LA-UR- 08-5792

Approved for public release; distribution is unlimited.

Title:	Assembling the Marine Metagenome, Once Cell at a Time
Author(s):	Gang Xie, Tanja Woyke, Alex Copeland, Jose Gonzalez, Cliff Han, Hajnalka Kiss, Jimmy Saw, Pavel Senin, Sourav Chatterji, Jan-Fang Cheng, Jonathan A. Eisen, Michael E. Sieracki, and Ramunas Stepanauskas
Intended for:	Science

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

Assembling the marine metagenome, one cell at a time

One-sentence summary: Genome sequencing from individual cells, combined with metagenomic data re-analysis, enabled reconstruction of metabolic properties and geographic distribution of two predominant, uncultured marine flavobacteria encoding proteorhodopsins

Tanja Woyke[1], Gary Xie[2], Alex Copeland[1], Jose Gonzalez[3], Cliff Han[3], Hajnalka Kiss[2], Jimmy Saw[2,6], Pavel Senin[2,7], Sourav Chatterji[1], Jan-Fang Cheng[1], Jonathan A. Eisen[1,4], Michael E. Sieracki[5] and Ramunas Stepanauskas[5*]

[1] DOE Joint Genome Institute, Walnut Creek, California; [2] Los Alamos National Laboratory, Los Alamos, New Mexico; [3] University of La Laguna, La Laguna, Tenerife, Spain; [4] University of California, Davis, California; [5] Bigelow Laboratory for Ocean Sciences, West Boothbay Harbor, Maine; [6] Dept. of Microbiology, University of Hawaii at Manoa, Honolulu, Hawaii; [7] Dept. of of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, Hawaii.

* To whom correspondence should be addressed. E-mail: rstepanauskas@bigelow.org

ABSTRACT

The uncultivability of most microorganisms and the complexity of natural microbial assemblages hinder genome reconstruction of representative taxa. We employed single cell genome sequencing to study two uncultured marine flavobacteria. In contrast to cultured strains, the single amplified genomes (SAGs) were excellent Global Ocean Sampling (GOS) metagenome fragment recruiters. The geographic distribution of GOS recruits along the coast of the Northwest Atlantic coincided with ocean surface currents. Composition of the two SAGs suggests genome streamlining and diversified energy sources, including biopolymer degradation, proteorhodopsin photometabolism, and H_2 oxidation. These features may explain the competitiveness of the two taxa in the ocean and the absence of their representatives in cultures. The combination of single cell genomics and metagenomics in this study revealed novel biological information about abundant, uncultured microorganisms.

TEXT

The metabolism of bacteria and archaea drives most of the biogeochemical cycles on Earth (*1*), has a tremendous effect on human health (*2*), and constitutes a largely untapped source of novel natural products (*3*). Recent advances in metagenomics revealed enormous diversity of previously unknown, uncultured microorganisms that predominate in the ocean, soil, deep subsurface, human body, and other environments (*2,*

43 *4-6*). However, the uncultivability of the vast majority of prokaryotes makes them
44 recalcitrant to whole genome studies. Metagenomic sequencing of microbial communities
45 has enabled genome reconstruction of only the most abundant members of extremely
46 simple assemblages (*7*). While novel isolation approaches resulted in significant progress
47 (*8*), they remain unsuited for high-throughput recovery of representative microbial taxa
48 from their environment. The lack of representative reference genomes is a major obstacle
49 in the interpretation of metagenomic data. For example, the Global Ocean Sampling
50 (GOS) produced 6.3 Gbp of shotgun DNA sequence data from surface ocean microbial
51 communities, but only a small fraction of the reads were closely related to known
52 genomes, while no novel genomes were assembled (*6*). These limitations of current
53 methods in microbiology are illustrated by the difficulty in determining the predominant
54 carriers of proteorhodopsins, which are abundant in marine metagenomic libraries and
55 likely provide a significant source of energy to the ocean food web (*6, 9, 10*). Thus, novel
56 research tools are necessary to complement cultivation and metagenomics-based studies
57 for the reconstruction of genomes, metabolic pathways, ecological niches, and
58 evolutionary histories of microorganisms that are representative of complex
59 environments.

60 Here we employed a novel approach to study the biology of two uncultured,
61 proteorhodopsin-containing flavobacteria, which are distant from existing cultured and
62 sequenced strains (fig. S1). Genomic sequencing from individual, uncultured cells
63 produced high quality draft genome assemblies, which, combined with our GOS
64 metagenome data re-analysis, enabled metabolic pathway reconstruction and examination

of the ecological adaptations and geographic distribution of these taxa. For single cell genome reconstruction, a water sample was collected from the Gulf of Maine in March 2006, microbial cells were separated by a fluorescence-activated cell sorter, and genomes of individual cells were amplified using multiple displacement amplification (MDA) (*11*). Droplet flow sorting has two significant advantages over sample dilution and microfluidics, methods which have been employed previously for single cell genome sequencing of environmental microorganisms (*12-14*): increased sample throughput, and decreased risk of DNA contamination from the sample matrix (*11*). The resulting single amplified genomes (SAGs) were shotgun-sequenced using a combination of Sanger and 454 technologies and subject to additional finishing steps and rigorous quality control to detect potential contaminants and amplification artifacts (table S1, figs. S2, S3). While 454 pyrosequencing provided low-cost, high coverage depth data without cloning biases, the addition of paired-end Sanger sequencing improved the genome assemblies and assisted resolving homopolymer regions (figs. S4-S5). The polished assemblies contain 1.9 Mbp in 17 contigs and 1.5 Mbp in 21 contigs for the SAGs MS024-2A and MS024-3C respectively, with contig length ranging 3-684 Kbp (Table 1). These major contigs recovered about 85% and 54% of the two genomes. Only about 0.7% of all sequence reads and only about 0.24% of the assembling sequences were contaminants or self-primed amplification products, and were removed from further analysis (fig. S3). The high genome recovery, low fragmentation, as well as negligible and carefully managed DNA contamination demonstrate major improvements in wetlab and bioinformatics protocols compared to prior single cell genome studies (*12-14*). We are currently

retrieving additional regions of the two SAGs with PCR-based genome finishing efforts. Further shotgun sequencing would be ineffective due to the significant overrepresentation of random genome regions in MDA products (figs. S4-S6). If a closed genome is the goal, an alternative approach may be combining shotgun sequences of multiple SAGs in a single assembly, given identical SAGs are available. Here we demonstrate that improved analytical procedures enable high-quality draft reconstruction of discrete genomes of uncultured taxa from complex communities.

We searched for the presence of MS024-2A- and MS024-3C-like DNA in the Global Ocean Sampling (GOS) data using metagenome fragment recruitment (*6*). The number of GOS fragments recruited by the two SAGs was higher, by at least an order of magnitude, than the recruitment by any of the eleven available genomes of cultured marine flavobacteria strains (Fig. 1, fig. S7). The GOS read recruitment by marine isolates, including those collected at or near GOS stations, was as low as the recruitment by the soil isolate *F. johnsoniae*. This suggests that existing flavobacteria cultures are poor representations of the predominant marine taxa. In contrast, the number of recruits at high DNA identity level was comparable for our two flavobacteria SAGs and the representatives of the ubiquitous marine genera *Prochlorochoccus*, *Synechococcus*, and *Pelagibacter*, which were previously identified as the only significant GOS fragment recruiters (*6*). This is quite remarkable, considering that the two SAGs are non-redundant genomes from a relatively small, pilot marine SAG library (*11*). Our results demonstrate the power of single cell genomics to reconstruct representative microbial genomes from complex communities, independent of their cultivability.

We further focused on the GOS recruits with >95% DNA identity to the two SAGs, as an operational demarcation of bacterial species (*15*). A total of 1,505 and 467 of >95% DNA identity recruits were obtained for MS024-2A and MS024-3C. Of these, only nine recruits encoding only two genes were shared by the two SAGs, demonstrating significant evolutionary distance between the two genomes. Interestingly, >99% of the recruits and the two SAGs themselves came from a distinct biogeographic region along the coast of the northwest Atlantic Ocean (Fig. 2A). The fraction of SAG-like DNA did not correlate with ambient temperature, salinity, and chlorophyll *a* concentrations but was highest at the two northern-most GOS stations. In Bedford Basin, Nova Scotia (GOS station #5), 1.3% and 1.2% of all metagenomic reads were >95% identical to MS024-2A and MS024-3C DNA respectively, and in the Bay of Fundy (GOS stn #6) 0.44% GOS reads matched to MS024-2A. No SAG recruits were found south of the GOS station #13 off Nags Head, North Carolina, including many tropical stations. This GOS recruit distribution corresponds to the coastal transport of the remnants of the Labrador Current, as illustrated by the ocean surface temperature during the GOS sampling (Fig. 2B). It appears that microbial taxa represented by MS024-2A and MS024-3C are most abundant in the coastal northwest Atlantic waters and may be transported southward and mixed into local bacterioplankton assemblages by surface currents along the coastline. Single cells (March 2006) and the GOS Atlantic coast stations (August-December 2003) were sampled over 3 years apart. Thus, MS024-2A and MS024-3C appear to represent two abundant marine flavobacteria taxa, which persist in particular geographic areas. To our

130 knowledge, this is the first time that biogeography of specific marine bacterioplankton
131 taxa has been linked to ocean circulation.

132 The abundance of MS024-2A- and MS024-3C-like bacterioplankton in the
133 intensely studied Atlantic coastal waters of U.S. and Canada raises two intriguing
134 questions: 1) what makes these organisms exceptionally competitive in their natural
135 environment and 2) why are they not represented in cultures? Here we propose some
136 answers, as based on the SAGs' genome composition, including genome streamlining,
137 energy-conserving metabolism, and diversified mixotrophy.

138 Genome streamlining was suggested as a nutrient and energy conserving
139 adaptation in the ubiquitous and hard-to-culture marine alphaproteobacteria clade SAR11
140 (*16*). Accordingly, MS024-2A- and MS024-3C have among the smallest genomes, the
141 lowest fraction of paralogous genes, and the lowest fraction of non-coding nucleotides
142 amongst the sequenced taxa of the Bacteroidetes phylum (Fig. 3). The significantly
143 reduced number of paralogs indicates that genome streamlining comes at a cost of
144 reduced biochemical potential. Thus, MS024-2A and MS024-3C may represent taxa well
145 adapted to a narrow ecological niche, which would explain their abundance in a specific
146 geographic area and difficulties in their laboratory cultivation.

147 Another similarity between SAR11 and the two flavobacteria SAGs is their
148 apparent inability to utilize oxidized sulfur compounds, which was recently demonstrated
149 for "*Candidatus* Pelagibacter ubique" (*17*). Both MS024-2A and MS024-3C lack genes
150 for the uptake and reduction of sulfate and sulfite. Both SAGs also lack reductases for
151 nitrate, nitrite, and nitrous oxide. The probability for a single gene being missing from

152 both SAGs due to the incomplete assemblies is only about 7%. Thus, it is likely that the
153 SAG-represented taxa rely solely on reduced N and S forms as an energy-saving strategy
154 in a C rather than N or S limited environment. The lack of nitrate and nitrite reductases is
155 a common feature in all currently available flavobacteria genomes, while sulfate and/or
156 sulfite reductases are found in 11 out of 17 of them. Although DMSP sulfur utilization
157 was suggested by microautoradiography for a subset of marine flavobacteria in a
158 community-level study (*18*), no significant homologs to known DMSP demethylases or
159 lyases were detected on any of the available flavobacteria genomes, including these two
160 SAGs. All available marine flavobacteria genomes also lack recognizable ureases, but
161 most, including MS024-2A (Flav2A_or0245, Flav2A_or0246), encode allophanate
162 hydrolases. Thus, allophanate, a breakdown product of urea, is a likely supplementary
163 source of N to many marine flavobacteria. Both SAGs contain phosphate permeases
164 (Flav2A_or0989, Flav3C_or0330) and polyphosphate kinases (Flav2A_or1736,
165 Flav3C_or0433), indicating their capacity for import and intracellular storage of
166 inorganic phosphorus. Like "*Candidatus* Pelagibacter ubique" and several *Roseobacter*
167 clade and marine flavobacteria isolates, the two SAGs lack key enzymes in the
168 production of biotin and cobalamin (biotin synthase and *cob*NST), providing evidence for
169 their nutritional reliance on fresh cellular material. This information indicates metabolic
170 dependencies in marine microbial communities and may facilitate future efforts of
171 bringing the predominant marine bacterioplankton taxa into culture.

172 The presence of proteorhodopsin genes (Flav2A_or1462, Flav3C_or0805) is yet
173 another similarity between the two SAGs and the *Pelagibacter* genomes.

Proteorhodopsins are light-driven proton pumps, which have recently been recognized for their abundance and likely biogeochemical significance in surface oceans (*6, 9, 10*). However, the hosts of the majority of marine proteorhodopsins remain unidentified. Two recent studies, utilizing single cell genomics and cultivation, demonstrated the presence of proteorhodopsins in marine bacteroidetes (*11, 19*). Our phylogenetic analysis of proteorhodopsin genes in the GOS database suggests that about 8% of them may be from organisms in the Bacteroidetes phylum (fig. S8). Bacteroidetes-like proteorhodopsins are also abundant in diverse freshwater environments (*20*). Interestingly, proteorhodopsin genes have not been detected on large environmental DNA inserts identifiable as bacteroidetes genetic material, although their presence was suggested by the gene composition of one of the Sargasso Sea metagenome scaffolds (*10*). This may be due to the large distance (in basepairs) between proteorhodopsin genes and taxonomic markers (e.g. 16S or 23S rRNA genes) in the genomes of species in this phylum. Among the currently available bacteroidetes genomes, the distance between proteorhodopsin and rRNA genes ranges 126-174 Kbp, which is larger than most metagenomic clones. In difference to metagenomics, single cell genomics enabled the linkage of phylogeny and function independent of the distance of diverse genes on a DNA molecule, thus providing unique and unbiased information about the metabolic potential of uncultured microorganisms.

Proteorhodopsin-containing microbial cultures currently include three alphaproteobacteria (*16, 21*), four bacteroidetes (*19*), and four SAR92 gammaproteobacteria (*22*). Despite extensive tests, light stimulation likely attributable to

196 proteorhodopsin activity has been detected in only one of these isolates (*19*). Thus the
197 ecological roles and expression conditions of marine proteorhodopsins remain enigmatic.
198 Intriguingly, marine planktonic bacteroidetes with proteorhodopsins have smaller
199 genomes and fewer paralogs compared to marine bacteroidetes without proteorhodopsins,
200 while non-marine bacteroidetes have larger genomes and more paralogs than their marine
201 counterparts (Fig. 3; $p<0.01$, t-test). Although the causality of this relationship is unclear,
202 the presence of proteorhodopsins on the streamlined genomes provides indirect evidence
203 for their adaptive significance. To elucidate proteorhodopsin relationships to other
204 biochemical pathways, we investigated what genes are present in all six available
205 proteorhodopsin-containing flavobacteria genomes but are absent in the remaining 13
206 flavobacteria genomes. Only three such genes were detected: proteorhodopsin, *blh*
207 (encoding β-carotene dioxygenase, which produces proteorhodopsin chromophore
208 retinal) and genes encoding DNA photolyase-like flavoproteins. The latter formed a
209 distinct phylogenetic cluster among photolyase-like genes of flavobacteria (fig. S9). It
210 may be speculated that photolyase-like flavoproteins regulate rhodopsin proton pump
211 expression or that both photometabolic systems are involved in synchronized
212 photosensing or energy production. These hypotheses may be experimentally tested using
213 pure cultures, metatranscriptome studies, or heterologous expression of SAG genes.
214 Uniquely among marine flavobacteria, MS024-2A possesses Ni,Fe-hydrogenase
215 genes *hyaA* and *hyaB* (Flav2A_or1764, Flav2A_or1770), raising the possibility that this
216 organism utilizes hydrogen as a supplementary source of energy. Potential sources of
217 hydrogen in the ocean photic zone include photochemical reactions (*23*), algal

218 metabolism (*24*), and heterotroph activity in anoxic microenvironments (*25*).
219 Hydrogenase genes are also harbored by the marine plankton *Roseobacter* clade isolates
220 *Roseovarius* sp. HTCC2601, *Roseovarius* sp. TM1035 and *Sagittula stellata* E-37, and
221 are abundant in GOS sequence data, which suggests a potentially widespread hydrogen
222 metabolism in the ocean photic zone. The physiological and ecological significance of
223 hydrogenases in marine bacterioplankton requires experimental verification. However,
224 the abundance of these genes among diverse taxonomic groups provides further evidence
225 for the significance of lithoheterotrophy in the ocean photic zone (*21*).

226 The H_2 oxidation and proteorhodopsin photometabolism may provide
227 supplementary energy and a competitive advantage in a carbon-limited environment.
228 However, the primary sources of energy and nutrients for MS024-2A and MS024-3C are
229 likely to be organic compounds. The two SAGs contain many genes involved in
230 biopolymer hydrolysis (table S2), and the import and degradation of hydrolysis products
231 (table S3). Both SAGs possess a substantial number of predicted proteins with domains
232 that have been implicated in cell-surface and cell-cell interaction (table S4). The
233 characteristic repetitive domain structures in adhesion proteins are known to bind calcium
234 ions, such as *Cadherin*, *FG-GAP* and *Thrombospondin type 3* repeats; or to bind cell
235 receptors and metal ions, such as *Fasciclin* and *Von Willebrand factor type A*. These cell
236 surface repetitive structures could play an important role in adhering to algal surface
237 mucilage to facilitate the colonization of the phytoplankton cells surface, in attaching to
238 the nutrient-rich marine snow particles, and in biofilm formation. These features are
239 consistent with the genome composition of other marine flavobacteria (*26*), with the

240 community-level evidence of marine flavobacteria proficiency in biopolymer hydrolysis
241 (*27*), and with the relative abundance of flavobacteria in algal blooms and in physical
242 associations with algal cells - the likely sources of these biopolymers (*27*).

243 In contrast to all sequenced marine flavobacteria isolates, MS024-2A contains
244 anti-sigma factor *rsb*W, its antagonist *rsb*V, an associated gene *rsb*U and a PAS domain
245 S-box (Flav2A loci or1527-or1530). This operon is probably controlled by the σ^{70} factor
246 (*rpo*D; Flav2A_ or1407) (fig. S10). The MS024-3C genome also encodes RsbW and a
247 fragment of RsbU at a tail of a contig. Other genes of the operon are missing, possibly
248 due to the incomplete MS024-3C assembly. It is likely that this cluster is involved in the
249 cellular regulation of the general stress response, as in the model organism *Bacillus*
250 *subtilis* with homologous genes (*28*) (table S5, fig. S10). The ability to switch between
251 active and dormant states in response to nutrient limitation or other stress is beneficial in
252 the ocean water column, which is characterized by a patchy availability of favorable
253 environments, with a major fraction of marine bacterioplankton being live but
254 metabolically inactive (*29*).

255 We demonstrate how the novel single cell genomics methodology enables
256 deciphering metabolic and ecological traits of the microbial "uncultured majority". This
257 approach overcomes the biases of cultivation and complements metagenomics with a
258 greatly improved genome assembly capability. Our approach may be of tremendous help
259 generating reference genomes and analyzing within-population genetic variation in
260 diverse environments, from oceans to the human microbiome.

261

REFERENCES AND NOTES

1. P. G. Falkowski, T. Fenchel, E. F. DeLong, *Science* 320, 1034 (2008).
2. P. J. Turnbaugh *et al.*, *Nature* 449, 804 (2007).
3. B. Haefner, *Drug Discovery Today* 8, 536 (2003).
4. N. R. Pace, D. A. Stahl, D. J. Lane, G. J. Olsen, *Advances In Microbial Ecology* 9, 1 (1986).
5. S. G. Tringe *et al.*, *Science* 308, 554 (Apr 22, 2005).
6. D. B. Rusch *et al.*, *PLoS Biology* 5, e77 (March 01, 2007, 2007).
7. T. Woyke *et al.*, *Nature* 443, 950 (2006).
8. U. Stingl, H. J. Tripp, S. J. Giovannoni, *ISME Journal* 1, 361 (2007).
9. O. Beja, E. N. Spudich, J. L. Spudich, M. Leclerc, E. F. DeLong, *Nature* 411, 786 (Jun 14, 2001).
10. J. C. Venter *et al.*, *Science* 304, 66 (Apr 2, 2004).
11. R. Stepanauskas, M. E. Sieracki, *Proceedings Of The National Academy Of Sciences Of The United States Of America* 104, 9052 (2007).
12. K. Zhang *et al.*, *Nature Biotechnology* 24, 680 (Jun, 2006).
13. Y. Marcy *et al.*, *Proceedings Of The National Academy Of Sciences Of The United States Of America* 104, 11889 (Jul, 2007).
14. T. Kvist, B. K. Ahring, R. S. Lasken, P. Westermann, *Applied Microbiology And Biotechnology* 74, 926 (Mar, 2007).

15. J. Goris *et al.*, *International Journal Of Systematic And Evolutionary Microbiology* 57, 81 (2007).

16. S. J. Giovannoni *et al.*, *Science* 309, 1242 (Aug 19, 2005).

17. H. J. Tripp *et al.*, *Nature* 452, 741 (2008).

18. M. Vila *et al.*, *Applied And Environmental Microbiology* 70, 4648 (Aug, 2004).

19. L. Gomez-Consarnau *et al.*, *Nature* 445, 210 (2007).

20. N. Atama-Ismael *et al.*, *The ISME Journal* 2, 656 (2008).

21. M. A. Moran, W. L. Miller, *Nature Reviews Microbiology* 5, 792 (2007).

22. U. Stingl, R. A. Desiderio, J. C. Cho, K. L. Vergin, S. J. Giovannoni, *Applied And Environmental Microbiology* 73, 2290 (Apr, 2007).

23. S. Punshon, R. M. Moore, *Marine Chemistry* 108, 215 (2008).

24. A. Melis, L. Zhang, M. Forestier, M. L. Ghirardi, M. Seibert, *Plant Physiology* 122, 127 (2000).

25. S. T. Braun, L. M. Proctor, S. Zani, M. T. Mellon, J. P. Zehr, *Fems Microbiology Ecology* 28, 273 (1999).

26. J. M. Gonzalez *et al.*, *Proceedings Of The National Academy Of Sciences Of The United States Of America* 105, 8724 (2008).

27. D. L. Kirchman, *Fems Microbiology Ecology* 39, 91 (Feb, 2002).

28. A. Petersohn *et al.*, *Journal Of Bacteriology* 183, 5617 (2001).

29. E. M. Smith, P. A. del Giorgio, *Aquatic Microbial Ecology* 31, 203 (Mar 13, 2003).

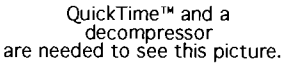
30. S. Kurtz *et al.*, *Genome Biology* 5, (2004).

ACKNOWLEDGEMENTS

We thank Lynne Goodwin for her efforts in managing this project, Mike Zhang for input on the chimer detection analysis, Natalia Ivanova for annotation advice, Nicole Poulton for flow cytometry, and Wendy Bellows for PCR analyses. The study was supported by the "DOE 2007 Microbes" program grant DOEM-78201 to RS; NSF grants EF-0633142 and MCB-0738232 to RS and MS, Maine Technology Institute grant to MS, Spanish Ministry of Science and Innovation grant CTM2007-63753-C02-01/MAR to JG; and DARPA grants HR0011-05-1-0057 and FA9550-06-1-0478 to SC. Part of this work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. The sequence data has been deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank) under project accessions XXXXX (MS024-2A) and YYYYYY (MS024-3C).

322 **TABLES**

323



324

325

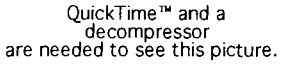
FIGURE LEGENDS

Fig. 1. Global Ocean Sampling (*6*) metagenome fragment recruitment by single amplified genomes MS024-2A and MS024-3C, the available marine flavobacteria isolate genomes, the non-marine *F. johnsoniae*, and the three best GOS fragment recruiters *Pelagibacter*, *Prochlorococcus* and *Synechocuccus*. Fragment recruitment was performed with MUMMER (*30*) and only ≥400 bp alignments were counted. *Psychroflexus torquis* ATCC 700755 was excluded from the analysis due to its poor genome assembly quality. The following marine flavobacteria genomes had fewer than 10 recruits and are not shown: *Croceibacter atlanticus* HTCC2559, *Robiginitalea biformata* HTCC2501, *Gramella forsetii* KT0803, *Kordia algicida* OT-1, isolate ALC-1, isolate HTCC2170, and isolate BBFL7. *Croceibacter atlanticus* HTCC2559 and *Robiginitalea biformata* HTCC2501 were originally collected at or near GOS sampling stations.

Fig. 2. A: Geographic distribution of the Global Ocean Sampling (GOS) metagenome fragments with >95% identity to MS024-2A and MS024-3C DNA. Numerals on the map indicate GOS station numbers.

B: Sea surface temperature in December 2003, which demonstrates hydrological separation of GOS aquatic samples collected north and south of Cape Hatteras (near GOS station 13). Provided is a composite Aqua-MODIS image for December 2003 (http://oceancolor.gsfc.nasa.gov). The GOS stations were numbered in

347 the order of their sampling, and stations 12, 13 and 14 were sampled on December 18, 19
348 and 20, 2003.
349
350 **Fig. 3.** Genome streamlining in MS024-2A and MS024-3C, evidenced by small genome
351 sizes, low fraction of genes in paralog families, and low fraction of non-coding bases.
352 Included are all available genomes of Bacteroidetes phylum. The number of genes in
353 paralog families was estimated using the BLASTCLUST tool from the NCBI BLAST
354 software (>30% sequence similarity, across >50% of their length and $E<10^{-6}$).
355



356

357

358

359

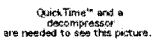
360

361

362

363

364 **Fig. 1.**



365

366

367

368

369

370

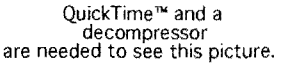
371

372

373

374

375 **Fig. 2.**



376

377

378

379

380

381

382 **Fig. 3.**