LETTERS

Bacterial carbon processing by generalist species in the coastal ocean

Xiaozhen Mou[1], Shulei Sun[1], Robert A. Edwards[2], Robert E. Hodson[1] & Mary Ann Moran[1]

The assimilation and mineralization of dissolved organic carbon (DOC) by marine bacterioplankton is a major process in the ocean carbon cycle[1]. However, little information exists on the specific metabolic functions of participating bacteria and on whether individual taxa specialize on particular components of the marine DOC pool[2]. Here we use experimental metagenomics to show that coastal communities are populated by taxa capable of metabolizing a wide variety of organic carbon compounds. Genomic DNA captured from bacterial community subsets metabolizing a single model component of the DOC pool (either dimethylsulphoniopropionate or vanillate) showed substantial overlap in gene composition as well as a diversity of carbon-processing capabilities beyond the selected phenotypes. Our direct measure of niche breadth for bacterial functional assemblages indicates that, in accordance with ecological theory, heterogeneity in the composition and supply of organic carbon to coastal oceans may favour generalist bacteria. In the important interplay between microbial community structure and biogeochemical cycling, coastal heterotrophic communities may be controlled less by transient changes in the carbon reservoir that they process and more by factors such as trophic interactions and physical conditions.

The composition of marine bacterioplankton communities can readily be accessed through culture-independent analyses of 16S ribosomal RNA, yet there have been only limited opportunities to determine metabolic roles of the individual taxa[3]. For the marine carbon cycle, the complexity of the DOC pool and the taxonomic diversity of the heterotrophic bacteria that process it has impeded efforts to establish informative taxon–function linkages. Thus, the discrete roles of the heterotrophic marine bacterioplankton groups in the mineralization, export and storage of organic matter in the oceans are not understood and cannot serve as the basis for predictive models of carbon cycling in a changing ocean.

Dimethylsulphoniopropionate (DMSP) and vanillate are components of the DOC pool in coastal sea water. DMSP is released from marine phytoplankton into surface sea water, where it supports up to 10% of bacterial carbon demand[4]. Vanillate and other lignin monomers are released during the microbial processing of vascular plant detritus, and they contribute significantly to the DOC pool in marsh-influenced coastal waters[5]. A coastal bacterial assemblage (0.2–3.0 μm size fraction) was amended with 100 nM DMSP or vanillate in the presence of the thymidine analogue bromodeoxyuridine (BrdU)[6,7], and newly synthesized DNA was separated by immunocapture of BrdU-labelled DNA after 12 h. The captured DNA represented metagenomes of functional subsets of the bacterial community able to metabolize DMSP or vanillate. Bacterial assemblages without an added model DOC compound served as controls. Pyrosequencing produced 190,872 total reads from duplicate DMSP-specific metagenomes and 115,807 reads from duplicate vanillate-specific metagenomes (96 ± 16 bp (mean ± s.d.) per read; Table 1).

To confirm that the immunocapture protocol was effective at targeting newly synthesized DNA, we performed taxonomic fingerprinting by terminal restriction-fragment length polymorphism (T-RFLP). T-RFLP analysis showed that the 16S rRNA gene composition of immunocaptured DNA was distinct from that of the DNA that remained uncaptured (Supplementary Fig. 1). Further, when immunocapture was performed for samples without amendment with BrdU, no measurable DNA was recovered.

In the metagenomes, 16S rRNA genes accounted for 0.14% of the sequences (Table 1), in accordance with the expected frequency in genomes of cultured marine prokaryotes (0.19%; Supplementary Table 1). Comparisons between the DMSP-specific and vanillate-specific metagenomes for 16S rRNA sequences identified to the Order level showed similar taxonomic compositions (Fig. 1). Typical coastal ocean bacterioplankton taxa (ref. 8 and Supplementary Fig. 2) dominated both types of functional assemblage, including γ-Proteobacteria (61% in DMSP and 53% in vanillate; primarily Alteromonadales and Oceanospirillales), α-Proteobacteria (16% and 12%; primarily Roseobacter clade) and β-Proteobacteria (9% and 11%; primarily Burkholderiales). To a smaller extent, environmental clusters characteristic of oligotrophic oceans were present (SAR11, SAR116, SAR86, SAR92, SAR432, OM60/241 and OM185;

Table 1 | Annotation statistics for experimental metagenomes

Parameter	*DMSP1*	*DMSP2*	*VAN1*	*VAN2*
Size of library (bp)	9,481,910	8,696,512	3,900,971	7,137,286
Number of sequences	99,096	91,776	41,552	74,255
Average sequence length (bp) (mean ± s.d.)	96 ± 16	95 ± 17	94 ± 17	96 ± 16
Number (%) of predicted 16S rRNA genes*	134 (0.1)	161 (0.2)	53 (0.1)	106 (0.1)
Number (%) of predicted functional genes*	31,114 (31)	23,860 (26)	11,055 (27)	26,259 (35)
Number (%) categorized by COG*	15,864 (16)	14,072 (15)	8,347 (20)	14,862 (20)
Number (%) categorized by KEGG pathway*	6,652 (7)	5,839 (6)	4,182 (10)	6,544 (9)
Number (%) categorized by SEED subsystem*	12,598 (13)	8,698 (9)	3,520 (8)	8,623 (12)

* Cutoff values for BLAST were determined by *in silico* analysis of randomly fragmented known genes. An *E* value of less than 10^{-5}, a hit length of more than 65 nt, and a similarity of more than 80% were used to predict 16S rRNA genes in BLASTN analysis against the Ribosomal Database Project. An *E* value of less than 10^{-2}, a hit length of more than 23 amino-acid residues and a similarity of more than 40% were used to predict functional genes in BLASTX analysis against the NCBI nr database. An *E* value of less than 10^{-1}, a hit length of more than 23 amino-acid residues and a similarity of more than 40% were used to predict functional genes in BLASTX analysis against the COG, KEGG and SEED databases.

[1]Department of Marine Sciences, University of Georgia, Athens, Georgia 30602, USA. [2]Department of Computer Science, San Diego State University, San Diego, California 92182, USA.

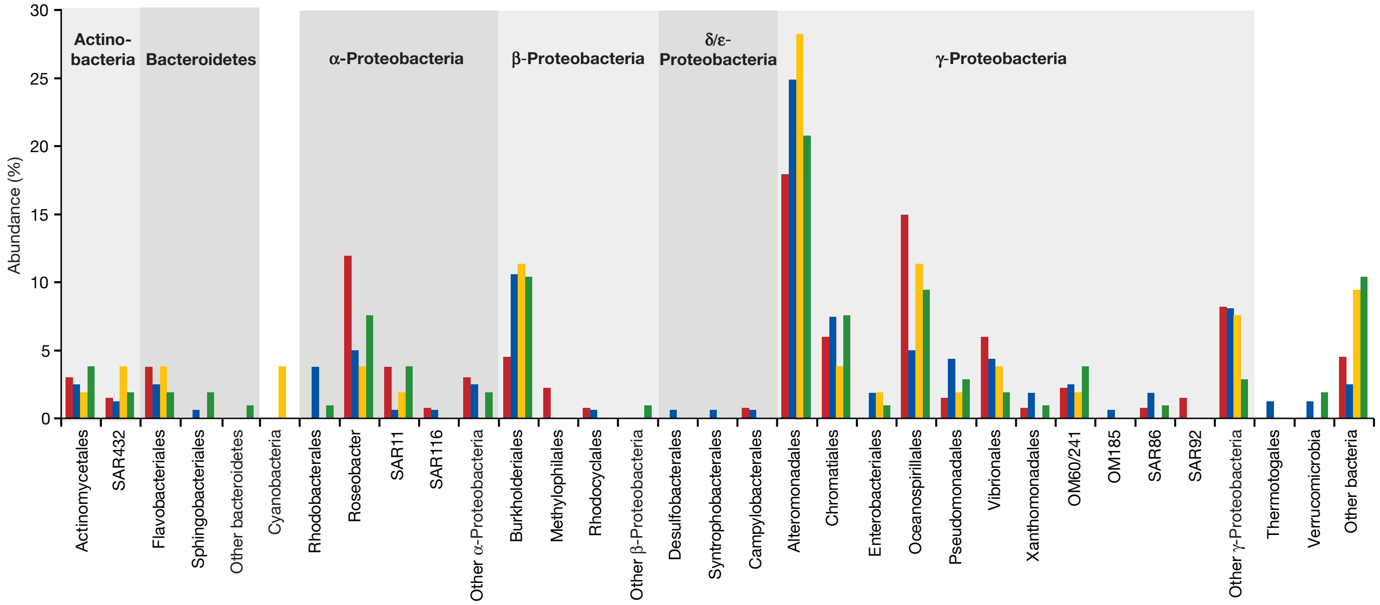

Figure 1 | Apparent taxonomic distribution of 16S rRNA-like gene sequences recovered from metagenomes. Red, *DMSP1*; blue, *DMSP2*; yellow, *VAN1*; green, *VAN2*.

1–4%), also with no apparent differences between metagenome types. Independent taxonomic analysis based on the amplification by PCR of 16S rRNA genes from the immunocaptured DNA confirmed the similar taxonomy of the DMSP-specific and vanillate-specific bacterioplankton assemblages, although the PCR-based methodology recovered fewer rare groups (Supplementary Fig. 3).

Similarity in 16S rRNA gene content of BrdU-labelled metagenomes might result if cell replication were to occur at the expense of organic matter pre-existing in the coastal sea water or released by experimental manipulations. However, cell numbers increased significantly for bacterioplankton communities amended with DMSP and vanillate (by 31% and 23%; two-tailed *t*-test, $P < 0.0001$ and $P = 0.0012$) but did not change in controls without carbon additions (Supplementary Fig. 4). Chemical analyses indicated that all added DMSP and vanillate was consumed within 12 h (Supplementary Fig. 5). Observed increases in cell numbers (3.4×10^5 ml^{-1}) were comparable to expected increases if all growth was at the expense of DMSP or vanillate (2.7×10^5 ml^{-1} if cells contain 10 fg C and grow at 35% efficiency), indicating that most cell replication was supported by the added model DOC. 16S rRNA clone libraries from the initial and 12-h controls without carbon addition had comparable fine-scale taxonomic compositions (Supplementary Fig. 6).

For a direct comparison of metabolic abilities represented in the DMSP-specific and vanillate-specific assemblages, metagenomic sequences were annotated on the basis of homology to protein-encoding genes. *In silico* analysis of random fragments of known genes was first used to establish criteria for the gene predictions from the pyrosequences by BLAST analysis (Supplementary Figs 7 and 8); self-hits were excluded to mimic annotation of environmental sequences. With these criteria, 30% of the pyrosequences were identified as fragments of protein-encoding genes on the basis of sufficient homology to entries in the National Center for Biotechnology Information's non-redundant protein (NCBI nr) database (Table 1). The remaining sequences could not be definitively identified as genes encoding protein or rRNA, and probably represent unidentified genes, poorly conserved regions of known genes, or intergenic regions. The taxonomic bins of identified bacterial protein homologues were similar to those of the 16S rRNA sequences (Supplementary Fig. 9). On the basis of BLAST hits to essential single-copy genes[9], we estimate that 22 and 17 genome equivalents were represented by the protein-encoding genes in the two DMSP-specific metagenomes (*DMSP1* and *DMSP2*, respectively, and 9 and 17 genome equivalents in the vanillate-specific metagenomes (*VAN1* and *VAN2*, respectively) (Fig. 2). Because of the short length of the pyrosequences, only about one-tenth of each gene (96 bp average sequence length per 1,000 bp average bacterial gene length) was covered by the metagenomic sequence (that is, 0.1× coverage of each genome equivalent).

Two genes known to mediate the catabolism of DMSP were not overrepresented in the DMSP-specific metagenomes relative to the vanillate-specific metagenomes. Homologues to genes encoding the demethylation (*dmdA*)[10] and cleavage (*dddD*)[11] of DMSP were present in numbers sufficient to be found in about 18% and about 11% of the bacterial cells overall, but there was no pattern to their distribution (Table 2). Similarly, homologues to two genes encoding the catabolism of vanillate (*vanB* and *pcaH*) were present in about 17% and about 13% of cells, but with no bias towards the vanillate-specific

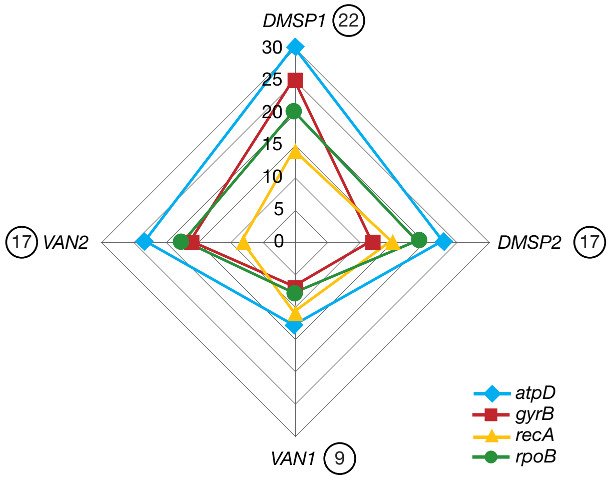

Figure 2 | Estimated genome equivalents in experimental metagenomic data sets based on homologue counts for single-copy genes. Raw counts were corrected for the effect of gene length on the probability of sampling by normalization to the length of *recA* (*recA*, 1,077 nt; *atpD*, 1,380 nt; *gyrB*, 2,418 nt; *rpoB*, 4,026 nt). The circled numbers show the estimated genome equivalent based on mean counts for the four single-copy genes. GenBank accession numbers for BLAST analysis: *atpD*, AAC76755; *gyrB*, ABG38537; *recA*, ABG32249; *rpoB* AAC76961.

Table 2 | Estimated percentages of cells with selected carbon-cycle genes

Gene*	Function	Percentage of cells†			
		DMSP1	*DMSP2*	*VAN1*	*VAN2*
dmdA	DMSP degradation	4	17	32	18
dddD	DMSP degradation	10	12	19	3
vanB	Vanillate degradation	10	33	12	14
pcaH	Protocatechuate degradation	7	9	16	18
chi	Chitin degradation	4	10	0	0
CoxL‡§	Carbon monoxide oxidation	50	73	44	46
PotA§	Polyamine transport	40	23	22	41
PR gene	Proteorhodopsin-based light harvesting	39	25	16	0
AAnP genes‖	Aerobic anoxygenic phototrophy	13	2	4	15
SoxB	Inorganic sulphur oxidation	20	18	7	23
Xsc	Taurine degradation	5	7	0	4
MtdB	Methylotrophy	0	10	0	0

* GenBank accession numbers of sequences used for BLAST analysis: AAV95190 (*dmdA*), EAV90715 and AAQ87407 (*dddD*), ABJ14290 (*vanB*), AAG03543 (*pcaH*), BAB21607 (*chi*), AAV94806 (*coxL*), AAV95654 (*coxL*), AAV94896 (*potA*), ZP_01054176, Q9F7P4 and EAQ40925 (proteorhodopsin), AAU00045 (*pufL*), ABN14037 (*pufM*), AAF24297 (*bchX*), AAV94301 (*soxB*), ABE35262 (*xsc*) and AAT02324 (*mtdB*).

† Raw homologue counts were corrected for differences in gene length by normalization to *recA* length as in Fig. 2. Corrected counts were converted to estimates of the percentage of cells containing homologues based on the equation percentage of cells = 100 × $H_{\mathrm{c,gene}}/G_{\mathrm{E}}$, where $H_{\mathrm{c,gene}}$ is the length-corrected number of homologues of a given gene and G_{E} is the estimated mean number of genome equivalents from Fig. 2.

‡ Sum of homologues from two clades of putative carbon monoxide dehydrogenases[29].

§ This may overestimate frequency because more than one copy of this gene has been found in genomes of some cultured marine bacteria.

‖ Average of homologue counts for three genes unique to aerobic anoxygenic phototrophy, namely *pufL*, *pufM* and *bchX* (ref. 30).

metagenomes. The relevance of these genes to DMSP and vanillate transformations by coastal marine bacteria has been demonstrated[10,12], although additional genes are likely to be involved[11].

Metagenomic sequences were assigned to COG (Clusters of Orthologous Groups) categories (2,620 categories), KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (173 pathways) and SEED subsystems (519 subsystems) (Table 1). Statistical comparisons of protein predictions between DMSP-specific and vanillate-specific metagenomes by resampling[13] identified only 15 COG groups, 2 KEGG pathways and 10 SEED subsystems that were overrepresented in either the DMSP-specific or vanillate-specific metagenomes at a confidence level of 95% (Supplementary Table 2). These numbers are considerably lower than previous comparisons of metagenomic libraries with a comparable approach[14,15]. We conducted simulations with the COG data to estimate the disparity in assignments necessary to achieve statistical significance. A 1.5-fold difference in frequency could be detected if a category contained at least 13 sequences in the smaller library; greater fold differences were required for less abundant categories (Supplementary Fig. 10). Non-metric multidimensional scaling analysis of sequence assignments to COG categories indicated that the composition of metagenomes was no more similar within treatments than between them (Supplementary Fig. 11).

We searched the annotated genes for the presence of additional carbon-cycle-relevant capabilities, including carbon monoxide oxidation (*coxL*), chitin degradation (*chi*), methylotrophy (*mtdB*), polyamine transport (*potA*), taurine oxidation (*xsc*) and supplementary energy acquisition through proteorhodopsin and aerobic anoxygenic phototrophy (*pufL*). Despite incomplete coverage (Fig. 2), homologues to most genes were identified in both types of metagenome (Table 2), often in numbers at least equivalent to the DMSP and vanillate catabolic genes. Thus bacteria contributing genomes to the immunocaptured DNA had the ability to mediate various carbon transformations in addition to the specific phenotypes targeted.

The incubation time (12 h) relative to the expected generation times for coastal marine bacteria (4–84 h)[16] provided limited opportunity for major community shifts. It is therefore unlikely that rare bacterioplankton taxa overwhelmed the community during the time required for the accumulation of BrdU-labelled DNA. Further, most sequences from the PCR-based 16S rRNA clone libraries (about 95%) had best hits in BLASTN analysis to sequences obtained by culture-independent methods. Some immunocaptured DNA might originate from bacterioplankton using metabolites of DMSP and vanillate released by other taxa, and active cells incapable of assimilating BrdU would be missed by the capture protocol. However, neither of these caveats would affect our conclusion that the observed metagenome compositions are inconsistent with substantial metabolic specialization within the bacterioplankton community. These results do not exclude the possibility that specialists for DMSP and vanillate are present as minor members of this coastal bacterioplankton community that could dominate under conditions of constant supply[17], or that compounds less ubiquitous than DMSP and vanillate are metabolized by specialist taxa. Metagenomic sequencing of the control BrdU-labelled DNA might have shown more evidence for these groups. Nevertheless, the bacterial taxa poised to metabolize low-concentration pulses of common components of the organic matter pool in this coastal ocean are a taxonomically broad collection of metabolic generalists rather than a limited number of metabolic specialists.

Ecological theory predicts that heterogeneous environments favour the establishment of generalist species with broad ecological niches[18], although quantitative measurements of niche breadth have hitherto been extremely difficult to obtain[19]. Our direct sampling of the genetic capabilities of bacterioplankton performing defined roles in DOC processing (that is, the 'fundamental niche' concept applied to a bacterial assemblage)[20] by using an experimental metagenomics approach is in agreement with this theory, given the heterogeneity in supply rate and composition of organic matter to this coastal system[21]. It remains to be seen whether generalist bacteria are typical of marine environments with more predictable delivery of organic matter, such as the deep ocean or cold seeps. Bacterial generalists have also been proposed to dominate open ocean surface waters, although driven by the diverse pool of dilute substrates in oligotrophic sea water[22] rather than by heterogeneity in DOC supply. For ocean waters dominated by heterotrophic generalists, transient changes in the DOC pool may be less important than selective viral[23] or protistan[24] mortality or physical conditions[2,25] in determining short-term dynamics in taxonomic and genomic composition. Predictive modelling of biogeochemical processes in a changing coastal ocean requires a knowledge of how carbon-cycle functionalities are packaged into[26,27] and regulated within[28] individual bacterial cells.

METHODS SUMMARY

Metagenomic DNA. Two sets of duplicate 20-litre microcosms were established with coastal sea water (collected in November 2005 at Sapelo Island, Georgia, geographical coordinates 81.2699° W, 31.3929° N) and amended with DMSP or vanillate (100 nM final concentration) and BrdU (10 μM). Two additional microcosms with BrdU only served as no-addition controls for manipulation effects. The BrdU-labelled DNA was extracted from microcosms after 12 h and was immunochemically purified with a modification of the method in ref. 6. Captured DNA from the treatment microcosms was sequenced by 454 Life Sciences.

Pyrosequence annotation. Unassembed pyrosequences were analysed by BLASTN against the Ribosomal Database Project II (RDPII) database to identify putative 16S rRNA sequences. Remaining sequences were analysed by BLASTX against the nr, COG, KEGG and SEED databases to identify putative protein-encoding sequences. *In silico* analyses of randomly fragmented nearly full-length 16S rRNA and protein-encoding genes from relevant bacterial taxa (Supplementary Table 3) were used to establish criteria for annotation (Supplementary Figs 7, 8 and 12).

Analysis of PCR-based 16S rRNA sequences. 16S rRNA gene sequences were amplified from initial, control and immunocaptured DNA by PCR. T-RFLP analysis was performed by digestion of carboxyfluorescein-labelled PCR amplicons with *Cfo*I (Roche). PCR amplicons were also cloned and sequenced from selected samples. Both T-RFLP profiles and 16S rRNA clone libraries showed no major composition shifts over 12 h in the absence of an added model compound (Supplementary Figs 6 and 13).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 20 September; accepted 30 November 2007.
Published online 27 January 2007.

1. Azam, F. Microbial control of oceanic carbon flux: The plot thickens. *Science* 280, 694–696 (1998).

2. Fuhrman, J. A. *et al.* Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc. Natl Acad. Sci. USA* **103,** 13104–13109 (2007).
3. Cottrell, M. T. & Kirchman, D. L. 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,** 1692–1697 (2000).
4. Kiene, R. P., Linn, L. J. & Bruton, J. A. New and important roles for DMSP in marine microbial communities. *J. Sea Res.* **43,** 209–224 (2000).
5. Moran, M. A. & Hodson, R. E. Dissolved humic substances of vascular plant origin in a coastal marine environment. *Limnol. Oceanogr.* **39,** 762–771 (1994).
6. Urbach, E., Vergin, K. L. & Giovannoni, S. J. Immunochemical detection and isolation of DNA from metabolically active bacteria. *Appl. Environ. Microbiol.* **65,** 1207–1213 (1999).
7. Hamasaki, K., Taniguchi, A., Tada, Y., Long, R. A. & Azam, F. Actively growing bacteria in the Inland Sea of Japan, identified by combined bromodeoxyuridine immunocapture and denaturing gradient gel electrophoresis. *Appl. Environ. Microbiol.* **73,** 2787–2798 (2007).
8. Crump, B. C., Armbrust, E. V. & Baross, J. A. Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia River, its estuary, and the adjacent coastal ocean. *Appl. Environ. Microbiol.* **65,** 3192–3204 (1999).
9. Santos, S. R. & Ochman, H. Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ. Microbiol.* **6,** 754–759 (2004).
10. Howard, E. C. *et al.* Bacterial taxa limiting sulfur flux from the ocean. *Science* **314,** 649–652 (2006).
11. Todd, J. D. *et al.* Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. *Science* **315,** 666–669 (2007).
12. Buchan, A., Collier, L. S., Neidle, E. L. & Moran, M. A. Key aromatic-ring-cleaving enzyme, protocatechuate 3,4-dioxygenase, in the ecologically important marine Roseobacter lineage. *Appl. Environ. Microbiol.* **66,** 4662–4672 (2000).
13. Rodriguez-Brito, B., Rohwer, F. & Edwards, R. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7,** 162 (2006).
14. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogenologic conditions. *BMC Genomics* **7,** 57 (2006).
15. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
16. Yokokawa, T. & Nagata, T. Growth and grazing mortality rates of phylogenetic groups of bacterioplankton in coastal marine environments. *Appl. Environ. Microbiol.* **71,** 6799–6807 (2005).
17. Vila, M. *et al.* Use of microautoradiography combined with fluorescence in situ hybridization to determine dimethylsulfoniopropionate incorporation by marine bacterioplankton taxa. *Appl. Environ. Microbiol.* **70,** 4648–4657 (2004).
18. Kassen, R. The experimental evolution of specialists, generalists, and the maintenance of diversity. *J. Evol. Biol.* **15,** 173–190 (2002).
19. Futuyma, D. J. & Moreno, G. The evolution of ecological specialization. *Annu. Rev. Ecol. Syst.* **19,** 207–233 (1988).
20. Hutchinson, G. E. Concluding remarks. *Quant. Biol.* **22,** 415–427 (1957).
21. Moran, M. A., Sheldon, W. M. & Sheldon, J. E. Biodegradation of riverine dissolved organic carbon in five estuaries of the Southeastern United States. *Estuaries* **22,** 55–64 (1999).
22. Button, D. K., Robertson, B., Gustafson, E. & Zhao, X. Experimental and theoretical bases of specific affinity, a cytoarchitecture-based formulation of nutrient collection proposed to supercede the Michaelis–Menten paradigm of microbial kinetics. *Appl. Environ. Microbiol.* **70,** 5511–5521 (2004).
23. Bouvier, T. & del Giorgio, P. A. Key role of selective viral-induced mortality in determining marine bacterial community composition. *Environ. Microbiol.* **9,** 287–297 (2007).
24. Beardsley, C., Pernthaler, J., Wosniok, W. & Amann, R. Are readily culturable bacteria in coastal North Sea waters suppressed by selective grazing mortality? *Appl. Environ. Microbiol.* **69,** 2624–2630 (2003).
25. Hewson, I., Steele, J. A., Capone, D. G. & Fuhrman, J. A. Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Mar. Ecol. Prog. Ser.* **311,** 67–77 (2006).
26. Doney, S. C., Abbott, M. R., Cullen, J. J., Karl, D. M. & Rothstein, L. From genes to ecosystems: the ocean's new frontier. *Frontiers Ecol. Environ.* **2,** 457–468 (2004).
27. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science* **315,** 1843–1846 (2007).
28. Su, Z. *et al.* Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102. *Nucleic Acids Res.* **34,** 1050–1065 (2006).
29. Moran, M. A. *et al.* Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432,** 910–913 (2004).
30. Yutin, N. *et al.* Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environ. Microbiol.* **9,** 1464–1475 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. Whitman, C. Lasher, B. Rodriguez-Brito, C. Reisch, J. Henriksen, E. Biers and W. Sheldon for advice or assistance with experimental or bioinformatic analyses; C. English for graphics; and J. T. Hollibaugh for discussions. This project was funded by grants from the National Science Foundation and the Gordon and Betty Moore Foundation.

Author Contributions X.M. and M.A.M. planned the project; X.M. conducted the experimental work; S.S. and R.A.E. conducted the bioinformatic and statistical analyses; X.M., R.E.H. and M.A.M. interpreted results; M.A.M. directed the project and wrote the paper.

Author Information Metagenomic sequences are deposited in the Genome Projects Database (http://www.ncbi.nlm.nih.gov/Genomes) under accession number 19145. 16S rRNA gene sequences are deposited in GenBank under accession numbers DQ880941–DQ881441 and EU167151–EU167496. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.A.M. (mmoran@uga.edu).

METHODS

Sample collection, processing, and labelling with BrdU. Surface water was collected in acid-washed Nalgene carboys and immediately filtered through 3.0-μm pore-size polycarbonate filters (Poretics Products) to exclude eukaryotes and large particles. Six microcosms were established in polycarbonate carboys consisting of filtered sea water amended with 10 μM (final concentration) of BrdU (Roche Applied Science). DMSP or vanillate (100 nM final concentration) was added to duplicate microcosms, and two served as no-addition controls. The time from water collection to the incubation start time was less than 2 h. Bacterial cells were collected from the initial sample and from the microcosms after 12 h (in the dark at 25 °C with occasional agitation) on 0.2-μm pore-size polycarbonate filters (Poretics Products). Subsamples were preserved in 4% paraformaldehyde and stored at 4 °C until cell enumeration by epifluorescence microscopy.

DNA extraction and immunocapture. Filters were processed with a PowerMax Soil Mega Prep DNA Isolation Kit (MoBio) to obtain genomic DNA. The BrdU-labelled DNA fraction was immunochemically purified with a modification of the method of ref. 31. In brief, herring sperm DNA (0.63 mg ml^{-1} in PBS; Promega) was heated at 98 °C for 5 min and then immediately cooled for 5 min in ethanol–solid CO$_2$. Denatured herring sperm DNA was mixed with monoclonal anti-BrdU antibodies (Zymed Laboratories) at a 9:1 ratio and incubated for 45 min at room temperature. A 35-μl aliquot of diluted community genomic DNA was denatured with the same procedure and mixed with 30 μl of the herring sperm DNA–antibody mixture before incubating for a further 1 h in the dark at room temperature, with agitation. Dynabeads coated with goat anti-mouse immunoglobulin G (Dynal Biotech ASA) were washed five times with 1 mg ml^{-1} acetylated BSA in PBS buffer with a magnetic particle concentrator (Dynal) and resuspended in BSA–PBS buffer. The washed Dynabeads were mixed with 65 μl of genomic DNA–herring sperm DNA–antibody mixture and incubated for 1 h at room temperature. Dynabeads were subsequently washed three times with 0.5 ml of BSA–PBS and incubated in 100 μl of 1.7 mM BrdU (in PBS–BSA) for 1 h in the dark at room temperature, with constant agitation. The BrdU-labelled DNA was separated from the beads with a magnetic particle concentrator and purified by ethanol precipitation and gel purification with a QIAquick gel extraction kit (Qiagen). BrdU-labelled genomic DNA from each duplicate DMSP and vanillate treatment was sequenced by 454 Life Sciences with the use of pyrosequencing technology.

Validation of gene prediction from pyrosequences. The nucleotide sequences of 100 known almost full-length (more than 1,200 nt) 16S genes encoding rRNA that covered major taxa of marine bacteria were selected from the RDPII database (Supplementary Table 3). Nine sequence fragments were randomly generated from each sequence (lengths from 20 to 500 nt) and used in a BLASTN analysis against the RDPII database. To mimic the conditions of annotating environmental sequences, self-hits were excluded and the taxonomic identities of the best not-to-self (BNTS) hits were assigned. As the length of the 16S rRNA gene fragments decreased, the expected value (E) increased (Supplementary Fig. 12). All fragments longer than 35 bp hit a 16S rRNA gene with an E value of 10^{-5} or less, and this value was used as the annotation cutoff criterion.

To establish criteria for further taxonomic assignments of 16S rRNA sequences, the known random fragments were analysed with the SIMO RDP Agent (http://www.simo.marsci.uga.edu/public_db/taxonomy.htm#rdpagent) and local Smith–Waterman alignment to reference sequences of marine bacteria[32]. The two programs retrieve classification information from the type sequences in the RDPII database and a local marine reference sequence data set, respectively, and perform pair-by-pair alignments to assign taxonomic ranks on the basis of similarity-weighted cutoffs. Most known 16S rRNA sequence fragments (more than 90%) were assigned to the correct Order when the query sequences were at least 65 bp long (about 93% of the pyrosequences) and at least 80% similar to the corresponding BNTS hits (Supplementary Fig. 12). On the basis of these analyses, criteria were established for accepting a taxonomic assignment at the Order level: $E \leq 10^{-5}$, overlapping length ≥ 65 nt and similarity ≥ 80% to the best hit to the RDPII database. To assign the 16S rRNA sequences further to typical marine clusters, a cutoff of at least 90% similarity in local Smith–Waterman alignments to reference sequences of marine bacteria was used.

A similar *in silico* prediction exercise was conducted for putative protein-encoding genes with the use of 100 known functional genes (Supplementary Table 3). The best hit in BLASTX analysis of the first 100 pyrosequences in the DMSP1 metagenomic library against the NCBI Clusters of Orthologous Groups (COG) database[33] was used to generate the *in silico* pool of full-length genes. Random sequence fragments generated from this pool were analysed by BLASTX against the NCBI nr and COG databases with self-hits excluded. Most BNTS hits belonged to the same role category as the parent sequence. However, the shorter the sequence fragment, the greater was the chance that no hit was returned with the default cutoff limits for the programs ($E \leq 10$). For sequence fragments at least 65 nt long and with an E value of 0.01 or less and a similarity of at least 40% to their BNTS hits in the NCBI nr database, more than 70% were assigned to the correct functional role (Supplementary Fig. 7), and these values were used as annotation criteria. At these cutoffs, the phylogenetic affiliation derived from the taxonomic bins of the BNTS hit protein homologue in the NCBI nr queries had more than 90% accuracy at the phylum/subphylum level. On the basis of similar analyses, the cutoff criteria for protein prediction against the COG database were established as $E \leq 0.1$, similarity ≥ 40% and overlapping length ≥ 65 nt (23 amino-acid residues) to the corresponding best hit. The COG cutoff criteria were also applied to BLASTX analysis results for KEGG[34]-pathway and SEED-subsystem predictions[35], because of the similar sizes of these databases.

No hits were returned from the BLASTX analysis against nr for fragmented 16S rRNA sequence fragments or the BLASTN analysis against the RDPII for fragmented protein fragments using the programs' default setting ($E \leq 10$), which was less stringent than our criteria. Thus, protein-encoding sequences and 16S rRNA sequences were distinct enough for the probability of cross-category misprediction from short sequence fragments to be very low.

Pyrosequence annotation and statistical analysis. 16S rRNA genes and protein-encoding sequences were identified and taxonomically categorized on the basis of the established cutoffs. Sequences not meeting the criteria were not considered further. For selected carbon-cycle genes, only those candidate sequences returning the correct protein name in the top three hits were considered orthologues to the gene of interest.

A pairwise comparison was performed to compare the gene content of metagenomic data sets (with pooled replicates) with a bootstrap resampling program[36] with a sample size of 20,000 sequences for COG groups, 20,000 repeated samplings, and a significance level of $P \leq 0.05$. The bootstrapping analysis was repeated for KEGG pathways and SEED subsystems with a sample size of 10,000. A simulation analysis was run with the COG data to estimate the fold difference in sequence abundance between metagenomes required for statistical significance. COGs with abundances in the smaller library varying from 1 to 65 were tested for statistical significance when the larger library had 1.1–10-fold higher abundances. Results showed that differences of 1.5-fold could be detected for categories with at least 13 sequences in the smaller sample, although this number increased for categories with fewer assigned sequences (Supplementary Fig. 10).

A multivariate comparison was performed to compare the overall composition of metagenomic data sets (with independent replicates) using non-metric multidimensional scaling (MDS; Primer5; Primer-E Ltd). The square-root-transformed Bray–Curtis similarity matrix of COG assignments was normalized for size differences between the data sets. A randomized permutation method (ANOSIM) was used to assess statistical significance between treatments using the same data matrix as for MDS (Primer5). R values were calculated on the basis of the difference of mean ranks within and between treatments and reported on a scale of 0 (average similarity between and within treatments is the same) to 1 (replicates within treatments are more similar to each other than to any from different treatments).

16S rRNA gene clone libraries. 16S rRNA genes were amplified by PCR with a FailSafe PCR Premix selection kit (Epicentre) with 0.4 μM concentrations of the 27F and 1492R primers[37]. A touchdown PCR programme was used with the annealing temperature decreasing by 1 °C per cycle, followed by 15 cycles at 52 °C. In each cycle, denaturing (at 95 °C), annealing (at 62–52 °C), and extension (at 72 °C) were 40 s in duration. An initial 4-min denaturation and a final 7-min extension step were also included in the programme. After electrophoretic analysis, PCR amplicons were excised from ethidium bromide-stained 1% agarose gels and cleaned (QIAquick kit; Qiagen). Amplicons from replicate samples were pooled and cloned (TA cloning kit; Invitrogen) with the pCR 2.1 vector. Cloned 16S rRNA genes were sequenced on an ABI Prism 3100 genetic analyser (Applied Biosystems) with the 27F primer.

Sequences were edited with Sequencher 4.1 (Gene Codes Corporation) and checked for chimaeras and potential vector contamination with the Pintail program[38] and vector search of EMBL-EBI (http://www.ebi.ac.uk/blastall/vectors.html). Bacterial taxonomic identities were assigned as detailed above. For phylogenetic tree building, the distance matrix of partial sequences (about 650 bp) was based on Kimura's two-parameter calculations with the MEGA 3.1 package[39]. Operational taxonomic units (OTUs) for each clone library were defined by using the DOTUR program with furthest-neighbour clustering at a distance of 0.03 (http://www.plantpath.wisc.edu/fac/joh/DOTUR.html)[40]. The distance matrix of all OTUs (with each OTU represented by only one sequence) was prepared with the same method as that described above. Bootstrapping with 1,000 replicates was used to assign confidence levels to the nodes.

16S rRNA T-RFLP analysis. 16S rRNA sequences were amplified as above except that the 27F primer was labelled with carboxyfluorescein. PCR amplicons were digested for 4.5 h with *Cfo*I (Roche) at 37 °C and precipitated in ethanol. The restricted amplicons were resuspended in 12 μl of deionized formamide plus 0.7 μl of DNA-fragment length standard (Gene-Scan-2500 TAMRA; Applied Biosystems). Terminal restriction fragment lengths were determined on an ABI PRISM 310 genetic analyser (Applied Biosystems).

Genome equivalents and carbon-cycle gene abundance calculations. Homologues to four single-copy essential genes (*recA*, *atpD*, *gyrB* and *rpoB*) were identified in the compound-specific metagenomic data sets to estimate genome equivalents. To correct for the effect of gene length on the probability of sampling by short-read pyrosequencing, raw numbers were corrected by dividing the length of each gene (which ranged from 1,077 to 4,026 nt) by the length of *recA* (1,077 nt). Raw numbers of carbon-cycle homologues were similarly normalized to *recA* gene length and divided by the genome equivalents to estimate the per-cell gene frequency in each metagenome. This approach overestimates the per-cell frequency if more than one copy of a gene is present in a genome.

31. Urbach, E., Vergin, K. L. & Giovannoni, S. J. Immunochemical detection and isolation of DNA from metabolically active bacteria. *Appl. Environ. Microbiol.* **65,** 1207–1213 (1999).
32. Moran, M. A. *et al.* Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432,** 910–913 (2004).
33. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29,** 22–28 (2001).
34. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
35. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33,** 5691–5702 (2005).
36. Rodriguez-Brito, B., Rohwer, F. & Edwards, R. An application of statistics to comparative metagenomics. *BMC Bioinform.* **7,** 162 (2006).
37. Delong, E. F., Wickham, G. S. & Pace, N. R. Phylogenetic stains—ribosomal RNA-based probes for the identification of single cells. *Science* **243,** 1360–1363 (1989).
38. Ashelford, K. E. *et al.* At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71,** 7724–7736 (2005).
39. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5,** 150–163 (2004).
40. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71,** 1501–1506 (2005).