Miniprimer PCR, a New Lens for Viewing the Microbial World

Thomas A. Isenbarger,1,2 Michael Finney,2‡ Carlos Rios-Velázquez,3
Jo Handelsman,1* and Gary Ruvkun2

Departments of Bacteriology and Plant Pathology, University of Wisconsin—Madison, 1550 Linden Drive, Madison,
Wisconsin 53706-1521;1 Department of Genetics, Harvard University, and Department of Molecular Biology,
Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114-2790; and
Department of Biology, University of Puerto Rico—Mayagüez, Mayagüez, Puerto Rico.

Received 22 August 2007/Accepted 29 November 2007

Molecular methods based on the 16S rRNA gene sequence are used widely in microbial ecology to reveal the
diversity of microbial populations in environmental samples. Here we show that a new PCR method using an
engineered polymerase and 10-nucleotide “miniprimers” expands the scope of detectable sequences beyond
those detected by standard methods using longer primers and Taq polymerase. After testing the method in
silico to identify divergent ribosomal genes in previously cloned environmental sequences, we applied the
method to soil and microbial mat samples, which revealed novel 16S rRNA gene sequences that would not have
been detected with standard primers. Deeply divergent sequences were discovered with high frequency and
included representatives that define two new division-level taxa, designated CR1 and CR2, suggesting that
miniprimer PCR may reveal new dimensions of microbial diversity.

Characterization of 16S rRNA gene sequences has become a central feature of microbial ecology. Frequently, these anal-
yses are initiated by using PCR to amplify 16S rRNA genes directly from environmental samples without culturing (4). In
deed, these types of studies have transformed our view of the microbial world. However, as the 16S rRNA sequence data-
base has grown, it has become evident that many sequences deviate within the most conserved regions targeted by “universal”
16S rRNA gene PCR primers (8, 9, 38). To accommodate these deviations, commonly used primers have been modified with
degenerate positions to enable the primers to target a wider range of 16S rRNA gene sequences. However, because polynucleotides used for PCR require primers of ~20 to 30 nu-
cleotides (nt), 16S rRNA gene primer design has been con-
strained to target conserved regions of those lengths. New
thermostable polymerases have recently become available (40), opening the possibility of changes in primer design. Pre-
liminary data suggested that some of these polymerases may be able to utilize primers shorter than the minimum length of ~20 to
30 nt typically recognized by the standard enzyme used for
PCR, Taq (Thermus aquaticus) DNA polymerase.

We posited that decreasing the length of 16S rRNA gene
PCR primers would broaden the scope of sequences detectable by PCR. In this study we employ a new polymerase, S-Thr, and short primers to demonstrate the utility of a “miniprimer”
PCR method, both computationally and with environmental samples.

MATERIALS AND METHODS

PCR conditions. The S-Thr (DyNamo II; Finnzymes Oy, Espoo, Finland) and
Taq (Roche, Indianapolis, IN) DNA polymerases were used for PCR. S-Thr is an
engineered polymerase where the N-terminal 5′-3′ exonuclease domain of Ther-
mus brockianus DNA polymerase I has been removed and replaced by the 7-kDa
double-stranded DNA-binding protein Soo7d from Sulfolobus solfataricus (40).
Oligonucleotide primers were synthesized by Operon (Huntsville, AL) or Inte-
grated DNA Technologies (Coralville, IA). PCR was performed in 20 or 30 μl
with sterile, nucleic acid-free water (MO BIO Laboratories, Carlsbad, CA); 1×
Taq polymerase reaction buffer (10 mM Tris-HCl, 1.5 mM MgCl2, 5 mM KCl,
pH 8.3 at 20°C), 200 μM each deoxyribonucleoside triphosphate, 0.75 U of poly-
merase, and primers at a concentration of 10 pmol/μl. PCR components were
mixed on ice before placing them in a thermocycler preheated to 80°C. PCR
samples were incubated at 95°C for 120 s; thermally cycled 30 times at 94°C for
5 s, 40 to 50°C for 40 s, and 72°C for 60 s; and finally incubated at 72°C for 600 s.

Unless otherwise noted in the text and figures, an annealing temperature of 40°C
was used. MJ Research (Waltham, MA) PTC-100, PTC-200, and Tetrad PTC-
225 thermocyclers were used for thermal cycling.

Environmental sample preparation. For PCR from soil, three ~1-g samples
were taken from the courtyard outside the Wellman building on the Massa-
chusetts General Hospital campus (Boston, MA) and boiled in 1 ml sterile,
nucleic acid-free water for 5 min. After centrifugation to pellet solid material, the supernatants were combined, diluted 1:10 in sterile nucleic acid-free water, and
used for PCR. A boiling method was used to prepare DNA quickly to test
miniprimers and was not intended to provide material for a comprehensive
analysis of a soil sample. The microbial mat samples were collected from a
mat mature community of the Candelaria lagoon, which is one of a series of
evaporative salt marshes in Cabo Rojo, southwest Puerto Rico (17°56′N, 67°11′W), and
immediately frozen for storage. The mat is a laminated structure of ~2.0 to
2.5 cm in thickness and has distinctive green, pink, and black layers. For a
detailed physical and geochemical description of similar mats from Cabo Rojo,
see reference 11. To prepare microbial DNA mat, samples were thawed on
ice and three ~1-cm-wide vertical sections were cut so that all mat layers were
present in the pieces removed. The sections were mixed to form a sample of ~20 g
and then mixed to homogeneity with 20 ml of sterile, nucleic acid-free water.
After centrifugation to collect the solid material, the liquid was removed and
three ~1-g samples were removed from the top, middle, and bottom of the
pellet. The nine samples were processed individually using a bead-beating
method (MO BIO PowerSoil kit; MO BIO Laboratories, Carlsbad, CA); the
protocol included the optional 70°C initial incubation in solution C1. Equal
volumes of the nine DNA samples were combined, and this mixture was used as
the template in the PCR. For each PCR method tested, three separate PCRs
were performed with the microbial mat template, and the products of the three

* Corresponding author. Mailing address: Department of Bacteriol-
ogy, University of Wisconsin—Madison, 1550 Linden Drive, Madison,
WI 53706-1521. Phone: (608) 263-8783. Fax: (608) 265-5289. E-mail: joh@bact.wisc.edu.
‡ Present address: Finney Capital, 489 Douglass St., San Francisco,
CA 94114.
† Supplemental material for this article may be found at http://aem
.asn.org/.
‡ Published ahead of print on 14 December 2007.
PCRs were combined for cloning and subsequent sequencing. All sample handling was performed in a clean, sterile laminar flow hood.

Cloning and sequencing. PCR products were extracted from ~1.0% agarose gels and cloned using the TOPO cloning kit for sequencing (Invitrogen, Carlsbad, CA). Escherichia coli genomic DNA was prepared using the QiAamp DNA mini kit (Qiagen, Valencia, CA) or purchased from USB Corporation (Cleveland, OH); to prepare Halobacterium salinarum genomic DNA, H. salinarum strain MPK1 (24) was grown in culture as previously described (12), followed by genomic DNA preparation using the QiAamp DNA mini kit. Automated DNA sequencing of clones was performed by the Massachusetts General Hospital DNA Sequencing Core Facility (Cambridge, MA) or by the University of Wisconsin Biotechnology Center (Madison, WI) using sequencing primers located outside the bounds of the cloned 16S rRNA gene inserts. Sequence reads were assembled using Phred and Phrap as implemented in XplorSeq (D. Frank, unpublished data). The Staden package, release 1.6.0, Soil study clones A01 and A04 were fully sequenced with two- to fivefold coverage by independent sequence reads and had at least threefold coverage of all highly conserved regions. Mallard was used as the primary tool to identify and remove anomalous or chimeric 16S rRNA gene sequences (7); for some ambiguous cases, Bellerophon (20), Pintail (6), BLAST, and manual checking were used to supplement results from Mallard analysis.

Informatics and computation. Sequences were aligned using the NAST aligner (16) and imported into Arb version 05.10.26 (27), using the full greengenes database (see below) as a 16S rRNA reference database (15). The Arb database of Cabo Rojo and reference sequences is available at plantpath.wisc.edu/~isen/CR. Distance matrices calculated in Arb using the Olsen correction and Lane mask (25) were used to cluster sequences into operational taxonomic units (OTUs) by pairwise identity and a furthest-neighbour algorithm (33) and for calculating similarity indices to compare libraries (34). Microbial mat sequences were inserted into the Arb dendrogram using the parsimony tool and Lane mask. Primary taxonomic assignments were informed by the resulting Arb dendrogram and the Lawrence Berkeley National Laboratory (LBNL) greengenes classifier. Novel taxa identified with colloquial names and for which references are not otherwise indicated in the text include BDS-13 (23), BRC1 (13), FW129 (10), NBE23 (28), OP11 (21), W26, and W36 (17); publications for AC1 and Eub66 were not identified by our search. Similarities to database records were calculated by a maximum likelihood method as implemented in PHYLP, using F84 correction, a transition:transversion ratio of 2.0, empirically determined base frequencies, and the Lane mask (25). Sequences that were not placed into previously described taxa or that formed monophyletic groups of greater than 0.15 distance unit from the closest previously isolated sequence in the full LBNL greengenes database were subjected to additional characterization as follows. First, novel sequences were aligned with reference sequences representing major defined lines of bacterial descent (logical intersection of the “core” and “GOLD_stamp” sequence sets from greengenes.arb) and with their nearest-neighbor sequences. Then, the alignments were used to infer phylogenies by neighbor joining using evolutionary distances corrected by the Kimura two-parameter or maximum-likelihood correction using empirically determined gamma distribution and base frequencies (PAUP*4b10), maximum parsimony using heuristic search (PAUP*4b10), and maximum likelihood using the general time-reversible model (35). Bootstrap analysis was used to assess the robustness of inferred topologies. Finally, taxonomic identifiers CR1 to CR6 were assigned to monophyletic clusters that were reproducibly unaffiliated with previously described taxa in all topologies and which also contained at least three sequences greater than 1,000 nt from independent libraries.

Database searches. The environmental sequence database “env_nt” was downloaded from NCBI on 9 May 2006 (1,047,779 sequences and 1,063,283,128 total nucleotides; the list of accession numbers is available on plantpath.wisc.edu/~isen/CR). The following 16S rRNA gene sequence databases were downloaded from LBNL (14): the nonredundant, nonchimeric (“revered”) database of full-length rRNA gene sequences (11 November 2005 version) and the full 16S rRNA gene database (greengenes.arb; 27 April 2007 version). For searches of the environmental sequence database, putative PCR primer binding sites were identified with BLAST and putative amplicons were identified by Perl scripts (available at plantpath.wisc.edu/~isen/CR). For each primer search, pertinent BLAST parameters were as follows: no gaps, $M = 1, N = -3, V = 1,000,000, and B = 1,000,000$, and both the word size ($W$) and score threshold ($S$) were set to the length of the primer sequence used for the search. Putative 16S rRNA gene amplicons found in the environmental sequence database were confirmed by a BLAST score ratio (31) of greater than 0.80, computed for each sequence by dividing the BLAST score for its best match to the revered 16S rRNA gene database by the BLAST score of the sequence compared to itself; default BLAST parameters were used for these searches. All BLAST searches were implemented using WU-BLAST (blastn 2.0MP-WashU [10 May 2005]; W. Gish, personal communication).

Minimum primer length for Taq and S-Tbr polymerases. We conducted studies to determine the minimum primer length required for PCR amplification by the S-Tbr and Taq polymerases. For initial testing, primers were designed based on the nondegenerate versions of the 27F (sometimes called 8F) and 1492R primers used to amplify bacterial 16S rRNA genes (designated 27P-F and 1492P-R in this work) (Table 1) (18). Beginning with a 20-nt forward primer sequence (Fig. 1A), we designed a series of progressively shorter primers with lengths from 20 to 8 nt (Fig. 1A). To determine the lower limit for primer length, PCRs were performed using a forward primer from the series of decreasing lengths paired with a 19-nt reverse primer; E. coli genomic DNA was used as the template (Fig. 1B). With S-Tbr polymerase, successful amplification resulted with a primer as short as 10 nt. In a parallel experiment, Taq required the primer to be longer than 14 nt for detectable amplification (Fig. 1B). PCR products were produced with forward and reverse primers of 10 nt at annealing temperatures as high as $49^\circ$C (Fig. 1C). To distinguish this new PCR method from traditional Taq PCR, we refer to the shortened primers as “miniprimers” and to PCR using them as “miniprimer PCR”.

Miniprimer design and specificity. Several more 16S rRNA gene miniprimer candidates of 9 to 11 nt were designed (Fig. 2A) and tested to assess their performance and specificity (Fig. 3).
2B). Miniprimers were designed by consulting 16S rRNA gene sequence alignments and published primer sequences to choose the most conserved regions within several long 16S rRNA gene primers. We based our designs on variants most often reported in the literature (Table 1). The 788F-10 and 797R-10 primers (Fig. 2A) were designed to target the 10-nt region found to be the longest string of totally conserved bases in an alignment of 500 bacterial 16S rRNA gene sequences (8). By use of several combinations of miniprimers, PCRs were performed to amplify 16S rRNA genes from Escherichia coli or Halobacterium salinarum genomic DNA (from the domains Bacteria and Archaea, respectively). First, we selected miniprimer pairs that produced amplicons of the expected sizes with a minimum of spurious products. Then, amplicons from several PCRs were sequenced to verify that they were regions of 16S rRNA genes. All sequences from amplicons of the expected sizes were from the correct regions of the 16S rRNA gene targeted by each particular miniprimer pair. This analysis yielded several pairs of miniprimers that performed well for the amplification of bacterial and archaeal 16S rRNA genes (Fig. 2); in particular, 27F-10, 524F-10, and 2F-10 paired with 1505R-10 have reproducibly performed well in our analyses.

**In silico evaluation of miniprimer targets in environmental sequences.** To determine whether miniprimer PCR might expand the phylogenetic breadth of 16S rRNA gene sequences detected in environmental samples, the environmental sequence database was searched for putative 16S rRNA gene amplicons. It is important to note that this database compiles sequences obtained not by PCR but by random shotgun cloning of DNA directly extracted from environmental samples. A modeled PCR search identified putative 16S rRNA gene amplicons delimited by forward and reverse primer sequences separated by an appropriate distance (Fig. 3). Our goal was not to determine the fraction of 16S rRNA sequences in the database identified by the primers but to compare experimental methods for 16S rRNA gene discovery. Each database sequence and its reverse complement were searched using several of the best-performing primer pairs (Fig. 2); additional primer pairs were also designed to identify smaller amplicons.
because the average sequence length of the environmental sequence database entries is ~1,000 nt, lowering the probability of identifying long amplicons. To compare the abundances of putative amplicons defined by pairs of long primers and miniprimers, matched primer pairs were used to target the same regions of the 16S rRNA gene (Fig. 3). Miniprimers identified 1,648 putative amplicons and the standard primers identified 448; surprisingly, of the 1,648 miniprimer sequences, 971 were not identified by any long-primer pair tested. By BLAST comparison against a database of 30,312 nonredundant, nonchimeric bacterial and archaeal 16S rRNA gene sequences of at least 1,350 nt, 1,068 miniprimer and 301 long-primer sequences were verified to be regions of 16S rRNA genes (Fig. 3). Importantly, each miniprimer pair identified 2- to 10-fold more putative 16S rRNA gene amplicons (Fig. 3). This analysis predicted that miniprimer PCR can amplify more 16S rRNA gene sequences in environmental samples than can standard long primers.

Next, the phylogenetic breadth represented by the putative 16S rRNA gene amplicons was determined. In sum, the miniprimer and long-primer pairs identified 1,648 putative amplicons and the standard primers identified 448; surprisingly, of the 1,648 miniprimer sequences, 971 were not identified by any long-primer pair tested. By BLAST comparison against a database of 30,312 nonredundant, nonchimeric bacterial and archaeal 16S rRNA gene sequences of at least 1,350 nt, 1,068 miniprimer and 301 long-primer sequences were verified to be regions of 16S rRNA genes (Fig. 3). Importantly, each miniprimer pair identified 2- to 10-fold more putative 16S rRNA gene amplicons (Fig. 3). This analysis predicted that miniprimer PCR can amplify more 16S rRNA gene sequences in environmental samples than can standard long primers.

The phylogenetic tree was constructed using the remaining sequences, and the resulting taxonomic distribution indicated that almost all the sequences originated from Bacteria, with Proteobacteria sequences in the majority (see S1 in the supplemental material). For almost every taxonomic class, miniprimers identified more sequences than long primers. Also, miniprimers identified sequences in some classes for which long primers identified no sequences. This generally occurred for classes in which few members were identified but also occurred for Sargasso Sea group 11 (SAR11), the most abundant class of marine bacteria currently known (29). In addition, miniprimers identified nine sequences from Archaea, a domain for which long primers identified no sequences. This suggests that miniprimer PCR can amplify more sequences in more phylogenetic groups than can standard long primers.

**Evaluation of miniprimer PCR with a soil sample.** To determine whether the computational analysis reflected experimental results, miniprimer PCR was used in a study to amplify 16S rRNA genes from soil by use of the primer pairs 27F-10/1505R-10, 12F-10/1505R-10, and 524F-10/1505R-10 (Table 1 and Fig. 2). In this study, we did not attempt a comprehensive analysis of a soil sample but only an assessment of the miniprimers and whether they could detect sequences with mismatches to the standard primers. After PCR and cloning, 32 clones from each primer pair were partially sequenced from one end to examine the primer binding regions. Products from the 27F-10/1505R-10 PCR were chosen for further study because 27F-10 comprises the first 10 nt from the 5’ half of the 27F suite of primers (Table 1) and thus enabled the identification...
TABLE 2. Candelaria microbial mat 16S rRNA gene
libraries constructed

<table>
<thead>
<tr>
<th>Library</th>
<th>Forward primer</th>
<th>Reverse primer</th>
<th>Polymerase</th>
<th>No. of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>27F-10</td>
<td>1505R-10</td>
<td>S-Tbr</td>
<td>255</td>
</tr>
<tr>
<td>P1</td>
<td>27F-P</td>
<td>1492R-P</td>
<td>Taq</td>
<td>147</td>
</tr>
<tr>
<td>H1</td>
<td>27F-HT</td>
<td>1492R-HT</td>
<td>Taq</td>
<td>150</td>
</tr>
<tr>
<td>M2</td>
<td>27F-10</td>
<td>1505R-10</td>
<td>S-Tbr</td>
<td>343</td>
</tr>
<tr>
<td>P2</td>
<td>27F-P</td>
<td>1492R-P</td>
<td>S-Tbr</td>
<td>202</td>
</tr>
<tr>
<td>H2</td>
<td>27F-HT</td>
<td>1492R-HT</td>
<td>S-Tbr</td>
<td>184</td>
</tr>
</tbody>
</table>

The Candelaria microbial mat sequence libraries comprised over 40 bacterial divisions, with the Choroflexi, Bacteroidetes, Halanaerobiales, and Planctomycetes divisions having the greatest representation (Fig. 4). All of the most highly populated divisions identified were represented in each of the libraries, though in different proportions (Fig. 4). Notably, miniprimers amplified a greater proportion of sequences that could not be classified at or below the division level (see below).

The miniprimers amplified more sequences with poor matches to previously isolated 16S rRNA gene sequences (Fig. 5). Miniprimer libraries contained a larger fraction than did long-primer libraries of sequences that matched the database at distances greater than 0.10 (Fig. 5); conversely, the miniprimer libraries matched many fewer sequences at distances less than or equal to 0.05. The distributions of database matches for the two long-primer libraries were very similar and
subtaxa were largely populated by long-primer sequences, sug-
which miniprimer sequences contributed; the other
to be the largest single taxonomic group from any division to
study and from the Guerrero Negro microbial mats (26) and
Eub6 is almost entirely composed of sequences cloned in this
previously defined taxonomic groups. The
Chloroflexi
1492R-HT libraries, respectively).
from the 27F-P/1492R-P and 1.2% from the 27F-HT/
the two long-primer libraries combined contributing (1.7%
5.9% of the miniprimer sequences contributing and 1.5% of
was met all the criteria for defining new division-level taxa. A
similar environments (Fig. 6). A few groups contain clusters of
miniprimer sequences, in particular a subgroup within WS6
and the CR1 and CR2 groups (Fig. 6). The clusters CR1 and
CR2 (Fig. 6; also see S3 in the supplemental material) formed
at distances of 0.26 and 0.24, respectively, from previously
isolated sequences and thus are likely to be representatives
of new division-level taxa. Another four clusters formed at
distances slightly below the division level: CR3 to CR5 at 0.18 and
CR6 at 0.17 (see S3 in the supplemental material). Interest-
ingly, seven of the eight sequences defining CR1 and CR2, and
25 of the 32 sequences defining CR3 to CR6, are miniprimer
sequences. Many other sequences also branched deeply at dis-
tances of 0.29 to 0.16 to their nearest neighbors but did not
meet all the criteria for defining new division-level taxa. A
small fraction of library sequences accounted for these putative
novel groups: approximately 3.5% of library sequences comprised
monophyletic groups at distances of at least 0.17, with
5.9% of the miniprimer sequences contributing and 1.5% of
the two long-primer libraries combined contributing (1.7%
from the 27F-P/1492R-P and 1.2% from the 27F-HT/
1492R-HT libraries, respectively).

Other Candelaria sequences expanded the membership of
previously defined taxonomic groups. The Chloroflexi group
Eub6 is almost entirely composed of sequences cloned in this
study and from the Guerrero Negro microbial mats (26) and
was the largest single taxonomic group from any division to
which miniprimer sequences contributed; the other Chloroflexi
subtaxa were largely populated by long-primer sequences, sug-
gesting the existence of a primer bias for particular groups of
Chloroflexi. Similar subdivision bias is demonstrated by the
Halanaerobiales sequences: most Halobacteroidaceae family se-
quencies were from miniprimer libraries, whereas most
Halanaerobiales family sequences were from long-primer li-
braries. A large fraction of miniprimer sequences was clustered
within the Planctomycetes group and sequences from all libraries
were distributed over many Deltaproteobacteria groups. Five sequences contributed to the
Deltaproteobacteria group GN04, which was recently defined based on sequences isolated from the
Guerrero Negro microbial mat. Several miniprimer sequences populated the
Spirochaetes group GN05-1, which was first defined as a group of Guerrero Negro sequences and is
currently exclusive to sequences cloned from microbial mats. Very few miniprimer sequences were from the Bacteroidetes
and Firmicutes divisions, though large fractions of long-primer sequences were classified in these divisions (Fig. 4). The
remaining divisions into which library sequences were placed
each contributed less than 5% to their respective libraries.

To evaluate the reproducibility of the three library construc-
tion methods and to compare the methods, the 16S rRNA gene
sequence libraries were compared to each other by calculating
similarity indices (34) using 97% sequence similarity to define
OTUs. Here, we focus on the classic incidence-based Sørenson
similarity index, a measure of membership, and the Clayton 0,
a nonparametric maximum likelihood estimator of community
structure that considers abundance (41). These calculations
demonstrate two trends (Table 3): first, duplicate libraries con-
structed with the same primers show high levels of similarity in
both membership and structure; and second, libraries con-
structed with long primers are typically more similar to each
other than to libraries constructed with miniprimers. These
trends are particularly evident when relative abundances are
considered (0, Table 3). Importantly, the data suggest no sig-
nificant dependence of library membership or structure on the
particular polymerase used, as demonstrated by the high sim-
ilarities of duplicate libraries constructed with the same primer
pairs, 27F-P/1492R-P or 27F-HT/1492R-HT, but amplified with
different polymerases (Tables 2 and 3). Moreover, the
miniprimer method is as reproducible as either of the other
methods tested here using long-primer pairs. Thus, these data
suggest that differences in compositions of the libraries con-
structed with miniprimers and long-primer libraries are not
due to the enzyme used for amplification.

The coverage and richness of the 1,281 Candelaria se-
quencies were estimated by calculating rarefaction curves and
the nonparametric diversity estimators Cha01 and Ace1 (see
S4 in the supplemental material). Rarefaction analysis sug-
gested the sequences represented a diversity of at least 423
phylogenotypes defined at an identity threshold of 97% but indi-
cated that sampling was far from complete at this threshold. At
an 80% identity threshold, the estimates appear to have
reached plateaus, suggesting that sampling is more complete at
approximately the division level. Cha01 and Ace1 suggest the
presence of 91 to 124 taxa (95% confidence intervals) at
the 80% threshold, and the rarefaction analysis indicates that
the libraries represent 87 taxa at this level. Thus, though most
division-level taxa have probably been identified in the micro-
bial mat sample, these analyses suggest a high degree of diver-

![Figure 5](https://via.placeholder.com/150)

**FIG. 5.** Distribution of microbial mat best database matches. The
fractions of sequences whose best database matches fall into the des-
nignated distance intervals are shown for the long-primer (P, 27F-P/
1492R-10; H, 27F-HT/1492R-HT) and miniprimer (M) libraries. Dis-
tances of 0.05, 0.10, and 0.20 approximately define the taxonomic levels
of genus, order, and division, respectively (32). Plotted values and
errors are the averages and standard deviations of six maximum like-
lihood calculations using six different taxonomies (15).
Lastly, we examined the miniprimer clone sequences to quantify mismatches within the regions that would be targeted by the long primers 27F-P, 27F-HT, 1492R-P, and 1492R-HT (Fig. 7). For the 27F binding region, 309 and 266 of the 598 miniprimer sequences deviated from the 27F-P and 27F-HT primer sequences, respectively, at one or more nucleotides; sequences with two mismatches to 27F-P and 27F-HT numbered 20 and 10, respectively, and 1 sequence had three mismatches to the 27F-P sequence. Though the miniprimer sequence consensus largely agreed with the sequences of the long 27F primers, the nucleotide frequency at position 12 did not agree with the degeneracy present at this position in 27F-HT. Position 12 of 27F-HT is designed to match A or C; however, while in the cloned miniprimer sequence libraries C was present in the majority of sequences, G was present more than four times as often as was A, and T was observed twice. Similar analysis of the 1492R binding region identified one sequence from the miniprimer clone libraries having a single mismatch to 1492R-P and 1492R-HT (data not shown).

**DISCUSSION**

In this work, we show that 10-nt miniprimers can be used to identify divergent 16S rRNA gene sequences from environmental samples. Miniprimer PCR amplified a greater proportion than did standard primers of sequences that were novel or that poorly matched a database of previously isolated 16S rRNA gene sequences. This confirmed a prediction based on a computational search of the NCBI environmental sequence database. Importantly, we show that miniprimer PCR is as reproducible as other methods and provides a different portrait of microbial communities than is attained with current methods.

**Potential amplification of nontarget sequences.** One may expect that the shorter size of miniprimers could increase promiscuous binding to targets other than the 16S rRNA gene. However, analysis of environmental sequences suggests this may not be problematic. Of the 1,129 predicted miniprimer amplicons in the NCBI environmental sequence database (Fig. 3), 61 (~5.4%) were found not to be 16S rRNA gene sequences; a similar calculation for long primers yields a similar false-positive rate (19/320 ~ 5.9%). In addition, the rates of identifying mitochondrial and chloroplast sequences (see Table S1 in the supplemental material) were also the same (8/685 ~ 3/205 ~ 1%) for both methods. In support of this, the microbial mat libraries contained no eukaryote, chloroplast, or mitochondrial sequences, though eukaryotes are known to associate with the Candelaria mats (11). Overall, the results suggest that miniprimers can identify more 16S rRNA gene sequences without an increased rate of amplifying false positives. However, much additional testing and sequencing of amplicons isolated from environmental samples should be completed to test this prediction.

**Benefits of miniprimer PCR.** Adding miniprimer PCR to the tools used for analyzing microbial communities may enable a more accurate measure of 16S rRNA gene sequences in environmental samples by expanding the sequences detectable by PCR. For example, in the 1 Gbp of randomly cloned sequence in the environmental sequence database, if one assumes a microbial genome size of 2 to 3 Mbp and an average of 2.5 16S rRNA genes per genome (2), it is expected that ~800 to 1,200 16S rRNA genes are present in the sample. When totaled, the miniprimer computational searches identified 1,068 putative 16S rRNA gene amplicons (Fig. 3), in close agreement with the estimate, while the searches with the longer primers totaled 301 (Fig. 3). If one assumes that decreasing the length of a PCR primer by 1 nt can increase by fourfold the number of perfectly matched sequence targets, then reducing the length of a primer from 20 nt to 10 nt could theoretically increase the possible targets by a factor of a million (~ 4^10). Sequence variation within the regions targeted by these particular miniprimers may render the theoretical limit meaningless in this situation at a greater taxonomic resolution and that many organisms have yet to be identified.

**TABLE 3. Estimation of library similarities**

<table>
<thead>
<tr>
<th>Library</th>
<th>M1</th>
<th>M2</th>
<th>P1</th>
<th>P2</th>
<th>H1</th>
<th>H2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.60 ± 0.09</td>
<td>0.64 ± 0.07</td>
<td>0.07 ± 0.02</td>
<td>0.08 ± 0.03</td>
<td>0.06 ± 0.02</td>
<td>0.08 ± 0.03</td>
</tr>
<tr>
<td>M2</td>
<td>0.58 ± 0.16</td>
<td>0.49 ± 0.14</td>
<td>0.06 ± 0.02</td>
<td>0.03 ± 0.01</td>
<td>0.06 ± 0.02</td>
<td>0.08 ± 0.03</td>
</tr>
<tr>
<td>P1</td>
<td>0.50 ± 0.12</td>
<td>0.37 ± 0.10</td>
<td>0.76 ± 0.15</td>
<td>0.52 ± 0.08</td>
<td>0.51 ± 0.11</td>
<td>0.56 ± 0.10</td>
</tr>
<tr>
<td>P2</td>
<td>0.42 ± 0.15</td>
<td>0.49 ± 0.16</td>
<td>0.88 ± 0.13</td>
<td>0.74 ± 0.15</td>
<td>0.73 ± 0.07</td>
<td>0.73 ± 0.07</td>
</tr>
<tr>
<td>H1</td>
<td>0.51 ± 0.12</td>
<td>0.55 ± 0.13</td>
<td>0.72 ± 0.10</td>
<td>0.60 ± 0.09</td>
<td>0.65 ± 0.11</td>
<td>0.65 ± 0.11</td>
</tr>
<tr>
<td>H2</td>
<td>0.50 ± 0.12</td>
<td>0.55 ± 0.13</td>
<td>0.72 ± 0.10</td>
<td>0.60 ± 0.09</td>
<td>0.65 ± 0.11</td>
<td>0.65 ± 0.11</td>
</tr>
</tbody>
</table>

*The Sørenson (S, lower left) and Clayton (H, upper right) indices were computed for each pair of libraries using an OTU definition of 97% sequence identity. For both indices, values can range from 0 to 1, with 1 indicating identity. Values are reported with their standard errors. See Table 2 for details regarding the six libraries.
instance; nonetheless, this upper limit could inform other applications of the method.

In addition to expanding the range of detectable targets, combining miniprimer PCR with standard techniques might increase the accuracy of environmental sampling by enabling estimates of microbial diversity to reflect sampled communities more closely. When PCR is used to analyze samples containing multiple heterogeneous templates, many mechanisms can result in a biased distribution of amplicons that does not accurately represent the distribution of templates in the sample (3, 22, 30, 36). Miniprimers may reduce these kinds of biases in two ways. First, because miniprimers are shorter and can target the most conserved regions of the 16S rRNA gene, the potential for mismatch is reduced compared with long primers. The degeneracy of 27F-HT was not appropriate for the nucleotide frequencies of the miniprimer sequences (Fig. 7), suggesting that particular communities may be better analyzed with a miniprimer such as 27F-10. Second, because miniprimers target smaller regions, the requirement for degenerate positions in primers can be eliminated or minimized, thus removing or decreasing preferential binding of particular primer species within a degenerate primer mixture. Obviously, the use of shortened primers may introduce different biases into PCR-based analyses of microbial communities, and more study is required to understand that potential, perhaps using known mixtures of bacterial DNAs. However, despite the biases of any PCR method, miniprimer PCR may be useful to supplement traditional methods either to reduce or make apparent the biases that occur in analysis of environmental samples and allow a more accurate description of microbial communities.

**Miniprimer design.** Though general rules have been established for *Taq* PCR primer design, it is not yet evident what parameters require optimization when primers are substantially shortened. Successful amplification was obtained using miniprimers whose G+C content covered a broad range from 30 to 90% (Fig. 2). Thus, product formation appeared unrelated to the G+C content of miniprimers. In our tests, a high rate of success has been obtained with 1505R-10 paired with many different forward primers (Fig. 2). Of the putative primer binding sequences identified in the environmental sequence database, the fewest sites were found for 1505R-10, suggesting that a productive miniprimer maximizes specificity for the target sequence in the context of the total template sequence present in the PCR. This may be a good rule with which to begin miniprimer design. It should be noted that the miniprimers used here are initial versions and that they should be optimized further. For instance, neither 27F-10 nor 27F perfectly matches *Chlamydia*, and 788F-10 does not match *Deltabaculocystina*. Thus, some problems associated with long primers remain concerns in miniprimer design, and regularly checking, curating, and updating miniprimer sequences are as required for miniprimers as for the longer primers.

**Future prospects for miniprimer PCR.** The use of miniprimers in PCR can benefit many areas of biological research in several ways. PCR with miniprimers may enable PCR strategies that are not possible when primer design is restricted to a minimum length of ~20 nt. Some strategies might benefit from using even shorter primers—preliminary data indicate that miniprimers can be as short as 8 nt when paired with a primer of 15 to 20 nt (data not shown). Also, miniprimer sequences provide a small amount of additional sequence data that may be helpful for making fine discriminations among closely related sequences. Finally, the combination of highly processive enzymes and shorter primers may decrease the cost of PCR. The increased processivity enables PCR to be performed with smaller amounts of enzyme (40), and shorter primers are less costly to synthesize. Although cost savings may be minimal for small projects, these considerations become important for projects requiring very large numbers of PCRs.

**Acknowledgments**

We gratefully acknowledge Tuomas Tenkanen (Finnzymes Oy) for providing DNA polymerases and for helpful technical discussions, Todd DeSantis (Lawrence Berkeley National Laboratory) and Phil Hugenholtz (DOE Joint Genome Institute) for assistance with 16S rRNA gene sequence databases, Kirk Harris (University of Colorado) and Joe Felsenstein (University of Washington) for help with constructing phylogenetic trees, and members of the Ruvkun laboratory (Massachusetts General Hospital) and the Handelsman laboratory (University of Wisconsin) for helpful discussion of this work. We also acknowledge the SETG team, especially George Church (Harvard Medical School) and Maria Zuber (Massachusetts Institute of Technology) for thoughtful discussions regarding this research and the Massachusetts Institute of Technology Center for Space Research for guidance.

This work was supported by grant number F32AI065067 from the National Institute of Allergy and Infectious Diseases, grant number NRA-01-01-ASTID-020 from the National Aeronautics and Space Administration, and grant numbers MCB-0132085 and MCB-0455620 from the National Science Foundation Microbial Observatory Program.

The content is solely our responsibility and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases, the National Institutes of Health, the National Aeronautics and Space Administration, or the National Science Foundation.