doi:10.1111/j.1462-2920.2004.00712.x

Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition

David T. Kysela,[†] Carmen Palacios and Mitchell L. Sogin*
The Josephine Bay Paul Center, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA.

Summary

Serial analysis of ribosomal sequence tags (SARST) is a novel technique for characterizing microbial community composition. The SARST method captures sequence information from concatemers of short 16S rDNA polymerase chain reaction (PCR) amplicons from complex populations of DNA. Here, we describe a similar method, serial analysis of V6 ribosomal sequence tags (SARST-V6), which targets the V6 hypervariable region of bacterial 16S rRNA genes. The SARST-V6 technique exploits internal primer sequences to generate compatible restriction digest overhangs, thereby improving upon the efficiency of SARST. Serial analysis of V6 ribosomal sequence tags of bacterial community composition in hydrothermal marine sediments from Guaymas Basin resembled results of cloning and sequencing of single, full-length PCR products from ribosomal RNA genes of the same microbial community. Both methods identified the same major bacterial groups, but only SARST-V6 recovered thermodesulfobacteria and γ-proteobacteria sequences, while only full-length PCR product cloning recovered candidate division OP11 sequences. There were differences in the relative frequencies of some phylotypes. The disparities reflect differences in the amplicon pool obtained during initial amplification that may result from different primer affinities or DNA degradation. These results demonstrate the utility of SARST-V6 in collecting taxonomically informative data for high-throughput analysis of microbial communities.

Introduction

The ability to characterize genes rapidly from natural populations of microorganisms has revolutionized microbial ecology. By sequencing the products of polymerase chain reaction (PCR) primed with oligonucleotides that target phylogenetically conserved sequences, it is possible to evaluate microbial community composition in a given environment without cultivating microorganisms in the laboratory. Analyses of PCR products from ribosomal RNA (rRNA) coding regions have proven particularly informative. These genes are conserved in all organisms, and molecular databases contain more than 90 000 rRNA sequences from diverse microbial forms (http://rdp.cme.msu.edu). Through proper primer design, it is possible to amplify rRNA genes from representatives from each of the primary lines of descent, of the major phyla, or of particular genera. This window on the microbial world has revealed new levels of largely unexplored microbial diversity not represented in laboratory cultures. Sequence studies continue to identify novel diversity within described taxa, as well as deep-branching, basal diversity in all three domains of life (Dojka *et al*., 1998,2000; Hugenholtz *et al*., 1998; Cifuentes *et al*., 2000; Lopez-Garcia *et al*., 2001; Moon-van der Staay *et al*., 2001; Dawson and Pace, 2002; Edgcomb *et al*., 2002; Teske *et al*., 2002).

Ribosomal RNA gene sequence analysis provides a detailed picture of microbial community composition. However, even in low-cost, high-throughput laboratories, the efficiency and output of sequencing methods are inadequate for population structure studies that seek to determine relative numbers of different kinds of microorganisms in an environmental sample. DNA fingerprinting techniques such as denaturing gradient gel electrophoresis (DGGE) (Muyzer *et al*., 1993) and terminal restriction fragment length polymorphism (T-RFLP) (Moeseneder *et al*., 1999) offer higher throughput and estimates of relative frequencies for different kinds of organisms, but little or no direct phylogenetic information (also see Cole *et al*., 2003). Taxonomic associations of particular bands or peaks in all commonly used fingerprinting studies ultimately rely on DNA sequencing to characterize phylotypes (Bending *et al*., 2003; McBain

Received 1 April, 2004; revised 26 July, 2004; accepted 27 July, 2004.
*For correspondence. E-mail sogin@evol5.mbl.edu; Tel. 508-289-7246; Fax 508-457-4727. [†]Present address: Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06511, USA.

et al., 2003; Nagashima *et al*., 2003). Probe hybridization methods, such as fluorescent *in situ* hybridization (FISH) (Amann *et al*., 1995) and microarrays (Loy *et al*., 2002), permit detection of targeted microbial taxa. However, these methods require careful probe selection and therefore typically examine only a limited subset of a complex microbial community.

A recently described method, serial analysis of ribosomal sequence tags (SARST) (Neufeld *et al*., 2004), draws upon information-rich DNA sequence analysis, while providing higher throughput and efficiency than standard sequencing techniques. The technique is similar to serial analysis of gene expression (SAGE), which describes relative expression levels for genomic tags in mRNA populations (Velculescu *et al*., 1995). The SARST method produces sequences of large concatemers of PCR-amplified ribosomal sequence tags (RSTs) from homologous V1 hypervariable regions in rRNA genes. Comparison against a comprehensive rRNA gene database identifies the taxonomic assignment of individual RSTs in the concatemers. Serial analysis of ribosomal sequence tags thus evaluates the diversity and relative numbers of different rDNA amplicons from a heterogeneous microbial population.

We have simultaneously and independently developed a SARST-like method, termed serial analysis of V6 ribosomal sequence tags (SARST-V6), which targets the 16S rDNA V6 hypervariable region. This approach also provides sequence data from short PCR product concatemers. However, unlike SARST, SARST-V6 generates compatible single-stranded DNA overhangs within the PCR priming regions. The SARST-V6 method therefore requires less enzymatic reactions and correspondingly fewer purification steps, thereby reducing cost and handling time. In order to evaluate the consistency between SARST-V6 and typical 16S rDNA PCR product sequencing, we compare SARST-V6 data and full-length 16S rDNA sequence data obtained from a hydrothermal vent community in Guaymas Basin (Gulf of California, Mexico).

Results

SARST-V6 overview

Figure 1 provides an overview of SARST-V6 and its application to studies of rRNA sequence variation in microbial populations. The biotinylated primers in Fig. 1 are complementary to conserved regions in 16S rRNA coding regions that flank hypervariable region V6. Each primer contains a linker with the recognition site (5′-3′: GTGCAG) for the type IIS restriction enzyme *Bsg*I. Subsequent to amplification, digestion with *Bsg*I cleaves 16 bases down-

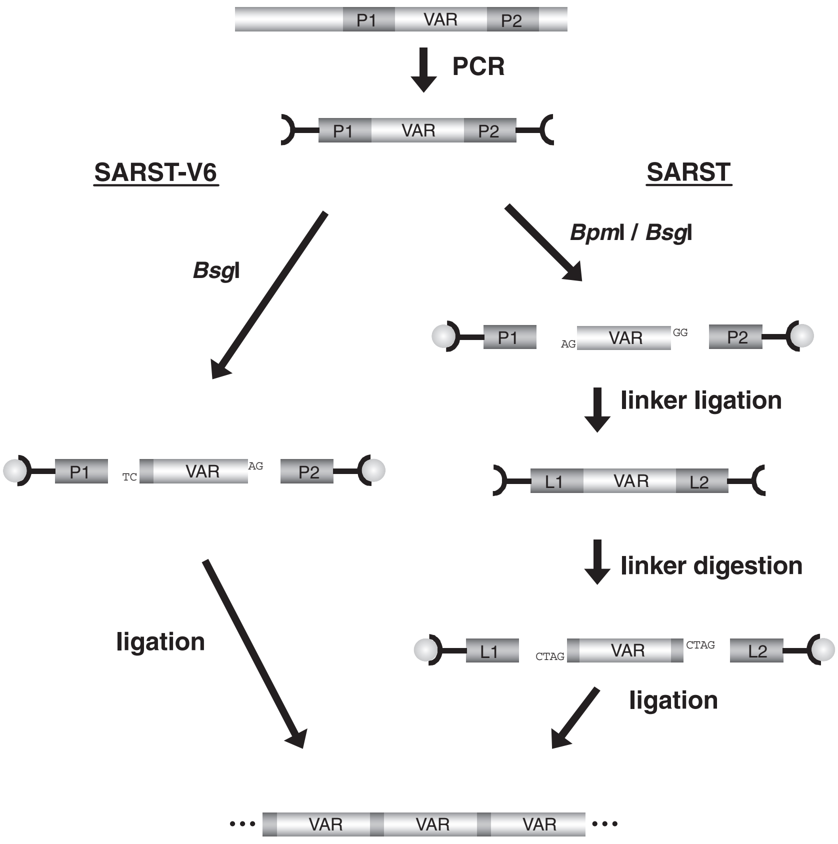

Fig. 1. Overview of the serial analysis of V6 ribosomal sequence tags (SARST-V6) method and comparison to SARST (Neufeld *et al*., 2004). Biotinylated PCR primers target conserved sites flanking the V6 (SARST-V6) or V1 (SARST) hypervariable region of the 16S rRNA gene. Primers include a 5′ linker region containing the recognition sequence for the type IIS restriction enzyme *Bsg*I or *Bpm*I. Type IIS restriction digestion results in cleavage downstream of the recognition sequence, leaving a 2-bp overhang at each end of the amplicon. Digested termini are purified away using magnetic streptavidin-coated beads. For SARST, but not SARST-V6, additional linker ligation and digestion steps followed by streptavidin purification are required in order to generate compatible overhangs. In both methods, ligation of digested amplicons yields concatemers with multiple, serially arranged PCR products. Sequence regions are not to scale.

stream on the sense strand and 14 bases downstream on the antisense strand. This step increases the information in each concatemer by as much as 30% relative to undigested PCR products and generates cohesive ends for ligation. To maximize ligation efficiency, treatment with streptavidin-coated beads removes undigested amplicons and digested priming region fragments. Treatment with DNA ligase assembles the digested amplicons into 2–3 kb concatemers that serve as templates in standard sequencing reactions. The embedded punctuation (11 residual bases from the primers) defines the boundaries of individual gene tags (hypervariable amplicons within the concatemers), which serve as query sequences in BLAST searches of the GenBank database. The taxonomic affiliation and frequency of recovery for a given tag provide an estimate of gene diversity and relative representation of specific amplicons from a DNA sample.

SARST-V6 Guaymas bacterial community analysis

We carried out two replicate SARST-V6 protocols for the A1 core sample of Guaymas Basin. These experiments generated 526 (274 and 252 respectively) complete RSTs from unidirectional sequence reads of 98 (56 and 42 respectively) cloned concatemers. All terminal RSTs in the concatemers lacked the residual primer sequence (AGAACCTTACC), which is consistent with either non-specific digestion or the presence of an internal restriction site for *Bsg*I within the original SARST-V6 amplicons. Our analyses did not include these terminal RST sequences. Based on BLAST searches of the GenBank 'nt' non-redundant nucleotide database, RSTs showed taxonomic affinities with hypervariable regions from 370 bacterial and nine archaeal rRNA genes. The average length of taxonomically assigned RSTs was 65.6 ± 8.5 bp (mean ± SD) with a range of 35–99 bp. Seventy-two of the RSTs were similar to bacterial sequence entries in GenBank that lacked descriptors at the level of genus and class. We inferred the taxonomy for 66 of these sequences from previously published phylogenies that included the matching GenBank sequences (Hugenholtz *et al*., 1998; Tanner *et al*., 2000; Derakshani *et al*., 2001; Madrid *et al*., 2001; Reed *et al*., 2002; Teske *et al*., 2002). We were unable to determine taxonomy for the remaining six GenBank entries. A total of 146 hypervariable amplicons did not retrieve database sequence matches that satisfied our search criteria (see *Experimental procedures*), although 95 of these sequences displayed matches to rRNA genes with *e*-values better than 10^{-2} in BLASTN searches of GenBank. Six of these unidentified rRNA gene phylotypes were recovered independently in both of the replicate SARST-V6 experiments, indicating that they were not the products of non-specific or chimeric amplification.

Figure 2 describes data from the SARST-V6 bacterial analysis of sediment sample A1. The most frequent sequence types include proteobacteria from the γ- and δ-subdivisions, as well as thermodesulfobacteria, green non-sulfur bacteria, and clones from the Guaymas bacterial group (Teske *et al*., 2002). Most RSTs affiliate with 'uncultured' or 'unassigned' GenBank entries. However, 93 of 96 γ-proteobacteria sequences matched with a single, cultured representative of the genus *Beggiatoa*

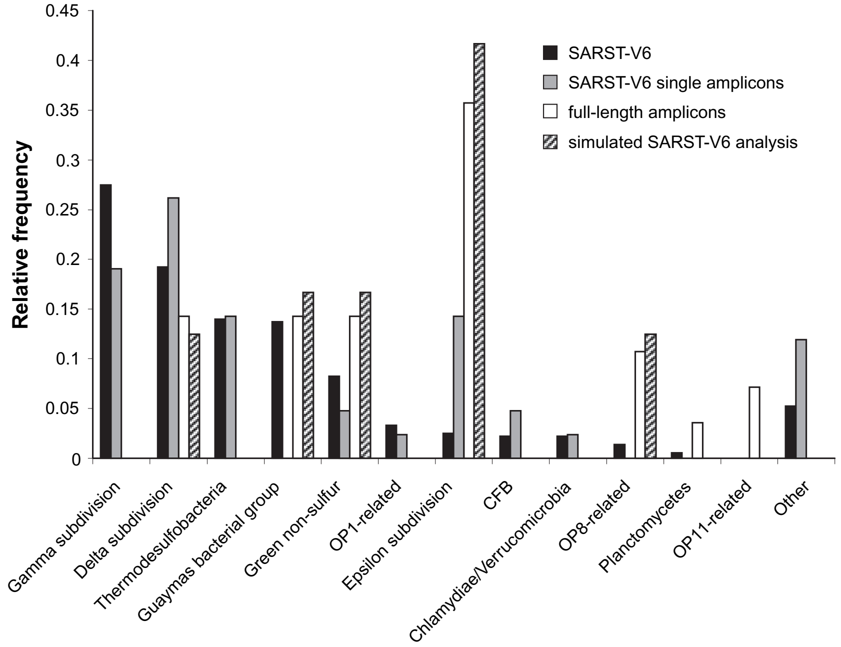

Fig. 2. Bacterial phylotype distributions obtained using SARST-V6 and other analysis methods. All data were collected on hydrothermally heated sediments from the Guaymas Basin (Gulf of California). Results of four analyses are shown: SARST-V6 (black; *n* = 364), single amplicon cloning using SARST-V6 primers (grey, *n* = 42), phylogenetic analysis of full-length 16S rRNA gene clone sequences (white, *n* = 28) and a simulated SARST-V6 analysis of full-length sequences (hatched, *n* = 24). Except for phylogenetic analysis of full-length gene sequences, taxonomic assignments were based on results of BLAST searches of the GenBank 'nt' database.

(AF035956), consistent with the presence of an abundant *Beggiatoa* mat phylotype. Distributions of RST types were consistent between the two trials of the SARST-V6 protocol (χ^2 similarity = 0.984; data not shown).

Comparisons of SARST-V6 with full-length 16S rRNA analyses

To evaluate potential bias of RSTs represented in larger concatemers, we constructed a clone library of individual RST amplicons generated using the primers SARST967F and SARST1046R. Sequence analysis of 64 RST single-amplicon clones revealed 42 bacterial sequences. Four of these bacterial sequences matched GenBank-entries-lacking class or genus-taxonomic identifiers. We assigned the taxonomy of these RSTs based on published phylogenies of the matching database sequences (LaPara *et al*., 2000; Alfreider *et al*., 2002; Teske *et al*., 2002). The resulting phylotype distribution closely resembled that obtained with SARST-V6 (χ^2 similarity = 0.958; Fig. 2).

In order to benchmark SARST-V6, we compared our results with full-length 16S rRNA amplicon cloning data collected (Teske *et al*., 2002) from the same microbial community. We determined the taxonomy of 28 full-length sequences from core A1 based on their published phylogeny (Teske *et al*., 2002). We also performed a simulated SARST-V6 analysis of the 28 full-length clone sequences by sampling only the portion of the 16S rRNA gene examined by SARST-V6. BLAST searches of the GenBank ‘nt’ database resulted in taxonomic assignments for 24 bacterial RSTs, while four sequences remained unassigned. Seven bacterial RSTs matched GenBank-sequences-lacking genus or class-taxonomic identifiers, and taxonomy was inferred from published phylogenies for these database sequences (Teske *et al*., 2002). This simulated SARST-V6 analysis produced a taxonomic profile similar to phylogenetic analysis of the complete gene sequences (χ^2 similarity = 0.935; Fig. 2). However, the simulated SARST-V6 analysis failed to detect the phylogenetic association of two sequences with candidate division OP11.

During this analysis, we identified one GenBank entry whose taxonomic assignment conflicted with phylogenetic analysis of the corresponding sequence from core A1: AB013265, classified in GenBank as a member of the ε-proteobacteria, affiliated with δ-proteobacteria clone A1_B009. Phylogenetic analysis using maximum parsimony and minimum evolution distance criteria was performed for this GenBank sequence as described by Teske and colleagues (2002). In all analyses, sequence AB013265 grouped within the δ-proteobacteria clade (data not shown). We adjusted the taxonomic assignment of this sequence accordingly in our analyses.

Although SARST-V6 and complete gene sequencing identified most of the same major bacterial groups (Fig. 2), SARST-V6 frequently recovered Thermodesulfobacteria and γ-proteobacteria sequences (14.5% and 27.3% relative frequency respectively), but full-length gene cloning studies failed to detect these taxonomic groups. Furthermore, sequences affiliated with candidate division OP11 (7.1%) in the phylogenetic analysis by Teske and colleagues (2002) did not appear in the SARST-V6 data set. Among the sequence types common to both SARST-V6 and full-length clones, certain differences emerged in the relative frequencies of major bacterial groups. ε-Proteobacteria in particular were more commonly recovered by phylogenetic analysis (35.7%) and simulated SARST-V6 analysis (41.7%) of full-length 16S rRNA gene clones than by SARST-V6 (2.5%). Candidate division OP8-related clones also appeared more frequently in full-length phylogenetic analysis (10.7%) and simulated SARST-V6 analysis (12.5%) than among SARST-V6 sequences (1.4%). Comparisons of phylotype distributions of SARST-V6 analysis with both phylogenetic and the simulated SARST-V6 analyses of full-length clone sequences revealed moderate similarity (χ^2 similarity = 0.893 and 0.893 respectively). Single amplicon cloning using SARST-V6 primers also yielded a phylotype distribution moderately similar to the phylogenetic (χ^2 similarity = 0.824) and the simulated SARST-V6 (χ^2 similarity = 0.817) analyses.

Discussion

Serial analysis of V6 ribosomal sequence tags closely resembles the recently described SARST method for characterizing environmental DNA samples (Neufeld *et al*., 2004). As demonstrated for SARST (Neufeld *et al*., 2004), SARST-V6 permits efficient, detailed and reproducible analysis of microbial community composition. By combining hypervariable sequences from multiple 16S rRNA genes into a single compound DNA template, we can generate descriptions of microbial biodiversity for many taxa in each sequencing reaction. In the current study, we observed an average of 5.4 sequence tags per concatemer, and improved concatemerization efficiency (i.e. longer concatemer templates) should yield even greater efficiency. Although fingerprinting methods such as DGGE and T-RFLP also afford high throughput, these techniques typically rely on follow-up sequencing for taxonomic assignments. Ribosomal sequence tags provide direct taxonomic information when matched to existing sequence database entries. Serial analysis of V6 ribosomal sequence tags thus provides more phylotype information than DNA fingerprinting methods, while affording higher throughput than sequencing full-length PCR products.

While both SARST-V6 and SARST employ type IIS restriction enzymes to eliminate primer sequences from PCR amplicons, SARST-V6 generates compatible, single-stranded DNA overhangs within the primer region (Fig. 1). This approach therefore requires a single restriction digest step, whereas SARST requires additional linker ligation, digestion and biotin purification prior to RST concatemerization (Neufeld *et al*., 2004).

Our simulated SARST-V6 of full-length 16S rRNA sequences demonstrates the taxonomic resolving power of the V6 hypervariable region. Despite the relatively short length of this gene segment, we obtained comparable phylotype distributions using both phylogenetic analysis of full-length rRNA gene sequences and a simulated SARST-V6 of the same data. This result indicates that the V6 hypervariable region is sufficient for identifying and estimating relative representation for taxonomic groups represented in an existing database.

Although SARST and SARST-V6 provide sequence-based phyletic information as RSTs, the taxonomic assignment of these data relies on information from similar sequences rather than direct phylogenetic analysis of RST sequences. The highly variable nature of these sequences and paucity of positions provide only low levels of evolutionary information for phylogenetic inferences. However, RST queries usually identify similar or nearly identical hypervariable sequences within the context of longer database entries. It is the phylogenetic affiliation of these longer sequences that allows us to assign phyletic positions for the corresponding RSTs. Accordingly, taxonomic assignments rely on the completeness and accuracy of the database used in BLAST searching. It will therefore prove difficult to identify sequences from poorly described taxa using SARST-V6, but future improvements in the comprehensiveness of sequence databases should simplify such assignments.

We are encouraged by the general overlap between microbial community profiles obtained using SARST-V6 and full-length 16S rRNA gene amplicon cloning. However, we observed some differences between the phylotype distributions obtained using these methods. Although Thermodesulfobacteria and *Beggiatoa* are known to inhabit hydrothermal vent sites in Guaymas Basin (Nelson *et al*., 1989; Jeanthon *et al*., 2002), sequences affiliated with these taxa were only recovered using SARST-V6 and not by full-length 16S rRNA gene cloning (Teske *et al*., 2002). Conversely, sequence types affiliated with ε-proteobacteria, candidate division OP8 and candidate division OP11 were prevalent in the full-length gene clone library but rare (ε-proteobacteria and OP8) or absent (OP11) from the SARST-V6 dataset*.* Although the small sample size of the full-length sequence dataset might explain some differences, similar patterns also emerge in a sample of 47 sequences from the A2 (1–2 cm sediment depth) sample (Teske *et al*., 2002). Database bias might underlie inconsistencies (discussed above), but the close overlap between phylogenetic analysis and simulated SARST-V6 of full-length sequences suggests little such effect. Instead, these disparities appear to result from amplification differences between the SARST-V6 and full-length primers, because single amplicon cloning using SARST-V6 primers yielded a community profile more consistent with the SARST-V6 data than with data obtained by full-length amplicon cloning.

Two explanations may account for such contrasting gene amplification patterns. First, there are likely to be differences in target specificity between the primers used in SARST-V6 and those used for full-length gene amplification. The priming regions used in Teske and colleagues (2002) for complete 16S rDNA amplification remain undetermined for Thermodesulfobacteria and *Beggiatoa* (Teske *et al*., 1996; Ahmad *et al*., 1999; Teske, 1999). The absence of these taxa among full-length sequence clones might reflect weak affinity of the primers for whole gene amplification. It is noteworthy that in a previous study, amplification of complete 16S rDNA from *Beggiatoa* using the terminal priming regions proved difficult (Teske, 1999). Conversely, the SARST-V6 primers contain mismatches to database sequences of ε-proteobacteria and candidate divisions OP8 and OP11 (data not shown). These mismatches may explain the lower recovery for these taxa using SARST-V6 than was observed among full-length sequence data.

Second, differential degradation of DNA from different organisms in an environmental sample (Blum *et al*., 1997) could also account for differences between the SARST-V6 and full-length sequence profiles. Because of a reduced target size, there will be fewer single-strand breaks within the initial templates defined by the SARST-V6 primers relative to the much larger full-length 16S rRNA genes. Therefore there is a higher probability of the *Taq* polymerase progressing through the entire SARST-V6 template. For instance, SARST-V6 may have amplified degraded DNA from dead *Beggiatoa* cells, whereas single-stand breakages might frequently interfere with full-length gene amplification of *Beggiatoa* DNA. To evaluate *active* communities, it may be possible to employ a SARST-V6 strategy and reverse transcription-PCR of extracted RNA (Freitag and Prosser, 2003) in a manner analogous to SAGE studies of mRNA populations.

Others have outlined potential pitfalls of PCR-based microbial community analyses, including selective nucleic acid recovery during extraction, varying rDNA copy number, formation of chimeric amplicons, uneven distributions of single-strand breaks in starting template populations, PCR contamination and gene amplification from dead or dormant species (Amann *et al*., 1995; Suzuki and Giovannoni, 1996; von Wintzingerode *et al*., 1997). Because of these artifacts, relative frequencies of PCR products for full-length rRNA genes do not always reflect organismal abundance in the natural environment. Many of these limitations are common to all PCR-based methods, but with SARST-V6 the small target size increases the probability of generating complete amplicons, reduces effects of differential amplification, and all but precludes the likelihood of forming chimeric molecules. At the same time, SARST-V6 offers a high-throughput means to sample rRNA diversity and representation of discrete coding regions in an environmental sample.

The SARST-V6 method also offers the potential for discovering novel microbial lineages. Most PCR-based studies of nearly full length 16S rRNAs rely on a relatively small number of sequenced genes for primer design (Eden *et al.*, 1991; Weisburg *et al.*, 1991). In contrast, the SARST-V6 primers target conserved regions of the 16S rRNA gene that have been sequenced from more than 30 000 organisms (http://rdp.cme.msu.edu). As SARST-V6 depends upon the use of primer sites that have been well characterized in thousands of rRNA genes, it is unlikely that novel sequences will be missed in these analyses. Furthermore, the high throughput possible with SARST-V6 also allows for more complete sampling of microbial communities, facilitating the detection of rare gene tags. Thus, SARST-V6 permits the detection of novel phylotypes that might be missed with less general primers or less thorough sampling.

Ribosomal sequence tags that do not show strong similarity to known rRNA sequences are candidates for novel phylotypes. Approximately one third of RSTs failed to return BLAST database matches that satisfied our search criteria. However, the majority of these taxonomically unassigned sequences still obtained best (lowest *e*-value) matches to rRNA sequences, indicating that they do not result from non-specific PCR amplification. Many of these unidentified phylotypes are likely to represent novel bacterial lineages warranting further study. By pairing primers that are complementary to novel RSTs with primers targeting conserved regions of the 16S rRNA gene, it becomes possible to amplify a larger, phylogenetically more informative amplicon from rRNAs of candidate novel phylotypes. Neufeld and co-workers successfully employed this approach to amplify, clone and sequence longer 16S rDNA sequences based on RST data (Neufeld *et al.*, 2004).

Although our current results were obtained from a single hypervariable region of the rRNA gene, type IIS restriction endonucleases permit the digestion of any primer sequence based on the inclusion of an appropriate 5′ primer recognition sequence (Szybalski *et al.*, 1991). The amplified target may therefore be adjusted according to the objectives of a particular study. While hypervariable gene segments provide fine-grain resolution of sequence types, more conserved regions will provide a better broad taxonomic overview and help to assign sequences to higher taxa in cases where hypervariable regions lack sufficient signal. Research on microbial community function may target other genes besides rRNA, such as genes involved in sulfur or nitrogen metabolism (e.g. Scala and Kerkhof, 1999; Perez-Jimenez *et al.*, 2001). The SARST-V6 technique may be readily adapted to these studies through the use of appropriate primers.

Experimental procedures

Study site and sample collection

Sediment core A is from the hydrothermally active sediments of Guaymas Basin (Gulf of California). The submersible Alvin (Woods Hole Oceanographic Institution) provided access to the sample core and concurrent thermoprobe temperature profiles. Core A (Alvin dive 3203, 26 April 1998) is from the Everest Mound area (27°1′388″N, 111°24′112″W). The temperature of the uppermost 5 cm ranged from 2°C to 74°C. After the core reached the surface (usually within 12 h of initial collection), we removed a thick (several centimetre) mat of *Beggiatoa* bacteria that overlay the sediment surface. Under anaerobic conditions, we sectioned the core into 0–1 and 1–2 cm layers, designated samples A1 and A2 respectively. We used bead beating and hot phenol procedures (Teske *et al.*, 2002) to extract nucleic acids from samples stored at –80°C.

PCR amplification

The 5′ biotin-TEG labelled, reverse-phase high-performance liquid chromatography (HPLC)-purified oligonucleotide primers SARST967F (5′-3′: TTATTTTAGTGCAGTTAATA CAACGCGAAGAACCTTACC) and SARST1046R (5′-3′: TTATTTTAAGTGCAGCAGCCATGCAVCACCT) (Operon Technologies, Alameda, CA) primed DNA synthesis in PCR amplifications of the V6 rRNA hypervariable regions (Wuyts *et al.*, 2002) in purified DNA of Guaymas Basin sample A1. Polymerase chain reactions in a 50-µl volume contained 1× PCR buffer [20 mM Tris (pH 8.55), 10 mM $(NH4)_2SO_4$, 2 mM $MgCl_2$, 30 mM KCl, 0.01% (w/v) gelatin, 0.05% (v/v) NP-40], 200 µM each dNTPs, 500 nM primer SARST967F, 500 nM primer SARST1046R, 2 units Taq DNA polymerase (Promega, Madison, WI) and 2 µl purified environmental DNA template. Amplification consisted of a single denaturation step at 94°C for 5 min, 30 amplification steps at 94°C for 20 s, 51°C for 30 s and 72°C for 30 s, and a final extension step at 72°C for 10 min. We purified PCR products using a QIAquick PCR Purification Kit (QIAGEN, Valencia, CA) according to the manufacturer's instructions.

Purification and restriction endonuclease digestion of PCR products

The binding of biotinylated PCR products to streptavidin-coated M-280 Dynabeads (Dynal, Oslo, Norway) removed trace amounts of starting template. The binding reactions contained 3–10 μg of QIAquick-purified PCR product and washed beads (600 μl of M-280 Dynabeads washed twice in 2 vol of B&W buffer [10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl]) in a 1.2-ml volume of 10 mM Tris-HCl (pH 7.5) and 1 mM EDTA. After incubation for 15 min at room temperature with constant mixing, a Dynal magnetic particle concentrator (MPC) (Dynal ASA, Oslo, Norway) separated beads with bound DNA from B&W buffer. Following three additional washes with B&W buffer, the beads with bound DNA were washed twice with 1× NEB4 buffer [50 mM KOAc, 20 mM Tris-OAc, 10 mM MgOAc, 1 mM DTT (pH 7.9)]. The restriction endonuclease reaction contained 80 μM S-adenosylmethionine, 1× NEB4 buffer, and 30 units of *Bsg*I (NEB, Beverly, MA) in a volume of 600 μl. Biotinylated DNA bound to Dynabeads was digested by overnight incubation at 37°C with end-over-end mixing. After incubation, the addition of 2 μl of 0.5 M EDTA (pH 8) stabilized single-stranded DNA overhangs. Treating the *Bsg*I digestion mixture for 2 min with a Dynal MPC removed biotinylated DNA fragments attached to the beads and allowed recovery of *Bsg*I restriction fragments in the supernatant. An additional Dynabead binding step using 100 μl of beads ensured the removal of residual undigested biotinylated DNA. The digested DNA was further purified using a QIAquick Gel Extraction Kit (QIAGEN, Valencia, CA) according to the manufacturer's directions using 43 μl of ddH$_2$O and 7 μl of EB buffer in the elution step and then concentrated to 7 μl in a Savant SVC 100H speedvac.

Ligation, cloning and sequencing

Incubation of digested SARST-V6 fragments at 16°C for 5 h in a 20-μl ligation reaction containing 1× T4 DNA ligase buffer [50 mM Tris-HCl (pH 7.6), 5 mM MgCl$_2$, 1 mM ATP, 1 mM DTT, 25% (w/v) polyethylene glycol-8000] and 5 units of T4 DNA ligase (Invitrogen, Carlsbad, CA) produced concatemers with multiple PCR amplicons of rRNA hypervariable regions. Heating at 70°C for 10 min inactivated the ligase. A 30-min incubation with 4.5 units of T4 DNA polymerase (NEB, Beverly, MA) in a 30-μl reaction volume of 1× NEB4 buffer, 100 μg ml^{-1} BSA, 100 μM each dNTPs removed 3′ overhangs from the concatemers. We isolated 1–2 kbp concatemers by agarose gel electrophoresis and subsequent purification using the QIAquick Gel Extraction Kit (QIAGEN, Valencia, CA).

To generate adenine overhangs for TA cloning, purified concatemers were A-tailed at 72°C for 10 min in a 10-μl reaction mixture containing 1× PCR buffer, 225 μM dATP and 1 unit of Taq polymerase. To eliminate 5′ phosphates, which interfere with terminal 3′ phosphates on the TOPO cloning vector, we supplemented the mixture with 5 units of calf intestinal alkaline phosphatase (NEB, Beverly, MA) and incubated at 37°C for 1 h. We then purified the DNA using the QIAquick PCR Purification Kit according to the manufacturer's directions, substituting 4 μl of EB buffer plus 26 μl of dH$_2$O in the final elution step. The product was concentrated as above until the volume of the solution had been reduced to 4 μl. We cloned the A-tailed DNA using the TOPO-TA or TOPO-XL Cloning Kit (Invitrogen, Carlsbad, CA) according to the manufacturer's directions.

DNA sequencing was performed on an ABI Prism 3700 capillary DNA sequencer using the BigDye cycle sequencing kit (Applied Biosystems, Foster City, CA) and M13 universal primer (–20) (5′-3′ GTAAAACGACGGCCAGT) (Operon Technologies, Alameda, CA). The computer programme Cross_match (Green, 1996) removed vector sequence from concatemer reads. A custom Perl script parsed the concatemers into separate hypervariable rRNA sequences based upon the identification of the residual primer sequence (5′-3′) AGAACCTTACC.

Single-amplicon cloning

In addition to sequencing concatemers of ligated amplicons, we cloned and sequenced individual, cloned PCR products amplified using SARST-V6 primers. We used primers 967F (5′-3′: CAACGCGAAGAACCTTACC) and 1046R (5′-3′: CAGCCATGCAVCACCT) to amplify partial 16S rRNA genes from sample A1 as described above. Polymerase chain reaction products were purified using the Qiaquick PCR Purification Kit (QIAGEN, Valencia, CA) and then cloned the DNA using the TOPO-TA Cloning Kit (Invitrogen, Carlsbad, CA). DNA sequencing and vector sequence removal were performed as described above.

DNA sequence analysis

The BLAST (Altschul *et al*., 1997) and SEALS (Walker and Koonin, 1997) software packages identified the taxonomic affiliation of sequences obtained using both SARST-V6 and single-amplicon cloning. We only retrieved sequences from GenBank that displayed BLAST *e*-values below 10^{-6} for more than 75% of the query sequence length. The tax_break programme from the SEALS package identified the taxonomic assignments of GenBank entries. The final taxonomic assignments were based on the best (lowest *e*-value) GenBank match for which tax_break returned a class or genus designation. If all GenBank sequences for a given query lacked these taxonomic identifiers, we examined published references to sequences of lowest BLAST *e*-value to determine taxonomic assignments.

To compare SARST-V6 with cloning and sequencing of individual, full-length PCR products, we performed a corresponding analysis of full-length 16S rRNA gene sequence data from sediment samples A1 and A2 (Teske *et al*., 2002). We extracted sequence positions 1019–1078 (*E. coli* numbering), corresponding to the SARST-V6 PCR amplicon, from the original full-length dataset. We then analysed these partial sequences as described above for SARST-V6, except that, for any given clone, a BLAST match to the previously published sequence of that same clone (Teske *et al*., 2002) was discarded.

Comparison of phylotype distributions

We grouped phylotypes into major bacterial groups in order to generate community profiles. The χ^2 similarity statistic (Legendre and Legendre, 1998) assessed the relationship between phylotype abundances obtained by the various analyses.

Acknowledgements

We thank Andreas Teske and Ginny Edgcomb for providing details about their previous work and for thoughtful input on the discussion. David Mark-Welch provided helpful comments about the manuscript. This research was supported by the NASA Astrobiology Institute (NCC-1054) to M.L.S., the Centro de Astrobiologia (Spain) to C.P. and the W. M. Keck and the Unger G. Vetlesen Foundations to the Josephine Bay Paul Center at the Marine Biological Laboratory.

References

Ahmad, A., Barry, J.P., and Nelson, D.C. (1999) Phylogenetic affinity of a wide, vacuolate, nitrate-accumulating Beggiatoa sp. from Monterey Canyon, California, with Thioploca spp. *Appl Environ Microbiol* **65:** 270–277.

Alfreider, A., Vogt, C., and Babel, W. (2002) Microbial diversity in an in situ reactor system treating monochlorobenzene contaminated groundwater as revealed by 16S ribosomal DNA analysis. *Syst Appl Microbiol* **25:** 232–240.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143–169.

Bending, G.D., Lincoln, S.D., Sorensen, S.R., Morgan, J.A., Aamand, J., and Walker, A. (2003) In-field spatial variability in the degradation of the phenyl-urea herbicide isoproturon is the result of interactions between degradative sphingomonas spp. & soil pH. *Appl Environ Microbiol* **69:** 827–834.

Blum, S.A.E., Lorenz, M.G., and Wackernagel, W. (1997) Mechanism of retarded DNA degradation and prokaryotic origin of DNases in nonsterile soils. *Syst Appl Microbiol* **20:** 513–521.

Cifuentes, A., Anton, J., Benlloch, S., Donnelly, A., Herbert, R.A., and Rodriguez-Valera, F. (2000) Prokaryotic diversity in Zostera noltii-colonized marine sediments. *Appl Environ Microbiol* **66:** 1715–1719.

Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., *et al.* (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res* **31:** 442–443.

Dawson, S.C., and Pace, N.R. (2002) Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc Natl Acad Sci USA* **99:** 8324–8329.

Derakshani, M., Lukow, T., and Liesack, W. (2001) Novel bacterial lineages at the (sub) division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. *Appl Environ Microbiol* **67:** 623–631.

Dojka, M.A., Hugenholtz, P., Haack, S.K., and Pace, N.R. (1998) Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl Environ Microbiol* **64:** 3869–3877.

Dojka, M.A., Harris, J.K., and Pace, N.R. (2000) Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. *Appl Environ Microbiol* **66:** 1617–1621.

Eden, P.A., Schmidt, T.M., Blakemore, R.P., and Pace, N.R. (1991) Phylogenetic analysis of Aquaspirillum magnetotacticum using polymerase chain reaction-amplified 16S rRNA-specific DNA. *Int J Syst Bacteriol* **41:** 324–325.

Edgcomb, V.P., Kysela, D.T., Teske, A., de Vera Gomez, A., and Sogin, M.L. (2002) Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proc Natl Acad Sci USA* **99:** 7658–7662.

Freitag, T.E., and Prosser, J.I. (2003) Community structure of ammonia-oxidizing bacteria within anoxic marine sediments. *Appl Environ Microbiol* **69:** 1359–1371.

Green, P. (1996) *Phrap Sequence Assembly Program.* Seattle, USA: University of Washington.

Hugenholtz, P., Pitulle, C., Hershberger, K.L., and Pace, N.R. (1998) Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* **180:** 366–376.

Jeanthon, C., L'Haridon, S., Cueff, V., Banta, A., Reysenbach, A.L., and Prieur, D. (2002) *Thermodesulfobacterium hydrogeniphilum* sp nov., a thermophilic, chemolithoautotrophic, sulfate-reducing bacterium isolated from a deep-sea hydrothermal vent at Guaymas Basin, and emendation of the genus *Thermodesulfobacterium*. *Int J Syst Evol Microbiol* **52:** 765–772.

LaPara, T.M., Nakatsu, C.H., Pantea, L., and Alleman, J.E. (2000) Phylogenetic analysis of bacterial communities in mesophilic and thermophilic bioreactors treating pharmaceutical wastewater. *Appl Environ Microbiol* **66:** 3951–3959.

Legendre, P., and Legendre, L. (1998) *Numerical Ecology*. Amsterdam, the Netherlands; New York, USA: Elsevier.

Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alio, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603–607.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68:** 5064–5081.

McBain, A.J., Bartolo, R.G., Catrenich, C.E., Charbonneau, D., Ledder, R.G., Rickard, A.H., *et al.* (2003) Microbial characterization of biofilms in domestic drains and the establishment of stable biofilm microcosms. *Appl Environ Microbiol* **69:** 177–185.

Madrid, V.M., Taylor, G.T., Scranton, M.I., and Chistoserdov, A.Y. (2001) Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. *Appl Environ Microbiol* **67:** 1663–1674.

Moeseneder, M.M., Arrieta, J.M., Muyzer, G., Winter, C., and Herndl, G.J. (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. *Appl Environ Microbiol* **65:** 3518–3525.

Moon-van der Staay, S.Y., De Wachter, R., and Vaulot, D. (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409:** 607–610.

Muyzer, G., de Waal, E.C., and Uitterlinden, A.G. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59:** 695–700.

Nagashima, K., Hisada, T., Sato, M., and Mochizuki, J. (2003) Application of new primer-enzyme combinations to terminal restriction fragment length polymorphism profiling of bacterial populations in human feces. *Appl Environ Microbiol* **69:** 1251–1262.

Nelson, D.C., Wirsen, C.O., and Jannasch, H.W. (1989) Characterization of large, autotrophic *Beggiatoa* spp. abundant at hydrothermal vents of the Guaymas Basin. *Appl Environ Microbiol* **55:** 2909–2917.

Neufeld, J.D., Yu, Z., Lam, W., and Mohn, W.W. (2004) Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities. *Environ Microbiol* **6:** 131–144.

Perez-Jimenez, J.R., Young, L.Y., and Kerkhof, L.J. (2001) Molecular characterization of sulfate-reducing bacteria in anaerobic hydrocarbon-degrading consortia and pure cultures using the dissimilatory sulfite reductase (dsrAB) genes. *FEMS Microbiol Ecol* **35:** 145–150.

Reed, D.W., Fujita, Y., Delwiche, M.E., Blackwelder, D.B., Sheridan, P.P., Uchida, T., and Colwell, F.S. (2002) Microbial communities from methane hydrate-bearing deep marine sediments in a forearc basin. *Appl Environ Microbiol* **68:** 3759–3770.

Scala, D.J., and Kerkhof, L.J. (1999) Diversity of nitrous oxide reductase (nosZ) genes in continental shelf sediments. *Appl Environ Microbiol* **65:** 1681–1687.

Suzuki, M.T., and Giovannoni, S.J. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625–630.

Szybalski, W., Kim, S.C., Hasan, N., and Podhajska, A.J. (1991) Class-IIS restriction enzymes – a review. *Gene* **100:** 13–26.

Tanner, M.A., Everett, C.L., Coleman, W.J., Yang, M.M., Youvan, D.C. (2000) Complex microbial communities inhabiting sulfide-rich black mud from marine coastal environments. *Biotechnology et alia* **8:** 1–16.

Teske, A. (1999) Phylogenetic position of a large marine Beggiatoa (Vol. 22, p. 39, 1999). *Syst Appl Microbiol* **22:** 492–492.

Teske, A., Ramsing, N.B., Kuver, J., and Fossing, H. (1996) Phylogeny of Thioploca and related filamentous sulfide-oxidizing bacteria. *Syst Appl Microbiol* **18:** 517–526.

Teske, A., Hinrichs, K.U., Edgcomb, V., de Vera Gomez, A., Kysela, D., Sylva, S.P., *et al.* (2002) Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Appl Environ Microbiol* **68:** 1994–2007.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science* **270:** 484–487.

Walker, D.R., and Koonin, E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *Proc Int Conf Intell Syst Mol Biol* **5:** 333–339.

Weisburg, W.G., Barns, S.M., Pelletier, D.A., and Lane, D.J. (1991) 16S ribosomal DNA amplification for phylogenetic study. *J Bacteriol* **173:** 697–703.

von Wintzingerode, F., Göbel, U.B., and Stackebrandt, E. (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21:** 213–229.

Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. (2002) The European database on small subunit ribosomal RNA. *Nucleic Acids Res* **30:** 183–185.