Applied and Environmental Microbiology, May 2009, p. 3263–3270 Vol. 75, No. 10
0099-2240/09/$08.00+0 doi:10.1128/AEM.01931-08

Effects of Experimental Choices and Analysis Noise on Surveys of the "Rare Biosphere"▽†

Timothy J. Hamp,[1] W. Joe Jones,[2] and Anthony A. Fodor[1]*

Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina[1] and Environmental Genomics Core Facility, University of South Carolina at Columbia, Columbia, South Carolina[2]

Received 19 August 2008/Accepted 25 February 2009

When planning a survey of 16S rRNA genes from a complex environment, investigators face many choices including which primers to use and how to taxonomically classify sequences. In this study, we explored how these choices affected a survey of microbial diversity in a sample taken from the aerobic basin of the activated sludge of a North Carolina wastewater treatment plant. We performed pyrosequencing reactions on PCR products generated from primers targeting the V1-V2, V6, and V6-V7 variable regions of the 16S rRNA gene. We compared these sequences to 16S rRNA gene sequences found in a whole-genome shotgun pyrosequencing run performed on the same sample. We found that sequences generated from primers targeting the V1-V2 variable region had the best match to the whole-genome shotgun reaction across a range of taxonomic classifications from phylum to family. Pronounced differences between primer sets, however, occurred in the "rare biosphere" involving taxa that we observed in fewer than 11 sequences. We also examined the results of analysis strategies comparing a classification scheme using a nearest-neighbor approach to directly classifying sequences with a naïve Bayesian algorithm. Again, we observed pronounced differences between these analysis schemes in infrequently observed taxa. We conclude that if a study is meant to probe the rare biosphere, both the experimental conditions and analysis choices will have a profound impact on the observed results.

For nearly 3 decades, investigations of the distribution of microbes in complex environments have focused on the use of rRNA genes (1, 2, 4, 11, 16, 18, 19, 22, 24). Because the full-length 16S rRNA sequence can be obtained with paired-end reads via traditional Sanger sequencing, until recently most studies of the 16S rRNA gene captured most or nearly most of the 16S sequence length. New pyrosequencing technologies, however, have recently been introduced that greatly reduce the per base cost of sequencing but with shorter read lengths than traditional Sanger sequencing (17). This new approach has proven powerful, yielding a previously unobtainable view of rare taxa (7, 12–14, 25).

The shorter reads produced by pyrosequencing require the choice of a particular region of the 16S rRNA gene to target for pyrosequencing as well as the choice of an algorithm to classify the taxonomy of the shorter reads. In their initial surveys of microbial diversity with pyrosequencing (12, 14, 25), Sogin and colleagues targeted the V6 variable region, in part because it is was small enough to be captured with the 100-bp reads of the pyrosequencing technology available at the time. Recently, the read length of 454 pyrosequencing machines has been increased to an average of ~250 bp. This allows for more flexibility in primer design and opens up the possibility of targeting regions of the 16S rRNA gene other than V6. In recent work, Huse et al. took advantage of this new capability to compare the classifications made for the human gut microbiome with the V6 and longer V3 regions (13). Plotting the taxonomic abundance of these two sequence sets against each other yielded an excellent correlation ($r^2 = 0.99$), suggesting that the choice of which variable region to target makes little difference. In this report, we introduce a data set examining the performance of sets of primers targeting the V1-V2, V6, and V6-V7 regions. By using a sample for which we have also generated a whole-genome shotgun sequencing run with 250 bp reads, we were able to compare the observed 16S rRNA genes in samples with and without an initial PCR step targeting the 16S rRNA gene. Our results demonstrate that experimental choices such as which region of the 16S rRNA gene to sequence and which algorithm to use to classify taxa are much more likely to affect observations of the "rare biosphere" than more commonly observed taxa.

MATERIALS AND METHODS

Sample collection. The Mallard Creek Water Reclamation Facility is located in Charlotte, North Carolina. The plant has an average daily inflow of 7.5 million gallons, and the wastewater is mostly domestic, with additional input from the University of North Carolina–Charlotte, University City Carolinas Medical Center hospital, and several industrial users (see reference 21 for additional details). On the morning of 20 March 2007, we collected a 50-ml sample from the aeration basin using a plastic dipper. The sample was decanted to remove as much foam as possible before the liquid was transferred to a sterile tube. DNA was extracted from the sample using a Mo Bio UltraClean Water DNA Kit. The sample tube was inverted several times to maximize homogeneity, and a 10-ml aliquot was removed and pipetted onto the provided filter (0.22 μm pore size). Filtrate was discarded, and the membrane was used for bacterial DNA extraction using the manufacturer's protocol. The final DNA extract was analyzed for purity and concentration using a NanoDrop ND-1000 spectrophotometer. Approximately 100 μl of extracted DNA was concentrated in a vacuum centrifuge and resuspended in about 12 μl of molecular-grade biology water. The final sample concentration was determined by a NanoDrop spectrophotometer. A whole-genome shotgun sequencing run was performed on this sample, and a detailed description of these data can be found in our previous paper (21). The DNA

* Corresponding author. Mailing address: Bioinformatics Resource Center, University of North Carolina—Charlotte, Cameron 212, 9201 University City Boulevard, Charlotte, NC 28223. Phone: (704) 687-8214. Fax: (704) 687-6610. E-mail: anthony.fodor@gmail.com.

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

▽ Published ahead of print on 6 March 2009.

TABLE 1. Primers used in initial 16S rRNA gene experiments[a]

Primer	Sequence	Source or reference
27F	gcctccctcgcgccatcagMGAGTTTGATCCTGGCTCAG	15
337R	gccttgccagcccgctcagGCTGCCTCCCGTAGGAGT	Fig. 1
967F	gcctccctcgcgccatcagCAACGCGAAGAACCTTACC	25
1046R	gccttgccagcccgctcagCGACAGCCATGCANCACCT	
967F	gcctccctcgcgccatcagCAACGCGAAGAACCTTACC	25
1177R	gccttgccagcccgctcagCGTCATCCCCACCTTCCT	Fig. 1

[a] The lowercase sequence represents the 454A and 454B primers while the uppercase sequence represents conserved regions of the 16S rRNA gene.

sample was stored for just over 1 year at −20°C before the experiments described in this report were performed.

PCR conditions. A 454 pyrosequencing reaction requires the presence of two oligonucleotide sequences, the 454A and 454B primers, for the emulsion PCR reaction that amplifies DNA prior to pyrosequencing. In a whole-genome shotgun sequencing experiment, these primer sequences are blunt-end ligated to sheared sequence prior to emulsion PCR. If a pyrosequencing sequencing run instead uses PCR to focus on a particular gene, these primer sequences can be incorporated into the primers used in the initial PCR. Table 1 shows the full primer sequences used in our initial PCR targeting the 16S rRNA gene. Primers were ordered from IDT and were not purified by high-performance liquid chromatography.

On 7 April 2008, four 50-μl PCR mixtures were set up using *Pfu* Ultra (Stratagene) high-fidelity DNA polymerase by following the manufacturer's directions. The PCR conditions were modified from Sekiguchi et al. (22) through trial and error to find cycling conditions that appeared to work well with all three primer sets based on analysis of PCRs with agarose gels. For the sequencing reactions described in this paper, our cycling conditions were 94°C for 5 min and 20 rounds of 94°C for 1 min, 60°C for 1 min (with this temperature dropping 1°C every second cycle), and 72°C for 1 min. The annealing temperature was then set to 55°C for 10 rounds (i.e., 94°C 1 min, 55°C 1 min, and 72°C 1 min). Finally, the samples were exposed to 72°C for 7 min and then cooled to 4°C. Our PCR buffer contained 5 μl of 10× *Pfu* Ultra buffer, 1 μl of deoxynucleoside triphosphate mix at 10 mM for each deoxynucleoside triphosphate, 1 μl of template DNA from the wastewater treatment plant at 58.7 ng/μl, 1 μl of *Pfu* Ultra polymerase at 2.5 U/μl, 40 μl of nuclease-free water, and 1 μl each of the forward and reverse primers at 10 μM. Samples were run on an agarose gel and gel purified with the Promega Wizard SV gel and PCR clean-up system by following the manufacturer's instructions.

Shannon sequence entropy. The aligned version of the Ribosomal Database Project ([RDP] version 9.59) was downloaded from http://rdp.cme.msu.edu/. A reference sequence was chosen (*Escherichia coli* J01695), and a new alignment was created in which all the columns of this alignment were numbered by the reference sequence. That is, if within the alignment the reference sequence *E. coli* J01695 had a gap, that column was removed from every sequence in the alignment. We then calculated the Shannon sequence entropy (Fig. 1) (23) for every column in the new alignment. This measure of conservation is defined as:

$$\sum_{x = \{A,C,G,T\}} p_x(i) \ln p_x(i)$$

where i is the column of interest, x are the four valid nucleotides, and P_x is the frequency of nucleotide x at column i.

Filtering of primer sequences. We followed the recommendations of Huse et al. (14) and removed all sequences (Table 2) that had any N residues anywhere in the sequence, did not start with the expected 5′ primer sequence, or for which the sequence read lengths (including nucleotides derived from the primer sequence) were below 150 bp (70 for primer 967F-1046R). For sequences that survived our filters, we removed regions containing the upstream and downstream primer sequences by simply removing the first and last 20 bp (since sequencing error tends to accumulate toward the end of pyrosequencing reads, we wished to treat sequences from all primer sets uniformly regardless of read length). Unfiltered sequence sets are available as File S1 in the supplemental material.

Whole-genome sequences and the RDP classification algorithm. Using BLASTN with an e-score cutoff of 0.01 (and with the filter parameter −*F* "*m D*"), we ran our 378,601 454-FLX whole-genome shotgun sequences from our 20 March 2007 sample of the Mallard Creek Wastewater Treatment Plant (24) against version 9.60 of the RDP database (22). This search yielded 739 hits of which 467 could be assigned by stand-alone version 2.0 of the RDP classification algorithm to either *Bacteria* (464) or *Archaea* (3) at the top of the phylogenetic tree with a confidence score of >80 (Fig. 2).

Compared to our PCR runs, we have a relatively small number of whole-genome shotgun sequences that can be assigned to taxa above the RDP threshold of >80% (Fig. 2). Because we had a modest number of whole-genome-derived 16S rRNA sequences, we did not remove whole-genome shotgun sequences that contained at least a single N as we did for our PCR runs. There are 17 sequences

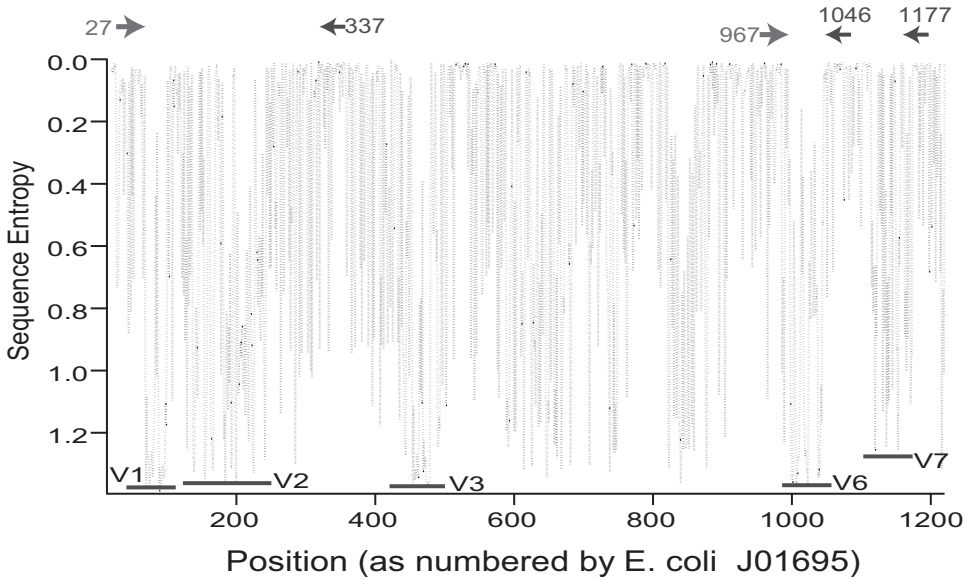

FIG. 1. Sequence conservation as a function of alignment position for the 489,840 sequences in version 9.59 of the RDP. The *x* axis shows the position in the alignment as numbered by the *E. coli* 16S rRNA gene. The *y* axis shows the Shannon sequence entropy (see Materials and Methods), a widely used measure of conservation in multiple sequence alignments (23). Highly conserved positions within the alignment have a sequence entropy close to zero and hence are shown toward the top of the *y* axis. Positions of the hypervariable regions V1-V3 and V6-V7 are derived from Chakravorty et al. (3).

TABLE 2. Sequences obtained via pyrosequencing of our 20 March 2007 aeration basin wastewater sample

Primer position	No. of sequences generated	Read length (bp)[a]	Fraction of sequences with correct 5′ primer sequence	Fraction of sequences with no N residues	No. of sequences remaining after filtering[b]
27F-337R	15,141	228.2 ± 57.0	0.96	0.88	11,316
967F-1046R	15,124	92.7 ± 13.6	0.95	0.95	12,277
967F-1177R	18,873	199.9 ± 41.6	0.96	0.76	11,645
Whole genome	378,601	250.4 ± 29.11			

[a] Read lengths are given as means ± standard deviations.
[b] The filters removed any sequences where the reads (including nucleotides derived from the primer sequence) were below a size cutoff (70 bp for 967F-1046R and 150 bp for all others), did not start with the exact 5′ primer sequence, or had any N residues anywhere in the sequence.

among the RDP-assigned 467 sequences that had at least one N. Of these, only seven were assigned as far as phylum (RDP classification level 3) with a confidence score of >80%. Differences in filtering strategy between the PCR and whole-genome runs, therefore, are unlikely to explain very much of our data, and the conclusions of our paper would be identical whether or not these seven sequences were included.

Implementation of JGast. A "nearest neighbor" algorithm maps a query sequence to a previously described full-length sequence in a database and then assigns the classification of the full-length sequence to the query sequence. There are several nearest-neighbor algorithms available as web servers (5, 6), but they have limitations in the number of sequences that can be uploaded, making them poorly suited to large pyrosequencing datasets. The code for the Global Alignment for Sequence Taxonomy (GAST) process (13) has been made publicly available (http://vamps.mbl.edu/resources/software.php) but requires creation of a database containing variable regions extracted from a full-length 16S rRNA alignment. As an alternative, we implemented a nearest-neighbor algorithm in Java (see File S2 in the supplemental material) that works directly on unaligned reference 16S rRNA sequences. Our implementation, which we call JGast, begins with classification of the 302,066 sequences in the Greengenes database (the file current_prokMSA_unaligned.fasta dated 16 December 2008 was downloaded from http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/on 7, January 2009) by the stand-alone RDP classification algorithm, version 2 (downloaded from http://sourceforge.net/projects/rdp-classifier/). In addition, a BLAST database was created from the current_prokMSA_unaligned-.fasta sequences using formatdb (with the default parameters except for "−p f"). For each query sequence, we performed a BLASTN search against this database. We used NCBI BLAST, version 2.2.18 for Linux, with the following parameters: −p blastn, v = 250, and −e 0.000001. For each query sequence, we took the top 250 sequences with the best BLAST match (i.e., lowest e-score) to the query sequence. For each of the 250 sequences found by BLAST, we performed a pairwise global alignment with the query sequence using the Needleman-Wunsch algorithm with a match score of 2, a mismatch score of −1, and an affine gap penalty of −2 for opening a gap and a score of 0 for extending the gap. Following Huse et al., we defined the distance of our pairwise Needleman-Wunsch-aligned sequences as the number of insertions, deletions, and mismatches, divided by the ungapped length of the query sequence. In determining the region of the target sequence to use for the global alignment, we took enough of the target sequence to ensure full coverage of the query sequence. For example, consider a BLAST alignment of 100 nucleotides between a query sequence of 175 nucleotides to a target sequence of 400 nucleotides that matched from nucleotides 200 to 300 of the target sequence. There are 75 nucleotides in the query sequence that were not present in the BLAST alignment. We wish to allow for the possibility that the missing 75 nucleotides could contribute to a global alignment on either the 5′ or 3′ side of the target sequence. We would therefore use the Needleman-Wunsch algorithm to align the full query sequence to a substring of 125 to 375 from the target sequence (plus an additional 10 nucleotides on either side to allow for gaps). Since trailing or leading gaps are treated as a single insertion event, the presence of the extra sequence does not substantially increase the calculated distance.

Of the 250 sequences aligned to the query sequence in this way, we followed Huse et al. and considered the taxonomy of the target sequence(s) with the best percent identity (i.e., lowest distance) from the Needleman-Wunsch alignment as reference sequences. We compared the classification of these reference sequences and generated a consensus taxonomy in which two-thirds of the reference sequences agreed on the taxonomy, with at least an 80% RDP confidence score on the full-length reference sequence. If two-thirds of the sequences did not agree on a classification with an 80% RDP confidence score at a given taxonomic level (e.g., genus), we moved to the next higher taxonomic level on the tree (e.g., family).

Although JGast shares many similarities with GAST, there are some differences that could yield different results. GAST starts with RDP classifications of the SILVA database. We instead used RDP classifications of the Greengenes database in part because the Greengenes database has a longer average sequence length (1,416 bp for Greengenes versus 1,005 for SILVA). GAST uses a blast database consisting of an alignment of just the target variable regions and uses a multiple sequence alignment program (MUSCLE) (8, 9) to align the query sequence with the 100 best BLAST hits. Distance scores are calculated on this global alignment. JGast, by contrast, calculates distance scores from a pairwise

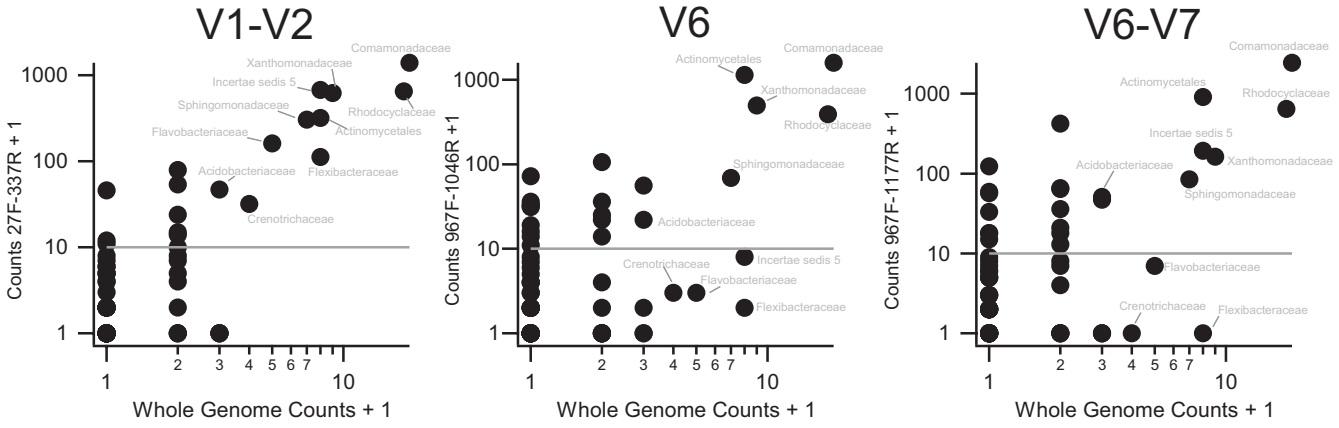

FIG. 2. Number of sequences assigned at the family classification level by the RDP classification algorithm to different sequences for the V1-V2, V6, and V6-V7 primers plotted against the number of 16S sequences assigned to the whole-genome shotgun sequence set. One has been added to each count to allow the data to be shown on a log-log plot.

TABLE 3. Number of sequences classified by RDP analysis scheme with a confidence score of >80%[a]

Region	No. of sequences assigned to:							
	Root	Kingdom	Phylum	Class	Order	Family	Genus	Species
V1-V2	11,316	11,313	10,646	9,824	6,979	4,647	2,287	205
V6	10,950	10,778	7,386	6,410	4,862	4,257	1,780	702
V6-V7	11,645	11,643	10,144	8,954	6,544	5,563	2,554	871
Whole genome	647	435	148	136	111	101	62	5

[a] See Fig. 2 and 3.

Needleman-Wunsch alignment of the query sequence to each of the 250 best BLAST hits found in a BLAST database consisting of full-length sequences of the Greengenes database. Finally, GAST considers the taxonomy of all sequences in the SILVA database that contain an exact match to the reference sequence substring containing the variable region. By contrast, JGast considers only the taxonomy of sequences among the 250 best BLAST hits that have an exact match to the substring of the target sequence used for the global alignment. We do not anticipate these different choices, employed for expediency or performance or to take advantage of an existing Java code base, would make large differences in the results that we obtain. Indeed, our results strongly suggest that different classifications that arise due to different choices made during construction of algorithms will mostly be limited to the rare biosphere.

In Fig. 1 and 2, we show sequences that are classified by the RDP classification algorithm with a confidence score of at least 80% (the number of sequences assigned in this way is shown in Table 3). In Fig. 3 and 4, we follow Huse et al. and show only assignments where the percent identity of the query sequence and reference sequence in the Greengenes database is >85% and where the reference sequence(s) has a consensus classification of a given taxonomic level with an RDP confidence score of at least 80% (the number of sequences assigned in this way is shown in Table 4). In order to ensure that the differences in Fig. 3 and 4 are not simply the result of different thresholds being applied to generate classifications of different sequence sets, for the direct RDP classifications of the query sequences in these figures we took the same set of sequences for which JGast had made a classification and used classifications from the RDP algorithm, regardless of confidence score. That is, we let JGast choose the sequences based on a maximum distance to the reference sequence of 0.15 and then took RDP assignments for this sequence set even if the RDP confidence score was less than 80%.

Linear regressions. All linear regressions were performed on log-transformed data with 1 added to each observed taxon before the log transformation. Taxa that were not present in either condition were not included in the regression.

Nucleotide sequence accession number. Sequences in this study are available at the NCBI short-read archive under accession number SRA001164 and as File S1 in the supplemental material.

RESULTS

Primer sets were chosen based on previous literature and an analysis of the RDP. Figure 1 shows conserved regions of the 16S rRNA gene. Based on our analysis in Fig. 1, we chose three sets of primers for evaluation (Table 1). We looked for pairs of primers from conserved regions that had melting temperatures, as predicted by Primer3 (20), of ~60°C (without the 454A and 454B sequences). To allow for easier comparison to other studies, where possible we chose primers that met our criteria that were also used in previous studies. We therefore included the primers targeting the V6 region that the Sogin group used in their first paper (25) (primer 967F-1046R). We also included sets of primers (967F-1177R) that start with the 5′ primer from Sogin et al. but extend further past the V7 region (967F-1177R) as well as primers that target the V1-V2 regions (27F-337R) (Table 1; Fig. 1).

Once the primers were chosen, we used as a template a DNA sample taken on 20 March 2007 from the aerobic basin of the Mallard Creek Wastewater Treatment Plant for which we had also performed a 454-FLX whole-genome shotgun sequencing reaction (21). Using the three sets of primers in Table 1, we performed 30 rounds of PCR on this sample (see Materials and Methods), gel purified the resulting amplicon, and submitted the resulting DNA for 454-FLX pyrosequencing at the Environmental Genomics Core Facility in Columbia, South Carolina. Each of the three reactions was run on 1/16 of an LR70 plate. The number of sequences that was generated for each primer pair is shown in Table 2.

Differences between whole-genome results and results from the 16S rRNA gene are pronounced for infrequently observed taxa. Although it is the basis of much of the biotechnology revolution, conventional PCR is known to have limits as a quantitative technology. Primer bias, saturation of amplicon at higher cycle numbers, and a stochasticity that is an inherent part of the PCR process can all limit the degree to which conventional PCR can be reliably used as a quantitative tool. A comparison of our whole-genome-derived 16S rRNA sequences with sequences derived from PCR targeting the 16S rRNA gene on the same DNA sample allows us to evaluate the degree of bias introduced by different PCR primers during the initial PCR step. Figure 2 shows this comparison for the family level for the V1-V2, V6, and V6-V7 primer sets (data for all taxonomic levels are given in File S3 in the supplemental material). For all primer sets, we see a rough correspondence, with a slightly better agreement for V1-V2. We note, however, that most of the differences occur in taxa that we observe 10 times or less in each PCR sequence set. If we limit our regression to sequences that we observed 11 times or more in the PCR sequence sets (Fig. 2), the correspondence between the whole-genome sequence sets and PCR sequence sets is reasonable. Figure 3 shows that, across taxonomic levels, linear regressions between the whole-genome and PCR sets involving only sequences we observed 11 times or more in the PCR sequence sets are much more positively correlated than regressions involving sequences we saw fewer than 11 times. We conclude that (i) different PCR primers yield profoundly different results for less abundant taxa and (ii) primers targeting the V1-V2 region do a somewhat better job of finding taxa that are also present in the whole-genome data set.

Differences between different analysis schemes also most significantly affect infrequently observed taxa. In Fig. 2 and 3, assignments were performed using the RDP classification algorithm (26). A popular alternative method is to use a nearest-neighbor approach (5, 13). In this approach, a query sequence is first mapped to a full-length sequence in an existing database, and this full-length sequence is used to assign taxonomy. Huse et al. used this approach, in an implementation they call

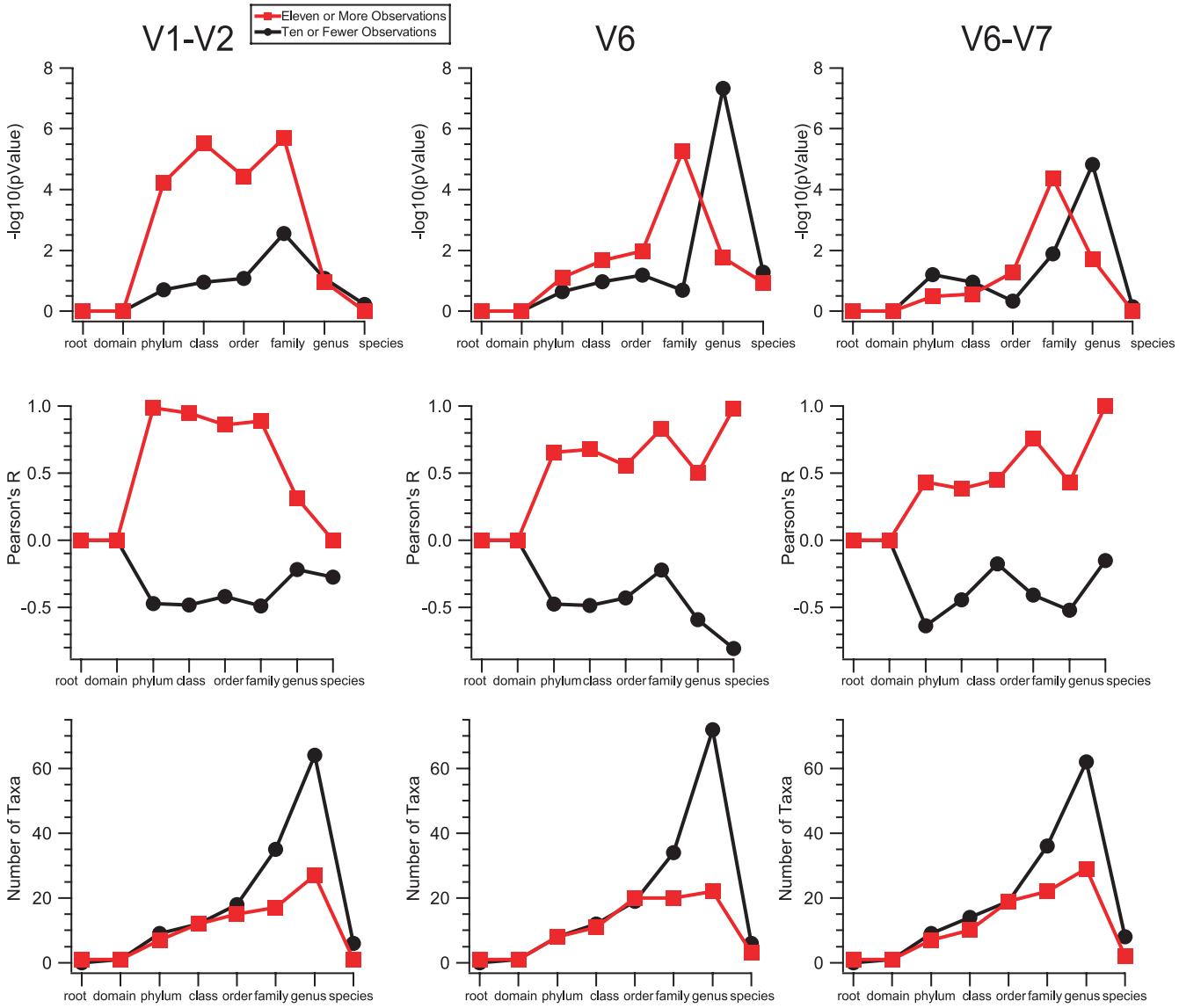

FIG. 3. Across taxonomic levels, the results of a linear regression on log-transformed data between sequences generated by PCR targeting the 16S rRNA gene and 16S sequences culled from our whole-genome shotgun sequence set. Assignments are by the RDP classification algorithm as in Fig. 2. For each sequence set, two separate regressions were constructed, one for the rare biosphere (circles) with taxa seen 10 or fewer times in that sequence set's PCRs and one for a common biosphere (squares) with taxa seen 11 or more times in the PCR reactions (Fig. 1, gray lines). The top panels show the −log10 of the *P* value of the null hypothesis that the slope of the regression equals zero. The middle panels show Pearson's *R* values while the bottom panel shows the number of taxa for which classifications are made. Note that a significant *P* value (top panel) can be produced by either a negative or positive correlation.

GAST, to compare pyrosequencing runs for the V3 and V6 regions (13). We implemented a broadly similar approach to GAST in Java that we call JGast (available in File S2 in the supplemental material). Figure 4 compares the JGast assignments for each taxon at the family level with the assignments taken from the RDP classification algorithm's direct assignment of taxonomy to the query sequence. We see generally reasonable agreement between the two approaches with, again, pronounced differences occurring in sequences that are observed fewer than 10 times. Figure 5 shows the results of regressions comparing JGast and RDP across taxonomic levels. Again, we see pronounced correlations when we consider a regression based only on taxa observed more than 11 times under either the RDP classification algorithm or JGast but generally very poor correlations when we consider taxa that are observed 10 or fewer times by both algorithms. We conclude that different classification approaches between different algorithms have a much more pronounced effect on the rare biosphere than on more commonly observed taxa.

DISCUSSION

Our comparison of 16S rRNA sequences derived from whole-genome sequence sets with PCR-generated 16S se-

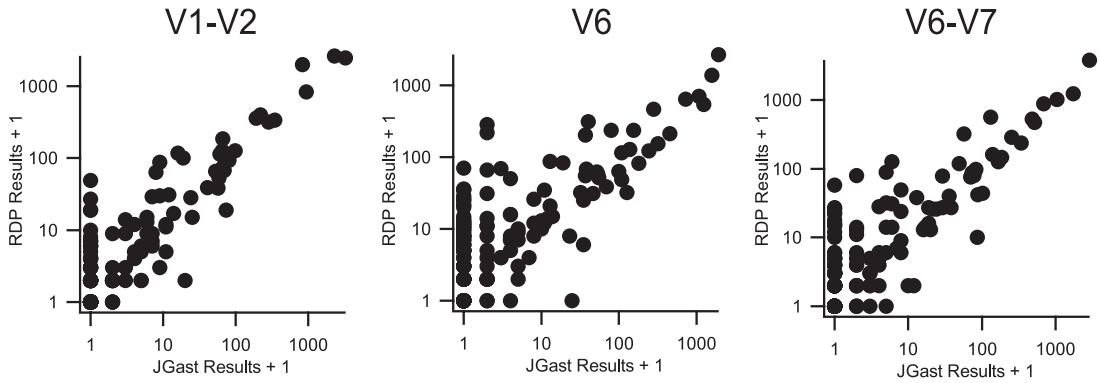

FIG. 4. Comparison of classifications at the family level made by JGast and the RDP classification algorithm.

quences yielded reasonable overall positive correlations (Fig. 2) but very poor to negative correlations for sequences seen fewer than 11 times (Fig. 3). Two of the causes of this poor correspondence are immediately obvious from inspection of Fig. 2. There are many low-abundance sequences detected by the PCR primers that are not found in our whole-genome sequence set. This reflects differences in the number of sequences generated; we have many thousands of 16S rRNA sequences in our PCR-based data sets but only a few hundred 16S rRNA sequences in our whole-genome set (despite their being over 350,000 sequences overall in this set). Presumably, more whole-genome sequencing would have detected more of these low-abundance taxa although this is a very inefficient way to discover 16S rRNA sequences. On the other hand, PCR bias causes the primer sequences to miss some of the taxa that are clearly present in the whole-genome sequence set, especially for the primers targeting V6 and V6-V7, which reported none or few sequences for taxa such as *Flexibacteraceae* that were easily detected in the V1-V2 and whole-genome sets (Fig. 2).

In addition to primer bias and differences in the number of 16S sequences as causes of the poor correspondence between whole-genome and PCR results for low-abundance taxa, we suggest a third, less immediately obvious, cause: low-abundance taxa are harder to classify, and hence results for these taxa are more sensitive to choices made during data analysis, what we here call "analysis noise" (Fig. 4 and 5). This assertion that analysis noise complicates descriptions of the rare biosphere while having less of an impact on more frequently observed taxa is consistent with previous observations that sequences that are infrequently observed are not as well represented in databases as more abundant taxa (7, 25) and hence will be more difficult to classify.

In a recent paper, Huse et al. performed a comparison of different primer sets (full-length, V3, and V6). Their regressions of the abundance of taxa produced by these different primer sets produced *r*-squared values down to genus of ~0.99 (13). These are substantially higher than the *r*-squared values we observed when we compared our whole-genome run to our runs targeting V1-V2, V6, and V7 (Fig. 2 and 3). We believe that these discrepancies may be explained by the following considerations. (i) The first is depth of sequencing coverage. Huse et al. generated 422,992 V3 tags, 441,894 V6 tags, and 7,215 full-length sequences. By contrast, our study consists of a much more modest number of tag sequences in the thousands and whole-genome 16S sequences in the hundreds (Table 1). Because there is more disagreement in the rare biosphere than in abundant taxa and because our regressions are on a log-log scale, we would expect a smaller *r*-squared value with less sequence coverage. That is, as the amount of sequencing goes up, the regression becomes more and more dominated by abundant taxa and therefore becomes tighter since primer bias primarily affects less abundant taxa. This effect is exacerbated by the fact that Huse et al. report their *r*-squared values based on an untransformed linear regression while we log transform our data before performing the regression. This increases the effect of the variability introduced by the rare biosphere in our analysis. (ii) Huse et al. generate their full-length sequences with PCR while we cull our 16S sequences from a whole-genome data set of 250 bp reads. This increases the analysis noise in our analysis as our taxonomic assignment of 250 bp fragments, randomly distributed throughout the 16S sequence, is certainly less accurate than the assignment of full-length sequences by Huse et al. (13). (iii) The data set of Huse et al. is from the human gut, which is likely a less complex ecosystem than our wastewater treatment plant. The additional complexity of our ecosystem likely increases the effect of the rare

TABLE 4. Number of sequences classified by JGast[a]

Region	No. of sequences assigned to:							
	Root	Kingdom	Phylum	Class	Order	Family	Genus	Species
V1-V2	11,316	11,316	11,137	10,968	10,104	9,513	6,907	424
V6	12,277	11,804	11,451	11,375	9,938	9,570	6,335	524
V6-V7	11,645	11,644	11,367	11,291	9,962	9,753	6,968	1,277

[a] Assignment of sequences required that there be at least 85% sequence identity between the query sequence and the reference sequence in the Greengenes file and that the nearest neighbor(s) be classified with at least an 80% RDP confidence at the given taxon.

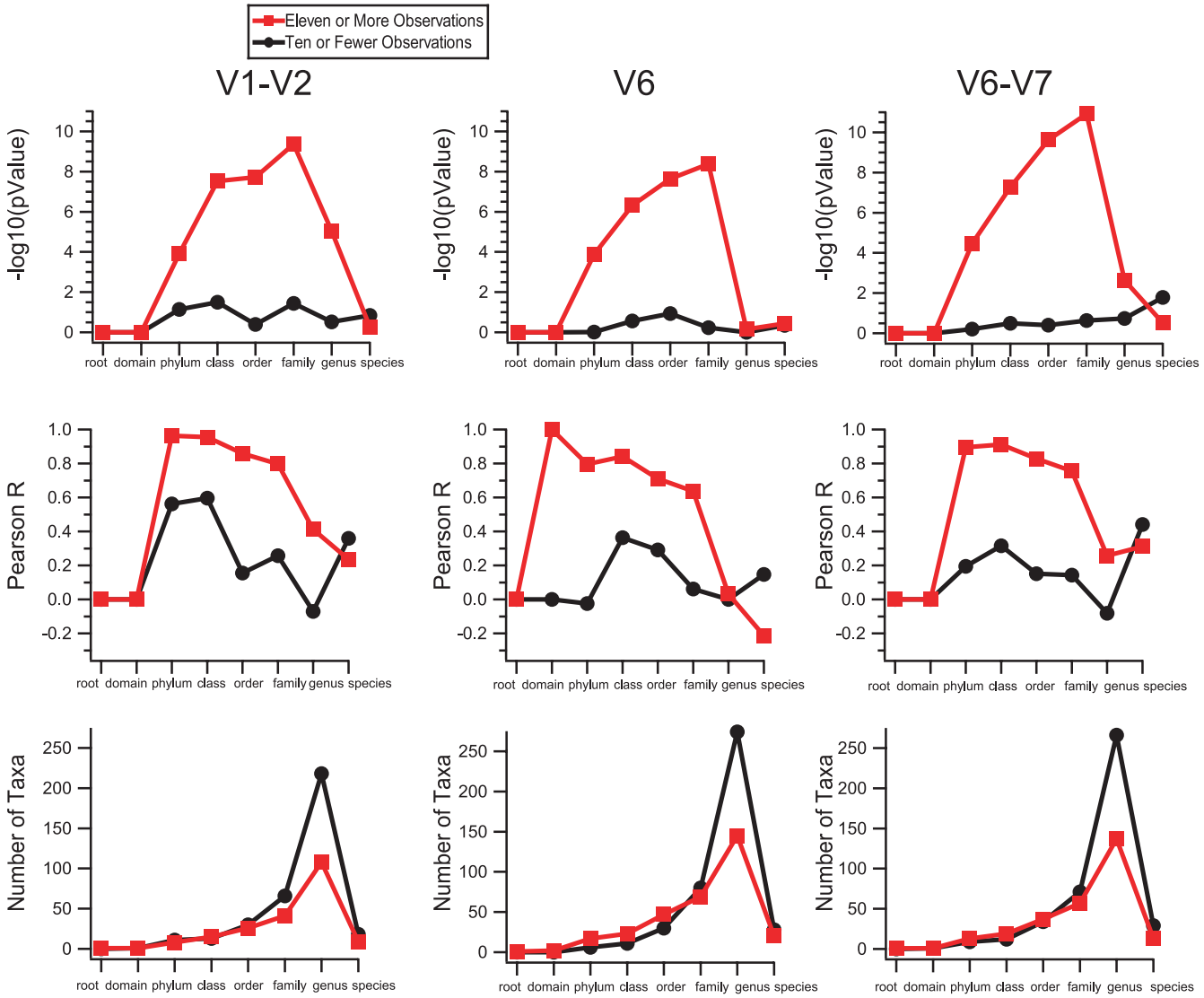

FIG. 5. Regressions across classification levels on log-transformed data showing the comparison between the RDP classification algorithm and JGast. Two regressions were constructed for each comparison: one for a common biosphere in which a taxon was observed 11 or more times under either JGast or RDP (squares) and one for a rare taxon in which fewer than 11 taxa were observed under both classification schemes.

biosphere, further suppressing the correspondence between our whole-genome and PCR sequence sets. (iv) Finally, PCR noise is a factor. PCR is an inherently stochastic process with many possible confounding effects including saturation of amplicon product. Nonquantitative PCR is not generally thought of as a quantitative technology. It would be surprising if the results of PCR runs on complex communities produced results that were reliably reproducible in quantitative detail. As the cost of sequencing continues to drop and as large pyrosequencing data sets become increasingly trivial to obtain, we anticipate that future studies will be able to discriminate the relative importance of these causes for discrepancies in whole-genome and PCR-based surveys of complex communities.

We note that there is not an immediately obvious superior choice between the two analysis schemes we examined. The RDP classification scheme offers simplicity, very fast run times, and a straightforward bootstrap mechanism for establishing significance. Algorithms like JGast (and GAST [13] and the Greengenes classification scheme [5]) offer potentially greater sensitivity but are also more susceptible to small changes in the database. For example, a single misclassified sequence inserted into the reference database can significantly change classifications for a large number of query sequences if it serves as the nearest neighbor. The RDP classification algorithm is less sensitive to small changes in the training set. We imagine that in the future as the read length of new sequencing technologies approaches the length of the 16S rRNA genes (10), nearest-neighbor approaches, which replace a short query sequence with a full-length database sequence, may be replaced by direct classification of full-length query sequences.

In the paper that introduced the RDP classification algorithm (26), an in silico analysis suggested that subsequences covering the V1-V2 region out-performed other variable regions in reproducing full-length classifications. Our results are

consistent with this finding, with the V1-V2 region slightly out-performing V6 and V6-V7 in matching our whole-genome results (Fig. 2 and 3). In addition, the V1-V2 region seemed slightly less susceptible to "algorithm noise." There are only three families which the RDP classification algorithm detects more than 10 times that are not detected by JGast (Fig. 4) while there are more such disagreements between the two approaches for V6 and V6-V7. These differences in performance between the primers sets are small, and it is reasonable to think that environments other than wastewater might yield different results. Nonetheless, all other things being equal, the V1-V2 region seems reasonable to target with the 250 bp available in the 454-FLX technology.

In their analysis of primer bias, Huse et al. conclude that "any effects of primer bias are limited to rare taxa." This analysis is largely in agreement with our results in which different primer sets produce somewhat similar results for abundant taxa while producing vast differences in the results generated for the rare biosphere. A principle argument for favoring pyrosequencing over traditional Sanger sequencing is that the much greater depth of pyrosequencing allows for investigations of the rare biosphere (25). Our results suggest that even more diversity in the rare biosphere will be uncovered when pyrosequencing experiments are repeated with different primer sets. Our results also suggest that longer sequence reads, which may be possible in sequencing technology to be introduced soon (10), will reduce analysis noise by eliminating the step in which a short query sequence is mapped to a full-length reference sequence. Such improvements in technology will be an important component of explorations of the rare biosphere while having much less impact on surveys of more abundant taxa.

ACKNOWLEDGMENT

We thank Mary Ann Moran for very helpful comments on a previous version of the manuscript.

REFERENCES

1. **Amann, R., W. Ludwig, and K. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143–169.
2. **Bond, P. L., P. Hugenholtz, J. Keller, and L. L. Blackall.** 1995. Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. Appl. Environ. Microbiol. **61:**1910–1916.
3. **Chakravorty, S., D. Helb, M. Burday, N. Connell, and D. Alland.** 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J. Microbiol. Methods **69:**330–339.
4. **Crump, B. C., G. W. Kling, M. Bahr, and J. E. Hobbie.** 2003. Bacterioplankton community shifts in an Arctic lake correlate with seasonal changes in organic matter source. Appl. Environ. Microbiol. **69:**2253–2268.
5. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
6. **DeSantis, T. Z., Jr., P. Hugenholtz, K. Keller, E. L. Brodie, N. Larsen, Y. M. Piceno, R. Phan, and G. L. Andersen.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. **34:**W394–W399.
7. **Dethlefsen, L., S. Huse, M. L. Sogin, and D. A. Relman.** 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. **6:**e280.
8. **Edgar, R. C.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113.
9. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
10. **Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner.** 2009. Real-time DNA sequencing from single polymerase molecules. Science **323:**133–138.
11. **Ginige, M. P., P. Hugenholtz, H. Daims, M. Wagner, J. Keller, and L. L. Blackall.** 2004. Use of stable-isotope probing, full-cycle rRNA analysis, and fluorescence in situ hybridization-microautoradiography to study a methanol-fed denitrifying microbial community. Appl. Environ. Microbiol. **70:**588–596.
12. **Huber, J. A., D. B. Mark Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97–100.
13. **Huse, S. M., L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin.** 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. **4:**e1000255.
14. **Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch.** 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. **8:**R143.
15. **Li, F., M. A. Hullar, and J. W. Lampe.** 2007. Optimization of terminal restriction fragment polymorphism (TRFLP) analysis of human gut microbiota. J. Microbiol. Methods **68:**303–311.
16. **Lyautey, E., B. Lacoste, L. Ten-Hage, J. L. Rols, and F. Garabetian.** 2005. Analysis of bacterial diversity in river biofilms using 16S rDNA PCR-DGGE: methodological settings and fingerprints interpretation. Water Res. **39:**380–388.
17. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
18. **Pace, N., D. Stahl, D. Lane, and G. Olsen.** 1985. Analyzing natural microbial populations by rRNA sequences. ASM News **51:**4–12.
19. **Rappe, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57:**369–394.
20. **Rozen, S., and H. Skaletsky.** 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. **132:**365–386.
21. **Sanapareddy, N., T. J. Hamp, L. C. Gonzalez, H. A. Hilger, A. A. Fodor, and S. M. Clinton.** 2009. Molecular diversity of a North Carolina wastewater treatment plant as revealed by pyrosequencing. Appl. Environ. Microbiol. **75:**1688–1696.
22. **Sekiguchi, H., M. Watanabe, T. Nakahara, B. Xu, and H. Uchiyama.** 2002. Succession of bacterial community structure along the Changjiang River determined by denaturing gradient gel electrophoresis and clone library analysis. Appl. Environ. Microbiol. **68:**5142–5150.
23. **Shenkin, P. S., B. Erman, and L. D. Mastrandrea.** 1991. Information-theoretical entropy as a measure of sequence variability. Proteins **11:**297–313.
24. **Snaidr, J., R. Amann, I. Huber, W. Ludwig, and K. H. Schleifer.** 1997. Phylogenetic analysis and in situ identification of bacteria in activated sludge. Appl. Environ. Microbiol. **63:**2884–2896.
25. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. USA **103:**12115–12120.
26. **Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. **73:**5261–5267.