APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Apr. 2009, p. 2221–2229 Vol. 75, No. 7
0099-2240/09/$08.00+0 doi:10.1128/AEM.02118-08

Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome▽†

Erin J. Biers,[1]* Shulei Sun,[1] and Erinn C. Howard[2]§

Department of Marine Sciences, University of Georgia, Athens, Georgia 30602,[1] and Department of Microbiology, University of Georgia, Athens, Georgia 30602[2]

Received 12 September 2008/Accepted 2 February 2009

The *Sorcerer II* Global Ocean Sampling (GOS) sequencing effort has vastly expanded the landscape of metagenomics, providing an opportunity to study the genetic potential of surface ocean water bacterioplankton on a global scale. Here we describe the habitat-based microbial diversity, both taxon evenness and taxon richness, for each GOS site and estimate genome characteristics of a typical free-living, surface ocean water bacterium. While *Alphaproteobacteria* and particularly SAR11 dominate the 0.1- to 0.8-μm size fraction of surface ocean water bacteria (43% and 31%, respectively), the proportions of other taxa varied with ocean habitat type. Within each habitat type, lower-bound estimates of phylum richness ranged between 18 and 59 operational taxonomic units (OTUs). However, OTU richness was relatively low in the hypersaline lagoon community at every taxonomic level, and open-ocean communities had much more microdiversity than any other habitat. Based on the abundance of single-copy eubacterial genes from the same data set, we estimate that the genome of an average free-living surface ocean water bacterium (sized between 0.1 and 0.8 μm) contains ~1,019 genes and 1.8 copies of the 16S rRNA gene, suggesting that these bacteria have relatively streamlined genomes in comparison to those of cultured bacteria and bacteria from other habitats (e.g., soil or acid mine drainage).

Marine bacterioplankton drive global biogeochemical processes. Most of what is known about these bacterioplankton, however, is limited to PCR-based, culture-independent studies of environmental consortia (16, 39) or studies of cultured organisms. Alternatively, shotgun metagenomic methods provide culture- and PCR-independent tools to study the genetic potential of a given system. Metagenomic shotgun libraries have been constructed for several ocean environments, including deep-sea sediments (20), whale falls (50), depth profiles within oligotrophic waters (8), the Sargasso Sea (52), and other habitats (42). Recently, the massive sampling and sequencing effort of the *Sorcerer II* Global Ocean Sampling (GOS) expedition (41) has produced the largest shotgun metagenome for any oceanic habitat to date. With 6.3 billion bp sequenced, the GOS is virtually an inexhaustible resource for the genetic study of surface ocean water biogeochemistry and microbial ecology.

The initial GOS investigations provided an overview of intraribotype (41), protein family (57), and kinome (26) diversity in surface ocean waters, but the GOS database can be interrogated to address many other hypotheses. One application of this data set, the identification of microbial diversity, is a fundamental question in microbial ecology. The taxonomic makeup of bacterioplankton in GOS samples was provided previously by Rusch et al. (41). However, only the diversity of the most abundant ribotypes and major taxonomic groups was reported, with diversity either reported qualitatively or pooled across the entire sampling range. While subtype classification is a valuable tool for comparing certain microbial populations, biogeography- or habitat-driven questions demand a quantitative assessment of environmentally relevant ribotypes at each site in the GOS metagenome.

In this study, we provide a quantitative, categorical assessment of microbial diversity—both operational taxonomic unit (OTU) evenness and OTU richness—in surface waters of the northwest Atlantic Ocean through the eastern tropical Pacific Ocean by analyzing 16S rRNA gene homologs in the GOS metagenomic library. Based on 16S rRNA gene sequences, we compared the bacterial diversities between sampled habitats and assessed the extent of diversity that was sampled. By normalizing the number of 16S rRNA gene hits (to gene length and the number of single-copy gene hits), we also report estimates for the average number of ribosomal gene copies as well as the number of genes present in the average surface water bacterial genome.

MATERIALS AND METHODS

GOS sampling and sequencing. GOS sampling procedures, sample sites, and sequencing methods have been described previously (41, 57). In short, surface waters mainly from oceanic sites were collected and sequentially filtered through 20-μm-, 3-μm-, 0.8-μm-, and 0.1-μm-pore-size filters. DNA mainly from the 0.1- to 0.8-μm particle size fraction was extracted, randomly sheared, size selected, inserted into pBR322 plasmid vectors, and electroporated into *Escherichia coli*. Shotgun sequencing was performed from both ends of the insert, creating paired reads. Individual sequences had an average trimmed length of 822 bp (41).

Retrieval of gene homologs. All Basic Local Alignment Search Tool (BLAST) searches for gene homologs in the unassembled GOS metagenome were performed through the community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) interface (44). To maximize homolog retrieval, the entire 1,542-bp sequence of the 16S rRNA gene (*rrsE*, locus tag b4007) from the *E. coli* K-12 genome (RefSeq accession no. NC_000913) was

* Corresponding author. Present address: Department of Environmental Health Sciences, University of South Carolina, Columbia, SC 29208. Phone: (803) 777-4553. Fax: (803) 777-3391. E-mail: ejbiers@mailbox.sc.edu.

§ Present address: Department of Marine Sciences, University of Georgia, Athens, GA 30602.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 6 February 2009.

set as the query sequence. Nucleotide sequences for 16S rRNA homologs were identified with BLASTn against "GOS: all metagenomic sequence reads (N)" (cutoff expect [E] value of $\leq 10^{-5}$; overlap length of $\geq$65 bp; similarity of $\geq$85%). Before taxonomic assignment of 16S rRNA gene homologs, the retrieved GOS sequences were trimmed to exclude any non-16S rRNA gene portion of the sequence read. Paired-read sequences, two sequence reads originating from the same insert (one forward read, one reverse read), within the 16S rRNA gene homologs were identified and then screened to verify that they were assigned into the same taxonomic bin (see method below). Then, one of the paired reads was removed from further analysis.

Single-copy gene hits, hits to genes found once per prokaryotic genome, were retrieved previously (23). Briefly, amino acid sequences for single-copy genes (*atpD*, *gyrB*, *dnaK*, *rpoB*, *tufA*, and *recA*) were identified with BLASTp against "GOS: all ORF peptides (P)" (cutoff E-value of $<10^{-20}$). Hits for each single-copy gene, excluding paired reads, were normalized to the relative length of the gene compared to the length of *recA* ($h_g^* = L_{recA}/L_g \times h_g$), and the average number of *recA*-normalized single-copy gene hits was used in this analysis. In this equation, h_g^* is the number of *recA* size-normalized hits for gene *g*, L_{recA} is the length of *recA* in bp, L_g is the length of gene *g* in bp, and h_g is the number of hits for gene *g*. In this study, the lengths of the query genes were used (for *recA*, L_{recA} is 1,062 bp; for the 16S rRNA gene, L_g is 1,542 bp).

Taxonomic assignment. The Ribosomal Database Project II Classifier (RDP Classifier) (55), a naïve Bayesian tool based upon taxonomic classifications in *Bergey's Taxonomic Outline of the Prokaryotes* (15), was initially used to classify 16S rRNA gene homologs (data not shown). However, certain abundant marine bacteria, especially taxa without cultured representatives, were poorly classified by this method, as they are not represented in Bergey's classification system. Therefore, reference sequences from marine bacteria spanning the known diversity of marine microbes were compiled as reference sequences (see Table S1 in the supplemental material) (34) and analyzed using the RDP Classifier. Many of the marine reference sequences were classified correctly by RDP. However, a lack of taxonomic assignment was found for ecologically relevant marine groups, such as SAR11, the *Roseobacter* clade, SAR116, SAR324, SAR86, and SAR406, accounting for 48% of all retrieved hits (see Table S2 in the supplemental material). Therefore, taxonomic assignment of the GOS 16S rRNA gene homologs was made by similarity binning with the marine reference sequences using a Smith-Waterman alignment (47). If the GOS sequence had $\geq$70% overlap and $\geq$90% identity with a reference sequence, taxonomy was assigned to the level of order, where possible, according to the reference sequence with the highest similarity.

Distance-based OTU and richness comparison (DOTUR). The partial 16S rRNA gene sequences we retrieved included sequences that were taxonomically classifiable by our methods, but because all sequences did not span the same location of the 16S rRNA gene (average read length of 822 bp, average 16S rRNA gene length of ~1,500 bp), these sequences would align poorly and artificially enhance predicted diversity. Therefore, statistical assessments of observed and predicted diversity were performed on a subset of 16S rRNA homologs that was retrieved from within the 16S rRNA gene homolog pool by using methods similar to those of Schloss and Handelsman (43). The forward and reverse complements of a modified universal 16S rRNA gene probe (variable 3 [V3] region; EUB338, GCN GCC NCC CGT AGG NGT) (2, 6) were used as query sequences against the 16S rRNA gene homologs we had retrieved previously. For each matching sequence, 150 bp upstream and 150 bp downstream of the probe was obtained so that each retrieved sequence started and ended at the same approximate location. These sequences were aligned using Greengenes, a tool that aligns 16S rRNA gene sequences with a core set of 16S rRNA gene templates (9, 10), and any sequences that did not align were removed. Sequences were trimmed to the same aligned region (positions 1699 to 2257) to maximize the number of overlapping sequences. Any trimmed sequences less than 170 bp (~2/3 the length of the longest sequences, excluding gaps) were discarded to minimize effects of gaps and poor alignment on diversity assessment.

To evaluate the diversity captured in the GOS sequencing effort, we used the V3 region 16S rRNA gene sequences to create rarefaction curves for several GOS niches: the entire GOS metagenome, open-ocean habitats, coastal habitats, estuarine habitats, and a hypersaline lagoon. The open-ocean, coastal, and estuary habitats were chosen because they were the only habitats sampled multiple times (14, 18, and 3 times, respectively), and the hypersaline lagoon was chosen because it was the most highly sequenced sample (sample 33). Distance matrices for sequences within each of these GOS niches were calculated from alignments using DNAdist within BioEdit (19).

These matrices were analyzed with DOTUR (43) for distance-based OTU and richness determination. For estimates of species richness, we used Chao1—a nonparametric estimator based on mark-release-recapture techniques—which is designed to estimate the lower limit of species richness (5). While Foggo et al. (13) and Kemp and Aller (27) found that asymptotic Chao1 values were appropriate estimates of species richness for intermediate sampling frequencies, Chao1 underestimates diversity when sample sizes are small (25). Regardless, Chao1 performs like other estimators, both parametric and nonparametric, when used as a relative measure of diversity between samples (45) and is used here as a best estimate of the lower bound of species richness.

Distance matrices retrieved from DNAdist were additionally analyzed with LIBSHUFF (46), a statistical tool that examines the similarity between two 16S rRNA libraries. LIBSHUFF comparisons were made between estuarine, open-ocean, and coastal habitats.

Hit normalization. To reconcile the effects of gene size on hit retrieval, the number of hits for each gene was normalized to the relative length of the gene compared to the length of *recA* (22, 23, 58) by using the equation $h_g^* = L_{recA}/L_g \times h_g$ (variables are described above).

Estimating the number of genes per genome. To approximate the number of genes (*g*) found in the average genome (*G*) of prokaryotes captured in the GOS, we used the following calculation: $g/G = [(N_R \times L_R)/L_g^*]/N_G$, where N_R is the total number of reads, L_R is the average read length, L_g^* is the average gene length, and N_G is the total number of genomes. For N_R, we used the total number of nonpaired reads at each site. For L_R, we used the average GOS read length of 822 bp (41). For L_g^*, we used the average gene length of prokaryotes of 924 bp (56). For N_G, we assumed that the abundance of single-copy genes, genes found in nearly all bacteria with only one copy per genome, is equivalent to the number of genomes sequenced. We used the single-copy gene numbers found by Howard et al. (23) for sites GS02 to GS51 and the methods described by Howard et al. to compute the numbers of single-copy genes in sites GS00 to GS001. These numbers include only nonpaired reads.

To approximate the number of 16S rRNA genes found in the average genome (*G*) of prokaryotes captured in the GOS, we used the calculation $h_g/G = h_g^*/N_G$, where h_g is the number of hits for gene *g*, h_g^* is the size-normalized number of hits for gene *g*, and N_G is the total number of genomes. For h_g^*, we used the size-normalized number of nonpaired reads.

RESULTS

To retrieve 16S rRNA gene sequences from the GOS metagenome, we queried the unassembled GOS sequences with a full-length 16S rRNA gene sequence using a low E-value (10^{-5}). Less than 1% of the retrieved sequences did not meet our further criteria for identifying 16S rRNA genes ($\geq$70% length overlap with a set of reference 16S rRNA gene sequences) (see Table S1 in the supplemental material) (34) and were thus removed from this analysis. After duplicate reads (from paired reads) were removed, 10,025 partial 16S rRNA gene sequences were retrieved from the GOS database, making up 0.24% of the unassembled GOS sequence reads. This is similar to the frequency of 16S rRNA gene homologs retrieved from a coastal marine shotgun metagenomic library sequenced by pyrosequencing (0.19%) (34).

Ribotypes retrieved. Of the 10,025 partial 16S rRNA gene sequences retrieved from the GOS database, 87% were classified into taxonomic bins by Smith-Waterman alignment to a set of reference 16S rRNA gene sequences (see Table S1 in the supplemental material). Similarly, Rusch et al. (41) assigned 88% of retrieved 16S rRNA gene sequences to ribotypes available in public databases. However, fewer sequences were reported in that study (4,125 primary assemblies). While the present study queried unassembled sequence data, Rusch et al. (41) queried primary assemblies which merge mated pairs, as well as sequences that overlap at >98% identity.

Observed taxon distributions are shown by GOS habitat and by sample in Table 1 and Table S3 in the supplemental material, respectively, while biogeographies of taxa are shown by phylum and by alphaproteobacterial order in Fig. 1 and Fig. S1 in the supplemental material, respectively. The dominant ri-

TABLE 1. Ribotype diversity for habitats sampled multiple times by GOS[a]

Classification	Coastal habitat		Estuary habitat		Open-ocean habitat		Entire GOS metagenome	
	No. of hits	% of hits	No. of hits	% of hits	No. of hits	% of hits	No. of hits	% of hits
Archaea	14	**0.4**	0	**0.0**	8	**0.2**	32	**0.3**
Bacteria								
Actinobacteria	230	**7.3**	92	**23.8**	188	**5.0**	824	**8.2**
CFB	252	**8.0**	28	**7.3**	126	**3.3**	603	**6.0**
Chlamydiae/Verrucomicrobia	12	**0.4**	4	**1.0**	9	**0.2**	41	**0.4**
Chloroflexi		**0.1**		**0.8**		**0.1**		**0.2**
Unclassified	2	0.1	3	0.8	1	0.0	15	0.1
SAR202	0	0.0	0	0.0	1	0.0	1	0.0
Cyanobacteria	167	**5.3**	0	**0.0**	255	**6.7**	532	**5.3**
Deinococcus/Thermus	0	**0.0**	0	**0.0**	0	**0.0**	4	**0.0**
Fibrobacteres/Acidobacteria	6	**0.2**	0	**0.0**	3	**0.1**	33	**0.3**
Firmicutes	4	**0.1**	1	**0.3**	5	**0.1**	22	**0.2**
OP11	0	**0.0**	0	**0.0**	0	**0.0**	0	**0.0**
OD1	0	**0.0**	0	**0.0**	0	**0.0**	2	**0.0**
SR1	0	**0.0**	0	**0.0**	0	**0.0**	0	**0.0**
Planctomycetales	1	**0.0**	2	**0.5**	4	**0.1**	19	**0.2**
Alphaproteobacteria		**51.3**		**36.8**		**40.8**		**43.1**
Unclassified	128	4.1	8	2.1	162	4.3	410	4.1
Caulobacterales	2	0.1	0	0.0	2	0.1	7	0.1
Rhizobiales	26	0.8	2	0.5	13	0.3	91	0.9
Rhodobacterales	39	1.2	0	0.0	35	0.9	122	1.2
Rhodobacterales, *Roseobacter* clade	118	3.8	1	0.3	40	1.1	258	2.6
Rhodospirillales	13	0.4	1	0.3	27	0.7	54	0.5
Rickettsiales	0	0.0	0	0.0	0	0.0	1	0.0
Rickettsiales, SAR11 cluster	1,149	36.7	127	32.9	1,178	31.0	3,093	30.9
SAR116 cluster	130	4.1	3	0.8	89	2.3	275	2.7
Sphingomonadales	3	0.1	0	0.0	3	0.1	10	0.1
Betaproteobacteria		**0.9**		**11.1**		**5.2**		**2.8**
Unclassified	8	0.3	9	2.3	1	0.0	23	0.2
Burkholderiales	20	0.6	34	8.8	195	5.1	260	2.6
Deltaproteobacteria		**0.2**		**0.0**		**0.8**		**0.4**
Unclassified	0	0.0	0	0.0	1	0.0	2	0.0
Bdellovibrionales	0	0.0	0	0.0	1	0.0	2	0.0
Desulfobacterales	0	0.0	0	0.0	0	0.0	0	0.0
Desulfovibrionales	0	0.0	0	0.0	0	0.0	1	0.0
Desulfuromonadales	0	0.0	0	0.0	0	0.0	1	0.0
Myxococcales	0	0.0	0	0.0	0	0.0	0	0.0
SAR324	5	0.2	0	0.0	27	0.7	38	0.4
Syntrophobacterales	0	0.0	0	0.0	0	0.0	0	0.0
Epsilonproteobacteria		**0.0**		**0.0**		**0.0**		**0.1**
Unclassified	0	0.0	0	0.0	0	0.0	0	0.0
Campylobacterales	0	0.0	0	0.0	0	0.0	4	0.1
Sulfurovum	0	0.0	0	0.0	0	0.0	2	0.0
Gammaproteobacteria		**14.0**		**4.9**		**21.2**		**17.3**
Unclassified	140	4.5	2	0.5	196	5.2	562	5.6
Aeromonadales	3	0.1	0	0.0	16	0.4	19	0.2
Alteromonadales	17	0.5	1	0.3	249	6.6	290	2.9
Chromatiales	1	0.0	0	0.0	3	0.1	5	0.0
Enterobacteriales	2	0.1	0	0.0	4	0.1	6	0.1
Methylococcales	5	0.2	1	0.3	7	0.2	15	0.1
Nitrincola	6	0.2	0	0.0	11	0.3	38	0.4
Oceanospirillales	9	0.3	0	0.0	15	0.4	47	0.5
Pasteurellales	0	0.0	0	0.0	0	0.0	0	0.0
Pseudomonadales	18	0.6	2	0.5	27	0.7	67	0.7
SAR86 cluster	231	7.4	12	3.1	269	7.1	636	6.3
Thiotrichales	0	0.0	1	0.3	0	0.0	8	0.1

Continued on following page

TABLE 1—*Continued*

Classification	Coastal habitat		Estuary habitat		Open-ocean habitat		Entire GOS metagenome	
	No. of hits	% of hits	No. of hits	% of hits	No. of hits	% of hits	No. of hits	% of hits
Vibrionales	7	0.2	0	0.0	7	0.2	40	0.4
Xanthomonadales	1	0.0	0	0.0	0	0.0	5	0.0
SAR406	32	**1.0**	0	**0.0**	66	**1.7**	106	**1.1**
Spirochaetes	0	**0.0**	0	**0.0**	0	**0.0**	0	**0.0**
Unclassified	20	**0.6**	0	**0.0**	43	**1.1**	67	**0.7**
Eukaryotes	9	**0.3**	0	**0.0**	15	**0.4**	28	**0.3**
Unclassified at >90%	303	**9.7**	52	**13.5**	494	**13.0**	1,304	**13.0**
Total	3,133		386		3,794		10,025	

[a] Coastal habitats were sampled 18 times, estuary habitats 3 times, and open-ocean habitats 14 times. Ribotype diversity for individual samples is provided in Table S3 in the supplemental material. The number of hits listed for each taxonomic bin is the raw, unnormalized count after duplicate reads (paired-end reads) were removed. Numbers in bold are total percentages of hits that were binned into the indicated domain/phylum/clade.

botypes are consistent with other assessments of marine microbial diversity (8, 16, 17, 36). *Alphaproteobacteria* dominate the data set (43.1% of 16S rRNA gene sequences), and the majority of these are SAR11-like bacteria (30.9% of 16S rRNA gene sequences) (see Fig. S1 in the supplemental material). *Gammaproteobacteria* are the second most abundant taxon (17.3%), followed by *Actinobacteria* (8.2%), *Cytophaga-Flavobacterium-Bacteroides* (CFB) (6.0%), *Cyanobacteria* (5.3%), *Betaproteobacteria* (2.8%), and SAR406 (1.1%). Notably, *Archaea* made up only 0.3% of the total diversity in the GOS. These phylum-level abundances are generally equivalent to those found by Rusch et al. (41) (see Table S4 in the supplemental material). While this study found comparatively low proportions of *Archaea*, *Bacteroidetes*, and *Firmicutes*, this study was able to classify more *Proteobacteria*, as well as taxa not mentioned by Rusch et al. (41) (*Chloroflexi*, *Deinococcus/Thermus*, *Fibrobacteres/Acidobacteria*, SAR406, mitochondrion/plastid).

Some taxa are more prevalent in certain habitats, which can be seen both statistically (Fig. 2; also see Fig. S2 in the supplemental material) and in gross assessments of taxon richness (Table 1 and Fig. 1). Statistically, ribotype diversities were significantly different between each of the following habitat types: estuarine, coastal, and open ocean (LIBSHUFF, $P < 0.001$) (see Fig. S2 in the supplemental material). Looking at specific taxa (Table 1 and Fig. 1), *Actinobacteria* and *Betaproteobacteria* are more dominant in estuarine communities than in coastal or open–ocean communities. On the other hand, *Cyanobacteria* and *Gammaproteobacteria* are less dominant in estuarine communities than in coastal or open–ocean communities. *Actinobacteria*, CFB, the *Roseobacter* clade, SAR11, and SAR116 are more abundant in coastal than in open–ocean communities, while *Cyanobacteria* and SAR406 are more abundant in open–ocean than in coastal communities (Table 1 and Fig. 1; also see Fig. S1 in the supplemental material). *Burkholderiales* and *Alteromonadales* were also found to be more abundant in open–ocean than coastal communities; however, *Burkholderia cepacia* (*Burkholderiales*) and *Shewanella oneidensis* (*Alteromonadales*) are probable contaminants (7).

Ribotype diversity. Since the average GOS sequence read length is 822 bp (41) and the average 16S rRNA gene is ~1,500 bp, the partial 16S rRNA gene sequences we retrieved span different regions of the 16S rRNA gene. Analyzing those sequences would result in enhanced estimates of ribotype diversity. Therefore, from within our 10,025 partial 16S rRNA gene sequences, we obtained similarly sized sequences (~300 bp) surrounding a 16S rRNA gene probe (EUB338) targeting the V3 region for use in statistical analyses of ribotype richness (43). We identified 3,404 V3 region 16S rRNA gene sequences, equaling 34% of the hits we retrieved by the full-length 16S rRNA gene query.

The number of OTUs observed in GOS samples was approximated by rarefaction curves of the V3 region 16S rRNA gene sequences (see Fig. S3a to e in the supplemental material). For the purpose of this paper, we define a phylum as a group of organisms with ≤20% 16S rRNA gene sequence dissimilarity and a species as a group of organisms with ≤3% 16S rRNA gene sequence dissimilarity (1, 24, 48). Based on rarefaction and LIBSHUFF collection curves, ribotype diversity at the phylum level has been sampled reasonably within the major ocean habitats, but saturation has not been met for other taxonomic levels at 10%, 3%, and 0% sequence dissimilarity (see Fig. S2 and S3a to e in the supplemental material).

Estimates of taxon diversity and evenness are shown by GOS habitat and by sample in Fig. 2a and b and in Table S5 in the supplemental material, respectively. While the numbers of V3 16S rRNA sequences at many individual sites were too few to analyze statistically, the number of sequences in a given habitat allowed for a comparison of diversities between open-ocean, coastal, estuarine, and hypersaline environments. Regardless of OTU definition, the highest diversity was found in open-ocean habitats, followed by coastal habitats (Fig. 2A).

We also predicted the lower bound of microbial richness within sampled habitats by using Chao1 richness estimates (Fig. 2; also see Fig. S3f to j in the supplemental material). At the phylum level, we found that the lower bounds of richness in coastal, open-ocean, and estuarine habitats were approximately equal (Fig. 2; also see Fig. S3j in the supplemental

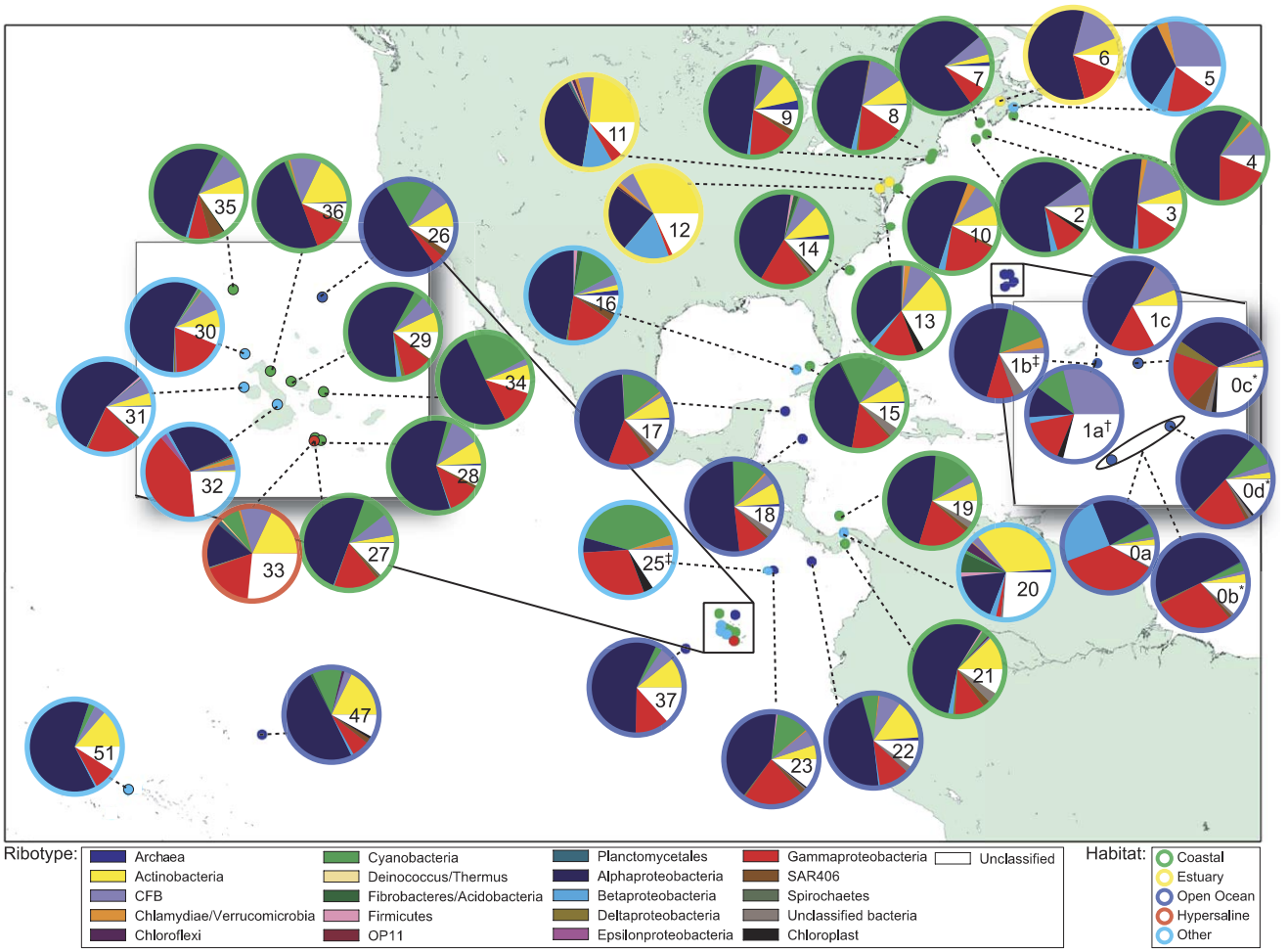

FIG. 1. Microbial diversity (16S rRNA) for all GOS samples. Dotted lines connect taxonomic data to sample locations (filled circles). Numbers within taxonomic plots indicate the sample identification numbers (preceded by "GS" as described by Rusch et al. [41]). The border color surrounding each taxonomic plot indicates the habitat type. The habitats classified as "Other" include the following: 5, embayment; 16, coastal sea; 20, fresh water; 25, fringing reef; 30, warm seep; 31, coastal upwelling; 32, mangrove; 51, coral reef atoll. All samples were collected from the 0.1- to 0.8-μm size fraction, unless indicated otherwise by the following symbols: ∗, from the 0.22- to 0.8-μm size fraction; †, from the 3- to 20-μm size fraction; and ‡, from the 0.8- to 3-μm size fraction.

material), with ~37 phyla in open-ocean and coastal habitats (95% confidence interval, 35 to 46). There was, however, potentially less phylum richness in the hypersaline lagoon habitat than in the open-ocean and coastal habitats (lower bound of richness, 23; 95% confidence interval, 22 to 34). A similar trend was found with the estimated number of species (Fig. 2; also see Fig. S3h in the supplemental material). In the open-ocean and coastal habitats, the estimated minimum number of species is ~420 (95% confidence interval, 336 to 565). The lowest bound of species richness was found in the hypersaline lagoon habitat (148 species; 95% confidence interval, 108 to 238). The minimum levels of richness of both phylum and species suggest that the hypersaline lagoon environment (GS33) contains a less complex bacterioplankton community, a conclusion also reached by Rusch et al. (41), based upon the relatively high degree of sequence assembly at this site. There is little difference in the predicted lower bounds of phylum and species numbers between open-ocean and coastal habitats.

When richness is based on unique sequences, however, the lower bounds of taxon richness within coastal and open-ocean communities differ. Populations of microbes in the open ocean have more unique sequences than those in all other habitats (Fig. 2; also see Fig. S3g in the supplemental material), suggesting more taxonomic microdiversity in the open ocean (45). Several major open-ocean taxa, such as SAR11 and *Prochlorococcus marinus*, have been shown to exhibit considerable intraclade diversity (12, 14, 31, 41, 53), so our observation of microdiversity within open-ocean bacteria is not as surprising as the relative lack within coastal habitats.

Average surface ocean water bacterium (0.1 to 0.8 μm). With this large metagenomic library, it is possible to approximate average genome characteristics for surface ocean water bacteria (in the 0.1- to 0.8-μm size fraction). In a previous study, Howard et al. (23) found the abundance of single-copy genes, genes found in nearly all bacteria with only one copy per genome, for each site in the GOS (see Table S6 in the supplemental material). Based on an average GOS read length of 822 bp (41) and an average gene length of 924 bp (56) and assum-

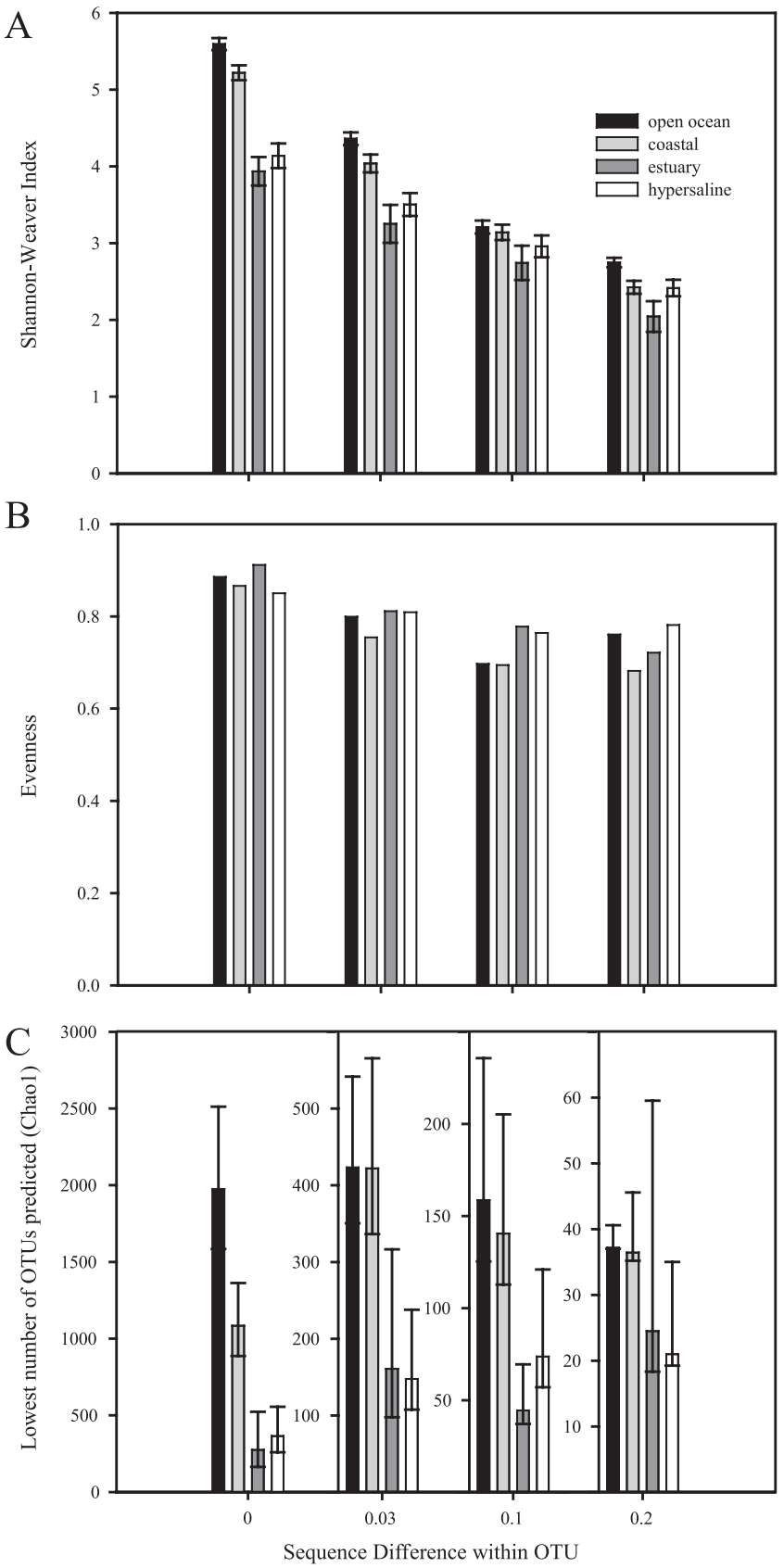

FIG. 2. Diversity, evenness, and richness indices for GOS habitats, as calculated by DOTUR. (A) Diversity indices were estimated using the Shannon-Weaver index of diversity. (B) Evenness scores were estimated from Shannon-Weaver indices of diversity as well as richness. (C) Lower bounds of richness were calculated using the Chao1 estimator. Error bars indicate 95% confidence intervals.

ing that the single-copy gene abundance is equivalent to the number of genome equivalents sequenced, we calculated the number of genes per genome at each GOS site $\{g/G = [(N_R \times L_R)/L_g^*]/N_G\}$ (see Table S6 in the supplemental material). By averaging across all sites, we estimate that the average genome of a marine surface water microorganism in the 0.1- to 0.8-μm size fraction is ~1 Mb and contains ~1,019 genes.

The scale of the GOS study also allows for an approximation of 16S rRNA gene abundance per bacterium ($h_g/G = h_g^*/N_G$). In this study, we found 10,025 16S rRNA gene hits within the GOS (6,904 when size normalized to *recA* by the equation $h_g^* = L_{recA}/L_g \times h_g$). Based on single-copy gene abundance, the average bacterium sequenced in the GOS (in the 0.1- to 0.8-μm size fraction) contains 1.8 16S rRNA genes, varying between 1.3 (GS00c, open ocean) and 2.8 (GS08, coastal). On average, coastal bacterioplankton genomes contain more 16S rRNA genes (1.9) than open-ocean or estuarine bacterioplankton (1.6 each).

DISCUSSION

Individual sequence reads (unassembled sequences) in metagenomic libraries are the fundamental unit of a metagenomic library. Scaffolds (assembled reads) are constructed by assembling multiple sequence reads with overlapping areas of highly similar sequences. Assembled reads provide longer sequences that are necessary for applications such as describing interribotype diversity (41), linking genes to phylogenetic markers (3, 4), discovering novel pharmaceuticals (21), or attempting to sequence the genome of an uncultured organism (41). However, there is an inherent nonquantitative aspect to assembled sequences and ambiguity when the assemblies are made across different environmental samples (57). In addition, an incomplete set of paired-end reads, where only some inserts were sequenced from both ends, can also bias results, rendering them semiquantitative. Therefore, our study of diversity and biogeography focused on the unassembled, nonpaired reads within the GOS metagenome.

Ribotype diversity. Our analyses agree with previous assessments of microbial diversity in surface ocean waters (16, 36) and further show that ribotype diversities are distinct between coastal, estuarine, and open-ocean habitats. However, because the current GOS metagenome focuses on the free-living (0.1- to 0.8-μm) size fraction, the diversity of larger, aggregate, or particle-bound microbes, such as those found in coastal sites, was likely underestimated (33, 58). This size bias can be seen by observing the microbial diversity in samples GS01a, GS01b, and GS01c (hydrostation S) from the Sargasso Sea pilot study (52), the only site to date for which sequences from three size fractions are available. The taxon composition in the 0.1- to 0.8-μm size fraction (GS01c) is typical of that of most open-ocean water samples in the GOS. However, even with a lower sequencing effort, major differences in represented taxa are evident for the larger size fractions (GS01a, GS01b). For example, within the *Alphaproteobacteria*, SAR11-like sequences dominate in the smaller size fraction, but the proportions of *Roseobacter clade*, SAR116-, and *Caulobacterales*-like sequences are more abundant in the larger (see Fig. S1 in the supplemental material). Unfortunately, the significantly smaller sequencing effort for those samples prevents a statistical comparison of levels of estimated diversity found in different size classes of the same water column. Even so, this illustrates that the genetic abundance and diversity retrieved in the GOS represent those of smaller, free-living, surface ocean water microbes and do not necessarily reflect the entire surface ocean water microbial gene pool. When the remaining size fractions are sequenced, the GOS metagenome will better reflect the entire microbial gene pool in surface ocean waters.

Typical surface ocean water bacterium. Assuming that there are, on average, 924 bp per gene, we estimate that the genome of a typical surface ocean water bacterium (sized 0.1 to 0.8 μm) is ~1 Mb and contains ~1,019 genes. Similarly, Raes et al. (38) estimated that a surface ocean water bacterium from within the Sargasso Sea sites (samples GS00 and GS01) has an effective genome size of 1.35 to 1.94 Mb. In contrast, a much larger genome size is estimated for sequenced, cultured, nonparasitic/symbiotic bacteria (average bacterium contains ~3,000 genes [18]) and for bacteria from soil, whale fall, and acid mine drainage environments (average bacterium genome contains 3.16 to 4.74 Mb [38]).

Pelagibacter ubique, a SAR11 clade member, contains 1,354 genes (18), remarkably close to the number of genes we found in a typical surface ocean water bacterium. This is not surprising considering that SAR11 clade members made up a significant fraction of the entire GOS 16S rRNA gene inventory (31%). However, the comparatively small genome of a surface ocean water bacterium (sized 0.1 to 0.8 μm) could indicate widespread genome streamlining, similar to that found in ubiquitous marine bacteria such as *Pelagibacter ubique* and *Prochlorococcus marinus* (11, 18, 40, 49). Genome streamlining is often found in conjunction with low G · C content (18, 51). A high G · C content requires more nitrogen during DNA synthesis, so in the ocean, where nitrogen is often limited, a low G · C content could provide a potential selective advantage. Indeed, in the GOS samples (sized 0.1 to 0.8 μm), where we found a surface ocean water bacterium's genome to be relatively small, Raes et al. (37) also found the G · C content to be relatively low (~35 to 45%).

In environmental metagenomic libraries such as the GOS, the presence of eukaryotic sequences may affect estimations of bacterial genome size (38). By increasing the number of sequence reads without adding bacterium-specific sequences containing genes such as *recA*, the presence of eukaryotic sequences tends to inflate estimations of bacterial genome size. Raes et al. (38) found that in the two Sargasso Sea sites where larger filtrate was sequenced (samples GS01a and GS01b; ≥0.8-μm size fractions), the effective genome size was 4.04 to 6.02 Mb, but after correction for bacterium-specific sequences, the effective genome size dropped to 1.71 to 1.94 Mb. We found the same result in our calculations, as the largest estimated genome sizes were found in GOS samples sequenced from size fractions larger than 0.8 μm (samples GS01a, GS01b, and GS25) (see Table S6 in the supplemental material). The presence of eukaryotic sequences in the larger size fractions as well as the presence of picoeukaryotic sequences in the 0.1- to 0.8-μm size fraction (35) indicates that our estimates of average bacterial genome size, while already relatively small, may be overestimates.

As further evidence for genome streamlining, regardless of habitat, the average surface ocean water bacterium in the 0.1-

to 0.8-μm size fraction contains only one to two rRNA operons. This is low in comparison to the average 3.9 rRNA operons found in cultured, sequenced prokaryotes (last checked on 14 August 2008) (29). A high number of rRNA operons is thought to allow a cell to respond quickly to transient improvements in environmental conditions (28, 30, 32, 54). This can be advantageous, especially in environments that receive pulses of nutrients. However, the tradeoff may be that maintaining multiple rRNA operons requires more energy (28). Our calculation that each genome contains one to two 16S rRNA genes implies that, while some cultured marine bacteria have a comparatively large number of rRNA operons per cell to quickly respond to the environment (e.g., *Roseobacter* clade members contain one to five operons per cell [32]), the average surface ocean water bacterium actually contains few, similarly to the GOS-dominant SAR11 group (*P. ubique* contains one operon per cell [18]). Based on our estimates of genome size and the number of 16S rRNA genes per genome, the average free-living marine bacterium may be constrained by energy conservation strategies more than by the need to adjust quickly to a changing environment.

ACKNOWLEDGMENTS

We thank J. Henriksen, R. S. Norman, W. Sheldon, X. Mou, P. Schloss, C. G. Fichot, and M. A. Moran for thoughtful discussions of the manuscript.

This research was supported by the National Science Foundation (OCE-0724017).

REFERENCES

1. **Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz.** 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature **430:**551–554.
2. **Amann, R. I., B. J. Binder, R. J. Olson, S. W. Chisholm, R. Devereux, and D. A. Stahl.** 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. Appl. Environ. Microbiol. **56:**1919–1925.
3. **Beja, O., L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong.** 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science **289:**1902–1906.
4. **Beja, O., M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong.** 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. **2:**516–529.
5. **Chao, A.** 1984. Nonparametric estimation of the number of classes in a population. Scand. J. Stat. **11:**265–270.
6. **Daims, H., A. Bruhl, R. Amann, K. H. Schleifer, and M. Wagner.** 1999. The domain-specific probe EUB338 is insufficient for the detection of all bacteria: development and evaluation of a more comprehensive probe set. Syst. Appl. Microbiol. **22:**434–444.
7. **DeLong, E. E.** 2005. Microbial community genomics in the ocean. Nat. Rev. Microbiol. **3:**459–469.
8. **DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl.** 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311:**496–503.
9. **DeSantis, T. Z., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. **34:**W394–W399.
10. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
11. **Dufresne, A., M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. T. de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf, and W. R. Hess.** 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc. Natl. Acad. Sci. USA **100:**10020–10025.
12. **Field, K. G., D. Gordon, T. Wright, M. Rappe, E. Urbach, K. Vergin, and S. J. Giovannoni.** 1997. Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. Appl. Environ. Microbiol. **63:**63–70.
13. **Foggo, A., M. J. Attrill, M. T. Frost, and A. A. Rowden.** 2003. Estimating marine species richness: an evaluation of six extrapolative techniques. Mar. Ecol. Prog. Ser. **248:**15–26.
14. **Fuhrman, J. A., and L. Campbell.** 1998. Marine ecology—microbial microdiversity. Nature **393:**410–411.
15. **Garrity, G. M., J. A. Bell, and T. G. Lilburn.** 2004. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, 2nd ed. Springer-Verlag, New York, NY.
16. **Giovannoni, S., and M. Rappè.** 2000. Evolution, diversity, and molecular ecology of marine prokaryotes, p. 47–84. *In* D. L. Kirchman (ed.), Microbial ecology of the oceans. Wiley-Liss, New York, NY.
17. **Giovannoni, S. J., and U. Stingl.** 2005. Molecular diversity and ecology of microbial plankton. Nature **437:**343–348.
18. **Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur.** 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science **309:**1242–1245.
19. **Hall, T.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41:**95–98.
20. **Hallam, S. J., N. Putnam, C. M. Preston, J. C. Detter, D. Rokhsar, P. M. Richardson, and E. F. DeLong.** 2004. Reverse methanogenesis: testing the hypothesis with environmental genomics. Science **305:**1457–1462.
21. **Hamann, M. T., R. Hill, and S. Roggo.** 2007. Marine natural products. Key advances to the practical application of this resource in drug development. Chimia **61:**313–321.
22. **Howard, E. C., J. R. Henriksen, A. Buchan, C. R. Reisch, H. Buergmann, R. Welsh, W. Y. Ye, J. M. Gonzalez, K. Mace, S. B. Joye, R. P. Kiene, W. B. Whitman, and M. A. Moran.** 2006. Bacterial taxa that limit sulfur flux from the ocean. Science **314:**649–652.
23. **Howard, E. C., S. Sun, E. J. Biers, and M. A. Moran.** 2008. Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. Environ. Microbiol. **10:**2397–2410.
24. **Huber, J. A., D. Mark Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97–100.
25. **Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan.** 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. Appl. Environ. Microbiol. **67:**4399–4406.
26. **Kannan, N., S. S. Taylor, Y. F. Zhai, J. C. Venter, and G. Manning.** 2007. Structural and functional diversity of the microbial kinome. PLoS Biol. **5:**467–478.
27. **Kemp, P. F., and J. Y. Aller.** 2004. Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? Limnol. Oceanogr. Methods **2:**114–125.
28. **Klappenbach, J. A., J. M. Dunbar, and T. M. Schmidt.** 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. **66:**1328–1333.
29. **Klappenbach, J. A., P. R. Saxman, J. R. Cole, and T. M. Schmidt.** 2001. RRNDB: the ribosomal RNA operon copy number database. Nucleic Acids Res. **29:**181–184.
30. **Krawiec, S., and M. Riley.** 1990. Organization of the bacterial chromosome. Microbiol. Rev. **54:**502–539.
31. **Moore, L. R., G. Rocap, and S. W. Chisholm.** 1998. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. Nature **393:**464–467.
32. **Moran, M. A., R. Belas, M. A. Schell, J. M. Gonzalez, F. Sun, S. Sun, B. J. Binder, J. Edmonds, W. Ye, B. Orcutt, E. C. Howard, C. Meile, W. Palefsky, A. Goesmann, Q. Ren, I. Paulsen, L. E. Ulrich, L. S. Thompson, E. Saunders, and A. Buchan.** 2007. Ecological genomics of marine roseobacters. Appl. Environ. Microbiol. **73:**4559–4569.
33. **Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Y. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. H. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward.** 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature **432:**910–913.
34. **Mou, X. Z., S. L. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran.** 2008. Bacterial carbon processing by generalist species in the coastal ocean. Nature **451:**708–711.
35. **Piganeau, G., Y. Desdevises, E. Derelle, and H. Moreau.** 2008. Picoeukaryotic sequences in the Sargasso Sea metagenome. Genome Biol. **9:**11.
36. **Pommier, T., B. Canback, L. Riemann, K. H. Bostrom, K. Simu, P. Lundberg, A. Tunlid, and A. Hagstrom.** 2007. Global patterns of diversity and community structure in marine bacterioplankton. Mol. Ecol. **16:**867–880.
37. **Raes, J., K. U. Foerstner, and P. Bork.** 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. Curr. Opin. Microbiol. **10:**490–498.

38. **Raes, J., J. O. Korbel, M. J. Lercher, C. von Mering, and P. Bork.** 2007. Prediction of effective genome size in metagenomic samples. Genome Biol. **8:**11.
39. **Rappè, M. S., P. F. Kemp, and S. J. Giovannoni.** 1997. Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. Limnol. Oceanogr. **42:**811–826.
40. **Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm.** 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature **424:**1042–1047.
41. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Y. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. **5:**398–431.
42. **Sabehi, G., O. Beja, M. T. Suzuki, C. M. Preston, and E. F. DeLong.** 2004. Different SAR86 subgroups harbour divergent proteorhodopsins. Environ. Microbiol. **6:**903–910.
43. **Schloss, P. D., and J. Handelsman.** 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. **71:**1501–1506.
44. **Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier.** 2007. CAMERA: a community resource for metagenomics. PLoS Biol. **5:**394–397.
45. **Shaw, A. K., A. L. Halpern, K. Beeson, B. Tran, J. C. Venter, and J. B. H. Martiny.** 2008. It's all relative: ranking the diversity of aquatic bacterial communities. Environ. Microbiol. **10:**2200–2210.
46. **Singleton, D. R., M. A. Furlong, S. L. Rathbun, and W. B. Whitman.** 2001. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. Appl. Environ. Microbiol. **67:**4374–4376.
47. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197.
48. **Stackebrandt, E., and B. M. Goebel.** 1994. A place for DNA-DNA reassociation and 16S ribosomal-RNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. **44:**846–849.
49. **Strehl, B., J. Holtzendorff, F. Partensky, and W. R. Hess.** 1999. A small and compact genome in the marine cyanobacterium Prochlorococcus marinus CCMP 1375: lack of an intron in the gene for tRNA(Leu)(UAA) and a single copy of the rRNA operon. FEMS Microbiol. Lett. **181:**261–266.
50. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
51. **Ussery, D. W., and P. F. Hallin.** 2004. Genome update: AT content in sequenced prokaryotic genomes. Microbiology **150:**749–752.
52. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Y. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
53. **Vergin, K. L., H. J. Tripp, L. J. Wilhelm, D. R. Denver, M. S. Rappe, and S. J. Giovannoni.** 2007. High intraspecific recombination rate in a native population of Candidatus Pelagibacter ubique (SAR11). Environ. Microbiol. **9:**2430–2440.
54. **von Wintzingerode, F., U. B. Gobel, and E. Stackebrandt.** 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol. Rev. **21:**213–229.
55. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.
56. **Xu, L., H. Chen, X. H. Hu, R. M. Zhang, Z. Zhang, and Z. W. Luo.** 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol. Biol. Evol. **23:**1107–1108.
57. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Z. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Y. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. F. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. **5:**432–466.
58. **Yutin, N., M. T. Suzuki, H. Teeling, M. Weber, J. C. Venter, D. B. Rusch, and O. Béjà.** 2007. 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.