Journal of Microbiological Methods 84 (2011) 81–87

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes

Minseok Kim [a], Mark Morrison [a,b], Zhongtang Yu [a,*]

[a] Department of Animal Sciences, The Ohio State University, 2029 Fyffe Road, Columbus, OH 43210, USA
[b] CSIRO Livestock Industries, St. Lucia, Australia

ARTICLE INFO

Article history:
Received 2 September 2010
Received in revised form 25 October 2010
Accepted 25 October 2010
Available online 31 October 2010

Keywords:
Diversity
16S rRNA gene
OTUs
Partial sequences

ABSTRACT

Operational taxonomic units (OTUs) are conventionally defined at a phylogenetic distance (0.03—species, 0.05—genus, 0.10—family) based on full-length 16S rRNA gene sequences. However, partial sequences (700 bp or shorter) have been used in most studies. This discord may affect analysis of diversity and species richness because sequence divergence is not distributed evenly along the 16S rRNA gene. In this study, we compared a set each of bacterial and archaeal 16S rRNA gene sequences of nearly full length with multiple sets of different partial 16S rRNA gene sequences derived therefrom (approximately 440–700 bp), at conventional and alternative distance levels. Our objective was to identify partial sequence region(s) and distance level(s) that allow more accurate phylogenetic analysis of partial 16S rRNA genes. Our results showed that no partial sequence region could estimate OTU richness or define OTUs as reliably as nearly full-length genes. However, the V1–V4 regions can provide more accurate estimates than others. For analysis of archaea, we recommend the V1–V3 and the V4–V7 regions and clustering of species-level OTUs at 0.03 and 0.02 distances, respectively. For analysis of bacteria, the V1–V3 and the V1–V4 regions should be targeted, with species-level OTUs being clustered at 0.04 distance in both cases.

1. Introduction

The difficulty in culturing most microbes present in natural or managed environments forces microbiologists to use the 16S rRNA gene as a phylogenetic marker in examining microbial diversity and classifying microbes. Even though the scarcity of well-characterized microbes and the lack of a reliable prokaryotic taxonomy system often make it difficult to classify microbes to species or sub-species level with certainty solely based on 16S rRNA gene sequences, 16S rRNA gene sequences can provide more objective and reliable classification of microbes than phenotyping (Schloss and Handelsman, 2004). Since Lane et al. (1985) first described the use of 16S rRNA gene for identifying and classifying uncultured microbes in the environment, PCR amplification, cloning and sequencing have been the primary technologies used in determining 16S rRNA gene sequences from various environments. During the past two decades, more than 1.3 million bacterial and 54,000 archaeal 16S rRNA gene sequences have been archived in RDP (as of March 20, 2010, Release 10, Update 18) (Cole et al., 2009). These sequences are curated and include 16S rRNA genes recovered from both cultured and uncultured prokaryotes, with the latter accounting for most of the sequences. The 16S rRNA gene sequences in RDP have been classified into genera among 35 bacterial phyla and 5 archaeal phyla, but many of these phyla are composed primarily or entirely of uncultured prokaryotes (Schloss and Handelsman, 2004).

The 16S rRNA gene sequences generated from microbiomes are typically clustered into operation taxonomic units (OTUs) at a few distance levels to determine species richness, diversity, composition, and community structure. Species, genus, family, and phylum are conventionally defined with distance values of 0.03, 0.05, 0.10 and 0.20, respectively, based on full-length (approximately 1540 bp) 16S rRNA gene sequences (Schloss and Handelsman, 2004). However, 16S rRNA gene sequences produced in most studies are partial sequences of 700 bp or shorter due to cost restraint (with the Sanger DNA sequencing technology) or technology limitations (with the next generation DNA sequencing technologies). Indeed, in the RDP database less than 44% of the bacterial and 15.3% of the archaeal sequences are longer than 1200 bp. Only a very small percentage of the sequences in RDP reached nearly full length. Therefore, most researchers used partial 16S rRNA gene sequences to make taxonomic assignments. Such a discord may create uncertainty in taxonomic placement of OTUs because of the following reasons: first, divergence among different 16S rRNA gene sequences is not distributed evenly along the 16S rRNA gene but concentrated primarily in the nine hypervariable (V) regions (Stackebrandt and Goebel, 1994). Second, some of the V regions are more variable than others (Youssef et al., 2009; Yu and Morrison, 2004). Third, some regions of the 16S rRNA genes produce more reliable taxonomic assignments than others (Liu

* Corresponding author. Tel.: +1 614 292 3057; fax: +1 614 292 2929.
E-mail address: yu.226@osu.edu (Z. Yu).

0167-7012/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2010.10.020

et al, 2007, 2008; Wang et al., 2007). We hypothesize that different V regions may produce different results with respect to estimates on species richness, diversity, and microbiome composition and structure, and some partial sequence regions may be better suited for microbiome analysis than others. A different taxonomic cutoff value, or distance level, may be required for a particular partial sequence region to give rise to similar results as nearly full-length sequences.

Recently, a number of studies used 454 pyrosequencing in comprehensive analysis of species richness and diversity present in complex microbiomes (Claesson et al., 2009; Sogin et al., 2006; Krober et al., 2009; Youssef et al., 2009). These studies generated large numbers of partial 16S rRNA gene sequences. By necessity, these partial sequences were clustered into OTUs using the same conventional distance values that were used for nearly full-length sequences. In some of these studies, two single V regions are compared between them and also against a set of nearly full-length sequences in estimating OTU richness (Claesson et al., 2009; Dethlefsen et al., 2008; Huse et al., 2008). It is recognized that the choice of V regions significantly affects estimates on OTU richness and diversity. One study also showed that the V1–V2 (approximately 350 bp) and the V8 regions produced different OTU evenness when a termite sample was analyzed (Engelbrektson et al., 2010). Another study compared eight V regions, either singular or dual, but the length of the partial sequence regions only ranged from 99 to 361 bp (Youssef et al., 2009). More importantly, in these studies conclusions were drawn from comparing short partial sequences recovered from one or a few habitats. As such, the conclusions derived from these studies may not be applied to broad environments. As the read length of pyrosequencing continues to increase, longer partial sequences (up to 800 currently) of 16S rRNA genes can be sequenced. Thus, there is a critical need to identify suitable partial sequence regions and phylogenetic distance cutoff values that can provide reliable analysis of microbiome. In this study, we systematically compared all the partial sequence regions (approximately 450 to 700 bp) delineated by commonly used domain-specific (bacterial or archaeal) PCR primers against nearly full-length 16S rRNA gene sequences archived in RDP that represent a broad taxonomy of both cultured bacteria and archaea. The comparisons were focused on observed OTU richness, parametric and nonparametric estimates of maximum OTU richness, accuracy of OTU clustering, and community structure. The objective was to identify a partial region(s) of 16S rRNA gene and a distance cutoff value(s) that enable analysis of 454 pyrosequencing reads and produce comparable results as nearly full-length sequences.

2. Materials and methods

2.1. Sequence collection, alignment, and clipping

All the sequences longer than 1200 bp were retrieved from the RDP database (Release 10, Update 18) in March 2010. All these sequences are of good quality as determined by RDP. For domain Bacteria, only the sequences recovered from type strains were selected, while for domain Archaea, sequences derived from both type and non-type strains were chosen because only a small number of archaeal type strains are archived in RDP. The bacterial and archaeal sequences were downloaded separately. The sequences that do not have nearly full-length (<1424 bp), as determined by the absence of the annealing sites of the domain-specific PCR primer pairs that anneal near the termini of 16S rRNA genes (A2Fa and U1510r for archaea, 27f and 1492r for bacteria), were removed manually from the datasets. The nearly full-length sequences were aligned using the NAST aligner according to a core set of alignment templates in the Greengenes database (DeSantis et al., 2006). Partial sequence regions that were delineated by the binding sites of domain-specific primer pairs (Tables 1 and 3) targeting different hypervariable regions (Baker et al., 2003; Yu and Morrison, 2004; Yu et al., 2008) were clipped out from the alignment of the nearly full-length sequences using the Geneious program (Biomatters Ltd., Auckland, New Zealand), with the original alignment being retained. The alignment of each partial sequence region and the full-length sequences was analyzed as a separate 'clone library'.

2.2. Diversity estimates

From the alignment of the nearly full-length sequences and each partial sequence region, a distance matrix at distances of 0.03, 0.05, and 0.10 was computed using the DNADIST program of the PHYLIP package (version 3.69, http://evolution.genetics.washington.edu/phylip.html) with the Jukes–Cantor correction applied. A distance matrix was also computed at 0.01 distance, which was suggested to be a new taxonomic cutoff value for species (Stackebrandt and Ebers, 2006). The DOTUR program (Schloss and Handelsman, 2005) was

Table 1
Estimates of species-level OTUs calculated from partial and full-length archaeal 16S rRNA gene sequences.

Primer set	V regions	Sequence length (bp)*	Distance level	# of OTUs ±	# of identical OTUs (%)‡	Maximum # of OTUs ±		
						Rarefaction	Chao1	ACE
A2Fa–U1510r	V1–V9	1435	0.03	363 (0.0)	363 (100)	462 (0.0)	590 (0.0)	660 (0.0)
A2Fa–519r	V1–V3	467	0.03	383 (5.5)	252 (65.8)	494 (6.9)	684 (15.9)	758 (14.8)
A2Fa–A693r	V1–V4	639	0.03	388 (6.9)	269 (69.3)	499 (8.0)	680 (15.3)	757 (14.7)
ARC344f–ARC915r	V3–V5	553	0.03	331 (−8.8)	231 (69.8)	401 (−13.2)	**567 (−3.9)**	581 (−12.0)
			0.02	384 (5.8)	256 (66.7)	495 (7.1)	696 (18.0)	726 (10.0)
U519f–UA1204r	V4–V7	702	0.03	311 (−14.3)	**235 (75.6)**	377 (−18.4)	513 (−13.1)	554 (−16.1)
			0.02	**367 (1.1)**	265 (72.2)	**473 (2.4)**	682 (15.6)	733 (11.1)
A679r§–UA1204r	V5–V7	527	0.03	279 (−23.1)	189 (67.7)	331 (−28.4)	450 (−23.7)	480 (−27.3)
			0.02	342 (−5.8)	231 (67.5)	432 (−6.5)	627 (6.3)	**663 (0.5)**
ARCH915–U1510r	V6–V9	585	0.03	307 (−15.4)	227 (73.9)	365 (−21.0)	465 (−21.2)	498 (−24.5)
			0.02	378 (4.1)	251 (66.4)	477 (3.2)	646 (9.5)	671 (1.7)
A1040f–U1510r	V7–V9	463	0.03	328 (−9.6)	231 (70.4)	394 (−14.7)	520 (−11.9)	555 (−15.9)
			0.02	377 (3.9)	243 (64.5)	478 (3.5)	660 (11.9)	696 (5.5)

The estimates for nearly full-length sequences and partial sequence regions at 0.03 are listed. For some partial sequence regions, the estimates at 0.02 or 0.04 are also listed when better estimates were obtained (same as in Table 3).

* Calculated from consensus sequences (same as in other tables).

± Values in parenthesis show the estimates relative to that of full-length sequences. Positive values designate overestimates, and negative values underestimates. The values that are both underlined and bolded are the best estimates, while the values that are only underlined are the second best estimates (same as in other tables).

‡ Number of OTUs that contain the same sequences as the corresponding OTUs clustered from the nearly full-length sequences. The values in parenthesis represent accuracy (%) of OTU clustering (same as in other tables).

§ The reverse complementary of primer A693r reported previously (17). Same as in Table 2.

Table 2
Estimates of genus- and family-level OTUs calculated from partial and full-length archaeal 16S rRNA gene sequences.

Primer set	V regions	Sequence length (bp)*	Distance level	# of OTUs ±	Maximum # of OTUs ±		
					Rarefaction	Chao1	ACE
A2Fa–U1510r	V1–V9	1435	0.05	273 (0.0)	322 (0.0)	469 (0.0)	462 (0.0)
A2Fa–519r	V1–V3	467		303 (11.0)	355 (10.2)	455 (−3.0)	495 (7.1)
A2Fa–A693r	V1–V4	639		**300 (9.9)**	**350 (8.7)**	**457 (−2.6)**	**477 (3.2)**
ARC344f–ARC915r	V3–V5	553		246 (−9.9)	278 (−13.7)	374 (−20.3)	392 (−15.2)
U519f–UA1204r	V4–V7	702		234 (−14.3)	268 (−16.8)	403 (−14.1)	399 (−13.6)
A679r§–UA1204r	V5–V7	527		206 (−24.5)	227 (−29.5)	321 (−31.6)	325 (−29.7)
ARCH915–U1510r	V6–V9	585		231 (−15.4)	255 (−20.8)	378 (−19.4)	368 (−20.3)
A1040f–U1510r	V7–V9	463		244 (−10.6)	271 (−15.8)	372 (−20.7)	366 (−20.8)
A2Fa–U1510r	V1–V9	1435	0.10	137 (0.0)	142 (0.0)	210 (0.0)	204 (0.0)
A2Fa–519r	V1–V3	467		168 (22.6)	174 (22.5)	265 (26.2)	238 (16.7)
A2Fa–A693r	V1–V4	639		165 (20.4)	172 (21.1)	235 (11.9)	**235 (15.2)**
ARC344f–ARC915r	V3–V5	553		**129 (−5.8)**	**130 (−8.5)**	193 (−8.1)	171 (−16.2)
U519f–UA1204r	V4–V7	702		122 (−10.9)	124 (−12.7)	**194 (−7.6)**	172 (−15.7)
A679r§–UA1204r	V5–V7	527		105 (−23.4)	107 (−24.6)	145 (−31.0)	143 (−29.9)
ARCH915–U1510r	V6–V9	585		115 (−16.1)	115 (−19.0)	150 (−28.6)	148 (−27.5)
A1040f–U1510r	V7–V9	463		123 (−10.2)	125 (−12.0)	169 (−19.5)	167 (−18.1)

used to cluster the sequences into OTUs (referred to as 'observed' OTUs) and determine the maximum number of OTUs represented by each 'clone library' using nonparametric Chao1 and ACE richness estimates. From the rarefaction output calculated by the DOTUR program, a parametric estimate of the maximum number of OTUs in each 'clone library' was also performed using the non-linear models procedure (PROC NLIN) of SAS (V9.1, SAS Inst. Inc., Cary, NC) as described previously (Larue et al., 2005). The number of OTUs defined by each partial sequence 'clone library' (referred to as observed OTU richness) and the maximum number of OTUs predicted from each partial sequence 'clone library' (referred to as maximum OTU richness) were compared to those defined by the corresponding nearly full-length sequence 'clone library'.

To identify a distance cutoff value that produces a better estimate on species-level OTUs than the commonly used 0.03 distance, all the partial sequence regions were also analyzed at 0.02 and 0.04 distances. Distance matrices at 0.02 and 0.04 were computed as described above. Each of the distance matrices was then used in clustering OTUs and estimating observed and maximum OTU richness as described above. The sequence composition was compared between respective OTUs defined from the nearly full-length 'clone library' and from each of the partial sequence 'clone library'. The 'accuracy' of OTU clustering was introduced as a percentage of partial sequence-based OTUs that have identical sequence composition as the OTUs defined by the nearly full-length sequences.

2.3. UniFrac analysis

The UniFrac program (Lozupone and Knight, 2005) was used to assess differences in 'clone libraries' represented by individual partial sequence datasets and the nearly full-length sequence dataset. The sequences from all the 'libraries' were aligned against the Greengenes database and then inserted into the ARB tree to build phylogenetic trees. The constructed trees were subjected to UniFrac significant test and P test.

2.4. Analysis of sequence datasets recovered from uncultured bacteria

To verify their applicability, one sequence dataset each recovered from rumen (Brulc et al., 2009) and deep-sea surface sediment (Schauer et al., 2010) was also analyzed as described for the composite RDP sequence datasets of bacterial strains. Each of the two sequence datasets was retrieved from the RDP database, aligned, and analyzed as mentioned above for the RDP sequence datasets. Because no individual studies reported large numbers of full-length archaeal sequences, this verification was not done for archaea.

2.5. Analysis of short partial sequence regions

Short partial sequences spanning 1 or 2 consecutive V regions (94–362 bp) were also evaluated using the composite RDP sequence

Table 3
Estimates of species-level OTUs calculated from partial and full-length bacterial 16S rRNA gene sequences.

Primer set	V regions	Sequence length (bp)*	Distance level	# of OTUs ±	# of identical OTUs (%) ‡	Maximum # of OTUs±		
						Rarefaction	Chao1	ACE
27f–1492r	V1–V9	1458	0.03	555 (0.0)	555 (100)	1105 (0.0)	1600 (0.0)	1809 (0.0)
27f–519r	V1–V3	484	0.03	602 (8.5)	459 (76.2)	1329 (20.0)	2126 (33.0)	2255 (24.7)
			0.04	**556 (0.2)**	459 (82.6)	1080 (−2.3)	1807 (12.9)	1859 (2.8)
27f–685r	V1–V4	652	0.03	603 (8.6)	478 (79.3)	1368 (23.8)	2085 (30.3)	2385 (31.8)
			0.04	543 (−2.2)	**469 (86.4)**	1015 (−8.1)	**1541 (−3.7)**	1681 (−7.1)
63f–519r	V1–V3	446	0.03	612 (10.3)	459 (75.0)	1372 (24.2)	2019 (26.2)	2235 (23.5)
			0.04	563 (1.4)	456 (81.0)	**1112 (0.6)**	1875 (17.2)	1950 (7.8)
63f–685r	V1–V4	614	0.03	606 (9.2)	482 (79.5)	1375 (24.4)	2133 (33.3)	2420 (33.8)
			0.04	557 (0.4)	472 (84.7)	1088 (−1.5)	1662 (3.9)	**1784 (−1.4)**
357f–907r	V3–V5	563	0.03	488 (−12.1)	402 (82.4)	823 (−25.5)	1244 (−22.0)	1313 (−27.4)
			0.02	552 (−0.5)	459 (83.2)	1093 (−1.1)	1670 (4.4)	1775 (−1.9)
533f–1100r	V4–V6	597	0.03	499 (−10.1)	408 (81.8)	866 (−21.6)	1335 (−17.0)	1396 (−22.8)
			0.02	567 (2.2)	468 (82.5)	1169 (5.8)	1813 (13.3)	2040 (12.8)
926f–1492r	V6–V9	605	0.03	505 (−9.0)	416 (82.4)	883 (−20.1)	1493 (−7.0)	1441 (−20.3)
			0.02	573 (3.2)	461 (80.5)	1219 (10.3)	1855 (15.9)	2072 (14.5)
968f–1492r	V6–V9	544	0.03	522 (−5.9)	432 (82.8)	952 (−13.8)	1554 (−3.0)	1551 (−14.3)
			0.02	587 (5.8)	462 (78.7)	1286 (16.4)	1946 (21.6)	2250 (24.4)

dataset for bacteria. Due to lack of primers that anneal to individual V regions of archaeal sequences, this evaluation was not done for the composite RDP sequences of archaea. The Illumina GAIIx system, which produces short reads of about 100 bp, has been used in microbiome analysis in a recent study (Caporaso et al., in press). To evaluate the reliability of such short sequence reads in estimating richness and diversity, both the 100 bp regions downstream of primer F515 and upstream of primer R806 (primers F515 and R806 were used in the study by Caporaso et al., in press) were also analyzed and compared to the nearly full-length sequences in the bacterial RDP dataset as described above.

3. Results

From the RDP database (Release 10 Update 18), 7450 bacterial sequences longer than 1200 bp were found that were derived from type strains. Of these sequences, 887 have a length of ≥1458 bp and contain the annealing sites of universal or domain-specific primers near both ends of 16S rRNA gene. These nearly full-length sequences represent a broad taxonomic spectrum, including 18 phyla, 25 class, 64 order, 165 family, 361 genera, and 4 unclassified groups above genus level. In total, 8375 archaeal sequences were found longer than 1200 bp. Among them, only 284 were derived from type strains and have nearly full length. To increase the taxonomic representation, the nearly full-length sequences derived from non-type strains of archaea were also included. The archaeal sequence dataset used in this study contained 1071 nearly full-length sequences (≥1435 bp), which represent all the 4 archaeal phyla, 9 class, 14 order, 25 family, 73 genera, and 14 unclassified groups above the phylum, class, order, family or genus level. We intentionally selected and used sequences from a broad taxonomic spectrum so that the results derived can be less biased toward taxa found in particular habitats and the conclusions drawn can be better applied to broad environments. The lengths of the partial sequence regions ranged from 463 to 702 bp for archaea and from 446 to 652 bp for bacteria. Each of the partial sequence regions was compared to the corresponding nearly full-length sequence dataset with respect to observed OTU richness, maximum OTU richness, community structure, and accuracy of OTU clustering.

3.1. Analysis of partial archaeal sequences

The different partial sequence 'clone libraries' (partial sequence regions) of the archaeal sequence dataset gave rise to varying observed OTU richness at 0.03 distance, and all the estimates differed from that computed from the full-length archaeal sequences (Table 1). The V1–V3 region resulted in the best estimates of both observed OTU richness (5.5% overestimate) and the rarefaction-predicted (6.9% overestimate) maximum OTU richness, followed by the V1–V4 region. For both the Chao1 and ACE estimates, the V3–V5 region afforded the best predictions, which were lower than (3.9 and 12%, respectively) that estimated from the nearly full-length sequences. The partial sequence region that generated the second best estimate at species level was the V7–V9 region for Chao1 estimate and the V1–V4 region for ACE estimate. Overall, the V1–V4 region yielded overestimates, while the downstream regions underestimated all the richness estimates.

None of the partial sequence regions faithfully recaptured all the OTUs that were defined by the nearly full-length archaeal sequences at any of the tested distances. At 0.04 distance, all the partial sequence regions produced worse estimates on both the observed and the maximum OTU richness measurements than at 0.03 distance. The V1–V3 and the V1–V4 regions rendered worse estimates on OTU richness at 0.02 distance than at 0.03 distance, but the downstream partial sequence regions gave rise to better estimates on both observed and maximum richness (Table 1). The only exception was the Chao1 estimate from the V3–V5 region. The V4–V7 region resulted in the best estimates of observed OTU richness and rarefaction-predicted maximum OTU richness at 0.02 distance, whereas the V3–V5 and the V5–V7 regions resulted in better Chao1 and ACE estimates at 0.03 and 0.02 distances, respectively. At 0.02 distance, the V4–V7 region produced the highest accuracy (72.2%) of OTU clustering among all the partial sequence regions, but OTU clustering based on this partial sequence region was 3.4% more accurate at 0.03 distance. At 0.02 distance, the partial sequence regions downstream of V1–V4 produced more accurate estimates on OTU richness. These results reflect the greater sequence divergence of the V1–V4 region than the downstream regions (Yu et al., 2008).

At 0.01 distance, a new distance cutoff value recommended by Stackebrandt and Ebers (2006) to define prokaryotic species, the V1–V4 and the V1–V3 were the first and second best partial sequence regions with respect to estimates of observed OTU richness and maximum OTU richness irrespective of prediction methods used (Suppl. Table 1).

At 0.05 distance (equivalent to genus), the V1–V4 region produced the best estimates on both observed and maximum OTU richness (Table 2). The V1–V3 region yielded the second best estimate of maximum OTU richness, though the observed OTU richness was overestimated. At 0.10 distance (equivalent to family), the V3–V5 region produced the best estimates on both the observed OTU richness and the rarefaction-predicted maximum OTU richness. For Chao1 estimate, the V4–V7 region gave rise to a more accurate estimate than other partial sequence regions, followed by the V3–V5 region. The V1–V4 region supported the best ACE estimate, followed by the V4–V7 region. It is also noted that the V1–V4 region (1–639 bp) tended to overestimate both observed and maximum OTU richness at 0.05 and 0.10 distances, while the downstream regions considerably underestimated all the estimates of OTU richness (Table 2).

3.2. Analysis of partial bacterial sequences

The estimates of OTU richness of the bacterial sequence dataset differed also among the different partial sequence regions analyzed (Table 3). The observed species-level OTU richness estimated from the V6–V9 region delineated by primers 968f–1492r was the closest (5.9% underestimate) to that calculated from the full-length sequences, followed by the V1–V3 region delineated by primers 27f–519r. For the rarefaction estimate of maximum species-level OTU richness, the V6–V9 region delineated by primers 968f–1492r is the best region (13.8% underestimate), followed by the V1–V3 region delineated by primers 27f–519r. The V6–V9 region delineated by primers 968f–1492r also generated the best Chao1 (3.0% underestimate) and ACE (14.3% underestimate) estimates, which was followed by the V6–V9 region delineated by primers 926f–1492r. As in the case of partial archaeal sequences, the upstream regions (V1–V4) consistently overestimated OTU richness, while the downstream regions underestimated OTU richness.

When different partial sequence regions were evaluated in clustering species-level OTUs at 0.02 and 0.04 distance levels, no partial sequence region completely recaptured the OTU estimates defined by the nearly full-length sequences either. However, the upstream regions (i.e. V1–V3 and V1–V4) produced more accurate estimates of observed and maximum OTU richness at 0.04 than at 0.03 distances, whereas the downstream partial sequence regions improved estimates at 0.02 distance, with a few exceptions (Table 3). The three estimates on maximum OTU richness based on the V6–V9 region delineated by primers 968f–1492r and the Chao1 estimate from the V6–V9 region delineated by primers 926f–1492r were more accurate at 0.03 than 0.02 distance. The V1–V3 region delineated by primers 27f–519r afforded nearly the same observed OTU richness at 0.04 distance as the nearly full-length sequences did at 0.03 distance. The V1–V4 region delineated by primers 63f–685r also generated a very close estimate of observed OTU richness at 0.04 distance. At this same distance, the V1–V3 region delineated by primers 63f–519r, the V1–V4 region delineated by primers

27f–685r, and the V1–V4 region delineated by primers 63f–685r also supported better estimate on maximum OTU richness by rarefaction, Chao1 and ACE, respectively, than other partial sequence regions. The V1–V4 region delineated by primers 27f–685r, however, produced the best accuracy (86.4%) of OTU clustering at 0.04 distance, while also producing rather accurate estimate (2.2% underestimate) on observed OTU richness.

At 0.01 distance, the V1–V4 region delineated by primers 27f–685r produced the best estimate on observed OTU richness, while the V1–V4 region delineated by primers 63f–685r generated the second best estimate (Suppl. Table 2). For rarefaction estimate of maximum OTU richness, the V6–V9 region delineated by primers 968f–1492r was the best region, while the V1–V4 region delineated by primers 27f–685r was the second best choice. For both Chao1 and ACE estimates, the V1–V4 region delineated by primers 63f–685r was the best region, while the V1–V4 region delineated by primers 27f–685r generated the second best estimate. When the rumen sequence dataset (Brulc et al., 2009) was analyzed at 0.01 distance, the V1–V4 region delineated by primers 27f–685r and 63f–685r also produced the first and second best estimates on observed and maximum OTU richness, respectively (data not shown). Therefore, the V1–V4 region is probably the best region for species richness and diversity estimates at 0.01 distance.

At genus level, the V6–V9 region delineated by primers 968f–1492r allowed the best estimates on both observed OTU richness and maximum OTU richness (Table 4). However, all these predictions were underestimated. The V1–V4 region delineated by primers 27f–685r yielded the second closest estimates of observed OTU richness and the maximum OTU richness that was predicted by rarefaction and Chao1, while the V3–V5 region delineated by primers 357f–907r resulted in the second best ACE estimate of maximum OTU richness. At family level, the V6–V9 region delineated by primers 968f–1492r again generated the closest estimates, all of which were underestimated (Table 4). The V3–V5 region delineated by primers 357f–907r produced the second closest estimates on both observed OTU richness and maximum OTU richness. Again, the upstream regions (V1–V4) overestimated OTU richness, while the downstream regions underestimated OTU richness (Table 4). These results corroborate the greater bacterial sequence divergence of the V1–V4 region than the downstream regions (Yu and Morrison, 2004) and are in general agreement with finding of Youssef et al. (2009).

3.3. UniFrac analysis

Both the UniFrac significance test and the P test showed that none of the 'microbiomes' represented by individual partial sequence regions was significantly different ($P \geq 0.25$) than that represented by the full-length sequences. These results suggest that the locations of the partial sequence regions of the analyzed lengths might not significantly affect comparison of microbiomes using UniFrac. This conclusion is consistent with the finding of several previous studies where several shorter partial sequence regions were analyzed (Huse et al., 2008; Liu et al., 2007; Wang et al., 2007). Therefore, any of the partial sequence regions analyzed in this study can depict a comparable microbiome structure as the full-length sequences.

3.4. Analysis of uncultured bacterial sequences

The rumen sequence dataset contained 1388 sequences (≥1438 bp) of rumen origin (Brulc et al., 2009) that represented 9 bacterial phyla but *Firmicutes*, *Bacteroidetes*, and *Proteobacteria* predominated. Various partial sequence regions were compared to their corresponding nearly full-length sequences with respect to estimates on OTU richness (data not shown). For observed OTU richness, the V1–V4 region delineated by primers 63f–685r produced a better estimate (0.3% overestimate) at 0.04 distance than at other distances, whereas the V1–V4 region delineated by primers 27f–685r gave rise to the second best estimate (2.4% underestimate) also at 0.04 distance. The V1–V4 region delineated by primers 27f–685r and 0.04 distance also provided the best Chao1 (1.8% overestimate) and ACE (0.3% overestimate) estimate and the second best rarefaction estimate (1.7% underestimate vs. 1.5% underestimate from the V4–V6 region). All other distances and partial regions produced worse estimates on either observed OTU richness or maximum OTU richness. When the bacterial sequence dataset (634 sequences of ≥1421 bp) recovered from deep-sea surface sediments (Schauer et al., 2010) was analyzed (data not shown), the V1–V4 region delineated by primers 27f–685r and 0.04 distance also produced the best estimate on OTU richness (0.3% overestimate) and the second best rarefaction estimate of maximum OTU richness (2.3% overestimate vs. 1.4% overestimate from the V6–V9 region delineated by primers 968f–1492r). Although not the best combination, this V1–V4 region and 0.04 distance only led to 7.6% and 6.5% overestimate for Chao1 or ACE estimates, respectively. Evidently, the V1–V4 region and 0.04 distance can provide accurate estimate on OTU richness from these two sets of uncultured bacterial sequences.

3.5. Analysis of short partial sequence regions

Short partial sequence regions delineated by primers used in a previous study (Youssef et al., 2009) were analyzed to evaluate their utility in estimate of species richness. These short sequences (94–

Table 4
Estimates of genus- and family-level OTUs calculated from partial sequence regions and full length of bacterial 16S rRNA gene sequences.

Primer set	Hypervariable regions	Sequence length (bp)*	Distance level	# of OTUs ±	Maximum # of OTUs ±		
					Rarefaction	Chao1	ACE
27f–1492r	V1–V9	1458	0.05	444 (0.0)	688 (0.0)	1087 (0.0)	1061 (0.0)
27f–519r	V1–V3	484		512 (15.3)	909 (32.1)	1452 (34.0)	1518 (43.1)
27f–685r	V1–V4	652		496 (11.7)	846 (23.0)	1282 (17.9)	1367 (28.8)
63f–519r	V1–V3	446		522 (17.6)	928 (34.9)	1434 (31.9)	1588 (49.7)
63f–685r	V1–V4	614		507 (14.2)	888 (29.1)	1399 (28.7)	1468 (38.4)
357f–907r	V3–V5	563		371 (−16.4)	521 (−24.3)	863 (−21.0)	831 (−21.7)
533f–1100r	V4–V6	597		384 (−13.5)	527 (−23.4)	749 (−31.0)	790 (−25.5)
926f–1492r	V6–V9	605		381 (−14.2)	525 (−23.7)	812 (−25.0)	818 (−22.9)
968f–1492r	V6–V9	544		**415 (−6.5)**	**608 (−11.6)**	**1007 (−7.0)**	**964 (−9.1)**
27f–1492r	V1–V9	1458	0.10	244 (0.0)	291 (0.0)	490 (0.0)	460 (0.0)
27f–519r	V1–V3	484		321 (31.6)	424 (45.7)	679 (39.0)	679 (47.6)
27f–685r	V1–V4	652		311 (27.5)	398 (36.8)	690 (40.8)	645 (40.2)
63f–519r	V1–V3	446		345 (41.4)	471 (61.9)	732 (49.4)	749 (62.8)
63f–685r	V1–V4	614		324 (32.8)	423 (45.4)	766 (56.3)	689 (49.8)
357f–907r	V3–V5	563		208 (−14.8)	239 (−17.9)	346 (−29.0)	359 (−22.0)
533f–1100r	V4–V6	597		194 (−20.5)	211 (−27.5)	295 (−40.0)	289 (−37.2)
926f–1492r	V6–V9	605		203 (−16.8)	220 (−24.4)	320 (−35.0)	307 (−33.3)
968f–1492r	V6–V9	544		**228 (−6.6)**	**255 (−12.4)**	**384 (−2.2)**	**360 (−21.7)**

362 bp) span 1 or 2 consecutive V regions (Suppl. Table 3). For observed OTU richness, the V6 region and 0.03 distance produced the best estimate, while the V1–V2 region gave rise to best rarefaction and ACE estimate and V7–V8 the best Chao1 estimate. Nevertheless, none of these short partial sequence region produced accurate estimate at any of the distances (0.01 to 0.05) examined. This also holds true for the rumen sequence dataset (data not shown). Because different V regions have been used in many different studies, β-diversity analysis using previously published datasets can be difficult.

The 100 bp region downstream of the forward primer F515 and the 100 bp region upstream of reverse primer R806, which were generated by the Illumina GAIIx system and used in analysis of several samples including human feces and soil, fresh water and freshwater sediments (Caporaso et al., in press), were compared to the nearly full-length sequences of RDP to assess the accuracy of such a short region in estimating species richness (data not shown). Except the 100 bp region downstream of F515 and 0.01 distance that produced an accurate estimate on observed OTU richness (3.7% underestimate), both short regions underestimated both observed OTU richness and maximum OTU richness by 13.5 to 66% at 0.01, 0.02, or 0.03 distances. Similar results were observed when the rumen sequence dataset (Brulc et al., 2009) was subjected to this analysis (data not shown). Thus, although the Illumina GAIIx can produce sequence reads more cost-effectively, the short sequences probably do not support accurate analysis of microbiomes.

4. Discussion

Defining the full diversity of microbiomes is essential for microbial ecologists to assess the functional significance of any bacterial or archaeal species or to determine if the major members have been accounted for in analysis of specific microbiomes. Pyrosequencing recently emerged as the enabling technology to comprehensively characterize complex microbiomes in natural environments (Gilbert et al., 2008), managed ecosystems (Krause et al., 2008; Liu et al., 2008; Zhang et al., 2009), or human and animal gut (Claesson et al., 2009; Dowd et al., 2008; Huse et al., 2008). It is difficult to assemble individual pyrosequencing reads into full-length 16S rRNA genes because of the conserved nature of this gene. Therefore, partial 16S rRNA gene sequences are directly used in microbiome analysis. Because different regions of the 16S rRNA gene have different divergence, the choice of partial sequence regions can significantly affect the analysis results (Engelbrektson et al., 2010; Liu et al., 2007; Youssef et al., 2009). Thus, it is important and useful to determine how a partial 16S rRNA gene sequence region can support characterization of microbiomes as 'reliably' as nearly full-length 16S rRNA genes.

Unlike other reported studies that compared single or dual V regions (94–360 bp), in this study we compared all partial sequence regions that span at least three consecutive V regions (463–702 bp for archaea, 446–652 bp for bacteria). All these partial sequence regions can be generated using domain-specific primer pairs so that they can be amplified and used in pyrosequencing analysis. In addition, instead of using uncultured bacterial sequences recovered from a particular habitat, we chose the sequences only recovered from cultured organisms (type strains for bacteria, both type and non-type strains for archaea). These sequences are better taxonomically characterized and are free of chimeric artifacts. Furthermore, because the sequences were not recovered from a particular habitat, the sequences used in this study represent a much broader taxonomy and diversity. As such, the results of this study might be applied to analysis of different microbiomes. To test this premise, two of the largest datasets of nearly full-length sequences were analyzed in parallel to verify if the best partial sequence region(s) and distance(s) can be applied to individual sequence datasets.

The 454 GS FLX systems is the primary technology used in most diversity studies of microbiomes (Droege and Hill, 2008). The previous 454 GS FLX system produces sequence reads about 250 bp, a length typically spanning single V regions of 16S rRNA gene. Such a length only allows for classification of 16S rRNA gene sequences to genus in RDP (Liu et al., 2008). Most studies reported so far pyrosequenced single V regions, and when compared, two different V regions typically produce different results (Claesson et al., 2009; Dethlefsen et al., 2008; Huse et al., 2008; Sogin et al., 2006; Youssef et al., 2009). Several of these studies also compared partial sequences to nearly full-length sequences in estimating OTU richness (Claesson et al., 2009; Huse et al., 2008; Youssef et al., 2009). However, few studies have assessed if partial sequences can be clustered into species-level OTUs as 'reliably' as nearly full-length sequences. Thus, this study is probably among the early studies in a continuum of research that lead to improved analysis of 16S rRNA gene sequences.

Different partial sequence regions produced different estimates of OTU richness, both observed and predicted maxima, at all the three conventional distance levels for either archaea or bacteria. However, none of the analyzed partial sequence regions of bacteria or archaea faithfully recaptured the richness estimates (observed or predicted) that were determined by the nearly full-length sequences at conventional 0.03, 0.05, or 0.10 distances. These results corroborate the finding of a recent study (Schloss, 2010). Therefore, estimates of OTU richness calculated from partial sequences should be interpreted with caution. As shown in this study, the V1–V4 region can provide improved estimates on species richness and accuracy of OTU clustering when clustered at 0.04 distance, while the downstream partial sequence regions need to be clustered at 0.02 distance (Tables 1 and 3).

As the 454 pyrosequencing technology is increasingly used in analysis of microbiomes and the sequence read length continues to increase, longer partial 16S rRNA gene sequences will be sequenced that will improve accuracy of microbiome analysis. If a common partial sequence region is targeted by different researchers, analysis results can be compared among laboratories. A common target region will also facilitate global analysis of microbial diversity in a particular type of environment of interest as well as β-diversity across multiple environments. As a beginning of this effort, for analysis of archaea we recommend the V1–V3 region to be targeted with species-level OTUs being clustered at 0.03 distance if the current FLX Titanium system is used, or the V4–V7 region to be targeted with species-level OTUs being clustered at 0.02 distance if the newest 454 FLX system that generates up to 800 bp sequence reads is used. For analysis of bacteria, the V1–V3 region should be targeted using the FLX Titanium system, while the V1–V4 should be targeted with the newest 454 FLX system, with species-level OTUs being clustered at 0.04 distance in both cases. Additionally, these partial sequence regions also provide better analysis of richness and diversity than other partial regions if 0.01 distance is used to define OTUs. It should be pointed out that if distance is set at thousandth or below partial sequence regions may produce similar richness estimates as nearly full-length sequence. However, it will be time consuming to compare millions of richness estimates (5×10^7 for archaea and 5×10^8 for bacteria sequences). In addition, most OTU clustering algorithms do not support thousandth distance.

The V1–V3 or the V1–V4 regions of bacterial 16S rRNA genes provide two additional advantages: first, the V1–V3 or the V1–V4 regions are more divergent and thus can provide more phylogenetic resolution than other regions. Greater resolution is especially important in analysis of microbiomes from specialized habitats, such as intestinal tract of animals and humans, the rumens, anaerobic digesters and biological wastewater treatment reactors, where great diversity exists at low taxa. Second, RDP and other databases stored more partial sequences that correspond to the V1–V4 region than the downstream regions. As such, partial sequences corresponding to this region will have more database sequences to compare to, greatly facilitating phylogenetic analysis.

Acknowledgements

This study was partially supported by an OARDC award (2010-007) to Z.Y.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.mimet.2010.10.020.

References

Baker, G.C., Smith, J.J., Cowan, D.A., 2003. Review and re-analysis of domain-specific 16S primers. J. Microbiol. Meth. 55, 541–555.

Brulc, J.M., Antonopoulos, D.A., Miller, M.E., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., Coutinho, P.M., Henrissat, B., Nelson, K.E., White, B.A., 2009. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc. Natl. Acad. Sci. USA 106, 1948–1953.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2010. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. USA, in press.

Claesson, M.J., O'Sullivan, O., Wang, Q., Nikkila, J., Marchesi, J.R., Smidt, H., de Vos, W.M., Ross, R.P., O'Toole, P.W., 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS ONE 4, e6669.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., Tiedje, J.M., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145.

DeSantis Jr., T.Z., Hugenholtz, P., Keller, K., Brodie, E.L., Larsen, N., Piceno, Y.M., Phan, R., Andersen, G.L., 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucl. Acids Res. 34, W394–W399.

Dethlefsen, L., Huse, S., Sogin, M.L., Relman, D.A., 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 6, e280.

Dowd, S.E., Callaway, T.R., Wolcott, R.D., Sun, Y., McKeehan, T., Hagevoort, R.G., Edrington, T.S., 2008. Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). BMC Microbiol. 8, 125.

Droege, M., Hill, B., 2008. The Genome Sequencer FLX™ System-Longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol. 136, 3–10.

Engelbrektson, A., Kunin, V., Wrighton, K.C., Zvenigorodsky, N., Chen, F., Ochman, H., Hugenholtz, P., 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. 4, 642–647.

Gilbert, J.A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., Joint, I., 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS ONE 3, e3042.

Huse, S.M., Dethlefsen, L., Huber, J.A., Mark Welch, D., Relman, D.A., Sogin, M.L., 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 4, e1000255.

Krause, L., Diaz, N.N., Edwards, R.A., Gartemann, K.H., Kromeke, H., Neuweger, H., Puhler, A., Runte, K.J., Schluter, A., Stoye, J., Szczepanowski, R., Tauch, A., Goesmann, A., 2008. Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. J. Biotechnol. 136, 91–101.

Krober, M., Bekel, T., Diaz, N.N., Goesmann, A., Jaenicke, S., Krause, L., Miller, D., Runte, K.J., Viehover, P., Puhler, A., Schluter, A., 2009. Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. J. Biotechnol. 142, 38–49.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., Pace, N.R., 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. USA. 82, 6955–6959.

Larue, R., Yu, Z., Parisi, V.A., Egan, A.R., Morrison, M., 2005. Novel microbial diversity adherent to plant biomass in the herbivore gastrointestinal tract, as revealed by ribosomal intergenic spacer analysis and *rrs* gene sequencing. Environ. Microbiol. 7, 530–543.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., Knight, R., 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 35, e120.

Liu, Z., DeSantis, T.Z., Andersen, G.L., Knight, R., 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36, e120.

Lozupone, C., Knight, R., 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71, 8228–8235.

Schauer, R., Bienhold, C., Ramette, A., Harder, J., 2010. Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. ISME J. 4, 159–170.

Schloss, P.D., 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. 6, e1000844.

Schloss, P.D., Handelsman, J., 2004. Status of the microbial census. Microbiol. Mol. Biol. Rev. 68, 686–691.

Schloss, P.D., Handelsman, J., 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl. Environ. Microbiol. 71, 1501–1506.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M., Herndl, G.J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA 103, 12115–12120.

Stackebrandt, E., Ebers, J., 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiol. Today 33, 152–155.

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846–849.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73, 5261–5267.

Youssef, N., Sheik, C.S., Krumholz, L.R., Najar, F.Z., Roe, B.A., Elshahed, M.S., 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. Appl. Environ. Microbiol. 75, 5227–5236.

Yu, Z., Morrison, M., 2004. Comparisons of different hypervariable regions of *rrs* genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. Appl. Environ. Microbiol. 70, 4800–4806.

Yu, Z., Garcia-Gonzalez, R., Schanbacher, F.L., Morrison, M., 2008. Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by archaea-specific PCR and denaturing gradient gel electrophoresis. Appl. Environ. Microbiol. 74, 889–893.

Zhang, H., Banaszak, J.E., Parameswaran, P., Alder, J., Krajmalnik-Brown, R., Rittmann, B.E., 2009. Focused-pulsed sludge pre-treatment increases the bacterial diversity and relative abundance of acetoclastic methanogens in a full-scale anaerobic digester. Water Res. 43, 4517–4526.