Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean

Brandon K. Swan[a], Ben Tupper[a], Alexander Sczyrba[b], Federico M. Lauro[c], Manuel Martinez-Garcia[d], José M. González[e], Haiwei Luo[f], Jody J. Wright[g], Zachary C. Landry[h], Niels W. Hanson[i], Brian P. Thompson[a], Nicole J. Poulton[a], Patrick Schwientek[j], Silvia G. Acinas[k], Stephen J. Giovannoni[h], Mary Ann Moran[f], Steven J. Hallam[g,i], Ricardo Cavicchioli[c], Tanja Woyke[j], and Ramunas Stepanauskas[a,1]

[a]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544; [b]Center for Biotechnology, Bielefeld University, 33615 Bielefeld, Germany; [c]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia; [d]Department of Physiology, Genetics and Microbiology, University of Alicante, 03080 Alicante, Spain; [e]Department of Microbiology, University of La Laguna, ES-38206 La Laguna, Tenerife, Spain; [f]Department of Marine Sciences, University of Georgia, Athens, GA 30602; [g]Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; [h]Department of Microbiology, Oregon State University, Corvallis, OR 97331; [i]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; [j]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; and [k]Department of Marine Biology and Oceanography, Institute of Marine Science, Consejo Superior de Investigaciones Científicas, ES-08003 Barcelona, Spain

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved May 28, 2013 (received for review March 7, 2013)

Planktonic bacteria dominate surface ocean biomass and influence global biogeochemical processes, but remain poorly characterized owing to difficulties in cultivation. Using large-scale single cell genomics, we obtained insight into the genome content and biogeography of many bacterial lineages inhabiting the surface ocean. We found that, compared with existing cultures, natural bacterioplankton have smaller genomes, fewer gene duplications, and are depleted in guanine and cytosine, noncoding nucleotides, and genes encoding transcription, signal transduction, and noncytoplasmic proteins. These findings provide strong evidence that genome streamlining and oligotrophy are prevalent features among diverse, free-living bacterioplankton, whereas existing laboratory cultures consist primarily of copiotrophs. The apparent ubiquity of metabolic specialization and mixotrophy, as predicted from single cell genomes, also may contribute to the difficulty in bacterioplankton cultivation. Using metagenome fragment recruitment against single cell genomes, we show that the global distribution of surface ocean bacterioplankton correlates with temperature and latitude and is not limited by dispersal at the time scales required for nucleotide substitution to exceed the current operational definition of bacterial species. Single cell genomes with highly similar small subunit rRNA gene sequences exhibited significant genomic and biogeographic variability, highlighting challenges in the interpretation of individual gene surveys and metagenome assemblies in environmental microbiology. Our study demonstrates the utility of single cell genomics for gaining an improved understanding of the composition and dynamics of natural microbial assemblages.

comparative genomics | marine microbiology | microbial ecology | microbial microevolution | operational taxonomic unit

Planktonic bacteria dominate surface ocean biomass and have a major impact on the global cycling of carbon, nitrogen, and other elements (1). Among the available pure cultures of marine bacterioplankton, only a limited number represent bacterioplankton that are abundant in the ocean, such as the cyanobacteria *Prochlorococcus* and *Synechococcus* and the Alphaproteobacteria *Pelagibacter* (collectively termed PSP cultures). This limits the scope of studies of the microbial metabolic processes and evolutionary changes that impact marine ecosystems and their geochemical cycles (2–6). Unusual nutritional requirements resulting from genome reduction may contribute to cultivation difficulties, as suggested by studies of the chemoheterotroph *Pelagibacter* (7, 8) and the methylotroph OM43 (9).

Although prevailing culture-independent tools, including microbial community shotgun sequencing, targeted gene surveys, and fluorescent in situ hybridization, have revealed the extent and significance of microbial diversity, they have not been able to provide the genome context information required for accurate metabolic reconstruction spanning organismal, population, and community levels of organization (10). As a result, the genomic repertoires, natural histories, and geographic distribution of even the most abundant taxonomic groups of marine bacterioplankton remain largely unknown (1, 11). Microbial studies in other environments, such as the human body and soils, face similar challenges (10). The recent development of robust protocols for single cell genomics provides a versatile, cultivation-independent approach for assessing natural microbial diversity with corresponding genome context information (12).

To determine whether genome streamlining is a prevalent feature among free-living marine bacterioplankton, and to analyze global patterns of surface ocean bacterioplankton distribution, we obtained draft genomes of 56 single amplified genomes (SAGs) (5, 13–15) and compared them with existing bacterioplankton cultures and metagenomes. The sequenced SAGs represent many ubiquitous surface ocean bacteria lineages, including Marine Group A, Verrucomicrobia, Actinobacteria, Bacteroidetes, and Proteobacteria lineages SAR86, ARCTIC96BD-19, SAR92, SAR116, and Roseobacter (*SI Appendix*, Fig. S1). The majority of these groups have few or no cultured representatives. Members of the PSP group were excluded from SAG selection, because their genome streamlining and environmental abundance have been demonstrated previously (1, 2, 4, 11). Samples for SAG generation were collected from the Gulf of Maine, the Mediterranean Sea, and the subtropical gyres of the North Pacific and South Atlantic Oceans (*SI Appendix*, Table S1). On average, 55% (range, 0.3–97.8%) of the genome was recovered from each analyzed cell (*SI Appendix*, Table S2). A subset of 41 SAGs, each >0.75 Mbp in size and with >30% estimated genome recovery, was used for our

Author contributions: B.K.S., B.T., A.S., and R.S. designed research; B.K.S., B.T., A.S., F.M.L., M.M.-G., J.M.G., H.L., J.J.W., Z.C.L., N.W.H., B.P.T., N.J.P., S.G.A., T.W., and R.S. performed research; B.K.S., B.T., A.S., T.W., and R.S. contributed new reagents/analytic tools; B.K.S., B.T., A.S., F.M.L., M.M.-G., J.M.G., H.L., J.J.W., Z.C.L., N.W.H., P.S., S.G.A., S.J.G., M.A.M., S.J.H., R.C., T.W., and R.S. analyzed data; and B.K.S. and R.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Whole-genome sequence data for single amplified genomes used for our analyses are available in the Joint Genome Institute's Integrated Microbial Genome database, http://img.jgi.doe.gov/cgi-bin/w/main.cgi (accession nos. 643886079, 643886118, 2228664025-26, 2228664028-29, 2228664032, 2228664034, 2228664052-53, 2228664055-56, 2236347001, 2236347003, 2236347013, 2236347015, 2236347017-19, 2236347021-24, 2236347026-27, 2236347030-33, 2236347035-36, 2236347039, 2236347041, 2236347043, 2236661010, 2236661014, 2236661017-18, 2507262045, 2507262047, and 2517572139).

[1]To whom correspondence should be addressed. E-mail: rstepanauskas@bigelow.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1304246110/-/DCSupplemental.

comparative genomics and biogeographic analyses. Our results demonstrate that genome streamlining is a prevalent evolutionary strategy among free-living bacterioplankton in the surface ocean. They also suggest that the global distribution of the majority of surface ocean bacterioplankton might not be limited by dispersal and is correlated with temperature and latitude.

Results and Discussion

Genomic Signatures of Streamlining and Oligotrophy Among Uncultured Marine Bacteria. A comparison of general genome features among marine bacterioplankton revealed that the majority of our SAGs clustered with cultures of *Prochlorococcus* and *Pelagibacter*, as well as with the SAR86 SAGs sequenced by Dupont et al. (6) (Fig. 1*A* and *SI Appendix*, Table S3). SAGs segregated from cultures along a principal component axis associated with low guanine and cytosine (GC) content, low percentage of noncoding nucleotides, low fraction of genes encoding periplasm and cytoplasm membrane proteins, and Clusters of Orthologous Groups (COG) categories K (transcription) and T (signal transduction). These genomic signatures have been identified as indicators of genome streamlining and oligotrophy (16, 17). All Verrucomicrobia and Bacteroidetes SAGs, one SAR92 SAG, and all Bacteroidetes cultures clustered separately from other SAGs and cultures (Fig. 1*A*). These genomes are associated with elevated frequency of genes encoding extracellular, outer membrane and multilocation proteins, and COG category V (defense mechanisms), corroborating the previously proposed role of Bacteroidetes (18, 19) and the recently suggested importance of Verrucomicrobia (14) in macromolecule degradation, a process requiring cell surface-associated or extracellular hydrolases. SAGs of the same taxonomic group but retrieved from different geographic locations had similar genomic signatures, indicating that the selection for these signatures operates in both the open ocean and coastal waters, and in diverse climate zones. In contrast, large differences in genomic signatures were found between SAGs and their cultured relatives within each taxonomic group that contains multiple SAGs and cultures, such as Roseobacter, SAR116, and Bacteroidetes (*SI Appendix*, Table S4).

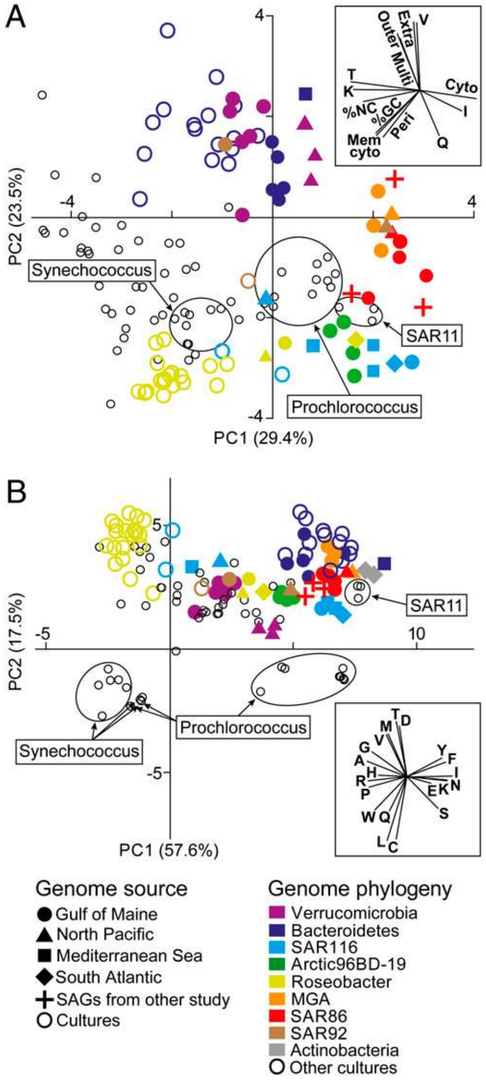


Fig. 1. Genomic differences between SAGs and cultured bacterioplankton. PCA of general genome characteristics (*A*) and encoded amino acid frequency (*B*) of SAGs (solid colored symbols) and cultures of marine bacterioplankton (open circles) are shown. Cultures belonging to the same taxonomic group as SAGs have the same color. The two Actinobacteria SAGs were excluded from the genome characteristics analysis because they are Gram-positive bacteria, which have a different cell wall architecture, and were not included in the development of the trophic strategy model of Lauro et al. (17). (*Insets*) Variable vectors corresponding to each PCA plot. The following input variables were used for the genome characteristic analysis: abundance of genes encoding proteins localized in the cytoplasm; cytoplasmic membrane, periplasm, outer membrane, extracellular, and multiple locations; COG categories I, K, Q, T, and V; %NC, % noncoding DNA.

Obligate oligotrophy has been proposed as a key factor leading to poor recovery of environmental microorganisms in pure cultures (19–21), and our study provides clear evidence for the predominance of a copiotroph lifestyle among existing marine cultures across taxonomic groups. Our data also suggest that oligotroph characteristics in surface ocean bacteria are not limited to members of *Prochlorococcus* and *Pelagibacter* in tropical regions, as previously thought (16, 22), but rather is a common trophic strategy among many bacterioplankton lineages around the globe.

As one of the variables contributing to genomic differences between SAGs and cultures (Fig. 1*A*), the average GC content of SAGs (37.9%) was significantly lower than that of the 101 marine bacterioplankton cultures (48.5%; *SI Appendix*, Fig. S2*A*). Although multiple displacement amplification (MDA) of mixed templates may introduce GC biases, here such biases were eliminated by performing MDA on individual cells, followed by high-coverage sequencing and de novo assembly, which have been demonstrated to accurately reconstruct GC of the analyzed genomes (23–25). The high similarity of the average GC content of SAGs (37.9%) and available surface ocean metagenomes (39.6%) provides further support for the representativeness of our SAG data (*SI Appendix*, Table S5). The difference in %GC between SAGs and cultures was significant in both coding and noncoding genome regions, suggesting GC content rather than protein composition as the primary adaptive trait (*SI Appendix*, Fig. S2*B*).

SAGs differed from cultures in the frequency of encoded amino acids (Fig. 1*B* and *SI Appendix*, Table S6), with SAGs being enriched in tyrosine, phenylalanine, isoleucine, glutamic acid, asparagine, lysine, and serine and depleted in valine, glycine, alanine, arginine, proline, histidine, and tryptophan. These two groups of amino acids were similar in terms of chemical properties, synthesis costs, and numbers of C and N atoms (*SI Appendix*, Table S7), but diverged in average GC content of the first two nucleotides of their codons (14% and 79%, respectively). This finding provides further evidence that differences in amino acid utilization between SAGs and cultures are driven primarily by differences in %GC. Recent experimental work suggests that high GC content may enhance bacterial growth in laboratory conditions (26). In contrast, low genomic GC content may be an adaptation to nitrogen limitation (27) or a result of mutational biases in the absence of effective DNA repair systems (16). It remains to be understood how the observed GC depletion in bacterioplankton and the resulting shifts in amino acid use impact surface ocean processes.

One predicted cost of genome streamlining in free-living bacteria is a reduction in physiological flexibility, leading to specialization in resource utilization. Accordingly, SAGs had fewer paralogs and smaller genomes compared with cultures from the same taxonomic groups, with the exception of SAR116 (Fig. 2). The low paralog frequency is not likely the result of incomplete genome recovery from SAGs, given that partial genes at the ends of contigs may be incorrectly assigned as paralogs, leading to overestimation of paralogs. This effect is evident in the substantially higher fraction of paralogs identified from highly fragmented SAR86 SAG assemblies sequenced by Dupont et al. (6) compared with the SAR86 SAGs reported here. This overall trend suggests that the small genome size and fewer gene duplications may provide an adaptive advantage to life in the oligotrophic ocean.

Comparisons of metabolic potential among taxonomic groups represented by multiple SAGs provide strong evidence for specialized resource utilization despite incomplete genome recovery from individual SAGs (*SI Appendix*, Figs. S3–S8 and Tables S8–S10). For example, Gammaproteobacteria lineages SAR86, SAR92, and ARCTIC96BD-19 encode a heterotrophic central metabolism but differ in terms of pathway completeness and variation. Moreover, genes encoding the oxidative component of the pentose phosphate metabolism are absent in most SAR86 SAGs, but this pathway was found to be complete in most ARCTIC96BD-19 SAGs (*SI Appendix*, Table S9). Evidence of autotrophic carbon fixation was found only in ARCTIC96BD-19 SAGs, which harbor the RuBisCO operon, as previously reported for SAGs of this lineage from the mesopelagic zone (15). Only the SAR116 SAGs encoded form I *coxL*, indicating a functional carbon monoxide dehydrogenase (*SI Appendix*, Fig. S7). Genes supporting various inorganic sulfur utilization pathways were common and lineage-specific, including polysulfide reductase (*psr*) in Marine Group A, the *sox* (sulfur oxidation) operon in SAR116, and adenylylsulfate reductase (*aprA*) among members of ARCTIC96BD-19. Proteorhodopsin genes were found consistently in Marine Group A and ARCTIC96BD-19 SAGs, expanding the taxonomic groups known to encode these photometabolic systems (*SI Appendix*, Fig. S4 and Table S10).

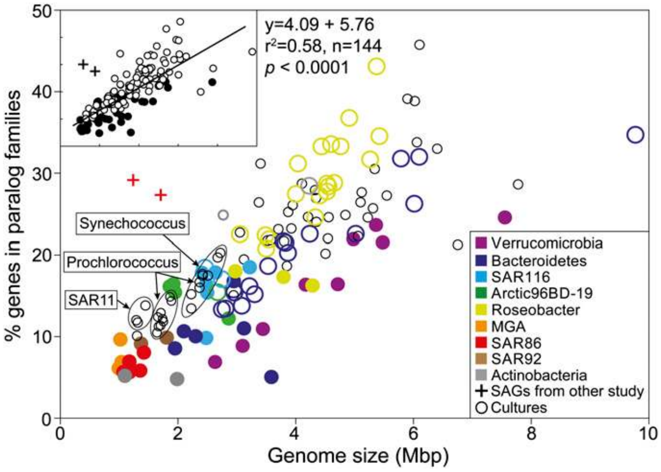


Fig. 2. Genome size and paralogous gene frequency of SAGs and bacterioplankton cultures. The percentages of genes belonging to paralog families in SAGs (solid colored circles) and cultures (open circles) were estimated using BLASTCLUST. Cultures belonging to the same taxonomic group as SAGs have the same color. (*Inset*) Results of least squares linear regression between genome size and paralog frequency.

The ubiquity of metabolic specialization and mixotrophy, as suggested by these data, may contribute to difficulties in cultivating marine bacterioplankton. Accordingly, a member of the ARCTIC96BD-19 lineage was recently cultured from the surface ocean and found to oxidize thiosulfate (28), as was suggested by genome information obtained from SAGs in our previous study (15). Thus, single cell genomics provides a means for the discovery of genes that can be unequivocally assigned to uncultured taxonomic groups, thereby providing critical knowledge about their biology, including clues for cultivation strategies.

Biogeography of Marine Bacterioplankton. We analyzed the global distribution of surface ocean bacterioplankton using SAGs as references in fragment recruitment (4–6) of publicly available metagenomes, which span diverse geographic regions and climate zones and contain 45 million sequence reads totaling 23 Gbp (*SI Appendix*, Table S5 and Fig. S9). Using the 95% genomic DNA identity threshold, an operational delineation of taxonomically defined microbial species (29), the combined set of our 41 SAGs recruited an average of 0.9% reads from each surface ocean metagenome (Figs. 3 and 4*A*). The available PSP genomes (*Prochlorococcus*, *Synechococcus*, and *Pelagibacter*; a total of 24) recruited 1.6%, whereas the remaining 82 genomes of marine bacterioplankton cultures recruited only 0.3% (Fig. 4*A*). Lowering the DNA identity threshold in fragment recruitment resulted in a linear increase in the fraction of recruited reads until BLAST effectiveness diminished at nucleotide identities <60%. At this relaxed threshold, which corresponds to ~94% identity of the small subunit (SSU) rRNA gene (30) and an approximate, operational delineation of taxonomic order (31), 5.2%, 12.0%, 4.7%, and 19.3% of marine metagenome reads were recruited by SAGs, PSP genomes, 82 other bacterioplankton cultures, and a combined set of all genomes, respectively. Although the majority of marine bacterioplankton remains genomically unexplored, single cell sequencing offers a practical solution for genome recovery of uncultivated environmental microorganisms.

Using the 95% genomic DNA identity threshold, all SAGs obtained from the Gulf of Maine recruited the highest fraction of metagenomes from temperate regions (average temperature, 11.7 °C; range, 4.0–18.2 °C), which are represented by the northeast and northwest coasts of North America, the Atlantic coast of Europe, and the Indian Ocean off New Zealand in available datasets (Fig. 3). In contrast, SAGs obtained from the two subtropical gyres recruited primarily from warm-water metagenomes in the Atlantic, Pacific, and Indian Oceans (average temperature, 25.7 °C; range, 18.6–29.3 °C; designated "tropical"). SAGs recovered from the Mediterranean Sea, which has an intermediate climate, recruited relatively evenly across temperate and tropical metagenomes. Metagenomes from the Southern Ocean (average temperature, −0.1 °C; range, −2.0 to 4.2 °C; designated "polar") recruited primarily to SAGs from the Gulf of Maine, although significantly less compared with temperate metagenomes. In contrast to recruitment to SAGs, metagenome fragment recruitment to the majority of marine cultures was limited, and fewer clear biogeographic patterns were apparent (*SI Appendix*, Figs. S10 and S11), in agreement with previous observations (4, 32).

The abundance of specific genotypes, determined by metagenome fragment recruitment, was most strongly correlated with surface water temperature and latitude (*SI Appendix*, Fig. S12). Chlorophyll *a* concentration, water column depth, and longitude were minor factors in the ordination, suggesting that phytoplankton abundance, proximity to the coast, and geographic distance among sampling stations are less important than latitude in determining the abundance of most analyzed genotypes. These findings corroborate recent reports of temperature as a major driver of the global distribution of marine algae (33, 34) and *Pelagibacter* (35). Temperature and latitude also have been

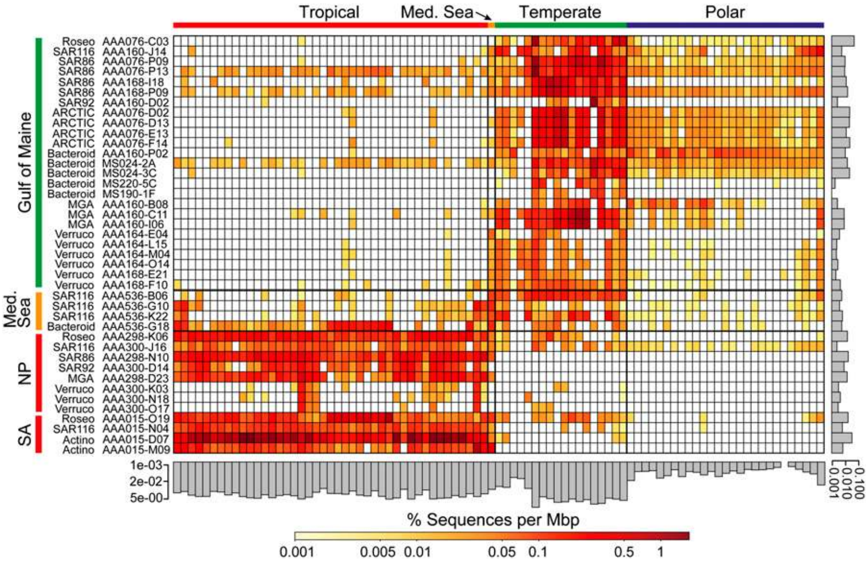


Fig. 3. Global distribution of SAG-related microorganisms, as determined by metagenomic fragment recruitment. SAGs are listed along the *y*-axis, where color bars indicate source locations. Color bars along the *x*-axis indicate the surface ocean climate zone (*SI Appendix*, Table S5 provides locations). Metagenomes are in the same order as presented in *SI Appendix*, Fig. S10 along the top *x*-axis. The scale bar indicates the percentage of aligned metagenome sequences with alignments ≥200 bp long and ≥95% identity, normalized by the length of each SAG assembly. Percentages of aligned sequences from each metagenome to all SAGs, and from all metagenomes to individual SAGs, are presented as gray bars on the *y*-axis and *x*-axis, respectively. Med. Sea, Mediterranean Sea; NP, North Pacific; SA, South Atlantic; Roseo, Roseobacter; ARCTIC, ARCTIC96-BD19 cluster; Bacteroid, Bacteroidetes; MGA, Marine Group A; Verruco, Verrucomicrobia; Actino, Actinobacteria. A threshold of ≥95% nucleotide sequence identity of alignments ≥200 bp was applied for the BLASTN-based recruitment.

identified as key determinants of less-specific descriptors of marine bacterioplankton biogeography, such as community richness (36) and the frequency of functionally related genes (37–39), for which our study provides extensive genomic context.

We estimated the ratio of metagenomic fragment recruitment from native versus nonnative climate zones, relative to SAG collection site, at various DNA identity intervals as proxies for evolutionary distance (Fig. 4*B*). In the case of temperate versus tropical zones, the ratio was highest (3,827) at 95–100% DNA identity, decreased to 154 at 90–95% identity, and declined to <10 at 80–85% identity. This pattern was similar for all taxonomic groups analyzed. The corresponding ratios were similar when comparing recruitment by temperate SAGs in temperate versus polar environments, but were higher when comparing recruitment by tropical SAGs in tropical versus polar environments. Thus, operationally defined species (>95% genomic DNA identity) were highly specific to their climate zones, but little geographic specificity was observed within phylogenetic groups that shared <80% genomic DNA identity, which corresponds to ∼97% identity of the SSU rRNA gene (31). Accordingly, several bacterioplankton cells analyzed in this study shared >97% identity of their SSU rRNA genes even though they originated from divergent climate zones and demonstrated contrasting geography in metagenome fragment recruitment; examples include SAR116 SAGs AAA158-M15 versus AAA015-N04 and SAR86 SAGs AAA298-N10 versus AAA076-P09 (*SI Appendix*, Fig. S1).

Whereas the SSU rRNA gene identities were high in these pairs of SAGs, the average nucleotide identity (29) was only 75% and 71%, respectively. The >97% identity of the SSU rRNA gene is the most widely used delineator of operational taxonomic units (OTUs) in microbial ecology. However, it is often overlooked that such OTUs encompass much broader phylogenetic groups than the currently accepted, operationally defined bacterial species, and may contain organisms with divergent adaptations. Thus, insufficient phylogenetic resolution might explain the difficulties encountered in earlier studies in detecting consistent differentiation of bacterioplankton along longitudinal gradients when using SSU rRNA gene surveys (35, 39, 40) or metagenome fragment recruitment with relaxed settings (32), although the more pronounced differences between polar and tropical bacterioplankton have been reported from such studies (35, 39, 40). Here, metagenome fragment recruitment using stringent settings and environmentally relevant, single cell genomes as references enabled us to identify previously undetected, community-wide genetic divergence among tropical, temperate, and polar marine bacterioplankton.

Assuming 1% divergence of the SSU rRNA gene every 50 Ma (41), we estimate that bacterioplankton genetic differences among the three climate zones might have accumulated over tens to hundreds of millions of years. Although such estimates contain significant uncertainties (42, 43), it is clear that the required evolutionary timeframe encompasses numerous overturns of the global ocean by surface currents and thermohaline circulation, which take 1,000–2,000 y each (44). These estimates corroborate the absence of longitudinal effects on fragment recruitment (Figs. 3 and 4) and suggest that the observed differences in bacterioplankton composition between nonpolar climate zones are not driven by dispersal limitations, but are defined by evolutionary innovation enabling certain genotypes to thrive in a specific climate zone. Given our lack of direct evidence for the genomic context of recruited metagenome fragments, how local populations of surface ocean bacterioplankton vary by their genome organization remains to be determined. Nevertheless, our data suggest that the global distribution of surface ocean bacterioplankton genes is not limited by dispersal at the time scales required for nucleotide substitution to exceed the current operational definition of bacteria species, thus adding some evolutionary constraints to the famous statement that "everything is everywhere, but the environment selects" (45).

Summary

Using large-scale single cell genomic sequencing and metagenome fragment recruitment, we have provided extensive, cultivation-independent insight into the genome-level diversity, metabolic potential, and biogeography of many abundant bacterial lineages

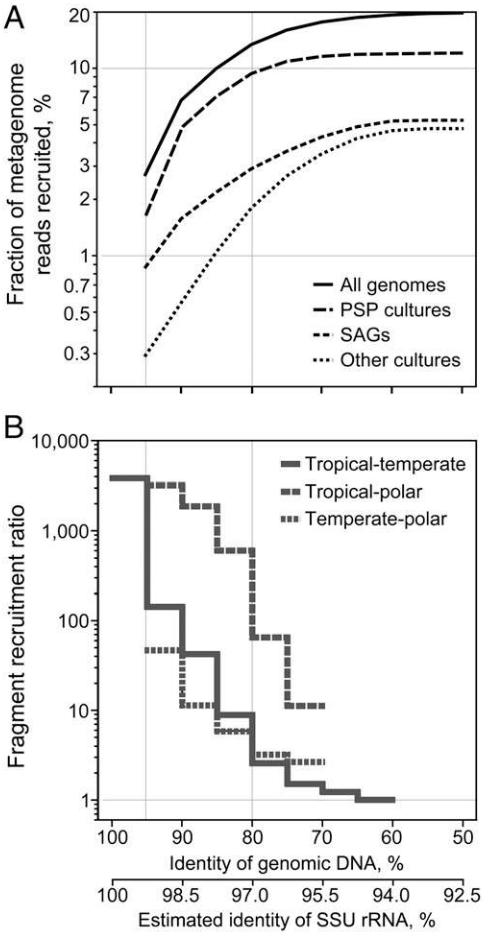


Fig. 4. Capacity of available genomes to represent surface ocean bacterioplankton assemblages, as related to genetic divergence and geographic differences. (*A*) Fraction of marine metagenome reads recruited by SAGs, genomes of bacterioplankton cultures, and the combined set of genomes using a range of genomic DNA identity thresholds. (*B*) Ratio of recruitment in the SAGs' native versus nonnative environment as a function of genomic DNA identity. Averages of values calculated for each metagenome (*A*) or genome (*B*) are provided. The scale of the SSU rRNA gene divergence was estimated using a Bacteria domain-wide correlation between SSU rRNA gene identity and the average nucleotide identity of available genomes (31). A threshold of ≥200-bp alignment was applied for the BLASTN-based recruitment.

inhabiting the surface ocean. Our data provide clear evidence that existing laboratory cultures consist mostly of copiotrophic genotypes, compared with free-living bacterioplankton that are streamlined for growth under resource-poor conditions. We also show that the global distribution of the majority of surface ocean bacterioplankton is correlated with temperature and latitude and is not likely limited by dispersal. Individual cells with highly similar SSU rRNA gene sequences exhibited significant genomic and biogeographic variability, highlighting challenges in the interpretation of individual gene surveys and metagenome assemblies in environmental microbiology. Our study demonstrates the utility of single cell genomics in providing a significantly improved understanding of the composition and dynamics of natural microbial assemblages in the ocean and other environments, which will be critical in predicting how ecosystems respond to large-scale environmental shifts, such as global warming and ocean acidification.

Materials and Methods

Collection and Construction of SAGs. Replicate, 1-mL aliquots of water collected for single cell analyses were cryopreserved with 6% glycine betaine (Sigma-Aldrich) and stored at −80 °C or in liquid nitrogen (46). Single cell sorting, whole-genome amplification, real-time PCR screens, and PCR product sequence analyses were performed at the Bigelow Laboratory Single Cell Genomics Center (www.bigelow.org/scgc), as described by Stepanauskas and Sieracki (13) for SAGs MS024-2A, MS024-3C, MS190-1F, and MS220-5C and by Swan et al. (15) and Martinez-Garcia et al. (14) for the remaining SAGs.

SSU rRNA gene sequences were edited using Sequencher v4.7 (Gene Codes) and compared with previously deposited sequences using the RDP v10 Classifier (SSU rRNA) and National Center for Biotechnology Information BLAST. SAG SSU rRNA sequences were aligned with selected database sequences using ClustalW. Alignment columns with > 90% gaps were removed, and a maximum likelihood tree (100 bootstrap replicates) was constructed using PhyML implemented in Geneious v6.0.5 (47). Details of SAG sequencing, assembly, and annotation are provided in the *SI Appendix*.

Genome Recovery Estimation of SAGs and Determination of Paralogs. To estimate the completeness of each assembled SAG genome, we analyzed all finished genome sequences of the taxonomic phyla Alphaproteobacteria (n = 145), Gammaproteobacteria (n = 317), Bacteroidetes (n = 22), and Actinobacteria (n = 131); the taxonomic phylum Verrucomicrobia (n = 4); and the taxonomic domain Bacteria (n = 1,023) available from the Integrated Microbial Genomes (IMG) database (48). Based on COG gene classifications, a set of conserved single copy genes (CSCGs) was extracted for each group of finished genomes from the IMG database. A CSCG was defined as a gene that occurs only once in each of 99% (95% in the case of the domain Bacteria) of the genomes contributing to the taxonomic group. The number of CSCGs for each group was as follows: Alphaproteobacteria, n = 58; Gammaproteobacteria, n = 47; Bacteroidetes, n = 86; Actinobacteria, n = 60; Verrucomicrobia, n = 330; Bacteria, n = 45. The ratio of the number of CSCGs observed for each SAG assembly and for the corresponding taxonomic group of finished genomes was used as a measure of genome recovery (*SI Appendix*, Table S2).

The frequency of paralog gene families within SAGs and marine cultures was determined using BLASTCLUST with the following settings: −L 0.5 −S 30.0 −e 1e-6. The number of paralogs out of the total number of protein coding genes was calculated for each genome.

Multivariate Analysis of SAG and Marine Culture Genome Signatures. The amino acid frequencies of 41 SAGs and bacterioplankton genomes were determined using Geneious v. 6.0.5, arcsin square root-transformed, and analyzed using principal components analysis (PCA) after standardization of values. Several genome characteristics found to separate marine prokaryotes by lifestyle (i.e., frequency of protein localizations and several COG categories) were calculated for SAGs and marine culture genomes as described previously (17), as was %GC and noncoding DNA, and these values were used as input for a second PCA analysis as described above. For this second PCA, the two Actinobacteria SAGs AAA015-D07 and AAA015-M09 were excluded. All PCAs were conducted using PRIMER v6.0.

Fragment Recruitment Analysis. The basic approach of Rusch et al. (4) was used to estimate the abundances of relatives of SAGs and bacterioplankton cultures within each metagenome. BLAST+ v2.2.25 was used to recruit metagenome sequences to each SAG assembly using default parameter values, except for the following: -evalue 0.0001 -reward 1 -penalty -1 -soft_masking true -lcase_masking -xdrop_gap 150. Genome contigs ≥2,000 kbp from each SAG were used in the fragment recruitment analysis. The 23S, 16S, 5S, and ITS regions were masked in each genome before recruitment. The percentage of unique recruits (≥200 bp long and matching at ≥95% identity) from each metagenome matching to each SAG was normalized by genome length. The percentage of unique reads for each metagenome–genome pair was also determined at 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, and 50% identity thresholds. SAG abundances from each metagenome were calculated from BLAST output and plotted using custom R scripts. Metagenomes used in fragment recruitment analysis were quality processed using PRINSEQ (49), and all sequences with the following characteristics were removed from further analysis: sequences <100 bp, sequences containing any ambiguities (Ns), all forms of replicate and duplicate sequences, and sequences with a minimum entropy value of 70 (applied to pyrosequencing datasets only).

Environmental and Sample Location Correlations with Fragment Recruitment Abundances. The influence of environmental factors on fragment recruitment-derived community composition was determined using nonmetric multidimensional scaling (MDS). MDS is an ordination technique that plots samples as points in low-dimensional space while attempting to maintain the relative distances between points as close as possible to the actual rank order of similarities between samples (50). Thus, metagenomes with similar community composition are plotted closer together in ordination space. A stress factor calculated for each MDS ordination indicates how well plotted configurations of sample distances agree with original rank orders calculated from the similarity matrices. SAG recruitment abundances were arcsin square root-transformed, and the Bray–Curtis distance was calculated for the MDS analysis. Sampling and environmental factors used for axis correlations were temperature, chlorophyll *a* concentration, water column depth at the sampling location (log-transformed), and latitude and longitude of the sampling location. All MDS calculations were performed using PC-ORD v6.08.

Calculation of Average Nucleotide Identity Between Genomes. Average nucleotide identity (ANI) values between the pairs of SAR116 SAGs AAA158-M15 and AAA015-N04 and SAR86 SAGs AAA298-N10 and AAA076-P09 were calculated following the method described by Goris et al. (29), using a custom Perl script. Each SAG served as a reference genome, and resulting ANI values were averaged.

ACKNOWLEDGMENTS. We thank David Emerson, William Hanage, Siv Anderson, and Jessica Labonte for valuable comments; Edward F. DeLong for metagenomes from Hawaii ocean time (HOT) station ALOHA; the officers and crew of the RV Ka'imikai-O-Kanaloa and the HOT team for sample collection at station ALOHA; and Jane Heywood and Michael Sieracki for South Atlantic field sample collection. This work was supported by National Science Foundation Grants EF-826924 (to R.S.), OCE-821374 (to R.S.), and OCE-1232982 (to R.S. and B.K.S.); US Department of Energy (DOE) JGI 2011 Microbes Program Grant CSP 387 (to R.S., B.K.S., S.G., M.A.M., F.M.L., R.C. and S.G.A.); the Gordon and Betty Moore Foundation (M.A.M.); Spanish Ministry of Science and Innovation Grant CGL2011-26848/BOS (to S.G.A) and CONSOLIDER-INGENIO2010 Program Grant CSD2008-00077 (to S.G.A. and J.M.G.); the Natural Sciences and Engineering Research Council of Canada (NSERC); the Canada Foundation for Innovation, and the Canadian Institute for Advanced Research (CIFAR; S.J.H.). J.J.W. was supported by NSERC. Research activities of R.C. and F.M.L. are supported by the Australian Research Council, and research activities of R.C. are supported by the Australian Antarctic Science program. Work conducted by the DOE Joint Genome Institute is supported by the DOE's Office of Science under Contract DE-AC02-05CH11231. This is contribution no. 006 of the Tara Oceans Expedition 2009–2012.

1. Glockner FO, et al. (2012) *Marine Microbial Diversity and Its Role in Ecosystem Functioning and Environmental Change*. Marine Board Position Paper 17. eds Calewaert JB, McDonough N (Marine Board, European Science Foundation, Ostend, Belgium).
2. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345(6270):60–63.
3. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59(1): 143–169.
4. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e77.
5. Woyke T, et al. (2009) Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4(4):e5299.
6. Dupont CL, et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6(6):1186–1199.
7. Tripp HJ, et al. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* 452(7188):741–744.
8. Carini P, Steindler L, Beszteri S, Giovannoni SJ (2013) Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. *ISME J* 7(3):592–602.
9. Halsey KH, Carter AE, Giovannoni SJ (2012) Synergistic metabolism of a broad range of C1 compounds in the marine methylotrophic bacterium HTCC2181. *Environ Microbiol* 14(3):630–640.
10. Temperton B, Giovannoni SJ (2012) Metagenomics: Microbial diversity through a scratched lens. *Curr Opin Microbiol* 15(5):605–612.
11. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394.
12. Stepanauskas R (2012) Single cell genomics: An individual look at microbes. *Curr Opin Microbiol* 15(5):613–620.
13. Stepanauskas R, Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* 104(21):9052–9057.
14. Martinez-Garcia M, et al. (2012) Capturing single cell genomes of active polysaccharide degraders: An unexpected contribution of *Verrucomicrobia*. *PLoS ONE* 7(4):e35314.
15. Swan BK, et al. (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* 333(6047):1296–1300.
16. Giovannoni SJ, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309(5738):1242–1245.
17. Lauro FM, et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106(37):15527–15533.
18. Cottrell MT, Kirchman DL (2000) Natural assemblages of marine proteobacteria and members of the *Cytophaga-Flavobacter* cluster consuming low- and high-molecular-weight dissolved organic matter. *Appl Environ Microbiol* 66(4):1692–1697.
19. Luo H (2012) Predicted protein subcellular localization bacterioplankton in dominant surface ocean. *Appl Environ Microbiol* 78(18):6550–6557.
20. Giovannoni S, Stingl U (2007) The importance of culturing bacterioplankton in the "omics" age. *Nat Rev Microbiol* 5(10):820–826.
21. Schut F, Prins RA, Gottschal JC (1997) Oligotrophy and pelagic marine bacteria: Facts and fiction. *Aquat Microb Ecol* 12:177–202.
22. Coleman ML, Chisholm SW (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* 15(9):398–407.
23. Rodrigue S, et al. (2009) Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* 4(9):e6864.
24. Raghunathan A, et al. (2005) Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* 71(6):3342–3347.
25. Marcy Y, et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* 104(29):11889–11894.
26. Raghavan R, Kelkar YD, Ochman H (2012) A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci USA* 109(36):14504–14507.
27. Grzymski JJ, Dussaq AM (2012) The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J* 6(1):71–80.
28. Marshall KT, Morris RM (2013) Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. *ISME J* 7(2):452–455.
29. Goris J, et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57(Pt 1):81–91.
30. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 102(7):2567–2572.
31. Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187(18):6258–6264.
32. Yooseph S, et al. (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 468(7320):60–66.
33. Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ (2010) Patterns of diversity in marine phytoplankton. *Science* 327(5972):1509–1511.
34. Thomas MK, Kremer CT, Klausmeier CA, Litchman E (2012) A global pattern of thermal adaptation in marine phytoplankton. *Science* 338(6110):1085–1088.
35. Brown MV, et al. (2012) Global biogeography of SAR11 marine bacteria. *Mol Syst Biol* 8:595.
36. Fuhrman JA, et al. (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* 105(22):7774–7778.
37. Gianoulis TA, et al. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* 106(5):1374–1379.
38. Jiang X, et al. (2012) Functional biogeography of ocean microbes revealed through non-negative matrix factorization. *PLoS ONE* 7(9):e43866.
39. Ghiglione JF, et al. (2012) Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc Natl Acad Sci USA* 109(43):17633–17638.
40. Pommier T, Pinhassi J, Hagstrom A (2005) Biogeographic analysis of ribosomal RNA clusters from marine bacterioplankton. *Aquat Microb Ecol* 41(1):79–89.
41. Ochman H, Wilson AC (1987) Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J Mol Evol* 26(1-2):74–86.
42. Kuo CH, Ochman H (2009) Inferring clocks when lacking rocks: The variable rates of molecular evolution in bacteria. *Biol Direct* 4:35.
43. Ho SYW, et al. (2011) Time-dependent rates of molecular evolution. *Mol Ecol* 20(15): 3087–3101.
44. Doos K, Nilsson J, Nycander J, Brodeau L, Ballarotta M (2012) The world ocean thermohaline circulation. *J Phys Oceanogr* 42:1445–1460.
45. Baas Becking LGM (1934) *Geobiologie of Inleiding tot de Milieukunde* (W.P. Van Stockum & Zoon, The Hague, The Netherlands).
46. Cleland D, Krader P, McCree C, Tang J, Emerson D (2004) Glycine betaine as a cryoprotectant for prokaryotes. *J Microbiol Methods* 58(1):31–38.
47. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ (2007) Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* 2:27.
48. Markowitz VM, et al. (2010) The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res* 38(Suppl. 1):D382–D390.
49. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6):863–864.
50. Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Aust J Ecol* 18:117–143.

Supplementary Appendix:

Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean

Brandon K. Swan, Ben Tupper, Alexander Sczyrba, Federico M. Lauro, Manuel Martinez-Garcia, José M. González, Haiwei Luo, Jody J. Wright, Zachary C. Landry, Niels W. Hanson, Brian P. Thompson, Nicole J. Poulton, Patrick Schwientek, Silvia G. Acinas, Stephen J. Giovannoni, Mary Ann Moran, Steven J. Hallam, Ricardo Cavicchioli, Tanja Woyke and Ramunas Stepanauskas

Supplementary Results and Discussion: Lineage-specific features of SAGs

The Marine Group A (MGA; also known as SAR406) is a phylum-level lineage, members of which are abundant throughout the ocean (1, 2). No MGA cultures are available, and their biology remains obscure. Here we sequenced five SAGs from the Gulf of Maine and the North Pacific Subtropical Gyre, all of which are affiliated with the subgroup ZA3312c (Fig. S3). General features of MGA SAGs, such as % GC, % non-coding DNA and frequencies of COG categories are similar to many of the Proteobacteria lineages and indicate genome streamlining and adaptations to oligotrophy (Fig. 2A). The presence of proteorhodopsin (Fig. S4) and polysulfide reductase genes in all five SAGs suggest that MGA supplement their heterotrophic energy sources by non-photosynthetic light harvesting and the oxidation of sulfur compounds.

Planktonic Verrucomicrobia are also widespread in surface ocean, constituting ~2% of heterotrophic bacterioplankton, yet lack cultured representatives (3). Recently, we found that certain Verrucomicrobia lineages specialize in the hydrolysis of polysaccharides (4). Here we report partial genomes of eight Gulf of Maine SAGs of class Verrucomicrobiae, four Gulf of Maine SAGs of Subdivision 3 and three SAGs of class Opitutae from the North Pacific subtropical gyre. All analyzed SAGs have elevated frequencies of genes encoding cell surface and extracellular proteins (Fig. 2A), and glycoside hydrolases (Fig. S5, Table S8), suggesting that specialization in polysaccharide degradation is a common feature among marine Verrucomicrobia. All SAGs from both the open ocean and coastal areas possessed a vast repertoire of glycoside hydrolases (Fig. S5), which would provide the metabolic machinery for the utilization of diverse and complex biopolymers (Table S7). Principal Component Analysis (PCA) of glycoside hydrolases from SAGs showed several Verrucomicrobia and Bacteroidetes from different geographical regions shared a similar set of these genes, suggesting that they might utilize similar polysaccharide substrates while others, such as AAA168-F10, may be more specialized (Fig. S6). Several Verrucomicrobia SAGs from the Gulf of Maine were found to contain phage-like DNA, indicating either infections, phage attachment on cell surface, or active uptake of phage DNA by the cell. The three Opitutae SAGs exhibited a unique biogeographic pattern, by recruiting metagenome reads almost exclusively from the centers of the two analyzed subtropical gyres (Figs. 1 and 3).

Two SAGs from marine Actinobacteria were sequenced, designated AAA015-M09 and AAA015-D07. Both SAGs had relatively low GC content (~32%). Both SAGs are closely related to the SAR432 group of marine Actinobacteria by SSU rRNA gene phylogeny (5). AAA015-D07 and AAA015-M09 were determined to be approximately 99% identical to each other and 98% identical to the original SAR432 clone, based on SSU rRNA gene comparisons. Gene annotations of coding sequences indicate an aerobic heterotrophic lifestyle. Both SAGs appear to have genes comprising large portions of the pentose phosphate pathway, suggesting that these organisms may be able to use sugars as a carbon or energy source. Additionally, a number of genes for glycolysis/gluconeogenesis are encoded. Both genomes also encode sequences for multiple cytochrome P450 proteins (6, 7), aromatic ring hydroxylases (8-14), and nitroreductases (12, 13, 15), suggesting that these organisms could play a role in the breakdown of recalcitrant dissolved organic matter in the ocean. Other metabolic genes of interest common to both genomes include short-chain alcohol dehydrogenases of unknown specificity and formate hydrogen lyases, indicating that these organisms may be able to utilize C1 compounds as a source of energy (16). Both genomes include annotated genes for low (*caa3*) and high (*cbb3*) affinity cytochrome C oxidases, suggesting adaptation to growth in a wide range of oxygen concentration (17, 18). These bacteria may periodically inhabit an environment with a reduced oxygen content, for instance, within a marine snow aggregate (18, 19). AAA015-M09 encodes candidate genes for a full TCA cycle and the AAA015-D07 contigs appear to contain most of the genes for a TCA cycle as well. Both SAGs also have predicted coding sequences for isocitrate lyase and malate synthase, confirming the presence of a glyoxalate bypass. These organisms appear to rely heavily on ABC transporters for transport, with AAA015-D07 containing coding sequences for 22 ABC transporter monomeric proteins and AAA015-M09 containing 32 sequences. Seven other transporters of varying other types were found in AAA015-D07 and nine in AAA015-M09. AAA015-D07 has transporters predicted to play a role in copper or nickel acquisition. A cobalt transporter is present in M09, suggesting that vitamin B12 may be a required cofactor for these organisms. This is supported by the inclusion of other genes related to cobalamin synthesis and modification, for instance, cobalamin adenosyltransferase and adenosyl cobanimide kinase, both found in AAA015-M09. Additionally, AAA015-D07 contains an adenosylcobalamin-dependent ribonucleoside-diphosphate reductase, suggesting that it also likely requires vitamin B12 as a growth factor. The SAGs also carry 17 and 16 glycosyltransferase genes in AAA015-D07 and AAA015-M09, respectively. Most likely, these enzymes play a role in cell wall biosynthesis, however alternative roles should be considered.

Expansion of glycosyltransferase paralogs in these genomes suggests an important role for them in the evolution of the SAR432 lineage.

The Alphaproteobacteria lineage SAR116 is ubiquitous in the surface ocean, and two genomes of cultures are publicly available (20, 21). In difference to these cultures, most of the 9 sequenced SAGs contain form I carbon monoxide (CO) dehydrogenase (Fig. S7), with the characteristic AYXCSFR motif, which has been suggested to be the only genuine CO dehydrogenase (22). One SAG from the relatively productive Gulf of Maine encodes a green-tuned rhodopsin, while three SAGs from the ultraoligotrophic South Atlantic subtropical gyre and the Mediterranean Sea encode blue-tuned rhodopsins (Fig. S4), which is in agreement with previously proposed rhodopsin adaptations to *in situ* light conditions (23). In contrast to the cultures, SAGs AAA015-N04 and AAA536-K22 encode *sox* operons, with similar organization to *sox* in two of the Roseobacter SAGs, indicating their capacity for S oxidation (Fig. S7). In AAA015-N04, this operon is adjacent to the *cox* operon and is in the vicinity of rhodopsin, ATPase and cytochrome c genes, indicating their metabolic importance and potential co-regulation (24). Thus, our data suggest a variety of previously reported and novel mixotrophy strategies within the SAR116 cluster, which resemble those found in the *Roseobacter* sister-cluster (25, 26).

The Roseobacter cluster within the Alphaproteobacteria is an abundant and among the best-studied lineages of marine bacterioplankton, with ~40 genomes currently available from cultures (25, 26). Yet, certain subclusters have resisted cultivation, and the cultivated subset of Roseobacter may be metabolically biased, as compared to the predominant relatives in the environment (26). Among the five Roseobacter SAGs, one (AAA076-C03) is closely related to the cultured strain HTCC2255 and represents a basal group in the lineage, whereas the other four (AAA015-O19, AAA076-E06, AAA298-K06, AAA300-J04) constitute a monophyletic clade in which no cultured representatives are found (Fig. S8). In agreement with cultured roseobacters, SAGs contained genes for DMSP and aromatic compound degradation, carbon monoxide oxidation, C1 utilization, C2 processing through the ethylmalonyl CoA pathway, and photoheterotrophy based on proteorhodopsin (AAA076-C03 only) in addition to the known aerobic anoxygenic phototrophy capability (AAA298-K06 only). The SAGs also contained 615 novel to Roseobacter ORFs, 70% of which encoded hypothetical or unnamed proteins, and the remaining novel ORF encoded toxin resistance, secondary metabolite biosynthesis, phage-related functions, and uncharacterized oxidases and hydrolases (the apparently phage-dominated

SAG AAA076-E06 was excluded from this analysis). The SAGs have 37-40% GC content, which is significantly lower than 49-70% found in cultures (except for 37% GC in HTCC2255). The SAGs have smaller genomes (estimated at 2.8-3.8 Mbp; Table S2) than cultured Roseobacter (median = 4.4 Mbp). Other genome features (Fig. 2) also differ between Roseobacter cultures and SAGs, in support of the recent suggestion (26) that the predominant Roseobacter in the environment have more streamlined genomes and are better adapted to oligotrophic conditions than the available cultures.

Genome analyses of several SAGs from the Gammaproteobacteria lineages SAR86, ARCTIC96BD-19 and SAR92 revealed significant metabolic flexibility, with each group possessing the genetic potential to utilize different pathways. Proteorhodopsin was identified in 11 out of the 13 SAGs analyzed (Table S10). Only the SAR92 SAGs contain a majority of genes required for the biosynthesis of retinal, which is required for proteorhodopsin functionality (27). Although the other Gammaproteobacteria SAGs were found to only possess geranylgeranyl diphosphate synthase (*crtE*), all proteorhodopsin genes within SAR86 have a dehydrogenase upstream, an arrangement noted in fosmids (28) and other SAR86 single cell genomes (29). It has been suggested that this dehydrogenase could convert retinal or ß-carotene to retinol, but this would require the pigment to be taken up from the environment (29). Two ARCTIC96BD-19 SAGs (AAA076-D13 and AAA076-F14) contain two copies of proteorhodopsin that are divergent and form separate phylogenetic clusters (Fig. S3). This is not surprising, as it is well documented that proteorhodopsin undergoes duplication and lateral transfer quite frequently (27). All SAG proteorhodopsin sequences are of the spectrally green tuned variant (30). Genes encoding near-complete Embden-Meyerhof-Parnas, pentose phosphate, and modified Entner-Doudoroff central metabolic pathways were detected within these Gammaproteobacteria SAGs, with some notable differences between groups (Table S9). The oxidative component of pentose phosphate pathway is not well represented within SAR86 but is complete in the majority of ARCTIC96BD-19 SAGs. Also, several SAR86 and SAR92 SAGs contain genes for the Entner-Doudoroff pathway, but the key genes for this pathway were not detected in ARCTIC96BD-19 SAGs. All SAGs contain a near-complete tricarboxylic acid cycle (TCA) with the exception that all SAR86 SAGs are lacking citrate synthase. Dupont et al. (29) also found this key gene missing from their SAR86 single cell genomes and suggested this group may use a combination of the TCA and methylTCA cycles, with the latter utilizing several methylcitrate enzymes. 2-Methylcitrate synthase, methylcitrate lyase, and methylcitrate dehydrogenase were detected in all SAR86 and some ARCTIC96BD-19 SAGs. Only the ARCTIC96BD-19 SAGs contain genes for inorganic carbon fixation, which was previously reported by Swan et al. (31) for genomes of this group from the mesopelagic. Only ARCTIC96BD-19 SAGs contain genes for sulfur oxidation (adenylylsulfate reductase, *aprA*; sulfite reductase), with no other reductases being found in any of the SAGs. Aside from the potential for chemoautotrophic growth by ARCTIC96BD-19, the potential for a heterotrophic metabolism appears to be dominant among these lineages.

Members of the Bacteroidetes phylum comprise 10-20% of the total marine heterotrophic bacterioplankton (32, 33). Metagenome fragment recruitment indicated that relatives of SAG AAA536-G18 are widely distributed in temperate and tropical waters, whereas a more restricted distribution to the temperate zone was found for the relatives of SAGs MS220-5C and MS190-1F (Fig. 3). Based on the 98% SSU rRNA gene identity to the culture *Polaribacter* sp. MED152, SAG AAA160-P02 may be considered a member of this genus, which appears to be abundant in both temperate and polar waters (34-37). Considering the high estimated genome completeness (84.9%; Table S2), and that it is a better recruiter of metagenome sequences than its cultured relatives, AAA160-P02 provides important information on this numerically important group. *Polaribacter* genomes, including AAA160-P02, encode green light-tuned proteorhodopsins, as indicated by Met105 (*38*). Proteorhodopsin was not detected in SAG AAA536-G18, but *blh* (β-carotene 15,15'-monooxygenase) and other genes needed to synthesize retinal are present. Similarly to cultured Bacteroidetes (37, 39, 40), the gene content of Bacteroidetes SAGs suggests specialization for growth on particles and high molecular weight compounds, including peptides and polysaccharides. Accordingly, we detected genes involved in gliding motility, exopolysaccharide biosynthesis and adhesion. The majority of genes required for glycolysis, gluconeogenesis and the TCA cycle were also detected. Furthermore, we found genes encoding PEP carboxylase in AAA160-P02 and AAA536-G18, malic enzyme in AAA160-P02, and pyruvate carboxylase in MS024-2A, indicating the potential for anaplerotic metabolism.

Materials and Methods

SAG genomic sequencing, assembly and annotation

With the exception of SAGs MS024-2A, MS024-3C, MS190-1F, and MS220-5C, draft genomes were generated at the DOE Joint genome Institute (JGI) using the Illumina technology (41). Illumina standard shotgun libraries were constructed and sequenced using the Illumina HiSeq 2000 platform.

All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. All raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts. The following steps were then performed for assembly: 1) filtered Illumina reads were assembled using Velvet v. 1.1.04 (42), 2) 1–3 kbp simulated paired end reads were created from Velvet contigs using wgsim (http://github.com/lh3/wgsim), 3) Illumina reads were assembled with simulated read pairs using Allpaths–LG v. r41043 (43). Parameters for assembly steps were: 1) Velvet: 63 -shortPaired and velvetg: -very clean yes -export -Filtered yes -min contig lgth 500 -scaffolding no -cov cutoff 10, 2) wgsim: -e 0 -1 100 -2 100 -r 0 -R 0 -X 0, 3) Allpaths: -LG PrepareAllpathsInputs: PHRED 64=1 PLOIDY=1 FRAG COVERAGE=125 JUMP COVERAGE=25 LONG JUMP COV=50, RunAllpathsLG: THREADS=8 RUN=std shredpairs TARGETS=standard VAPI WARN ONLY=True OVERWRITE=True.

The draft genomes of Flavobacteria sp. MS190-1F and MS220-5C were generated at the DOE Joint genome Institute (JGI) using a combination of Illumina (44) and 454 technologies (45). For the MS190-1F genome, we constructed and sequenced an Illumina GAii shotgun library which generated 13,362,482 reads totaling 481 Mbp, a 454 Titanium standard library which generated 446,098 reads and 2 paired end 454 libraries with an average insert size of 5 kbp which generated 753,634 reads totaling 145.8 Mbp of 454 data. For MS220-5C genome, we constructed and sequenced an Illumina GAii shotgun library which generated 11,376,334 reads totaling 409.5 Mbp, a 454 Titanium standard library which generated 559,605 reads and 2 paired end 454 libraries with an average insert size of which generated 530,819 reads totaling 127.6 Mbp of 454 data. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. The initial draft assembly of MS190-1F and MS220-5C contained 1317 contigs in 33 scaffolds, and 1066 contigs in 149 scaffolds, respectively. The 454 Titanium standard data and the 454 paired end data were assembled together with Newbler, version 2.3-PreRelease-6/30/2009. The Newbler consensus sequences were computationally shredded into 2 kbp overlapping fake reads (shreds). Illumina sequencing data was assembled with VELVET, version 1.0.13 (46), and the consensus sequence were computationally shredded into 1.5 kbp overlapping fake reads (shreds). We integrated the 454 Newbler consensus shreds, the Illumina VELVET consensus shreds and the read pairs in the 454 paired end library using parallel phrap, version SPS - 3.65 (High Performance Software, LLC). The software Consed (47, 48) was used in the following finishing process. Illumina data was used to correct potential base errors and increase consensus quality using the software Polisher developed at JGI. Possible mis-assemblies were corrected using gapResolution, Dupfinisher (49), or sequencing cloned bridging PCR fragments with subcloning. Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR primer walks. The estimated genome size of MS190-1F is 2.4 Mbp and the final assembly is based on 63.6 Mbp of 454 draft data which provides an average 26.5x coverage of the genome and 480.3 Mbp of Illumina draft data which provides an average 200.1x coverage of the genome. The estimated genome size of MS220-5C is 1.6 Mbp and the final assembly is based on 33.9 Mbp of 454 draft data which provides an average 21.2x coverage of the genome and 408.2 Mbp of Illumina draft data which provides an average 255.1x coverage of the genome. Sequencing and assembly details of draft genomes of *Flavobacteria* sp. MS024-2A and MS024-3C are published in Woyke et al. (50).

Genes were identified using Prodigal (51). The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database (nr), UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScan-SE tool (52) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (53). Other non–coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching genomes for the corresponding Rfam profiles using INFERNAL (54). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) (55) platform developed by the Joint Genome Institute, Walnut Creek, CA, USA (http://img.jgi.doe.gov).

SAG whole genome sequence quality control

Each raw sequence data set was screened against all finished bacterial and archaeal genome sequences (downloaded from NCBI) and the human genome to identify potential contamination in the sample. Reads were mapped against reference genomes with bwa version 0.5.9 (56) using default parameters (96% identity threshold). None of the libraries showed significant contamination. Additionally, gene sequences of the final assemblies (see below) were compared against the GenBank nr database by BLASTX and taxonomically classified using MEGAN (57).

To further verify the absence of contaminating sequences in the assemblies, tetramer frequencies were extracted from all scaffolds using two alternative settings: 1) sliding window of 1000 bp and 100 bp step size and 2) sliding window of 5000 bp and 500 bp step size. Reverse-complementary tetramers were combined and the frequencies represented as a N×136 feature matrix, where N is the number of windows and each column of the matrix corresponds to the frequency of one of the 136 possible tetramers. Principal component analysis (PCA) was then used to extract the most important components of this high dimensional feature matrix. The analysis produced unimodal distribution along the first four PCs for the majority of SAGs, suggesting homogenous DNA sources. Scaffolds representing extremes on the first four PCs were identified and manually examined for their closest TBLASTX hits against NCBI nt database.

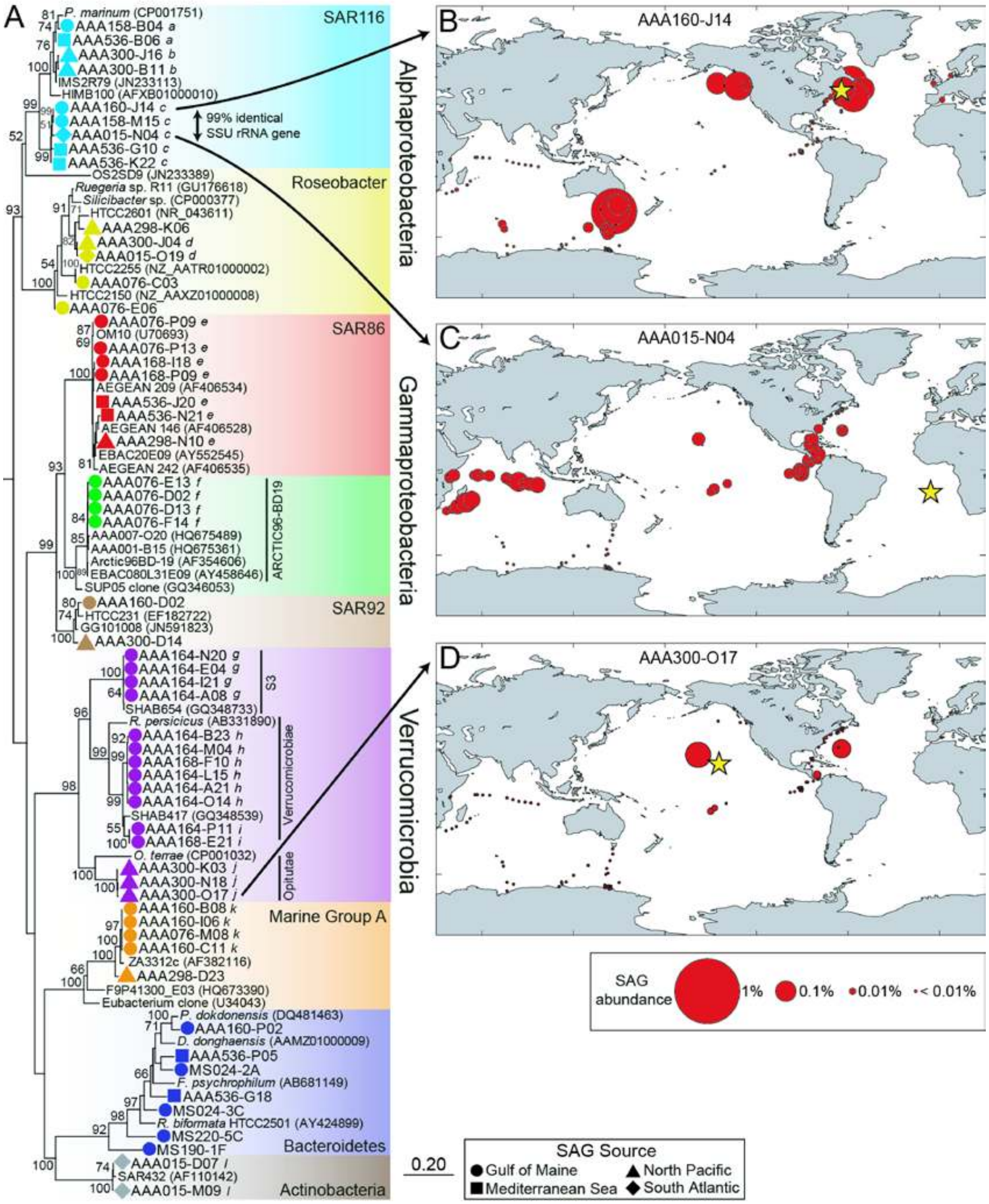


Figure S1. Phylogenetic and geographic distribution of single amplified genomes (SAGs). Phylogenetic tree of SSU rRNA gene sequences from single amplified genomes (SAGs; color symbol) and closely related cultures and environmental clones (*A*); and geographic distribution of selected SAGs, as inferred from metagenomic fragment recruitment (*B-D*). The phylogenetic tree was inferred using maximum likelihood in PhyML, with bootstrap values ≥50% indicated at nodes. *Thermococcus peptonophilus* was used as the outgroup. Lower-case letters to the right of the genome's name indicate SAGs with SSU rRNA identities ≥97%. A threshold of ≥95% nucleotide sequence identity of alignments ≥200 bp was applied in BLASTN-based fragment recruitment. The estimated SAG genotype abundance indicates the fraction of aligned metagenome sequences, normalized by SAGs' estimated genome size. The SSU rRNA sequences of SAR116 SAGs AAA160-J14 and AAA015-N04 are ≥99% identical. The stars in panels B-D indicate the SAG sampling locations.

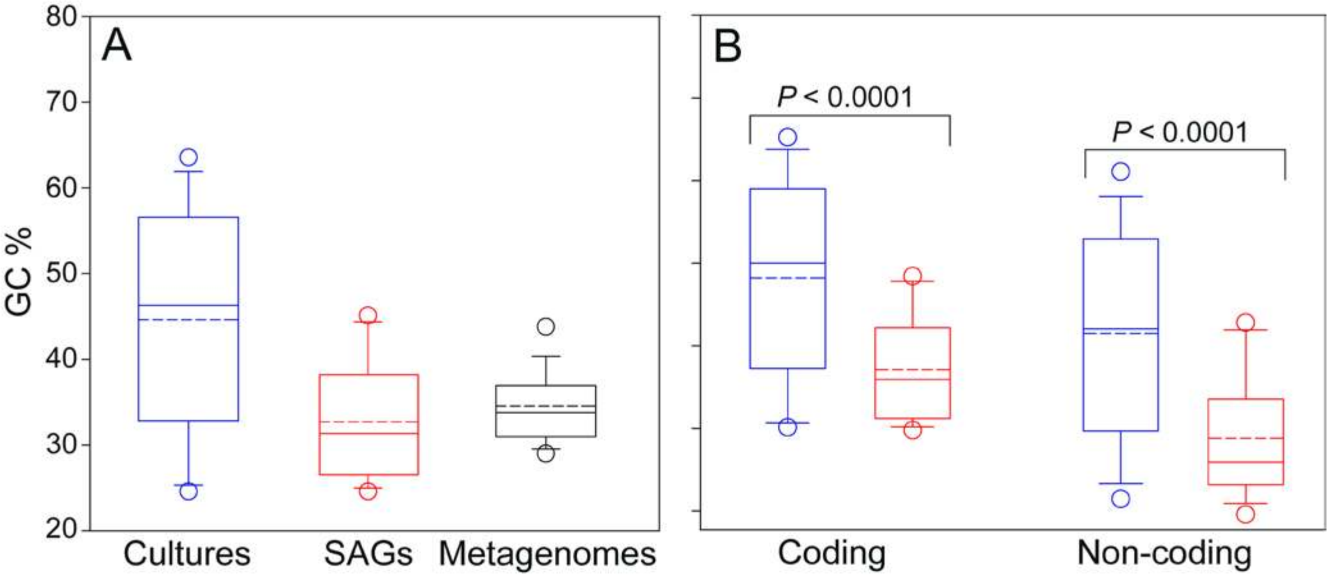


Figure S2. GC content comparisons. GC content differences between single amplified genomes (SAGs), cultured bacterioplankton, and metagenome sequences (*A*). GC content of coding and non-coding genome regions of cultures (n=101; blue box plots) and SAGs (n=41; red box plots) (*B*). Box plots show median (solid line), mean (dashed line), 75th and 25th percentiles (top and bottom of box, respectively), 90th and 10th percentiles (top and bottom bar), and 95th and 5th percentiles (open circles) of GC content. T-tests were used to determine statistical significance of differences.

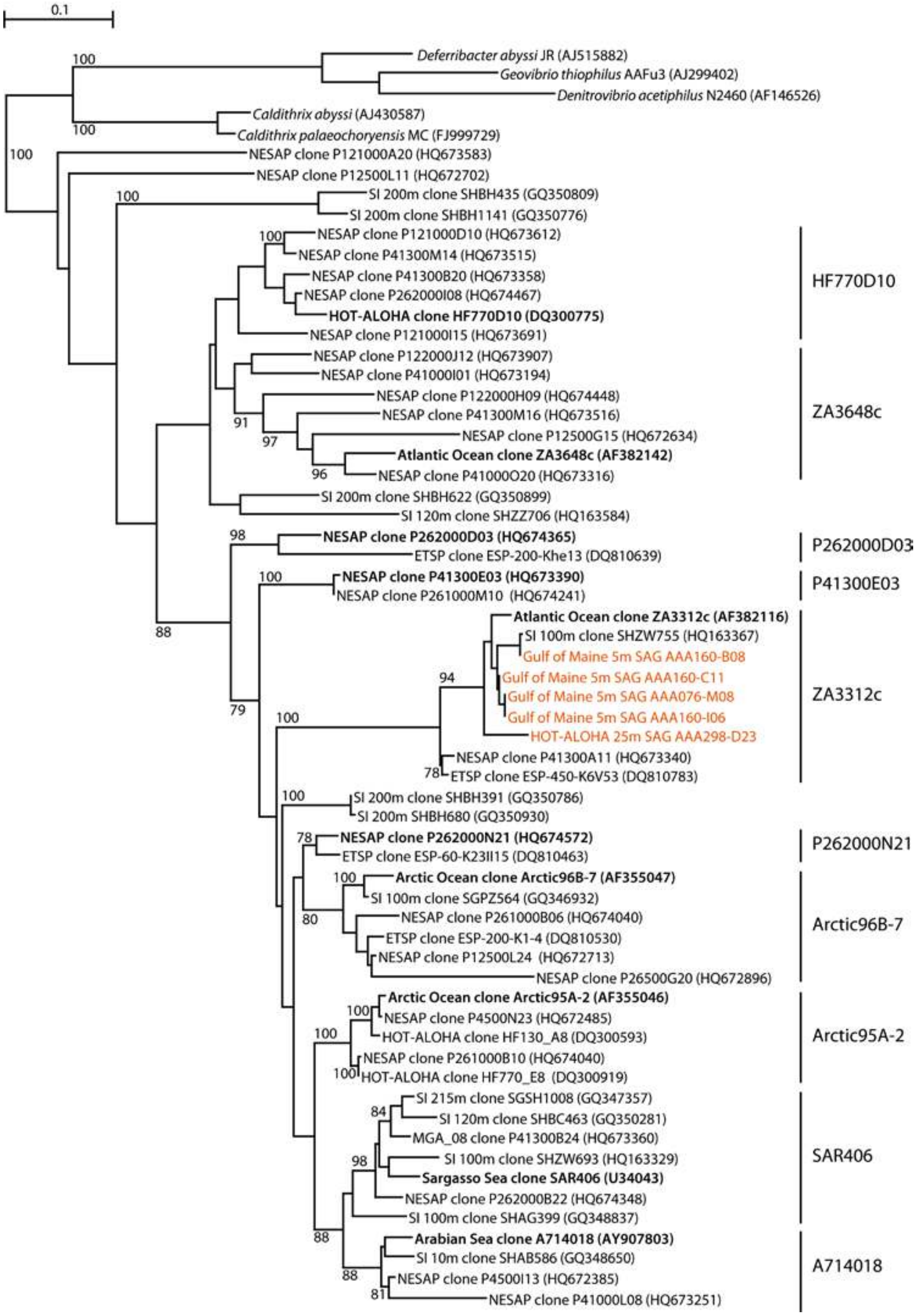

0.1
100
100
100
Deferribacter abyssi JR (AJ515882)
Geovibrio thiophilus AAFu3 (AJ299402)
Denitrovibrio acetiphilus N2460 (AF146526)
Caldithrix abyssi (AJ430587)
Caldithrix palaeochoryensis MC (FJ999729)
NESAP clone P121000A20 (HQ673583)
NESAP clone P12500L11 (HQ672702)
100
SI 200m clone SHBH435 (GQ350809)
SI 200m clone SHBH1141 (GQ350776)
100
NESAP clone P121000D10 (HQ673612)
NESAP clone P41300M14 (HQ673515)
NESAP clone P41300B20 (HQ673358)
NESAP clone P262000I08 (HQ674467)
HOT-ALOHA clone HF770D10 (DQ300775)
NESAP clone P121000I15 (HQ673691)
HF770D10
NESAP clone P122000J12 (HQ673907)
NESAP clone P41000I01 (HQ673194)
NESAP clone P122000H09 (HQ674448)
NESAP clone P41300M16 (HQ673516)
91
97
NESAP clone P12500G15 (HQ672634)
Atlantic Ocean clone ZA3648c (AF382142)
96
NESAP clone P41000O20 (HQ673316)
ZA3648c
SI 200m clone SHBH622 (GQ350899)
SI 120m clone SHZZ706 (HQ163584)
98
NESAP clone P262000D03 (HQ674365)
ETSP clone ESP-200-Khe13 (DQ810639)
P262000D03
100
NESAP clone P41300E03 (HQ673390)
NESAP clone P261000M10 (HQ674241)
P41300E03
88
Atlantic Ocean clone ZA3312c (AF382116)
SI 100m clone SHZW755 (HQ163367)
Gulf of Maine 5m SAG AAA160-B08
94
Gulf of Maine 5m SAG AAA160-C11
Gulf of Maine 5m SAG AAA076-M08
Gulf of Maine 5m SAG AAA160-I06
HOT-ALOHA 25m SAG AAA298-D23
ZA3312c
79
100
NESAP clone P41300A11 (HQ673340)
78
ETSP clone ESP-450-K6V53 (DQ810783)
100
SI 200m clone SHBH391 (GQ350786)
SI 200m SHBH680 (GQ350930)
78
NESAP clone P262000N21 (HQ674572)
ETSP clone ESP-60-K23II15 (DQ810463)
P262000N21
100
Arctic Ocean clone Arctic96B-7 (AF355047)
SI 100m clone SGPZ564 (GQ346932)
80
NESAP clone P261000B06 (HQ674040)
ETSP clone ESP-200-K1-4 (DQ810530)
NESAP clone P12500L24 (HQ672713)
NESAP clone P26500G20 (HQ672896)
Arctic96B-7
Arctic Ocean clone Arctic95A-2 (AF355046)
100
NESAP clone P4500N23 (HQ672485)
100
HOT-ALOHA clone HF130_A8 (DQ300593)
NESAP clone P261000B10 (HQ674040)
100
HOT-ALOHA clone HF770_E8 (DQ300919)
Arctic95A-2
SI 215m clone SGSH1008 (GQ347357)
84
SI 120m clone SHBC463 (GQ350281)
MGA_08 clone P41300B24 (HQ673360)
98
SI 100m clone SHZW693 (HQ163329)
Sargasso Sea clone SAR406 (U34043)
NESAP clone P262000B22 (HQ674348)
SI 100m clone SHAG399 (GQ348837)
SAR406
88
Arabian Sea clone A714018 (AY907803)
SI 10m clone SHAB586 (GQ348650)
88
NESAP clone P4500I13 (HQ672385)
81
NESAP clone P41000L08 (HQ673251)
A714018

Figure S3. Phylogenetic analysis of MGA. Unrooted phylogenetic tree based on SSU rRNA gene sequences derived from clone libraries and SAGs, showing the phylogenetic affiliation of MGA SAGs (orange) identified in this study. The tree was inferred using maximum likelihood implemented in PhyML using an HKG + 4G + I model of nucleotide evolution where the parameter of the G distribution, the proportion of invariable sites, and the transition/transversion ratio were estimated for each dataset. The confidence of each node was determined by assembling a consensus tree of 100 bootstrap replicates. Bootstrap values below 60% are not shown. The bar represents 1% estimated sequence divergence.

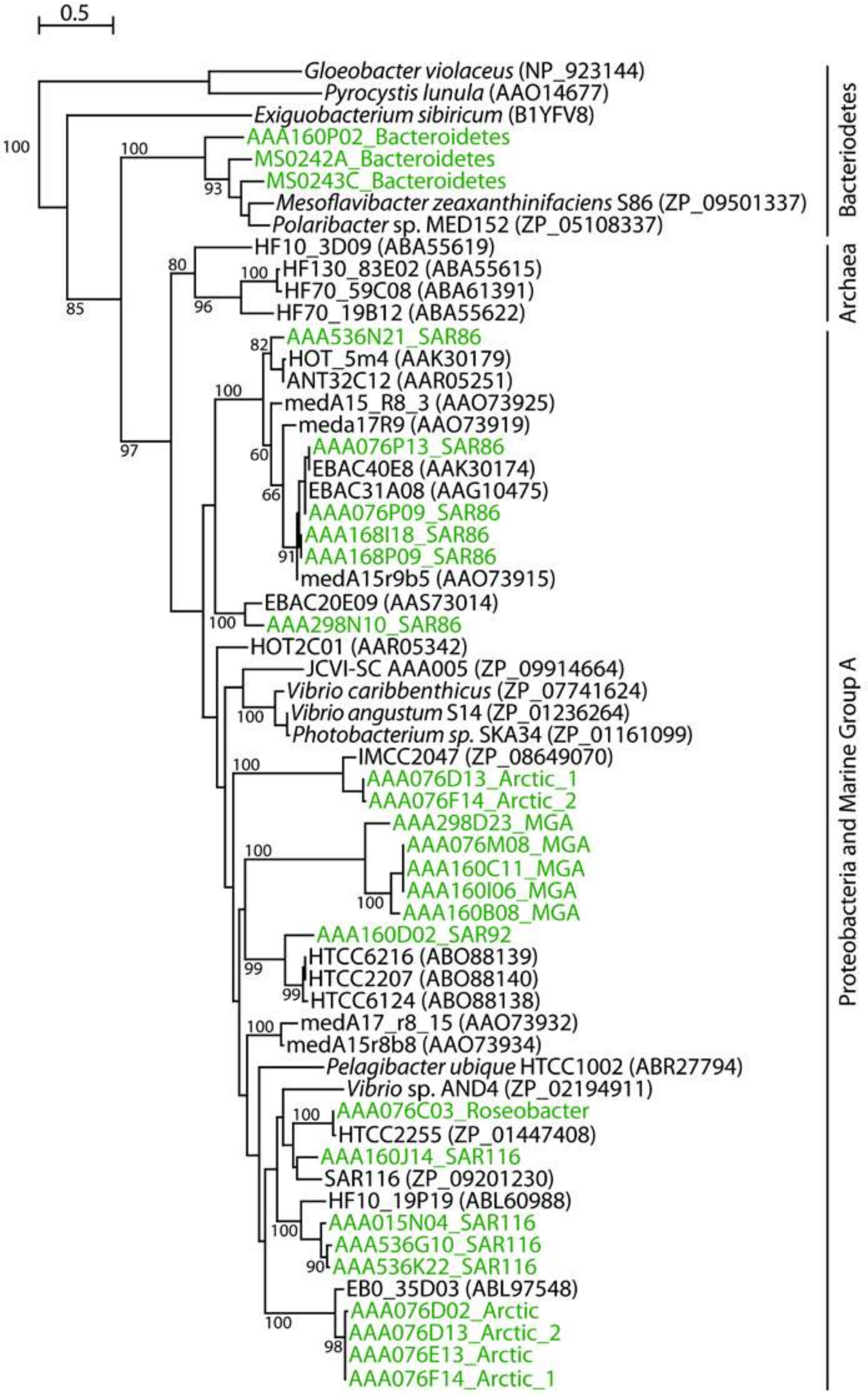


Figure S4. Phylogenetic analysis of proteorhodopsin genes from SAGs, cultures, and environmental clones. Unrooted phylogenetic tree based on proteorhodopsin gene protein sequences showing the phylogenetic affiliation of putative proteorhodopsin sequences identified on surface ocean SAGs. The tree was inferred using maximum likelihood implemented in PhyML (100 bootstrap replicates). The bar represents amino acid substitutions per site.

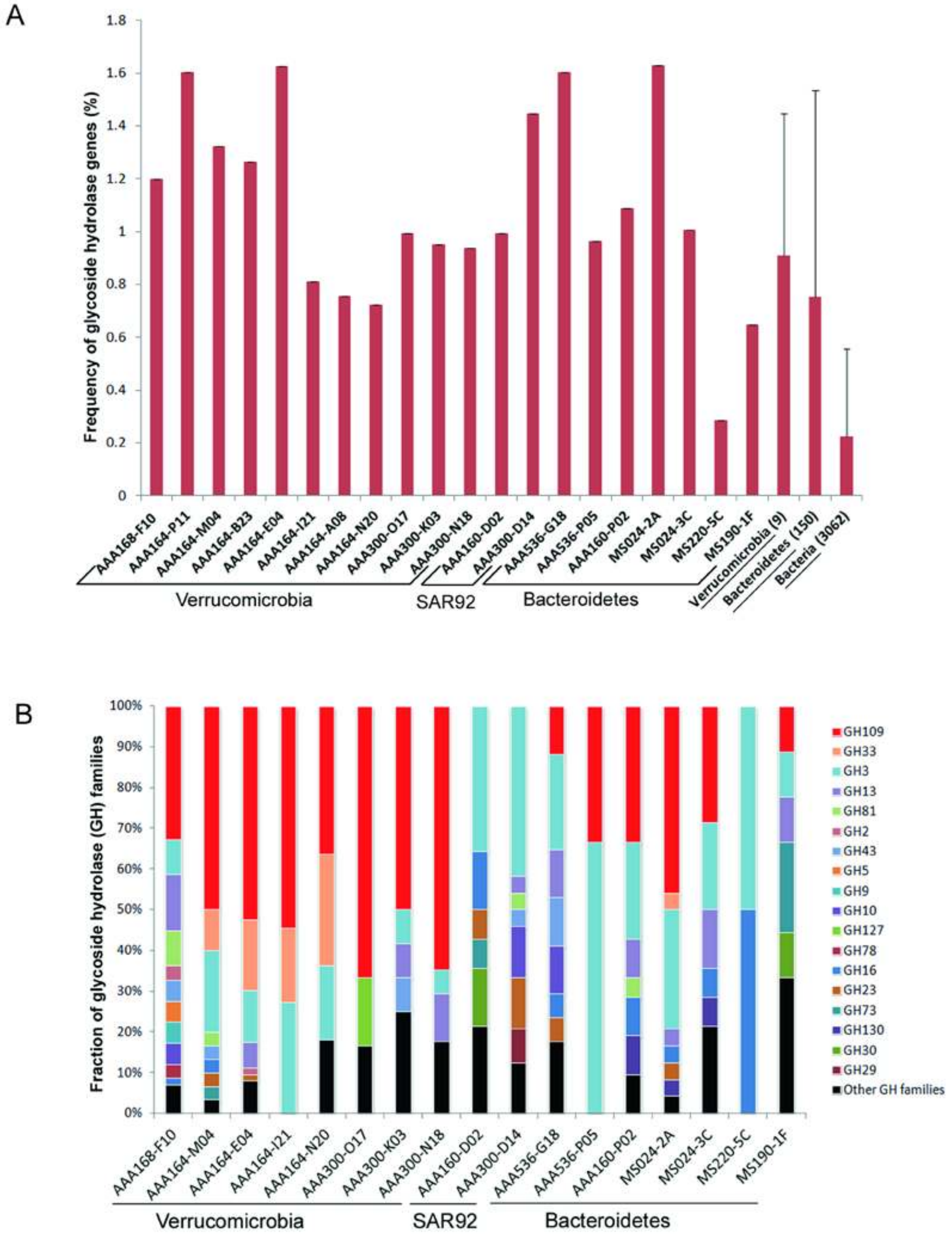


Figure S5. Genomic comparison of glycoside hydrolases in those SAGs showing elevated frequency of genes encoding extracellular proteins. Frequency of glycoside hydrolase genes involved in polysaccharide hydrolysis in Verrucomicrobia, Bacteroidetes and SAR92 SAG genomes (*A*). Frequency was estimated by dividing the total number of genes annotated as glycoside hydrolases by the total number of genes annotated. Bioinformatic resources of the Integrated Microbial Genomes (IMG) system were used to estimate the frequency of glycoside hydrolase (E.C. 3.2.1.x; see CAZy database (58) in the publicly available prokaryote genomes. Fraction of glycoside hydrolase (GH) families detected for each SAG (*B*). Fraction of the different GH families (according to nomenclature in CAZy database (58) was obtained by dividing the number of glycoside hydrolase genes belonging to a specific family by the total number of glycoside hydrolase genes annotated for each SAG. Glycoside hydrolase families were automatically annotated by CAZymes Analysis Toolkit applying the association rule learning algorithm (59) and then, the resulting annotation was carefully revised. Other GH families in figure legend are: 2, 17, 18, 20, 26, 31, 32, 65, 75, 84, 92, 94, 97, 103, 114 and 125.

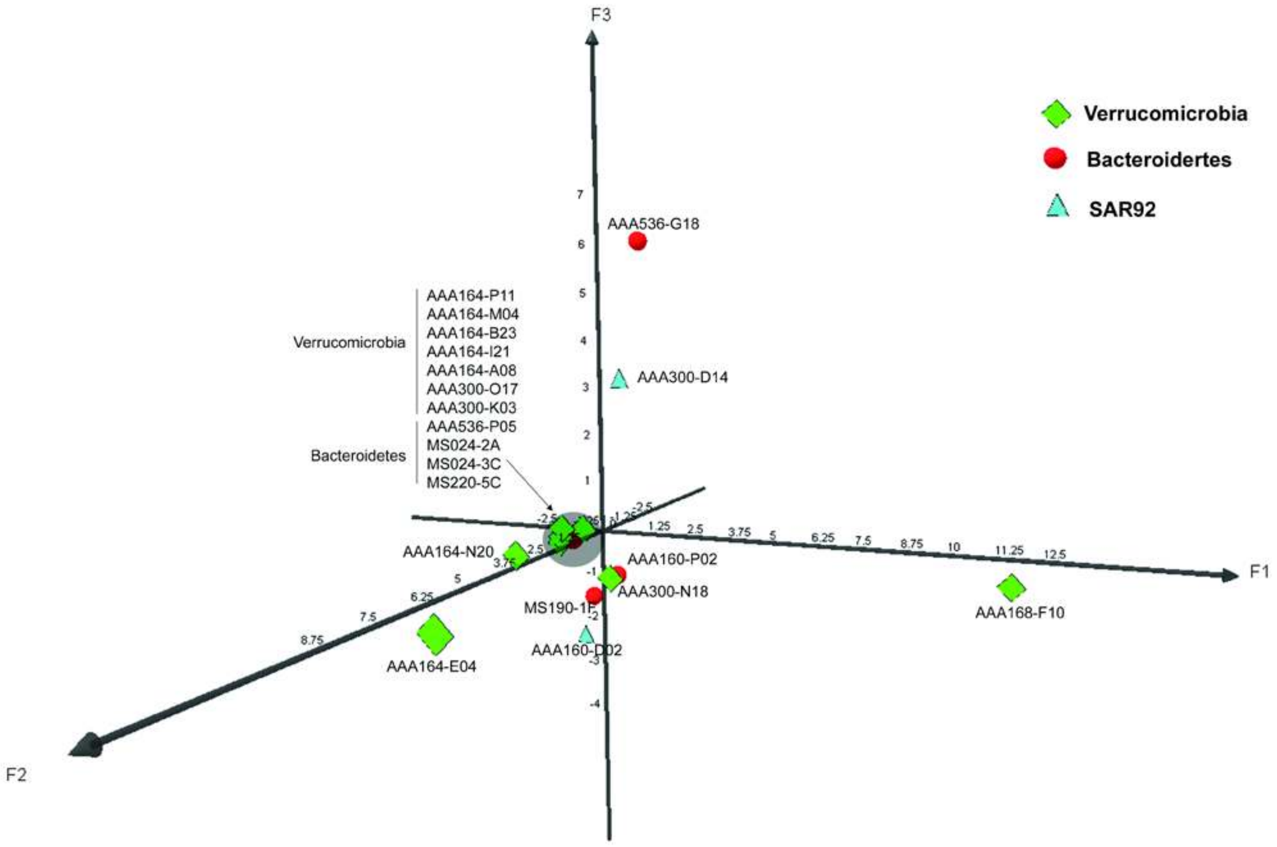


Figure S6. Principal component analysis (PCA) of repertory of glycoside hydrolases found in each SAG. Data used for the PCA analysis is derived from Figure S4B.

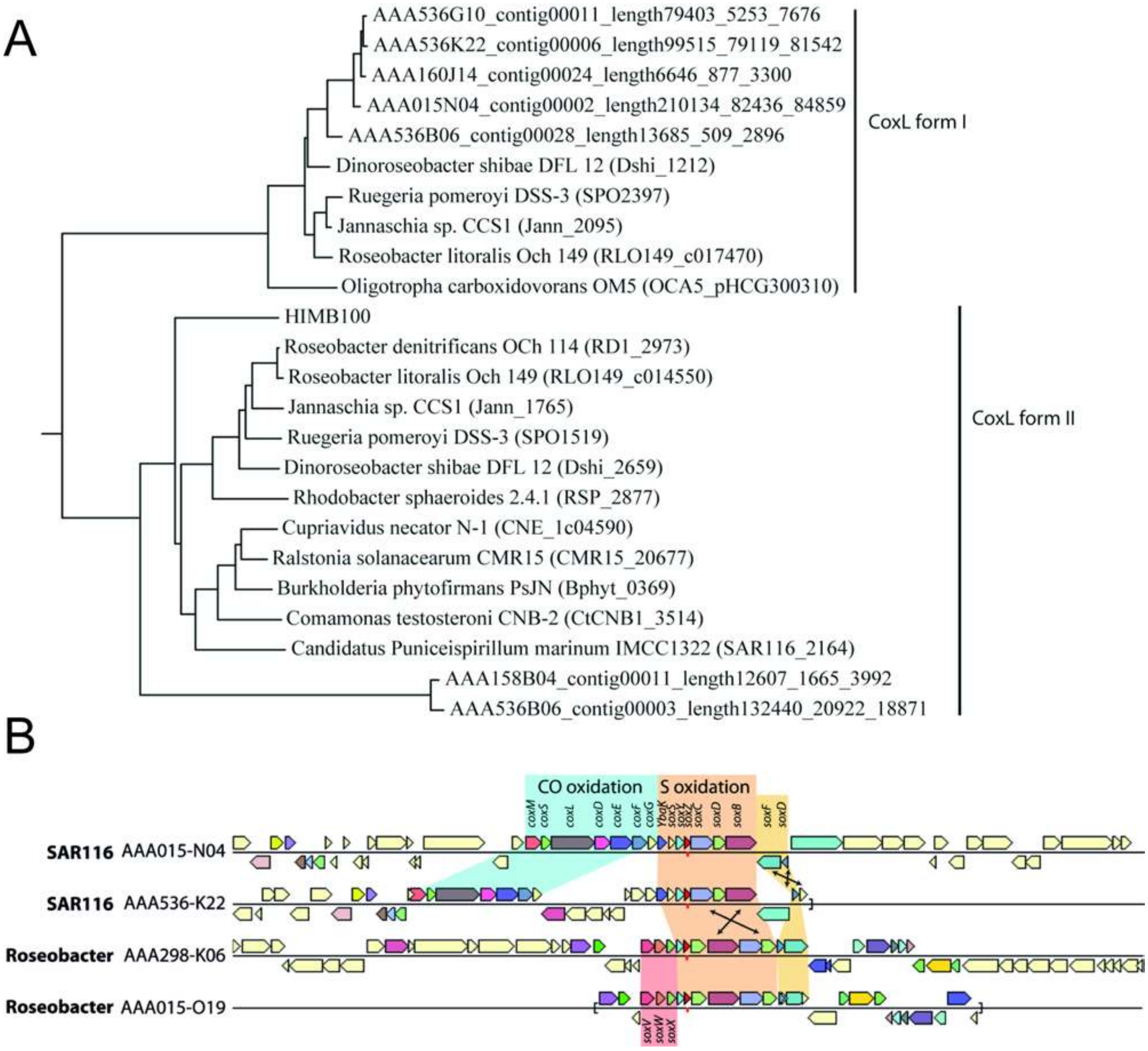

A
AAA536G10_contig00011_length79403_5253_7676
AAA536K22_contig00006_length99515_79119_81542
AAA160J14_contig00024_length6646_877_3300
AAA015N04_contig00002_length210134_82436_84859
AAA536B06_contig00028_length13685_509_2896
Dinoroseobacter shibae DFL 12 (Dshi_1212)
Ruegeria pomeroyi DSS-3 (SPO2397)
Jannaschia sp. CCS1 (Jann_2095)
Roseobacter litoralis Och 149 (RLO149_c017470)
Oligotropha carboxidovorans OM5 (OCA5_pHCG300310)
CoxL form I
HIMB100
Roseobacter denitrificans OCh 114 (RD1_2973)
Roseobacter litoralis Och 149 (RLO149_c014550)
Jannaschia sp. CCS1 (Jann_1765)
Ruegeria pomeroyi DSS-3 (SPO1519)
Dinoroseobacter shibae DFL 12 (Dshi_2659)
Rhodobacter sphaeroides 2.4.1 (RSP_2877)
Cupriavidus necator N-1 (CNE_1c04590)
Ralstonia solanacearum CMR15 (CMR15_20677)
Burkholderia phytofirmans PsJN (Bphyt_0369)
Comamonas testosteroni CNB-2 (CtCNB1_3514)
Candidatus Puniceispirillum marinum IMCC1322 (SAR116_2164)
AAA158B04_contig00011_length12607_1665_3992
AAA536B06_contig00003_length132440_20922_18871
CoxL form II
B
CO oxidation
S oxidation
coxM
coxS
coxL
coxD
coxE
coxF
coxG
YbaK
soxS
soxY
soxZ
soxC
soxD
soxB
soxF
soxD
SAR116 AAA015-N04
SAR116 AAA536-K22
Roseobacter AAA298-K06
Roseobacter AAA015-O19
soxV
soxW
soxX

Figure S7. Phylogenetic and synteny analysis of chemoautotrophy genes. Phylogenetic analysis of carbon monoxide dehydrogenase (*coxL*) genes (*A*) and synteny of chemoautotrophy genes (*B*) of Alphaproteobacteria SAGs.

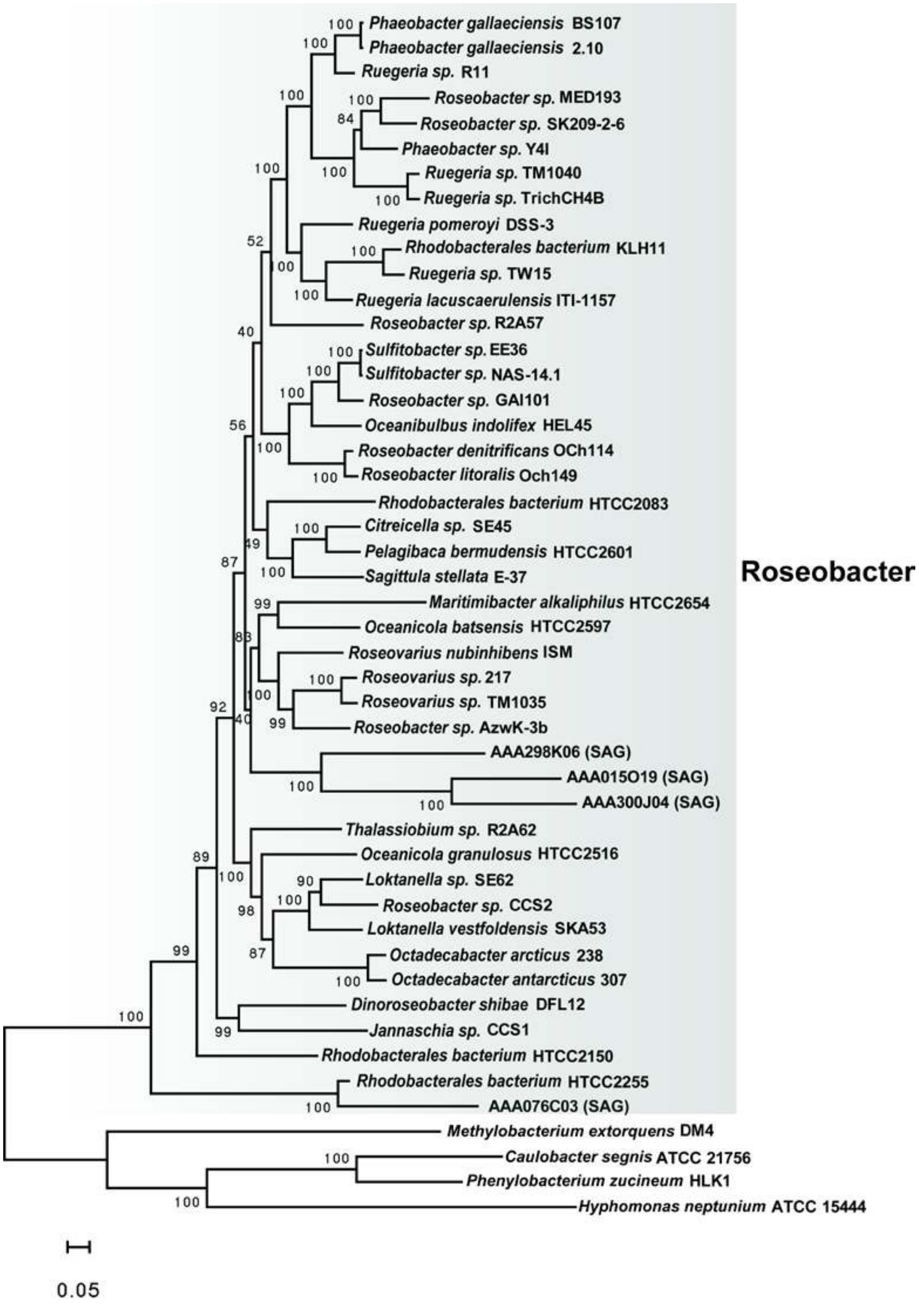

Phaeobacter gallaeciensis BS107
Phaeobacter gallaeciensis 2.10
Ruegeria sp. R11
Roseobacter sp. MED193
Roseobacter sp. SK209-2-6
Phaeobacter sp. Y4I
Ruegeria sp. TM1040
Ruegeria sp. TrichCH4B
Ruegeria pomeroyi DSS-3
Rhodobacterales bacterium KLH11
Ruegeria sp. TW15
Ruegeria lacuscaerulensis ITI-1157
Roseobacter sp. R2A57
Sulfitobacter sp. EE36
Sulfitobacter sp. NAS-14.1
Roseobacter sp. GAI101
Oceanibulbus indolifex HEL45
Roseobacter denitrificans OCh114
Roseobacter litoralis Och149
Rhodobacterales bacterium HTCC2083
Citreicella sp. SE45
Pelagibaca bermudensis HTCC2601
Sagittula stellata E-37
Maritimibacter alkaliphilus HTCC2654
Oceanicola batsensis HTCC2597
Roseovarius nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
Roseobacter sp. AzwK-3b
AAA298K06 (SAG)
AAA015O19 (SAG)
AAA300J04 (SAG)
Thalassiobium sp. R2A62
Oceanicola granulosus HTCC2516
Loktanella sp. SE62
Roseobacter sp. CCS2
Loktanella vestfoldensis SKA53
Octadecabacter arcticus 238
Octadecabacter antarcticus 307
Dinoroseobacter shibae DFL12
Jannaschia sp. CCS1
Rhodobacterales bacterium HTCC2150
Rhodobacterales bacterium HTCC2255
AAA076C03 (SAG)
Methylobacterium extorquens DM4
Caulobacter segnis ATCC 21756
Phenylobacterium zucineum HLK1
Hyphomonas neptunium ATCC 15444
Roseobacter
100
100
100
100
84
100
100
100
52
100
100
100
40
100
100
100
56
100
100
100
49
87
100
99
88
100
100
99
92
40
100
100
89
100
90
100
98
87
100
99
100
99
100
100
100
100
0.05

Figure S8. Maximum likelihood phylogeny of the Roseobacter clade using 49 concatenated orthologous protein sequences. The tree was constructed using RAxML 7.3.0 software with data partition model which allows each protein alignment to have its own evolutionary model. Values at the nodes show the number of times the clade defined by that node appeared in the 100 bootstrapped data sets. Grey shading indicates the Roseobacter clade. Tree is rooted using species associated with Rhizobiales, Hyphomonadaceae, and Caulobacterales. Although the branching order of several major clades is not resolved, the three SAGs (AAA298-K06, AAA015-O19, AAA300-J04) constitute a well-supported clade in which no cultured relatives are found.

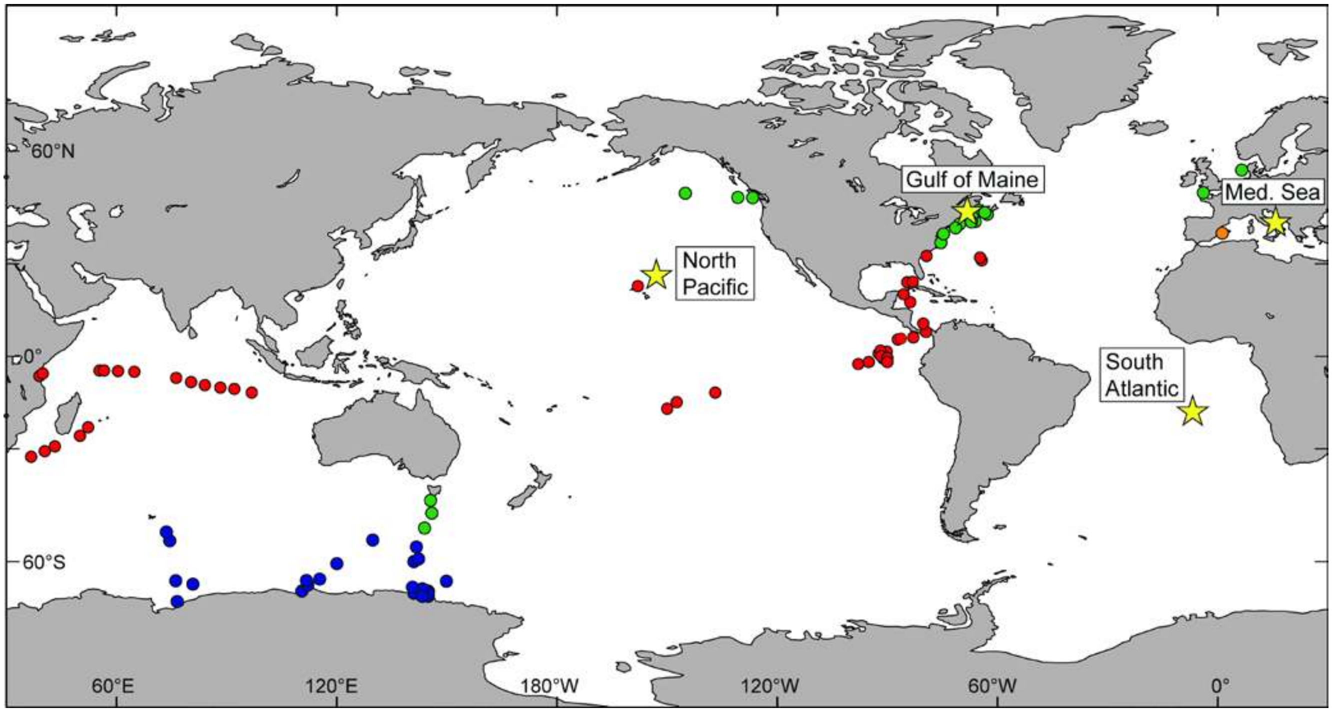


Figure S9. SAG collection and metagenome sample locations. Colored circles indicate locations and climate zone of metagenomes used for fragment recruitment, and stars represent the four SAG sampling locations. Red, tropical zone; blue, polar zone; green, temperate zone; orange, Mediterranean Sea.

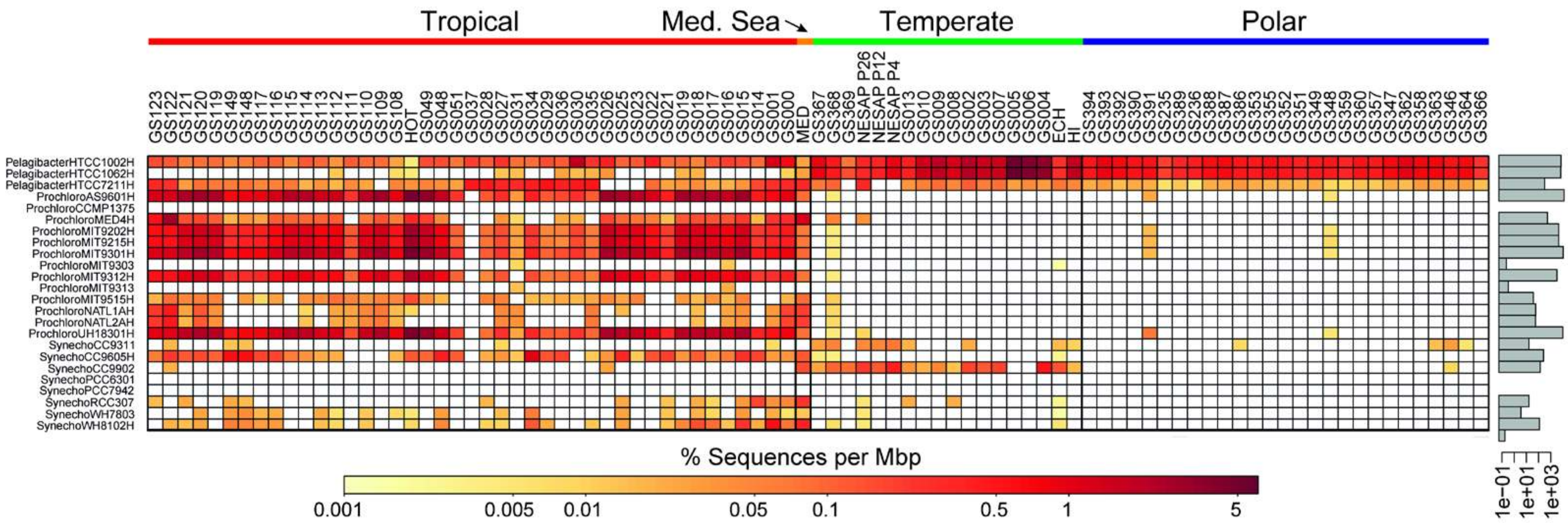


Figure S10. Metagenome fragment recruitment of 24 PSP (*Prochlorococcus-Synechococcus-Pelagibacter*) cultures. Fragment recruitment was carried out as described in Fig. 3. Percentages of aligned sequences from all metagenomes to individual SAGs are presented as grey bars on the y-axis. Metagenomes used in fragment recruitment are listed along the top x-axis, color bars indicate the surface ocean climate zone, and cultures are listed along the y-axis. HOT, HOT Station ALOHA; MED, Mediterranean Sea; NESAP, Northeast subarctic Pacific Ocean LineP stations; ECH, English Channel; HI, Helgoland Island.

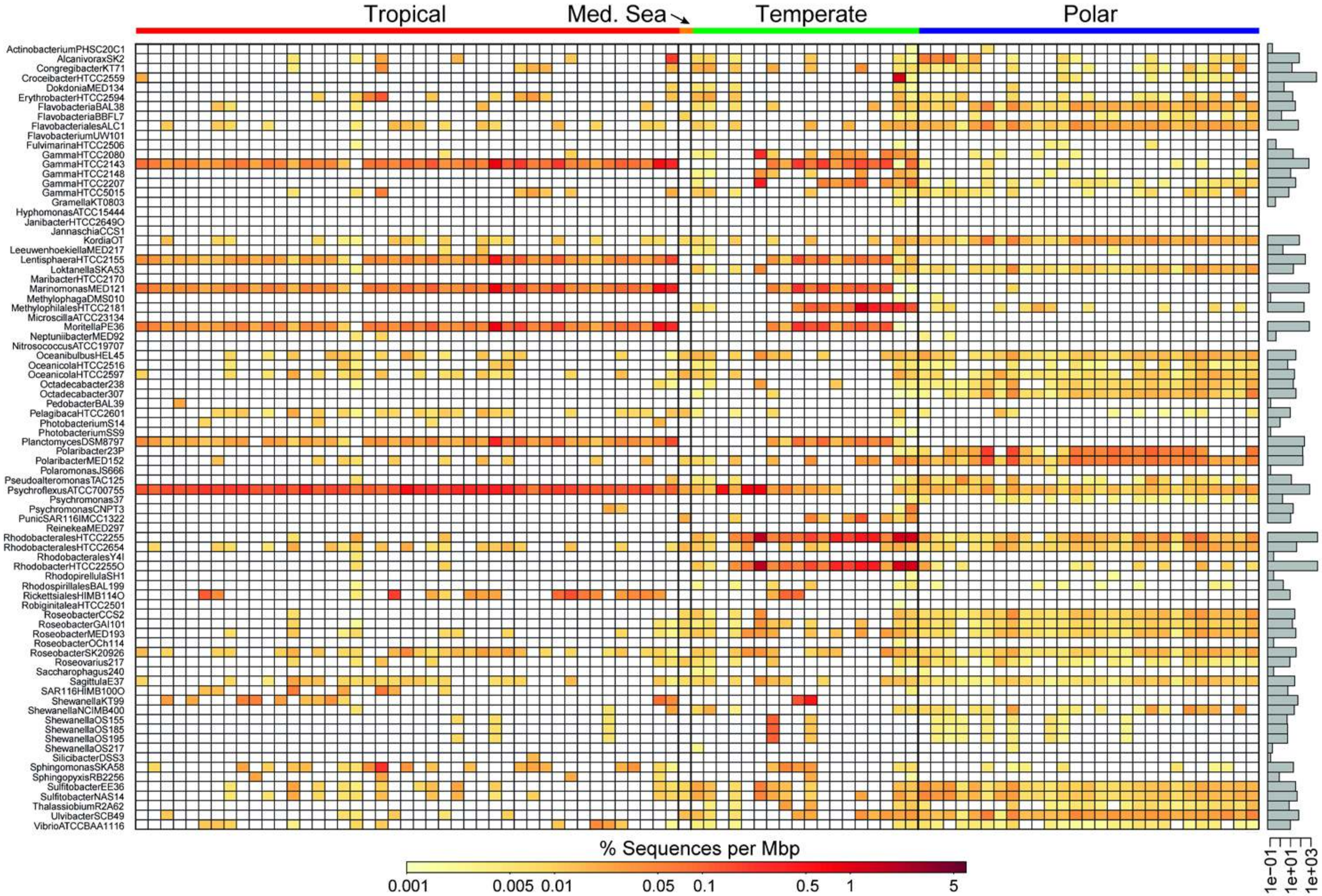
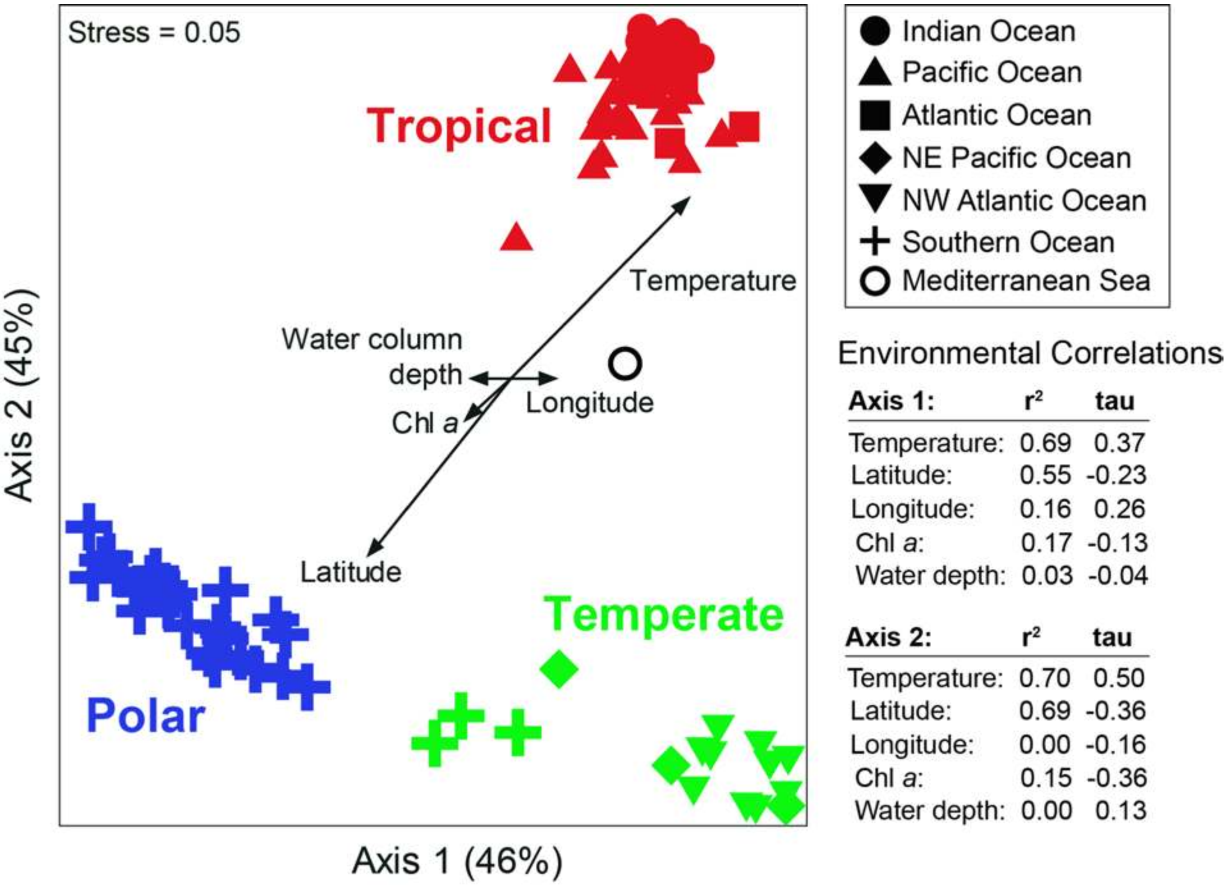


Figure S11. Metagenome fragment recruitment of 82 marine cultures. Fragment recruitment was carried out as described in Fig. 3. Percentages of aligned sequences from all metagenomes to individual SAGs are presented as grey bars on the y-axis. Cultures are listed along the y-axis and color bars indicating the surface ocean climate zone. Metagenomes are in the same order as presented in Fig. S10 along the top x-axis.



Environmental Correlations

Axis 1:	**r^2**	**tau**
Temperature:	0.69	0.37
Latitude:	0.55	-0.23
Longitude:	0.16	0.26
Chl *a*:	0.17	-0.13
Water depth:	0.03	-0.04

Axis 2:	**r^2**	**tau**
Temperature:	0.70	0.50
Latitude:	0.69	-0.36
Longitude:	0.00	-0.16
Chl *a*:	0.15	-0.36
Water depth:	0.00	0.13

Figure S12. Clustering of metagenomes from climatic zones as a function of SAG fragment recruitment. Metagenome samples are colored by their climatic zone, and symbol shapes indicate geographic location. Non-metric multidimensional analysis was used to analyze Bray-Curtis dissimilarities of SAG recruitment abundances (arcsin square-root transformed). Pearson (r^2) and Kendall (tau) correlation coefficients were calculated for each environmental parameter. Chl *a*, chlorophyll *a* concentration; water column depth, depth of water column at each sampling location.

Table S1. Sources of samples used for single amplified genome (SAG) generation. The Mediterranean Sea sample was collected at the deep chlorophyll *a* maximum. NA, not available; Verruco, Verrucomicrobia.

Date	Latitude	Longitude	Depth (m)	T (°C)	S (PSU)	DO (mL L^{-1})	SAG labels	Lineages
Gulf of Maine								
03/28/06	43°50’39.87” N	69°38’27.49” W	1	7.0	33.0	NA	MS024	SAR116 (3)
							MS190	Roseobacter (2)
							MS220	SAR86 (4)
08/16/09	43°50’39.87” N	69°38’27.49” W	1	22.3	30.0	NA	AAA076	SAR92 (1)
							AAA158	Arctic96BD-19 (4)
							AAA160	Bacteroidetes (5)
							AAA164	Marine Group A (4)
							AAA168	Verruco-Verruco (8)
								Verruco-S3 (4)
								Thaumarchaeota (1)
North Pacific subtropical gyre (HOT station ALOHA)								
09/09/09	22°45’00” N	158°00’00” W	25	26.5	35.5	4.69	AAA298	SAR116 (2)
							AAA300	Roseobacter (2)
								SAR86 (1)
								SAR92 (1)
								Marine Group A (1)
								Verruco-Opitutae (3)
South Atlantic subtropical gyre								
12/01/07	12°29’41.40” S	4°59’55.20” W	10	21.9	36.4	4.70	AAA015	SAR116 (1)
								Roseobacter (1)
								Actinobacteria (2)
Mediterranean Sea								
11/18/09	42°12’19.26” N	17°42’50.46” E	56	15.5	38.5	NA	AAA536	SAR116 (3)
								SAR86 (2)
								Bacteroidetes (2)

Table S2. SAG sequencing and assembly characteristics. Verruco, Verrucomicrobia.

SAG	Cluster	Sequencing effort (Mbp)	No. of contigs	Assembly size (Mbp)	% Genome recovery	Estimated genome size (Mbp)	Protein coding genes	%GC
Gulf of Maine								
AAA158-B04	SAR116	3,017.45	70	0.52	12.1	4.28	547	46.4
AAA158-M15	SAR116	2,734.96	52	0.40	13.8	2.92	460	31.1
AAA160-J14	SAR116	2,691.90	47	0.94	37.9	2.48	799	31.0
AAA076-C03	Roseobacter	3,029.55	107	2.00	67.2	2.97	1988	37.9
AAA076-E06	Roseobacter	2,962.94	23	0.22	1.7	12.77	322	38.1
AAA076-P09	SAR86	2,809.56	58	1.00	85.1	1.17	1074	33.3
AAA076-P13	SAR86	2,897.79	39	1.30	91.5	1.42	1369	33.6
AAA168-I18	SAR86	2,885.60	35	0.96	87.2	1.10	1014	32.5
AAA168-P09	SAR86	2,398.16	46	1.30	95.7	1.36	1390	33.0
AAA160-D02	SAR92	3,067.74	117	0.88	63.8	1.38	904	43.1
AAA076-D02	Arctic96BD-19	2,746.07	55	1.80	95.7	1.88	1787	38.1
AAA076-D13	Arctic96BD-19	2,782.59	81	1.70	87.2	1.95	1730	38.0
AAA076-E13	Arctic96BD-19	2,941.49	88	0.98	34.0	2.87	1045	37.3
AAA076-F14	Arctic96BD-19	2,643.54	48	1.80	93.6	1.92	1788	36.9
MS024-2A	Bacteroidetes	[1]112.45	17	1.91	91.0	2.10	1780	36.0
MS024-3C	Bacteroidetes	[1]130.72	21	1.52	78.0	1.95	1388	39.0
MS190-1F	Bacteroidetes	[1]626.80	38	1.52	48.8	3.12	1391	36.1
MS220-5C	Bacteroidetes	[1]537.10	22	0.71	19.8	3.59	696	39.4
AAA160-P02	Bacteroidetes	4,226.04	157	2.50	84.9	2.95	2390	31.6
AAA076-M08	Marine Group A	3,154.46	49	0.45	73.3	0.61	513	32.7
AAA160-B08	Marine Group A	2,102.06	47	0.94	84.4	1.11	999	33.1
AAA160-C11	Marine Group A	4,036.86	64	0.96	91.1	1.05	1069	32.6
AAA160-I06	Marine Group A	4,507.13	78	0.97	95.6	1.02	1097	32.6
AAA164-A21	Verruco-Verruco	2,439.26	318	1.10	23.6	4.65	1196	48.6
AAA164-B23	Verruco-Verruco	5,231.47	30	0.12	0.3	4.00	158	46.2
AAA164-L15	Verruco-Verruco	3,974.62	225	2.50	50.0	5.00	2222	48.8
AAA164-M04	Verruco-Verruco	4,204.91	282	2.50	53.0	4.71	2308	48.5
AAA164-O14	Verruco-Verruco	4,335.31	522	3.30	61.5	5.36	3117	48.5
AAA164-P11	Verruco-Verruco	4,286.52	49	0.29	5.2	5.56	368	49.7
AAA168-E21	Verruco-Verruco	2,713.90	367	2.40	57.6	4.17	2265	48.6
AAA168-F10	Verruco-Verruco	1,971.36	560	4.50	58.2	7.73	4057	47.3
AAA164-A08	Verruco-S3	4,187.21	15	0.09	0.9	9.87	132	37.3
AAA164-E04	Verruco-S3	3,008.07	506	4.10	74.9	5.48	3776	47.5
AAA164-I21	Verruco-S3	3,947.13	389	1.10	29.7	3.70	1341	46.0
AAA164-N20	Verruco-S3	3,552.69	461	1.40	36.7	3.82	1638	45.8
North Pacific subtropical gyre (HOT station ALOHA)								
AAA300-B11	SAR116	4,558.86	51	0.18	0.0	NA	230	44.4
AAA300-J16	SAR116	8,454.80	214	1.00	31.0	3.22	1181	45.4
AAA298-K06	Roseobacter	4,069.72	231	1.70	39.7	4.29	1931	39.9
AAA300-J04	Roseobacter	4,123.54	77	0.62	22.4	2.77	688	39.1

SAG	Cluster	Sequencing effort (Mbp)	No. of contigs	Assembly size (Mbp)	% Genome recovery	Estimated genome size (Mbp)	Protein coding genes	%GC
AAA298-N10	SAR86	2,183.29	189	1.00	87.2	1.15	1197	32.9
AAA300-D14	SAR92	2,604.59	154	1.50	83.0	1.81	1449	38.0
AAA298-D23	Marine Group A	5,060.55	50	1.00	97.8	1.02	1081	31.9
AAA300-K03	Verruco-Opitutae	2,603.89	182	1.20	38.8	3.09	1243	43.0
AAA300-N18	Verruco-Opitutae	2,423.84	260	1.70	49.4	3.44	1685	43.8
AAA300-O17	Verruco-Opitutae	2,551.41	203	1.10	41.8	2.63	1192	43.2
South Atlantic subtropical gyre								
AAA015-N04	SAR116	3,538.14	132	1.70	69.0	2.46	1894	30.8
AAA015-O19	Roseobacter	6,383.59	159	1.70	44.8	3.79	1861	38.5
AAA015-D07	Actinobacteria	3,943.43	42	0.63	31.7	1.99	710	32.3
AAA015-M09	Actinobacteria	2,635.47	67	0.67	61.7	1.08	773	34.2
Mediterranean Sea								
AAA536-B06	SAR116	7,329.83	182	1.60	63.8	2.51	1754	41.3
AAA536-G10	SAR116	2,144.80	148	2.20	91.4	2.41	2303	30.8
AAA536-K22	SAR116	1,609.29	92	2.00	75.9	2.64	2150	31.6
AAA536-J20	SAR86	2,136.24	94	0.36	17.0	2.10	448	33.0
AAA536-N21	SAR86	2,005.02	47	0.44	70.2	0.62	483	32.9
AAA536-G18	Bacteroidetes	3,891.41	137	1.20	52.3	2.30	1246	31.2
AAA536-P05	Bacteroidetes	3,427.22	55	0.26	48.1	3.21	311	38.6

[1] Sanger and 454, or combined with Illumina sequencing was employed for these SAGs, and these were not used in the averages and ranges presented.

Average	3,445.52	149	1.32	55.4	3.10	1,356	38.6
Range	1,609-8,455	15-560	0.09-4.50	0.3-97.8	0.61-12.77	132-4,057	30.8-49.7

Table S3. Genome characteristics used to compare marine cultures and SAGs with PCA. GC content and non-coding DNA percentages were extracted from IMG. Protein localization category values “Multi”, “Cytoplasmic”, “Cytoplasmic membrane”, “Periplasmic”, “Outer membrane”, and “Extracellular” were calculated according to Lauro et al. (60). The frequency of COG categories T (Signal transduction mechanisms), V (Defense mechanism), K (Transcription), Q (Secondary metabolites biosynthesis, transport and catabolism), and I (Lipid transport and metabolism) were also calculated according to Lauro et al. (60).

Genome	%GC	% Non-coding DNA	Multi	Cytoplasmic	Cytoplasmic membrane	Periplasmic	Outer membrane	Extracellular	T	V	K	Q	I
Alcanivorax_borkumensis_SK2	54.7	12.0	0.02613	0.48566	0.22323	0.01887	0.02142	0.00690	0.04869	0.01069	0.05899	0.02890	0.05740
Candidatus_Pelagibacter_HTCC7211	29.0	8.0	0.02350	0.56531	0.19074	0.01797	0.00553	0.00829	0.02186	0.00859	0.02888	0.02810	0.03981
Candidatus_Pelagibacter_ubique_HTCC1002	29.0	4.1	0.02010	0.55779	0.19239	0.01292	0.00861	0.00431	0.01997	0.00915	0.03577	0.02829	0.04409
Candidatus_Pelagibacter_ubique_HTCC1062	29.7	3.9	0.01625	0.56721	0.19719	0.01403	0.00960	0.00517	0.02068	0.00910	0.03639	0.03226	0.04301
Candidatus_Puniceispirillum_IMCC1322	48.9	10.1	0.02123	0.46716	0.22021	0.02241	0.00590	0.01022	0.02356	0.00873	0.04756	0.03272	0.03709
Congregibacter_KT71	57.7	10.3	0.03426	0.44177	0.20655	0.01599	0.02208	0.00787	0.04127	0.01651	0.05175	0.03873	0.05460
Croceibacter_atlanticus_HTCC2559	33.9	8.2	0.03163	0.43472	0.18647	0.00993	0.03935	0.01912	0.03350	0.01483	0.05656	0.03075	0.04887
Dokdonia_MED134	37.3	8.5	0.03736	0.42833	0.19192	0.00883	0.04008	0.01291	0.04420	0.01872	0.06188	0.02236	0.03900
Erythrobacter_litoralis_HTCC2594	63.1	8.5	0.03188	0.45799	0.18931	0.01827	0.02192	0.00565	0.03751	0.01236	0.04817	0.04646	0.07246
Flavobacteria_bacterium_BAL38	31.5	8.4	0.03331	0.45176	0.17841	0.00919	0.03522	0.01953	0.02947	0.01842	0.04665	0.02026	0.04052
Flavobacteria_bacterium_BBFL7	35.4	9.4	0.04784	0.46682	0.18017	0.00810	0.03858	0.02199	0.04293	0.02118	0.05037	0.02576	0.04408
Flavobacteriales_ALC1	32.7	7.3	0.05515	0.42409	0.19390	0.00697	0.04122	0.02206	0.04942	0.02751	0.07226	0.02611	0.03357
Flavobacteriales_HTCC2170	37.0	9.0	0.04140	0.45457	0.18516	0.01121	0.03249	0.00920	0.03359	0.02153	0.06029	0.02670	0.04048
Flavobacterium_johnsoniae_UW101	34.1	12.2	0.03588	0.43911	0.16006	0.01495	0.04405	0.01515	0.06204	0.02026	0.07787	0.03197	0.04115
Fulvimarina_pelagi	61.1	12.6	0.02877	0.43021	0.21364	0.02318	0.00719	0.00773	0.04082	0.00745	0.05118	0.03110	0.03758
Gamma_HTCC2080	52.0	10.0	0.03422	0.45840	0.19906	0.02009	0.02229	0.01005	0.03235	0.01209	0.04301	0.06114	0.08745
Gamma_HTCC2143	47.2	18.0	0.02430	0.45112	0.19033	0.01502	0.01666	0.00737	0.04310	0.01852	0.04151	0.05364	0.08301
Gamma_HTCC2148	53.0	11.3	0.03057	0.47557	0.18160	0.01516	0.02430	0.00784	0.03674	0.01464	0.05287	0.06183	0.08871
Gamma_HTCC2207	49.4	12.0	0.02471	0.50461	0.18467	0.01298	0.02052	0.01089	0.02989	0.01087	0.05480	0.03623	0.06114
Gamma_HTCC5015	54.1	13.0	0.02725	0.47987	0.19317	0.01139	0.02277	0.01179	0.05634	0.01207	0.05634	0.02565	0.04728
Gramella_forsetii_KT0803	36.6	9.2	0.04241	0.46261	0.18331	0.00865	0.02930	0.00809	0.04187	0.01763	0.05333	0.02909	0.04187
Hyphomonas_neptunium_ATCC_15444	61.9	9.5	0.03110	0.42967	0.20114	0.01598	0.02397	0.00942	0.03568	0.01634	0.06669	0.04368	0.06469
Jannaschia_CCS1	62.3	9.2	0.02802	0.43894	0.23091	0.02031	0.00584	0.00724	0.03214	0.00783	0.07406	0.04136	0.04584
Kordia_algicida_OT-1	34.3	11.6	0.04231	0.40297	0.15995	0.00820	0.02658	0.01949	0.05693	0.01969	0.06849	0.03425	0.03467
Leeuwenhoekiella_blandensis_MED217	40.0	9.0	0.04016	0.43079	0.18340	0.01285	0.04257	0.00857	0.03916	0.02267	0.05565	0.02391	0.03504
Lentisphaera_araenosa_HTCC2155	41.0	11.3	0.05741	0.43907	0.13636	0.01509	0.01038	0.01078	0.05571	0.01354	0.07695	0.01754	0.02555
Loktanella_vestfoldensis_SKA53	59.8	8.1	0.02347	0.42797	0.23533	0.01956	0.00652	0.00945	0.03496	0.00977	0.05752	0.03609	0.04774
Marinomonas_MED121	40.8	12.3	0.02178	0.45539	0.20602	0.02075	0.01307	0.00830	0.08019	0.00700	0.10773	0.02778	0.03478
Methylophaga_DSM010	46.9	11.1	0.01766	0.44285	0.18461	0.02166	0.01966	0.00467	0.06548	0.01233	0.05315	0.01658	0.02679
Methylophilales_bacterium_HTCC2181	38.0	5.0	0.02317	0.56726	0.17115	0.01570	0.01644	0.00598	0.02278	0.01058	0.03743	0.01871	0.03417
Microscilla_marina_ATCC23134	40.8	18.0	0.02560	0.32179	0.14725	0.00841	0.02176	0.02320	0.10237	0.03026	0.09884	0.03177	0.03833
Moritella_PE36	41.0	13.2	0.01924	0.39239	0.23552	0.02301	0.02343	0.01380	0.06455	0.01281	0.08094	0.02126	0.03125
Nitrosococcus_oceani_ATCC_19707	50.0	15.0	0.02320	0.46039	0.22207	0.01690	0.01525	0.00862	0.04529	0.02141	0.03664	0.02223	0.02923
Oceanibulbus_indolifex	60.0	10.6	0.02504	0.41849	0.22177	0.02745	0.00698	0.00674	0.04385	0.00717	0.06907	0.03726	0.04586
Oceanicola_batsensis	66.2	10.8	0.03205	0.48433	0.20655	0.01994	0.00712	0.00736	0.02935	0.00866	0.06596	0.04807	0.07323
Oceanicola_granulosus	70.2	8.5	0.04341	0.46764	0.22610	0.02716	0.00666	0.00879	0.03357	0.00816	0.06622	0.03840	0.04264
Oceanospirillum_sp_MED92	46.6	8.9	0.02467	0.49295	0.22451	0.01871	0.01220	0.00759	0.10688	0.00829	0.07310	0.02488	0.03563
Octadecabacter_238	50.8	18.2	0.01166	0.38721	0.14724	0.01423	0.00291	0.00566	0.02363	0.00716	0.05107	0.02983	0.03508
Octadecabacter_307	57.8	16.6	0.01565	0.35632	0.17307	0.01856	0.00437	0.00637	0.02569	0.00823	0.05661	0.03392	0.03915
Pedobacter_BAL39	45.2	9.2	0.02607	0.37436	0.18405	0.01725	0.04567	0.00764	0.08993	0.02060	0.08819	0.02901	0.03800
Photobacterium_angustum_S14	38.3	14.8	0.01865	0.41553	0.24133	0.02479	0.01975	0.01316	0.05288	0.01511	0.07983	0.01864	0.02896
Photobacterium_profundum_SS9	41.2	17.2	0.02186	0.41847	0.22554	0.02368	0.01840	0.00929	0.05734	0.01788	0.07826	0.02093	0.02703
Planctomyces_maris_DSM8797	49.7	12.8	0.04367	0.36929	0.17037	0.01543	0.00725	0.00972	0.06189	0.01882	0.06874	0.03793	0.03451
Polaribacter_irgensii_23-P	34.8	12.9	0.01760	0.44310	0.18224	0.00860	0.03246	0.00665	0.02989	0.01614	0.06097	0.02271	0.04124
Polaribacter_MED152	30.6	6.9	0.03919	0.47294	0.17581	0.00672	0.04143	0.01157	0.03738	0.01264	0.05992	0.02364	0.04508
Polaromonas_JS666	62.5	12.4	0.02219	0.42032	0.20576	0.04273	0.00935	0.00568	0.04657	0.00927	0.08112	0.03877	0.04888
Prochlorococcus_marinus_AS9601	31.3	8.8	0.01510	0.46226	0.19313	0.00364	0.00677	0.00312	0.01801	0.01488	0.03211	0.02662	0.03211
Prochlorococcus_marinus_CCMP1375	36.4	10.8	0.01593	0.42804	0.20393	0.00584	0.00637	0.00425	0.02147	0.01610	0.03374	0.02224	0.03221

Genome	%GC	% Non-coding DNA	Multi	Cytoplasmic	Cytoplasmic membrane	Periplasmic	Outer membrane	Extracellular	T	V	K	Q	I
Prochlorococcus_marinus_MED4	30.8	11.6	0.01514	0.49330	0.19045	0.00582	0.00641	0.00349	0.02062	0.01427	0.03410	0.02379	0.03172
Prochlorococcus_marinus_MIT_9215	31.1	10.4	0.01412	0.45436	0.18507	0.00403	0.00555	0.00353	0.01845	0.01384	0.03228	0.02613	0.03305
Prochlorococcus_marinus_MIT_9301	31.3	8.8	0.01311	0.45779	0.18773	0.00420	0.00734	0.00210	0.01947	0.01324	0.03505	0.02414	0.03193
Prochlorococcus_marinus_MIT_9303	50.0	15.5	0.01168	0.29997	0.20354	0.01001	0.00934	0.00934	0.02630	0.01520	0.03974	0.03507	0.02864
Prochlorococcus_marinus_MIT_9312	31.2	10.4	0.01768	0.48729	0.19282	0.00552	0.00829	0.00552	0.02103	0.01324	0.03271	0.02492	0.03349
Prochlorococcus_marinus_MIT_9515	30.8	11.1	0.01522	0.44911	0.19832	0.00525	0.00577	0.00315	0.02034	0.01330	0.03208	0.02739	0.03286
Prochlorococcus_marinus_MIT9313	50.7	17.8	0.01587	0.37241	0.22829	0.01454	0.00705	0.00441	0.02792	0.01923	0.04156	0.02792	0.02730
Prochlorococcus_marinus_NATL1A	35.1	12.7	0.01322	0.39444	0.19745	0.00502	0.00456	0.00319	0.02347	0.01287	0.03482	0.02725	0.03028
Prochlorococcus_marinus_NATL2A	35.0	14.4	0.01268	0.45455	0.19979	0.00740	0.00793	0.00529	0.02221	0.01302	0.03446	0.02680	0.03063
Pseudoalteromonas_haloplanktis_TAC125	40.2	11.6	0.02438	0.41595	0.22002	0.01922	0.03127	0.00947	0.06254	0.01345	0.06826	0.02555	0.04338
Psychroflexus_ATCC700755	32.8	14.5	0.02518	0.40898	0.15168	0.00785	0.02118	0.01555	0.03222	0.02562	0.04617	0.03196	0.04820
Psychromonas_CNPT3	38.6	14.2	0.01907	0.44334	0.22956	0.02127	0.01870	0.00807	0.04915	0.00991	0.05989	0.01735	0.02230
Psychromonas_ingrahamii_37	40.1	21.2	0.01354	0.45388	0.24062	0.02454	0.01326	0.01241	0.05589	0.01092	0.06232	0.02217	0.02377
Reinekea_MED297	52.2	9.9	0.02265	0.42473	0.23478	0.02336	0.01156	0.01321	0.08560	0.01263	0.08111	0.02273	0.03452
Rhodobacterales_HTCC2654	38.0	11.0	0.02377	0.44270	0.19822	0.01868	0.00531	0.00764	0.03196	0.00778	0.07310	0.04586	0.06476
Rhodobacterales_Y4I	67.8	13.8	0.02420	0.43963	0.19913	0.02541	0.00968	0.00871	0.05071	0.01217	0.07447	0.03999	0.04173
Rhodopirellula_baltica_SH1	55.4	5.0	0.03932	0.28287	0.15372	0.01078	0.00491	0.00846	0.06406	0.02156	0.06621	0.03141	0.03234
Rhodospirillales_BAL199	65.0	15.0	0.02252	0.45692	0.18424	0.01942	0.00669	0.00979	0.03699	0.00861	0.05794	0.05598	0.04463
Robiginitalea_biformata_HTCC2501	55.3	8.1	0.04926	0.45942	0.19207	0.00898	0.02912	0.00960	0.03613	0.02159	0.05162	0.02863	0.04411
Roseobacter_CCS2	55.2	9.0	0.02186	0.40000	0.22951	0.02186	0.00820	0.01093	0.03305	0.00964	0.06575	0.03029	0.04406
Roseobacter_denitrificans_OCh_114	59.0	10.6	0.02398	0.43134	0.22185	0.02930	0.00751	0.00751	0.03632	0.01307	0.06479	0.03544	0.04213
Roseobacter_GAI101	59.0	15.0	0.01856	0.44421	0.22103	0.02641	0.00857	0.00928	0.04028	0.00946	0.06002	0.04731	0.05515
Roseobacter_MED193	57.5	10.9	0.02029	0.43484	0.20750	0.02271	0.00617	0.00728	0.04092	0.00969	0.08318	0.04226	0.05357
Roseobacter_SK209-2-6	57.0	11.2	0.01763	0.45669	0.19771	0.02226	0.00882	0.00617	0.04162	0.00779	0.08969	0.03759	0.04431
Roseovarius_217	61.1	9.8	0.02368	0.43671	0.21563	0.02619	0.00775	0.00775	0.03669	0.00904	0.07649	0.04083	0.04910
Roseovarius_HTCC2601	66.5	11.3	0.02696	0.42920	0.22084	0.02806	0.00752	0.00844	0.03495	0.00754	0.07423	0.04043	0.04500
Saccharophagus_degradans_2-40	45.8	13.0	0.03343	0.38423	0.20734	0.02221	0.03069	0.02720	0.08411	0.01289	0.07154	0.02514	0.03029
Sagittula_stellata	64.9	11.7	0.02566	0.44188	0.21591	0.02743	0.00632	0.00710	0.04228	0.00874	0.07204	0.04086	0.04771
SAR116_HIMB100	50.5	8.1	0.02399	0.47772	0.20523	0.02228	0.00557	0.01200	0.01731	0.00655	0.03789	0.04537	0.05472
SAR86C	32.8	6.4	0.01328	0.45863	0.15015	0.00613	0.02145	0.00409	0.00857	0.02938	0.03550	0.03060	0.07099
SAR86D	31.5	11.2	0.01728	0.47290	0.14925	0.01257	0.01100	0.00393	0.00683	0.01463	0.04780	0.03610	0.06927
SAR86E	36.2	6.3	0.02584	0.56066	0.16798	0.00933	0.02441	0.00431	0.01180	0.01101	0.03619	0.05980	0.09127
Shewanella_baltica_OS155	46.3	15.6	0.02540	0.41301	0.22611	0.02718	0.02428	0.01225	0.07071	0.01729	0.07252	0.02065	0.02994
Shewanella_baltica_OS185	46.3	15.3	0.02777	0.39099	0.23009	0.03004	0.02777	0.01161	0.07423	0.01748	0.07800	0.02179	0.02959
Shewanella_baltica_OS195	46.3	15.1	0.02858	0.38588	0.22099	0.03029	0.02816	0.01237	0.07398	0.01474	0.07708	0.02173	0.02845
Shewanella_denitrificans_OS217	45.1	14.6	0.03010	0.40783	0.21417	0.02318	0.02291	0.02025	0.07011	0.01332	0.06631	0.02538	0.03236
Shewanella_frigidimarina_NCIMB_400	41.6	14.8	0.02457	0.41201	0.23951	0.02556	0.02805	0.01092	0.06885	0.01439	0.06913	0.02511	0.03640
Shewanella_KT99	46.0	16.9	0.02007	0.38465	0.19339	0.01700	0.01983	0.00826	0.05963	0.01461	0.05993	0.02087	0.03488
Silicibacter_pomeroyi_DSS-3	64.2	9.8	0.02893	0.48236	0.21731	0.03034	0.00776	0.00635	0.03154	0.01156	0.08830	0.04442	0.05125
Sphingomonas_SKA58	62.5	9.5	0.02351	0.38886	0.19520	0.02708	0.02657	0.00613	0.04466	0.01235	0.06335	0.03801	0.05353
Sphingopyxis_alaskensis_RB2256	65.5	9.4	0.03818	0.43505	0.19280	0.01972	0.02441	0.00626	0.03849	0.00990	0.06452	0.04289	0.06525
Sulfitobacter_sp_EE36	60.3	9.0	0.02303	0.43638	0.22337	0.02763	0.01036	0.00691	0.03433	0.00990	0.06504	0.03830	0.05612
Sulfitobacter_sp_NAS141	60.0	10.0	0.02196	0.43311	0.21227	0.02726	0.00984	0.00707	0.03284	0.00874	0.06870	0.03616	0.04640
Synechococcus_CC9311	52.4	12.8	0.01487	0.32365	0.22891	0.00864	0.00692	0.00795	0.03306	0.01322	0.04242	0.03306	0.03085
Synechococcus_CC9605	59.2	13.1	0.01323	0.36106	0.20870	0.01134	0.00378	0.00378	0.02887	0.01414	0.03889	0.02534	0.02534
Synechococcus_CC9902	54.2	10.0	0.02037	0.39489	0.21630	0.00867	0.00477	0.00564	0.02556	0.01434	0.04052	0.02805	0.02930
Synechococcus_elongatus_PCC_6301	55.5	12.0	0.01662	0.38702	0.28097	0.01504	0.00435	0.00791	0.05189	0.01455	0.04413	0.02473	0.02085
Synechococcus_elongatus_PCC_7942	55.5	10.8	0.01615	0.37716	0.27536	0.01503	0.00488	0.00864	0.05165	0.01483	0.04495	0.02439	0.02104
Synechococcus_RCC307	60.8	5.8	0.01460	0.32702	0.23511	0.01420	0.00473	0.00552	0.02839	0.01419	0.04258	0.02957	0.02839
Synechococcus_sp_WH8102	59.4	9.7	0.01866	0.38547	0.20365	0.00913	0.00476	0.00635	0.02850	0.01513	0.04072	0.03083	0.02734
Synechococcus_WH_7803	60.2	6.6	0.02053	0.35452	0.23885	0.01579	0.00513	0.00632	0.03164	0.01525	0.03898	0.02768	0.02881
Thalassobium_R2A62	55.2	10.0	0.01893	0.41915	0.20092	0.02109	0.00649	0.00946	0.03150	0.01004	0.05884	0.03219	0.04119
Ulvibacter_SCB49	34.0	10.4	0.03494	0.41757	0.18725	0.00611	0.03596	0.01357	0.04247	0.02043	0.05538	0.02312	0.04301
Vibrio_harveyi_ATCC_BAA-1116	45.5	14.0	0.01784	0.39141	0.18960	0.02164	0.01932	0.01470	0.05646	0.01489	0.07135	0.01934	0.02378
AAA076C03	37.9	7.2	0.02162	0.52640	0.20362	0.02061	0.00654	0.00553	0.02474	0.00895	0.05000	0.03158	0.04211
AAA160J14	31.0	7.6	0.01121	0.48941	0.17061	0.00996	0.00498	0.00623	0.02340	0.01248	0.02340	0.06240	0.09204
AAA076P09	33.3	3.8	0.02249	0.52484	0.16963	0.01218	0.02437	0.00375	0.01564	0.01564	0.02682	0.02793	0.06704
AAA076P13	43.1	7.3	0.02214	0.54686	0.17491	0.01402	0.02509	0.00295	0.01444	0.01274	0.03229	0.03568	0.06797

Genome	%GC	% Non-coding DNA	Multi	Cytoplasmic	Cytoplasmic membrane	Periplasmic	Outer membrane	Extracellular	T	V	K	Q	I
AAA168I18	32.5	4.1	0.01876	0.57256	0.16782	0.01086	0.02073	0.00296	0.01019	0.01246	0.03511	0.02605	0.07475
AAA168P09	33.0	4.5	0.02321	0.54170	0.17549	0.00870	0.01813	0.00580	0.01373	0.01288	0.03519	0.03176	0.06953
AAA160D02	48.8	8.0	0.02137	0.49719	0.16873	0.01350	0.02587	0.02362	0.02304	0.01626	0.04743	0.01084	0.02846
AAA076D02	38.1	5.7	0.01811	0.57159	0.20374	0.01868	0.00170	0.00566	0.01478	0.00887	0.03548	0.03548	0.04258
AAA076D13	38.0	5.8	0.01983	0.54169	0.20875	0.01691	0.00466	0.00583	0.01673	0.01053	0.03532	0.02664	0.04027
AAA076E13	37.3	5.9	0.01927	0.54335	0.18882	0.01445	0.00385	0.00674	0.01505	0.01183	0.03226	0.03333	0.03333
AAA076F14	36.9	5.5	0.01925	0.57644	0.20102	0.01755	0.00340	0.00566	0.01524	0.00879	0.03751	0.03048	0.04279
AAA160P02	31.6	7.7	0.02894	0.44128	0.16946	0.00755	0.03440	0.01552	0.02941	0.01726	0.04987	0.03005	0.04923
MS0242A	36.1	14.9	0.03233	0.47286	0.17898	0.00924	0.03406	0.00808	0.01699	0.01467	0.03629	0.02934	0.05019
MS0243C	35.7	6.1	0.02305	0.45279	0.18736	0.01041	0.02825	0.01115	0.02004	0.01603	0.04910	0.02505	0.03607
MS1901F	39.4	6.4	0.03153	0.42342	0.14790	0.01126	0.03529	0.00901	0.02104	0.01733	0.04455	0.02599	0.04208
MS2205C	29.0	8.0	0.03418	0.47548	0.17088	0.01189	0.02080	0.02377	0.02607	0.00948	0.03791	0.03791	0.03791
AAA160B08	33.1	3.7	0.01816	0.50050	0.21090	0.01312	0.02119	0.01009	0.00653	0.01436	0.03133	0.02350	0.07180
AAA160C11	32.6	3.6	0.02547	0.47642	0.17547	0.01038	0.02170	0.01604	0.00927	0.01457	0.02649	0.02649	0.06358
AAA160I06	32.6	3.3	0.02319	0.48887	0.19017	0.01206	0.01948	0.01206	0.00963	0.01564	0.03490	0.03008	0.06980
AAA164E04	47.5	10.0	0.03104	0.43485	0.14628	0.01291	0.00922	0.00953	0.03014	0.02160	0.03918	0.03918	0.03064
AAA164L15	48.6	8.3	0.03748	0.44477	0.09911	0.01874	0.01824	0.02071	0.01933	0.02285	0.04482	0.02109	0.03427
AAA164M04	48.5	8.5	0.03832	0.43567	0.09981	0.01088	0.01608	0.01798	0.01243	0.01865	0.04707	0.01954	0.02753
AAA164O14	48.5	8.2	0.04355	0.41791	0.09990	0.01417	0.01590	0.02005	0.01786	0.01323	0.04431	0.02315	0.03307
AAA168E21	39.9	8.1	0.03707	0.43220	0.10878	0.01512	0.01317	0.01805	0.02946	0.01733	0.04853	0.02600	0.03553
AAA168F10	47.3	8.1	0.04166	0.43264	0.12391	0.01371	0.02188	0.02267	0.02089	0.01778	0.04622	0.02267	0.02978
AAA536B06	30.8	6.7	0.02397	0.43559	0.18933	0.02277	0.00839	0.00419	0.01806	0.00650	0.03035	0.03468	0.04335
AAA536G10	38.7	6.6	0.01253	0.51015	0.17840	0.01512	0.00950	0.00821	0.02186	0.00661	0.02745	0.04728	0.07219
AAA536K22	31.6	7.1	0.01438	0.50788	0.17904	0.01623	0.00928	0.00928	0.01939	0.00862	0.02800	0.04900	0.06947
AAA536G18	31.2	4.7	0.04130	0.45263	0.16275	0.00648	0.04211	0.02672	0.02005	0.01128	0.04010	0.02381	0.05013
AAA298K06	45.4	16.6	0.01961	0.49285	0.16799	0.01908	0.00530	0.00901	0.02089	0.00633	0.03734	0.03291	0.05063
AAA300J16	30.8	6.9	0.01625	0.36733	0.14711	0.01895	0.00632	0.01986	0.03937	0.00919	0.04068	0.06562	0.04331
AAA298N10	32.9	3.4	0.02423	0.52715	0.17293	0.01170	0.02339	0.00835	0.01515	0.01515	0.03333	0.03636	0.06869
AAA300D14	38.0	5.3	0.02861	0.57582	0.13662	0.01359	0.02575	0.01288	0.01932	0.01208	0.03543	0.04348	0.07246
AAA298D23	31.9	2.9	0.02146	0.51213	0.18470	0.00933	0.02705	0.00840	0.00847	0.01332	0.03390	0.02179	0.06053
AAA300K03	43.0	5.9	0.02500	0.50268	0.13393	0.00625	0.01964	0.01161	0.02153	0.01615	0.04441	0.03499	0.02423
AAA300N18	43.8	5.7	0.03119	0.49104	0.13006	0.00995	0.01725	0.01725	0.02037	0.02342	0.03157	0.02546	0.02953
AAA300O17	43.2	6.4	0.03474	0.47887	0.12113	0.00563	0.01502	0.01596	0.02453	0.01587	0.03608	0.03608	0.03319
AAA015O19	38.5	7.3	0.01964	0.52209	0.16530	0.01691	0.00818	0.00764	0.02366	0.00872	0.03238	0.04421	0.07098
AAA015N04	33.6	3.9	0.01263	0.50105	0.17316	0.02105	0.00684	0.00947	0.02257	0.00976	0.02441	0.05552	0.07810

Table S4. Pairwise percent SSU rRNA gene similarities between SAGs and taxonomically related cultures.

IMG ID	**Culture**	**SAG**					
	Bacteroidetes (80-95%)	**AAA536G18**	**AAA160P02**	**MS0242A**	**MS0243C**	**MS1901F**	**MS2205C**
648028020	*Croceibacter atlanticus* HTCC2559	89	88	87	90	82	84
638341059	*Dokdonia* MED134	90	90	90	90	82	84
640612204	Flavobacteria bacterium BAL38	89	89	90	89	81	85
638341093	Flavobacteria bacterium BBFL7	88	87	85	89	81	82
641380439	Flavobacteriales ALC1	92	92	90	91	83	84
648028027	Flavobacteriales HTCC2170	89	88	89	91	81	84
644736369	*Flavobacterium johnsoniae* UW101	88	89	88	88	81	84
639633025	*Gramella forsetii* KT0803	89	88	89	90	82	85
641380434	*Kordia algicida* OT-1	90	91	91	89	82	84
638341115	*Leeuwenhoekiella blandensis* MED217	90	87	89	92	81	85
640196209	*Microscilla marina* ATCC23134	80	81	80	80	81	81
640963036	*Pedobacter* BAL39	81	82	83	83	81	83
638341152	*Polaribacter irgensii* 23-P	88	94	89	89	81	84
638341218	*Polaribacter* MED152	89	95	89	89	81	84
638341165	*Psychroflexus* ATCC700755	89	89	89	89	81	83
646311950	*Robiginitalea biformata* HTCC2501	88	89	89	91	82	85
640963037	*Ulvibacter* SCB49	91	90	90	90	80	84
	SAR116 (89-97%)	**AAA160J14**	**AAA536K22**	**AAA536B06**	**AAA015N04**	**AAA300J16**	**AAA536G10**
2503113005	SAR116 HIMB100	90	90	95	90	93	89
646564516	*Cand.* 'Puniceispirillum marinum' IMCC1322	91	91	97	91	96	90
	Rosebacter (90-95%)	**AAA015O19**	**AAA076C03**	**AAA298K06**			
637000137	*Jannaschia* CCS1	93	92	91			
638341119	*Loktanella vestfoldensis* SKA53	92	92	92			
638341139	*Oceanicola batsensis*	94	91	94			
638341140	*Oceanicola granulosus*	92	92	92			
647533189	*Octadecabacter* 238	93	92	91			
647533190	*Octadecabacter* 307	93	92	91			
648276686	*Rhodobacterales* HTCC2654	94	92	93			
640612221	*Roseobacter* CCS2	93	92	92			
639633056	*Roseobacter denitrificans* OCh 114	94	91	93			

IMG ID	**Culture**	**SAG**		
647533205	*Roseobacter* GAI101	94	92	92
638341182	*Roseobacter* sp. MED193	95	92	94
640612220	*Roseobacter* sp. SK209-2-6	95	92	94
638341184	*Roseovarius* 217	93	92	94
648276709	*Roseovarius* HTCC2601	95	91	94
640612219	*Sagittula stellata*	93	90	92
637000267	*Silicibacter pomeroyi* DSS-3	94	92	95
638341211	*Sulfitobacter* sp. EE36	94	92	93
638341212	*Sulfitobacter* sp. NAS141	94	92	93
647533238	*Thalassobium* R2A62	94	93	92
	SAR92 (95-96%)	**AAA300D14**	**AAA160D02**	
638341247	Gamma HTCC2207	95	96	
	Actinobacteria (79-81%)	**AAA015D07**	**AAA015M09**	
638341107	*Janibacter* sp. HTCC2649	79	81	
638341246	Marine actinobacterium PHSC20C1	80	80	

Table S5. Metadata for metagenomes used in fragment recruitment analysis, in the order presented in the heatmap (Fig. 3). Library sequencing type designation: S, sanger; P, pyrosequencing (454 Titanium); F, fosmid. Physio-chemical abbreviations: T, temperature; S, salinity; Depth, water column depth; N, nitrate; P, phosphate; Si, silicate; Chl *a*, chlorophyll *a* concentration. Metadata for GOS samples was taken from Yilmaz, et al. (61). NA, not available; NDE, not detectable; ND, not determined.

						Size			**S**	**Depth**	**N (µmol**	**P (µmol**	**Si (µmol**	**Chl *a***
Sample	**Date**	**Latitude**	**Longitude**	**Library**	**No. Seqs**	**(Mbp)**	**%GC**	**T (°C)**	**(PSU)**	**(m)**	**L^{-1})**	**L^{-1})**	**L^{-1})**	**(µg kg^{-1})**
GS123	10/01/05	-32.399	36.592	S	107,966	115.6	36.0	20.40	35.80	1860	0.15	0.22	3.55	0.23
GS122	09/30/05	-30.898	40.420	S	151,654	157.9	41.5	20.20	35.80	4921	1.01	0.15	2.71	0.15
GS121	09/29/05	-29.349	43.216	S	110,720	119.4	34.8	23.10	35.40	4309	0.72	0.17	3.52	0.14
GS120	09/27/05	-26.035	50.123	S	46,052	45.7	34.2	22.50	35.60	5081	0.12	0.20	3.18	0.12
GS119	09/26/09	-23.216	52.306	S	60,987	65.1	33.9	23.80	35.40	2995	0.17	0.17	2.93	0.08
GS149	09/12/05	-6.117	39.117	S	110,984	111.2	37.5	21.27	29.28	5	NA	0.16	3.00	NA
GS148	09/11/05	-6.317	39.009	S	107,741	107.6	38.7	21.27	29.28	1	NA	0.16	3.00	NA
GS117	09/09/05	-4.614	55.509	S	397,561	394.6	40.8	26.40	35.50	4513	0.25	0.19	2.29	0.21
GS116	08/17/05	-4.635	56.836	S	60,932	64.2	36.1	26.20	33.10	2150	0.18	0.13	2.93	0.29
GS115	08/16/05	-4.663	60.523	S	61,020	64.2	35.2	27.90	33.20	3220	0.13	0.27	4.15	0.14
GS114	08/15/05	-4.990	64.977	S	348,823	345.3	34.9	28.20	33.10	3649	0.11	0.23	2.75	0.14
GS113	08/09/05	-7.008	76.331	S	109,700	118.3	35.0	27.50	33.30	4573	0.30	0.16	4.37	0.24
GS112*	08/08/05	-8.505	80.376	S	151,899	157.5	41.3	26.60	32.50	4573	0.20	0.05	2.97	0.13
GS112_454	08/08/05	-8.505	80.376	P	410,687	227.0	37.0	-	-	4573	-	-	-	-
GS111	08/07/05	-9.597	84.198	S	59,080	62.1	34.8	26.40	32.30	3841	0.15	0.08	2.60	0.20
GS110	08/06/05	-10.446	88.303	S	148,885	153.7	41.3	27.00	32.70	1220	0.12	0.11	3.44	0.13
GS109	08/05/05	-10.944	92.059	S	59,813	62.8	34.4	27.20	32.60	4573	0.03	0.13	2.15	0.14
GS108*	08/04/05	-12.093	96.882	S	101,382	104.4	42.2	25.80	32.40	7	0.02	0.21	1.43	0.11
GS108_454	08/04/05	-12.093	96.882	P	529,447	295.6	33.2	-	-	7	-	-	-	-
HF10	10/07/02	22.750	-158.000	F	7,829	13.1	48.6	26.40	35.08	4790	0.01	0.04	1.05	0.08
HOT215	09/24/09	22.750	-158.000	P	943,226	351.4	35.2	26.48	35.48	4790	0.01	0.04	1.11	0.06
GS049	05/17/04	-17.453	-149.799	S	92,501	94.4	34.7	28.80	32.60	900	0.01	0.19	0.80	0.10
GS048*	05/17/04	-17.476	-149.812	S	138,207	143.8	45.7	28.90	35.10	34	0.01	0.19	0.80	0.10
GS051	05/22/04	-15.144	-147.435	S	128,982	140.5	36.5	27.30	34.20	10	0.08	0.24	0.73	NA
GS037	03/17/04	-1.974	-95.015	S	65,670	68.7	37.3	28.00	34.38	3334	5.61	0.56	4.83	0.21
GS028	02/04/04	-1.217	-90.320	S	189,052	205.0	36.1	25.22	34.39	156	3.17	0.52	6.37	0.35
GS027	02/04/04	-1.216	-90.423	S	222,080	237.3	37.3	25.50	34.90	2	3.20	0.52	6.42	0.40
GS031	02/10/04	-0.301	-91.652	S	436,401	461.7	34.4	18.60	29.07	20	0.87	0.12	0.50	0.35

Sample	Date	Latitude	Longitude	Library	No. Seqs	Size (Mbp)	%GC	T (°C)	S (PSU)	Depth (m)	N (µmol L^{-1})	P (µmol L^{-1})	Si (µmol L^{-1})	Chl *a* (µg kg^{-1})
GS034	02/19/04	-0.383	-90.280	S	134,347	142.2	40.2	27.50	34.23	35	1.95	0.48	6.45	0.36
GS029	02/08/04	-0.200	-90.835	S	131,529	143.8	35.6	26.20	34.50	12	1.95	0.48	6.45	0.40
GS036	03/02/04	-0.021	-91.198	S	77,538	85.8	37.4	25.80	34.60	67	2.69	0.64	1.10	0.65
GS030	02/09/04	0.272	-91.633	S	359,152	391.7	35.2	26.90	34.40	19	4.95	0.48	6.91	NA
GS035	03/01/04	1.389	-91.817	S	140,814	151.8	36.4	21.80	34.50	71	2.55	0.50	3.22	0.28
GS026	02/01/04	1.264	-90.295	S	102,708	109.0	34.9	27.80	32.60	2386	0.69	0.28	3.02	0.22
GS025	01/28/04	5.553	-87.088	S	120,671	129.8	45.9	28.30	31.40	30	6.63	0.71	4.42	0.11
GS023	01/21/04	5.640	-86.565	S	133,051	143.6	35.7	28.70	32.60	1139	6.50	0.69	4.50	0.07
GS022	01/20/04	6.493	-82.904	S	121,662	131.1	35.6	29.30	32.30	2431	1.86	0.33	2.17	0.33
GS021	01/19/04	8.129	-79.691	S	131,798	143.5	39.0	27.60	30.70	76	0.01	0.21	2.71	0.50
GS019	01/12/04	10.716	-80.254	S	135,325	146.4	35.5	27.70	35.40	3336	0.00	0.05	2.25	0.23
GS018	01/10/04	18.037	-83.785	S	142,743	156.5	36.1	27.40	35.40	4470	0.45	0.10	2.21	0.14
GS017	01/09/04	20.523	-85.414	S	257,581	42.1	36.0	27.00	35.80	4513	0.31	0.14	1.82	0.13
GS016	01/08/04	24.175	-84.344	S	127,122	137.5	37.0	26.40	35.80	3333	0.59	0.04	1.32	0.16
GS015	01/08/04	24.488	-83.070	S	127,362	138.0	36.1	25.00	36.00	47	0.95	0.04	1.22	0.20
GS014	12/20/03	32.507	-79.264	S	128,885	139.9	36.9	18.60	36.04	31	0.15	0.20	1.14	1.70
GS001a	05/15/03	32.167	-64.500	S	142,352	143.3	50.0	22.90	36.70	4200	0.10	0.05	0.91	0.10
GS001b	05/15/03	32.167	-64.500	S	90,905	91.0	48.1	-	-	4200	-	-	-	-
GS001c	05/15/03	32.167	-64.500	S	92,351	92.7	35.6	-	-	4200	-	-	-	-
GS000b	02/26/03	31.175	-64.324	S	317,180	321.0	36.1	20.50	36.70	4200	0.24	0.06	0.81	0.17
GS000c	02/26/03	32.175	-64.010	S	368,835	371.7	37.3	19.80	36.70	4200	0.38	0.06	0.96	0.17
GS000d	02/26/03	31.175	-64.324	S	332,240	335.9	36.5	20.00	36.60	4200	0.11	0.06	0.79	0.17
MED	10/15/07	38.069	0.232	P	157,230	88.5	39.5	15.90	38.60	200	NDE	NDE	NA	3.48
GS367	01/08/25	-48.249	145.805	P	1,204,979	482.9	41.0	10.9	3.44	3490	NA	NA	NA	0.20
GS368	01/08/26	-44.718	145.755	P	661,063	246.2	37.1	14.2	3.47	3201	NA	NA	NA	1.30
GS369	01/09/24	-77.680	166.009	P	957,060	340.7	42.9	-2.0	3.35	300	NA	NA	NA	4.46
P26_j	06/14/09	50.000	-145.000	F	8,373	7.6	49.4	9.53	32.57	4300	10.75	1.05	17.50	0.75
P26_a	08/27/09	50.000	-145.000	F	5,767	3.4	51.1	12.55	32.46	4300	7.89	0.76	12.50	0.52
P12_j	06/09/09	48.970	-130.667	F	7,031	4.3	40.2	11.23	32.42	3300	6.15	0.83	11.80	0.72
P12_a	08/23/09	48.970	-130.667	F	6,402	3.9	43.0	15.79	32.27	3300	1.20	0.48	11.30	1.02
P12_f	02/06/10	48.970	-130.667	F	685	1.1	47.7	8.41	32.35	3300	6.89	0.87	10.60	NA
P4_j	06/08/09	48.650	-126.667	F	7,238	4.6	43.6	12.30	32.12	1300	0.00	0.37	2.20	0.78

Sample	Date	Latitude	Longitude	Library	No. Seqs	Size (Mbp)	%GC	T (°C)	S (PSU)	Depth (m)	N (µmol L^{-1})	P (µmol L^{-1})	Si (µmol L^{-1})	Chl *a* (µg kg^{-1})
P4_a	08/29/09	48.650	-126.667	F	7,273	4.3	43.8	12.33	32.24	1300	5.19	0.72	18.50	1.72
P4_f	02/04/10	48.650	-126.667	F	530	0.9	46.0	9.81	32.44	1300	6.26	0.79	8.60	NA
GS013	12/19/03	36.004	-75.395	S	138,033	149.0	43.9	9.30	33.95	20	1.06	0.25	1.27	3.00
GS010	11/18/03	38.940	-74.685	S	78,304	82.4	38.6	12.00	31.00	10	1.95	0.48	1.02	2.00
GS009	11/17/03	41.091	-71.602	S	79,303	84.3	38.3	11.00	31.00	32	1.39	0.63	0.83	4.00
GS008	11/16/03	41.486	-71.351	S	129,655	137.7	45.3	9.40	26.50	12	0.34	0.60	0.76	2.20
GS002	08/21/03	42.503	-67.240	S	121,590	128.8	36.6	18.20	29.20	106	0.29	0.21	2.18	1.40
GS003	08/21/03	42.853	-66.217	S	61,605	66.9	37.4	11.70	29.90	119	0.33	0.21	2.21	1.40
GS007	08/25/03	43.632	-66.847	S	50,980	55.4	40.7	17.90	31.70	139	0.35	0.28	2.86	1.40
GS005	08/22/03	44.690	-63.637	S	61,131	66.0	41.0	15.00	30.20	64	0.07	0.12	0.71	6.00
GS006	08/23/03	45.112	-64.947	S	59,679	64.6	35.5	11.20	28.90	11	0.07	0.13	0.81	2.80
GS004	08/22/03	44.137	-63.644	S	52,959	56.9	39.6	13.86	28.30	142	0.05	0.09	0.55	0.40
ECH1_4444077	04/22/08	50.252	-4.209	P	513,568	193.1	36.4	9.7	35.12	50	4.02	0.40	2.60	2.20
ECH2_4444083	08/27/08	50.252	-4.209	P	262,800	97.8	36.8	15.7	33.30	50	0.90	0.06	0.22	8.17
ECH3_4445065	08/26/08	50.252	-4.209	P	426,931	161.5	37.3	15.9	32.10	50	0.08	0.03	0.12	9.24
ECH4_4445066	08/27/08	50.252	-4.209	P	406,423	153.7	36.8	15.8	33.20	50	0.09	0.10	0.15	11.91
ECH5_4445067	04/22/08	50.252	-4.209	P	387,691	143.2	36.8	9.6	35.00	50	3.75	0.32	2.70	1.32
ECH6_4445068	08/26/08	50.252	-4.209	P	499,348	185.4	37.5	15.8	33.30	50	0.90	0.08	0.33	9.80
ECH7_4445069	01/28/08	50.252	-4.209	P	591,615	216.5	38.3	10.1	33.33	50	10.9	0.53	6.01	0.81
ECH8_4445070	01/28/08	50.252	-4.209	P	627,119	234.8	37.7	10.1	34.20	50	10.0	0.52	5.75	0.85
ERS095011	08/21/08	54.184	7.900	P	171,339	98.7	40.2	NA	NA	ND	NA	NA	NA	NA
ERS095012	02/11/09	54.184	7.900	P	308,889	167.7	44.4	4.0	34.26	ND	6.14	0.40	2.76	0.25
ERS095013	03/31/09	54.184	7.900	P	290,819	155.7	43.8	4.0	33.39	ND	9.87	0.19	4.23	0.92
ERS095014	04/07/09	54.184	7.900	P	382,770	199.5	43.4	5.8	32.17	ND	11.01	0.04	0.25	4.63
ERS095015	04/14/09	54.184	7.900	P	897,396	538.6	41.3	6.4	32.90	ND	4.92	0.01	0.17	3.62
ERS095018	06/16/09	54.184	7.900	P	177,797	92.9	44.9	13.2	31.55	ND	1.66	0.03	0.61	3.09
ERS095019	09/01/09	54.184	7.900	P	1,078,370	583.9	40.3	18.0	32.55	ND	NA	0.43	12.61	12.13
GS394	12/08/17	-53.025	73.375	P	758,197	252.8	40.8	2.4	3.39	100	NA	NA	NA	0.60
GS393	12/08/15	-55.265	74.256	P	1,002,776	385.1	38.7	2.0	3.39	2246	NA	NA	NA	0.50
GS392	12/08/13	-64.198	76.457	P	988,765	377.3	40.9	-1.5	3.36	3847	NA	NA	NA	0.04

Sample	Date	Latitude	Longitude	Library	No. Seqs	Size (Mbp)	%GC	T (°C)	S (PSU)	Depth (m)	N (µmol L^{-1})	P (µmol L^{-1})	Si (µmol L^{-1})	Chl *a* (µg kg^{-1})
GS390	10/08/30	-64.831	80.724	P	741,120	261.2	41.1	-1.7	3.43	116	NA	NA	NA	0.32
GS391	12/08/12	-68.396	76.680	P	529,491	199.6	43.0	-1.4	3.42	378	NA	NA	NA	5.00
GS235	01/01/07	-66.270	110.533	P	792,452	278.3	37.4	-0.5	3.39	60	NA	NA	NA	8.60
GS389	10/08/22	-64.803	112.380	P	832,650	322.3	43.1	-1.8	3.47	500	NA	NA	NA	0.14
GS236	01/07/07	-63.891	112.073	P	1,133,502	408.0	38.8	-0.2	3.37	2500	NA	NA	NA	12.10
GS388	10/08/20	-63.818	115.173	P	741,703	304.6	39.3	-1.7	3.40	2500	NA	NA	NA	1.50
GS387	10/08/19	-60.503	120.048	P	717,796	290.8	40.5	-1.5	3.45	3200	NA	NA	NA	0.22
GS386	10/08/17	-54.948	129.620	P	806,943	324.5	42.0	2.0	3.38	3200	NA	NA	NA	0.22
GS353	12/07/30	-67.052	144.669	P	940,823	380.2	40.6	-1.8	3.45	178	NA	NA	NA	0.00
GS355	01/03/08	-66.762	144.334	P	1,116,030	436.7	39.7	-0.9	3.40	891	NA	NA	NA	8.40
GS352	12/07/29	-66.765	143.291	P	1,254,021	499.4	40.6	-0.8	3.40	169	NA	NA	NA	1.00
GS351	12/07/28	-66.559	143.337	P	1,402,873	493.8	43.0	-0.7	3.40	597	NA	NA	NA	1.60
GS349	12/07/27	-66.566	142.317	P	901,998	346.9	42.4	-1.3	3.40	365	NA	NA	NA	3.70
GS348	12/07/24	-66.339	142.988	P	837,796	347.9	40.5	-0.6	3.42	649	NA	NA	NA	12.60
GS359	01/08/12	-66.190	143.492	P	1,327,129	435.7	42.8	0.1	3.41	364	NA	NA	NA	2.40
GS360	01/08/13	-66.582	140.881	P	838,841	307.0	41.3	-0.7	3.41	308	NA	NA	NA	7.50
GS357	01/05/08	-66.172	142.935	P	1,212,316	460.7	40.9	-0.5	3.42	533	NA	NA	NA	2.70
GS347	12/07/23	-66.021	142.666	P	915,367	292.8	41.5	-0.7	3.40	443	NA	NA	NA	3.20
GS362	01/08/19	-65.537	140.723	P	938,200	358.0	38.8	0.8	3.62	1027	NA	NA	NA	0.20
GS358	01/09/08	-64.300	150.006	P	818,549	302.1	41.1	0.0	3.35	3561	NA	NA	NA	0.30
GS363	01/08/22	-60.000	141.234	P	945,021	330.0	41.6	3.5	3.37	4473	NA	NA	NA	0.10
GS346	12/07/20	-59.312	142.463	P	873,249	325.2	44.5	2.9	3.37	3294	NA	NA	NA	0.30
GS364	01/08/23	-56.695	141.869	P	914,798	355.0	42.2	4.2	3.37	3693	NA	NA	NA	0.50
GS366	01/08/24	-52.023	144.066	P	901,102	335.1	37.9	7.7	3.38	3180	NA	NA	NA	0.30
GS346	12/07/20	-59.312	142.463	P	873,249	325.2	44.5	2.9	3.37	3294	NA	NA	NA	0.30

*Sanger sequences obtained from all size fractions (0.1µm, 0.8µm, and 3.0µm) were used in the analysis.

Total	45,138,685	22,890.8	
Average	395,953	200.8	39.6
Range	530- 1,402,873	0.9- 583.9	33.2- 51.1

Table S6. Percentage of encoded amino acids in genomes from marine cultures and SAGs.

Genome	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Alcanivorax_borkumensis_SK2	10.0	1.0	5.8	6.0	3.6	7.8	2.3	4.9	3.8	11.1	2.6	3.3	4.7	4.5	6.2	5.7	5.1	7.2	1.5	2.6
Candidatus_Pelagibacter_sp_HTCC7211	5.6	1.0	5.3	6.0	5.2	6.3	1.6	9.6	10.4	9.3	2.2	6.7	3.3	2.7	3.2	6.9	4.8	5.6	1.0	3.4
Candidatus_Pelagibacter_ubique_HTCC1002	5.7	1.0	5.4	6.1	5.2	6.3	1.6	9.5	10.3	9.4	2.2	6.5	3.2	2.6	3.1	6.9	5.1	5.6	0.9	3.4
Candidatus_Pelagibacter_ubique_HTCC1062	5.7	1.0	5.4	6.0	5.2	6.3	1.6	9.5	10.2	9.4	2.2	6.5	3.2	2.6	3.1	6.9	5.1	5.7	0.9	3.3
Candidatus_Puniceispirillum_marinum_IMCC1322	10.9	1.0	6.6	5.0	4.0	7.8	2.3	6.6	4.2	9.5	3.1	3.6	4.4	3.4	5.4	6.2	5.6	6.8	1.2	2.5
Congregibacter_KT71	10.8	0.9	6.1	6.4	3.7	7.9	2.1	4.9	3.2	10.7	2.4	3.0	4.7	3.7	6.7	6.4	5.1	7.1	1.4	2.7
Croceibacter_atlanticus_HTCC2559	6.5	0.7	5.9	6.6	5.0	6.4	1.7	7.6	7.3	9.3	2.1	6.3	3.4	3.5	3.4	6.6	6.4	6.4	1.0	4.0
Dokdonia_MED134	7.3	0.7	6.1	6.5	4.9	6.7	1.7	7.4	6.7	9.1	2.2	5.4	3.5	3.5	3.7	6.2	6.6	6.6	1.0	4.0
Erythrobacter_litoralis_HTCC2594	12.3	0.8	6.3	6.6	3.7	8.8	1.9	5.1	3.3	9.6	2.5	2.6	5.1	3.2	6.9	5.3	5.2	6.9	1.4	2.2
Flavobacteria_bacterium_BAL38	6.2	0.8	5.2	6.6	5.6	6.2	1.6	8.3	8.1	9.0	2.2	6.7	3.3	3.4	3.0	6.6	6.0	6.3	1.0	4.1
Flavobacteria_bacterium_BBFL7	6.8	0.7	6.3	6.2	4.8	6.5	1.9	7.8	6.6	9.1	2.4	6.0	3.4	3.7	3.6	6.6	6.1	6.3	1.0	4.1
Flavobacteriales_ALC1	6.1	0.7	6.0	6.4	5.2	6.3	1.7	8.2	7.8	9.2	2.1	6.6	3.2	3.2	3.2	6.7	6.0	6.1	1.0	4.2
Flavobacteriales_HTCC2170	6.3	0.7	5.8	6.7	5.1	6.9	1.9	7.7	7.6	9.2	2.4	5.9	3.6	3.2	3.5	6.5	5.6	6.3	1.2	3.9
Flavobacterium_johnsoniae_UW101	6.4	0.8	5.4	6.4	5.3	6.2	1.6	8.0	8.2	9.1	2.1	6.5	3.4	3.4	3.2	6.7	5.9	6.0	1.1	4.2
Fulvimarina_pelagi	11.7	0.8	6.0	6.6	4.0	8.6	1.9	5.5	3.4	9.6	2.5	2.7	4.9	2.9	7.0	5.9	5.5	7.2	1.2	2.2
Gamma_HTCC2080	10.5	1.0	5.9	6.1	3.8	8.1	2.1	5.3	3.3	10.4	2.4	3.4	4.7	4.0	5.7	6.4	5.6	7.2	1.5	2.6
Gamma_HTCC2143	9.2	1.0	5.9	6.0	4.0	7.5	2.1	6.4	4.5	10.1	2.6	3.9	4.2	4.1	5.2	6.8	5.3	7.0	1.3	2.9
Gamma_HTCC2207	9.7	1.0	5.9	6.1	3.9	7.6	2.0	6.2	4.3	10.3	2.6	3.8	4.2	4.3	5.0	6.9	5.2	7.1	1.2	2.8
Gamma_proteobacterium_HTCC2148	10.0	1.1	6.0	6.4	3.8	8.0	2.1	5.4	3.6	10.3	2.6	3.5	4.5	4.1	5.6	6.5	5.1	7.0	1.4	2.9
Gamma_proteobacterium_HTCC5015	9.8	1.0	6.0	6.7	3.7	7.4	2.4	5.2	4.2	10.3	2.5	3.4	4.3	4.6	6.1	6.3	4.9	7.0	1.4	2.8
Gramella_forsetii_KT0803	6.2	0.7	5.8	7.6	5.3	6.5	1.7	7.9	7.7	9.2	2.3	6.0	3.4	3.3	3.7	6.6	5.2	5.9	1.0	3.9
Hyphomonas_neptunium_ATCC_15444	12.7	0.8	5.7	6.2	3.8	8.7	1.8	5.2	3.4	9.8	2.6	2.6	5.3	3.1	6.7	5.6	5.4	6.9	1.4	2.3
Janibacter_sp_HTCC2649	36.7	8.4	25.4	23.5	16.8	31.0	14.9	20.1	15.2	32.0	13.8	14.2	23.7	16.8	26.8	24.3	26.0	30.8	12.7	13.5
Jannaschia_CCS1	12.6	0.9	6.4	5.6	3.8	8.8	2.1	5.2	2.6	9.9	2.9	2.5	5.3	3.2	6.5	5.1	5.9	7.3	1.4	2.2
Kordia_algicida_OT-1	6.3	0.8	5.6	6.7	5.2	5.9	1.8	8.0	8.1	8.9	2.1	6.3	3.2	3.6	3.4	6.3	6.4	6.0	1.0	4.2
Leeuwenhoekiella_blandensis_MED217	7.4	0.7	5.7	7.0	5.1	6.5	1.8	7.0	6.9	9.7	2.1	5.5	3.6	3.9	3.8	6.2	5.9	6.2	1.1	4.1
Lentisphaera_araneosa_HTCC2155	6.9	1.2	5.8	6.7	4.6	6.3	2.3	6.6	8.0	9.8	2.5	4.9	3.9	3.8	4.3	7.0	4.8	5.5	1.3	3.6
Loktanella_vestfoldensis_SKA53	13.0	0.9	6.5	4.6	3.7	8.5	2.1	5.6	3.0	10.0	2.9	2.6	4.9	3.6	6.4	4.8	5.8	7.4	1.3	2.2
Marine_actinobacterium_PHSC20C1	36.1	7.1	24.5	24.3	18.5	29.4	13.8	23.5	16.2	32.3	13.8	16.5	22.1	17.1	25.2	26.2	25.4	29.9	11.9	14.5
Marinomonas_MED121	8.5	1.0	5.6	6.3	4.3	6.5	2.2	6.7	5.8	10.9	2.6	4.4	3.7	4.6	4.1	7.1	5.1	6.4	1.2	3.1
Methylophaga_DSM010	9.1	0.9	6.2	6.3	3.8	6.9	2.3	6.2	4.9	10.3	2.7	4.0	4.0	4.8	5.0	6.1	5.5	6.8	1.3	2.8
Methylophilales_bacterium_HTCC2181	7.1	0.9	5.6	6.2	4.6	6.7	2.1	8.6	7.5	9.8	2.6	5.3	3.7	3.4	3.9	7.0	5.0	6.0	1.0	3.0
Microscilla_marina_ATCC23134	7.2	0.8	4.9	5.8	4.7	6.3	2.2	6.5	7.8	9.8	2.2	5.5	3.7	5.0	4.1	5.9	6.0	6.4	1.2	4.1
Moritella_PE36	8.7	1.1	5.6	5.6	4.2	6.5	2.1	7.1	5.3	10.5	2.6	4.8	3.7	4.5	4.1	6.8	5.8	6.8	1.1	3.1
Nitrosococcus_oceani_ATCC_19707	9.5	1.0	4.7	6.6	3.9	7.6	2.4	5.8	4.2	11.2	2.2	3.2	5.0	4.5	6.6	5.7	5.0	6.5	1.5	2.9
Oceanibulbus_indolifex	12.2	0.9	6.1	6.0	3.7	8.5	2.1	5.2	3.3	10.1	2.9	2.7	5.0	3.4	6.5	5.2	5.5	7.2	1.3	2.2
Oceanicola_batsensis	12.3	0.8	6.3	6.3	3.6	9.0	2.0	5.0	2.7	9.9	2.7	2.4	5.2	3.0	7.3	5.0	5.5	7.3	1.3	2.2
Oceanicola_granulosus	13.6	0.8	6.2	6.3	3.5	9.2	2.0	4.5	2.2	10.4	2.4	2.2	5.3	2.7	7.5	4.7	5.3	7.5	1.4	2.1
Oceanospirillum_sp_MED92	8.8	1.1	5.7	7.0	4.0	7.0	2.2	6.3	5.1	10.7	2.6	3.9	4.0	4.5	4.9	6.6	5.0	6.8	1.2	2.8
Octadecabacter_238	11.4	1.1	6.1	5.4	3.8	8.0	2.3	5.6	4.1	9.5	2.9	3.0	4.7	3.4	6.5	5.4	5.8	7.0	1.5	2.3
Octadecabacter_307	11.3	1.0	6.2	5.2	3.9	8.3	2.2	5.8	3.7	9.6	3.0	3.1	4.7	3.4	6.3	5.5	5.9	7.2	1.4	2.3

Genome	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pedobacter_BAL39	7.7	0.7	5.4	5.8	4.9	6.9	1.8	6.9	6.7	9.6	2.4	5.4	3.7	3.8	4.2	6.6	5.7	6.4	1.1	4.1
Photobacterium_angustum_S14	8.3	1.1	5.6	5.8	4.2	6.6	2.3	6.9	5.6	10.2	2.6	4.7	3.9	4.5	4.2	6.6	5.6	6.8	1.2	3.2
Photobacterium_profundum_SS9	8.4	1.2	5.5	6.0	4.1	6.7	2.3	6.7	5.5	10.3	2.7	4.5	3.8	4.5	4.5	6.7	5.6	6.8	1.3	3.1
Planctomyces_maris_DSM8797	8.1	1.1	5.6	6.5	4.0	7.1	2.2	5.8	4.8	10.1	2.2	3.8	5.0	4.8	5.5	6.7	5.7	6.6	1.5	2.8
Polaribacter_irgensii_23-P	6.9	0.7	5.0	6.5	5.5	6.3	1.8	8.3	8.4	9.4	2.2	6.0	3.1	3.3	3.4	6.4	5.8	6.2	1.0	3.8
Polaribacter_MED152	6.3	0.7	5.6	6.6	5.4	6.1	1.6	8.2	8.4	9.2	2.0	6.8	3.1	3.3	3.2	6.5	5.9	6.2	1.0	4.0
Polaromonas_JS666	12.3	1.0	5.0	5.2	3.6	8.2	2.3	4.6	3.8	10.7	2.6	2.7	5.2	4.0	6.5	5.7	5.3	7.6	1.4	2.3
Prochlorococcus_marinus_AS9601	5.4	1.1	5.2	6.7	4.9	6.3	1.5	9.1	8.6	10.7	1.8	6.5	3.6	3.0	3.9	7.7	4.5	5.2	1.3	2.9
Prochlorococcus_marinus_CCMP1375	6.9	1.2	4.9	6.4	4.1	6.9	1.8	7.8	6.6	11.5	2.0	5.0	4.3	3.7	4.9	7.7	4.7	5.7	1.5	2.4
Prochlorococcus_marinus_MED4	5.4	1.2	5.1	6.6	4.9	6.3	1.5	9.2	8.6	10.8	1.8	6.6	3.6	3.0	3.9	7.7	4.5	5.3	1.3	2.9
Prochlorococcus_marinus_MIT_9215	5.4	1.2	5.1	6.7	5.0	6.2	1.5	9.1	8.7	10.8	1.8	6.5	3.6	3.0	3.9	7.7	4.4	5.2	1.3	2.9
Prochlorococcus_marinus_MIT_9301	5.4	1.2	5.1	6.7	5.0	6.3	1.5	9.0	8.6	10.8	1.9	6.5	3.6	3.0	3.9	7.7	4.4	5.3	1.3	2.8
Prochlorococcus_marinus_MIT_9303	9.4	1.3	5.1	5.8	3.3	7.7	2.2	5.1	3.7	12.3	2.2	3.5	5.1	4.8	6.3	7.0	4.8	6.6	1.7	2.1
Prochlorococcus_marinus_MIT_9312	5.4	1.2	5.1	6.7	5.0	6.3	1.5	9.1	8.5	10.7	1.8	6.5	3.6	3.0	3.9	7.7	4.4	5.2	1.3	2.9
Prochlorococcus_marinus_MIT_9313	9.6	1.3	5.1	5.8	3.3	7.8	2.2	4.9	3.5	12.5	2.1	3.2	5.2	4.9	6.5	6.8	4.7	6.8	1.7	2.0
Prochlorococcus_marinus_MIT_9515	5.3	1.2	5.0	6.6	5.0	6.2	1.5	9.3	8.7	10.8	1.9	6.6	3.6	3.0	3.9	7.7	4.4	5.2	1.3	2.9
Prochlorococcus_marinus_NATL1A	6.4	1.1	5.2	6.5	4.3	6.8	1.7	8.1	7.2	11.1	2.0	5.3	4.1	3.4	4.6	7.8	4.7	5.6	1.4	2.6
Prochlorococcus_marinus_NATL2A	6.5	1.1	5.3	6.5	4.3	6.9	1.7	8.0	7.1	11.1	2.0	5.3	4.1	3.5	4.5	7.8	4.7	5.6	1.4	2.5
Pseudoalteromonas_haloplanktis_TAC125	9.0	1.0	5.4	5.7	4.3	6.4	2.2	6.8	5.8	10.5	2.4	5.0	3.7	4.8	4.0	6.6	5.6	6.6	1.1	3.2
Psychroflexus_ATCC700755	6.0	0.8	5.8	6.7	5.3	6.3	1.8	8.0	7.9	9.5	2.2	5.9	3.4	3.4	3.7	7.1	5.4	5.9	1.1	3.7
Psychromonas_CNPT3	8.1	1.2	5.4	5.6	4.4	6.2	2.2	7.6	6.5	11.0	2.6	4.7	3.4	4.6	4.0	6.9	5.2	6.3	1.0	3.1
Psychromonas_ingrahamii_37	8.3	1.1	5.4	5.9	4.4	6.5	2.1	7.4	6.2	10.7	2.4	4.8	3.6	4.4	4.0	6.7	5.4	6.4	1.1	3.1
Reinekea_MED297	9.2	0.9	6.2	6.1	4.0	7.0	2.3	5.6	3.7	10.8	2.5	3.7	4.3	4.8	5.6	6.4	5.6	7.1	1.5	2.8
Rhodobacterales_HTCC2654	12.1	0.8	6.5	6.1	3.8	8.9	2.0	5.1	3.2	9.5	2.8	2.6	5.1	2.9	6.7	4.9	5.7	7.5	1.4	2.2
Rhodobacterales_Y4I	12.9	1.0	5.6	6.2	3.7	8.8	2.1	4.8	3.4	10.2	2.7	2.6	5.1	3.6	6.6	5.2	5.1	6.9	1.4	2.3
Rhodopirellula_baltica_SH1	9.3	1.3	6.2	6.0	3.7	7.5	2.3	4.9	3.4	9.3	2.4	3.4	5.3	4.0	6.9	7.4	5.9	7.0	1.5	2.1
Rhodospirillales_BAL199	12.5	0.9	6.2	5.4	3.5	8.7	2.2	4.8	2.8	10.0	2.4	2.3	5.3	2.8	7.9	5.1	5.5	8.0	1.5	2.1
Robiginitalea_biformata_HTCC2501	8.2	0.8	5.9	6.9	4.6	7.9	1.9	6.0	4.7	9.9	2.3	4.2	4.5	3.5	5.9	5.9	5.3	6.5	1.2	3.7
Roseobacter_CCS2	11.9	0.9	6.6	5.4	3.9	8.4	2.0	5.7	3.4	9.6	3.0	3.0	4.8	3.5	5.9	5.1	5.9	7.4	1.4	2.3
Roseobacter_denitrificans_OCh_114	12.0	1.0	6.2	5.6	3.9	8.4	2.1	5.4	3.3	9.8	2.8	2.8	4.9	3.5	6.3	5.4	5.7	7.3	1.4	2.2
Roseobacter_GAI101	11.9	0.9	6.3	5.4	3.8	8.5	2.1	5.5	3.6	9.9	3.0	2.8	4.9	3.4	6.1	5.4	5.7	7.2	1.3	2.2
Roseobacter_sp_MED193	11.7	1.0	5.8	6.1	3.8	8.4	2.1	5.2	3.5	10.3	2.8	2.8	4.9	3.8	6.2	5.7	5.3	6.9	1.4	2.3
Roseobacter_sp_SK209-2-6	11.5	1.0	5.5	6.6	3.9	8.5	2.1	5.2	3.7	10.4	2.8	2.8	4.9	3.8	6.2	5.8	5.1	6.7	1.4	2.3
Roseovarius_217	12.2	0.9	6.0	5.8	3.7	8.7	2.1	5.2	2.9	10.2	2.8	2.6	5.0	3.2	6.9	5.1	5.6	7.3	1.4	2.2
Roseovarius_HTCC2601	12.6	0.9	5.9	6.4	3.6	8.9	2.0	4.9	2.8	10.3	2.8	2.3	5.3	3.1	7.0	5.3	5.4	7.2	1.4	2.1
Saccharophagus_degradans_2-40	9.5	1.0	5.6	6.1	4.0	7.1	2.1	5.9	4.9	9.8	2.2	4.7	4.1	4.1	4.5	6.9	5.7	7.1	1.4	3.2
Sagittula_stellata	12.4	0.9	6.3	5.9	3.7	8.9	2.1	4.8	3.0	10.0	2.9	2.5	5.2	3.0	6.9	5.1	5.7	7.5	1.4	2.2
SAR116_HIMB100	11.0	1.1	6.1	5.4	4.0	8.0	2.3	5.8	3.9	10.2	2.7	3.2	4.5	4.1	5.5	6.3	5.6	6.8	1.2	2.4
SAR86_cluster_bacterium_SAR86C	6.3	1.0	6.0	6.5	4.7	6.8	1.7	8.7	7.2	9.7	2.5	5.8	3.6	2.9	3.7	7.9	4.8	6.1	1.0	3.2
SAR86_cluster_bacterium_SAR86D	6.3	1.0	5.9	6.5	4.9	6.4	1.8	9.0	7.8	9.6	2.5	6.1	3.4	2.9	3.5	7.6	4.7	5.8	0.9	3.3
SAR86_cluster_bacterium_SAR86E	7.0	0.9	5.8	6.6	4.9	6.9	1.9	8.3	6.8	9.9	2.4	5.4	3.8	3.4	3.7	7.7	4.7	5.8	1.0	3.0
Shewanella_baltica_OS155	9.6	1.0	5.4	5.7	4.0	6.8	2.3	6.0	5.1	10.9	2.6	4.0	4.0	4.9	4.7	6.6	5.4	6.8	1.3	3.0
Shewanella_baltica_OS185	9.6	1.1	5.5	5.7	4.0	6.8	2.3	6.1	5.0	10.8	2.6	4.1	4.0	4.9	4.5	6.6	5.5	6.8	1.3	3.0
Shewanella_baltica_OS195	9.5	1.1	5.5	5.7	4.0	6.8	2.3	6.1	5.1	10.8	2.6	4.1	4.0	4.9	4.5	6.7	5.5	6.7	1.3	3.0
Shewanella_denitrificans_OS217	9.3	1.0	5.5	5.6	4.0	6.8	2.3	6.2	5.2	10.9	2.5	4.3	3.9	5.0	4.3	7.0	5.3	6.6	1.2	3.0

Genome	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Shewanella_frigidimarina_NCIMB_400	8.9	1.0	5.7	5.5	4.1	6.6	2.3	6.7	5.1	10.5	2.7	4.5	3.8	5.0	4.2	6.7	5.6	6.9	1.2	3.0
Shewanella_KT99	8.8	1.1	5.6	6.0	3.9	6.9	2.3	6.6	5.4	10.7	2.7	4.0	3.8	4.5	4.7	6.9	5.3	6.6	1.2	3.0
Silicibacter_pomeroyi_DSS-3	12.5	1.0	5.9	5.7	3.7	8.9	2.1	5.0	2.8	10.3	2.8	2.5	5.2	3.3	7.0	5.0	5.3	7.2	1.4	2.2
Sphingomonas_SKA58	13.2	0.8	6.3	5.1	3.4	8.8	2.1	5.2	3.0	9.9	2.7	2.5	5.3	3.4	7.4	5.3	5.1	6.9	1.4	2.2
Sphingopyxis_alaskensis_RB2256	14.0	0.8	6.2	5.3	3.5	8.9	2.0	5.1	2.9	9.8	2.5	2.4	5.4	2.9	7.7	5.0	5.0	7.0	1.5	2.1
Sulfitobacter_sp_EE36	12.2	0.9	6.3	5.5	3.7	8.6	2.0	5.3	3.5	9.8	2.9	2.8	4.9	3.5	6.2	5.2	5.7	7.3	1.3	2.3
Sulfitobacter_sp_NAS141	12.1	0.9	6.3	5.6	3.7	8.5	2.1	5.3	3.5	9.8	2.9	2.7	4.9	3.4	6.3	5.3	5.7	7.3	1.4	2.3
Synechococcus_CC9311	9.6	1.3	5.3	5.8	3.3	8.0	2.1	4.9	3.2	12.3	2.2	3.1	5.2	4.7	6.7	7.0	4.9	6.9	1.7	1.9
Synechococcus_CC9605	10.3	1.3	5.5	5.9	3.2	8.1	2.1	4.3	3.0	12.1	2.2	2.8	5.5	4.8	7.1	6.4	4.8	7.0	1.8	1.9
Synechococcus_CC9902	9.8	1.3	5.5	5.7	3.2	8.1	2.2	4.7	3.1	12.2	2.1	3.0	5.3	4.7	6.9	6.6	4.9	7.0	1.7	1.9
Synechococcus_elongatus_PCC_6301	10.5	1.1	5.1	5.7	3.5	7.2	1.7	5.4	2.6	12.3	1.6	2.8	5.5	6.1	6.8	6.0	5.2	6.6	1.7	2.5
Synechococcus_elongatus_PCC_7942	10.5	1.1	5.1	5.7	3.5	7.2	1.7	5.4	2.6	12.3	1.6	2.8	5.5	6.1	6.8	6.0	5.2	6.6	1.8	2.5
Synechococcus_RCC307	11.0	1.2	4.9	5.8	3.1	8.5	2.0	4.0	2.6	12.8	2.0	2.6	5.7	5.5	6.9	6.4	4.3	6.9	1.8	1.8
Synechococcus_sp_WH8102	10.2	1.2	5.7	5.7	3.1	8.2	2.1	4.3	2.8	12.3	2.1	2.8	5.5	4.8	7.2	6.5	5.0	7.0	1.7	1.9
Synechococcus_WH_7803	10.6	1.2	5.4	5.6	3.1	8.3	2.1	4.2	2.5	12.6	2.0	2.6	5.6	4.7	7.3	6.5	4.9	7.1	1.8	1.8
Thalassobium_R2A62	11.4	0.9	6.5	5.6	3.9	8.4	2.1	5.6	3.5	9.5	3.0	3.1	4.6	3.4	6.1	5.5	5.9	7.3	1.4	2.3
Ulvibacter_SCB49	6.8	0.8	5.8	6.7	5.1	6.4	1.7	7.8	7.6	9.2	2.2	6.0	3.3	3.3	3.3	6.4	6.3	6.4	1.0	4.0
Vibrio_harveyi_ATCC_BAA-1116	8.4	1.1	5.5	6.6	4.1	6.6	2.3	6.1	5.7	10.1	2.7	4.2	3.9	4.5	4.7	6.6	5.4	7.0	1.3	3.2
AAA076C03	8.3	1.0	5.6	5.9	4.3	7.5	1.9	8.0	6.1	9.7	2.8	4.9	4.1	3.0	4.6	6.8	5.1	6.3	1.2	2.7
AAA160J14	6.0	1.1	5.3	5.8	4.8	6.5	1.7	9.4	8.5	9.7	2.4	6.5	3.6	2.8	3.5	7.3	4.9	5.9	1.1	3.1
AAA076P09	6.3	0.9	6.1	6.5	4.9	6.6	1.8	8.7	7.3	9.5	2.4	6.0	3.5	2.8	3.5	8.0	4.8	5.9	1.0	3.4
AAA076P13	6.4	0.9	6.1	6.5	4.9	6.6	1.8	8.7	7.3	9.5	2.4	6.0	3.5	2.9	3.5	8.0	4.9	6.0	1.0	3.3
AAA168I18	6.2	1.0	6.0	6.6	4.9	6.5	1.8	9.0	7.6	9.4	2.3	6.2	3.5	2.9	3.5	7.8	4.7	5.8	1.0	3.4
AAA168P09	6.2	0.9	6.0	6.6	4.9	6.5	1.8	9.1	7.6	9.5	2.3	6.2	3.5	2.9	3.5	7.8	4.8	5.8	1.0	3.4
AAA160D02	8.7	1.0	6.3	6.3	3.9	7.2	1.9	6.7	5.1	9.9	2.3	4.5	3.9	4.6	4.3	7.3	5.3	7.0	1.1	2.8
AAA076D02	7.0	1.1	5.7	6.3	4.6	6.9	2.0	8.1	6.6	9.7	2.7	5.3	3.6	3.3	3.8	7.8	4.8	6.5	1.1	3.1
AAA076D13	7.1	1.1	5.7	6.3	4.5	7.0	2.0	8.1	6.6	9.7	2.7	5.2	3.6	3.3	3.7	7.7	4.9	6.6	1.1	3.1
AAA076E13	6.8	1.1	5.8	6.4	4.6	6.8	2.0	8.2	6.8	9.7	2.6	5.4	3.5	3.3	3.7	7.8	4.8	6.3	1.1	3.2
AAA076F14	7.1	1.1	5.7	6.4	4.5	6.9	2.0	8.1	6.7	9.7	2.7	5.2	3.6	3.3	3.7	7.7	4.8	6.5	1.1	3.1
AAA160P02	6.1	0.7	5.4	6.4	5.4	6.4	1.8	8.5	8.2	9.2	2.1	6.2	3.2	3.3	3.3	6.9	5.9	6.0	1.0	3.9
MS0242A	6.5	0.7	5.3	6.5	5.3	6.8	1.9	7.8	7.1	9.8	2.2	5.6	3.8	3.7	3.6	7.0	5.6	6.0	1.2	3.7
MS0243C	8.0	0.7	5.2	6.0	5.1	7.0	1.9	7.2	7.0	9.8	2.3	5.2	3.8	3.7	3.4	6.6	5.7	6.5	1.1	3.7
MS1901F	6.3	0.9	6.4	6.2	4.6	6.8	1.8	8.3	6.6	9.0	2.6	5.7	3.4	3.2	3.8	7.0	5.6	6.3	1.1	4.3
MS2205C	6.7	0.9	5.8	6.6	5.2	7.0	1.8	7.7	6.3	9.3	2.5	5.4	3.5	3.3	3.8	7.4	5.4	6.5	1.1	3.7
AAA160B08	5.8	0.9	6.1	6.1	5.1	6.5	1.8	8.9	7.6	9.3	2.5	6.0	3.3	3.0	3.4	7.7	5.1	6.0	1.0	3.8
AAA160C11	6.0	0.8	6.3	6.2	5.0	6.7	1.7	8.6	7.7	9.0	2.4	6.0	3.3	3.0	3.5	7.7	5.3	6.1	1.1	3.7
AAA160I06	5.8	0.8	6.2	6.1	5.2	6.5	1.8	8.9	7.8	9.2	2.5	6.0	3.3	3.0	3.4	7.7	5.0	6.0	1.0	3.7
AAA164E04	7.8	1.1	5.7	6.2	4.4	7.7	2.5	5.9	5.1	9.9	2.6	4.0	4.8	4.0	5.5	6.9	5.2	6.3	1.7	2.7
AAA164L15	8.2	0.9	5.9	6.7	4.4	8.1	2.1	5.8	5.7	9.7	2.1	4.1	4.8	3.3	5.6	6.8	5.4	6.2	1.6	2.7
AAA164M04	8.2	0.9	5.8	6.8	4.4	7.9	2.0	6.0	5.8	9.8	2.1	4.0	4.7	3.3	5.5	6.9	5.4	6.3	1.5	2.6
AAA164O14	8.2	0.9	5.9	6.5	4.4	8.1	2.1	5.9	5.8	9.6	2.1	4.2	4.9	3.3	5.4	6.8	5.6	6.2	1.5	2.7
AAA168E21	8.3	0.9	5.9	6.7	4.4	8.1	2.1	5.9	5.7	9.7	2.0	4.1	4.8	3.3	5.5	6.8	5.5	6.2	1.5	2.6
AAA168F10	7.9	0.8	5.9	6.7	4.5	7.9	2.1	6.1	5.5	9.7	2.1	4.3	4.6	3.5	5.2	7.0	5.6	6.3	1.5	2.9
AAA536B06	10.3	1.1	6.0	5.1	4.2	7.9	2.2	6.8	5.0	9.8	2.9	3.9	4.3	3.3	5.2	6.2	5.1	6.7	1.2	2.6
AAA536G10	5.9	1.0	5.2	5.9	5.0	6.6	1.7	9.4	8.6	9.8	2.3	6.6	3.5	2.8	3.5	7.3	4.8	5.7	1.0	3.1

Genome	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
AAA536K22	6.0	1.0	5.3	5.9	5.0	6.6	1.7	9.4	8.5	9.8	2.4	6.6	3.5	2.8	3.5	7.2	4.9	5.8	1.1	3.1
AAA536G18	4.9	0.8	6.0	6.4	5.5	6.3	1.5	9.3	8.1	8.9	2.1	7.6	3.1	2.7	3.0	7.9	5.0	5.6	1.0	4.3
AAA298K06	8.3	1.1	5.6	6.3	4.5	7.5	1.9	7.3	5.9	9.8	2.6	4.6	4.1	3.1	4.7	7.0	5.1	6.7	1.3	2.7
AAA300J16	9.5	1.1	5.9	5.8	4.4	7.6	2.0	6.8	5.1	9.6	2.7	4.2	4.0	3.2	5.2	6.6	5.5	6.9	1.1	2.7
AAA298N10	5.8	0.9	6.2	6.8	5.1	6.5	1.7	9.1	7.9	9.2	2.3	6.4	3.4	2.7	3.5	7.9	4.7	5.9	1.0	3.5
AAA300D14	6.7	1.1	6.0	6.4	4.6	6.8	1.9	8.0	6.5	9.7	2.4	5.3	3.6	3.2	4.4	8.0	5.1	6.3	1.1	3.1
AAA298D23	5.5	0.8	6.0	6.4	5.5	6.3	1.7	8.9	8.0	9.4	2.3	6.5	3.2	3.0	3.4	7.8	4.7	6.0	1.0	3.6
AAA300K03	6.5	1.1	5.5	6.8	5.0	7.1	2.1	6.8	6.6	10.2	2.1	4.8	4.3	3.5	4.8	7.6	4.8	6.0	1.3	3.0
AAA300N18	6.7	1.1	5.6	6.8	4.9	7.4	2.1	6.4	6.4	10.0	2.1	4.7	4.4	3.6	4.9	7.5	4.9	6.1	1.4	3.0
AAA300O17	6.5	1.1	5.6	6.9	4.9	7.2	2.0	6.6	6.5	10.0	2.1	4.9	4.3	3.5	4.7	7.9	5.0	6.0	1.3	3.0
AAA015O19	8.1	1.0	5.6	6.5	4.6	7.3	1.8	7.5	6.5	9.8	2.5	4.7	4.0	3.2	4.5	7.2	5.1	6.4	1.2	2.7
AAA015N04	5.9	1.0	5.2	5.9	5.1	6.5	1.7	9.5	8.9	9.8	2.3	6.7	3.5	2.7	3.5	7.2	4.8	5.7	1.1	3.1
AAA015D07	5.2	0.7	5.8	7.3	5.7	6.1	1.4	9.3	7.4	9.9	1.9	6.5	3.2	2.7	3.3	7.6	5.1	6.3	1.0	3.7
AAA015M09	5.2	0.7	5.9	7.2	5.5	6.3	1.4	9.2	7.4	9.8	1.9	6.4	3.3	2.8	3.4	7.6	5.2	6.2	1.0	3.7

Table S7. Properties of amino acids which are overrepresented and underrepresented in SAGs, as compared to marine cultures. Abbreviations: % GC, fraction of GC in the first two codon positions; ATP cost, the metabolic cost for aerobic synthesis by *E. coli* (62); pI, isoelectric point; pK, dissociation constants of respective groups.

Amino acid	**Short**	**%GC**	**ATP cost**	**Atoms C**	**Atoms N**	**pI**	**pK1 (α-COOH)**	**pK2 (α-$^{+}$NH$_3$)**	**Polar**	**Aromatic or aliphatic**	**pH**	**Hydrophobic**
Enriched in SAGs												
Tyrosine	Y	0	-8	9	1	5.64	2.2	9.21	X	Aromatic	weak acidic	-
Phenylalanine	F	0	-6	9	1	5.49	2.2	9.31	-	Aromatic	-	X
Isoleucine	I	0	7	6	1	6.05	2.32	9.76	-	Aliphatic	-	X
Glutamic acid	E	50	-7	5	1	3.15	2.1	9.47	X	-	acidic	-
Asparagine	N	0	3	4	2	5.41	2.14	8.72	X	-	weak basic	-
Lysine	K	0	5	6	2	9.6	2.16	9.06	X	-	basic	-
Serine	S	50	-2	3	1	5.68	2.19	9.21	X	-	weak acidic	-
Average		**14**	**-1.14**	**6.00**	**1.29**	**5.86**	**2.19**	**9.25**				
Depleted in SAGs												
Valine	V	50	-2	5	1	6	2.39	9.74	-	Aliphatic	-	X
Glycine	G	100	-2	2	1	6.06	2.35	9.78	-	-	-	X
Alanine	A	100	-1	3	1	6.01	2.35	9.87	-	-	-	X
Arginine	R	100	5	6	4	10.76	1.82	8.99	X	-	strongly basic	-
Proline	P	100	-2	5	1	6.3	1.95	10.64	-	-	-	X
Histidine	H	50	1	6	3	7.6	1.8	9.33	X	Aromatic	weak basic	-
Tryptophan	W	50	-7	11	2	5.89	2.46	9.41	-	Aromatic	weak basic	-
Average		**79%**	**-1.14**	**5.43**	**1.86**	**6.95**	**2.16**	**9.68**				
***P*-value (t-test)**		**0.0005**	**1.0000**	**0.6893**	**0.2707**	**0.2925**	**0.8137**	**0.0885**				

Table S8. The most common glycoside hydrolase families detected in Verrucomicrobia, Bacteroidetes and Gammaproteobacteria SAR92 SAGs.

CAZy family	**Putatuve enzyme/s**	**Putative substrates**
GH109	α-N-acetylgalactosaminidase (EC 3.2.1.49)	*N*-acetylgalactosamine (glycoproteins from cellular surface and cell wall)
GH33	sialidase or neuraminidase (EC 3.2.1.18)	glycosidic linkages of terminal sialic residues in oligosaccharides, glycoproteins and glycolipids,
GH3	(2) β-N-acetylhexosaminidase (EC 3.2.1.52) (2) β-glucosidase (EC 3.2.1.21) (1) glucan 1,4-β-glucosidase (EC 3.2.1.74)	Hexosamines beta-D-glucosides (β-D-galactosides, α-L-arabinosides, β-D-xylosides, β-D-fucosides) 1,4-β-D-glucans and related oligosaccharides
GH13	(1) cyclomaltodextrinase (EC 3.2.1.54) (7) α-amylase (EC 3.2.1.1)	cyclomaltodextrin starch, glycogen and related polysaccharides and oligosaccharides
GH81	endo-β-1,3-glucanase (EC 3.2.1.39)	laminarin
GH2	(2) β-glucuronidase (EC 3.2.1.31)	β-D-glucuronic acid (glycosaminoglycans/ mucopolysaccharides)
GH43	xylosidase/arabinosidase (EC 3.2.1.37, EC 3.2.1.55)	xylan , arabinans
GH5	Cellulase family A (including endo-1,4-β-xylanase EC 3.2.1.8)	Cellulose and hemicellulose
GH9	Cellobiohydrolase (EC 3.2.1.91)	cellulose and cellotetraose
GH10	endo-1,4-β-xylanase (EC 3.2.1.8)	xylan
GH127 /121	arabinofuranosidase (EC 3.2.1.55)	L-arabinofuranosides (pectins, hemicelluloses and others)
GH78	a-L-rhamnosidase (EC 3.2.1.40)	α-L-rhamnoside

Table S9. Presence (X) and absence of central carbon metabolism pathway genes identified in Gammaproteobacteria SAGs.

	SAR86					ARCTIC				SAR92	
Enzyme/Pathway	AAA076P09	AAA076P13	AAA168I18	AAA168P09	AAA298N10	AAA076D02	AAA076D13	AAA076E13	AAA076F14	AAA160D02	AAA300D14
Central Carbon Metabolism											
Glycolysis											
glucokinase (2.7.1.2)							X	X	X	X	X
glucose-6-phosphate isomerase / aldose 1-epimerase (5.3.1.9 / 5.1.3.3)		X	X	X		X	X	X	X	X	X
6-phosphofructokinase (2.7.1.11)	X	X	X	X	X	X			X		
fructose 1,6-bisphosphate aldolase (4.1.2.13)	X	X	X	X	X	X	X	X	X		
triosephosphate isomerase (5.3.1.1)	X	X	X	X	X	X	X		X	X	X
glyceraldehyde-3-phosphate dehydrogenase (1.2.1.12)	X	X	X	X	X	X	X	X	X	X	X
phosphoglycerate kinase (2.7.2.3)		X	X	X	X	X	X	X	X		X
phosphoglycerate mutase (5.4.2.1)						X	X		X		X
enolase (4.2.1.11)	X	X	X	X	X	X	X	X	X	X	X
pyruvate kinase (2.7.1.40)	X	X		X	X	X	X	X	X	X	X
Pentose Phosphate Pathway											
glucose-6-phosphate dehydrogenase (1.1.1.49)	X	X					X	X	X	X	X
6-phosphogluconolactonase (3.1.1.31)	X	X					X	X	X	X	X
6-phosphogluconate dehydrogenase (1.1.1.44)							X	X	X		
ribulose-phosphate 3-epimerase (5.1.3.1)	X	X		X	X	X	X		X		X
ribose-5-phosphate isomerase (5.3.1.6)	X	X	X	X	X	X	X		X		X
transketolase (2.2.1.1)	X	X	X	X	X	X	X	X	X		X
transaldolase (2.2.1.2)											X
Entner-Doudoroff Pathway (modified or semi-phosphorylated)											
glucokinase (2.7.1.2)							X	X	X	X	X
glucose-6-phosphate dehydrogenase (1.1.1.49)	X			X			X	X	X		X
6-phosphogluconate dehydratase (4.2.1.12)	X	X		X						X	X
2-dehydro-3-deoxygluconokinase (2.7.1.45)		X		X							
2-dehydro-3-deoxy-phosphogluconate aldolase (4.1.2.14)	X	X		X						X	X
glyceraldehyde-3-phosphate dehydrogenase (1.2.1.12)	X	X	X	X		X	X	X	X	X	X

	SAR86					ARCTIC				SAR92	
Enzyme/Pathway	AAA076P 09	AAA076P1 3	AAA168I1 8	AAA168P0 9	AAA298N 10	AAA076D 02	AAA076D 13	AAA076E 13	AAA076F 14	AAA160D 02	AAA300D 14
phosphoglycerate kinase (2.7.2.3)		X	X	X	X	X	X	X	X		
phosphoglycerate mutase (5.4.2.1)				X		X	X		X		X
enolase (4.2.1.11)	X	X	X	X	X	X	X	X	X	X	X
pyruvate kinase (2.7.1.40)	X	X		X	X	X	X	X	X	X	X
*unique to this pathway; all others shared by glycolysis and the PPP pathways											
Fate of Pyruvate											
Pyruvate Dehydrogenase Reaction											
pyruvate dehydrogenase (1.2.4.1)	X	X	X	X	X	X	X	X	X	X	X
Citric Acid Cycle (TCA)											
citrate synthase (2.3.1.1)						X	X		X	X	X
aconitate hydratase (4.2.1.3)	X	X	X	X		X	X	X	X		X
isocitrate dehydrogenase (1.1.1.42)			X	X	X	X	X		X	X	X
oxoglutarate dehydrogenase (1.2.4.2)		X	X	X			X	X		X	X
succinyl CoA-synthetase (6.2.1.5)		X	X	X		X	X		X	X	X
succinate dehydrogenase (1.3.99.1)		X	X	X		X	X		X		X
fumurate hydratase (4.2.1.2)	X	X	X	X	X	X	X		X	X	X
malate dehydrogenase (1.1.1.37)	X	X	X	X	X	X	X		X		X

Table S10. Presence (X) and absence of photoheterotrophy pathway genes identified in Gammaproteobacteria SAGs.

	SAR86					**ARCTIC**				**SAR92**	
Enzyme	AAA076 P09	AAA076 P13	AAA168I 18	AAA168 P09	AAA298 N10	AAA076 D02	AAA076 D13	AAA076 E13	AAA076 F14	AAA160 D02	AAA300 D14
Proterorhodopsin	X	X	X	X	X	X	X	X	X	X	
Retinal Biosynthesis Pathway											
geranylgeranly pyrophosphate synthase (*crtE*)	X	X	X	X	X	X	X	X	X	X	X
phytoene desaturase (*crtI*)										X	X
phytoene synthase (*crtB*)										X	X
lycopene cyclase (*crtY*)										X	
flanked by dehydrogenase	X	X	X	X	X						

Supplementary References

1. Fuhrman JA, McCallum K, Davis AA (1993) Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Appl Environ Microb* 59(5):1294–1302.
2. Gordon DA, Giovannoni SJ (1996) Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific oceans. *Appl Environ Microb* 62(4):1171–1177.
3. Freitas S, et al. (2012) Global distribution and diversity of marine Verrucomicrobia. *ISME J* 6(8):1499–1505.
4. Martinez-Garcia M, et al. (2012) Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. *PLoS ONE* 7:e35314.
5. Rappé MS, Gordon DA, Vergin KL, Giovannoni SJ (1999) Phylogeny of actinobacteria small subunit (SSU) rRNA gene clones recovered from marine bacterioplankton. *Syst Appl Microbiol* 22:106–112.
6. Wong LL (1998) Cytochrome P450 monooxygenases. *Curr Opin Chem Biol* 2:263–268.
7. Harayama S (1997) Polycyclic aromatic hydrocarbon bioremediation design. *Curr Opin Biotech* 8:268–273.
8. Demanèche S, et al. (2004) Identification and functional analysis of two aromatic-ring-hydroxylating dioxygenases from a Sphingomonas strain that degrades various polycyclic aromatic hydrocarbons. *Appl Environ Microb* 70:6714–6725.
9. Evans WC (1963) The microbiological degradation of aromatic compounds. *J Gen Microbiol* 32:177–184.
10. Gibson DT, Parales RE (2000) Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr Opin Biotech* 11:236–243.
11. Juhasz AL, Naidu R (2000) Bioremediation of high molecular weight polycyclic aromatic hydrocarbons: a review of the microbial degradation of benzo [a] pyrene. *Int Biodeter Biodegr* 45:57–88.
12. Roldán MD, Pérez-Reinado E, Castillo F, Moreno-Vivián C (2008) Reduction of polynitroaromatic compounds: the bacterial nitroreductases. *FEMS Microbiol Rev* 32:474–500.

13. Symons ZC, Bruce NC (2006) Bacterial pathways for degradation of nitroaromatics. *Nat Prod Rep* 23:845.

14. Ullrich R, Hofrichter M (2007) Enzymatic hydroxylation of aromatic compounds. *Cell Mol Life Sci* 64:271–293.

15. de Oliveira IM, Bonatto D, Henriques JAP (2010) Nitroreductases: Enzymes with environmental, biotechnological and clinical importance. *Reactions* 3:6.

16. Sun J, et al. (2011) One carbon metabolism in SAR11 pelagic marine bacteria. *PLOS ONE* 6:e23973.

17. Peters A, Kulajta C, Pawlik G, Daldal F, Koch H-G (2008) Stability of the cbb3-type cytochrome oxidase requires specific CcoQ-CcoP interactions. *J Bacteriol* 190:5576–5586.

18. Pitcher R, Brittain T, Watmough N (2002) Cytochrome cbb(3) oxidase and bacterial microaerobic metabolism. *Biochem Soc T* 30:653–658.

19. Alldredge AL, Cohen Y (1987) Can microscale chemical patches persist in the sea? Microelectrode study of marine snow, fecal pellets. *Science* 235:689–691.

20. Oh H-M, et al. (2010) Complete genome sequence of "Candidatus Puniceispirillum marinum" IMCC1322, a representative of the SAR116 clade in the *Alphaproteobacteria*. *J Bacteriol* 192(12):3240–3241.

21. Grote J, et al. (2011) Draft genome sequence of strain HIMB100, a cultured representative of the SAR116 clade of marine *Alphaproteobacteria*. *Stand Genomic Sci* 5(3):269–278.

22. King GM, Weber CF (2007) Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Microbiol* 5(2):107–118.

23. Sabehi G, et al. (2007) Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. *ISME J* 1(1):48–55.

24. Junier I, Martin O, Képès F (2010) Spatial and topological organization of DNA chains induced by gene co-localization. *PLOS Comput Biol* 6(2):e1000678.

25. Buchan A, Gonzalez JM, Moran MA (2005) Overview of the marine Roseobacter lineage. *Appl Environ Microb* 71(10):5665–5677.

26. Luo H, Löytynoja A, Moran MA (2012) Genome content of uncultivated marine Roseobacters in the surface ocean. *Environ Microbiol* 14(1):41–51.

27. McCarren J, DeLong EF (2007) Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ Microbiol* 9(4):846–858.

28. Sabehi G, Béjà O, Suzuki MT, Preston CM, DeLong EF (2004) Different SAR86 groups harbour divergent proteorhodopsins. *Environ Microbiol* 6:903–910.

29. Dupont CL, et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6(6):1186–1199.

30. Man D, et al. (2003) Diversification and spectral tuning in marine proteorhodopsins. *EMBO J* 22(8):1725–1731.

31. Swan BK, et al. (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* 333(6047):1296–1300.

32. Cottrell MT, Kirchman DL (2000) Community composition of marine bacterioplankton determined by 16S rRNA gene clone libraries and fluorescence *in situ* hybridization. *Appl Environ Microb* 66(12):5116–5122.

33. Glöckner FO, Fuchs BM, Amann R (1999) Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence *in situ* hybridization. *Appl Environ Microb* 65(8):3721–3726.

34. Nikrad MP, Cottrell MT, Kirchman DL (2012) Abundance and single-cell activity of heterotrophic bacterial groups in the western Arctic Ocean in summer and winter. *Appl Environ Microb* 78(7):2402–2409.

35. Malmstrom RR, Straza TRA, Cottrell MT, Kirchman DL (2007) Diversity, abundance, and biomass production of bacterial groups in the western Arctic Ocean. *Aquat Microb Ecol* 47:45–55.

36. Alonso-Sáez L, Sánchez O, Gasol JM, Balagué V, Pedrós-Alio C (2008) Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environ Microbiol* 10(9):2444–2454.

37. González JM, et al. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (*Flavobacteria*). *Proc Natl Acad Sci USA* 105(25):8724–8729.

38. Gómez-Consarnau L, et al. (2007) Light stimulates growth of proteorhodopsin-containing marine *Flavobacteria*. *Nature* 445(7124):210–213.

39. Bauer M, et al. (2006) Whole genome analysis of the marine *Bacteroidetes* 'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol* 8(6):2201–2213.

40. González JM, et al. (2011) Genomics of the proteorhodopsin-containing marine flavobacterium *Dokdonia* sp. strain MED134. *Appl Environ Microb* 77(24):8676–8686.

41. Béjà O, et al. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microb* 68(1):335–345.

42. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* 178(3):591–599.

43. Hsiao WWL, et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLOS Genet* 1(5):e62.

44. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5(4):433–438.

45. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.

46. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Gen Res* 18(5):821–829.

47. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Gen Res* 8:186–194.

48. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Gen Res* 8:195–202.

49. Han C, Chain P (2006) Finishing repeat regions automatically with Dupfinisher. *Proceeding of the 2006 international conference on bioinformatics & computational biology*, ed Valafar HRAH (CSREA Press), pp 141–146.

50. Woyke T, et al. (2009) Assembling the marine metagenome, one cell at a time. *PLOS ONE* 4(4):e5299.

51. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11(1):119.

52. Hacker J, Kaper JB (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 54(1):641–679.

53. Pruesse E, et al. (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucl Acids Res* 35(21):7188–7196.

54. Makarova KS, et al. (1999) Comparative genomics of the *Archaea* (*Euryarchaeota*): Evolution of conserved protein families, the stable core, and the variable shell. *Gen Res* 9(7):608–628.

55. Markowitz VM, et al. (2008) IMG/M: a data management and analysis system for metagenomes. *Nucl Acids Res* 36(suppl_1):D534-538.

56. Philippot L (2002) Denitrifying genes in bacterial and Archaeal genomes. *BBA-Gene Struct Expr* 1577(3):355–376.

57. Beaumont HJE, et al. (2002) Nitrite reductase of *Nitrosomonas europaea* is not essential for production of gaseous nitrogen oxides and confers tolerance to nitrite. *J Bacteriol* 184(9):2557–2560.

58. Cantarel BL, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucl Acids Res* 37:D233–238.

59. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC (2010) CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* 20:1574–1584.

60. Lauro FM, et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106(37):15527–15533.

61. Tully BJ, Nelson WC, Heidelberg JF (2011) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* 14(1):254–267.

62. Phillips R, Kondev J, Theriot J (2008) *Physical biology of the cell.*