Undocumented potential for primary productivity in a globally-distributed bacterial photoautotroph

Authors: E.D. Graham[1], J.F. Heidelberg[1,2], B.J. Tully[*1,2]

Author Affiliations:

[1]Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

[2]Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA, USA

*Corresponding Author: Benjamin Tully, tully.bj@gmail.com

Keywords: autotrophy; marine carbon cycle; metagenomics; Alphaproteobacteria; aerobic anoxygenic phototrophs

Abstract: Aerobic anoxygenic phototrophs (AAnPs) are common in the global oceans and are associated with photoheterotrophic activity. To date, AAnPs have not been identified in the surface ocean that possess the potential for carbon fixation. Using the *Tara Oceans* metagenomic dataset, we have reconstructed high-quality genomes of four bacteria that possess the genomic potential for anoxygenic phototrophy, carbon fixation via the Calvin-Benson-Bassham cycle, and the oxidation of sulfite and thiosulfate. Forming a monophyletic clade within the *Alphaproteobacteria* and lacking cultured representatives, the organisms compose minor constituents of local microbial communities (0.1-1.0%), but are globally distributed, present in multiple samples from the North Pacific, Mediterranean Sea, the East Africa Coastal Province, and the South Atlantic. These organisms represent a shift in our understanding of microbially-mediated photoautotrophy in the global oceans and provide a previously undiscovered route of primary productivity.

Significance Statement: In examining the genomic content of organisms collected during the *Tara Oceans* expedition, we have identified a novel clade within the *Alphaproteobacteria* that has the potential for photoautotrophy. Based on genome observations, these organisms have the potential to couple inorganic sulfur compounds as electron donors to fix carbon into biomass. They are globally distributed, present in samples from the North Pacific, Mediterranean Sea, East Africa Coastal Current, and the South Atlantic. This discovery may require re-examination of the microbial communities in the global ocean to understand and constrain the impacts of this group of organisms on the global carbon cycle.

Introduction

It has been understood for decades that the basis of the global marine carbon cycle are oxygenic photoautotrophs that perform photoautotrophic processes. Two additional phototrophic processes are common in the ocean and are mediated by proteorhodopsin-containing microorganisms and aerobic anoxygenic phototrophs (AAnPs). Proteorhodopsins and AAnPs have historically been associated with photoheterotrophy[1,2], a process that supplements additional energy to microorganisms beyond what is obtained as part of a heterotrophic metabolic strategy. AAnPs utilize bacteriochlorophyll (BChl *a*), are globally distributed[3], and have been identified in phylogenetically diverse groups of microorganisms[4-6]. Though anaerobic microorganisms with BChl *a* are known to fix CO_2[7] and marine AAnPs can incorporate inorganic carbon via anaplerotic reactions[8], marine AAnPs have not been linked to carbon fixation[9]. The identification of marine AAnPs capable of carbon fixation adds to our understanding of microbial photosynthesis in the global oceans and represents a previously undiscovered route of photoautotrophy.

The *Tara Oceans* expedition generated microbial metagenomes during a circumnavigation of the global oceans[10,11]. *Tara Oceans* samples were collected from 63 sites in 10 major ocean provinces, with most sites contributing multiple sampling depths (generally, surface, deep chlorophyll maximum [DCM], and mesopelagic) and multiple size fractions (generally, 'viral', 'girus' [giant virus], 'bacterial', and 'protistan') from each depth (Supplementary Information 1). We independently assembled each sample and assemblies from all samples within a province were combined and subjected to binning techniques to reconstruct microbial genomes (Fig. 1 and Extended Data Fig. 1). Microbial genomes reconstructed from eight of ten provinces (Mediterranean, Red Sea, Arabian Sea, Indian Monsoon Gyre, East Africa Coastal, South Atlantic, and North Pacific; 36 sites, 134 samples) were annotated using the KEGG Ontology (KO) system[12] and examined for the genes and pathways of interest.

Results and Discussion

From 1,774 metagenome-assembled genomes (MAGs), 53 genomes possessed the genes encoding the core subunits of type-II photochemical reaction centers (PufLM). Of those 53, four genomes (MED800, EAC638, SAT68, NP970; Fig. 1) also contained genes for ribulose-1,5-bisphosphate carboxylase (Rubisco; RbsLS; Fig. 2). Rubisco has four major forms, of which three (Types I, II, and III) have been shown to fix CO_2 and two are known to participate in the Calvin-Benson-Bassham (CBB) cycle (Types I and II[13,14]. Phylogenetic placement of the Rubisco large subunits recovered from the genomes revealed them to be of the Type IC/D subgroup[13], suggesting that the identified proteins represent *bona fide* Rubiscos capable of carbon fixation (Fig. 2). Within the Type IC/D subgroup, the Rubisco sequences from the four analyzed genomes formed a distinct cluster with environmental sequences derived from the Global Ocean Survey (GOS) metagenomes[15,16], but lacking sequences from reference organisms.

Similarly, the PufM sequences from the analyzed *Tara* genomes did not cluster with reference sequences, instead grouping with sequences from the GOS metagenomes[15]. Sequences from MED800, SAT68, and NP970 were group together in one cluster, while EAC638 was located in a separate cluster (Fig. 3). The MED800/SAT68/NP970 clade is basal to the previously identified phylogroups E and F, while the EAC638 clade is basal to the *Roseobacter*-related phylogroup G[17]. As MED800, EAC638, SAT68, and NP970 branch in distinct clades on both the RbsL and PufM trees that consist of entirely environmental sequences, it may be

possible that these clades represent a phylogenetically coherent group of organisms with the potential for both phototrophy and carbon fixation.

The genomes were of high-quality (66-85% complete; <5.5% duplication; Table 1) with sufficient phylogenetic markers for accurate placement (Extended Data Table 1). The four organisms form a monophyletic clade basal to the Family *Rhodobacteraceae* (Fig. 4). The relationship between the genomes would suggest that NP970, SAT68, and MED800 are phylogenetically more closely related to each other than either are to EAC638. As is common with assembled metagenomic sequences, the recovered genomes lack a distinguishable 16S rRNA gene sequence. However, based on the observed phylogenetic distance in the concatenated marker tree, we suggest that these organisms represent a new clade within the *Rhodobacteraceae*, and possibly a family-level clade previously without a reference sequence within the *Alphaproteobacteria*. We propose that NP970, SAT68, and MED800 represent three species within the same genus (tentatively named, '*Candidatus Luxescamonas taraoceani'*), with EAC638 as a representative of a species in a sister genus (tentatively named, '*Candidatus Luxescabacter africus'*).

In addition to Rubisco, all four genomes contained genes encoding phosphoribulokinase, an essential gene of the CBB cycle, and 50-89% of the genes necessary to perform complete carbon fixation (Fig. 5). The BChl *a* genes in MED800, EAC638, SAT68, and NP970 were accompanied by bacteriochlorophyll biosynthesis and light-harvesting genes (Supplementary Information 2). This complement of bacteriochlorophyll biosynthesis and essential carbon fixation genes support a role for autotrophy within these organisms beyond the identified marker genes.

All four genomes possessed ATP-binding cassette (ABC) type transporters for spermidine/putrescine and L- and branched-chain amino acids. These transporters are indicative of the utilization of organic nitrogen compounds, as spermidine and putrescine are nitrogen rich organic compounds, while the scavenging of amino acids reduces the overall nitrogen demands of the cell. Further, the genomes lacked transporters and degradation enzymes for many of the saccharides common in the marine environment[18] (Fig. 5). However, MED800, SAT68, and NP970 possessed an ABC-type α-glucoside transporter, an annotated β-glucosidase in SAT68, and D-xylose and D-ribose ABC-type transporters in EAC638. While autotrophs are generally considered to not require external sources of organic carbon, saccharide transporters are commonly observed in classical photoautotrophic organisms[19,20], including the specific example of α-glucoside transporters in strains of *Synechocystis*[21]. For the four genomes, the minimal number of carbon transport and degradation genes may suggest that the organisms have a limited capacity to utilize dissolved organic carbon compounds, but are capable of heterotrophic growth under certain conditions. As such, the genomic potential of these organisms suggest that NP970, SAT68, MED800, and EAC638 are likely facultative autotrophs or mixotrophs.

In oxygenic photosynthesis, electrons are donated as a result of the oxidation of water. Lacking photosystem II, BChl *a*-containing organisms are incapable of oxidizing water and would require an alternative electron donor for autotrophic processes. EAC638 and SAT68 contained the full/partial suite of genes necessary for thiosulfate oxidation, while MED800, NP970, and SAT68 possessed the genes for oxidizing sulfite. The oxidation of sulfur compounds has previously been linked to autotrophy in the marine environment[22]. The oxidation of organic sulfur compounds, like dimethyl sulfide, has been shown to be a source of thiosulfate[23] and sulfite[24] in the marine environment. Electrons derived from inorganic sulfur sources (thiosulfate

122 and/or sulfite) could be transferred directly through cytochromes or membrane-bound quinone
123 dehydrogenases to the electron transport chain. Electrons could be shuttled to reaction centers to
124 generate proton motive force necessary to convert NADH to NADPH via transhydrogenase
125 (reverse electron flow) and generate ATP for the CBB cycle (Fig. 5).

126 The reconstruction of four genomes from the same novel family in four different
127 provinces (North Pacific, Mediterranean, East Africa coastal current, and South Atlantic)
128 suggests that the observed genomes represent an *in situ* microbial population from the surface
129 marine environment. Each of the genomes recruit metagenomic reads from multiple sampling
130 sites in each province and are present at >0.1% of the microbial relative abundance (range: 0.1-
131 1.04%; mean: 0.286%) in 20 samples (Fig. 1; Supplemental Information 1). Predominantly, the
132 genomes were present in samples are located in the surface (n = 5) or DCM (n = 12). These
133 organisms were collected at depths where light was available for photosynthesis and less
134 frequently identified at deeper depths (n = 3 mesopelagic samples). When >0.1% relative
135 abundance, the genomes tend to be more abundant in the 'bacterial'/'girus' size fraction (n = 14),
136 though were also observed in 'protistan' (n = 4), and 'viral' (n = 2) size fractions (Supplemental
137 Information 1). The nature of the 'bacterial' size fractions suggests that these organisms are
138 generally not particle attached and <1.6μm in size. The occurrence in the protistan fraction may
139 be due to slightly larger cells or attachment to particles, but this data are difficult to interpret as
140 the 'protistan' and 'bacterial' size fractions can overlap (0.8-1.6μm). As members of the free-
141 living bacterioplankton, these organisms should be poised to grow in aerobic conditions. All four
142 genomes possessed the genes encoding for cytochromes involved in aerobic metabolisms (aa_3-
143 and bc_1-type), and lacked the genes for cytochromes involved in microaerobic metabolisms and
144 alternative electron acceptors. Further, all four genomes encoded the gene for an oxygen-
145 dependent ring cyclase (*acsF*), a necessary component in bacteriochlorophyll biosynthesis for
146 which there is alternative that is oxygen-independent (*bchE*) and used by anaerobic organisms.

147 With this discovery, the potential photosynthesis in the ocean has expanded beyond
148 organisms harboring chlorophyll *a* to include *Alphaproteobacteria* with BChl *a*. Though these
149 organisms have not been cultivated or sequenced before, both PufM and Rubisco in MED800,
150 EAC638, NP970, and SAT68 are phylogenetically related to GOS-derived proteins, lending
151 credence to the fact that these organisms may be a persistent element of oceanic carbon fixation.
152 As such, clades of environmentally sampled genes (*rbsL* and *pufM*) can now be linked to a
153 previously unrecognized source of marine primary productivity. Because carbon fixation
154 estimates are broadly estimated by the concentration of chlorophyll compounds (specifically
155 chlorophyll *a*), current values are likely incomplete and do account for the contributions of
156 bacteriochlorophyll-containing, carbon fixing AAnPs. The identification of a globally distributed
157 clade of AAnPs in the ocean capable of carbon fixation continues to expand our understanding of
158 the marine carbon cycle.

159 **Materials and Methods:**

160 *Assembly*

161 All sequences for the reverse and forward reads from each sampled station and depth within the
162 *Tara Oceans* dataset were accessed from European Molecular Biology Laboratory (EMBL)[10,11].
163 Typically, *Tara* sampling sites have multiple metagenomic samples, representing different
164 sampling depths and size fractions. The common size fractions were used during sampling were:
165 'bacterial' (0.22-1.6μm) (includes Mediterranean 'girus' samples), 'protistan' (0.8-5.0μm),

166 'girus' (0.45-0.8μm) and 'viral' (<0.22μm). Surface samples were collected at ~5-m depth, while
167 deep chlorophyll maximum (DCM) and mesopelagic depths were variable depending the
168 physiochemical features of the site. Paired-end reads from different filter sizes from each site and
169 depth (e.g., TARA0007, girus filter fraction, sampled at the DCM) were assembled using
170 Megahit[25] (v1.0.3; parameters: --preset, meta-sensitive) (Supplementary Information 1). All of
171 the Megahit assemblies from each province were pooled in to two tranches based on assembly
172 size, <2kb and ≥2kb. Longer assemblies (≥2kb) with ≥99% semi-global identity were combined
173 using CD-HIT-EST[26] (v4.6; -T 90 -M 500000 -c 0.99 -n 10). The reduced set of contiguous
174 DNA fragments (contigs) ≥2kb was then cross-assembled using Minimus2[27] (AMOS v3.1.0;
175 parameters: -D OVERLAP=100 MINID=95).

176 *Binning*

177 Contigs from each province were initially clustered into tentative genomic bins using
178 BinSanity[28]. Due to computational limitations, the South Atlantic, East African Coastal province,
179 and Mediterranean Sea were initially run with contig size cutoffs of 11.5kbp, 7.5kbp, and 7kbp,
180 respectively. The BinSanity workflow was run iteratively three times using variable preference
181 values (v.0.2.5.5; parameters: -p [(1) -10, (2) -5, (3) -3] -m 4000 -v 400 -d 0.95). Between each
182 of the three main clustering steps, refinement was performed based on sequence composition
183 (parameters: -p [(1) -25, (2) -10, (3) -3] -m 4000 -v 400 -d 0.95 -kmer 4). After refinement and
184 before the next pass with BinSanity, bins were evaluated using CheckM[29] (v.1.0.3; parameters:
185 lineage_wf, default settings) for completion and redundancy. High-quality genomes were
186 considered those that were >90% complete with <10% contamination, >80% complete with <5%
187 contamination, or >50% complete with <2% contamination. Bins representing high-quality
188 genomes were removed from subsequent rounds of clustering. After identification of the four
189 genomes of interest (initially 51-84% complete, <7.0% contamination), binning was performed
190 with CONCOCT[30] (v.0.4.1; parameters: -c 800 -I 500) on contigs >5kb from each province that
191 had a produced a genome of interest. To improve completion estimates, overlapping CONCOCT
192 and BinSanity bins were visualized using Anvi'o[31] (v.2.1.0) and manually refined to improve
193 genome completion and minimize contamination estimates (Extended Data Fig. 3-6).

194 *Annotation*

195 Putative DNA coding sequences (CDSs) were predicted for each genome using Prodigal[32]
196 (v.2.6.2; -m -p meta). Putative CDS were submitted for annotation by the KEGG database using
197 BlastKOALA[12] (taxonomy group, Prokaryotes; database, genus_prokaryotes +
198 family_eukaryotes; Accessed March 2017) (Supplementary Information Table 3). Assessment of
199 pathways and metabolisms of interest were determined using the script KEGG-decoder.py
200 (www.github.com/bjtully/BioData/tree/master/KEGGDecoder). Genomes of interested were
201 determined based on the presence of genes assigned as the M subunit of type-II photochemical
202 reaction center (PufLM) and ribulose-1,5-bisphosphate carboxylase (RbsLS). After confirmation
203 of the genes of interest (see below), additional annotations were performed for the genomes
204 using the Rapid Annotation using Subsystem Technology (RAST) service (Classic RAST default
205 parameters - Release70)[33].

206 *Phylogeny*

207 An initial assessment of phylogeny was conducted using pplacer[34] within CheckM. The
208 Prodigal-derived CDSs were searched for a collection of single-copy marker genes that was
209 common to all four *Tara* assembled genomes using hidden Markov models collected from the

Pfam database[35] (Accessed March 2017) and HMMER [36] (v3.1b2; parameters: hmmsearch -E 1e-10 --notextw). 17 marker genes were identified that met this criteria[37-39] (Extended Data Table 1). The 17 markers were identified in 2,889 reference genes from complete and partial genomes accessed from NCBI Genbank[40] (Supplementary Information 4). If a genome contained multiple copies of a single marker gene both were excluded from the final tree. Only genomes containing ≥10 markers were used for phylogenetic placement. Each marker set was aligned using MUSCLE[41] (v3.8.31; parameter: -maxiters 8) and trimmed using TrimAL[42] (v.1.2rev59; parameter: -automated1). Alignments were then manually assessed and concatenated in Geneious[43]. An approximate maximum likelihood tree was generated using FastTree[44] (v.2.1.3; parameters: -gtr -gamma; Supplementary Information 5). A simplified version of this phylogenetic tree was constructed using the same protocol, but with 160 reference genomes for Fig. 2 (Supplementary Information 6 and 7).

Phylogenetic tree – Rubisco and Bacteriochlorophyll

RbsL and PufM sequences representing previously described lineages were collected[13,17] (Supplementary Information 8 and 9). Additional reference PufM sequences were collected from environmentally generated bacterial artificial chromosomes[5] and Integrated Microbial Genomes (IMG; Accessed Feb 2017)[45]. Protein sequences from IMG were assessed based on genomes with KEGG Ontology (KO) annotations[46] for the reaction center subunit M (K08929). PufM sequences from Prodigal predicted CDS (as above) of Global Ocean Survey (GOS) assemblies[16] were identified using DIAMOND[47] (v.0.8.36.98; parameters: BLASTP, default settings), where all reference and *Tara* genome sequences were used as a query. Two separate phylogenetic trees were constructed (RbsL and PufM) using the following methodology. Sequences were aligned using MUSCLE[41] (parameter: -maxiters 8) and automatically trimmed using TrimAL[42] (parameter: -automated1) (Supplementary Information 10 and 11). After manual assessment, trimmed alignments were used to construct approximately-maximum-likelihood phylogenetic trees using FastTree[44] (parameters: -gtr -gamma) (Supplementary Information 12 and 13).

Relative abundance of genomes in each sample

Reads from each sample were recruited against all assemblies ≥2kb from the same province using Bowtie2[48] (parameters: default settings), under the assumptions that contigs <2kb would include, low abundance bacteria and archaea, bacteria and archaea with high degrees of repeats/assembly poor regions, fragmented picoeukaryotic genomes, and problematic read sequences (low quality, sequencing artefacts, etc.). For the four sets of contigs (North Pacific, Mediterranean, East Africa Coastal province, and South Atlantic), putative CDS were determined via Prodigal (parameters: see above). In order to estimate the relative abundance of the four analyzed genomes within the bacteria and archaea portion of the total microbial community (excluding eukaryotes and viruses), single-copy marker genes were identified using a collection of previously identified HMMs[49,50] and searched using HMMER[36] (hmmsearch --notextw --cut_tc). Markers belonging to the four genomes were isolated from the total set of environmental markers. The number of reads aligned to each marker was determined using BEDTools[51] (v2.17.0; multicov default parameters). Length-normalized relative abundance values were determined for each genome (Supplementary Information 1):

$$\frac{\sum \text{Reads bp}^{-1} \text{ genome markers}}{\sum \text{Reads bp}^{-1} \text{ metagenome contig markers}} \times 100$$

Data availability

Data is available… submission to NCBI is ongoing. [Currently data is available at FigShare, including high resolution copies of figures, contig and protein sequences, and all supplementary information files: **https://figshare.com/s/9f603e9bbef71164e61b**]

References:

1. Béjá, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411,** 786–789 (2001).
2. Harashima, K., Kawazoe, K., Yoshida, I. & Kamata, H. Light-Stimultated Aerobic Growth of Erythrobacter Species Och-114. *Plant and Cell Physiology* **28,** 365–374 (1987).
3. Schwalbach, M. S. & Fuhrman, J. A. Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. *Limnology and Oceanography* **50,** 620–628 (2005).
4. Koblížek, M. Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiol. Rev.* **39,** 854–870 (2015).
5. Béjá, O. *et al.* Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415,** 630–633 (2002).
6. Kang, I. *et al.* Genome Sequence of Fulvimarina pelagi HTCC2506T, a Mn(II)-Oxidizing Alphaproteobacterium Possessing an Aerobic Anoxygenic Photosynthetic Gene Cluster and Xanthorhodopsin. *J. Bacteriol.* **192,** 4798–4799 (2010).
7. Frigaard, N.-U. Biotechnology of Anoxygenic Phototrophic Bacteria. *Adv. Biochem. Eng. Biotechnol.* **156,** 139–154 (2016).
8. Hauruseu, D. & Koblížek, M. Influence of Light on Carbon Utilization in Aerobic Anoxygenic Phototrophs. *Appl. Environ. Microbiol.* **78,** 7414–7419 (2012).
9. Moran, M. A. *et al.* Deciphering ocean carbon in a changing world. *Proceedings of the National Academy of Sciences* **113,** 3143–3151 (2016).
10. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2,** 150023–16 (2015).
11. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359–1261359 (2015).
12. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428,** 726–731 (2016).
13. Tabita, F. R. *et al.* Function, Structure, and Evolution of the RubisCO-Like Proteins and Their RubisCO Homologs. *Microbiol. Mol. Biol. Rev.* **71,** 576–599 (2007).
14. Badger, M. R. & Bek, E. J. Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO2 acquisition by the CBB cycle. *Journal of Experimental Botany* **59,** 1525–1541 (2007).
15. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Biol* **5,** e77 (2007).
16. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
17. Yutin, N. *et al.* Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ. Microbiol.* **9,** 1464–1475 (2007).
18. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3,** 14 (2015).
19. Gómez-Baena, G. *et al.* Glucose Uptake and Its Effect on Gene Expression in Prochlorococcus. *PLoS ONE* **3,** e3416–11 (2008).
20. Michelou, V. K., Cottrell, M. T. & Kirchman, D. L. Light-Stimulated Bacterial Production and Amino Acid Assimilation by Cyanobacteria and Other Microbes in the North Atlantic Ocean. *Appl. Environ. Microbiol.* **73,** 5539–5546 (2007).
21. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3,** 109–136 (1996).

22. Swan, B. K. *et al.* Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science* **333,** 1296–1300 (2011).
23. de Zwart, J., Nelisse, P. N. & Kuenen, J. G. Isolation and characterization of Methylophaga sulfidovorans sp nov: An obligately methylotrophic, aerobic, dimethylsulfide oxidizing bacterium from a microbial mat. *FEMS Microbiol. Ecol.* **20,** 261–270 (1996).
24. Kelly, D. P., Baker, S. C., Trickett, J., Davey, M. & Murrell, J. C. Methanesulphonate utilization by a novel methylotrophic bacterium involves an unusual monooxygenase. *Microbiology* **140,** 1419–1426 (1994).
25. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102,** 3–11 (2016).
26. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
27. Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* **Chapter 11,** Unit 11.8 (2011).
28. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5,** e3035–19 (2017).
29. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
30. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Meth* **11,** 1144–1146 (2014).
31. Eren, A. M. *et al.* Anvi‘o: an advanced analysis and visualization platform for ’omics data. *PeerJ* **3,** e1319 (2015).
32. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28,** 2223–2230 (2012).
33. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).
34. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11,** 538 (2010).
35. Bateman, A. *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **30,** 276–280 (2002).
36. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
37. Wu, D., Jospin, G. & Eisen, J. A. Systematic Identification of Gene Families for Use as ‘Markers’ for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS ONE* **8,** e77033–11 (2013).
38. Santos, S. R. & Ochman, H. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ. Microbiol.* **6,** 754–759 (2004).
39. Alexandre, A., Laranjo, M., Young, J. P. W. & Oliveira, S. dnaJ is a useful phylogenetic marker for alphaproteobacteria. *Int. J. Syst. Evol. Microbiol.* **58,** 2839–2849 (2008).
40. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28,** 15–18 (2000).
41. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
42. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
43. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
44. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5,** e9490 (2010).
45. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,** D344–8 (2006).
46. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).
47. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* **12,** 59–60 (2014).
48. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
49. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31,** 533–538 (2013).

50. Tully, B. J. & Heidelberg, J. F. Potential Mechanisms for Microbial Energy Acquisition in Oxic Deep-Sea Sediments. *Appl. Environ. Microbiol.* **82,** 4232–4243 (2016).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

Acknowledgments:

We would like to acknowledge and thank Drs. Eric Webb and William Nelson for providing invaluable comments and critiques in the early stages of this research. We are indebted to the *Tara Oceans* consortium for their commitment to open-access data that allows data aficionados to indulge in the data and attempt to add to the body of science contained within. And we thank the Center for Dark Energy Biosphere Investigations (C-DEBI) for providing funding to BJT and JFH (OCE-0939654). This is C-DEBI contribution number ###.

Author contributions: BJT conceived of the research plan, performed analysis, and wrote the manuscript. EDG performed analysis and wrote the manuscript. JFH provided funding, provided guidance, and edited the manuscript.

Competing financial interests: The authors declare no conflict of interest.

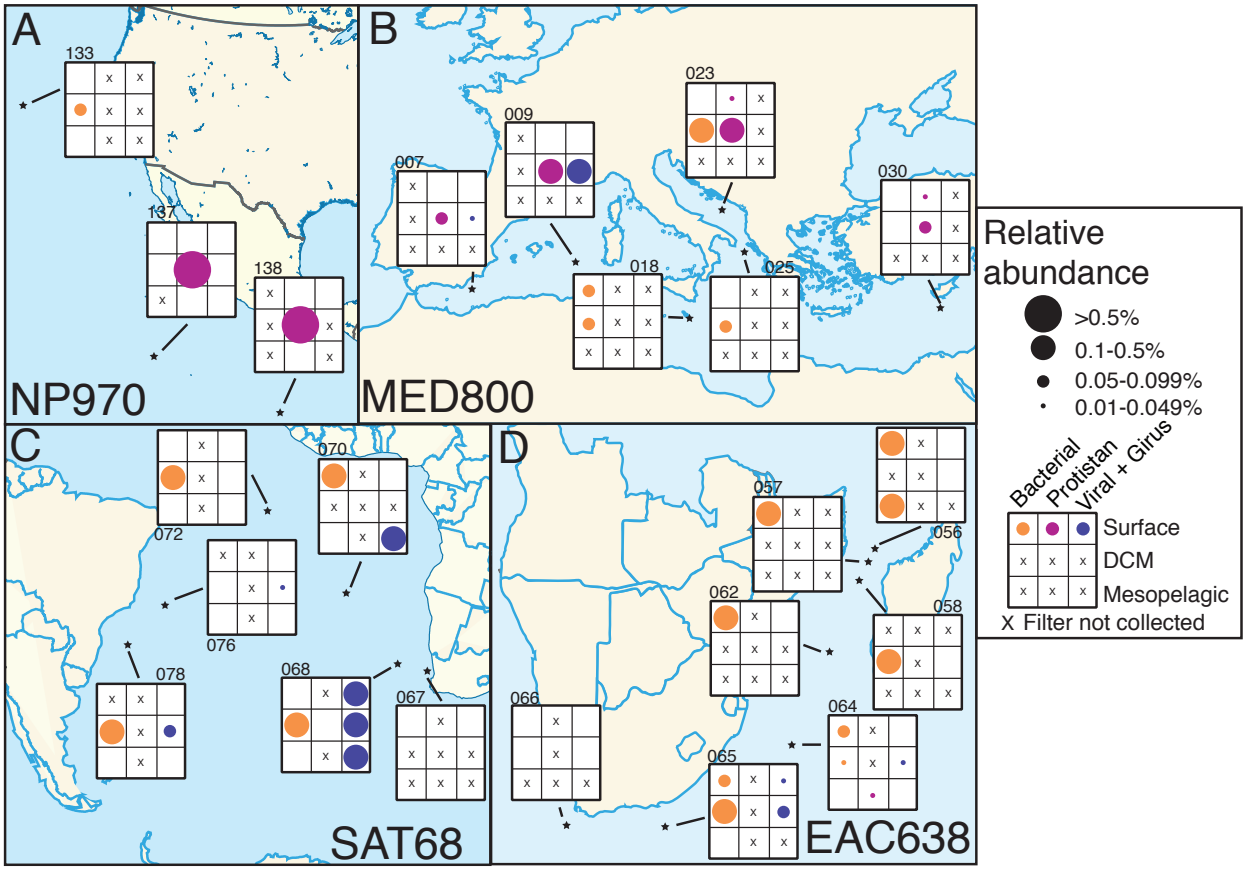


Fig. 1. The approximate locations of *Tara Oceans* sampling sites used to generate metagenomes incorporated in to this study. Each grid represents the three possible sample depths and filter fractions (top row: surface, middle row: DCM, bottom row: mesopelagic). An 'X' denotes that no sample was collected for that depth and size fraction at the site. Circle size represents relative abundance. (A) North Pacific – NP 970, (B) Mediterranean Sea – MED800, (C) South Atlantic – SAT68, and (D) East Africa current – EAC638. Size fraction: orange, 'bacterial' (0.22-1.6µm); purple, 'protistan' (0.8-5.0µm); blue, 'girus'+'viral' (<0.22-0.8µm).

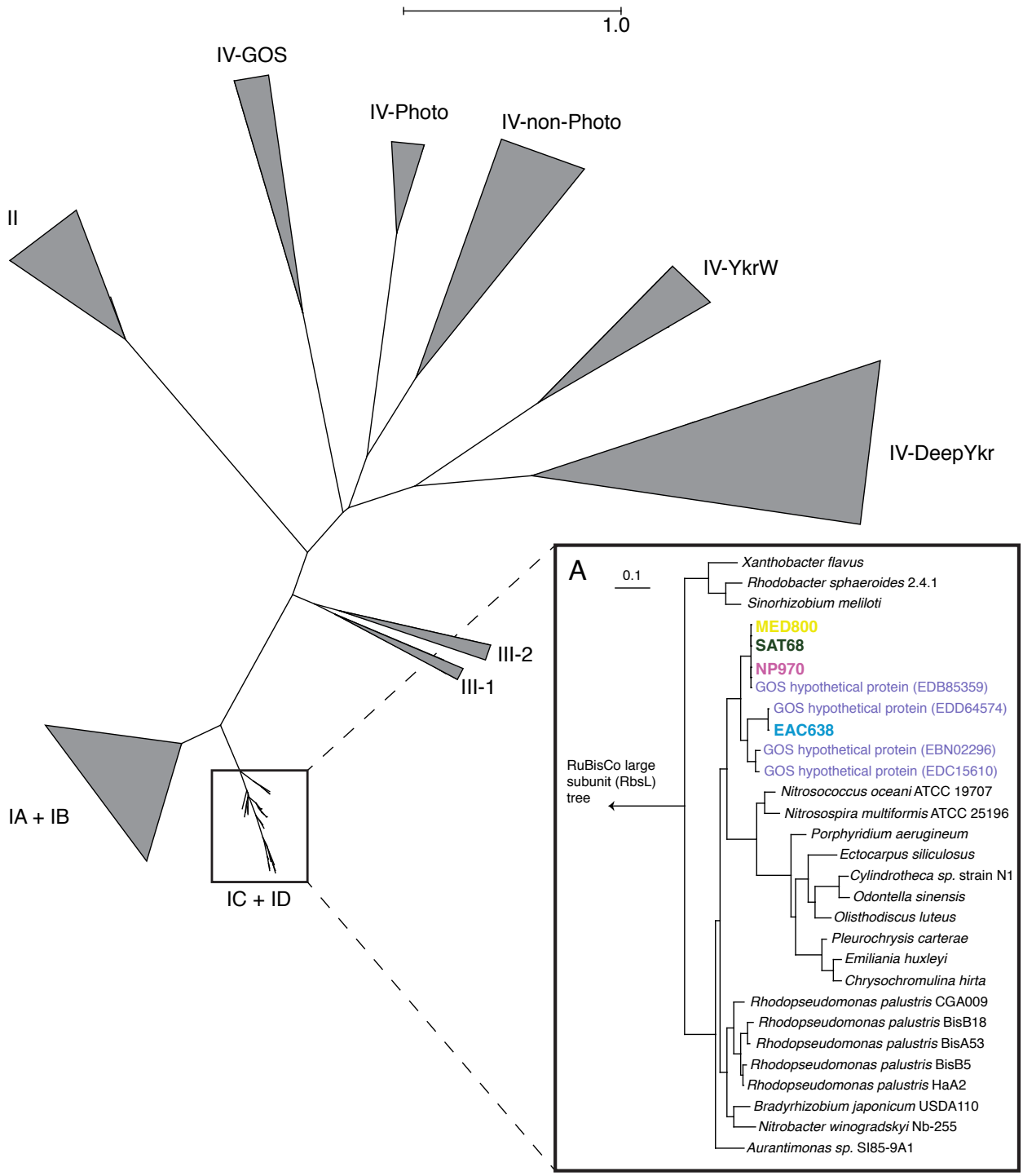
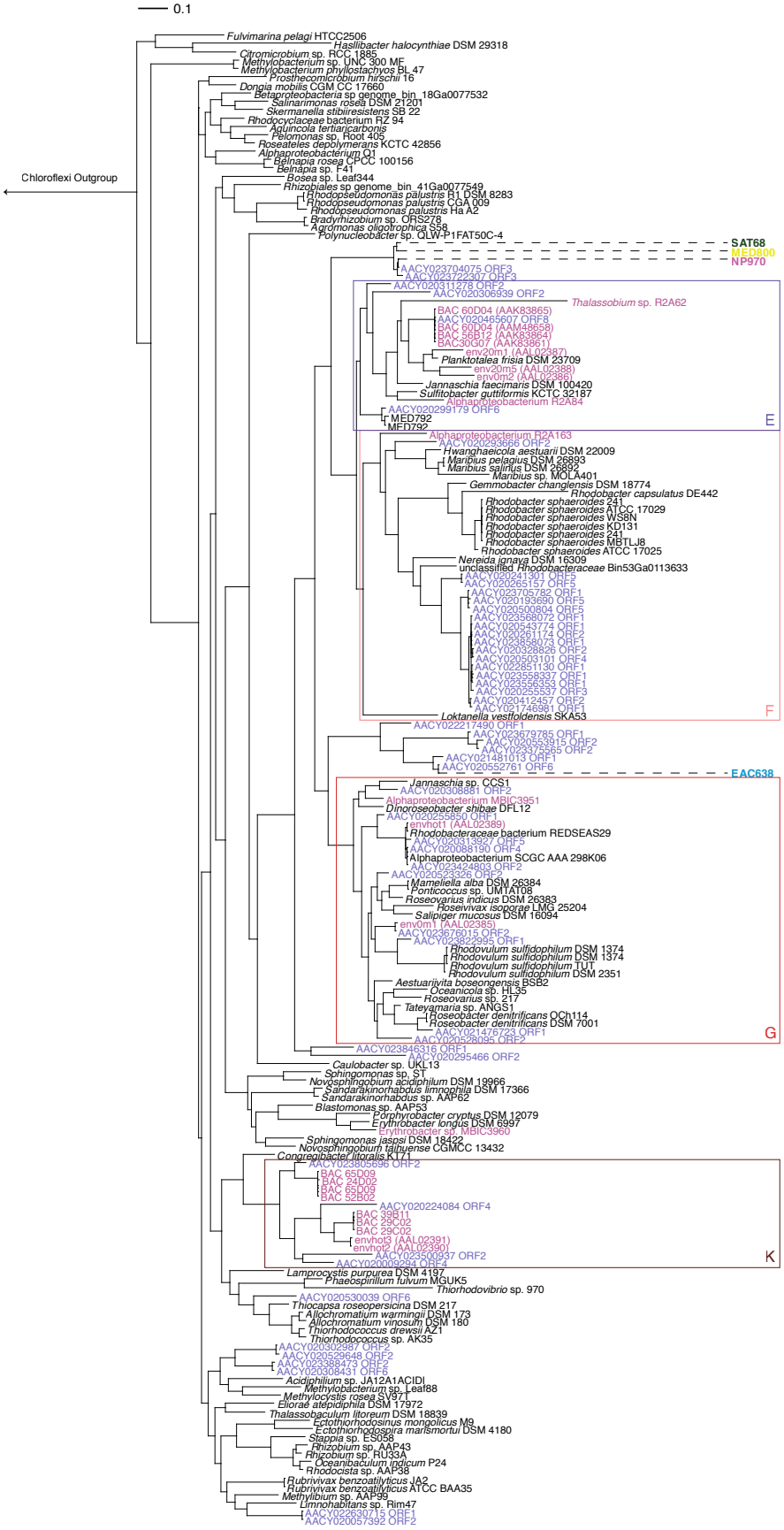


Fig. 2. Phylogenetic tree of the ribulose-1,5-bisphosphate carboxylase large subunit (Rubsico, RbsL) with the major forms denoted. Inset (**A**) – A zoomed in view of the Type 1C/D Rubsico subgroup. Purple sequence names denote RbsL proteins from the Global Ocean Survey. Sequences used for this tree can be found in Supplementary Information 10. Phylogenetic distances and local support values can be found in Supplementary Information 12. Sequence information, including accession numbers and assignments can be found in Supplementary Information 8.


0.1
Chloroflexi Outgroup
Fulvimarina pelagi HTCC2506
Hasllibacter halocynthiae DSM 29318
Citromicrobium sp. RCC 1885
Methylobacterium sp. UNC 300 MF
Methylobacterium phyllostachyos BL 47
Prosthecomicrobium hirschii 16
Dongia mobilis CGM CC 17660
Betaproteobacteria sp genome_bin_18Ga0077532
Salinarimonas rosea DSM 21201
Skermanella stibiireesistens SB 22
Rhodocyclaceae bacterium RZ 94
Aquincola tertiaricarbonis
Pelomonas sp. Root 405
Roseateles depolymerans KCTC 42856
Alphaproteobacterium QI
Belnapia rosea CPCC 100156
Belnapia sp. F41
Bosea sp. Leaf344
Rhizobiales sp genome_bin_41Ga0077549
Rhodopseudomonas palustris R1 DSM 8283
Rhodopseudomonas palustris CGA 009
Rhodopseudomonas palustris Ha A2
Bradyrhizobium sp. ORS278
Agromonas oligotrophica S58
Polynucleobacter sp. QLW-P1FAT50C-4
SAT68
MED800
NP970
AACY023704075 ORF3
AACY023722307 ORF3
AACY020311278 ORF2
AACY020306939 ORF2
Thalassobium sp. R2A62
BAC 60D04 (AAK83865)
AACY020465607 ORF8
BAC 60D04 (AAM48658)
BAC 56B12 (AAK83864)
BAC30G07 (AAK83861)
env20m1 (AAL02387)
Planktotalea frisia DSM 23709
env20m5 (AAL02388)
env0m2 (AAL02386)
Jannaschia faecimaris DSM 100420
Sulfitobacter guttiformis KCTC 32187
Alphaproteobacterium R2A84
AACY020299179 ORF6
MED792
MED792
E
Alphaproteobacterium R2A163
AACY020293666 ORF2
Hwanghaeicola aestuarii DSM 22009
Maribius pelagius DSM 26893
Maribius salirius DSM 26892
Maribius sp. MOLA401
Gemmobacter changlensis DSM 18774
Rhodobacter capsulatus DE442
Rhodobacter sphaeroides 241
Rhodobacter sphaeroides ATCC 17029
Rhodobacter sphaeroides WS8N
Rhodobacter sphaeroides KD131
Rhodobacter sphaeroides 241
Rhodobacter sphaeroides MBTLJ8
Rhodobacter sphaeroides ATCC 17025
Nereida ignava DSM 16309
unclassified rhodobacteraceae Bin53Ga0113633
AACY020241301 ORF5
AACY020265157 ORF5
AACY023705782 ORF1
AACY020193690 ORF5
AACY020500804 ORF5
AACY023568072 ORF1
AACY020543774 ORF1
AACY020261174 ORF2
AACY023858073 ORF1
AACY020328826 ORF2
AACY020509101 ORF4
AACY022851130 ORF1
AACY023558337 ORF1
AACY023556353 ORF1
AACY020255537 ORF3
AACY020412457 ORF2
AACY021746981 ORF1
Loktanella vestfoldensis SKA53
F
AACY022217490 ORF1
AACY020379785 ORF1
AACY020553915 ORF2
AACY023375565 ORF2
AACY021481013 ORF1
AACY020552761 ORF6
EAC638
Jannaschia sp. CCS1
AACY020308861 ORF2
Alphaproteobacterium MBIC3951
Dinoroseobacter shibae DFL12
AACY020255850 ORF1
envhot1 (AAL02389)
Rhodobacteraceae bacterium REDSEAS29
AACY020313927 ORF5
AACY020088190 ORF4
Alphaproteobacterium SCGC AAA 298K06
AACY023424809 ORF2
AACY020523326 ORF2
Mameliella alba DSM 26384
Ponticoccus sp. UMTA106
Roseovarius indicus DSM 26383
Roseovivax isopogae LMG 25204
Salipiger mucosus DSM 16094
env0m1 (AAL02385)
AACY023676015 ORF2
AACY023822995 ORF1
Rhodovulum sulfidophilum DSM 1374
Rhodovulum sulfidophilum DSM 1374
Rhodovulum sulfidophilum TUT
Rhodovulum sulfidophilum DSM 2351
Aestuarijivita boseongensis BSB2
Oceanicola sp. HL35
Roseovarius sp. 217
Tateyamaria sp. ANGS1
Roseobacter denitrificans OCh114
Roseobacter denitrificans DSM 7001
AACY021475723 ORF1
AACY020528085 ORF2
G
AACY023846316 ORF1
AACY020295466 ORF2
Caulobacter sp. UKL13
Sphingomonas sp. ST
Novosphingobium acidiphilum DSM 19966
Sandarakinorhabdus limnophila DSM 17366
Sandarakinorhabdus sp. AAP62
Blastomonas sp. AAP53
Porphyrobacter cryptus DSM 12079
Erythrobacter longus DSM 6997
Erythrobacter sp. MBIC3960
Sphingomonas jaseai DSM 18422
Novosphingobium taihuense CGMCC 13432
Congregibacter litoralis KT71
AACY023806696 ORF2
BAC 65D09
BAC 24D02
BAC 65D09
BAC 52B02
AACY020224084 ORF4
BAC 39B11
BAC 28C02
BAC 29C02
envhot3 (AAL02391)
envhot2 (AAL02390)
AACY023500937 ORF2
AACY020009294 ORF4
K
Lamprocystis purpurea DSM 4197
Phaeospirillum fulvum MGUK5
Thiorhodovibrio sp. 970
AACY020530039 ORF6
Thiocapsa roseopersicina DSM 217
Allochromatium warmingii DSM 173
Allochromatium vinosum DSM 180
Thiorhodococcus drewsii AZ1
Thiorhodococcus sp. AK35
AACY020302987 ORF2
AACY020529648 ORF2
AACY023388473 ORF2
AACY020308431 ORF6
Acidiphilium sp. JA12A1ACIDI
Methylobacterium sp. Leaf88
Methylocystis rosea SV97
Eliorae atepidiphila DSM 17972
Thalassobaculum litoreum DSM 18839
Ectothiorhodosinus mongolicus M9
Ectothiorhodospira marismortui DSM 4180
Stappia sp. ES058
Rhizobium sp. AAP43
Rhizobium sp. RU33A
Oceanibaculum indicum P24
Rhodocista sp. AAP38
Rubrivivax benzoatilyticus JA2
Rubrivivax benzoatilyticus ATCC BAA35
Methylibium sp. AAP99
Limnohabitans sp. Rim47
AACY020263715 ORF1
AACY020057392 ORF2

Extended Data Fig. 3. Phylogenetic tree of the M subunit of type-II photochemical reaction center (PufM). Environmental sequences obtained from the Global Ocean Survey (purple, •) and *Béjá et al.* (2002) (pink, •) are highlighted. Boxes illustrate approximate positions of phylogroups previously assigned by Yutin *et al.* (2007). Sequences used for this tree can be found in Supplementary Information 11. Phylogenetic distances and local support values can be found in Supplementary Information 13. Sequence information, including accession numbers and taxonomies can be found in Supplementary Information 9.

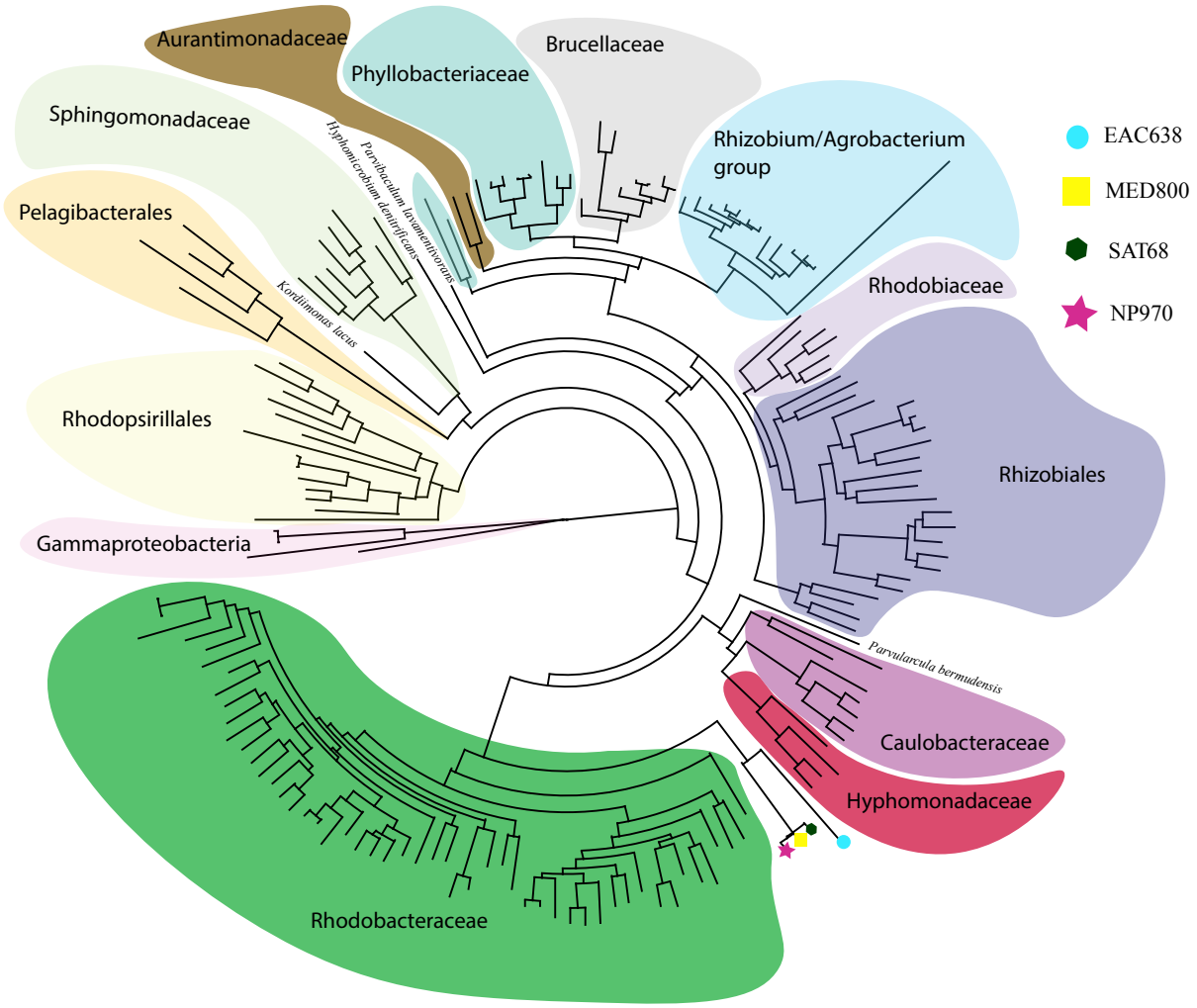


Fig. 4. Approximate maximum likelihood *Alphaproteobacteria* phylogenetic tree of 17 concatenated single-copy marker genes for the *Tara* assembled and 160 reference genomes. Reference sequences from the *Gammaproteobacteria* used as an outgroup. Reference genome information, including accession numbers, can be found in Supplementary Information 4. Phylogenetic distances and local support values can be found in Supplementary Information 5.

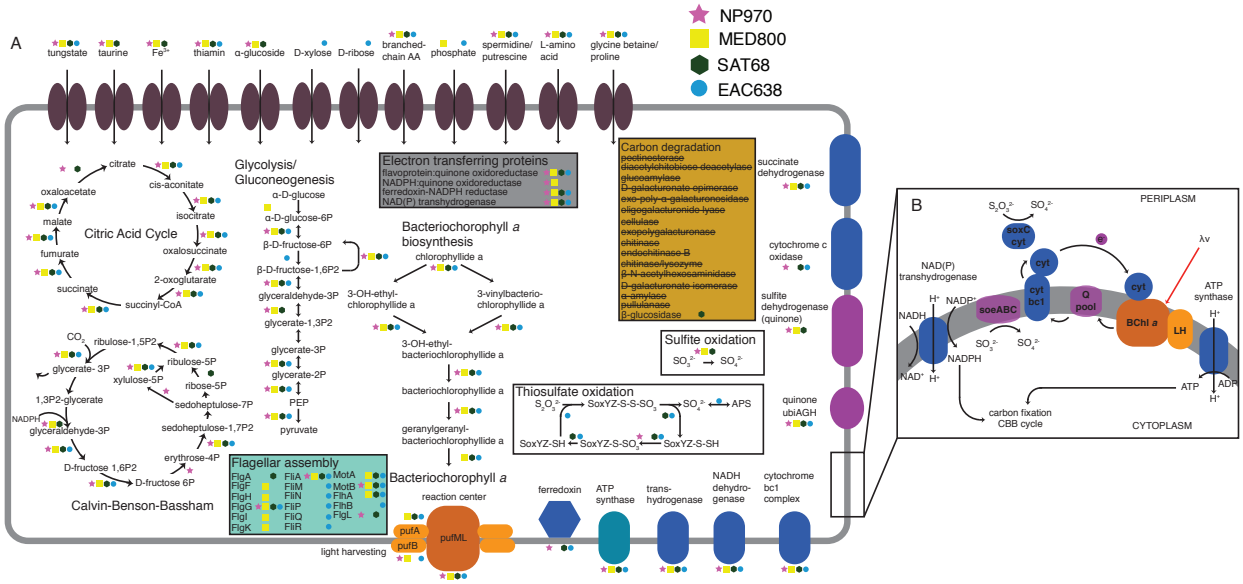


Fig. 5. Cellular schematic of the four reconstructed genomes. (**A**) The presence of a gene(s) in a genome is represented by a yellow square (MED800), pink star (NP970), green hexagon (SAT68), and/or blue circle (EAC638). Schematic illustrates predicted membrane bound proteins, but does accurately represent cellular localization. (**B**) A detailed view of the proposed flow of electrons from donors to photosynthesis and carbon fixation. Abbreviations: cyt, cytochrome; Q pool, quinone pool; BChl *a*, bacteriochlorophyll a; LH, light-harvesting proteins; soeABC, sulfite dehydrogenase (quinone); CBB, Calvin-Benson-Bassham.

420 **Table 1**. Statistics of the four *Tara* assembled genomes.

Genome ID	No. of contigs	Total length (bp)	Max. contig length (bp)	N50	Mean length (bp)	GC (%)	No. of predicted CDS◊	Est. Completeness (%)*	Est. duplication/redundancy (Est. strain heterogeneity) (%)*
MED800	**174**	**2,140,579**	**55,951**	**13,831**	**12,302**	**36.7**	**2,184**	**66.98**	**2.36 (100.0)**
SAT68	**192**	**2,295,506**	**63,422**	**13,440**	**11,956**	**36.8**	**2,334**	**85.02**	**5.37 (75.0)**
EAC638	**121**	**1,868,409**	**80,916**	**19,502**	**15,441**	**30.7**	**1,939**	**68.88**	**4.75 (75.0)**
NP970	**120**	**2,585,639**	**84,649**	**32,137**	**21,547**	**36.1**	**2,445**	**82.04**	**3.04 (33.33)**

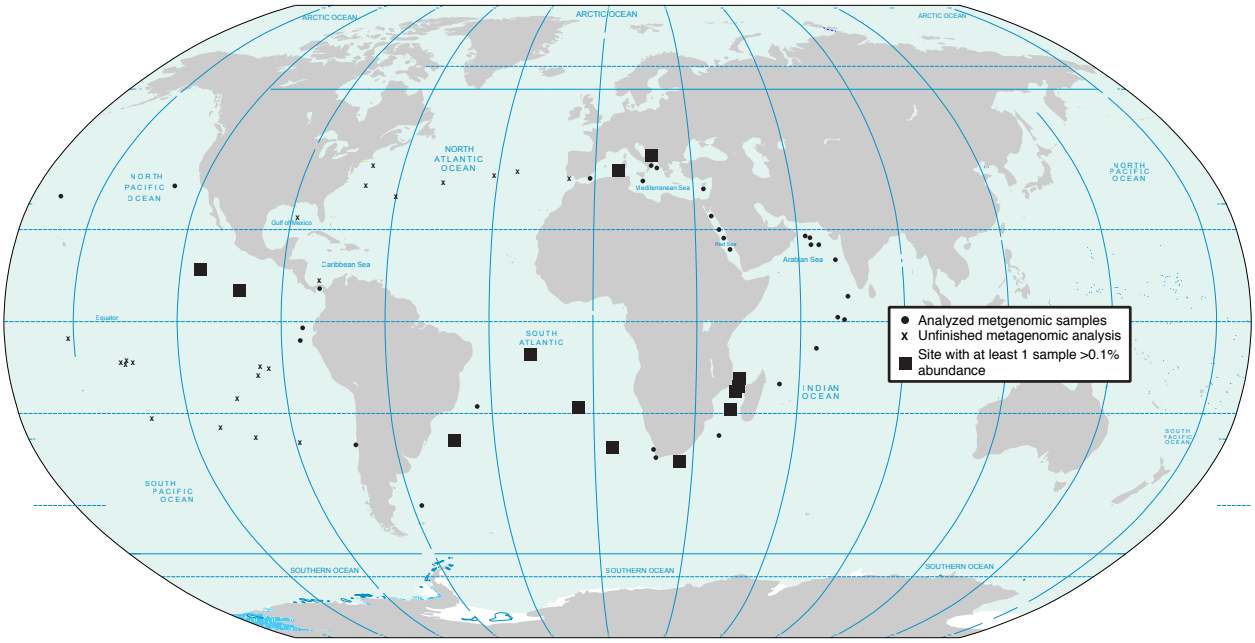
421 ◊, as determined using Prodigal
422 *, as determine using CheckM with the Alphaproteobacteria markers (225 markers in 148 sets)
423 N50 - length of contig for which all contigs longer in length contain half of the total genome
424 CDS - coding DNA sequence
425

Supplementary Materials:

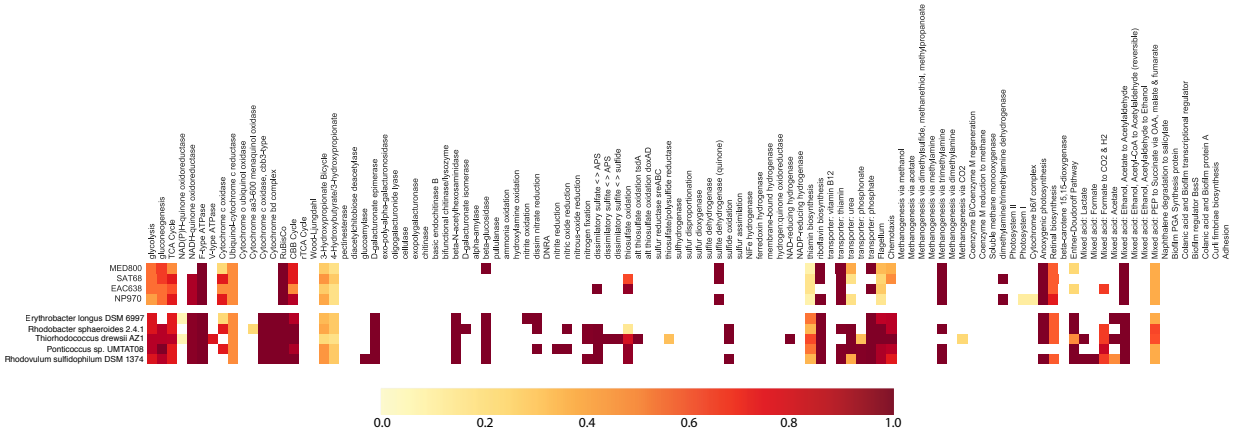
Extended Data Figures 1-6

Extended Data Table 1

Supplementary Information 1-13



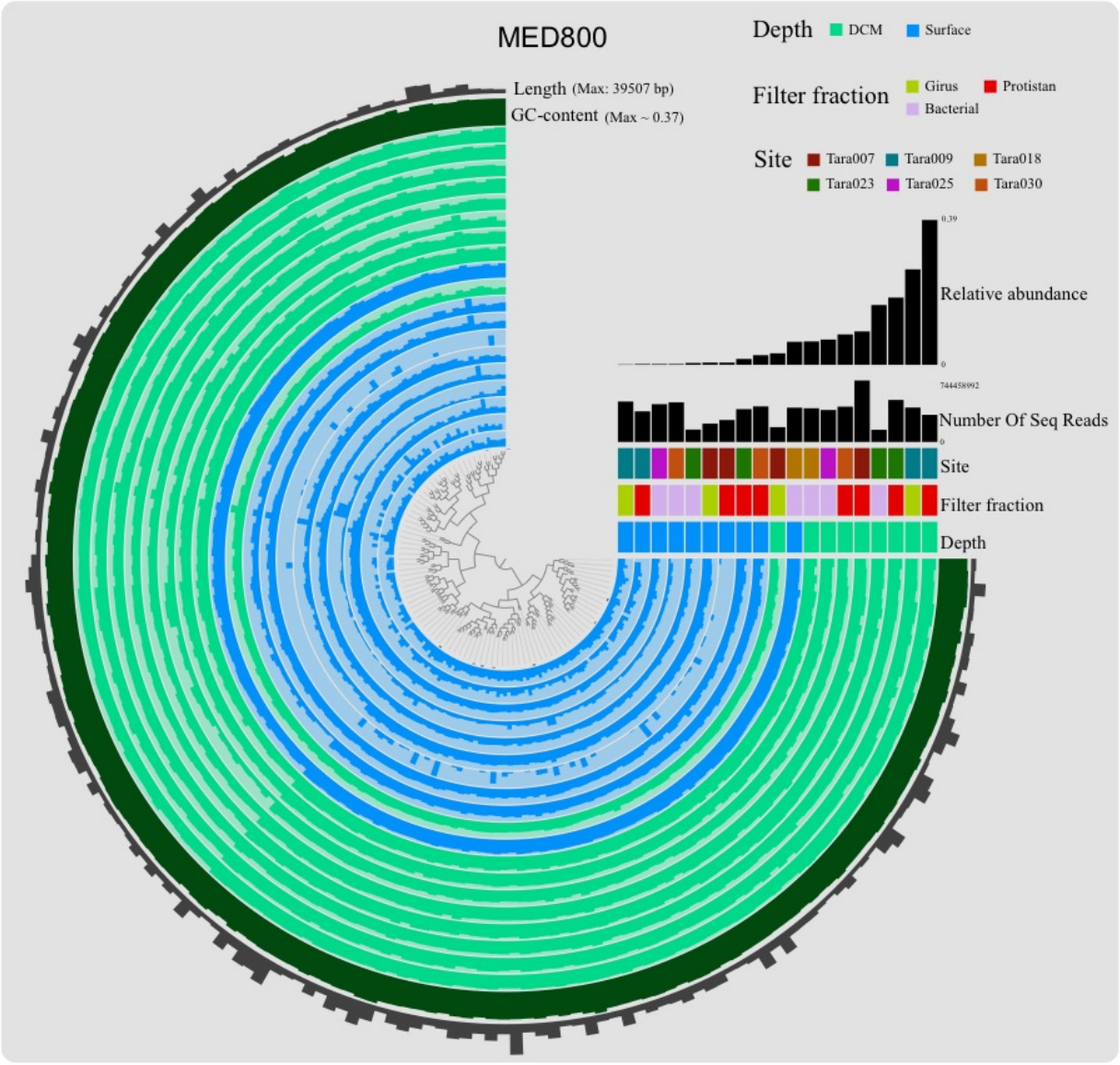
Extended Data Fig. 1. A map illustrating the approximate locations of the *Tara Oceans* sampling stations. Stations and samples with >0.1% relative abundance of the *Tara* assembled genomes are emphasized. Circles - stations for which metagenomes have been assembled, genomes have been reconstructed, and cursory functional analysis has been performed. Exes – stations awaiting assembly, binning, and analysis. Squares – stations where at least 1 sample had >0.1% abundance of the *Tara* assembled genomes.



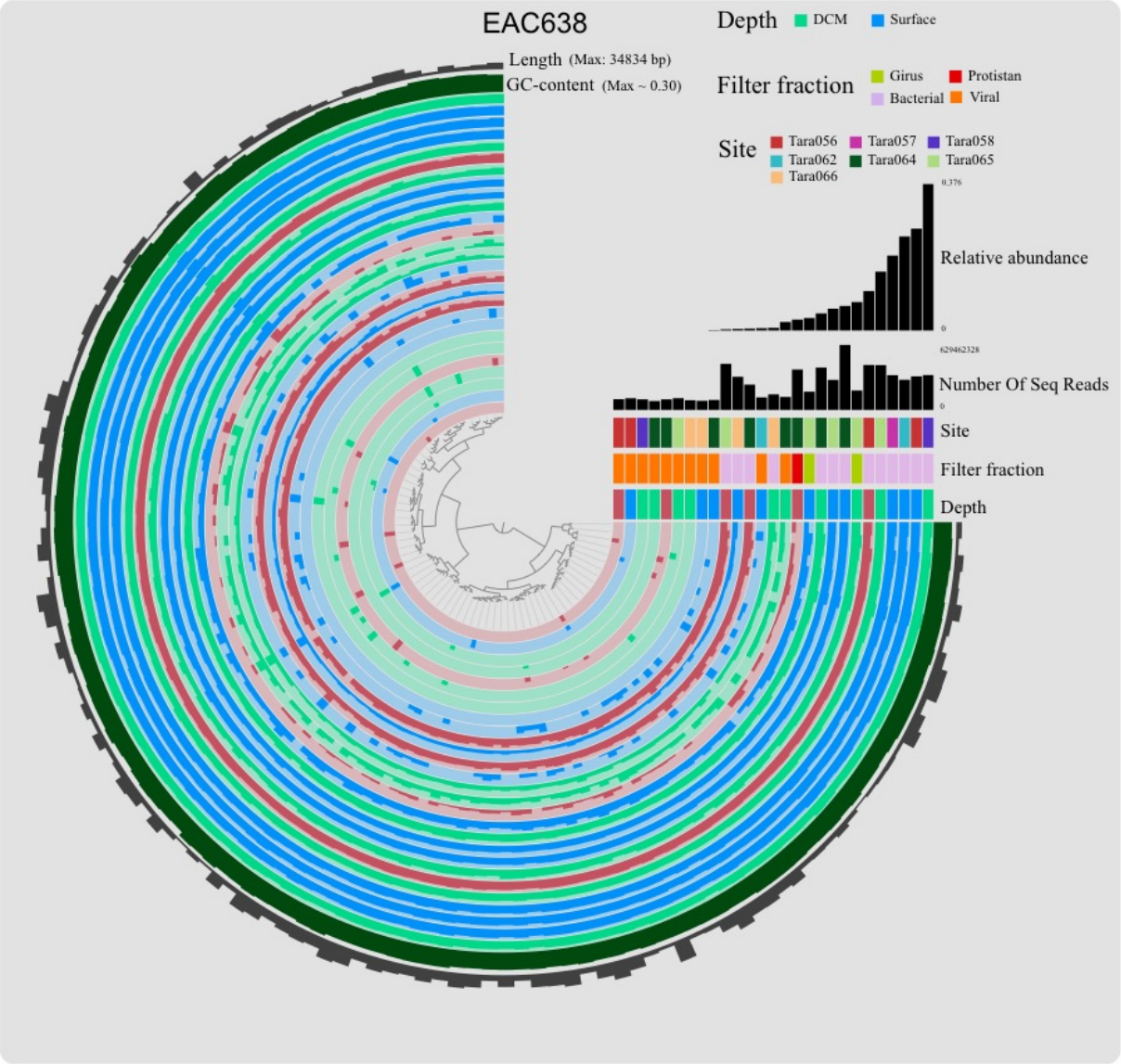
438

439 **Extended Data Fig. 2**. A heat map demonstrating the degree of completeness of various
440 pathways and genes determined for the four *Tara* assembled genomes. Additionally, the same
441 process was applied to genomes of confirmed anaerobic anoxygenic photoautotrophs for
442 comparison. Pathways are linked to KEGG Ontology identifiers and the figure was generated
443 using the script KEGG-decoder.py.

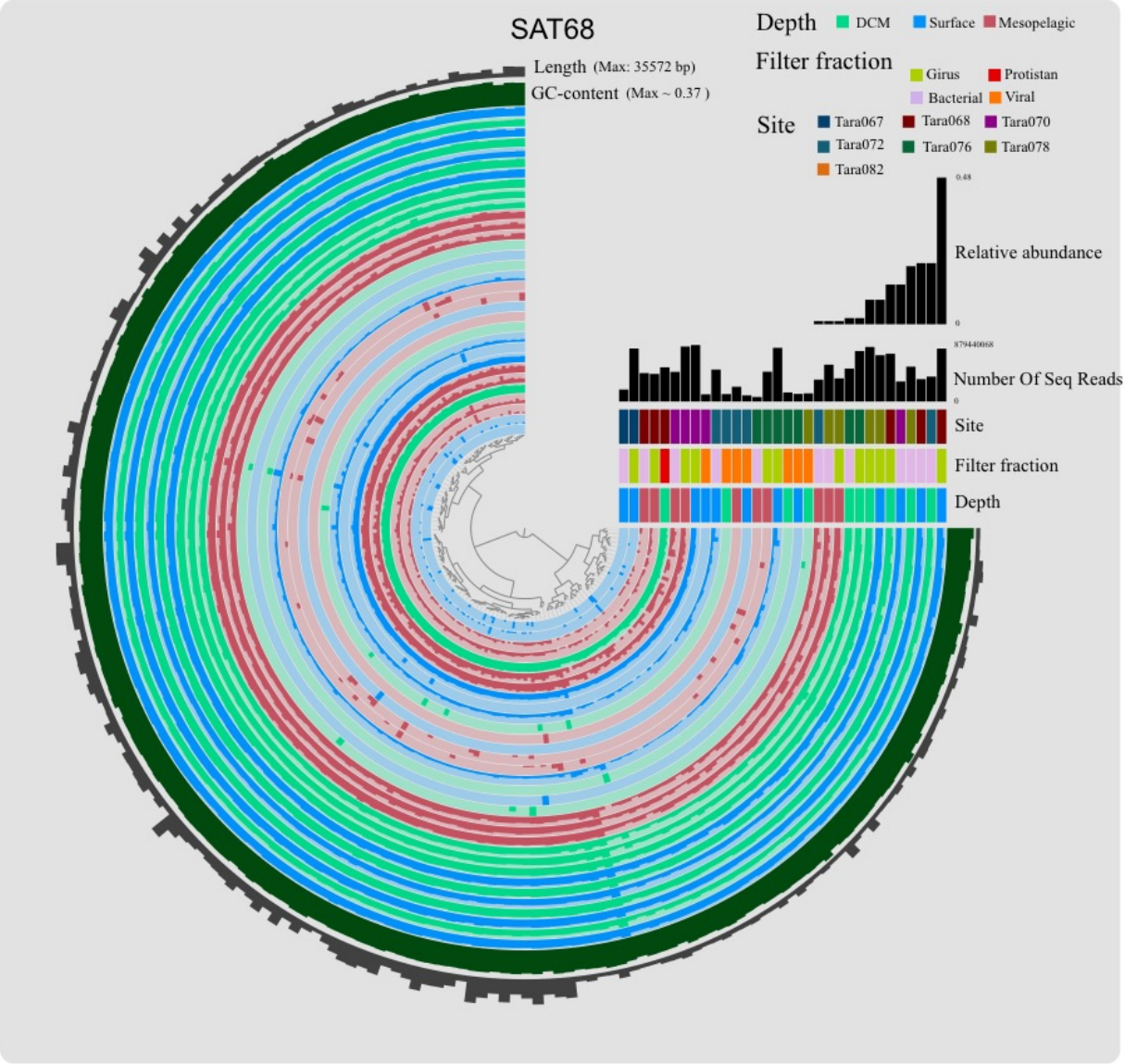
444



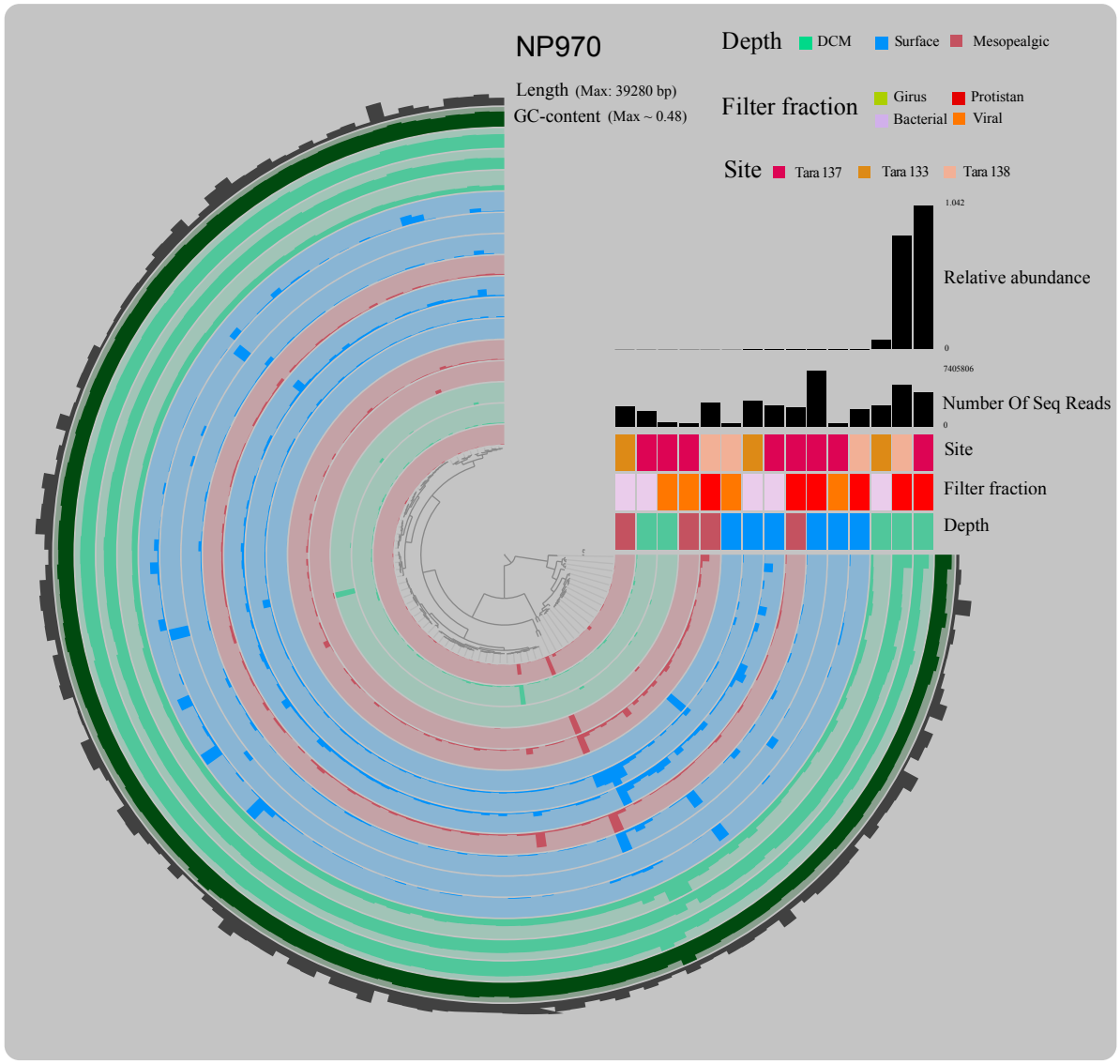
Extended Data Fig. 3. Visualization of MED800 using Anvi'o. Contigs are clustered in Anvi'o using Ward distance hierarchical clustering combining sequence composition and coverage information. Standardized coverage is shown for each sample. Relative abundance, number of sequencing reads generated, site, filter fraction, and depth are all shown. Completion, contamination, and strain heterogeneity were produced using CheckM.



Extended Data Fig. 4. Visualization of EAC638 using Anvi'o. Contigs are clustered in Anvi'o using Ward distance hierarchical clustering combining sequence composition and coverage information. Standardized coverage is shown for each sample. Relative abundance, number of sequencing reads generated, site, filter fraction, and depth are all shown. Completion, contamination, and strain heterogeneity were produced using CheckM.



Extended Data Fig. 5. Visualization of SAT68 using Anvi'o. Contigs are clustered in Anvi'o using Ward distance hierarchical clustering combining sequence composition and coverage information. Standardized coverage is shown for each sample. Relative abundance, number of sequencing reads generated, site, filter fraction, and depth are all shown. Completion, contamination, and strain heterogeneity were produced using CheckM.



Extended Data Fig. 6. Visualization of NP970 using Anvi'o. Contigs are clustered in Anvi'o using Ward distance hierarchical clustering combining sequence composition and coverage information. Standardized coverage is shown for each sample. Relative abundance, number of sequencing reads generated, site, filter fraction, and depth are all shown. Completion, contamination, and strain heterogeneity were produced using CheckM.

475

476 **Extended Data Table 1**. Single-copy marker genes identified in each of the putative AAnP and
477 reference genomes with corresponding Pfam identifiers.

478

Marker Gene Names	
Aminoacyl tRNA synthetase class II, N-terminal domain	**PF02912**
GTP1/OBG	**PF01018**
RecA	**PF00154**
Ribosomal L17	**PF01196**
Ribosomal L25p family	**PF01386**
Ribosomal L28 protein	**PF00830**
Ribosomal protein L20	**PF00453**
Ribosomal protein L31	**PF01197**
Ribosomal protein L35	**PF01632**
Ribosomal protein L6	**PF00347**
Ribosomal protein S11	**PF00411**
Ribosomal protein S13/S18	**PF00416**
Ribosomal protein S14	**PF00253**
Ribosomal protein S15	**PF00312**
Ribosomal protein S4/S9 N-terminal domain	**PF00163**
Ribosomal protein S5, N-terminal domain	**PF00333**
RNA polymerase Rpb3/Rpb11 dimerisation domain	**PF01193**

479

Supplementary Information 1. Information and statistics of the Tara stations and samples from the North Pacific, Mediterranean Sea, East Africa Coastal province, and South Atlantic. Includes, station, size fraction, depth, number of paired-end reads assembled, number of assemblies generated by Megahit, and relative abundance values for each genome.

Supplementary Information 2. Information about genes specifically referenced in manuscript and figures. Includes, KO identifiers, gene names, and gene abbreviations.

Supplementary Information 3. BlastKOALA annotations of the Prodigal predicted genes for each of the four genomes.

Supplementary Information 4. List of 2,889 reference genomes used to construct the *Alphaproteobacteria* phylogenetic tree.

Supplementary Information 5. Newick format of the *Alphaproteobacteria* phylogenetic tree containing 2,889 reference genomes. Contains branch lengths and local support values.

Supplementary Information 6. List of 85 reference genomes used to construct the *Alphaproteobacteria* phylogenetic tree in Figure 2.

Supplementary Information 7. Newick format of the *Alphaproteobacteria* phylogenetic tree containing X reference genomes. Contains branch lengths and local support values.

Supplementary Information 8. List of reference sequences used to construct the ribulose-1,5-bisphosphate carboxylase large subunit phylogenetic tree. Sequences collected from Tabita *et. al.* (2007).

Supplementary Information 9. List of reference sequences used to construct the photosynthetic reaction center M subunit phylogenetic tree. Sequences collected from Yutin *et. al.* (2007).

Supplementary Information 10. Sequences used to construct the ribulose-1,5-bisphosphate carboxylase large subunit phylogenetic tree in FASTA format.

Supplementary Information 11. Sequences used to construct the photosynthetic reaction center M subunit phylogenetic tree in FASTA format.

Supplementary Information 12. Newick format of the ribulose-1,5-bisphosphate carboxylase large subunit phylogenetic tree. Contains branch lengths and local support values.

Supplementary Information 13. Newick format of the photosynthetic reaction center M subunit phylogenetic tree. Contains branch lengths and local support values.