means needed for experimental work (e.g. metatranscriptomics, qPCR analysis or mesocosms) in order to test emerging hypotheses concerning the metabolic differentiation between closely related members of the same or closely related clades. Future studies could also focus on recovering a broader diversity of high-quality or complete genomes from these clades, facilitating continued comparative studies of SAR11, inclusive of other subclades such as IIa.B and IIb, whose specific metabolic adaptations are yet to be described.

Description of Ca. *Mesopelagibacter carboxydoxydans and* Ca. *Anoxipelagibacter denitrificans*

Candidatus Mesopelagibacter carboxydoxydans [Me.so.pe.la.gi.bac'ter. Gr. masc. adj. *mesos*, medium, middle; L. neut. n. *pelagus* the open sea; N.L. masc. n. *bacter*, a rod; N.L. masc. n. Mesopelagibacter, a rod from the mesopelagic regions of the ocean; car.bo.xyd.o'xy.dans. N.L. neut. n. *carboxydum*, carbon monoxide; N.L. pres. part. *oxydans*, oxidizing; N.L. part. adj. carboxydoxydans, oxidizing carbon monoxide. This representative is named after its abundance in the mesopelagic regions of the ocean and its encoded potential capacity to oxidize carbon monoxide. We propose to name SAR11 subclade Ic members as *Candidatus* Mesopelagibacter, defined according to recently proposed genomic, phylogenetic and ecological standards [59,60]. Members of this genus are heterotrophic, mesopelagic-dwelling bacteria having the capacity to thrive in oxygen minimum zones and oxygenated, deeper sea areas contrary to their closest named relative *Ca.* Pelagibacter ubique. The type species is *Candidatus* Mesopelagibacter carboxydoxydans, which has the potential for oxidation of carbon monoxide, nitrate reduction and incorporation of taurine, which allow it to be a versatile member of the SAR11 clade, colonizing environmentally different ocean regions. The genome size is predicted to be approximately 1.2 Mbp with a G+C% content of approximately 30%. Based on the genomes available, we estimate there are five genomospecies within this genus to date. The designated type material is MAG ETNP_2013_S10_SV74_125m_MAG_02. Genome sequences can be found under NCBI BioSample accession number SAMN15305926.

Candidatus Anoxipelagibacter denitrificans [An.o.xy.pe.la.gi.bac'ter. Gr. pref. *an-*, without; Gr. masc. adj. *oxys*, acid or sour and in combined words indicating oxygen; L. neut. n. *pelagus*, the open sea; N.L. masc. n. *bacter*, a rod; N.L. masc. n. Anoxypelagibacter, a rod from oxygen-depleted areas in the ocean; de.ni.tri'fi.cans. N.L. part. adj. *denitrificans*, denitrifying]. We propose to name SAR11 subclade IIa.A members as *Candidatus* Anoxipelagibacter, due to their abundance in ocean oxygen minimum zones where oxygen is below detection levels. The type species is *Candidatus* Anoxipelagibacter denitrificans, named because of its capacity to use nitrate as an electron

acceptor, which is a trait not present in its closest named surface relative *Ca.* P. ubique. In general, *Ca.* A. denitrificans exhibits small rod-shaped cells of approximately 1 × 0.2 μm. The genome size is predicted to be approximately 1.2 Mbp with a G+C% content of approximately 30%. The designated type material is MAG ENTP2013_S02_SV82_300m_MAG_01 and its genome sequence can be found under NCBI BioSample accession number SAMN15305928.

Acknowledgements

This research was supported, in part, by the US National Science Foundation (awards DBI 1759831 and OCE 1416673 to KTK and 1151698 to FJS) and a Community Science Program from the US Department of Energy (DOE; to FJS and KTK). We thank Ramon Rosselló-Móra for his help with CARD-FISH and Janet Hatt for proofreading and providing valuable suggestions. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.syapm.2021.126185.

References

[1] Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C. (2014) Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146.
[2] Arai, W., Taniguchi, T., Goto, S., Moriya, Y., Uehara, H., Takemoto, K., Ogata, H., Takami, H. (2018) MAPLE 2.3.0: an improved system for evaluating the functionomes of genomes and metagenomes. Biosci. Biotechnol. Biochem. 82, 1515–1517.
[3] Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., Ogata, H. (2020) KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36, 2251–2252.
[4] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
[5] Bertagnolli, A.D., Padilla, C.C., Glass, J.B., Thamdrup, B., Stewart, F.J. (2017) Metabolic potential and *in situ* activity of marine *Marinimicrobia* bacteria in an anoxic water column. Environ. Microbiol. 19, 4392–4416.
[6] Biller, S.J., Berube, P.M., Lindell, D., Chisholm, S.W. (2015) *Prochlorococcus*: the structure and function of collective diversity. Nat. Rev. Microbiol. 13, 13–27.
[7] Boeuf, D., Audic, S., Brillet-Gueguen, L., Caron, C., Jeanthon, C. (2015) MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. Database (Oxford) 2015, bav080.
[8] Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., Madden, T.L. (2019) Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinform. 20, 405.
[9] Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G., Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., Garrity, G.M., Dodsworth, J.A., Yooseph, S., Sutton, G., Glockner, F.O., Gilbert, J.A., Nelson, W.C., Hallam, S.J., Jungbluth, S.P., Ettema, T.J.G., Tighe, S., Konstantinidis, K.T., Liu, W.T., Baker, B.J., Rattei, T., Eisen, J.A., Hedlund, B., McMahon, K.D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G.W., Rinke, C., Genome Standards, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D.H., Eren, A.M., Schriml, L., Banfield, J.F., Hugenholtz, P., Woyke, T. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731.
[10] Brown, M.V., Lauro, F.M., DeMaere, M.Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M.J., Fuhrman, J.A., Andrews-Pfannkoch, C., Hoffman, J.M., McQuaid, J.B., Allen, A., Rintoul, S.R., Cavicchioli, R. (2012) Global biogeography of SAR11 marine bacteria. Mol. Syst. Biol. 8, 595.
[11] Buckel, W., Golding, B.T. (2006) Radical enzymes in anaerobes. Annu. Rev. Microbiol. 60, 27–49.
[12] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. (2009) BLAST+: architecture and applications. BMC Bioinform. 10, 421.
[13] Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F., Revsbech, N.P., Ulloa, O. (2010) A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. Science 330, 1375–1378.
[14] Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.
[15] Carini, P., Steindler, L., Beszteri, S., Giovannoni, S.J. (2013) Nutrient requirements for growth of the extreme oligotroph '*Candidatus* Pelagibacter ubique' HTCC1062 on a defined medium. ISME J. 7, 592–602.
[16] Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. 3, 283–295.
[17] Caro-Quintero, A., Konstantinidis, K.T. (2012) Bacterial species may exist, metagenomics reveal. Environ. Microbiol. 14, 347–355.
[18] Carolan, M.T., Smith, J.M., Beman, J.M. (2015) Transcriptomic evidence for microbial sulfur cycling in the eastern tropical North Pacific oxygen minimum zone. Front. Microbiol. 6, 334.
[19] Castro, J.C., Rodriguez, R.L., Harvey, W.T., Weigand, M.R., Hatt, J.K., Carter, M.Q., Konstantinidis, K.T. (2018) imGLAD: accurate detection and quantification of target organisms in metagenomes. PeerJ 6, e5882.
[20] Coleman, M.L., Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. Proc. Natl. Acad. Sci. USA 107, 18634–18639.
[21] Conte, L., Szopa, S., Séférian, R., Bopp, L. (2019) The oceanic cycle of carbon monoxide and its emissions to the atmosphere. Biogeosciences 16, 881–902.
[22] Darriba, D., Taboada, G.L., Doallo, R., Posada, D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27, 1164–1165.
[23] de Kok, A., Hengeveld, A.F., Martin, A., Westphal, A.H. (1998) The pyruvate dehydrogenase multi-enzyme complex from Gram-negative bacteria. Biochim. Biophys. Acta 1385, 353–366.
[24] Delmont, T.O., Kiefl, E., Kilinc, O., Esen, Ö.C., Uysal, I., Rappé, M.S., Giovannoni, S., Eren, A.M. (2017) The global biogeography of amino acid variants within a single SAR11 population is governed by natural selection. bioRxiv, 170639.
[25] DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., Chisholm, S.W., Karl, D.M. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311, 496–503.
[26] Diender, M., Stams, A.J., Sousa, D.Z. (2015) Pathways and bioenergetics of anaerobic carbon monoxide fermentation. Front. Microbiol. 6, 1275.
[27] Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.J., Richter, R.A., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., Haft, D.H., Halpern, A.L., Lasken, R.S., Nealson, K., Friedman, R., Venter, J.C. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. 6, 1186–1199.
[28] Eddy, S.R. (2011) Accelerated profile HMM searches. PLoS Comput. Biol. 7, e1002195.
[29] Eichhorn, E., van der Ploeg, J.R., Kertesz, M.A., Leisinger, T. (1997) Characterization of alpha-ketoglutarate-dependent taurine dioxygenase from *Escherichia coli*. J. Biol. Chem. 272, 23031–23036.
[30] Eloe, E.A., Fadrosh, D.W., Novotny, M., Zeigler Allen, L., Kim, M., Lombardo, M.J., Yee-Greenbaum, J., Yooseph, S., Allen, E.E., Lasken, R., Williamson, S.J., Bartlett, D.H. (2011) Going deeper: metagenome of a hadopelagic microbial community. PLOS ONE 6, e20388.
[31] Elser, J.J., Bracken, M.E., Cleland, E.E., Gruner, D.S., Harpole, W.S., Hillebrand, H., Ngai, J.T., Seabloom, E.W., Shurin, J.B., Smith, J.E. (2007) Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol. Lett. 10, 1135–1142.
[32] Eren, A.M., Esen, O.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data'. PeerJ 3, e1319.
[33] Fuchs, G. (2011) Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? Annu. Rev. Microbiol. 65, 631–658.
[34] Furdui, C., Ragsdale, S.W. (2000) The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the Wood-Ljungdahl pathway. J. Biol. Chem. 275, 28494–28499.
[35] Ganesh, S., Bristow, L.A., Larsen, M., Sarode, N., Thamdrup, B., Stewart, F.J. (2015) Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. ISME J. 9, 2682–2696.
[36] Ganesh, S., Parris, D.J., DeLong, E.F., Stewart, F.J. (2014) Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. ISME J. 8, 187–211.
[37] Ghai, R., Martin-Cuadrado, A.B., Molto, A.G., Heredia, I.G., Cabrera, R., Martin, J., Verdu, M., Deschamps, P., Moreira, D., Lopez-Garcia, P., Mira, A., Rodriguez-Valera, F. (2010) Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. ISME J. 4, 1154–1166.
[38] Giovannoni, S.J. (2017) SAR11 bacteria: the most abundant plankton in the oceans. Ann. Rev. Mar. Sci. 9, 231–255.
[39] Giovannoni, S.J., Bibbs, L., Cho, J.-C., Stapels, M.D., Desiderio, R., Vergin, K.L., Rappé, M.S., Laney, S., Wilhelm, L.J., Tripp, H.J., Mathur, E.J., Barofsky, D.F. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature 438, 82–85.
[40] Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappe, M.S., Short, J.M., Carring-

ton, J.C., Mathur, E.J. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242–1245.
[41] Glass, J.B., Kretz, C.B., Ganesh, S., Ranjan, P., Seston, S.L., Buck, K.N., Landing, W.M., Morton, P.L., Moffett, J.W., Giovannoni, S.J., Vergin, K.L., Stewart, F.J. (2015) Meta-omic signatures of microbial metal and nitrogen cycling in marine oxygen minimum zones. Front. Microbiol. 6, 998.
[42] Glockner, F.O., Fuchs, B.M., Amann, R. (1999) Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence *in situ* hybridization. Appl. Environ. Microbiol. 65, 3721–3726.
[43] Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., Rappe, M.S. (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. MBio 3.
[44] Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., Fornas, O., Martinez-Garcia, M., Lopez-Perez, M. (2020) Ecogenomics of the SAR11 clade. Environ. Microbiol. 22, 1748–1763.
[45] Henson, M.W., Lanclos, V.C., Faircloth, B.C., Thrash, J.C. (2018) Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME J. 12, 1846–1860.
[46] Herlemann, D.P., Woelk, J., Labrenz, M., Jurgens, K. (2014) Diversity and abundance of "Pelagibacterales" (SAR11) in the Baltic Sea salinity gradient. Syst. Appl. Microbiol. 37, 601–604.
[47] Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Burgmann, H., Welsh, R., Ye, W., Gonzalez, J.M., Mace, K., Joye, S.B., Kiene, R.P., Whitman, W.B., Moran, M.A. (2006) Bacterial taxa that limit sulfur flux from the ocean. Science 314, 649–652.
[48] Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 119.
[49] Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T., Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9, 5114.
[50] Jimenez-Infante, F., Ngugi, D.K., Vinu, M., Blom, J., Alam, I., Bajic, V.B., Stingl, U. (2017) Genomic characterization of two novel SAR11 isolates from the Red Sea, including the first strain of the SAR11 Ib clade. FEMS Microbiol. Ecol. 93.
[51] Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359.
[52] Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., Pesant, S., Reynaud, E.G., Sardet, C., Sieracki, M.E., Speich, S., Velayoudon, D., Weissenbach, J., Wincker, P., Tara Oceans, C. (2011) A holistic approach to marine eco-systems biology. PLoS Biol. 9, e1001177.
[53] Katoh, K., Misawa, K., Kuma, K., Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
[54] Katoh, K., Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform. 9, 286–298.
[55] Kim, M., Oh, H.S., Park, S.C., Chun, J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64, 346–351.
[56] King, G.M. (2006) Nitrate-dependent anaerobic carbon monoxide oxidation by aerobic CO-oxidizing bacteria. FEMS Microbiol. Ecol. 56, 1–7.
[57] King, G.M., Weber, C.F. (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. Nat. Rev. Microbiol. 5, 107–118.
[58] Konstantinidis, K.T., DeLong, E.F. (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2, 1052–1065.
[59] Konstantinidis, K.T., Rosselló-Móra, R. (2015) Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal. Syst. Appl. Microbiol. 38, 223–230.
[60] Konstantinidis, K.T., Rosselló-Móra, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISME J. 11, 2399–2406.
[61] Konstantinidis, K.T., Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA 102, 2567–2572.
[62] Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549.
[63] Lam, P., Lavik, G., Jensen, M.M., van de Vossenberg, J., Schmid, M., Woebken, D., Gutierrez, D., Amann, R., Jetten, M.S., Kuypers, M.M. (2009) Revising the nitrogen cycle in the Peruvian oxygen minimum zone. Proc. Natl. Acad. Sci. USA 106, 4752–4757.
[64] Larsen, M., Lehner, P., Borisov, S.M., Klimant, I., Fischer, J.P., Stewart, F.J., Canfield, D.E., Glud, R.N. (2016) *In situ* quantification of ultra-low O_2 concentrations in oxygen minimum zones: application of novel optodes. Limnol. Oceanogr. Methods 14, 784–800.
[65] Lee, M.D. (2019) GToTree: a user-friendly workflow for phylogenomics. Bioinformatics 35, 4162–4164.
[66] Lo, C.C., Chain, P.S. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinform. 15, 366.
[67] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Forster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., Konig, A., Liss, T., Lussmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H. (2004) ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.
[68] Meyer, O., Schlegel, H.G. (1983) Biology of aerobic carbon monoxide-oxidizing bacteria. Annu. Rev. Microbiol. 37, 277–310.
[69] Meziti, A., Tsementzi, D., Rodriguez, R.L., Hatt, J.K., Karayanni, H., Kormas, K.A., Konstantinidis, K.T. (2019) Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. ISME J. 13, 767–779.
[70] Miller, B.G., Raines, R.T. (2004) Identifying latent enzyme activities: substrate ambiguity within modern bacterial sugar kinases. Biochemistry 43, 6387–6392.
[71] Miller, B.G., Raines, R.T. (2005) Reconstitution of a defunct glycolytic pathway via recruitment of ambiguous sugar kinases. Biochemistry 44, 10776–10783.
[72] Moran, M.A., Buchan, A., Gonzalez, J.M., Heidelberg, J.F., Whitman, W.B., Kiene, R.P., Henriksen, J.R., King, G.M., Belas, R., Fuqua, C., Brinkac, L., Lewis, M., Johri, S., Weaver, B., Pai, G., Eisen, J.A., Rahe, E., Sheldon, W.M., Ye, W., Miller, T.R., Carlton, J., Rasko, D.A., Paulsen, I.T., Ren, Q., Daugherty, S.C., Deboy, R.T., Dodson, R.J., Durkin, A.S., Madupu, R., Nelson, W.C., Sullivan, S.A., Rosovitz, M.J., Haft, D.H., Selengut, J., Ward, N. (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432, 910–913.
[73] Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182–W185.
[74] Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420, 806–810.
[75] Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29, 2933–2935.
[76] Nayfach, S., Pollard, K.S. (2015) Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 16, 51.
[77] Nei, M., Kumar, S. 2000 Molecular Evolution and Phylogenetics, Oxford University Press, Oxford, New York, pp 333.
[78] Neumann, S., Wynen, A., Truper, H.G., Dahl, C. (2000) Characterization of the *cys* gene locus from *Allochromatium vinosum* indicates an unusual sulfate assimilation pathway. Mol. Biol. Rep. 27, 27–33.
[79] Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.
[80] O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745.
[81] Oh, H.M., Kang, I., Lee, K., Jang, Y., Lim, S.I., Cho, J.C. (2011) Complete genome sequence of strain IMCC9063, belonging to SAR11 subgroup 3, isolated from the Arctic Ocean. J. Bacteriol. 193, 3379–3380.
[82] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055.
[83] Patel, M.S., Roche, T.E. (1990) Molecular biology and biochemistry of pyruvate dehydrogenase complexes. FASEB J. 4, 3224–3233.
[84] Pernthaler, A., Pernthaler, J., Amann, R. (2002) Fluorescence *in situ* hybridization and catalyzed reporter deposition for the identification of marine bacteria. Appl. Environ. Microbiol. 68, 3094–3101.
[85] Pruesse, E., Peplies, J., Glockner, F.O. (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28, 1823–1829.
[86] Quaiser, A., Zivanovic, Y., Moreira, D., Lopez-Garcia, P. (2011) Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. ISME J. 5, 285–304.
[87] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596.
[88] Rappé, M.S., Connon, S.A., Vergin, K.L., Giovannoni, S.J. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418, 630–633.
[89] Rodriguez-R, L., Konstantinidis, K. (2014) Bypassing cultivation to identify bacterial species. Microbe 9, 111–118.
[90] Rodriguez-R, L.M., Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints 4, e1900v1901.
[91] Rodriguez, R.L., Gunturu, S., Harvey, W.T., Rosselló-Móra, R., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T. (2018) The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level. Nucleic Acids Res. 46, W282–W288.
[92] Rothery, R.A., Workun, G.J., Weiner, J.H. (2008) The prokaryotic complex iron-sulfur molybdoenzyme family. Biochim. Biophys. Acta 1778, 1897–1929.
[93] Salter, I., Galand, P.E., Fagervold, S.K., Lebaron, P., Obernosterer, I., Oliver, M.J., Suzuki, M.T., Tricoire, C. (2015) Seasonal dynamics of active SAR11

ecotypes in the oligotrophic Northwest Mediterranean Sea. ISME J. 9, 347–360.
[94] Schada von Borzyskowski, L., Severi, F., Kruger, K., Hermann, L., Gilardet, A., Sippel, F., Pommerenke, B., Claus, P., Cortina, N.S., Glatter, T., Zauner, S., Zarzycki, J., Fuchs, B.M., Bremer, E., Maier, U.G., Amann, R.I., Erb, T.J. (2019) Marine *Proteobacteria* metabolize glycolate via the beta-hydroxyaspartate cycle. Nature 575, 500–504.
[95] Schattenhofer, M., Fuchs, B.M., Amann, R., Zubkov, M.V., Tarran, G.A., Pernthaler, J. (2009) Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environ. Microbiol. 11, 2078–2093.
[96] Schwalbach, M.S., Tripp, H.J., Steindler, L., Smith, D.P., Giovannoni, S.J. (2010) The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. Environ. Microbiol. 12, 490–500.
[97] Sekar, R., Pernthaler, A., Pernthaler, J., Warnecke, F., Posch, T., Amann, R. (2003) An improved protocol for quantification of freshwater *Actinobacteria* by fluorescence *in situ* hybridization. Appl. Environ. Microbiol. 69, 2928–2935.
[98] Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F. (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 3, 836–843.
[99] Smith, D.P., Nicora, C.D., Carini, P., Lipton, M.S., Norbeck, A.D., Smith, R.D., Giovannoni, S.J. (2016) Proteome remodeling in response to sulfur limitation in "*Candidatus* Pelagibacter ubique". mSystems 1, e00068-16.
[100] Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30, 1312–1313.
[101] Stingl, U., Tripp, H.J., Giovannoni, S.J. (2007) Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. ISME J. 1, 361–371.
[102] Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., Landry, Z.C., Giovannoni, S.J. (2011) One carbon metabolism in SAR11 pelagic marine bacteria. PLoS ONE 6, e23973.
[103] Sun, J., Todd, J.D., Thrash, J.C., Qian, Y., Qian, M.C., Temperton, B., Guo, J., Fowler, E.K., Aldrich, J.T., Nicora, C.D., Lipton, M.S., Smith, R.D., De Leenheer, P., Payne, S.H., Johnston, A.W., Davie-Martin, C.L., Halsey, K.H., Giovannoni, S.J. (2016) The abundant marine bacterium *Pelagibacter* simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. Nat. Microbiol. 1, 16065.
[104] Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans, c., 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, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P. (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science 348, 1261359.
[105] Swinnerton, J.W., Linnenbom, V.J., Lamontagne, R.A. (1970) The ocean: a natural source of carbon monoxide. Science 167, 984–986.
[106] Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit, Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA 102, 13950–13955.
[107] The UniProt, C. (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169.
[108] Thrash, J.C., Temperton, B., Swan, B.K., Landry, Z.C., Woyke, T., DeLong, E.F., Stepanauskas, R., Giovannoni, S.J. (2014) Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. ISME J. 8, 1440–1451.
[109] Tripp, H.J. (2013) The unique metabolism of SAR11 aquatic bacteria. J. Microbiol. 51, 147–153.
[110] Tripp, H.J., Kitner, J.B., Schwalbach, M.S., Dacey, J.W., Wilhelm, L.J., Giovannoni, S.J. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature. 452, 741–744.
[111] Tsementzi, D., Rodriguez-R, L.M., Ruiz-Perez, C.A., Meziti, A., Hatt, J.K., Konstantinidis, K.T. (2019) Ecogenomic characterization of widespread, closely-related SAR11 clades of the freshwater genus "*Candidatus* Fonsibacter" and proposal of *Ca*. Fonsibacter lacus sp. nov. Syst. Appl. Microbiol. 42, 495–505.
[112] Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez, R.L., Burns, A.S., Ranjan, P., Sarode, N., Malmstrom, R.R., Padilla, C.C., Stone, B.K., Bristow, L.A., Larsen, M., Glass, J.B., Thamdrup, B., Woyke, T., Konstantinidis, K.T., Stewart, F.J. (2016) SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature 536, 179–183.
[113] Ulloa, O., Canfield, D.E., DeLong, E.F., Letelier, R.M., Stewart, F.J. (2012) Microbial oceanography of anoxic oxygen minimum zones. Proc. Natl. Acad. Sci. USA 109, 15996–16003.
[114] van Dongen, S., Abreu-Goodger, C. (2012) Using MCL to extract clusters from networks. Methods Mol. Biol. 804, 281–295.
[115] Vaser, R., Pavlovic, D., Sikic, M. (2016) SWORD – a highly efficient protein database search. Bioinformatics 32, i680–i684.
[116] Vergin, K.L., Beszteri, B., Monier, A., Thrash, J.C., Temperton, B., Treusch, A.H., Kilpert, F., Worden, A.Z., Giovannoni, S.J. (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J. 7, 1322–1332.
[117] Viklund, J., Martijn, J., Ettema, T.J., Andersson, S.G. (2013) Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. PLoS ONE 8, e78858.
[118] Williams, T.J., Long, E., Evans, F., Demaere, M.Z., Lauro, F.M., Raftery, M.J., Ducklow, H., Grzymski, J.J., Murray, A.E., Cavicchioli, R. (2012) A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. ISME J. 6, 1883–1900.
[119] Wu, Y.W., Simmons, B.A., Singer, S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607.
[120] Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. 12, 635–645.
[121] Yilmaz, L.S., Parnerkar, S., Noguera, D.R. (2011) mathFISH, a web tool that uses thermodynamics-based mathematical models for *in silico* evaluation of oligonucleotide probes for fluorescence *in situ* hybridization. Appl. Environ. Microbiol. 77, 1118–1122.
[122] Yoch, D.C. (2002) Dimethylsulfoniopropionate: its sources, role in the marine food web, and biological degradation to dimethylsulfide. Appl. Environ. Microbiol. 68, 5804–5815.
[123] Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. 19, 153.
[124] Zhang, X., Liu, X., Yang, F., Chen, L. (2018) Pan-genome analysis links the hereditary variation of *Leptospirillum ferriphilum* with its evolutionary adaptation. Front. Microbiol. 9, 577.
[125] Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J.G., Belda-Ferre, P., Al-Ghalith, G.A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J.B., Huang, S., Salam, N., Jiao, J.Y., Wu, Z., Xu, Z.Z., Cantrell, K., Yang, Y., Sayyari, E., Rabiee, M., Morton, J.T., Podell, S., Knights, D., Li, W.J., Huttenhower, C., Segata, N., Smarr, L., Mirarab, S., Knight, R. (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains *Bacteria* and *Archaea*. Nat. Commun. 10, 5477.