Environmental Microbiology (2004) **6**(2), 131–144 doi:10.1046/j.1462-2920.2003.00547.x

Serial analysis of ribosomal sequence tags (SARST): a high-throughput method for profiling complex microbial communities

Josh D. Neufeld,[1] Zhongtang Yu,[2] Wan Lam[3] and William W. Mohn[1]*

[1]*Department of Microbiology and Immunology, University of British Columbia, 6174 University Boulevard, Vancouver, British Columbia, V6T 1Z3, Canada.*
[2]*Department of Animal Sciences (Microbiology), The Ohio State University, 2027 Coffey Road, Columbus, Ohio, 43210, USA.*
[3]*British Columbia Cancer Research Center, 601 West 10th Avenue, Vancouver, British Columbia, V5Z 1L3, Canada.*

Summary

Two decades of culture-independent studies have confirmed that microbial communities represent the most complex and concentrated pool of phylogenetic diversity on the planet. There remains a need for innovative molecular tools that can further our knowledge of microbial diversity and its functional implications. We present the method and application of serial analysis of ribosomal sequence tags (SARST) as a novel tool for elucidating complex microbial communities, such as those found in soils and sediments. Serial analysis of ribosomal sequence tags uses a series of enzymatic reactions to amplify and ligate ribosomal sequence tags (RSTs) from bacterial small subunit rRNA gene (SSU rDNA) V1-regions into concatemers that are cloned and sequenced. This approach offers a significant increase in throughput over traditional SSU rDNA clone libraries, as up to 20 RSTs are obtained from each sequencing reaction. To test SARST and measure the bias associated with this approach, RST libraries were prepared from a defined mixture of pure cultures and from duplicate arctic soil DNA samples. The actual RST distribution reflected the theoretical composition of the original defined mixture. Data from duplicate soil libraries (1345 and 1217 RSTs, with 525 and 505 unique RSTs, respectively) indicated that replication provides a strongly correlated RST profile (r^2 = 0.80) and division-level distribution of RSTs (r^2 = 0.99). Using sequence data from abundant soil RSTs, we designed specific primers that successfully amplified a larger portion of the SSU rDNA for further phylogenetic analysis. These results suggest that SARST is a powerful approach for reproducible high-throughput profiling of microbial diversity amenable to medical, industrial or environmental microbiology applications.

Received 11 July, 2003; revised 24 September, 2003; accepted 2 October, 2003. *For correspondence. E-mail wmohn@interchange.ubc.ca; Tel. (+1) 604 822 4285; Fax (+1) 604 822 6041.

Introduction

Complex microbial communities such as those found in soil, sediment and activated sludge are commonly characterized using a diverse set of culture-independent tools. Of those that rely on SSU rDNA sequence heterogeneity, methods usually involve either DNA fingerprinting or sequencing of individual PCR-amplified fragments. Although useful for quick comparisons of multiple communities, the drawbacks to fingerprint-based methods include a lack of resolution provided by gel-based separation and also difficulty in assigning phylogenetic information to the complex banding patterns that are usually obtained. Whereas SSU rDNA clone libraries provide useful phylogenetic information that is reflective of community composition and relative distributions of organisms, small sample sizes prevent adequate representation of microbial community phylotypes because of cost and labour limitations. Some of the largest clone libraries generated have been limited to the sequencing of several hundred clones per library (Kroes *et al*., 1999; McCaig *et al*., 1999; Suau *et al*., 1999; Nogales *et al*., 2001; Axelrood *et al*., 2002; Chow *et al*., 2002). Larger sample sizes, perhaps by an order of magnitude, are required for exploring and comparing the high diversity common to most microbial communities and for adequate microbial community comparisons (Tiedje *et al*., 1999).

Because the number of clones that can be sequenced in a clone library is limited by cost and sequencing capacity, a number of methods have attempted to increase the throughput of clone library characterization. Amplified ribosomal DNA restriction analysis (ARDRA) of cloned SSU rDNA inserts (Pace *et al*., 1986) and a variety of hybridization-based screening methods (Snaidr *et al*.,

1997; Ravenschlag *et al*., 1999; Valinsky *et al*., 2002) have reduced the number of clones that require sequencing. However, the number of organisms screened is limited to the number of clones that can be processed.

Toward solving an analogous challenge, the advent of serial analysis of gene expression (SAGE) greatly facilitated the comparison of mRNA abundances in eukaryotic cells by generating large libraries of sequence tags from individual mRNA transcripts (Velculescu *et al*., 1995). In prokaryotes, a recently developed approach generates large libraries of short genomic sequence tags (GSTs) derived from bacterial genomic DNA (Dunn *et al*., 2002). Here we present serial analysis of ribosomal sequence tags (SARST) for high-throughput determination of short ribosomal sequence tags (RSTs) from the genomic DNA of any microbial community. Serial analysis of ribosomal sequence tags is similar in concept to an approach proposed by Borneman *et al*. (1997) but differs in the methodology, which is based on SAGE. Serial analysis of ribosomal sequence tags uses PCR to specifically amplify a short and hypervariable region (V1 region; 65–103 *Escherichia coli* numbering) of the SSU rDNA from extracted genomic DNA. Through sequential enzymatic manipulations, the RSTs are isolated with compatible ends and specific border sequences that allow ligation of individual RSTs into concatemers that are cloned and sequenced. The power of this novel method is a significant increase in throughput, as RSTs from many organisms can be read serially in each sequencing reaction. These large RST libraries provide high-resolution profiles that reflect the sampled community composition.

In addition to presenting the method, we quantify some of the potential bias associated with SARST. Whereas bias is inherent to any method used to measure the abundance of microorganisms in communities, molecular methods have a unique set of challenges that include DNA extraction, PCR copy fidelity, PCR artifacts and sequencing error (Wintzingerode *et al*., 1997). To assess the extent of some of these biases using SARST, we generated large RST libraries from a defined mixture of pure cultures with known RST sequences and operon copy numbers. Along with tests for copy fidelity and DNA extraction bias, we tested two polymerase mixtures with different reported fidelities to measure the frequency of misincorporated nucleotides during PCR. Also, by preparing large duplicate RST libraries from arctic soil genomic DNA, we demonstrate the reproducibility of SARST-generated data. The RST sequences obtained from the arctic soil genomic DNA provided enough specificity for primer design to amplify a larger portion of the corresponding SSU rDNA for further phylogenetic analysis. We demonstrate that SARST facilitates high-throughput SSU rDNA screening of complex microbial communities.

Results

Overview of SARST

Figure 1 illustrates the SARST procedure for generating RSTs. The PCR primers we designed have 3′ regions

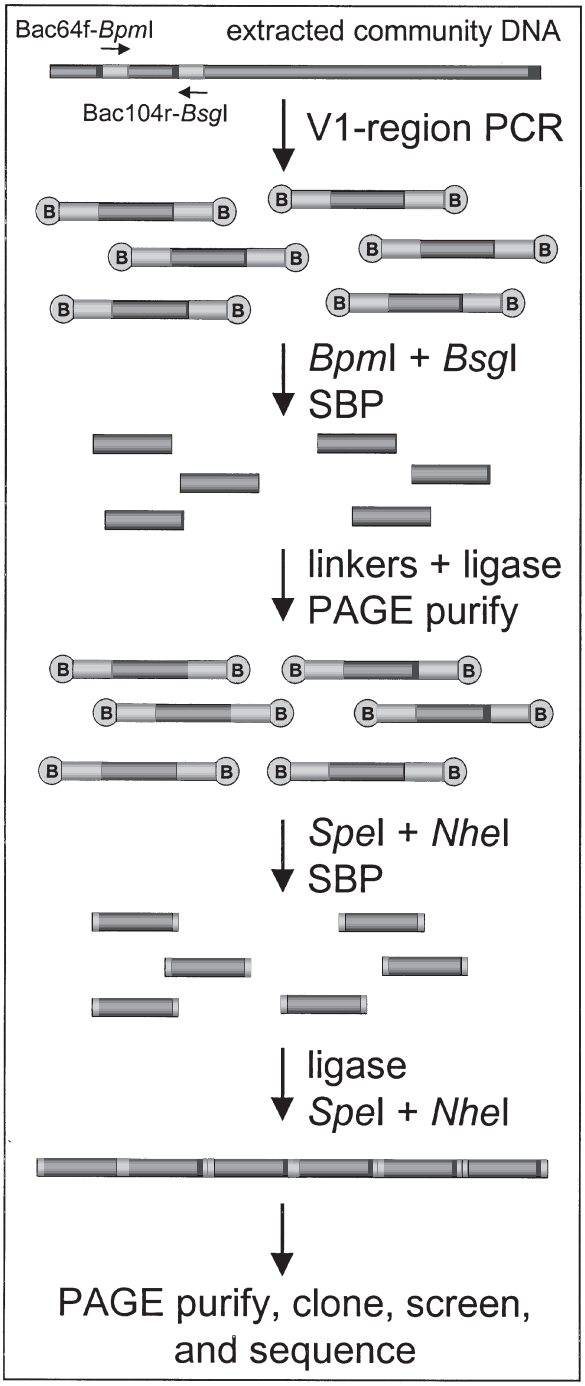

Fig. 1. Flow diagram of serial analysis of ribosomal sequence tags (SARST). Polyacrylamide gel electrophoresis (PAGE) purifications and streptavidin bead purifications (SBP) using biotin label (B) isolate desired DNA intermediates for subsequent steps.

almost identical to primers designed by Bertilsson *et al.* (2002), except Bac104r differs from theirs in having less degeneracy at one position. Whereas the 3′ regions are complimentary to almost all known bacterial SSU rDNA sequences in the RDP-II alignment, our primers have 5′ extensions that contain recognition sites for subsequent removal with type IIS restriction enzymes (*Bpm*I on Bac64f and *Bsg*I on Bac104r). The resulting cut sites from digestion with these enzymes leave specific two-nucleotide overhangs immediately flanking the RSTs. DNA linkers are ligated to each side of the RSTs. Enzymatic digestion of these linkers creates compatible overhangs for RST ligation into concatemers. Concatemers contain predicted border sequences that demarcate the position and polarity of RSTs. Digesting the RST-linker DNA with *Spe*I (cuts within Linker A) and *Nhe*I (cuts within Linker B) frees the RSTs for purification and concatenation.

A critical step in the SARST protocol is the concatemer-forming ligation. *Spe*I and *Nhe*I digest overhangs are eligible for 'like' (*Spe*I-*Spe*I; *Nhe*I-*Nhe*I) or 'unlike' (*Spe*I-*Nhe*I) end ligation. Like ligations have proven problematic in previous methods for SARST, since the RSTs on either side of a like border are similar reverse complements. These complementary RSTs can form hairpins that frustrate single-strand reactions such as PCR screening and sequencing. This problem would be especially pronounced when preparing libraries from relatively simple communities such as defined mixtures of pure cultures, in which similar or identical RSTs more frequently ligate adjacently. The current protocol concatenates RSTs with T4 DNA ligase in the presence of both *Spe*I and *Nhe*I, which recognize and cut borders formed between like ligations. As a result, the concatemer-forming reaction is a series of ligations and digestions occurring simultaneously. This particular step has greatly decreased unsuccessful sequencing reactions and increased the number of RSTs obtained per sequencing reaction.

RST specificity

Because the RST region surveyed by SARST is short (17–55 bp), the specificity of these regions is critical to understanding the phylogenetic resolution this approach offers. Towards testing RST region specificity, we analysed RSTs from a dataset of 3850 type-strain sequences from the most recent release of the RDP-II (April 2003). Each of the RST regions was removed from the dataset and then screened against the entire dataset to assess its specificity. Approximately 78% of the RSTs were specific to the genus level, which was the highest phylogenetic level that could be reliably tested using this Sequence Match approach (B. Chai and J. Cole, pers. comm.). We also obtained sequence data from a previously published forest soil SSU rDNA clone library ($n = 709$) that spans the V1-V3 variable regions (Chow *et al.*, 2002). We aligned the sequences (~450 base pair) using POA (Lee *et al.*, 2002) and removed sequences from the analysis that aligned poorly. In order to organize the SSU rDNA clones and RSTs into like groups, we used a program called Fastgroup (Seguritan and Rohwer, 2001). Fastgroup clusters sequences based on a user-defined similarity value and was developed with the intention of managing SSU rDNA clone-library sequence data. Fastgroup clustered the set of sequences with 95% grouping, as this is the lowest threshold typically used to demarcate operational taxonomic units (OTUs) with SSU rDNA sequences (Hughes *et al.*, 2001). We then grouped the RSTs derived from this same dataset using identical criteria. Out of 702 sequences, the longer SSU rDNA sequences formed 413 unique groups whereas the RSTs formed 312 unique groups. Even with 100% grouping, the RSTs formed 360 unique groups, which is substantially less than the number of groups formed with the longer sequences. Thus, RSTs provide somewhat lower resolution than longer rDNA fragments.

Defined community analysis

In order to test the ability of SARST data to reflect the actual composition of a community, we constructed and analysed a simple, defined community (Table 1). Five organisms were chosen because they represent phylogenetic groups that differ in their susceptibility to lysis due to their cell wall compositions. Also, the five strains have

Table 1. Organisms selected for a defined community.

Organism	rrn copies	RST code	5′-3′ RST sequence (inter-operon differences bolded)
Bacillus subtilis strain 168 (ATCC 23857)	8	BS1	AGCG**G**ACAG**A**TGGGAGCTTGCTCCCTGATGTTAGC
	1	BS2	AGCG**A**ACAG**A**TGGGAGCTTGCTCCCTGATGTTAGC
	1	BS3	AGCG**G**ACAG**G**TGGGAGCTTGCTCCCTGATGTTAGC
Escherichia coli K12 (ATCC 10798)	2	EC1	AACGGTAACAGGAA**AC**AGCTTGCT**GT**TT**C**GCTGACGAGT
	1	EC2	AACGGTAACAGGAA**GC**AGCTTGCT**GC**TT**C**GCTGACGAGT
	4	EC3	AACGGTAACAGGAA**GA**AGCTTGCT**TC**TT**T**GCTGACGAGT
Staphylococcus aureus NCTC 8325	5	SA	AGCGAACGGACGAGAAGCTTGCTTCTCTGATGTTAGC
Caulobacter crescentus CB15 (ATCC 19089)	2	CC	AACGGATCCTTCGGGATTAGT
Burkholderia sp. LB400	1	LB400	AACGGCAGCACGGGGGCAACCCTGGTGGCGAGT

sequenced genomes, and are known to have a range of ribosomal copy numbers. One of the difficulties in measuring bias is that different sources of bias are difficult to isolate and measure separately. We selected *Escherichia coli* K12 and *Bacillus subtilis* strain 168 as these organisms have multiple *rrn* operons with RST sequences differing between some of the individual operons (Table 1). As a result, the ratio of the different *rrn* operons for these organisms reflects PCR copy fidelity independent of other sources of bias. The ratio of the RSTs between the different organisms indicates the extent to which other bias (DNA extraction in particular) might have affected the results. To simulate the treatment of soil samples for SARST, a soil FastDNA SPIN kit (Qbiogene, Carlsbad, CA) in conjunction with a beadbeater was used to extract DNA from an equal-ratio mixture of the five strains. We used SARST to generate RST libraries using *Taq* polymerase (Taq) and a high fidelity enzyme mixture for PCR (HiFi). Testing these two enzyme mixtures enabled PCR-generated errors and artifacts to be measured (see below). Figure 2 shows the predicted and actual distribution of RSTs in the SARST libraries. The Taq ($n = 341$ RSTs) and HiFi ($n = 276$ RSTs) library distributions were positively correlated with the expected distribution, assuming an equal ratio of cells from each species ($r^2 = 0.62$ and 0.52, respectively). The distribution of RSTs in the Taq and HiFi libraries were also strongly correlated ($r^2 = 0.96$), which indicates that the data generated by SARST are reproducible. The multiple-operon ratios between the RST types from *E. coli* and *B. subtilis* were close to the expected ratios, which suggests that the PCR bias was not significant given the conditions and templates chosen here. Whereas the overall RST ratios between the different organisms vary somewhat from the expected ratios, they do reflect the overall composition and abundance. The *B. subtilis* and *Burkholderia* sp. LB400 RST abundances were notably different from those expected. For *B. subtilis*, this difference is likely explained by a poorer yield of DNA than from the other organisms, due to its Gram-positive cell wall. We tested the DNA extraction method on each of these cultures individually and found the DNA yield was substantially less from *B. subtilis* than from the other pure cultures (data not shown). In addition, counting cells with AODC is somewhat subjective because cells are in various stages of cell division. Thus, some of the discrepancy may result from cell counting, rather than from SARST.

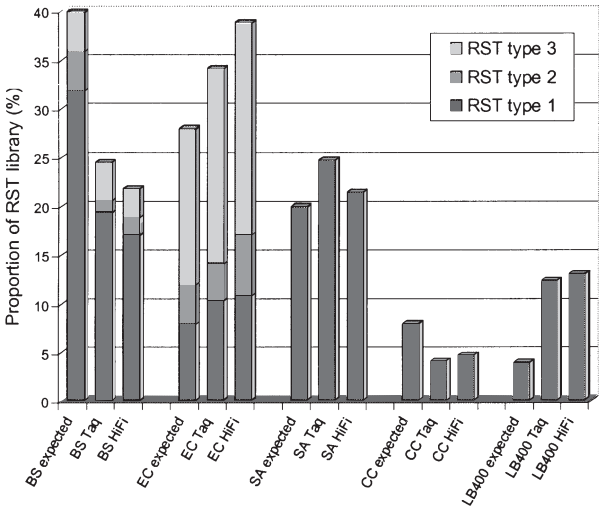

Fig. 2. Expected RST abundance in the defined community compared to the observed distributions in libraries prepared with *Taq* polymerase (Taq; $n = 341$ RSTs) and a high fidelity enzyme mixture (HiFi; $n = 276$ RSTs). Also indicated in different shades are the expected and observed proportions of the different RST types for *Escherichia coli* and *Bacillus subtilis*. See Table 1 for a list of bacterial strains represented by RST codes.

By sequencing a large number of RSTs from libraries prepared with *Taq* polymerase and a high fidelity enzyme mixture we measured the number of PCR-generated errors and artifacts in the form of misincorporated nucleotides and chimeric sequences. We selected Platinum *Taq* DNA polymerase High Fidelity (Invitrogen), as this enzyme mixture provided a comparable amount of reaction product with the defined community DNA as template for PCR. The error rates for the Taq and HiFi enzymes mixture were 1 in 1760 (5.7×10^{-4}) and 1 in 1430 (7.0×10^{-4}) nucleotides respectively. This translates into approximately 1 in 50 and 1 in 40 RSTs with an erroneous base respectively. These results were surprising since the high fidelity enzyme mix is reported to have a sixfold higher fidelity than *Taq* polymerase due to the 3′-5′ exonuclease activity provided by the addition of *Pyrococcus* species GB-D polymerase. The PCR conditions used in this experiment were not identical to the manufacturer's protocol for the high fidelity enzyme mixture. The presence of BSA and higher extension temperature (72°C instead of 68°C) were chosen to provide consistent conditions between the two reaction mixtures but may have negatively affected the proofreading ability of the *Pyrococcus* species GB-D polymerase. While not optimal, 72°C is within the range of acceptable temperatures for this enzyme (Sambrook and Russel, 2001). Regardless, the literature reports a wide range of error rates for *Taq* polymerase (Wintzingerode *et al.*, 1997) and those we observed are consistent with those estimates. Because *Taq* polymerase provides the highest yield of amplified DNA with all templates used in this study, this enzyme is appropriate and sufficient for SARST. In addition to misincorporated nucleotides, PCR reactions generate chimeric sequences formed between amplified sequences from multiple organisms. From a total of 617 RSTs in our defined community libraries, we obtained a single chimeric sequence. Although this chimera was detectable using CHECK_CHIMERA, an online tool

at the RDP-II (Maidak *et al*., 2001), RST sequences are too short to routinely and reliably screen with this method. The presence of chimeras has received attention in the literature (Kopczynski *et al*., 1994; Wang and Wang, 1996; Wang and Wang, 1997; Qiu *et al*., 2001; Speksnijder *et al*., 2001; Hugenholtz and Huber, 2003) with the reported frequency of chimeric molecules in SSU rDNA libraries from model communities ranging between undetectable (Qiu *et al*., 2001) to over 30% (Wang and Wang, 1997). Chimera formation, caused by incomplete extension of the polymerase, decreases with decreasing PCR cycles, with increasing extension times, and with decreasing sequence similarity of the mixed templates (Wang and Wang, 1996; 1997; Qiu *et al*., 2001). Because the PCR template was a defined community and the number of PCR cycles was limited, it is not surprising that chimeric sequences were almost undetectable. There remains the possibility that chimeric RSTs might be more numerous in libraries generated from communities with a gradient of similar V1 regions. However, Borneman *et al*. (1996) reported a low frequency of chimera formation in their soil-derived SSU rDNA library comprised of ~200 bp sequences. The short length of the RST region should minimize the abundance of chimeric molecules in SARST-generated RST libraries.

An artifact we detected in the RST sequences is a missing nucleotide at the 3′ end of some RSTs. This occurred in 1–2% of the RSTs in our Taq and HiFi libraries. Nucleotides cleaved from the 3′ end might be caused by exonuclease contamination during enzymatic reactions or from incorrect cleavage of the DNA by the star activity of restriction endonucleases used in SARST. Nonetheless, these sequences were correctly grouped with their corresponding full-length sequences, with either a 100% or 95% similarity threshold setting with Fastgroup. We also detected a single unexpected RST that likely represents a low-frequency PCR contaminant. The 39 bp RST (5′-AACGGCAGCACAGAGGAGCTTGCTCCTTGGGTGGCGAGT) has perfect matches to several *Xanthomonadaceae (*e.g. *Lysobacter* sp. C3, GenBank accession AY074793) within the γ–*Proteobacteria.* This confirms that extreme caution is required for manipulation of all components of the SARST PCR reaction.

Although Fastgroup successfully clusters RST sequences, the appropriate similarity cutoff for grouping RSTs was uncertain. When full-length or partial SSU rDNA sequences are grouped, commonly used thresholds for OTUs range from 95%-99% (Hughes *et al*., 2001), ideally providing species-level grouping. As RSTs usually provide genus-level specificity, grouping RSTs with 100% sequence similarity would theoretically provide at least genus-level grouping. However, as Fig. 3 illustrates, when a grouping criterion of 100% identity was selected for all 617 RSTs in the defined communities, ungrouped multiple operons from individual organisms and PCR-generated

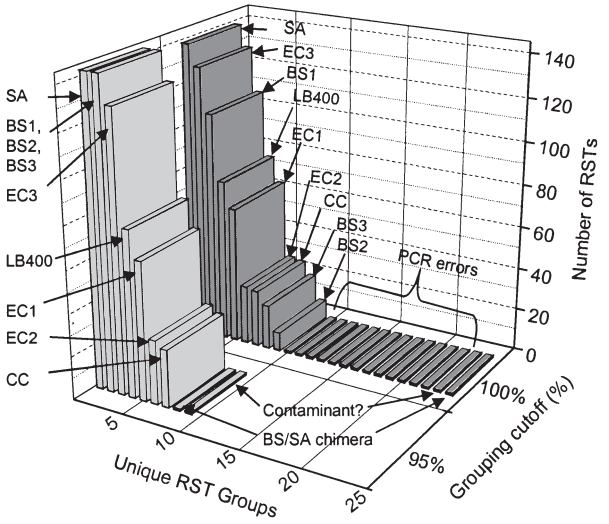

Fig. 3. Relationship between the number of unique RST groups formed using Fastgroup with either a 100% or 95% grouping cutoff on the combined Taq/HiFi RST libraries ($n = 617$ RSTs). See Table 1 for a list of bacterial strains represented by RST codes.

errors led to an overestimation of the total richness of RST groups contained in the community. Thus, we also tested a grouping threshold of 95% identity, which allows for one or two base differences depending on the RST length (Fig. 3). This grouping threshhold correctly grouped RST sequences with PCR errors as well as the multiple operons of *Bacillis subtilis*. Therefore, we chose 95% grouping as a means of organizing RSTs into their respective OTUs. Whereas the grouping threshold can be varied to balance between resolution and compensation for errors and *rrn* operon variability, consistency is necessary for comparing the composition and diversity of multiple communities.

Arctic soil community analysis

In order to test SARST on a complex microbial community, we analysed an arctic soil sample (FOX-B) for SARST. We screened concatemer inserts of transformant colonies in two 96-well plates for duplicate, composite samples. Because the number of suitable inserts was high (>80%), we sequenced PCR products from all clones, regardless of size. From this sequence data, we generated duplicate RST libraries: FOX-B1 and FOX-B2 with 1345 and 1217 RSTs respectively. By preparing libraries from separate DNA extractions we were able to assess how well a DNA sample represents the soil sample from which it was extracted. After grouping all the RSTs in each library with Fastgroup, using a 95% sequence similarity threshold, we determined that the FOX-B1 and FOX-B2 duplicate libraries have broadly similar composition. Figure 4 demonstrates that most of the abundant RST groups (>0.5% of total) are well represented by both libraries. A simple

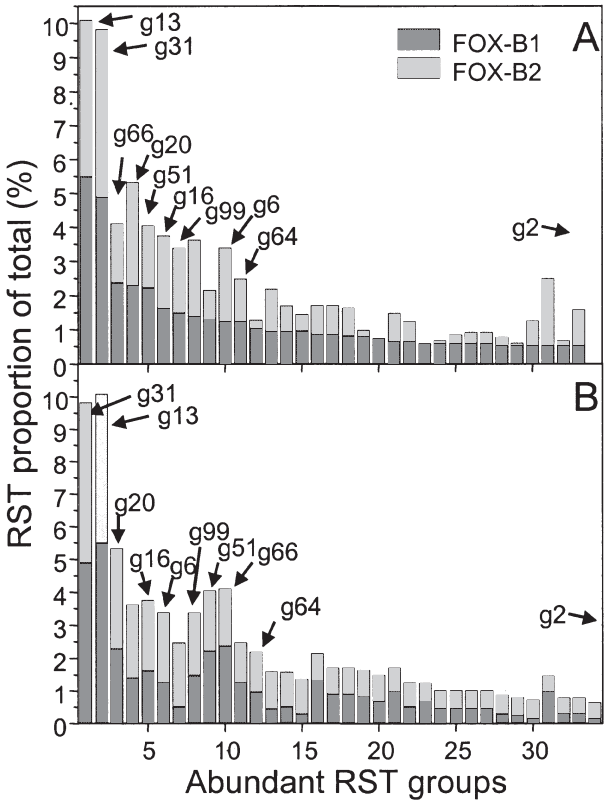

Fig. 4. A comparison of the proportion of FOX-B1 (n = 1345 RSTs) and FOX-B2 (n = 1217 RSTs) RST sequences in their respective libraries sorted by the most abundant RST groups (>0.5% representation in library) in the FOX-B1 (A) and in the FOX-B2 (B) libraries. Also indicated are the groups that were selected for RST primer design.

correlation analysis of all RST groups in the FOX-B1 and FOX-B2 libraries indicates that they are strongly correlated (r^2 = 0.80). The division-level distribution of RSTs in these two libraries (Fig. 5) is also very similar (r^2 = 0.99). Divisions were assigned to individual RSTs based on the reference sequence in the RDP-II version 9.0 with the highest Sequence Match similarity score. The most abundant divisions in this RST library are the *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Planctomycetes*, *Acidobacteria,* and *Verrucomicrobia*. Some RSTs defined as putative chimeras were not chimeric themselves, but are affiliated with full-length database sequences that were recently determined to be of chimeric origin (Hugenholtz and Huber, 2003). A problem with using Sequence Match for phylogenetic affiliation is that some affiliations are incorrect, because the database is lacking coverage and because some database sequences that are potentially affiliated with known groups have not yet been classified. However, the bias with this procedure was similar for both the FOX-B libraries, and the data are still useful for the purpose of comparison. Until more efficient and accurate means exist for determining the phylogenetic affiliations for hundreds or thousands of query sequences simultaneously, some degree of error will always be present.

Using EstimateS to analyse a modified Fastgroup outfile, we calculated the Chao1 richness estimate and Shannon-Weiner diversity index for each of the individual FOX-B datasets (Table 2). The Chao1 richness estimator provides an estimate of the total number of RSTs that are predicted to exist in the sample, based on number of singletons and doubletons in the library (Chao, 1984). The Shannon-Weiner index, when applied to clone library analysis, was recently described as a measure of the difficulty of predicting the next clone drawn at random (Hill *et al.*, 2003). This index is positively correlated with species diversity, which includes both richness and evenness. As expected, the data for the duplicate libraries have similar Chao1 and Shannon-Weiner values. By pooling the RSTs from the two libraries it becomes clear that the FOX-B arctic soil is still inadequately sampled, since the proportion of singletons in the library to the total number of groups is similar. For the combined library, the Chao1 richness estimate and Shannon diversity index were both greater than for the individual libraries. This suggests that more RSTs are required to adequately characterize the diversity of this particular arctic soil microbial community.

RST-based primer design

We selected nine RSTs, reflecting a range of phylogenetic groups, that were abundant in both the FOX-B1 and FOX-B2 libraries and designed specific primers to PCR amplify

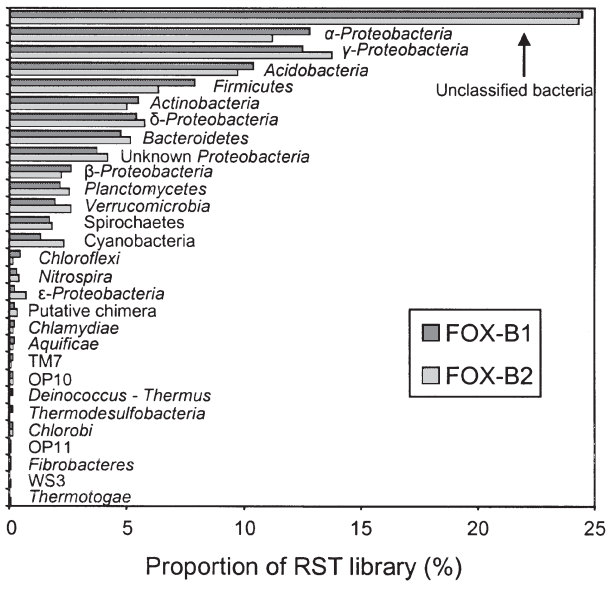

Fig. 5. Division-level distribution of RST sequences in the FOX-B duplicate libraries. The proportions indicate the percent representation of each division in the respective library.

Table 2. Summary of FOX-B arctic soil RST libraries.

Library	Total RSTs	RST groups[a]	Singletons	Chao1 (SD)	Shannon index
FOX-B1	1345	525	342	1269 (121)	5.53
FOX-B2	1217	505	338	1206 (113)	5.50
Combined	2562	806	510	1904 (145)	5.72

a. Unique RST groups based on Fastgroup 95% similarity threshold.

a larger portion of the corresponding 16S rDNA sequences (Table 3). All primers amplified sequences of the expected size using the 907r reverse primer. By cloning and sequencing PCR products from the highest annealing temperature that yielded product, we were able to identify the correct RST 3′ end for eight of the nine RSTs. The shortest and least abundant RST selected for primer design, g2, did not successfully amplify the correct 3′ sequence in the nine inserts screened. However, we did detect 3′ ends that differed by only 1 bp from that expected. The short RST length of g2 may have prevented sufficient primer specificity, and the low abundance (0.16%) of this RST may have increased the likelihood of amplifying similar non-target templates. With the eight successful amplifications, we prepared double-pass sequence data for all inserts with the correct 3′ RST end. We generated a phylogenetic tree to compare the ~450 bp amplicons (Fig. 6). For each RST primer, most of the resulting PCR products cluster closely within the same clade, suggesting that the sequences originate within the same organism (with small differences likely due to multiple operons having sequence variability) or from closely related organisms. One apparent exception is g6-B11, which likely represents an RST shared between two different species within the *Acidobacteria* division. The *Acidobacteria* primer could conceivably amplify RSTs from two different groups in our RST library that contained RSTs differing by a single base close to the 3′ end. Based on their abundance in the RST library, we correctly predicted that the more abundant RST group (g31) would also be the SSU-rDNA sequence that we would preferentially amplify using the RST specific primer. Finally, the g64 RST, which shares identity with rDNA sequences from several different divisions, seems to originate in an organism with no affiliation to known bacterial divisions.

Discussion

We have demonstrated a method that efficiently produces large libraries of V1-region sequence tags from bacterial communities of defined and unknown compositions. This method provides a more efficient way of surveying a larger proportion of microbial communities than previously feasible with traditional rDNA clone library analysis. Bacteria-specific SARST primers can easily be modified to prepare RST libraries from the Archaea or Eukaryotes. High throughput is the strength of this method, as up to 20 OTUs are obtained with the effort required to obtain a single OTU in a conventional rDNA clone library. Although the RST region is the most variable region of the SSU rDNA, it is shorter and has less specificity than the full gene sequence. A benefit of lower specificity is more comprehensive sampling of a community with a given number of sequences. Thus, lower resolution of RSTs, relative to larger SSU rDNA fragments, contributes to the objective of a relatively comprehensive survey of a microbial community. The results of this study show that SARST is functional, reproducible, and that the distribution of RSTs in the defined community library reflects the composition of the original sample.

Although the defined community RST library was similar to the predicted distribution of RSTs, it is clear that the distribution of sequences in RST or traditional rDNA clone libraries do not quantitatively measure the abundance of

Table 3. PCR primers designed on the basis of RST sequences.

Group	RST phylogenetic affiliation (%)[a]	Abundance (%)	Primer and RST sequence[b]
g13	Uncultured γ-*Proteobacteria* (100)	5.07	GTCG**AGCGGTAACAGGTGTAG**CAATACATGCTGACGAGC
g6/g31	Uncultured *Acidobacteria* (100)	1.68/4.96	CAAGTCG**AACGAGAAAGTGGAG**CAATCCATGAGTA(C/A)AGT
g51	α-*Proteobacteria* 100	2.03	TGCAAGTCG**AACGCCGTAGCAATACGGAGT**
g66	α-*Proteobacteria* (100)	2.07	CAAGTCG**AGCGGGCGTAG**CAATACGTCAGC
g20	Uncultured bacteria clone (100)	2.65	TCG**AGCGGTAACGCGGGAG**CAATCCTGGCGACGAGC
g16	γ-*Proteobacteria* (100)	1.88	CG**AACGGCAGCACAGAGG**AGCTTGCTCCTTGGGTGGCGAGT
g2	α-*Proteobacteria* (100), Cyanobacteria (100)	0.16	GCAAGTCG**AACGCACCTTC**GGGTGAGT
g99	*Actinobacteria* (100)	1.68	TCG**AGCGGAAAGGCCCTTC**GGGGTACTCGAGC
g64	*Firmicutes* (100), Uncultured α- (100), γ-*Proteobacteria* (100)	1.25	GCAAGTCG**AACGAGGTAG**CAATACCTAGT

a. Percent similarity to closest BLAST hit in GenBank
b. Primer sequence is underlined and RST sequence is in bold.

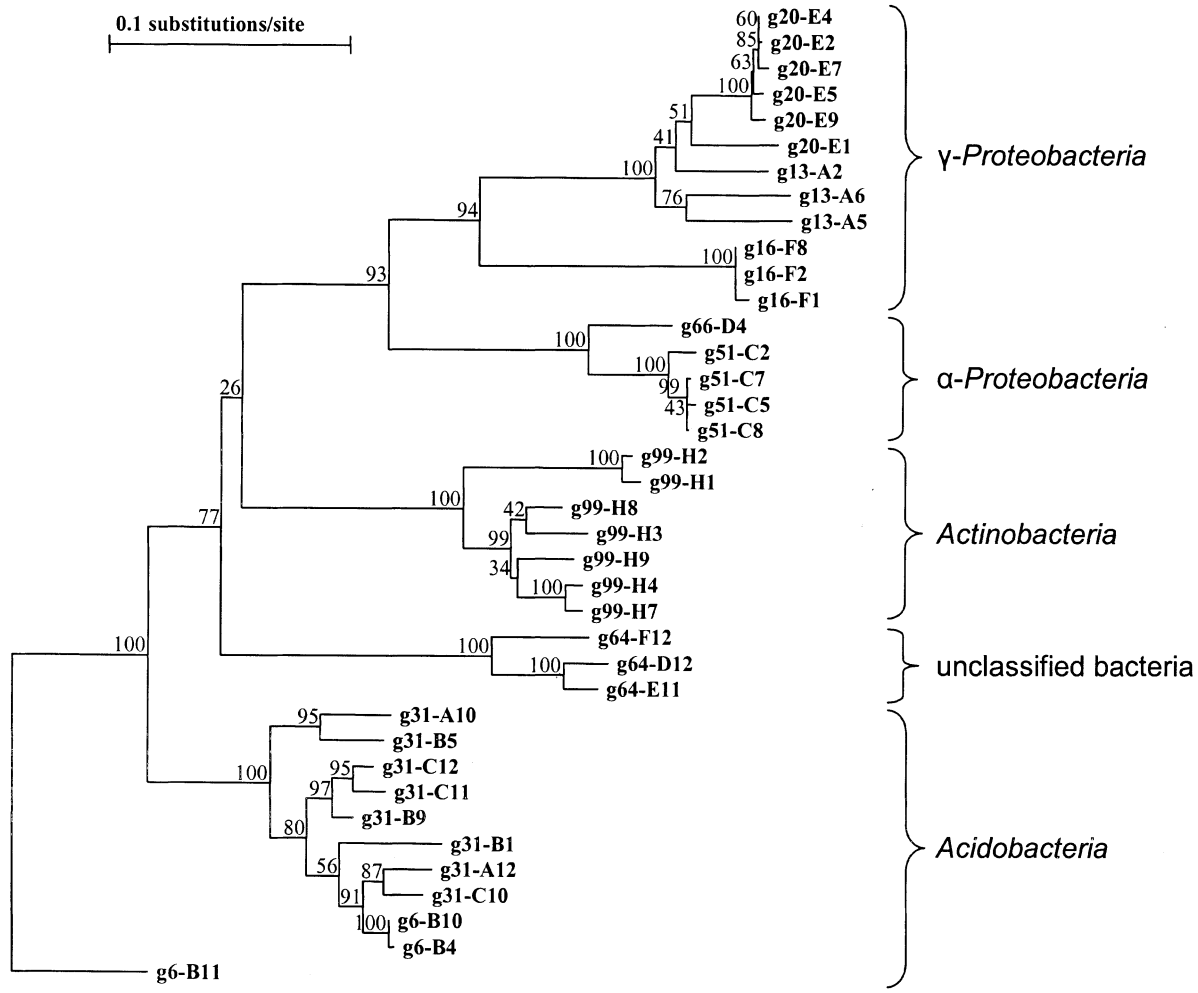

Fig. 6. Phylogenetic tree of the 38 partial SSU rDNA sequences amplified using RST-specific primers. These sequences all originated from the FOX-B arctic soil. The tree was rooted with an *Acidobacteria* sequence (g6-B11) and bootstrap values were calculated from 100 bootstraps.

different species in the sample. Because the number of *rrn* operons varies widely between different organisms (Fogel *et al.*, 1999), the distributions of sampled sequences will be biased by multiple operons (Farrelly *et al.*, 1995; Klappenbach *et al.*, 2000; Crosby and Criddle, 2003). Sequence heterogeneity of operons within the same organism (Clayton *et al.*, 1995) and similarities between SSU rDNA sequences of different species (Fox *et al.*, 1992) further frustrate accurate quantitative inferences. However, clone libraries are generally used for making comparisons between samples or treatments. Although the absolute abundance of SSU rDNA sequences may be inaccurate, the increase or decrease in specific sequence representation between libraries should reflect real population trends.

The FOX-B arctic soil RST library is dominated by bacterial divisions that are well represented in other soil clone library analyses. The relative abundance of phylogenetic groups varies from soil to soil, but the predominance of SSU rDNA sequences belonging to *Proteobacteria*, *Acidobacteria*, *Planctomycetes*, *Verrucomicrobia*, and Gram-positive bacteria (primarily *Actinobacteria*) is commonly reported (Dunbar *et al.*, 1999; McCaig *et al.*, 1999; 2001a, b; Nogales *et al.*, 1999; 2001; Smit *et al.*, 2001; Axelrood *et al.*, 2002; Chow *et al.*, 2002). This is in agreement with the distribution of RSTs in our soil library. Division-level profiling of SSU rDNA sequence libraries is one of the strengths of traditional clone libraries. Even with limited sample sizes, division-level trends become apparent and rarefaction curves of division-level diversity come close to reaching an asymptote (Dunbar *et al.*, 2002). However, the implications of division-level diversity are ambiguous for soil ecosystem function (Dunbar *et al.*, 2002). SARST facilitates a more in-depth analysis of the diversity present within each of the sampled divisions. By grouping like sequences from multiple RST libraries, the fine-scale abundance of individual RSTs will help to reveal populations in flux as a response to environmental parameters.

Clone libraries are frequently generated from genomic DNA extracted from soil and yet the question of representation is unanswered as SSU rDNA clone based studies do not prepare replicate libraries from the same soil sample. Generating clone libraries in replicate is not usually feasible, yet comparison of multiple samples is often an objective. Using SARST, our results indicate that, provided soil composites are carefully prepared and a rigorous DNA extraction protocol is employed, clone library-generated data is reproducible. However, there were some differences in the representation within the most abundant groups. The number of groups and the statistical estimators employed all suggest that the FOX-B2 RST library might be somewhat less diverse than the FOX-B1 library (Table 2). In addition, FOX-B2 is missing RST representation in two of the most abundant RST groups (Fig. 5). So while the replicates were highly correlated, there were still some differences that were presumably caused by the soil sample heterogeneity. As a result, caution must precede interpretation of sequence-based library comparisons, and replication should be included whenever possible.

Using SARST, we have collected over 2500 RSTs from a single composite soil sample. To date, this is the largest number of SSU rDNA sequences collected from a single microbial community. However, from the RST data we collected from the FOX-B arctic soil it is clear that even with numerous sequences accounted for, the microbial community requires further sampling to achieve a more comprehensive coverage. The presence of numerous singletons is common with soil library SSU rDNA clone library screening (Hughes *et al*., 2001), and it remains a characteristic of the FOX-B community where more than half of the RST groups are represented by a single RST (Table 2). Inadequately sampled clone libraries tend to underestimate the richness of the samples being studied. Perhaps not surprisingly, the Shannon index and Chao1 richness estimator both increased when the individual FOX-B replicate data was combined and neither reached a maximum (data not shown). Larger libraries of RSTs would be required to more completely, or exhaustively, profile the RST sequences in the FOX-B library. FOX-B was prepared as a composite soil sample and several DNA extractions were combined before SARST PCR. These steps were taken to ensure that the resulting RST library was representative of this particular arctic soil microbial community. Whereas the sample may be representative, the RST richness may have been compounded by mixing composites and multiple DNA extracts. For other soil samples where representation of a large soil area is not required, or for less diverse microbial communities, such as the rumen or activated sludge, we suspect that SARST will permit extensive or possibly complete coverage of microbial RST diversity. The actual number of sequences that must be sampled to allow reliable field-scale comparisons of multiple soil samples has recently been predicted. Estimates range between hundreds or thousands of clones (Tiedje *et al*., 1999; Hughes *et al*., 2001), to tens or hundreds of thousands of clones (Dunbar *et al*., 2002). With thousands of bacterial species present in every gram of soil (Rossello-Mora and Amann, 2001), the number of clones that must be sequenced must at least be in the thousands. SARST provides a high-throughput way to generate RST sample sizes of this order.

Because the RST sequences generate specific primers, this indicates that the specificity of these V1 regions could also be sufficient for the design of hybridization probes. Bertilsson *et al*. (2002) have explored the possibility of using the bacterial V1-region for hybridization with initial success. Clone-library based methods are not easily amenable to the comparison of numerous samples with replication. But, the successful use of microarrays would permit comparisons of many samples with a desired level of replication. SARST data provides abundant RST sequence information, specific to a microbial community of interest, that could facilitate the design of community-specific microarrays.

Experimental procedures

Pure cultures and soil samples

Escherichia coli K12 (ATCC 10798), *Staphylococcus aureus* NCTC 8325, *Bacillus subtilis* strain 168 (ATCC 23857) were grown in Tripticase soy broth media (Difco). *Caulobacter crescentus* CB15 (ATCC 19089) and *Burkholderia* sp. LB400 were grown on peptone-yeast extract (PYE) medium (0.2% peptone, 0.1% yeast extract, 0.01% $CaCl_2$, 0.02% $MgSO_4$). All cultures were grown at 30°C to early stationary phase then placed on ice. Aliquots of serial dilutions were plated in five replicates on their corresponding solid medium. Using these same aliquots, acridine orange direct counts (AODC) were performed as previously described (Hobbie *et al*., 1977). Culture aliquots were immediately frozen at −80°C until DNA extraction. Mean CFU counts from the solid media and viable cell counts from AODC were averaged. These cell count estimates were used to mix 2.3×10^8 cells from each culture, making up a total volume of 0.5 ml for DNA extraction.

Soil was sampled beside Nadluardjuk Lake, at a former auxiliary Defense Early Warning site called FOX-B, on the west side of Baffin Island (73°12′ N, 68°37′ W) in Nunavut, Canada. Ten replicate soil samples were taken from a defined grid and stored at 4°C during transport to our lab. The 10 soil samples were sieved (5 mm) and an equal weight of each sample was mixed to form a composite. The composite was mixed well before DNA extraction.

DNA extraction

The FOX-B composite soil sample was split into duplicate portions (FOX-B1, FOX-B2) and DNA was extracted from

triplicate 0.5 g subsamples from each portion using the soil FastDNA SPIN kit in conjunction with a FastPrep Instrument (Qbiogene, Carlsbad, CA), with the following modification to the manufacturer's protocol. Sodium phosphate buffer and MT buffer were added again to each of the triplicates after removing the supernatant from the first lysis and centrifugation steps and these steps were repeated to ensure maximal yields of DNA. Combining triplicate DNA solutions from both the first and second lysis steps produced one DNA extract for SARST. DNA extractions from 0.5 ml of individual pure cultures or from the defined mixed culture were done without replication but with the same modification described above. All extracted DNA was quantified in a 1% agarose gel stained with ethidium bromide using an AlphaImager 1200 (Alpha Innotech, San Leandro, CA) and diluted in sterile dH_2O to give equal concentrations (5 ng μl^{-1}) for PCR.

PCR primer design

SARST PCR primers were designed using Alignment Scanner (Neufeld and Mohn, 2002a), a software tool we developed to analyse ribosomal alignment slices from RDP-II. By downloading 1 bp slices of bacterial sequences from the RDP-II (release 8.1) Alignment Scanner determined the nucleotide representation of each position flanking the RST region (data not shown). Degenerate bases were introduced into the primer if the percent representation of a nucleotide was greater than 1%. The 3′ primer region is complimentary to almost all known bacterial SSU rDNA sequences. Both primers have 5′ extensions with type IIS restriction enzyme recognition sites (*Bpm*I on Bac64f and *Bsg*I on Bac104r) for subsequent enzymatic removal. Both primers were synthesized with dual 5′ biotin label and were HPLC purified by Integrated DNA Technologies (Coralville, IA).

PCR

A schematic representation of SARST is shown in Fig. 1. Each SARST library requires a total of eighteen 50 μl PCR reactions. The PCR reactions were performed as described previously (Yu and Mohn, 2001), except for using Bac64f-*Bpm*I (5′-Dual biotin-**TTT ACC TGG AGC C**TW ANR CAT GCA AGT CG; bold is extension) and Bac104r-*Bsg*I (5′-Dual biotin-**TTG CTG TGC AG**T ACK CAC CCG TBY GCC) as the primers, increased $MgCl_2$ (2.0 mM) and decreased amount of DNA template (5 ng per reaction). For the defined community, we prepared RST libraries using both *Taq* DNA polymerase (Taq) and a high fidelity enzyme mixture (HiFi). For the HiFi library we used the PCR buffer, magnesium, and enzyme from the supplier (Platinum *Taq* Polymerase High Fidelity, Invitrogen, Burlington, ON) but otherwise used identical reaction component concentrations and cycling conditions as used for the Taq library. Simplified PCR hot-start was used to increase specificity (Yu and Mohn, 2001).The PCR reaction consisted of an initial denaturation at 95°C for 2 min, followed by 25 cycles at 94°C for 30 s, 50°C for 30 s, and 72°C for 10 s, with a final extension of 72°C for 5 min. All PCR products for each library were pooled and the products were checked visually by running 5 μl on a 12% PAGE gel (19 acrylamide:1 bisacrylamide) for 150 V for 50 min on a Mini-PROTEAN II Cell (Bio-Rad, Mississauga, ON) with 300 ng 10 bp ladder (Invitrogen) to check for the expected 75–110 bp products. Unless indicated otherwise, all 12% PAGE gels in the SARST protocol were run with 300 ng 10-bp ladder. The PCR products were divided into three 300 μl aliquots in 1.5 ml microfuge tubes and extracted with equal volumes of phenol/chloroform (P/C). The used P/C tubes were extracted again with 100 μl LoTE (3 mM Tris-HCl, 0.2 mM EDTA, pH 7.5) and the aqueous phases from the three secondary extractions were pooled. The DNA was precipitated by adding 5 μl glycogen, 133 μl of 7.5 M ammonium acetate, and 1 ml 100% ethanol and leaving on ice for 30 min after mixing. After centrifuging for 15 min at 4°C, the supernatant was carefully aspirated. The pellets were washed twice with 80% ethanol and centrifuged for 5 min after each wash. The pellets were dried for 10–15 min at room temperature before being suspended in 8.75 μl of LoTE each. The DNA was dissolved for 30 min on ice and then PCR products were combined into one tube.

Digest with BpmI *and* BsgI

To enzymatically cleave primers from RST sequences, the following components were added to the sample DNA solution: 5 μl of 10 × NEBuffer 3 [New England Biolabs (NEB), Pickering, ON], 2.5 μl of fresh 20 × S-adenosylmethionine (SAM; SAM is supplied from NEB at 400 ×), 2.5 μl 20 × BSA (supplied at 100 × from NEB), 2.5 μl (5 U) *Bpm*I, and 2.5 μl (7.5 U) of *Bsg*I. This reaction was incubated at 37°C for 2–3 h. To complete the digestion the following was added to the reaction tube: 35 μl LoTE, 5 μl of 10 × NEBuffer 3, 2.5 μl of 20 × SAM, 2.5 μl 20 × BSA, 2.5 μl *Bpm*I and 2.5 μl of *Bsg*I. This mixture was incubated overnight at 37°C. One μl of reaction mixture was run on a 12% PAGE gel (19 : 1) to check the completeness of the digestion. An intense ~30 bp band at the bottom of the gel corresponded to the biotinylated primer.

Remove primers with streptavidin beads

To each of three microfuge tubes, 200 μl (2 mg) Dynal M-280 beads (Dynal Biotech, Lake Success, NY) were added and washed twice with 1 × magnetic bead binding buffer (5 mM Tris-HCl at pH 7.5, 1 mM EDTA, 1 M NaCl) using an MPC magnet (Dynal Biotech) for 2 min to collect the beads after each wash. Following addition of 100 μl 2 × magnetic bead binding buffer (Dynal Biotech) to the double-digested sample, sequential purification with 15 min incubations in each tube was conducted according to the manufacturer's protocol. A 100 μl wash with 1 × magnetic bead binding buffer added sequentially to each of the three tubes maximized RSTs recovery. The 300 μl purified sample was extracted with an equal volume of P/C and the DNA was precipitated by adding 5 μl glycogen, 100 μl of 7.5 M ammonium acetate, and 1 ml 100% ethanol. Incubation on ice, centrifugation, washing and drying of the DNA was done as described above. The dried pellet was suspended in 42.5 μl LoTE. A 0.5 μl aliquot of the DNA solution was run on a 12% PAGE gel (19 : 1) to ensure that undigested product was removed.

Ligate the RST and linkers

Linker oligonucleotides were synthesized with either dual 5′ biotin label (Linkers A2, B2) or with 5′ phosphorylation (Linkers A1, B1) and were HPLC purified by Integrated DNA Technologies. Double-stranded linker A was generated by mixing equal volumes of 90 μM linker A1 (5′-P-CTA GTA CGT GCT GGT) and 90 μM linker A2 (5′-Dual biotin-AAC ACC AGC ACG TAC TAG TC), heating to 95°C for 30 s in a thermal cycler, then turning off the instrument and allowing slow cooling for 1 h before storage in aliquots at −20°C. Linker B was generated from linker B1 (5′-P-CTA GCA ACG TGC TGG T) and linker B2 (5′-Dual biotin-AAC ACC AGC ACG TTG CTA GCC) using the same method. To ligate linkers to each end of the RSTs, the following was added to the 42 μl RST solution: 10 μl of annealed linker A, 10 μl of annealed linker B, and 18 μl 5 × ligase buffer (Invitrogen). The tube was heated for 2 min at 40°C then allowed to sit at room temperature for 15 min. After adding 10 μl T4 DNA ligase HC (50 U), the reaction was incubated at 23°C for 4 h or overnight. One μl of this reaction was run on a 12% (19 : 1) PAGE gel to check the ligation efficiency.

Purify RST-linker ligation products with PAGE

PAGE purification isolates RSTs with linker A and linker B attached. The entire sample was loaded in a large well (6–7 cm wide) on a 12% PAGE (19 : 1), with 1 μg 10 bp ladder as standard in a separate lane. The large well was made by taping around eight teeth of a Protean II mini comb. The gel was run at 75 V for 150 min. After staining with ethidium bromide and under long-wave UV a gel slice corresponding to ~70–100 bp was carefully cut from the gel. The bottoms of three 0.5-ml tubes were pierced with a 22-guage needle and placed into 2 ml tubes. The gel slice was divided into the three 0.5-ml tubes and centrifuged at maximum speed for 2 min to slurry the gel slices. To disassociate DNA from the gel, 250 μl LoTE and 50 μl 7.5 M ammonium acetate was added to each 2 ml tube and vortexed briefly. The tubes were then incubated at 50°C for 10 min, on ice for 10 min, and then at 50°C for a further 10 min (the tubes were then stored at 4°C overnight). Gel slurries were transferred onto 3 Spin-X columns (Corning, Corning, NY) and centrifuged at maximum speed for 5 min. The eluates were transferred into new microfuge tubes. After adding 84 μl LoTE and 17 μl 7.5 M ammonium acetate to the gel material in each Spin-X column, the heating and cooling cycles were repeated. The columns were then centrifuged at high speed for 5 min and these three eluates were pooled in a fresh 1.5 ml tube. The DNA in each tube was precipitated with 5 μl glycogen, 83 μl 7.5 M ammonium acetate, and 1 ml 100% ethanol. Incubation on ice, centrifugation, washing and drying of the DNA was done as described above. Each pellet was suspended in 9.5 μl of LoTE. After dissolving DNA for at least 30 min on ice the RST solutions were pooled. A 0.5 μl aliquot was run on a 12% PAGE gel (19 : 1) to check purity.

Release RSTs from linkers A and B with Spe*I and* Nhe*I and purification*

To the remaining 37.5 μl of the RST solution, 5 μl 10 × NE Buffer 2, 2.5 μl 20 × BSA, 2.5 μl *Spe*I (25 U), and 2.5 μl *Nhe*I (25 U) was added. The reaction was incubated at 37°C for 4 h. To further this reaction, 37.5 μl LoTE, 5 μl 10 × NEBuffer 2, 2.5 μl 20 × BSA, 2.5 μl *Spe*I, and 2.5 μl *Nhe*I was added. This reaction was incubated at 37°C overnight. Two μl was run on a 12% PAGE gel (19 : 1) to check if the digestion was complete. To purify RSTs, aliquot 200 μl Dynal M-280 beads (2 mg) to each of two 1.5 ml tubes. Prepare the beads and remove the linkers as described above, including the 100 μl wash. The 300 μl purified sample was extracted with an equal volume of P/C and 5 μl of extract was run on a 12% (19 : 1) PAGE gel to check removal of linkers. The 300 μl purified sample was extracted with an equal volume of P/C and the DNA was precipitated by adding 5 μl glycogen, 100 μl of 7.5 M ammonium acetate, and 1 ml 100% ethanol. Incubation on ice, centrifugation, washing and drying of the DNA was done as described above. The dried pellet was suspended in 10 μl LoTE and DNA was dissolved for 30 min on ice.

Ligation of RSTs to form concatemers

In a fresh 0.5 ml tube, the following was combined: 4.38 μl of purified RSTs, 1.25 μl of 5× ligation buffer, 0.25 μl *Spe*I, 0.25 μl *Nhe*I, and then 0.125 μl T4 DNA ligase HC. The contents were gently mixed and incubated for 10 min at 23°C. The tube was then transferred to 65°C for 10 min to stop the reaction and then placed at room temperature for 5 min. To the tube was added: 1 μl of 5 × ligation buffer, 3.5 μl H2O, 0.25 μl *Spe*I and 0.25 μl *Nhe*I (NEB). This digestion was incubated for 1 h at 37°C. To this mixture, 1.5 μl 10× loading buffer was added before transfering the tube to 65°C for 15 min, then cooling on ice for 10 min. This final heating step was added to ensure that compatible concatemer ends were not annealed prior to electrophoresis (Kenzelmann and Mühlemann, 1999). The entire sample was loaded in one lane of an 8% PAGE gel (37.5 : 1) at 100 V for 90 min. In a separate lane, 500 ng of a 1 kb DNA ladder (Invitrogen) was run. The gel was stained with GelStar (Cambrex, Rockland, ME) and regions of the gel corresponding to 500–700 bp, and 700–1000 bp were carefully cut out and placed into separate 0.5 ml tubes with bottoms pierced with a 22-gauge needle. We have since found that cloning the 300–500 bp region can also provide a successful set of suitable inserts. In fact, smaller size ranges typically provide smaller, but more reliable insert sizes for us and observations of SAGE data seems to confirm this (Koehl *et al*., 2003). Gels were slurried by placing the 0.5 ml tubes in 2 ml tubes and centrifuging for 2 min at maximum speed. To the 2 ml tubes, 300 μl of LoTE were added and the tubes were vortexed briefly. The tubes were heated to 65°C for 10 min. The contents of each tube was divided into two Spin-X tubes and centrifuged at full speed for 5 min. For each size range the eluates were combined into one new 1.5 ml tube. DNA was precipitated by adding 3 μl glycogen, 133 μl of 7.5 M ammonium acetate, and 1 ml 100% ethanol. Incubation on ice, centrifugation, washing and drying of the DNA was done as described above. The dried pellet was suspended in 7 μl of LoTE and DNA was dissolved for 30 min on ice.

Cloning, screening and sequencing of concatemers

Approximately half of the RST solution was used for ligation into *Spe*I-cut pZErO-2 (Invitrogen) according to manufacturer's protocols. The ligation products were purified with the MinElute Reaction Cleanup Kit (Qiagen) and eluted in 11 μl sterile dH_2O. Electrocompetent TOP10 cells (Invitrogen) were electroporated using 5 μl of ligation product and rescued in SOC medium according to manufacturer's protocols. Blue/white selection using low salt LB/Kan (50 μg ml^{-1}) plates with 40 μl of 20 mg ml^{-1} X-gal (Invitrogen) in dimethyl formamide spread on the surface of each plate possibly helps to indicate which inserts are of appropriate size for screening and sequencing (Angelastro *et al*., 2002). White colonies were screened by colony PCR in 10 μl reactions using modM13f (5′-CGCCAGGGTTTTCCCAGTCACGA) and modM13r (5′-AGCGAATAACAATTTCACACAGGA) primers. Adding 5% DMSO to the PCR reaction helped prevent secondary structures from affecting the PCR amplification. Reaction tubes were loaded into a thermocycler at 95°C. The PCR reaction consists of an initial denaturation at 95°C for 5 min, followed by 30 cycles at 95°C for 1 min, 55°C for 1 min, and 72°C for 1 min 30 s, with a final extension of 72°C for 7 min. Five μl of each reaction were run on a 1.5% agarose gel to check insert sizes.

Remaining reaction product with suitable size PCR products (500–1000 bp) were diluted to 20 μl with sterile dH_2O and cleaned before sequencing using Centrisep 96-well purification plates (Princeton Separations, Adelphia, NJ) by refilling the plates with autoclaved 7% (w/vol H_2O) Sephadex G-50 Fine (Amersham Biosciences, Baie d'Urfé, QC). Aliquots of 0.5–1.0 μl PCR product were sequenced in 5 μl reactions with 1 μl BigDye Terminator v3.1 (Applied Biosystems, Streetsville, ON) and M13r primer (5′-CAGGAAACAGCTATGACC). Sequencing reactions were diluted and cleaned using refillable Centrisep 96-well purification plates before drying, resuspension, and analysing on a Basestation 51 (MJ Research, Scarborough, ON) according to the manufacturer's protocols.

Sequence data

DNA sequences were manually verified for base-calling accuracy using Chromas vs. 2.23 (Technelysium, Queensland, AU) and RSTs were extracted from the resulting sequence text files using SARSTeditor (Neufeld and Mohn, 2002b). Using the M13r primer for sequencing inserts from the pZERO-II plasmid, SARSTeditor looks for and recognizes the folowing default border sequences: vector to head, GATCCACTAGTCG; vector to tail, GATCCACTAGCC; tail to head, GGCTAGTCG; head to tail, CGACTAGCC; tail to tail, GGCTAGCC; head to head, CGACTAGTCG; tail to end, GGCTAGTAACGG; head to end, CGACTAGTAACGG, and a user-aborted end to poor quality sequence data, TAGTAG TAG. All grouping of individual RSTs was done using Fastgroup (Seguritan and Rohwer, 2001) with the following changes to the default settings. The entire sequence was used for grouping, the search window size was set to five nucleotides, the per cent sequence identity for RSTs was set to either 100% or 95%, and the minimum sequence length was set to 10. The resulting Fastgroup data files (e.g. coverage.txt) were organized in an Excel spreadsheet and modified to serve as an infile for EstimateS. RST accumulation curves and richness estimators were generated using EstimateS [version 5.0.1; R. Colwell, University of Connecticut (http://viceroy.eeb.uconn.edu/estimates)] as previously described (Hughes *et al*., 2001). Divisions were assigned to individual RSTs based on the phylogenetic affiliation of the closest database hit in the RDP-II version 9.0 (Cole *et al*., 2003). Ungrouped FOX-B RST libraries are available for download (http://www.microbiology.ubc.ca/Mohn/SARST.htm).

RST primer design and conditions

Secondary structure prediction of RSTs prior to primer design was done online using GeneBee software (www.genebee.msu.su/genebee.html). RST-specific primers were designed to be approximately 20 nucleotides long. The 5′ end was positioned within the conserved region targeted by the Bac64f primer and the 3′ end targeted half of the RST, flanking the loop of the V1-region hairpin. For each RST primer, we tested a range of annealing temperatures (59.3, 61.0, 63.5 and 65.0°C) on a PTC-200 thermal cycler (MJ Research) using a mixture of FOX-B1 and FOX-B2 DNA as template and the 907r reverse primer (5′-CCGTCAATTC MTTTGAGTTT). The PCR reaction consisted of an initial denaturation at 95°C for 2 min, followed by 30 cycles at 94°C for 30 s, the selected annealing temperature for 30 s, and 72°C for 1 min, with a final extension of 72°C for 7 min. Three μl of each reaction were run on a 1.5% agarose gel. Aliquots from the highest annealing temperatures that yielded visible product were cloned into TOPO TA plasmids (Invitrogen) and electrocompetent TOP10 cells were transformed according to the manufacturer's protocols. Nine transformant colonies were screened by PCR using M13f and M13r primers and all PCR products were sequenced from both ends. The CHECK_CHIMERA program of the RDP-II (Maidak *et al*., 2001) failed to detect chimeric artifacts in the partial sequences of the 38 clones. DNA sequences with correct corresponding 3′ RST regions corresponding to approximate *E. coli* positions 83–534 were aligned using CLUSTALX (Thompson *et al*., 1997). A neighbour-joining phylogenetic tree was generated with 100 bootstraps using TreeCon for Windows 1.3b (Van de Peer and De Wachter, 1994). All alignment positions were used with insertions and deletions taken into account while calculating Jukes-Cantor distance. Classification of partial SSU rDNA sequences were based on the closest matching sequences in GenBank (Benson *et al*., 2000). The sequences generated with RST primers were deposited in GenBank with accession numbers AY330232-AY330269.

Acknowledgements

This work was supported by a scholarship to J.D.N. from the Natural Sciences and Engineering Research Council (NSERC) of Canada. We thank B. Chai and J. Cole of the RDP-II for invaluable help with RST specificity tests and with batch Sequence Match analysis. P. Axelrood is thanked for providing us with 16S rDNA clone library sequences. We thank W. A. Visser and P. Sue for contributing programming

skills. R. Hancock is thanked for providing access to DNA sequencing facilities. B. R. Steen, M. Murphy, J. F. Nomellini, J. Smit, and P. Fortin are thanked for helpful suggestions. We acknowledge J. Davies who provided us with *Staphylococcus aureus* NCTC 8325, J. Smit who provided us with *Caulobacter crescentus* CB15 and J. Tiedje for sending us *Burkholderia* sp. LB400.

References

Angelastro, J.M., Ryu, E.J., Törõcsik, B., Fiske, B.K., and Greene, L.A. (2002) Blue-white selection step enhances the yield of SAGE concatemers. *Biotechniques* **32:** 486.

Axelrood, P.E., Chow, M.L., Radomski, C.C., McDermott, J.M., and Davies, J. (2002) Molecular characterization of bacterial diversity from British Columbia forest soils subjected to disturbance. *Can J Microbiol* **48:** 655–674.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2000) Genbank. *Nucleic Acids Res* **28:** 15–18.

Bertilsson, S., Cavanaugh, C.M., and Polz, M.F. (2002) Sequencing-independent method to generate oligonucleotide probes targeting a variable region in bacterial 16S rRNA by PCR with detachable primers. *Appl Environ Microbiol* **68:** 6077–6086.

Borneman, J., Skroch, P.W., O'Sullivan, K.M., Palus, J.A., Rumjanek, N.G., Jansen, J.L., *et al.* (1996) Molecular microbial diversity of an agricultural soil in Wisconsin. *Appl Environ Microbiol* **62:** 1935–1943.

Borneman, J., Austin, S., and Triplett, E.W. (1997) Method development to assess microbial diversity in soil. In *Biotechnology Risk Assessment Symposium*. Levin, M., Grim, C. and Angle, J.S., (eds). Maryland, USA: University of Maryland Biotechnology Institute, pp. 10–16.

Chao, A. (1984) Non-parametric estimation of the number of classes in a population. *Scand J Stat* **11:** 265–270.

Chow, M.L., Radomski, C.C., McDermott, J.M., Davies, J., and Axelrood, P.E. (2002) Molecular characterization of bacterial diversity in Lodgepole pine (*Pinus contorta*) rhizosphere soils from British Columbia forest soils differing in disturbance and geographic source. *FEMS Microbiol Ecol* **42:** 347–357.

Clayton, R., Sutton, G., Hinkle, P. Jr, Bult, C., and Fields, C. (1995) Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa. *Int J Syst Bacteriol* **45:** 595–599.

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.

Crosby, L.D., and Criddle, C.S. (2003) Understanding bias in microbial community analysis techniques due to *rrn* operon copy number heterogeneity. *Biotechniques* **34:** 790–802.

Dunbar, J., Takala, S., Barns, S.M., Davis, J.A., and Kuske, C.R. (1999) Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Appl Environ Microbiol* **65:** 1662–1669.

Dunbar, J., Barns, S.M., Ticknor, L.O., and Kuske, C.R. (2002) Empirical and theoretical bacterial diversity in four Arizona soils. *Appl Environ Microbiol* **68:** 3035–3045.

Dunn, J.J., McCorkle, S.R., Praissman, L.A., Hind, G., Van Der Lelie, D., Bahou, W.F., *et al.* (2002) Genomic signature tags (GSTs): a system for profiling genomic DNA. *Genome Res* **12:** 1756–1765.

Farrelly, V., Rainey, F., and Stackebrandt, E. (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* **61:** 2798–2801.

Fogel, G.B., Collins, C.R., Li, J., and Brunk, C.F. (1999) Prokaryotic genome size and SSU rDNA copy number: estimation of microbial relative abundance from a mixed population. *Microbiol Ecol* **38:** 93–113.

Fox, G., Wisotzkey, J., and Jurtshuk, P. Jr (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42:** 166–170.

Hill, T.C.J., Walsh, K.A., Harris, J.A., and Moffett, B.F. (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol* **43:** 1–11.

Hobbie, J.E., Daley, R.J., and Jasper, S. (1977) Use of nucleopore filters for counting bacteria by fluorescence microscopy. *Appl Environ Microbiol* **33:** 1225–1228.

Hugenholtz, P., and Huber, T. (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int J Syst Evol Microbiol* **53:** 289–293.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67:** 4399–4406.

Kenzelmann, M., and Mühlemann, K. (1999) Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. *Nucleic Acids Res* **27:** 917–918.

Klappenbach, J.A., Dunbar, J.M., and Schmidt, T.M. (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66:** 1328–1333.

Koehl, A., Friauf, E., and Nothwang, H.G. (2003) Efficient cloning of SAGE tags by blunt-end ligation of polished concatemers. *Biotechniques* **34:** 692–694.

Kopczynski, E.D., Bateson, M.M., and Ward, D.M. (1994) Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. *Appl Environ Microbiol* **60:** 746–748.

Kroes, I., Lepp, P.W., and Relman, D.A. (1999) Bacterial diversity within the human subgingival crevice. *Proc Natl Acad Sci USA* **96:** 14547–14552.

Lee, C., Grasso, C., and Sharlow, M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics* **18:** 452–464.

Maidak, B.L., Cole, J.R., Lilburn, T.G., Parker, C.T., Saxman, P.R., Farris, R.J., *et al.* (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29:** 173–174.

McCaig, A.E., Glover, L.A., and Prosser, J.I. (1999) Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl Environ Microbiol* **65:** 1721–1730.

McCaig, A.E., Glover, L.A., and Prosser, J.I. (2001a) Numerical analysis of grassland bacterial community structure under different land management regimens by using 16S

ribosomal DNA sequence data and denaturing gradient gel electrophoresis banding patterns. *Appl Environ Microbiol* **67:** 4554–4559.

McCaig, A.E., Grayston, S.J., Prosser, J.I., and Glover, L.A. (2001b) Impact of cultivation on characterisation of species composition of soil bacterial communities. *FEMS Microbiol Ecol* **35:** 37–48.

Neufeld, J.D., and Mohn, W.W. (2002a) Alignment Scanner vs. 1.0. [WWW document] URL http://www.microbiology.ubc.ca/Mohn/SARST.

Neufeld, J.D., and Mohn, W.W. (2002b) SARSTeditor vs. 1.0. In [WWW document] URL http://www.microbiology.ubc.ca/Mohn/SARST.

Nogales, B., Moore, E.R., Abraham, W.R., and Timmis, K.N. (1999) Identification of the metabolically active members of a bacterial community in the polychlorinated biphenyl-polluted moorland soil. *Environ Microbiol* **1:** 199–212.

Nogales, B., Moore, E.R., Llobet-Brossa, E., Rossello-Mora, R., Amann, R., and Timmis, K.N. (2001) Combined use of 16S ribosomal DNA and 16S rRNA to study the bacterial community of polychlorinated biphenyl-polluted soil. *Appl Environ Microbiol* **67:** 1874–1884.

Pace, N.R., Stahl, D.A., Lane, D.J., and Olsen, G.J. (1986) The analysis of natural microbial populations by ribosomal RNA sequences. *Adv Microbial Ecol* **9:** 1–55.

Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., and Zhou, J. (2001) Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* **67:** 880–887.

Ravenschlag, K., Sahm, K., Pernthaler, J., and Amann, R. (1999) High bacterial diversity in permanently cold marine sediments. *Appl Environ Microbiol* **65:** 3982–3989.

Rossello-Mora, R., and Amann, R. (2001) The species concept for prokaryotes. *FEMS Microbiol Rev* **25:** 39–67.

Sambrook, J., and Russel, D.W. (2001) *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Seguritan, V., and Rohwer, F. (2001) FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics* **2:** 9.

Smit, E., Leeflang, P., Gommans, S., van den Broek, J., van Mil, S., and Wernars, K. (2001) Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. *Appl Environ Microbiol* **67:** 2284–2291.

Snaidr, J., Amann, R., Huber, I., Ludwig, W., and Schleifer, K.-H. (1997) Phylogenetic analysis and in situ identification of bacteria in activated sludge. *Appl Environ Microbiol* **63:** 2884–2896.

Speksnijder, A.G.C.L., Kowalchuk, G.A., De Jong, S., Kline, E., Stephen, J.R., and Laanbroek, H.J. (2001) Microvariation artifacts introduced by PCR and cloning of closely related 16S rRNA gene sequences. *Appl Environ Microbiol* **67:** 469–472.

Suau, A., Bonnet, R., Sutren, M., Godon, J.J., Gibson, G.R., Collins, M.D., and Doré, J. (1999) Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut. *Appl Environ Microbiol* **65:** 4799–4807.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **24:** 4876–4882.

Tiedje, J.M., Asuming-Brempong, S., Nüsslein, K., Marsh, T.L., and Flynn, S.J. (1999) Opening the black box of soil microbial diversity. *Appl Soil Ecol* **13:** 109–122.

Valinsky, L., Della Vedova, G., Scupham, A.J., Alvey, S., Figueroa, A., Yin, B., *et al.* (2002) Analysis of bacterial community composition by oligonucleotide fingerprinting of rRNA genes. *Appl Environ Microbiol* **68:** 3243–3250.

Van de Peer, Y., and De Wachter, R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* **10:** 569–570.

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

Wang, G.C., and Wang, Y. (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiol* **142:** 1107–1114.

Wang, G., and Wang, Y. (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol* **63:** 4645–4650.

Wintzingerode, F.V., 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.

Yu, Z., and Mohn, W.W. (2001) Bacterial diversity and community structure in an aerated lagoon revealed by ribosomal intergenic spacer analysis and 16S ribosomal DNA sequencing. *Appl Environ Microbiol* **67:** 1565–1574.