APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Jan. 2010, p. 609–617 Vol. 76, No. 2
0099-2240/10/$12.00 doi:10.1128/AEM.01258-09

Deep Sequencing of a Dimethylsulfoniopropionate-Degrading Gene (*dmdA*) by Using PCR Primer Pairs Designed on the Basis of Marine Metagenomic Data▽†

Vanessa A. Varaljay,[1] Erinn C. Howard,[2] Shulei Sun,[2]‡ and Mary Ann Moran[2]*

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

Received 1 June 2009/Accepted 16 November 2009

***In silico* design and testing of environmental primer pairs with metagenomic data are beneficial for capturing a greater proportion of the natural sequence heterogeneity in microbial functional genes, as well as for understanding limitations of existing primer sets that were designed from more restricted sequence data. PCR primer pairs targeting 10 environmental clades and subclades of the dimethylsulfoniopropionate (DMSP) demethylase protein, DmdA, were designed using an iterative bioinformatic approach that took advantage of thousands of *dmdA* sequences captured in marine metagenomic data sets. Using the bioinformatically optimized primers, *dmdA* genes were amplified from composite free-living coastal bacterioplankton DNA (from 38 samples over 5 years and two locations) and sequenced using 454 technology. An average of 6,400 amplicons per primer pair represented more than 700 clusters of environmental *dmdA* sequences across all primers, with clusters defined conservatively at >90% nucleotide sequence identity (~95% amino acid identity). Degenerate and inosine-based primers did not perform better than specific primer pairs in determining *dmdA* richness and sometimes captured a lower degree of richness of sequences from the same DNA sample. A comparison of *dmdA* sequences in free-living versus particle-associated bacteria in southeastern U.S. coastal waters showed that sequence richness in some *dmdA* subgroups differed significantly between size fractions, though most gene clusters were shared (52 to 91%) and most sequences were affiliated with the shared clusters (~90%). The availability of metagenomic sequence data has significantly enhanced the design of quantitative PCR primer pairs for this key functional gene, providing robust access to the capabilities and activities of DMSP demethylating bacteria *in situ*.**

Dimethylsulfoniopropionate (DMSP) is an abundant organic sulfur compound produced by marine phytoplankton as an osmolyte and for antioxidant purposes (5, 19, 27, 34, 36, 38). Upon cell lysis, DMSP and its degradation products are released into the surrounding seawater, thus providing bacterial communities with reduced organic carbon and sulfur (20) as well as contributing significantly to ocean-atmosphere sulfur flux (1, 24). Marine organisms capable of DMSP degradation can use either of two environmentally significant pathways. One route, known as the cleavage pathway, can lead to degassing of DMSP-derived sulfur from surface waters in the form of dimethylsulfide (DMS), an important catalyst in cloud formation. The second, a bacterium-specific route known as the demethylation pathway, results in DMSP-derived sulfur compounds (such as methylmercaptopropionate [MMPA] and methanethiol [MeSH]) that typically remain within the marine microbial food web. Studies show that certain groups of bacteria can mediate either or both competing pathways (11, 35), although the predominant route of DMSP degradation is through demethylation (18, 20, 21). Significant biogeochemical data for bacterially mediated DMSP flux are now available (21, 33) and have allowed us to establish a framework for understanding this process in the marine environment (32). Yet the underlying genetic basis by which bacterioplankton perform and regulate these globally important sulfur transformations is relatively unknown.

The identification of *dmdA* (15), the gene encoding a DMSP demethylase that mediates the first step in the demethylation pathway, provides a key genetic tool for understanding the fate of DMSP in ocean waters. *dmdA* is highly abundant in marine metagenomic data sets, with thousands of homologs (15, 16) identified in the Global Ocean Survey (GOS) Sargasso Sea (37) and 2007 GOS (29) data sets. These findings indicate an important ecological role for *dmdA* in natural bacterioplankton communities. Two pressing areas for gene-based research include characterizing the diversity, abundance, and distribution of demethylating bacteria in the marine environment and determining how bacterial communities regulate DMSP fate via demethylation.

Here we describe our strategy for designing and testing *dmdA* primers to study the diversity of DMSP demethylating bacterial genes in marine environments. We took advantage of the non-PCR-amplified *dmdA* homolog sequence reads identified in the 2007 GOS release to design universal and clade-specific primer pairs for *dmdA* sequences. An *in silico* primer-testing pipeline checked specificity against metagenomic reads and identified mismatches to iteratively improve primer design. Primer pairs were tested empirically on free-living bacterial communities in nearshore waters of Sapelo Island, GA,

* Corresponding author. Mailing address: Department of Marine Sciences, University of Georgia, Athens, GA 30602. Phone: (706) 542-6481. Fax: (706) 542-5888. E-mail: mmoran@uga.edu.

‡ Present address: Center for Research in Biological Systems, University of California San Diego, 9500 Gilman Drive #0446, La Jolla, CA 92093-0446.

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

▽ Published ahead of print on 30 November 2009.

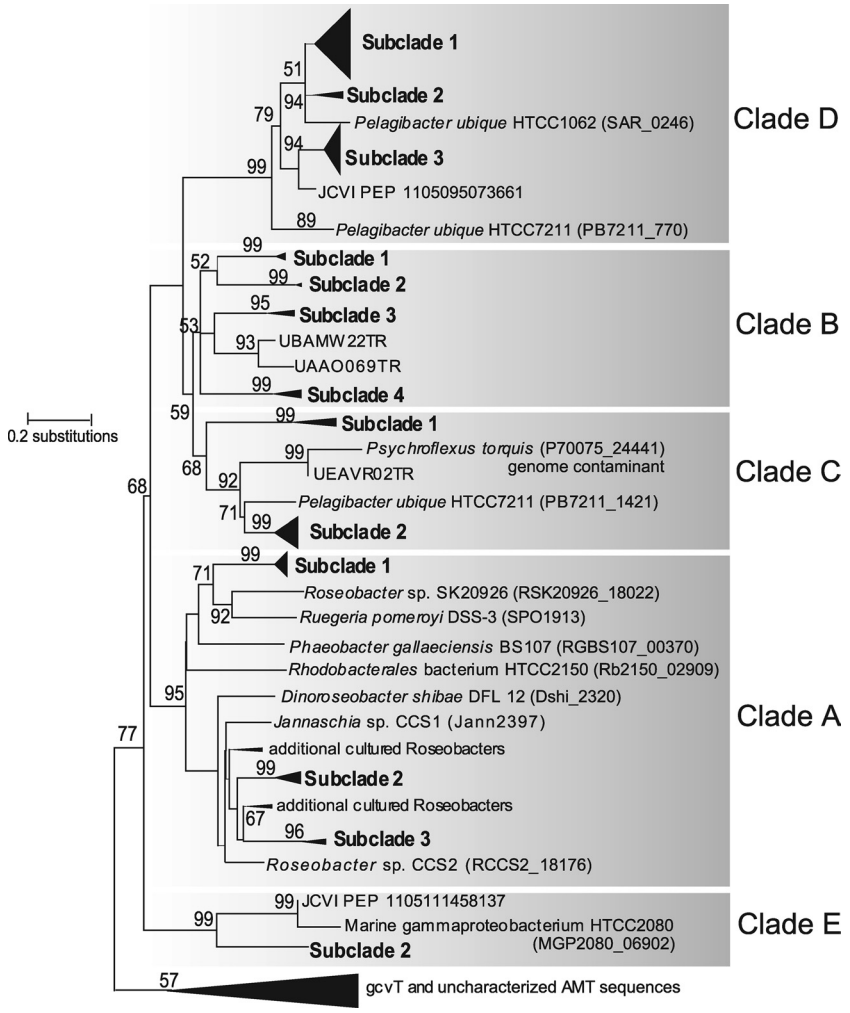

FIG. 1. Amino acid tree of representative GOS DmdA sequences. The wedge size is approximately proportional to the number of sequences within the group. Selected DmdA homologs from cultured marine bacteria are included. "Additional cultured Roseobacters" includes *Roseobacter denitrificans* Och114, *Roseobacter* sp. Azwk3b, *Roseobacter* sp. MED193, *Roseovarius* sp. 217, *Roseovarius nubinhibens* ISM, *Roseovarius* sp. TM1035, and *Ruegeria* sp. TM1040. Related glycine cleavage T (gcvT) and aminomethyltransferase (AMT) sequences serve as outgroups. Bootstrap values of <50 have been removed for clarity. The neighbor-joining tree was made with Jones-Taylor-Thornton distances. The exact position of the cluster designated clade C/1 can vary depending on the sequences included in the tree (data not shown).

using pyrosequencing to examine the deep diversity of *dmdA* amplicons. Selected primer pairs were then used to compare *dmdA* richness in gene reservoirs of the free-living and particle-associated communities.

MATERIALS AND METHODS

Design of *dmdA* primer pairs. Metagenomic reads used in *dmdA* primer design were obtained from the Global Ocean Sampling (GOS) metagenome (29), with *dmdA* homologs in each of the five major clades (A through E) (Fig. 1) identified as previously described (16). DmdA sequences that were not in one of the major clades (11% of 1,701 total sequences) were labeled as unclassified. These were used in primer design for the universal primer but not for the clade or subclade primers. To identify subclades, the nucleotide sequences from the five major clades were clustered using MEGA version 3.1 (pairwise alignment, Jukes-Cantor algorithm, neighbor-joining model, 100 bootstrap replicates) (16) or Geneious Pro 3.5.6 (9) Tree-Builder (Tree global alignment: cost matrix 65% similarity [5.0/−4.0], gap open penalty 12, and gap extension penalty 3, with Jukes-Cantor algorithm, neighbor-joining model). Glycine cleavage T protein (*gcvT*) and related aminomethyl transferase (AMT) sequences served as outgroups. Subclades were defined as sequence clusters with bootstrap values of ≥50% which captured at least 10% of the reads in a clade. However, not all subclades had conserved regions appropriate for primer design, and the ones that did not could not be considered further (see below).

Subsets of *dmdA* nucleotide sequences were globally aligned with BioEdit sequence alignment editor (14) and Geneious Pro 3.5.6 (9) programs using ClustalW. Primers were either designed manually or with the aid of Beacon Designer (Premier Biosoft International, Palo Alto, CA) primer design software. Primer pairs were designed to target amplicons without degeneracies ("specific" primer pairs) or included degenerate or inosine (a nucleoside that pairs indiscriminately) bases ("degenerate" and "inosine" versions) to accommodate common mismatches between primers and GOS reads that emerged from *in silico* testing (see below).

Bioinformatic pipeline: *in silico* primer tests. All primer pairs were iteratively tested *in silico* for specificity against the 1,701 *dmdA* sequences from the 2007 GOS release (see Fig. S1 in the supplemental material). Each GOS *dmdA* read was aligned to the *dmdA* gene from *Ruegeria pomeroyi* DSS-3 (SPO1913; 1,095 bp) to determine whether it contained the full region targeted by a given primer pair. Those that did (designated "reads in range") were used for primer testing; those that did not were excluded. Primer pair specificity was then quantitatively assessed against GOS reads using an exact sequence and pattern (ESP) search program (http://web.chemistry.gatech.edu/~doyle/espsearch/) to determine the

percentage of reads successfully targeted by the primer pair. Sequences with mismatches were mined for number, location, and base of the mismatch. As a quality control check, the pipeline also determined if primers would bind nonspecifically to sequences in nontarget *dmdA* clades (including unclassified *dmdA* sequences).

A separate *in silico* test of nonspecific binding of primers was also carried out against GOS metagenomic reads from three southeastern U.S. coastal sites (JCVI sites GS13, GS14, and GS15 [29]). All *dmdA* sequence reads were removed from these samples, and the remaining 394,170 reads were queried, allowing up to six total mismatches for forward-plus-reverse primers.

Primer pairs were either accepted or rejected based on results of the *in silico* testing and, if rejected, were iteratively redesigned. Degenerate and inosine bases were incorporated into some finalized primer pairs if there were common mismatches, especially at a position away from the 5′ end.

DNA samples. Surface water was collected between October 2000 and April 2005 at two sampling sites at the Sapelo Island Microbial Observatory (SIMO) (http://simo.marsci.uga.edu) in coastal Georgia. The Dean Creek site is a salt marsh tidal creek, and the Doboy Sound site is a coastal ocean inlet. To obtain each DNA sample, approximately 20 liters of water was filtered sequentially through 8.0-μm-, 1.0-μm-, and 0.2-μm-pore-size polycarbonate membrane filters (Poretics Corp., Livermore, CA), with two replicate samples obtained at each location on each date. Cells captured on the 1.0-μm filter (particle associated) and the 0.2-μm filter (free living) were stored at −20°C until DNA extraction with the PowerMax Soil DNA Isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA). A total of 76 DNA extracts, representing 38 samples of each size fraction (free living, 0.2 to 1.0 μm; particle associated, 1.0 to 8.0 μm), were used in this study (see Table S1 in the supplemental material). These samples were separately pooled by size fraction in equal amounts to produce composite free-living and particle-associated DNA samples. Each composite sample encompassed temporal (seasonal/yearly) and spatial (tidal creek and coastal sound) variability at the SIMO site.

PCR amplicon preparation and sequencing. Primer pairs giving single amplicons of the correct size from the composite SIMO DNA were chosen for analysis by sequencing. Amplicons suitable for 454 sequencing were prepared by modifying each primer pair with an adaptor sequence at the 5′ end of the forward primer according to the method of Huber et al. (17). Additional four-base key sequences in between the adaptor and primer sequence were used to distinguish inosine and degenerate primer sequences.

The typical PCR mix consisted of 0.5 U of Invitrogen (Carlsbad, CA) high-fidelity platinum *Taq* polymerase, 0.2 mM deoxynucleoside triphosphates (dNTPs), and 2 mM $MgSO_4$, although modifications of the $MgSO_4$ concentrations were used for some primer pairs. Primer concentrations ranged from 0.2 to 0.8 μM in final concentrations in a 25.0-μl reaction volume. PCR conditions were as follows: initial denaturing at 94°C for 2.0 min, 30 to 40 cycles of denaturing at 94°C for 20 s, annealing at various temperatures (Table 1) for 30 s, extension at 68°C for 30 s, and a final extension at 68°C for 5.0 min. All PCRs were carried out in duplicate using 24 ng template DNA and then pooled before sequencing. For the clade C/2 inosine primer pair, four PCRs were pooled because of low amplicon abundance. Pooled products were cleaned (QIAquick PCR purification kit; Qiagen, Valencia, CA) and stored at −20°C; for some products, an additional gel excision step was included (QIAquick gel extraction kit; Qiagen, Valencia, CA). Amplicons were cleaned using the AMPure purification method (Agencourt Bioscience Corp., Beverly, MA) according to the 454 Life Sciences protocol (Roche Diagnostics Corp., Branford, CT), with modifications to the volume of purified PCR products (30.0 μl) and AMPure beads (50.4 μl). Products were quantified spectrophotometrically and combined in equal concentrations in four separate pools based on primer and size fraction. Four-region 454 FLX LR70 sequencing was carried out at the University of South Carolina EnGenCore facility.

Clustering and clade designations. After removal of low-quality reads (quality score, <20; ≤0.03% of sequences), primer sequences were stripped from the remaining 252,319 reads. For the universal primer pair, sequence data were obtained for the first ~250 bases. For the other primer pairs, the full amplicon was sequenced. Within a primer pair (including specific, degenerate, and inosine versions when applicable) sequences were clustered based on 90% nucleotide identity (Cd-hit clustering [23]). Given an error rate for 454 sequencing of 0.3% (25), sequencing errors should not change cluster assignments, but would inflate estimates of unique sequences.

Amplicon sequences were annotated by BLASTx analysis using a default maximum E value of 10 against an in-house database which consisted of DmdA and related non-DmdA sequences from the GOS metagenome and cultured organisms. This analysis was used to distinguish correct target sequences from closely related paralogous sequences and to classify amplicons by clade. The high E value cutoff reflected the short length of the query sequences (e.g., 39 bp for the clade D/1 amplicons after primer sequences were stripped), but most hits had percent similarities of >90%. The BLAST database consisted of 3,280 total protein sequences (assembled from the Sargasso Sea GOS data set, the 2007 GOS data set, the Indian Ocean GOS data set, and cultured organisms; see references 15 and 16 and http://camera.calit2.net), including sequences from clade A (n =146), clade B (n =76), clade C (n =407), clade D (n =1,792), and clade E (n =19), as well as unclassified DmdA sequences (n =217), and nontarget gcvT and aminomethyltransferase sequences (n =623). Of the 3,280 sequences in the database, ~20 were DmdA sequences from cultured organisms.

Richness and shared sequence analyses. To account for differences in the number of amplicons sequenced for each primer pair (ranging from 2,000 to 12,000 sequences), a resampling approach was used in which 1,000 sample populations of the same size were randomly drawn from the amplicon pools being compared. This approach was used to normalize the number of 90% *dmdA* clusters in comparisons between primer pairs and size fractions. Statistical significance was assigned based on the distribution of pairwise differences between the 1,000 random populations using a 95% confidence interval (12). Rarefaction curves for a primer pair was based on 90% sequence clusters using EcoSim 7.0 (13) with 1,000 resamplings.

Nucleotide sequence accession number. The nucleotide sequences of *dmdA* 454-sequenced PCR amplicons were deposited in the GenBank Short Read Archive (SRA) under the accession number SRA008804.8.

RESULTS

***In silico dmdA* primer design.** The 1,701 *dmdA* sequences identified from the 2007 GOS metagenome (16) served as the database for designing hierarchical PCR primer pairs for the DMSP demethylase gene (Fig. 1). Primer design efforts focused on a universal primer pair, to capture as many *dmdA* sequences as possible from marine environmental samples, as well as on clade and subclade primer pairs to capture conserved sequence subsets within the five known clades of *dmdA*. Multiple alignments of a subset of target sequences (up to 50) were used for initial primer design. We avoided AT-rich regions (particularly problematic for clades C and D), long nucleotide repeats, sequences that might lead to primer dimers, and regions with high degrees of similarity to glycine cleavage T genes or other related non-*dmdA* genes. Primer pairs were tested *in silico* against the remaining sequences, followed by design optimization to complement the greatest number of identified *dmdA* sequences. The pipeline (see Fig. S1 in the supplemental material) simultaneously checked for matches to nontarget sequences, including sequences in the incorrect *dmdA* clade or subclade, or sequences of paralogous genes (i.e., *gcvT* and related aminomethyltransferases; Fig. 1).

While the original goal was to design all primers for use in quantitative PCR (qPCR), sufficiently conserved primer areas flanking a small (≤250-bp) region of the gene could not be identified for a universal primer pair. However, a universal *dmdA* primer pair amplifying a larger region (537 bp) from sequences in all five protein clades and targeting ≥90% of 2007 GOS *dmdA* reads in range, with ≤2 mismatches per primer when degeneracies were included, was identified (Table 1).

A clade-specific qPCR primer pair was designed for clade D; clades A, B, C, and E were highly diverse at the nucleotide level and primers were targeted instead to the abundant subclades (Table 1 and Fig. 1). Although smaller subsets of diverse sequences were not considered in primer design with this approach, they accounted for only ~20% of the 1,701 GOS *dmdA* sequences. In order to accommodate as many sequences as possible, clade and subclade primer pairs were designed with-

TABLE 1. Eighteen *dmdA* primer pairs (including degenerate and inosine versions) targeting 10 sequence groups and results of *in silico* testing against the 2007 GOS data set

Primer name	Primer version	*dmdA* position[a]	Amplicon length (bp)	Primer sequence[b]	Annealing temp (°C)	No. of target GOS reads	No. of target GOS reads in range[d]	No. (%) of reads in range binding primers	
								≤4 mismatches	≤6 mismatches
dmdAU	ND[e]	157–694	537	dmdAUF160: GTICARITITGGGAYGT dmdAUR697: TCIATICKITCIATIAIRTTDGG	32 and 41[c]	1,701	1,041, 1,093	993, 991 (93)	ND
A/1	Specific	368–596	228	A/1-spFP: ATGGTGATTTGCTTCAGTTTCT A/1-spRP: CCCTGCTTTGACCAACC	53	30	16	13 (81)	15 (94)
A/2	Specific	339–486	147	A/2-spFP: CGATGAACATTGGTGGGTTTCTA A/2-spRP: GCCATTAGGTCGTCTGATTTTGG	59	16	10	4 (40)	7 (70)
	Degenerate	339–486	147	A/2-dgFP: YGATGAWCATTGGTGGGTTTCKA A/2-dgRP: GCCATYARGTCGTCYGATTTTGG	58	16	10	8 (80)	9 (90)
	Inosine	339–486	147	A/2-inoFP: IGATGAICATTGGTGGGTTTCIA A/2-inoRP: GCCATIAIGTCGTCIGATTTTGG	57	16	10	8 (80)	9 (90)
B/3	Specific	169–323	154	B/3-spFP: GATGTCTCCTGCCAACGTCAGG TCGA B/3-spRP: ACCGGGTCATTGATCATGCCTGCG	62	4	3	3 (100)	3 (100)
B/4	Specific	361–553	192	B/4-spFP: ATTGCCGACTCGGATGTTCT B/4-spRP: CAAGAAGGTCAAACATGGCAAAC	58	5	4	4 (100)	4 (100)
C/2	Specific	291–482	191	C/2-spFP: AGATGAAAATGCTGGAATGATA AATG C/2-spRP: AAATCTTCAGACTTTGGACCTTG	50	141	94	19 (20)	44 (47)
	Degenerate	291–482	191	C/2-dgFP: AGATGAAAATGCWGGRATGATA AATG C/2-dg RP: AAWTCTTCAGAYTTTGGACCTTG	52	141	94	44 (47)	55 (60)
	Inosine	291–482	191	C/2-inoFP: AGATGAAAATGCIGGIATGATA AATG C/2-inoRP: AAITCTTCAGAITTTGGACCTTG	52	141	94	44 (47)	57 (61)
D/1	Specific	268–356	89	D/1-spFP: AGATGTTATTATTGTCCAATAATT GATG D/1-spRP: ATCCACCATCTATCTTCAGCTA	49	402	268	110 (41)	189 (71)
D/3	Specific	347–473	126	D/3-spFP: AATGGTGGATTTCTATTGCAG ATAC D/3-spRP: GATTTTGGACCTTGTACAGCCA	54	262	155	94 (61)	116 (75)
	Degenerate	347–473	126	D/3-dgFP: AATGGTGGRTTTCTATTGCWG ATWC D/3-dgRP: GATTTWGGMCCTTGYACAGCCA	56	262	155	113 (73)	137 (88)
D/all	Specific	984–1089	105	D/all–spFP: TATTGGTATAGCTATGAT D/all–spRP: TAAATAAAAGGTAAATCGC	42	1,125	457	190 (42)	320 (70)
	Degenerate	984–1089	105	D/all-dgFP: TATTGGTATWGCWATGAT D/all-dgRP: TAAATRAAAGGYAAATCGC	41	1,125	457	324 (71)	394 (86)
	Inosine	984–1089	105	D/all-inoFP: TATTGGTATIGCIATGAT D/all-inoRP: TAAATIAAAGGIAAATCGC	48	1,125	457	346 (76)	417 (91)
E/2	Specific	154–287	133	E/2-spFP: CATGTTCAGATCTGGGACGT E/2-spRP: AGCGGCACATACATGCACT	57	4	2	2 (100)	2 (100)
	Degenerate	154–287	133	E/2-dgFP: CATGTTCAGATMTGGGAYGT E/2-dgRP: AGCGGCAYATACATGCACT	56	4	2	2 (100)	2 (100)

[a] Position numbers based on the full-length *dmdA* sequence in *Ruegeria pomeroyi* DSS-3 (SPO1913).
[b] Degenerate codes are as follows: R, A or G; Y, C or T; W, A or T; M, A or C; K, G or T.
[c] Two annealing temperatures were used in separate PCRs.
[d] "Reads in range" refers to sequences that span the full region between the forward and reverse primers, allowing both to be tested for complementarity. In the case of the universal primer pair, the larger amplicon size required that the forward and reverse primers be tested with different subsets of reads, resulting in different numbers of reads in range for each primer.
[e] ND, not done.

out degeneracies ("specific" primer pairs) or included degenerate or inosine (a nucleoside that pairs indiscriminately) bases ("degenerate" and "inosine" versions) to accommodate common mismatches. When primer design was completed, the clade and subclade primer pairs targeted an average of 70% (with ≤4 mismatches) or 80% (with ≤6 mismatches) of *dmdA* reads in the correct clade (Table 1; see Table S2 in the supplemental material), although the success rate was as low as 20% for one primer pair. Preliminary subclade C/1 and D/2 primers targeted few sequences based on results of the bioinformatic analyses and were not considered further.

An *in silico* check for nonspecific primer binding was carried out against non-*dmdA* metagenomic reads from three coastal sites in the 2007 GOS (sites GS13, GS14, and GS15 [29]); these were selected because they are geographically closest to the source of environmental DNA used in this study (see below). Fewer than 150 of the ~350,000 non-*dmdA* metagenomic reads were complementary to both primers in any pair, even

TABLE 2. BLASTx and clustering results for *dmdA* amplicons of the free-living size fraction from southeastern U.S. coastal seawater[a]

Primer name	Clade	Subclade	% with correct clade(s) targeted	% with correct subclade (of correct clade) targeted	% with incorrect clade targeted	% not *dmdA*[b]	No. of sequences resampled	Normalized no. of *dmdA* clusters[c]
dmdAU	All	All	94.0	N/A	N/A	6.0	400	51
A/1-sp	Clade A	Subclade 1	99.2	99.8	0.5	0.3	2,500	30
A/2-sp	Clade A	Subclade 2	98.7	97.8	0.3	1.0	3,500	25
A/2-dg	Clade A	Subclade 2	99.1	99.4	0.1	0.8	3,500	24
A/2-ino	Clade A	Subclade 2	99.4	99.4	0.05	0.5	3,500	20*
B/3-sp	Clade B	Subclade 3	97.6	97.9	1.5	0.9	5,500	46
B/4-sp	Clade B	Subclade 4	33.6	99.3	65.4	0.9	1,500	20
C/2-sp	Clade C	Subclade 2	92.5	68.8	6.3	1.2	1,200	23
C/2-dg	Clade C	Subclade 2	64.2	81.8	33.8	2.0	1,200	35*
C/2-ino	Clade C	Subclade 2	71.7	98.8	26.2	2.2	1,200	20
D/1-sp	Clade D	Subclade 1	88.4	97.8	0.5	11.2	6,000	200
D/3-sp	Clade D	Subclade 3	99.6	91.5	0.10	0.3	4,300	30
D/3-dg	Clade D	Subclade 3	95.3	96.7	4.6	0.1	4,300	32
D/all-sp	Clade D	All	99.3[d]	N/A	0.2	0.5	4,500	82
D/all-dg	Clade D	All	99.9[d]	N/A	0.1	0	4,500	68*
D/all-ino	Clade D	All	99.8[d]	N/A	0	0.2	4,500	74*
E/2-sp	Clade E	Subclade 2	96.65	99.99	1.58	1.77	3,000	43
E/2-dg	Clade E	Subclade 2	98.97	99.78	0.51	0.51	3,000	35*

[a] For particle-associated data, see Table S3 in the supplemental material. N/A, not applicable.
[b] Includes sequences with hits to *gcvT* and those with no hits.
[c] Average of 1,000 resamplings (see Materials and Methods) using the population sizes indicated in the "No. of sequences resampled" column. Cluster numbers marked with an asterisk were significantly different ($P < 0.05$) from that obtained by the specific version of that primer pair.
[d] For the D/all-sp primer pair, 16.2% of the hits were to subclade D/1 and 2.5% to subclade D/3; for the D/all-dg primer pair, 13.8% of the hits were to subclade D/1 and 6.7% to subclade D/3; and for the D/all-ino primer pair, 4.4% of the hits were to subclade D/1 and 6.3% to subclade D/3. The remaining correct hits were to clade D sequences not classified within a subclade.

with an allowance of six mismatches per primer pair, and none of these would produce an amplicon of the correct size. Overall, final primer designs from the bioinformatic pipeline resulted in 22 primer pairs (which included degenerate and inosine versions where applicable) to 14 target groups: one universal target group, one clade-specific target group (clade D), and 12 subclade-specific target groups (three in clade A, four in clade B, one in clade C, two in clade D, and two in clade E) (Table 1; see Table S2 in the supplemental material).

Experimental primer testing. All *in silico*-tested primer pairs (including degenerate and inosine versions) were tested experimentally using composite DNA from free-living bacterioplankton communities (0.2- to 1.0-μm size fractions) collected over 5 years at the Sapelo Island Microbial Observatory (SIMO; http://simo.marsci.uga.edu) (see Table S1 in the supplemental material). DNA from 38 different samples was combined in order to capture the temporal and spatial variability of *dmdA* sequences at this coastal site, while keeping the number of amplicon pools to a reasonable level for sequencing. Of the 14 target groups, *dmdA* primer pairs to four (A/3, B/1, B/2, and E/1) did not produce amplicons from the composite DNA samples. Since these primers passed all bioinformatic criteria, they are described in the supplemental material (see Table S2) for potential use in PCR-based analyses of *dmdA* sequences in other marine environments. The remaining 10 groups were targeted by 18 primer pairs (including degenerate and inosine versions [Table 1]) that successfully produced amplicons from the composite DNA sample.

Amplicons were sequenced using 454 pyrosequencing technology and annotated based on the best hit in BLASTx analysis against our 3,280-member *dmdA* database (Table 2). An *in silico* test of known *dmdA* sequences with priming sites trimmed indicated that the BLAST analysis was accurate in assigning sequences to clades despite the short amplicons produced by some primer pairs (e.g., clade D/1 primers produce a 39-bp trimmed amplicon). For each primer pair, we determined (i) the percentage of correct sequences retrieved by the primers (as opposed to sequences with best hits to the wrong clade or to *dmdA* paralogs, or sequences that had no hit; some of these might include novel *dmdA* genes) and (ii) the richness of *dmdA* sequence clusters retrieved by the primers, defining clusters at a ≥90% nucleotide (~95% amino acid) identity level and using a resampling approach to normalize for differences in the number of sequences between primer pairs (see Materials and Methods).

For the universal primer pair, the majority of the sequences were *dmdA* (94%), with only a small number having better homology to paralogous genes or having no hits in the BLAST analysis (6%) (see Table 1; two different annealing temperatures were tested for the universal primer pair, but both yielded similar numbers of correct *dmdA* sequences). Cluster analysis indicated that 116 *dmdA* clusters were retrieved from the composite free-living bacterioplankton DNA, and these sequences represented all five major clades (Fig. 2). Clade A and D amplicons were the most abundant in terms of both numbers of sequences and numbers of clusters (Fig. 2).

For most specific clade and subclade primer pairs, at least 90% of the sequences were *dmdA* from the correct target clade (Table 2). The majority of nonspecific hits were to unclassified *dmdA* sequences, and fewer than 1% of the hits were to paralogous proteins. For most subclade primers, ~98% of the amplicons hitting the correct clade also hit the correct subclade (Fig. 2). Summing across all specific primer pairs for the targeted clades and subclades, cluster analysis indicated that 600

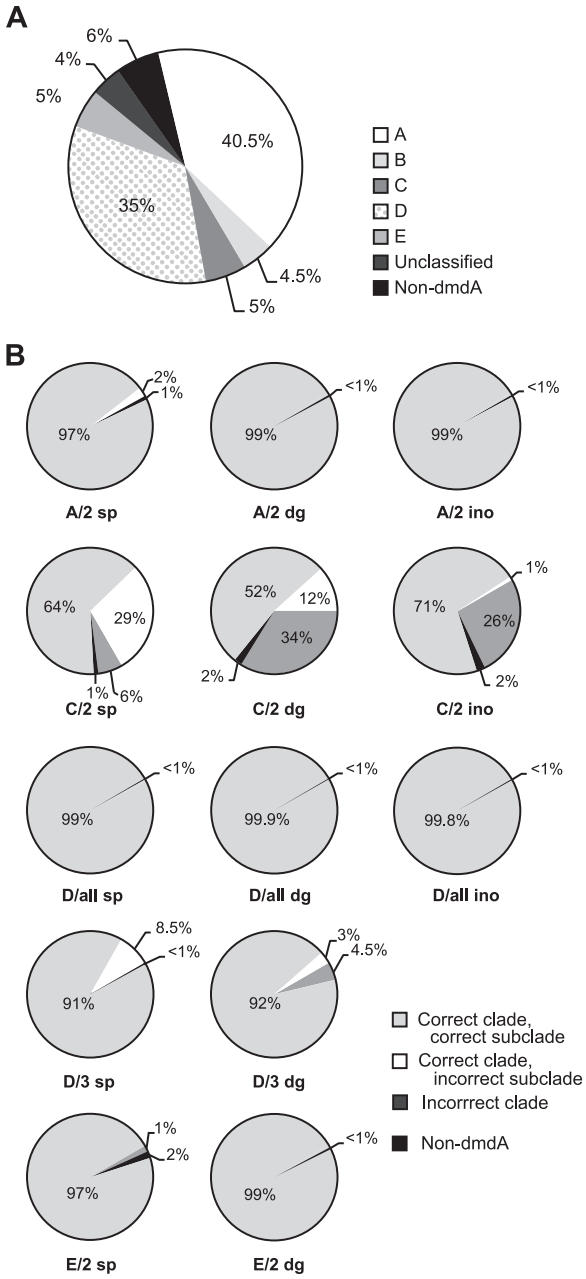

FIG. 2. Annotation of free-living (0.2- to 1.0-μm) amplicon sequences from *dmdA* primer pairs based on best hits in a BLASTx analysis against known *dmdA* sequences. (A) Universal primer pair. (B) Clade and subclade primer pairs, including specific, degenerate, and inosine versions.

total clusters and up to 17,203 unique nucleotide sequences were retrieved (from a total of 62,606 sequences). *dmdA* richness cannot be compared between clades or subclades using these primer pairs, however, because the regions of the gene targeted by the primers differ.

Specific versus degenerate primer pairs. Primer pairs with degenerate or inosine positions were included for some target groups if the bioinformatic pipeline indicated that they might substantially improve retrieval of *dmdA* diversity. The degenerate/inosine primer pairs were no more likely to retrieve incorrect sequences than the specific primers (Fig. 2), indicating that the modifications did not cause undue problems with nonspecific amplification. However, they were also no more likely to retrieve a higher degree of richness of *dmdA* sequences than the specific primers (as defined by 90% nucleotide sequence clusters) (Table 2) except for clade C/2 inosine primers. Moreover, most of the *dmdA* sequences retrieved with modified primers were the same as those retrieved with the specific primers (see Fig. S2 in the supplemental material), and a similar percentage of unique clusters were captured with the modified and specific primers. Thus, for this particular functional gene, primers modified with degenerate or inosine bases did not retrieve a richer sequence library. Based on the similar performances of these primer types and potential complications of using modified primers in future qPCR applications, only amplicons of the specific versions of the primer pairs were used in a subsequent comparative analysis of free-living versus particle-associated bacterioplankton communities.

***dmdA* in free-living and particle-associated bacterial communities.** The *dmdA* sequences amplified with the universal primer pair from southeastern U.S. coastal waters had comparable clade distributions in both the particle-associated (1.0- to 8.0-μm) and free-living (0.2- to 1.0-μm) size fractions. Clades A and D made up the majority of sequences in both fractions (Fig. 2A; see Table S3 footnote in the supplemental material). The universal primer pair targeted a higher number of apparent non-*dmdA* sequences in the particle-associated fraction (19%) than in the free-living fraction (6%) (Table 2; Table S3) but also showed a higher richness of correct *dmdA* clusters in the particle-associated community (Fig. 3). The clusters shared between the two communities accounted for most of the sequences (91%), and unique clusters were small (~2 sequences per cluster).

Amplicon richness and composition for clades and subclades of *dmdA* retrieved with the specific primer pairs were also comparable for free-living and particle-associated bacteria (see Table S3 in the supplemental material). An average of 60% of the clusters were shared across size fractions (Table 3). While four of nine subclade primer pairs showed a significant difference in the number of unique clusters retrieved between size fractions (Table 3), the degree of richness was higher in the particle-associated fraction in some cases (clade C/2) and in the free-living fraction in some cases (clades A/2, B/3, and D/1) (Fig. 3). However, unique clusters typically had few sequences and, as with the universal primer pair, an average of 90% of the *dmdA* sequences obtained with clade and subclade primer pairs were members of clusters shared across the size fractions.

DISCUSSION

The advent of metagenomic sequencing offers a significant advantage in environmental primer design. Previously, sequences from cultured organisms or small environmental clone libraries formed the basis for primer sequences. Yet how well those primers targeted the full natural gene diversity, and therefore captured gene abundance, distribution, and expression in complex bacterial communities (3, 31), was not known. Ecologically relevant sequences from metagenomic data are now available for designing primers for field studies (6). Here we made use of the thousands of *dmdA* homologs from marine

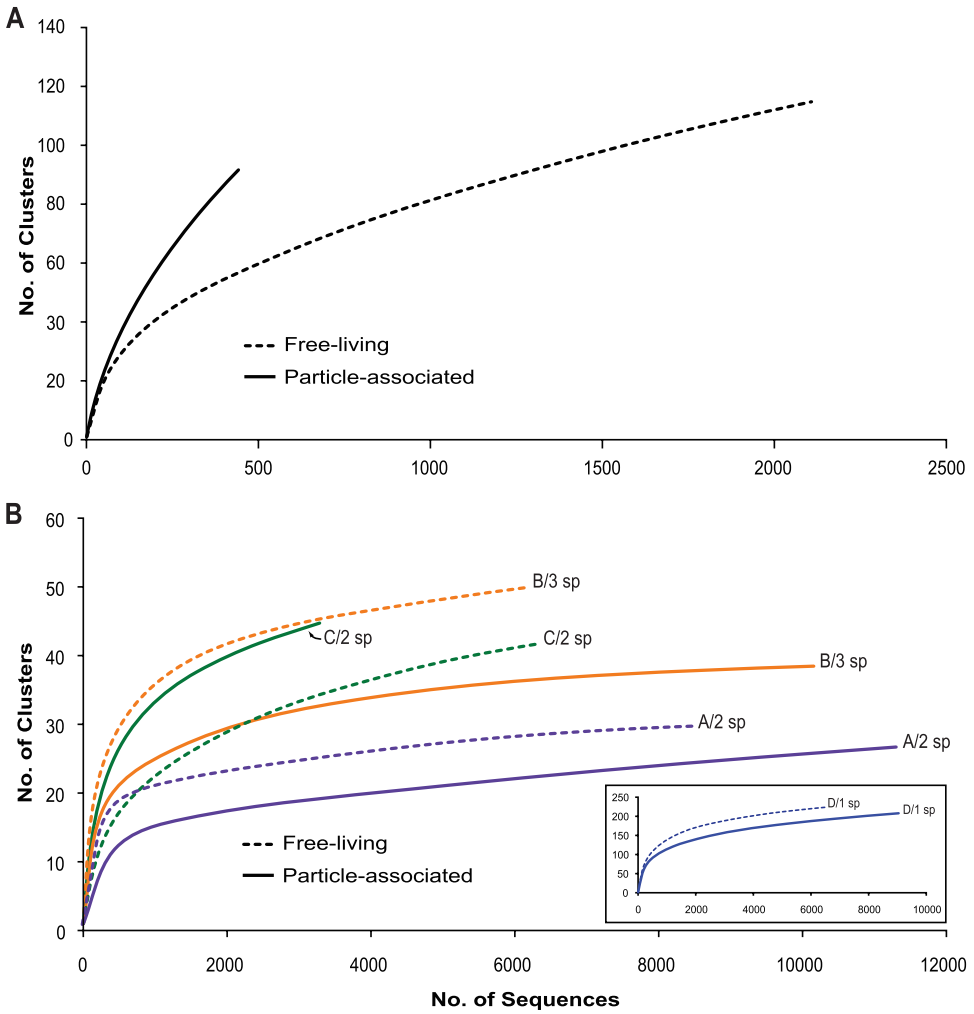

FIG. 3. Rarefaction curves of *dmdA* amplicons from free-living and particle-associated bacterioplankton communities based on 90% nucleotide identity clusters. (A) Universal *dmdA* primer pair. (B) Selected subclade primer pairs.

TABLE 3. Unique and shared clusters and percent shared sequences between size fractions for 10 *dmdA* primer pairs

Primer name	% Unique clusters[a]		% Shared clusters	% Sequences in unique clusters		% Sequences in shared clusters
	Free living	Particle associated		Free living	Particle associated	
Univ	24	43*	33	3	6	91
A/1-sp	7	15	78	<1	<1	99
A/2-sp	33*	10	57	7	<1	93
B/3-sp	31*	5	64	27	<1	73
B/4-sp	10	6	84	<1	<1	>99
C/2-sp	23	33*	44	2	<1	98
D/1-sp	29*	17	54	21	<1	79
D/3-sp	19	15	66	<1	<1	>99
D/all-sp	21	22	57	11	<1	88
E/2-sp	15	13	72	5	<1	95

[a] Average of 1,000 resamplings (see Materials and Methods) using population sizes indicated in Table 2. Cluster numbers marked with an asterisk were significantly higher ($P < 0.05$) than those obtained for the other size fraction.

metagenomic data to design optimized primer pairs and then systematically assessed the primers by deep sequencing of amplicon populations. The substantial nucleotide sequence diversity in the GOS data set for this single gene made it necessary to target groups at the subclade level. Similarly high levels of richness have been found for another widespread and abundant marine bacterial gene, the gene for proteorhodopsin (3).

When primers were tested on coastal DNA, more than 90% of the amplicons were from the correct *dmdA* target group. The universal primer pair captured all five clades, with a significant proportion of correct sequences classified as clade A (43%) or D (37%). These two clades harbor genes from cultured roseobacters and SAR11 members, respectively, and were also abundant among *dmdA* genes retrieved from coastal and open ocean sites in the GOS data set (16). Other primer pairs for clades and subclades of *dmdA* were likewise highly specific in targeting correct sequences. Overall, the *dmdA* amplicons formed hundreds of clusters at ≥90% nucleotide identity (~95% amino acid identity, based on manual alignments of translated sequences from a subset of clusters) and did not

reach full saturation even after ~6,400 sequences per primer pair. Since a composite DNA preparation from 38 samples was used to assess primer performance (to increase the likelihood of target genes for each primer pair being tested), we do not yet know how abundance and composition of the *dmdA* pool vary over time and space; these vetted qPCR primers now provide a robust tool to address *dmdA* dynamics in this and other locations.

The modification of primer sequences with degenerate bases or inosine has been used previously to improve PCR primer annealing when target sequences are heterogeneous (3, 10, 22, 31, 39). For environmental primers, such modifications might allow more of the natural diversity of a functional protein to be captured (39) although potentially at the expense of nonspecific binding. In this study, modified primers were no more prone to nonspecific amplification than specific primers. Yet while we expected that amplicons from the unmodified parent primers would be a subset of those from the modified primers, surprisingly this was not the case for this study. Generalizing across the primer pairs tested, the degenerate and inosine primers captured an equally diverse but slightly different suite of sequences compared to those captured by the specific primers. These empirical results guided us toward the use of the specific clade and subclade primers in subsequent analyses. We did not design or test a specific version of the universal *dmdA* primer.

In the first use of these primer pairs, we asked whether the composition of the *dmdA* reservoir (based on 38 pooled samples spanning 5 years) differs between free-living and particle-attached bacterial communities in southeastern U.S. coastal waters. The GOS metagenomic data set, which comprises the largest collection of environmental *dmdA* sequences to date, is heavily biased toward free-living cells (defined as $\leq$0.8 μm in diameter), providing little information on representation of the major clades and subclades of *dmdA* in particle-associated communities. DMSP concentrations are locally higher in marine particle "microenvironments" than in bulk seawater (20), since the primary source of DMSP is phytoplankton cells, raising the issue of whether particle-associated demethylation is driven by a different suite of *dmdA* orthologs. Differences in *dmdA* composition between the two size fractions could reflect ecological advantages conferred by differing kinetic parameters of the major clades (e.g., K_m and k_{cat}) (28). Alternatively, taxonomic differences between marine bacterial size classes, as has been shown previously (8), may drive differences in the composition of the *dmdA* reservoirs. In either case, gene composition might provide insights into rates of, or controls on, DMSP demethylation. DMSP lyase activity (i.e., the competing pathway for DMSP degradation) has been shown to be greater in particle-associated microbial communities than in free-living microbial communities (4, 30).

Here, we used a 1.0-μm-pore-size filter to operationally separate free-living from particle-associated bacteria, and we conducted a comparative analysis of their *dmdA* reservoirs. While the universal primer pair suggested greater overall sequence richness in the particle-attached communities (Table 3), results were mixed for individual clade and subclade primer pairs: one primer pair also displayed a significantly higher degree of richness in the particle-associated fraction, three displayed significantly higher richness in the free-living fraction, and five showed no difference. Since clade D primers likely target *dmdA* sequences in SAR11 populations (15), we predicted greater richness for this clade of planktonic oligotrophs (26) in the free-living size fraction, and this was the case (Fig. 3). Since clade A primers target *dmdA* sequences in *Roseobacter* cells (and other taxa), we predicted greater richness for this clade of surface colonizers (2, 7) in the particle-associated fraction, but this was not the case. Yet despite these significant differences in cluster richness for some primer pairs, the vast majority of sequences were assigned to clusters that were shared between free-living and particle-attached cells (Table 3). Since our primers were designed from the GOS metagenome, which includes mostly free-living bacterioplankton in the 0.2- to 0.8-μm size range, we cannot rule out the possibility that they systematically miss *dmdA* diversity in particle-associated bacteria. Better metagenomic coverage of larger size classes of marine particles in future sequencing efforts will provide a mechanism to check, and if necessary redesign, *dmdA* primers.

The availability of metagenomic sequence data has greatly improved our ability to design qPCR primers to assess abundance, diversity, and expression of microbial functional genes in the environment. In the case of the DMSP demethylase, knowledge of how *dmdA* genes vary over time and space, and how their expression changes in response to DMSP dynamics and environmental drivers, will increase understanding of the marine bacterial communities that regulate sulfur emission from the ocean surface.

ACKNOWLEDGMENTS

We thank W. Ye and C. Lasher for DNA samples; S. Sharma for bioinformatic expertise; C. English for graphics assistance; R. Newton, W. Whitman, A. Karls, S. Gifford, and J. Edmonds for helpful discussions; and J. Jones at the University of South Carolina EnGenCore Sequencing Facility for sequencing expertise.

This research was supported by grants from the National Science Foundation (OCE-0724017) and the Gordon and Betty Moore Foundation.

REFERENCES

1. **Andreae, M. O.** 1990. Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. Mar. Chem. **30:**1–29.
2. **Buchan, A., J. M. González, and M. A. Moran.** 2005. Overview of the marine *Roseobacter* lineage. Appl. Environ. Microbiol. **71:**5665–5677.
3. **Campbell, B. J., L. A. Waidner, M. T. Cottrell, and D. L. Kirchman.** 2008. Abundant proteorhodopsin genes in the North Atlantic Ocean. Environ. Microbiol. **10:**99–109.
4. **Cantin, G., M. Levasseur, S. Schultes, and S. Michaud.** 1999. Dimethylsulfide (DMS) production by size-fractionated particles in the Labrador Sea. Aquat. Microb. Ecol. **19:**307–312.
5. **Chambers, S. T., C. M. Kunin, D. Miller, and A. Hamada.** 1987. Dimethylthetin can substitute for glycine betaine as an osmoprotectant molecule for *Escherichia coli*. J. Bacteriol. **169:**4845–4847.
6. **Contreras-Moreira, B., B. Sachman-Ruiz, I. Figueroa-Palacios, and P. Vinuesa.** 1 July 2009. Primers4clades: a web server that uses phylogenetic trees to design lineage-specific PCR primers for metagenomic and diversity studies. Nucleic Acids Res. **37:**W95–W100. doi:10.1093/nar/gkp377.
7. **Dang, H., and C. R. Lovell.** 2000. Bacterial primary colonization and early succession on surfaces in marine waters as determined by amplified rRNA gene restriction analysis and sequence analysis of 16S rRNA genes. Appl. Environ. Microbiol. **66:**467–475.
8. **DeLong, E. F., D. G. Franks, and A. L. Alldredge.** 1993. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. Limnol. Oceanogr. **38:**924–934.
9. **Drummond, A., B. Ashton, M. Cheung, J. Heled, M. Kearse, R. Moir, S. Stones-Havas, T. Thierer, and A. Wilson.** 2008. Geneious, v4.0. Biomatters Ltd., Auckland, New Zealand.
10. **Ehlen, T., and L. Dubeau.** 1989. Detection of ras point mutations by polymerase chain reaction using mutation-specific, inosine-containing oligonucleotide primers. Biochem. Biophys. Res. Commun. **160:**441–447.

11. **González, J. M., R. P. Kiene, and M. A. Moran.** 1999. Transformation of sulfur compounds by an abundant lineage of marine bacteria in the α-subclass of the class *Proteobacteria*. Appl. Environ. Microbiol. **65:**3810–3819.
12. **Good, P. I.** 2005. Introduction to statistics through resampling methods and R/S-PLUS. John Wiley and Sons, Inc., Hoboken, NJ.
13. **Gotelli, N. J., and G. L. Entsminger.** 2008. EcoSim: null models software for ecology. Version 7. Acquired Intelligence Inc. & Kesey-Bear, Jericho, VT.
14. **Hall, T. A.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41:**95–98.
15. **Howard, E. C., J. R. Henriksen, A. Buchan, C. R. Reisch, H. Buergmann, et al.** 2006. Bacterial taxa that limit sulfur flux from the ocean. Science **314:**649–652.
16. **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.
17. **Huber, J. A., D. B. M. 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.
18. **Kiene, R. P.** 1996. Production of methanethiol from dimethylsulfoniopropionate in marine surface waters. Mar. Chem. **54:**69–83.
19. **Kiene, R. P., and L. J. Linn.** 2000. Distribution and turnover of dissolved DMSP and its relationship with bacterial production and dimethylsulfide in the Gulf of Mexico. Limnol. Oceanogr. **45:**849–861.
20. **Kiene, R. P., L. J. Linn, and J. A. Bruton.** 2000. New and important roles for DMSP in marine microbial communities. J. Sea Res. **43:**209–224.
21. **Kiene, R. P., L. J. Linn, J. Gonzalez, M. A. Moran, and J. A. Bruton.** 1999. Dimethylsulfoniopropionate and methanethiol are important precursors of methionine and protein-sulfur in marine bacterioplankton. Appl. Environ. Microbiol. **65:**4549–4558.
22. **Knoth, K., S. Roberds, C. Poteet, and M. Tamkun.** 1988. Highly degenerate, inosine-containing primers specifically amplify rare cDNA using the polymerase chain reaction. Nucleic Acids Res. **16:**10932.
23. **Li, W., and A. Godzik.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658–1659.
24. **Lovelock, J. E., R. J. Maggs, and R. A. Rasmussen.** 1972. Atmospheric dimethyl sulphide and the natural sulphur cycle. Nature **237:**452–453.
25. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
26. **Morris, R. M., M. S. Rappe, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, and S. J. Giovannoni.** 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature **420:**806–810.
27. **Pichereau, V., J.-A. Pocard, J. Hamelin, C. Blanco, and T. Bernard.** 1998. Differential effects of dimethylsulfoniopropionate, dimethylsulfonioacetate, and other S-methylated compounds on the growth of *Sinorhizobium meliloti* at low and high osmolarities. Appl. Environ. Microbiol. **64:**1420–1429.
28. **Reisch, C. R., M. A. Moran, and W. B. Whitman.** 2008. Dimethylsulfoniopropionate-dependent demethylase (DmdA) from *Pelagibacter ubique* and *Silicibacter pomeroyi*. J. Bacteriol. **190:**8018–8024.
29. **Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, et al.** 2007. The *Sorcerer* II global ocean sampling expedition: Northwest Atlantic through Eastern tropical Pacific. PLoS Biol. **5:**e77.
30. **Scarratt, M., G. Cantin, M. Levasseur, and S. Michaud.** 2000. Particle size-fractionated kinetics of DMS production: where does DMSP cleavage occur at the microscale? J. Sea Res. **43:**245–252.
31. **Schwalbach, M. S., and J. A. Fuhrman.** 2005. Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. Limnol. Oceanogr. **50:**620–628.
32. **Simo, R.** 2004. From cells to globe: approaching the dynamics of DMS(P) in the ocean at multiple scales. Can. J. Fish. Aquat. Sci. **61:**673–684.
33. **Simo, R.** 2001. Production of atmospheric sulfur by oceanic plankton: biogeochemical, ecological and evolutionary links. Trends Ecol. Evol. **16:**287–294.
34. **Sunda, W., D. J. Kieber, R. P. Kiene, and S. Huntsman.** 2002. An antioxidant function for DMSP and DMS in marine algae. Nature **418:**317–320.
35. **Taylor, B. F., and D. C. Gilchrist.** 1991. New routes for aerobic biodegradation of dimethylsulfoniopropionate. Appl. Environ. Microbiol. **57:**3581–3584.
36. **Vairavamurthy, A., M. O. Andreae, and R. L. Iverson.** 1985. Biosynthesis of dimethylsulfide and dimethylpropiothetin by *Hymenomonas carterae* in relation to sulfur source and salinity variations. Limnol. Oceanogr. **30:**59–70.
37. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, et al.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
38. **Yoch, D. C.** 2002. Dimethylsulfoniopropionate: its sources, role in the marine food web, and biological degradation to dimethylsulfide. Appl. Environ. Microbiol. **68:**5804–5815.
39. **Yutin, N., M. T. Suzuki, and O. Beja.** 2005. Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. Appl. Environ. Microbiol. **71:**8958–8962.