Applied and Environmental Microbiology, Jan. 1997, p. 63–70 Vol. 63, No. 1
0099-2240/97/$04.00+0

Diversity and Depth-Specific Distribution of SAR11 Cluster rRNA Genes from Marine Planktonic Bacteria

K. G. FIELD,* D. GORDON, T. WRIGHT, M. RAPPÉ, E. URBACH, K. VERGIN, and S. J. GIOVANNONI

Department of Microbiology, Oregon State University, Corvallis, Oregon 97331

Received 25 March 1996/Accepted 11 October 1996

Small-subunit (SSU) ribosomal DNA (rDNA) gene clusters are phylogenetically related sets of SSU rRNA genes, commonly encountered in genes amplified from natural populations. Genetic variability in gene clusters could result from artifacts (polymerase error or PCR chimera formation), microevolution (variation among *rrn* copies within strains), or macroevolution (genetic divergence correlated with long-term evolutionary divergence). To better understand gene clusters, this study assessed genetic diversity and distribution of a single environmental SSU rDNA gene cluster, the SAR11 cluster. SAR11 cluster genes, from an uncultured group of the α subclass of the class *Proteobacteria*, have been recovered from coastal and midoceanic waters of the North Atlantic and Pacific. We cloned and bidirectionally sequenced 23 new SAR11 cluster 16S rRNA genes, from 80 and 250 m in the Sargasso Sea and from surface coastal waters of the Atlantic and Pacific, and analyzed them with previously published sequences. Two SAR11 genes were obviously PCR chimeras, but the biological (nonchimeric) origins of most subgroups within the cluster were confirmed by independent recovery from separate gene libraries. Using group-specific oligonucleotide probes, we analyzed depth profiles of nucleic acids, targeting both amplified rDNAs and bulk RNAs. Two subgroups within the SAR11 cluster showed different highly depth-specific distributions. We conclude that some of the genetic diversity within the SAR11 gene cluster represents macroevolutionary divergence correlated with niche specialization. Furthermore, we demonstrate the utility for marine microbial ecology of oligonucleotide probes based on gene sequences amplified from natural populations and show that a detailed knowledge of sequence variability may be needed to effectively design these probes.

Bacterial 16S rRNA gene sequences cloned from natural populations are used to identify population members without the necessity of first cultivating the microorganisms (29, 35, 45, 46). An early example of this kind of analysis unexpectedly uncovered 16S ribosomal RNA gene clusters: sets of environmental 16S rDNA sequences more closely related to each other than to any sequence from a described microorganism (16). Many studies of marine microbial populations have now observed gene clusters (10, 15, 41), and rRNA gene clusters have been recovered from a variety of other habitats ranging from peat bogs (18) to the hindguts of termites (27), in libraries prepared both with and without the PCR (1, 3, 6, 10, 15, 30). From a phylogenetic perspective, gene clusters resemble gene clades from cultivated organisms, but because gene cluster sequences come from uncultivated organisms, it is not known whether they represent separate cellular lineages. Generally, the origin and significance of gene clusters from natural populations are unknown.

Oligonucleotide probes targeting rRNA genes are increasingly used in studies involving uncultured bacteria. Decisions about the design of these probes often rest on assumptions about the relationship between sequence diversity and ecological and evolutionary specialization. A recent study utilizing oligonucleotide probes specific for activated sludge bacteria provided evidence that gene clusters may represent diversified cellular lineages but also showed that the patterns of genetic diversity among these bacteria were far more complex than assumed (1). Understanding the mechanisms, both biological and artifactual, that determine the observed genetic diversity in environmental rRNA sequences is therefore essential for probe design.

Genetic variability in gene clusters could result from artifacts, microevolution, or macroevolution. Polymerase error and formation of chimeric genes (shuffle genes [43], which are PCR amplicons containing regions copied from two different template genes) could introduce artifactual variations in genes amplified from mixed nucleic acid samples. Chimeras between phylogenetically distant 16S ribosomal DNA (rDNA) sequences have been recovered at low frequencies in several gene libraries prepared by PCR from environmental samples (3, 7, 20, 36). Phylogenetic trees constructed separately from the 3′ and 5′ ends of chimeras formed from distantly related sequences will place them in different lineages (e.g., see references 20 and 23). However, chimeras between closely related sequences are difficult to detect. Using CHECK_CHIMERA of the Ribosomal Database Project (22, 24) or a similar analysis (36), Robison-Cox and colleagues estimated that chimeric sequences are detected at the 95% confidence level only when the two parental sequences are no more than 84% similar; the probability of detecting a chimera between sequences that are 96% similar is only 50% (36).

Microevolutionary variation in rRNA gene clusters could be produced by sequence variability among ribosomal operons within a single strain. The copy number of rRNA genes in bacteria varies from 1 to at least 13 (reviewed in reference 42). In general, multiple gene copies within a genome are homogenized by concerted evolution (48), thought to occur via gene conversion and unequal crossing over. However, multiple heterogeneities among ribosomal operons within a single strain have been found in *Escherichia coli* (8), *Mycoplasma* spp. (4, 31, 32), *Rhodobacter spheroides* (11), and *Haloarcula marismor-*

* Corresponding author. Mailing address: Department of Microbiology, Oregon State University, Corvallis, OR 97331. Phone: (541) 737-1837. Fax: (541) 737-0496. E-mail: fieldk@bcc.orst.edu.

TABLE 1. Times and dates of collections at the BATS

BATS sample no.	Date (mo-day-yr)	Time	Surface temp (°C)	DCM[a] (m)
35	8-12-91	1912	29	100–150
38	11-11-91	2355	24	100
41	2-13-92	1000	19	0–50
42	3-10-92	0830	20	30–70
45	6-17-92	2355	24	80–100
48	9-15-92	1420	27	90
52	1-12-93	2345	20	
53	2-9-93	1415	20	50
54	3-10-93	1724	19	70
58	7-13-93	1730	27	121

[a] DCM, deep chlorophyll maximum.

tui (26), although not in *Haemophilus influenzae* (14). Since unless genes are chimeric, rRNA genes cloned from environmental DNAs represent single operons, variability among ribosomal operons within single strains could contribute to variability within a gene cluster from the environment.

rRNA genes are widely used for phylogenetic studies, because they are relatively conserved over evolutionary time scales (22, 47); areas of functional significance in the molecule are conserved, and substitutions in rRNA sequences are thought to be largely selectively neutral. Thus, we would expect that even the level of rRNA gene variation found in gene clusters could occur in the context of long-term evolutionary (macroevolutionary) events such as speciation or the evolution of higher taxa.

The goal of this study was to investigate the significance of gene clusters in natural populations through a detailed study of the genetic diversity, phylogenetic relationships, and ecological distribution of a particular cluster, the SAR11 cluster. SAR11 cluster genes can constitute a significant component of the 16S rRNA genes from the bacterioplankton community at the surface of the Sargasso Sea (6, 16). Since the original discovery of the cluster, lineages related to the SAR11 cluster have been found in surface waters at ALOHA Station in the north central Pacific (41); at depths of 100 and 500 m in the California Current, North Pacific (15); and in coastal waters of the Santa Barbara Channel (10). The widespread occurrence of the group suggests its worldwide importance in subtropical bacterioplankton communities (25). Although a member of the α subclass proteobacterial phylum, the SAR11 cluster has no close relatives among the α subclass of the *Proteobacteria* (25) but represents a novel, previously unknown bacterial line of descent (16). Similarities among the original SAR11 cluster sequences ranged from 94 to 97% (16). The similarity between the 16S rRNA genes of *E. coli* and *Salmonella enteridis* is 98%; thus, some SAR11 cluster gene sequences show the same (or greater) evolutionary separation from one another as do strains recognized as separate species in other groups.

Our approach was to isolate, clone, and sequence SAR11-related 16S rRNA genes from a variety of locales, analyzing sequences for evidence of chimera formation. We made a detailed phylogenetic analysis of new and published SAR11 cluster gene sequences and used the sequences to design oligonucleotide probes specific to subgroups within the SAR11 cluster. Since SAR11 cluster genes initially recovered from surface and deep waters appeared to belong to different lineages, we hypothesized that some SAR11 cluster lineages might have depth-specific distributions. Therefore, we used the probes to study the distribution of SAR11 rRNAs in the water column.

We found that total variation among genes in the SAR11 cluster was large: pairwise similarities between complete SAR11 cluster genes ranged from 89.9 to 99.3%, with some partial sequences from coastal libraries representing even more distant SAR11 cluster lineages. Variation within the SAR11 cluster genes was not continuous but, instead, defined discrete subgroups, some of which were not strongly supported by bootstrap analyses. Therefore, to test whether artifacts such as chimeric genes had produced some of the variation defining subgroups within the cluster, we searched for independent confirmation of the subgroups. Independent recovery of the same subgroup from separate gene libraries confirmed most of the subgroups. Finally, using group-specific oligonucleotide probes to map the distribution of the subgroups against one ecological variable, depth, we found that two of the subgroups had highly depth-specific distributions.

MATERIALS AND METHODS

Sample collection. Picoplankton samples were collected from the Bermuda Atlantic Time Series Station (BATS) at 31°50′N, 64°10′W in the Sargasso Sea, Atlantic Ocean, from depths of 0, 40, 80, 120, 160, 200 and 250 m (Table 1), by pumping water through 0.2-μm-pore-size filters as previously described (17). On 28 April 1993, picoplankton samples were collected from a depth of 10 m at 44°39.1′N, 124°10.6′W, off the Pacific Ocean coast of Newport, Oreg. (17). Water samples were collected at 35°59′N, 75°08′W, off the coast of Cape Hatteras, N.C., Atlantic Ocean, as previously described (33).

Nucleic acid extraction. Total nucleic acids were extracted by cell lysis with sodium dodecyl sulfate and proteinase K followed by phenol-chloroform extraction, as previously described (6). Cellular RNAs and DNAs were separated by isopycnic centrifugation in cesium trifluoroacetate (17).

Gene amplification. Bacterial 16S rRNA primers 27F and 1522R (Table 2) (12) were used in PCR (37) to amplify 16S rDNAs from DNA preparations. We used *Taq* DNA polymerase (Promega Corporation, Madison, Wis.) to amplify the Sargasso Sea samples, as described previously (17). We amplified 16S rRNA genes from the Oregon and Cape Hatteras samples with *Pfu* polymerase (Stratagene) as previously described (33).

Clone library construction. Clone libraries were constructed from the Sargasso Sea samples with the TA Cloning System (Invitrogen Corporation, San Diego, Calif.) as previously described (33). For the Oregon Coast (OCS) and Cape Hatteras Ocean Margins (OM) libraries, we blunt-end cloned into pBluescript KS− as previously described (16). We isolated recombinant plasmid DNAs from clones by a standard alkaline lysis method (5, 39) or with Magic Minipreps (Promega) or Prep-A-Gene plasmid purification kits (Bio-Rad, Richmond, Calif.).

DNA sequencing. Double-stranded plasmid DNAs were sequenced by dye-terminator chemistry and an ABI 373A automated sequencer (Applied Biosystems, Foster City, Calif.) or by conventional dideoxy-terminated sequencing (40) with Sequenase (U.S. Biochemical Co., Cleveland, Ohio) and ^{35}S-dATP. To screen clone libraries, we sequenced one end of each insert with a standard M13 sequencing primer and subjected these sequences to a preliminary phylogenetic analysis. SAR11 cluster genes selected for further analysis were bidirectionally sequenced with standard M13 sequencing primers and rRNA primers (21).

Phylogenetic analysis. We manually aligned sequences with the Genetic Data Environment, version 2.0, sequence analysis software (provided by Steven Smith, Millipore Corporation, Marlborough, Mass.). Using the PHYLIP, version 3.5, software package (13), we estimated evolutionary distances with the Kimura two-parameter model for nucleotide change (19) and a transition-transversion ratio of 2.0. Phylogenetic trees were constructed from distance estimates by

TABLE 2. Oligonucleotides used in this study

Name	*E. coli* numbering	Sequence
27F	8–27	5′-AGAGTTTGATCMTGGCTCAG-3′
338R	338–355	5′-GCTGCCTCCCGTAGGAGT-3′
1492R	1492–1510	5′-GGTTACCTTGTTACGACTT-3′
1522R	1522–1541	5′-AAGGAGGTGATCCANCCRCA-3′
SAR11-A1	179–209	5′-AAGCTTTCTCCGTAAAGACTTAT-3′
SAR11-A2	220–237	5′-GCAGGCTCATCCAATGGT-3′
SAR11-B2	220–236	5′-CGGGCTCATCTTTCGGC-3′
SAR11-B3	1244–1265	5′-CTCTTCGCDTCTCAYTGTAAGT-3′
SAR11-D4	1113–1135	5′-AATGTTAGTAACTAAACGTAGGG-3′
SAR11-G1	171–193	5′-CCTGTGAAGGCTTATTCAGTATT-3′

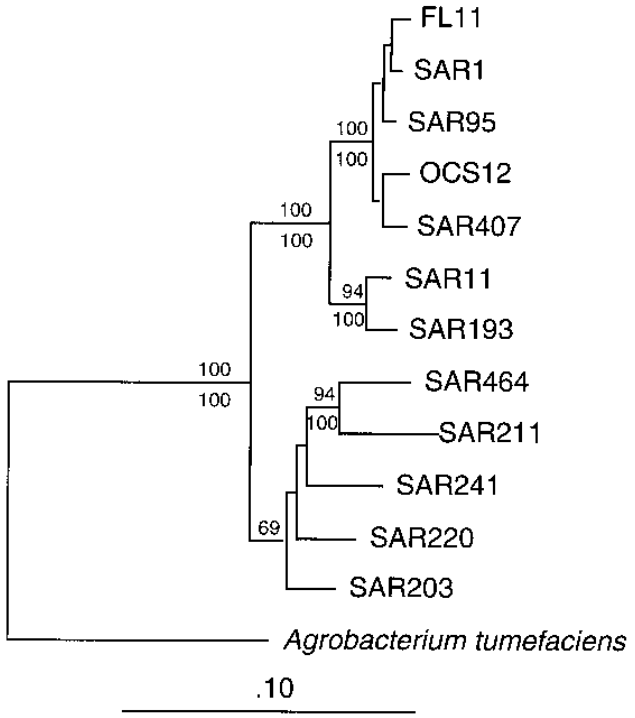

FIG. 1. Phylogenetic relationships among SAR11 cluster 16S rRNA genes, inferred by neighbor joining (38) from *E. coli* positions 9 through 1005 (the sequence positions found in the shortest gene, OCS12). The gene sequence from *Agrobacterium tumefaciens* was used to root the tree. The same branching order was recovered by the method of Wagner parsimony (44). The numbers above the internal segments are the percentages of bootstrap replicates which supported the branching order for the neighbor-joining tree; bootstrap values for the parsimony analysis are shown below the segments. Bootstrap values below 60% are not shown. SAR1, -11, and -95, surface; SAR193, -203, -211, -220, and -241, 250 m; SAR407 and -464, 80 m. SAR, Sargasso Sea; OCS12, Oregon coast; FL11, California coast (10).

neighbor joining (38). We inferred parsimony trees with the heuristic search option of PAUP (44). The bootstrap (12) with 100 replicates was used to estimate the robustness of branches in both neighbor-joining and parsimony trees. We edited phylogenetic trees with the program Treetool, provided by Mike Maciukenas, Ribosomal Database Project (RDP [22, 24]).

Identification of chimeric genes. We obtained secondary structure models for rDNAs with the program gRNAID, version 1.4 (46a), and examined them for location and nature of mutations and compensatory base changes across helices. Gene sequences were sent to the program CHECK_CHIMERA of the RDP.

Probe design and testing. We designed group-specific oligonucleotide probes (Table 2) by reference to aligned sequences and phylogenetic trees. The probes were screened for possible cross-reaction with unrelated sequences by sending them to CHECK_PROBE of the RDP. The 5′ terminus of each of the oligonucleotide probes was ^{32}P labeled with T4 polynucleotide kinase, as described previously (16). Labeled probes were tested for specificity by hybridizing them to control rDNAs amplified from known clones. Using the amplification conditions given above, we amplified template DNAs with 1492R and 27F primers (21), blotted the DNAs, and hybridized them to labeled probes as described previously (17). Washes at 5° temperature increments were used to empirically establish t_h, the stringent wash temperature, for each probe (probes SAR11-A1, -A2, -D4, and -G1, 37°C; SAR11-B2 and -B3, 45°C).

Depth profiles. After amplification with 1492R and 27F primers, we blotted Sargasso Sea 16S rRNA genes from 0, 40, 80, 120, 160, 200 and 250 m (from the BATS sample no. 35 water samples [Table 1]) as described above and sequentially hybridized the blots to the SAR11 oligonucleotide probes. Hybridization was quantified with an Ambis Mark II Radioanalytic System (Automated Microbiological Systems, San Diego, Calif.). Bulk RNAs prepared from depth profile samples (Table 1) were resuspended, denatured, blotted, and probed as reported previously (17).

Nucleotide sequence accession numbers. Nucleotide sequences were filed in GenBank under the accession numbers listed in Table 3.

RESULTS

To assess genetic diversity and ecological distribution of SAR11 cluster rRNA genes, we sequenced new SAR11 cluster SSU rRNA genes from gene libraries prepared by PCR of bulk nucleic acids collected from several depths at sites in the coastal and open-ocean North Atlantic and Pacific. We analyzed the sequences phylogenetically, designed phylogenetic-group-specific probes targeting 16S rRNAs, and hybridized the probes to nucleic acid samples from various depths to look for depth-specific distributions of SAR11 cluster subgroups. We did not use in situ hybridization fluorescent probes for this purpose, because these organisms cannot quantitatively be detected by this approach, possibly because they are small or growing slowly.

To select clones from clone libraries for sequencing, we examined preliminary phylogenetic trees inferred from sequences obtained from a single primer (M13 forward) (data not shown) and chose clones to represent the genetic variability revealed by these analyses. We bidirectionally sequenced eight complete SAR11 cluster-cloned genes: SAR193, SAR203, SAR211, SAR220, and SAR241 from a depth of 250 m in the Sargasso Sea, Atlantic Ocean; SAR407 and SAR464 from a depth of 80 m in the Sargasso Sea; and OCS12 from a depth of 10 m off the Oregon Coast, Pacific Ocean. We also analyzed 15 new partial SAR11 cluster sequences.

Parsimony and neighbor-joining analyses recovered essentially the same phylogenetic branching orders, and trees constructed from either full-length sequences alone or both full- and nearly full-length sequences were also the same (Fig. 1). We used the bootstrap (12) to estimate robustness of the tree branches for both neighbor-joining and parsimony analyses. A

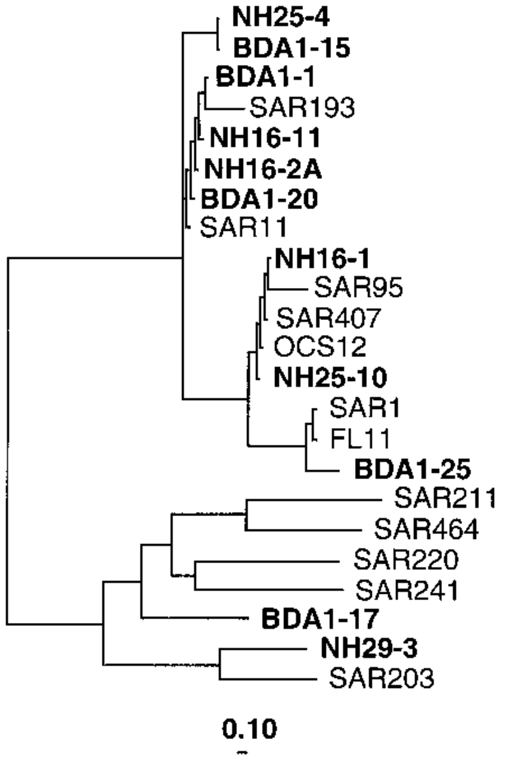

FIG. 2. Phylogenetic relationships of SAR11 cluster gene fragments (in boldface type) from Atlantic and Pacific ocean clone libraries (15). This is an unrooted tree inferred by neighbor joining from positions corresponding to *E. coli* 537 to 741. The implied root was set to match that in Fig. 1.

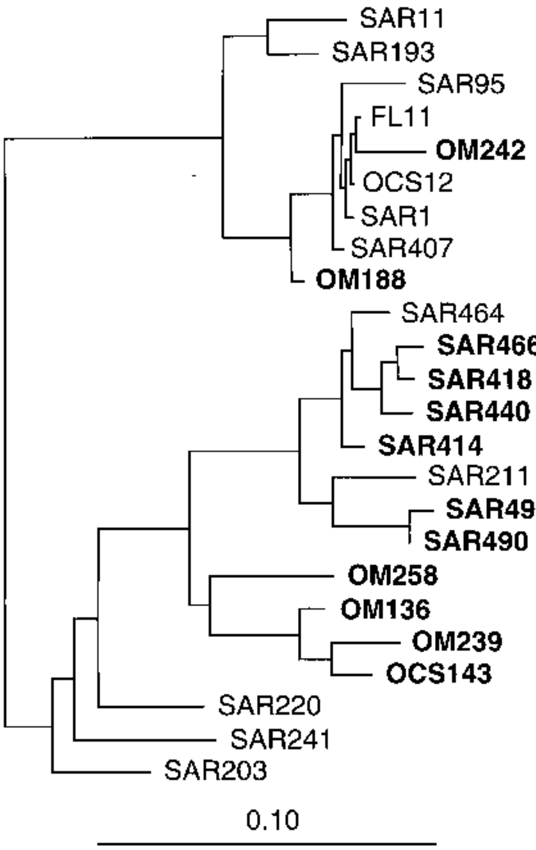

FIG. 3. Phylogenetic relationships of partial sequences (in boldface type) from the Sargasso Sea 80-m library (SAR400), OCS library, and OM library. This is an unrooted tree inferred by neighbor joining from positions 106 to 319 (*E. coli* numbering). The implied root was set to match that in Fig. 1.

bootstrap value of 100% provided strong support for the monophyly of the entire SAR11 cluster lineage (Fig. 1). Within the cluster, the sequences were divided into two major subclusters. The subcluster containing SAR11 was supported by a high bootstrap value, as were the two lineages within this subcluster. The second major subcluster was supported by moderate bootstrap values. Within the second subcluster, the subgroup comprising SAR211 and SAR464 was moderately well supported (Fig. 1). The rest of the branches in the phylogenetic tree were not strongly supported by bootstrapping.

Gene sequence fragments from different studies often don't overlap, making phylogenetic comparisons difficult. Using neighbor joining, we inferred separate phylogenetic trees for sets of nonoverlapping fragments in comparison to complete sequences (Fig. 2 and 3). SAR11 cluster gene fragments recovered from both the Atlantic and Pacific Oceans by Fuhrman and his colleagues (15) fell into the same phylogenetic groups as the full-length sequences (Fig. 2), as did sequences from the ALO library from the Central Pacific (data not shown) (41). Partial sequences from the OM library, the OCS library, and the 80-m Sargasso Sea library considerably expanded the subgroup that included SAR211 and SAR464 (Fig. 3); the coastal sequences formed a new, deep branch within this subgroup.

To eliminate misincorporation errors as a major source of variations, we used secondary structure models to examine the nature and position of sequence variations. If base substitutions had been caused by polymerase error, we would expect them to be randomly distributed throughout the sequences. Instead, most substitutions were confined to highly variable areas of the sequences and did not disturb the highly conserved secondary structures.

To screen for chimeras, we analyzed secondary structures of gene sequences for disturbed (mismatched) pairing across helices; we did not find such evidence. We also sent the gene sequences to the CHECK_CHIMERA program of the RDP (22, 24). CHECK_CHIMERA uncovered one chimeric gene among our clones: OCS12. The 5′ end of the OCS12 sequence showed a high similarity to the SAR1 and SAR95 sequences; however, the 3′ end was related to the *Methylomonas methylotrophus* sequence, in the γ subclass of the *Proteobacteria* (19, 47a). Therefore, only the first 1,005 bases of the OCS12 sequence were used in subsequent analyses. CHECK_CHIMERA found a second chimeric gene among published SAR11 cluster sequences: FL1, recovered by DeLong and colleagues (10). The 5′ end of FL1 (1,114 bases) was closely related to the SAR11 cluster sequence, but the 3′ end was related to the *Terrabacter tumescens* sequence, in the gram-positive division (9). The other SAR11 cluster sequences were most closely related to other SAR11 cluster sequences.

The probability of CHECK_CHIMERA detecting a chimera formed from closely related sequences is low (36). Because formation of a particular chimera is a very precise process involving both specific breakpoints and the joining of two (or more) specific sequences chosen from the diverse templates available, the probability of formation of identical or similar chimeras in different gene libraries is low. Therefore, we analyzed SAR11 cluster sequences from different clone libraries to test for independent recovery of each of the SAR11 cluster phylogenetic subgroups (Table 3). We considered a subgroup unlikely to be of chimeric origin if it contained full-length sequences from more than one gene library. The phylogenetic subgroup comprising SAR1 and its relatives contained five complete or nearly complete sequences and 10 partial sequences, representing seven independently constructed gene libraries from separate DNA samples. The subgroup comprising SAR11 and its relatives contained two full- or nearly full-length sequences from two independent gene libraries and six fragments from two more gene libraries. The SAR211 subgroup contained 2 full-length sequences and 14 partial sequences, from six different gene libraries, covering both ends of the gene. The SAR203 subgroup contained only one full-length sequence and one fragment; thus only a 230-bp segment of this sequence type was independently recovered.

Another way to confirm the existence in nature of subgroups within a gene cluster is to show that these subgroups are separated in space or time: in other words, that they occupy different niches. To do this, we examined the distribution of SAR11 cluster genes across a depth profile, since we had hypothesized a depth-specific distribution for some of the sequence types. We designed phylogenetic-group-specific probes to target different parts of the phylogenetic tree. Because we could not predict what parts of the phylogenetic tree might turn out to have specific distributions, our probes (Table 2) were designed with a nested range of specificities (Fig. 4) (34). We tested the probes for specificity by hybridizing them to control DNAs from sequenced clones (Fig. 5).

We used the SAR11 probes to analyze a depth profile of rDNAs amplified from bulk nucleic acid samples collected at the BATS on 12 August 1991. DNA samples from seven depths were amplified with primers for the domain *Bacteria*, and the resulting products were hybridized sequentially to the SAR11 probes and to a universal probe. The proportion of genes that hybridized to each SAR11 probe was expressed as a ratio, with

TABLE 3. Phylogenetic subgroups among SAR11 cluster 16S rRNA genes[a]

Name (reference)	Accession no.[c]	Origin	Depth (m)	Position (*E. coli* numbering)
SAR1 (16)	X52280	Atlantic, Hydrostation S	Surface	20–1191
SAR95 (16)	M63812	Atlantic, Hydrostation S	Surface	49–1406
SAR407 (16)	U75253	Atlantic, BATS	80	8–1541
SAR425[b]	U75261	Atlantic, BATS	80	1161–1540
BDA1-25 (15)	L11942	Atlantic, near Bermuda	10	537–815
OM242[b]	U70689	Atlantic, Cape Hatteras	10	56–467
OM188[b]	U70687	Atlantic, Cape Hatteras	10	56–464
OCS-12[b]	U75252	Pacific, Oregon Coast	10	9–1005
FL11 (10)	L10935	Pacific, Santa Barbara Channel	10	14–1448
FL1 (10)	L10934	Pacific, Santa Barbara Channel	10	8–1114
ALO21 (41)	M64525	Pacific, ALOHA Station	Surface	307–499
ALO38 (41)	M64532	Pacific, ALOHA Station	Surface	307–499
ALO39 (41)	M64533	Pacific, ALOHA Station	Surface	307–499
NH16-1 (15)	L11949	Northeast Pacific	100	537–761
NH25-10 (15)	L11967	Northeast Pacific	100	537–817
SAR11 (16)	X52172	Atlantic, Hydrostation S	Surface	22–1191
SAR193[b]	U75649	Atlantic, BATS	250	54–613
BDA1-1 (15)	L11934	Atlantic, near Bermuda	250	537–815
BDA1-20 (15)	L11941	Atlantic, near Bermuda	10	537–765
BDA1-15 (15)	L11939	Atlantic, near Bermuda	10	537–752
NH16-2A (15)	L11961	Northeast Pacific	10	537–817
NH16-11 (15)	L11951	Northeast Pacific	100	537–763
NH25-4 (15)	L11974	Northeast Pacific	100	537–857
SAR211[b]	U75256	Atlantic, BATS	250	8–1542
SAR464[b]	U75254	Atlantic, BATS	80	8–1541
SAR466[b]	U75263	Atlantic, BATS	80	8–357
SAR440[b]	U75262	Atlantic, BATS	80	8–357
SAR414[b]	U75259	Atlantic, BATS	80	8–355, 1161–1540
SAR490[b]	U75264	Atlantic, BATS	80	8–357, 1161–1540
SAR418[b]	U75260	Atlantic, BATS	80	8–350, 1180–1540
SAR492[b]	U75265	Atlantic, BATS	80	54–516
BDA1-27 (15)	L11943	Atlantic, near Bermuda	10	537–741
OM239[b]	U70688	Atlantic, Cape Hatteras	10	49–319
OM136[b]	U70684	Atlantic, Cape Hatteras	10	49–494
OM258[b]	U70691	Atlantic, Cape Hatteras	10	60–551
OCS143[b]	U75266	Pacific, Oregon Coast	10	106–411
NH49-1 (15)	L11987	Northeast Pacific	500	537–756
SAR220[b]	U75257	Atlantic, BATS	250	8–1541
SAR241[b]	U75258	Atlantic, BATS	250	8–1542
BDA1-17 (15)	L11940	Atlantic, near Bermuda	10	537–790
SAR203[b]	U75255	Atlantic, BATS	250	537–772
NH29-3 (15)	L11982	Northeast Pacific	100	537–756

[a] Genes in boldface type are also shown in the tree in Fig. 1. Line spaces separate phylogenetic subgroups.

[b] This gene was first reported in this paper.

[c] GenBank.

the hybridization to the universal 338R probe providing the denominator.

We observed three patterns of distribution with depth: common at shallow depths only (SAR11-A1 and SAR11-A2), common in deep water only (SAR11-G4), and evenly distributed throughout the water column (all others). Figure 6 shows the results of hybridizing depth-specific amplification products to probes SAR11-A1, SAR11-A2, and SAR11-G4. The SAR11 rDNAs that hybridized to SAR11-A1 and SAR11-A2 were at maximal levels in surface samples. Below the mixed layer, levels dropped rapidly. The abundance of the genes that hybridized to SAR11-G1 was very low at depths of 120 m and above but increased to a maximum at 200 m.

The accuracy with which amplified rDNAs represent the cell types present in a particular collection, such as a depth sample, depends on whether all templates amplify with equal efficiency. We tested the hybridization results independently by hybridizing the SAR11 probes to environmental RNAs. The RNA sample set consisted of pairs of consecutive monthly samples collected from 0 and 200 m (Table 1). In addition, 10 depth profiles consisting of seven samples (some samples were lost) spanning the region from 0 to 250 m were collected during the same period.

The three depth profiles in Fig. 7 reflected variation in the relative abundance of the A1, D4, and G1 subgroups. These data confirmed the previous results based on amplified rDNAs. A1 was found to be most abundant at the surface, D4 abundance was constant with depth with the exception of one date, and G1 hybridization was greatest at the bottom of the ocean surface layer (250 m). Time series data provided statistical

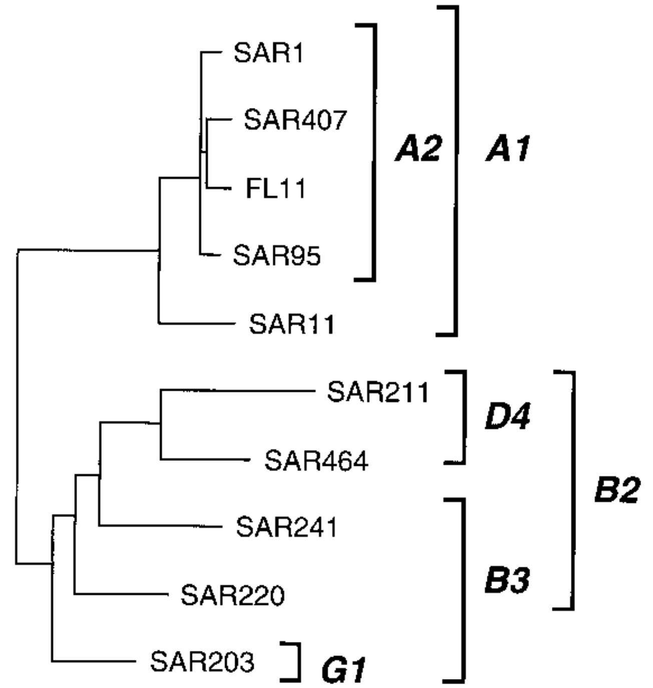

FIG. 4. Specificity of SAR11 cluster oligonucleotide probes to lineages in the phylogenetic tree.

support for these conclusions. The hypothesis that A1 accounted for a higher proportion of total rRNA at 0 m than at 200 m was supported at $P = 2.7 \times 10^{-5}$ in a one-tailed *t* test, assuming unequal variances. There was no statistically significant variation in the distribution of D4 hybridization with depth, although on one date (BATS sample no. 52) it was much more abundant in the lower region of the surface layer. The hypothesis that G1 accounted for a greater proportion of the total rRNA at 200 m than at the surface was supported at $P = 0.005$.

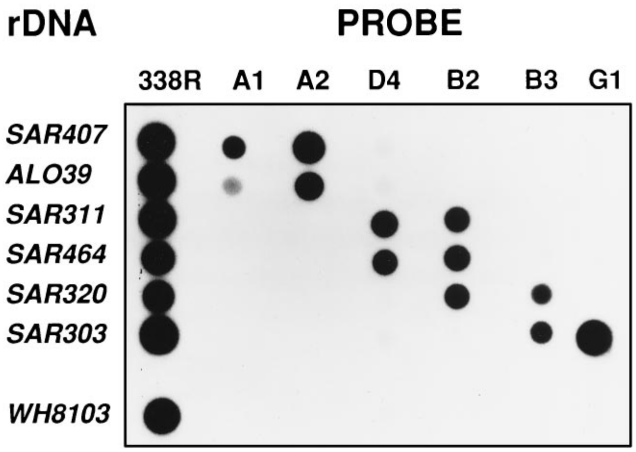

FIG. 5. Probe specificity determined by dot blot hybridization. Thirty nanograms of each control DNA (PCR of 16S rRNA from a known clone) per dot was immobilized on a nylon support membrane and hybridized to ^{32}P-labeled probes. Clones SAR407, ALO39, SAR211, SAR464, SAR220, and SAR203 are members of the SAR11 cluster. WH103 is a marine *Synechococcus*. Probe 338R, a positive control, is a universal bacterial probe.

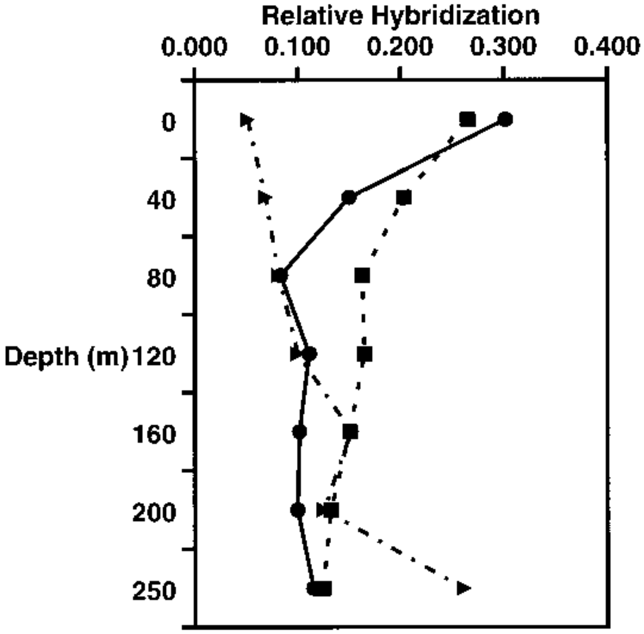

FIG. 6. The distribution of SAR11 cluster rDNAs in the upper 250 m of the water column at BATS during a stratified period in August 1991, as shown by hybridization of rDNA amplification products from depth samples to group-specific oligonucleotide probes SAR11-A1, SAR11-A2, and SAR11-G1. Hybridization is expressed relative to hybridization to a universal bacterial probe, 338R or 1406R; thus, these values represent the proportion of the specific SAR11 cluster genes among the total amplified bacterial genes. ■, SAR11A1 and 338R; ●, SAR11A2 and 338R; ▲, SAR11G1 and 1406R/20.

DISCUSSION

Genetic variation within a gene cluster recovered by PCR from environmental DNA could result from artifacts and errors during sequence recovery or from genetic and evolutionary processes. Possible sources of errors in genes recovered by PCR include nucleotide misincorporation by DNA polymerase and chimera formation. We ruled out polymerase error as an important cause of variation in the SAR11 cluster genes by examination of the location of mutations and secondary structure analysis. Chimera analyses revealed only two instances of chimeras formed between a SAR11 cluster sequence and an unrelated sequence. The generally low bootstrap values supporting some of the branching orders within the phylogenetic trees suggest the possibility of PCR chimeras among closely related sequences as well. However, the likelihood that identical chimeras could be created independently is very low, unless one postulates an unknown systematic mechanism of formation. Therefore, if the same gene has been recovered more than once, it is likely to represent an extant lineage, not a PCR chimera. Some of the SAR11 lineages have repeatedly been recovered in independent clone libraries. Other lineages (e.g., SAR203) have only been recovered more than once over a short portion of sequence, leaving the question of chimera formation in the full-length sequence unanswered for them.

Because the SAR11 cluster sequences were obtained by cloning, each sequence represents a single *rrn* operon. Members of a gene family are homogenized by the processes of concerted evolution, including unequal crossing over and gene conversion (e.g., see references 2 and 28), which could produce a recombination-like pattern of sequence diversity. We calculated the pairwise similarities between seven operon sequences from a single strain of *E. coli* (8); they ranged from 98.9 to

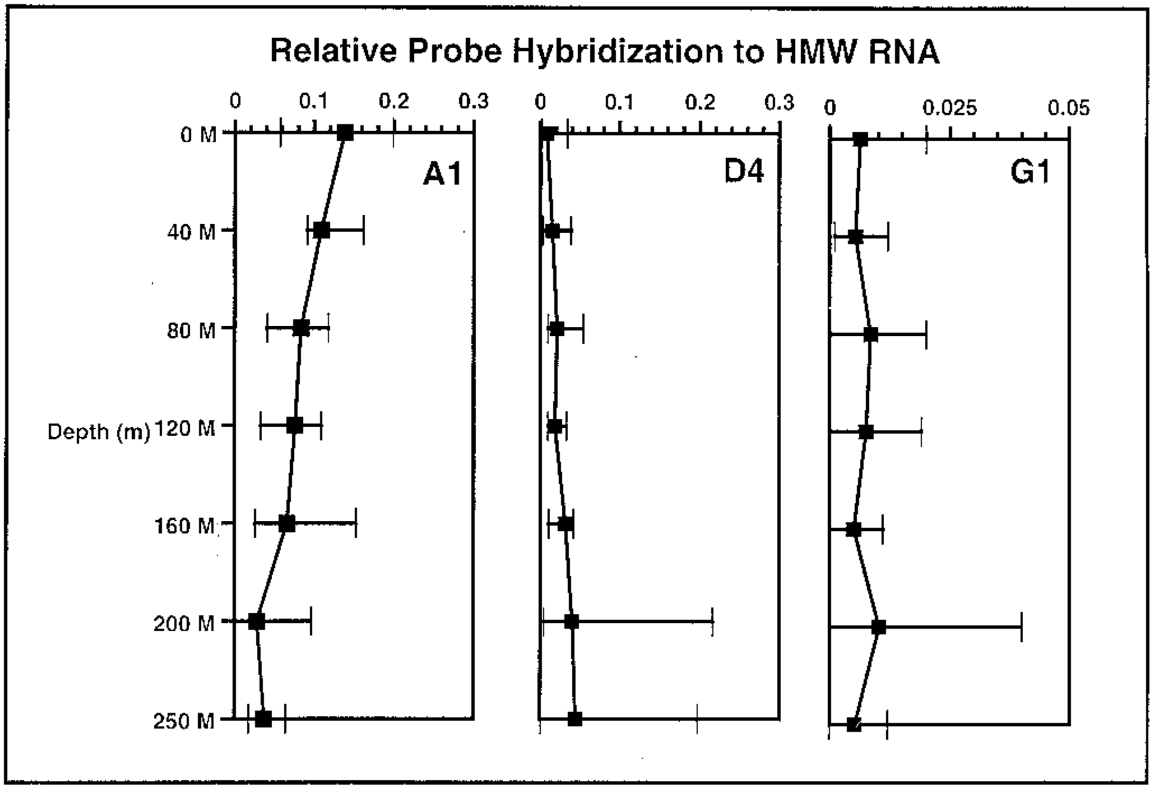

FIG. 7. Average hybridization of subcluster-specific probes to picoplankton high-molecular-weight (HMW) RNA from 10 depth profiles collected between August 1991 and July 1993 in the western Sargasso Sea. Relative hybridization values are the ratios of subcluster-specific probe hybridization to the universal bacterial (338R) probe hybridization. The error bars indicate standard deviations between profiles and, thus, include environmental variation.

99.9%. The variation in these operons was localized in a few regions, and the distribution of variation among operons was mosaic, suggesting that gene conversions had occurred in random directions over very short domains within each gene (8). For comparison, pairwise similarities between the 12 SAR11 cluster sequences shown in Fig. 1 ranged from 89.9 to 99.3%. If the amount of microheterogeneity among *E. coli* operon sequences is typical, then genetic variation within the SAR11 cluster is far greater than would be expected due to within-cell microheterogeneity alone. It seems likely that some of the most similar SAR11 gene cluster sequences represent different *rrn* operons within single SAR11 strains. The more divergent sequences are likely to represent different strains and species, whose rRNA genes have undergone repeated random fixation of mutations and gene conversions during macroevolutionary divergence.

To test whether genetic variation in the SAR11 gene cluster was correlated with macroevolution, we looked for niche partitioning among sequence types. We used group-specific oligonucleotide probes for these experiments. The impact of chimeric genes and recombination on the design of probes for environmental studies is quite different from their effect on phylogenetic tree reconstruction, because oligonucleotide probes hybridize to very small sequence domains. Since recombination is a function of the distance between loci, small sequence domains are less affected by this phenomenon than large sequence domains. Furthermore, the probes we describe are all based on multiple gene sequences (except for SAR11-G3, for which only one sequence was available in the targeted region).

Our initial comparisons of genes from surface and 250-m samples suggested that depth might be an important environmental variable in the evolution of SAR11 cluster groups. Therefore, the first environmental variable we examined was depth. We probed two types of depth profiles from the Sargasso Sea: amplified rDNAs and bulk RNAs. In both, we found two SAR11 cluster groups with strong depth-specific distributions. PCR chimeras and recombination among SAR11 cluster lineages would not affect these results, since the probes target real marker sequences that exist in the ocean. The probe data showed the existence and relative importance of these marker sequences in depth profiles. Differential, depth-specific expression of members of an rRNA gene family within a single species could not explain the depth-specific distribution, because hybridization to PCR products, which would not be influenced by levels of gene expression, supported the same distributions.

The depth profiles strongly suggest niche partitioning; thus, some of the genetic variability within the SAR11 rRNA gene cluster may been fixed during evolution due to selection.

We conclude, first, that some of the variation observed within gene clusters recovered from environmental DNA corresponds to bacterial "speciation" with depth. Second, although artifacts caused by PCR and microheterogeneities among gene families constitute important, poorly understood limitations to PCR-based approaches to bacterial phylogeny, it is possible to uncover ecologically significant variation in gene sequences by PCR and gene cloning approaches. Molecular studies, therefore, remain the most effective means of understanding the natural history of uncultured microbial groups such as the SAR11 gene cluster.

ACKNOWLEDGMENTS

We are grateful to the Bermuda Biological Station BATS group and the crew of the RV *Weatherbird II* for support. We thank Nanci Adair for technical assistance. Lee Kerkhoff and Paul Kemp of Brookhaven National Laboratories provided DNA for the OM (Cape Hatteras) clone library.

This work was supported by National Science Foundation grant OCE 9016373 and Department of Energy grant FG0693ER61697.

REFERENCES

1. **Amann, R., J. Snaidr, M. Wagner, W. Ludwig, and K.-H. Schleifer.** 1996. In situ visualization of high genetic diversity in a natural microbial community. J. Bacteriol. **178:**3496–3500.
2. **Arnheim, N.** 1983. Concerted evolution of multigene families, p. 38–61. *In* M. Nei and R. K. Koehn (ed.), Evolution of genes and proteins. Sinauer Associates, Sunderland, Mass.
3. **Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace.** 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA **91:**1609–1613.
4. **Bascuñana, C. R., J. G. Mattsson, G. Bölske, and K. E. Johansson.** 1994. Characterization of the 16S rRNA genes from *Mycoplasma* sp. strain F38 and development of an identification system based on PCR. J. Bacteriol. **176:**2577–2586.
5. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7:**1513–1523.
6. **Britschgi, T. B., and S. J. Giovannoni.** 1991. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. **57:**1313–1318.
7. **Choi, B. K., B. J. Paster, F. E. Dewhirst, and U. B. Gobel.** 1994. Diversity of cultivable and uncultivable oral spirochaetes from a patient with severe destructive periodontitis. Infect. Immun. **62:**1889–1895.
8. **Cilia, V., B. Lafay, and R. Christen.** 1996. Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. Mol. Biol. Evol. **13:**451–461.
9. **Collins, M. D., M. Dorsch, and E. Stackebrandt.** 1989. Transfer of *Pimelobacter tumescens* to *Terrabacter* gen. nov. as *Terrabacter tumescens* comb. nov. and of *Pimelobacter jensenii* to *Nocarioides* as *Nocarioides jensonii* comb. nov. Int. J. Syst. Bacteriol. **39:**1–6.
10. **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.
11. **Dryden, S. C., and S. Kaplan.** 1990. Localization and structural analysis of the ribosomal RNA operons of *Rhodobacter spheroides*. Nucleic Acids Res. **18:**7267–7277.
12. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evol. **39:**783–791.
13. **Felsenstein, J.** 1991. PHYLIP. University of Washington, Seattle.
14. **Fleischmann, R. D., et al.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496–512.
15. **Fuhrman, J. A., K. McCallum, and A. A. Davis.** 1993. Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific oceans. Appl. Environ. Microbiol. **59:**1294–1302.
16. **Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field.** 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature **345:**60–63.
17. **Gordon, D. A., and S. J. Giovannoni.** 1996. Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific oceans. Appl. Environ. Microbiol. **62:**1171–1177.
18. **Hales, B. A., C. Edwards, D. A. Ritchie, G. Hall, R. W. Pickup, and J. R. Saunders.** 1996. Isolation and identification of methanogen-specific DNA from blanket bog peat by PCR amplification and sequence analysis. Appl. Environ. Microbiol. **62:**668–675.
19. **Kimura, M.** 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.
20. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** 1994. Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. Appl. Environ. Microbiol. **60:**746–748.
21. **Lane, D. J.** 1991. 16S/23S rRNA sequencing, p. 115–148. *In* E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. John Wiley and Sons, New York, N.Y.
22. **Larsen, N., G. J. Olsen, B. L. Maidak, M. J. McCaughey, R. T. Overbeek, J. Macke, T. L. Marsh, and C. R. Woese.** 1993. The Ribosomal Database Project. Nucleic Acids Res. **21:**3021–3023.
23. **Liesack, W., H. Weyland, and E. Stackebrandt.** 1991. Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria. Microb. Ecol. **21:**191–198.
24. **Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K. Fogel, J. Blandy, and C. R. Woese.** 1994. The Ribosomal Database Project. Nucleic Acids Res. **22:**3485–3487.
25. **Mullins, T. D., T. B. Britschgi, R. L. Krest, and S. J. Giovannoni.** 1995. Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. Limnol. Oceanogr. **40:**148–158.
26. **Mylvaganam, S., and P. P. Dennis.** 1992. Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaeobacterium *Haloarcula marismortui*. Genetics **130:**399–410.
27. **Ohkuma, M., and T. Kudo.** 1996. Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. Appl. Environ. Microbiol. **62:**461–468.
28. **Ohta, T.** 1991. Multigene families and the evolution of complexity. J. Mol. Evol. **33:**34–41.
29. **Olsen, G. J., D. L. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl.** 1986. Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. **40:**337–366.
30. **Pedersen, K., J. Arlinger, L. Hallbeck, and C. Pettersson.** 1996. Diversity and distribution of subterranean bacteria in groundwater at Oklo in Gabon, as determined by 16S rRNA gene sequencing. Mol. Ecol. **5:**427–436.
31. **Pettersson, B., K. E. Johansson, and M. Uhlen.** 1994. Sequence analysis of 16S rRNA from mycoplasmas by direct solid-phase sequencing. Appl. Environ. Microbiol. **60:**2456–2461.
32. **Pettersson, B., T. Leitner, M. Ronaghi, G. Bölske, M. Uhlen, and K. E. Johansson.** 1996. Phylogeny of the *Mycoplasma mycoides* cluster as determined by sequence analysis of the 16S rRNA genes from the two rRNA operons. J. Bacteriol. **178:**4131–4142.
33. **Rappé, M. S., P. F. Kemp, and S. J. Giovannoni.** 1995. Chromophyte plastid 16S ribosomal RNA genes found in a clone library from Atlantic Ocean seawater. J. Phycol. **31:**979–988.
34. **Raskin, L., J. M. Stromley, B. E. Rittman, and D. A. Stahl.** 1994. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl. Environ. Microbiol. **60:**1232–1240.
35. **Reysenbach, A. N., G. S. Wickham, and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. Appl. Environ. Microbiol. **60:**2113–2119.
36. **Robison-Cox, J. F., M. M. Bateson, and D. M. Ward.** 1995. Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. Appl. Environ. Microbiol. **61:**1240–1245.
37. **Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich.** 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:**487–491.
38. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
39. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
40. **Sanger, F., S. Nicklen, and R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
41. **Schmidt, T. E., E. F. DeLong, and N. R. Pace.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. **173:**4371–4378.
42. **Schmidt, T. M.** Multiplicity of ribosomal RNA operons on prokaryotic genomes. *In* F. J. deBruijn, J. R. Lupski, and G. Weinstock (ed.), Bacterial genomes: physical structure and analysis, in press.
43. **Schuldiner, A. R., A. Nirula, and J. Roth.** 1989. Hybrid DNA artifact from PCR of closely related target sequences. Nucleic Acids Res. **17:**4409.
44. **Swofford, D. L.** 1991. PAUP, version 3. Illinois Natural History Survey Champaign.
45. **Ward, D. M., M. M. Bateson, R. Weller, and A. L. Ruff-Roberts.** 1992. Ribosomal RNA analysis of microorganisms as they occur in nature. Adv. Microb. Ecol. **12:**219–286.
46. **Ward, D. M., R. Weller, and M. Bateson.** 1990. 16S rRNA sequences reveal numerous uncultured organisms in a natural community. Nature **345:**63–65.

46a. **Whitmore, S.** Unpublished results.

47. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.

47a. **Woese, C. R.** Unpublished sequence in the Ribosomal Database Project.

48. **Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson.** 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proc. Natl. Acad. Sci. USA **77:**2158–2162.