REVISED

Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers¹

Phillip R. Myer ^{a,*}, MinSeok Kim ^b, Harvey C. Freetly ^b, and Timothy P.L. Smith ^b

^a Department of Animal Science, University of Tennessee Institute of Agriculture, University of Tennessee, Knoxville, TN 37996

^b USDA-ARS, U.S. Meat Animal Research Center, Clay Center, NE 68933²

tim.smith@ars.usda.gov; mkim2276@korea.kr; harvey.freetly@ars.usda.gov

* Corresponding author - pmyer@utk.edu

¹Mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that may be suitable.

²USDA is an equal opportunity provider and employer

ABSTRACT

Next generation sequencing technologies have vastly changed the approach of sequencing of the 16S rRNA gene for studies in microbial ecology. Three distinct technologies are available for large-scale 16S sequencing. All three are subject to biases introduced by sequencing error rates, amplification primer selection, and read length, which can affect the apparent microbial community. In this study, we compared short read 16S rRNA variable regions, V1-V3, with that of near-full length 16S regions, V1-V8, using highly diverse steer rumen microbial communities, in order to examine the impact of technology selection on phylogenetic profiles. Short paired-end reads from the Illumina MiSeq platform were used to generate V1-V3 sequence, while long "circular consensus" reads from the Pacific Biosciences RSII instrument were used to generate V1-V8 data. The two platforms revealed similar microbial operational taxonomic units (OTUs), as well as similar species richness, Good's coverage, and Shannon diversity metrics. However, the V1-V8 amplified ruminal community resulted in significant increases in several orders of taxa, such as phyla Proteobacteria and Verrucomicrobia ($P < 0.05$). Taxonomic classification accuracy was also greater in the near full-length read. UniFrac distance matrices using jackknifed UPGMA clustering also noted differences between the communities. These data support the consensus that longer reads result in a finer phylogenetic resolution that may not be achieved by shorter 16S rRNA gene fragments. Our work on the cattle rumen bacterial community demonstrates that utilizing near full-length 16S reads may be useful in conducting a more thorough study, or for developing a niche-specific database to use in analyzing data from shorter read technologies when budgetary constraints preclude use of near-full length 16S sequencing.

KEYWORDS: 16S rRNA gene, MiSeq, OTUs, PacBio, Rumen Microbiome

1. INTRODUCTION

Next generation sequencing has taken a central role in studies of microbial ecology, especially with regard to culture-independent methods based on molecular phylogenies of the small-subunit ribosomal RNA gene (16S rRNA gene). The ability to relate trends at the species or genera level to host/environmental parameters using 16S profiling has proven powerful (Hamady et al., 2009). These amplicon-based studies depend on annealing of amplification primers to conserved regions that flank variable regions (V1-V9) of the 16S rRNA gene. However, bias associated with amplicon sequencing and variable region selection can confound the quantitative assessment of bacterial community dynamics, regardless of sequencing depth (Pinto, 2012). This issue has been and continues to be the focus of study, with the goal to minimize the disadvantages of PCR-based estimates of microbial diversity (Forney et al., 2014; von Wintzingerode et al., 1997). Primer choice also affects apparent phylogeny, and there is yet no consensus on which subset 16S rRNA variable region(s) to use for community analysis (Kim et al., 2011; Liu et al., 2007). As a gold standard (Vinje et al., 2015), full-length 16S rRNA genes (about 1500 bp) can be used for accurate taxonomic identification, although cost constraints usually lead to targeting of one or a few variable regions. It is well established that the choice of variable region can affect the taxonomic classification, OTU (operational taxonomic unit) richness, and OTU diversity (Huse et al., 2008; Claesson et al., 2009). These studies make it clear that use of full-length sequence, or preliminary study to identify the best

variable region(s) for the ecological niche under study, would improve effectiveness of the 16S-based profiling technique. Near full-length reads of high quality can be generated on the Pacific Biosciences RSII instrument, which uses long-read SMRT sequencing in circular consensus mode. However, the cost per read is much higher than on short-read platforms, which may limit its use for large-scale comparisons of many samples. It is common for current studies to rely on smaller regions, specifically V1-V3 or V3-V5, produced on the Illumina MiSeq or Life Technologies Ion Torrent platforms, for reasons of economy.

The microbial ecosystem of the cattle rumen has been extensively studied (Bath et al., 2013; Myer et al., 2015a; Pitta et al., 2010), and has a complex and diverse structure that is essential for the host to digest plant material. Recent research has focused on characterization of the rumen microbiome among animals differing in feed efficiency (Jami, 2014; McCann et al., 2014). These studies characterized the bacterial community by sequencing 16S rRNA regions V1-V3 or V1/V3 individually. Analysis of these variable regions was reported to demonstrate trends, with some significant changes in specific taxa, but for the most part, these studies lack phylogenetic resolution. Our previous studies have demonstrated genus and species-level associations of bacteria with extremes in feed efficiency in steers (based on feed intake and body weight gain), but noted that in order to better determine the bacterial communities associated with feed efficiency, a finer resolution analysis must be completed, either via metagenomic (shotgun sequence) analyses, a better defined preliminary reference niche, or full-length 16S rRNA analyses (Myer et al., 2015a,b,c; Myer et al., 2016).

Our previous data regarding the impact of microbiome variation on feed efficiency in steers indicated the necessity for increasing the resolution of the analysis. Therefore, this study aimed at contrasting the microbial profiles of the rumen bacterial community produced from 16S

rRNA regions V1-V3 and V1-V8, using the Illumina MiSeq platform (V1-V3) and the Pacific Biosciences RSII instrument (V1-V8). We hypothesized that differences in taxa identification, number of OTUs, OTU richness, phylogeny, and reference database hits would be apparent as a function of partial or near full-length 16S sequence analysis. We also wished to determine if a database of near full-length 16S rRNA sequences from cattle rumen, produced by V1-V8 sequencing, would assist in more accurately determining the microbial community compared to public rRNA databases, when used as the reference for V1-V3 phylogenetic assignment. Finally, we wanted to determine the number of sequences required to generate adequate coverage of all relevant taxa by either the larger or smaller 16S rRNA amplicon(s).

2. MATERIALS AND METHODS

2.1 Experimental design and rumen sampling

This experiment was approved by the U.S. Meat Animal Research Center Animal Care and Use Committee. Feed efficiency was determined as referenced by Myer et al., 2015a, and utilized. Three steers displaying an equivalent feed efficiency phenotype ($ADG_{Greater} - ADFI_{Less}$) and with the least deviation among each other were selected and sampled for the study (see Figure 1 in Myer et al., 2016 for data corroboration).

2.2 DNA extraction, amplification and sequencing

DNA was extracted from rumen samples using a repeated bead beating plus column (RBB+C) method (Yu and Morrison, 2004). Cell lysis was achieved by bead beating 0.15 g of the resuspended sample in ZR BashingBead Lysis Tubes (Zymo Research Corp, Santa Ana, CA, USA) using the TissueLyser II system (Qiagen, Hilden, Germany) for 3 min at 21Hz, in the presence of 4% (w/v) sodium dodecyl sulfate (SDS), 500 mM NaCl, and 50 mM EDTA.

Genomic DNA concentration was determined using a Nanodrop 1000 spectrophotometer (ThermoScientific, Wilmington, USA).

DNA library preparation was performed by PCR amplification of the V1–V3 region of 16S rRNA gene, using modified universal primers 27F (5'- Adapter / Index / AGAGTTTGATCCTGGCTCAG) (Lane, 1991) and 519R (5' Adapter / Index / GTATTACCGCGGCTGCTG) (Turner et al., 1999) including TruSeq® adapters sequences and indices for the Illumina platform. Polymerase chain reaction (PCR) amplification and DNA library preparation of the V1–V8 region was performed using universal primers 27F (5'- AGAGTTTGATCCTGGCTCAG) and 1392R (5'-GACGGGCGGTGTGTAC) for the Pacific Biosciences instrument. Amplification consisted of 20 cycles for both platforms, with an annealing temperature of 58°C. Products were purified using AmPure® bead purification (Agencourt, Beverly MA) and all libraries were quantified by the PicoGreen® dsDNA quantitation kit (Invitrogen, Carlsbad, CA) and by real-time PCR on the LightCycler 480 system (Roche, Mannheim, Germany). The PCR amplicon libraries were sequenced using the 2x300, v3 600-cycle kit and the Illumina MiSeq® sequencing platform (Illumina, San Diego, CA) for the V1–V3 region libraries, and the Pacific Biosciences RSII instrument for the V1–V8 region libraries.

Isolated metagenomic DNA was sheared to 350bp (Covaris, Woburn, MA) and used to create TruSeq® PCR Free libraries for sequencing using the 2x150 NextSeq 500 high output kit and the Illumina NextSeq 500® sequencing platform (Illumina, San Diego, CA). In total, 147,995,952 (22.2 Gb) shotgun reads were generated from one sequencing run.

2.3 Sequence read processing and analysis

All sequences were processed using the QIIME-1.9.1 software package (Caporaso et al., 2010) and Mothur version 1.36.1 (Schloss et al., 2009). For Illumina reads, paired reads were joined using fastq-join (Aronesty, 2011) and filtered for quality ($\geq Q30$) using Mothur. Sequences that contained read lengths shorter than 400bp were removed and adapters/index sequences were trimmed. Pacific Biosciences reads were parsed so that quality scores of zero were interpreted as corresponding to an ambiguous base call, and then filtered for quality ($\geq Q30$) using Mothur. Sequences that contained read lengths shorter than 1200bp were removed. Read directionality was checked and corrected where necessary. For all reads, homopolymers >7 were discarded and chimeric sequences were checked using ChimeraSlayer (Haas et al., 2011). All cleaned sequences were classified into taxa using the Greengenes 16S rRNA Gene Database, 13_8 release (DeSantis et al., 2006). Operational taxonomic units (OTUs) were calculated using the uclust program at 0.03 dissimilarity (Edgar, 2010). Singletons were removed from analysis. Based on rarefaction curves, the number of OTUs was normalized via subsampling 25,000 sequences from each rumen sample for MiSeq data and 40,000 sequences for Pacific Biosciences data. A phylogenetic tree was built with FastTree (Price et al., 2010) to determine alpha and beta diversity metrics.

Shotgun metagenomic reads were similarly cleaned as described above, and mapped against consensus 16S rRNA V1-V3 and V1-V8 sequences, yielding 23,379 and 103,064 reads, respectively. Reads mapping to the respective variable regions were used for analysis and classification using the Greengenes 16S rRNA Gene Database, 13_8 release.

2.4 Analysis

The mean abundances of data metrics and classification accuracies were compared among the variable region groups using one-way ANOVA followed by the Tukey's test. Significant difference was determined at $P < 0.05$. Beta diversity was estimated by determining weighted and unweighted UniFrac distances between samples using the constructed trees and QIIME. The constructed trees were subjected to UniFrac significant test and P test. Samples were clustered based on their between-sample distances using UPGMA (unweighted pair group method with arithmetic mean), and the robustness of the UPGMA tree was estimated using jackknife based on 500 replicates with replacement at a depth of 40,000 and 25,000 sequences per PacBio and MiSeq samples, respectively. Principal coordinates analysis (PCoA) was performed using weighted and unweighted UniFrac distances and analyses (Lozupone and Knight, 2005). Rarefaction curves were calculated using the Ribosomal RNA database project's pyrosequencing pipeline. The cutoff value was equal to or greater than 0.03 dissimilarity.

3. RESULTS

3.1 Metrics

Preliminary data on the rumen in steers determined that 25,000 reads for the final analysis was adequate for coverage of the rumen microbial community using the V1-V3 regions of the 16S rRNA gene. The slopes of the rarefaction curves plotting the number of sequences by the number of OTUs were compared between the two amplicon data sets (Figure 1). The slope at 25,000 reads ($y=0.0213x + 453.22$) for the V1-V3 amplicon was applied to the V1-V8 curve, and indicated that approximately 40,000 reads per sample for the V1-V8 regions was sufficient for comparative analysis. With a greater sequencing length for the V1-V8 region, and taking into consideration that only a minute percentage of the sequences in 16S databases reach near full-length (Kim and Yu, 2010), it was anticipated that a greater amount of sequences would be required to obtain similar coverages compared to shorter reads. Therefore, rarified samples to 25,000 and 40,000 sequences/sample for the V1-V3 and V1-V8 regions, respectively, represented adequate depth (Jami et al., 2012), and were used for comparison between amplicons.

Figure 1. Rarefaction curves of operational taxonomic units (OTUs; $\geq 97\%$ sequence similarity) for Illumina and PacBio-derived sequences.

Total DNA was extracted from the strained rumen contents of three steers (Myer et al., 2015a). All three DNA samples were amplified with both V1-V3, and V1-V8 primer sets in individual reactions (6 total reactions; three each with V1-V3 or V1-V8). Amplification products were processed into libraries for MiSeq or RSII sequencing, respectively. There were 890,086 reads from the MiSeq that passed filtering for quality and for proper 16S amplification; there were 138,180 reads from the RSII that passed corresponding filters. Within animal, these reads were normalized to 25,000 or 40,000 reads for MiSeq or RSII, respectively. Across amplicons, the assigned phylogeny and family/genus/species counts were analyzed via the mean among the three animals. The total number of normalized reads and the mean length of the reads obtained of each sample are presented in Table 1. Reads were examined via bacterial diversity (Shannon Diversity), richness (Chao-1), and coverage (Good's coverage estimator) (Table 1). The average number of OTUs from the normalized groups did not differ ($P > 0.05$) between the two amplicon datasets, which had an average of 1116 ± 269 (V1-V8) and 1010 ± 446 (V1-V3) OTUs per group. Additionally, no significant differences were found within richness metrics ($P > 0.05$) among the number of OTUs, which were determined at 1652 ± 729 and 1511 ± 655 for regions V1-V8 and V1-V3, respectively. Though some variability was observed, the diversity indices for each normalized group were not significantly different. Each group was adequately covered, up to 98.85% in the V1-V8 normalized group and 98.47 % in the V1-V3 normalized group.

3.2 Taxonomic Composition

Relative abundances of bacterial phyla and genera representing $\geq 1\%$ of total reads are presented in Figure 2. The MiSeq V1-V3 normalized reads were classified by the Greengenes 16S rRNA Gene Database (DeSantis et al., 2006) into 20 phyla, 28 classes, 39 orders, 64 families, and 73 genera. Bacteroidetes (68.64%) and Firmicutes (21.58%) dominated the phylum dataset within each normalized group (in terms of percent of the total reads, Figure 2A), and together with the unclassified sequences represented over 97% of all the data. Therefore, the other 18 phyla cumulatively represented less than 3% of the sequences. The V1-V8 normalized reads were classified into 18 phyla, 27 classes, 44 orders, 65 families, and 81 genera. Bacteroidetes (62.18%) and Firmicutes (23.38%) dominated the phylum dataset, similar to the shorter amplicon, although these two phyla only represented 85% of the sequences compared with 90% for V1-V3. In the near full-length 16S amplicon data, the proportion of data not represented by the two dominant phyla was significantly lower, with the other 16 phyla representing over 12% of the sequences. The relative abundance within the phylum Proteobacteria was different ($P < 0.05$) between V1-V8 and V1-V3 amplicons (9.59 vs 0.51%, respectively), and there was a significant difference between the amplicons for the phylum Verrucomicrobia (0.3 vs 0.03%, respectively). In addition, the percentage of unassigned reads was much greater in the V1-V3 sequence data (7% vs. 3% in V1-V8).

Classification and assignment to the genus level was possible for 77% of the shorter amplicon sequences and 72% of the longer amplicon (Figure 2B). Figure 2B represents sequences that were classifiable to the genus level; able to be classified to genus level, but may or may not be designated an assignment. The figure does not include reads that were

Figure 2. The taxonomic profiles for the relative phylum-level (A) and genus-level (B) abundance of each region classified by representation at $\geq 1\%$ of total sequences. Taxonomic composition of the ruminal

microbiota between the two variable regions was compared based on the relative abundance (reads of a taxon/total reads in a sample).

unclassifiable at any taxonomic level. This classification revealed that nearly all (98% or 97% according to the V1-V3 and V1-V8 amplicons, respectively; data not shown) of the Bacteroidetes taxon was composed of the genus *Prevotella*, suggesting that the overall prevalence of this genus may be as high as 67% in these rumen samples. The second most abundant genus, according to the larger amplicon, were members of the family Succinivibrionaceae, although the shorter amplicon did not place this family among the most abundant. Overall, several lower orders of taxa had significant differences in their relative abundances between amplicons (Figure 2B). Respectively, families *Succinivibrionaceae* (9.01; 0.18%) and *Paraprevotellaceae* (1.36; 0.51%), and genera *Succiniclasticum* (3.53; 1.52%) and *Succinivibrio* (0.26; 0.01%), represented differences between the PacBio V1-V8 and Miseq V1-V3 amplicons ($P < 0.05$), which have been identified in the rumen at abundances greater than 0.1% (Myer et al., 2015a; Pitta et al., 2014).

Community composition estimates and analyses are based on PCR-amplified DNA fragments, which may suffer from potential primer bias (Logares et al., 2014; Giovannoni and Suzuki, 1996; Polz and Cavanaugh, 1998). To examine this potential bias, alternative V1-V3 amplicons were generated by trimming the PacBioV1-V8 amplicons to 492bp (V1-V3) and classified by the Greengenes 16S rRNA Gene Database, as performed previously. The resultant taxonomic profiles observed in Figure 2 demonstrated differences between the trimmed PacBio V1-V3 and the MiSeq V1-V3, as well as the PacBio V1-V8 amplicons, at both the phylum and genus levels. Indication of potential primer bias was observed, as the taxonomic profiles at

phylum-level between the PacBio V1-V8 and the trimmed PacBio V1-V3 amplicons were remarkably similar in composition contrasted to the MiSeq V1-V3 amplicons, especially with regard to the abundance and classification of Proteobacteria (9.5% V1-V8, 10.3% V1-V3). The profiles at the genus-level were not as conclusive. Unassigned taxa at the genus level from both V1-V3 datasets were similar at 7.5% and 8.0% for the MiSeq and PacBio amplicons, respectively, compared to 2.8% for the V1-V8 amplicon. This was also seen at the phylum level.

To further examine the potential for primer bias, direct sequencing of metagenomic DNA was used to circumvent the primer-associated bias. In these data, metagenomic DNA was mapped against consensus 16S rRNA V1-V3 and V1-V8 regions, and the resultant mapped reads were classified by the Greengenes 16S rRNA Gene Database. Rarefaction indicated adequate depth for each amplicon (see Figure 2 in Myer et al., 2016). Relative abundances of bacterial phyla and genera representing $\geq 1\%$ of total reads are presented in Figure 3. At the phylum level (Figure 3A), though there were noted differences between the metagenomic and amplicon profiles, especially with regard to the potential overestimation of Bacteroidetes derived from the MiSeq V1-V3 region, the aforementioned amplicon data were distinct from the remaining profiles with respect to the abundances of Proteobacteria. Unassigned reads were reduced in the metagenomic 16S dataset to 1.3% and 1.5% for the V1-V3 and V1-V8 regions, respectively. Similar associations were demonstrated at the genus level and noted within *Prevotella* and family Succinivibrionaceae, reaffirming the genus-level abundances (Figure 3B, C).

3A.

3B.

3C.

Figure 3. Metagenomic 16S comparison of the taxonomic profiles for the relative phylum-level (A) and genus-level (B/C) abundance of each region classified by representation at $\geq 1\%$ of total sequences. Due to major shifts in taxa, the amplicon (B) and metagenomic (C) profiles are represented within separate figures. Taxonomic composition of the ruminal microbiota was compared based on the relative abundance (reads of a taxon/total reads in a sample).

Observed differences in abundance of unclassified reads between the V1-V3 and V1-V8 combined variable regions were further investigated by comparing the proportion of 16S rRNA PacBio and Illumina-generated sequences **successfully** classified and assigned at five taxonomic levels (Figure 4). These data additionally include sequences that do not classify at any level. This analysis demonstrated that the classification success of the V1-V8 combined variable regions was significantly greater ($P < 0.05$) than the V1-V3 combined variable regions, at all five taxonomic levels, using the Greengenes 16S rRNA Gene Database to perform the comparison (DeSantis et al., 2006). As expected, all variable region samples decreased in classification success as the taxonomic levels approached the genus classification, with the

greatest at the phylum-level (73.21; 85.91%) and least at the genus-level (38.52; 58.07%), for V1-V3 and V1-V8 reads, respectively.

Figure 4. Proportion of 16S rRNA PacBio and Illumina-generated sequences successfully classified and assigned at five taxonomic levels. Error bars represent standard errors of the mean. * $P < 0.05$

3.3 Beta Diversity

Beta diversity is useful in the measurement of the degree of similarity between communities. Phylogenetic metrics can be used to document shifts in the phylogenetic composition of the communities, rather than focusing on inter-community shifts. Therefore, the rumen samples were compared among each other via UniFrac distance matrices using jackknifed UPGMA (unweighted pair-group method with arithmetic mean) clustering (Lozupone and Knight, 2005), based either on the longer or shorter amplicon sequences. Note that long and short amplicon sequences are compared independently among rumen samples, rather than comparing short and long amplicons within rumen sample. The resulting UPGMA trees for short or long

amplicons were compared. Although nodes were less resolved, clustering of the samples differed when comparing the V1-V8 and V1-V3 16S regions (Figure 5) in both unweighted and weighted distance matrices. Branch lengths differed in the unweighted measure. Furthermore, although each amplicon dataset consisted of only 3 samples, the corresponding samples in each dataset split differently in the weighted measure, based on the amplicon used for analysis. To highlight this contrast, principal coordinates analysis was performed, to better illustrate the differences compared to UPGMA clustering (Figure 6).

Figure 5. UPGMA tree of unweighted (A) and weighted (B) UniFrac distances between rumen samples determined using the PacBio-derived V1/V8 sequence data (left) as well as the MiSeq-derived V1/V3 sequence data (right). Nodal supports were assessed by jackknife using 500 replicates of 40000 and 25000 sequences per sample A and B, respectively.

Unweighted and weighted UniFrac significance tests and the P test were used to determine differences among the bacterial communities by calculating distances between rumen

samples. These tests are commonly-used phylogenetic approaches to assess community differentiation (Lozupone and Knight, 2005; Martin, 2002). The results showed that rumen microbial communities identified by near full-length and partial 16S sequence regions were significantly different ($P < 0.01$).

Figure 6. Unweighted (A) and weighted (B) UniFrac principal coordinates analysis (PCoA) using sequences from each sample. Samples from PacBio-derived V1/V8 are Rumen 1 (purple), Rumen 2 (green), and Rumen 3 (orange). Samples from MiSeq-derived V1/V3 are Rumen 1 (yellow), Rumen 2 (blue), and Rumen 3 (red). Data are analyzed from rarefied subsets of 40,000 and 25,000 sequences from each sample V1/V8 and V1/V3, respectively.

4. DISCUSSION

Numerous factors such as budget, niche, organismal coverage, and sequencing platform contribute towards the selection of 16S rRNA gene fragment selected for sequencing when interrogating microbial community profiles. Comparing time to run and number of bases per run between both platforms can be difficult to compare directly. For example, the MiSeq takes three days to run a 2x300 paired-end run, but can sequence many libraries in parallel, whereas each cell of the PacBio RSII is an independent library and takes 4 hours run time. The MiSeq can

generate 25 million reads per run, however one cannot achieve that with entirely 16S sequence, as one has to dilute it with some other more complex library, with the best reference being the Illumina guidelines that suggest use of 50% PhiX control if conducting 16S sequencing. This effectively results in approximately 12 million reads per run. The RSII cells each produce about 80-100,000 reads, which is far less than the MiSeq instrument, but in excess and far adequate for an accurate microbiome determination. Regardless of instrument, this study examined the differences in the ruminal bacterial community and its phylogenetic diversity as defined by two differing 16S rRNA amplicons. However, these considerations must be made when ultimately evaluating experimental design.

Data regarding each specific ecological community can be confounded and biased based upon the variable region(s) selected (Baker et al., 2003; Kim et al., 2011). Further, the extent of sequence diversity within the 16S rRNA gene varies between regions, which also may affect the accurate profiling of bacterial communities (Chakravorty et al., 2007). The primary aim of this study was to evaluate the use of partial or near full-length 16S sequence for analysis of the rumen bacterial community in steers, rather than focus on sequencing technology. When examining the alpha diversity of the sample from the two differing amplicons, the results revealed no differences between any of the metrics measured, specifically that of OTU count and richness (Table 1). These data also validated the method of comparison from the OTU rarefaction curves, which were normalized to 40,000 reads/sample for PacBio (V1-V8) reads, compared to the MiSeq reads at 25,000 reads/sample (V1-V3). Additionally, the study was able to recover approximately 98% of all OTUs calculated at 0.03 dissimilarity, as determined by Good's coverage estimator, which demonstrated adequate depth established in previous studies

(Myer et al., 2015a; Jami and Mizrahi, 2012). This was expected, as the aim was to accurately compare the two groups at their phylogenetic and taxonomic levels, whereas differences in alpha diversity indices are more useful for pinpointing errors in the methodology. Although there were no changes regarding the alpha diversity among the samples from the two groups examined, the results do not necessarily indicate a lack of associated differences resulting from the choice of amplicon sequenced. For example, variations in the assignment and distribution of taxa may be evident, as demonstrated in Figure 2. This has also been demonstrated in other studies that focus on bacterial community diversity analyses (Ibarbalz et al., 2014).

The relative bacterial taxonomic abundances produced by the amplicon groups were consistent with previous studies of rumen content microbial communities, where the current consensus abundance of *Prevotella* has been estimated between 42% and 60%, based on bacterial 16S rRNA gene sequences (Stevenson and Weimer, 2007; Pitta et al., 2010). Also, in agreement with the current study, *Prevotella* has been reported as the most abundant ruminal genus (Stevenson and Weimer, 2007). This was important in verifying the reliability of the dataset, especially with regard to the V1-V8 regions, which to date has not been utilized to survey the bacterial community within the rumen of steers. The data show that the bacterial community identified based on the PacBio V1-V8 regions followed the generally accepted consensus of the ruminal bacterial community.

Commonly used primer pairs can produce differences in coverage and taxa identification (Klindworth et al., 2013). The regions used in this study held no exception. When determining the number of organisms identified at each level of classification, three out of the five taxonomic levels examined were enriched, showing greater numbers in identified taxa using the V1-V8 amplicon for 16S identification. This was anticipated, as

longer sequence reads have been shown to provide a greater phylogenetic resolution (Fichot and Norman, 2013). Longer amplicons identified a greater number of phyla, such as Proteobacteria and Verrucomicrobia. Although Proteobacteria are not regarded as a major phylum in the rumen of steers, it is evident that classification using only the V1-V3 hypervariable region may have contributed towards an underestimation of its abundance in this study.

Functional, procedural, and platform-specific differences exist between the two sequencing methodologies, and artifacts of such differences may be reflected within the datasets. Bacterial community composition estimates from 16S rRNA gene amplicons include some degree of primer bias. In addition, the PacBio instrument may preferentially load smaller fragments, which may also be a function of PCR product exhaustion during sequencing. We found no evidence that this particular bias influenced the results, as demonstrated by (1) reduced amplification and pooling of several PCR products to produce one PacBio V1-V8 library (see Figure 3 in Myer et al., 2016), (2) the use of trimmed PacBio V1-V3 and metagenomic 16S reads permitted examination of primer bias within the PacBio reads, and (3) the assessment of taxa assignment without primer bias. The trimmed data suggested that some degree of primer bias may exist within the data derived from the PacBio reads (Figure 2), indicated by shifts in the relative taxonomic abundances of Proteobacteria and Bacteroiodes compared to the full V1-V8 amplicon sequences. The data support an underestimation of Proteobacteria by the MiSeq-derived V1-V3 region, however, these data alone may or may not indicate primer bias, as other hypervariable regions have been found to better identify and classify Proteobacteria (Li et al., 2009). Additionally, the abundance of unassigned taxa in the MiSeq V1-V3 profile is similar to that of the PacBio V1-V3 unassigned taxa, which would be expected from the shorter amplicon.

The metagenomic 16S data aided in the understanding of potential primer bias and the taxonomic abundances of Proteobacteria and Bacteroidetes (Figure 3). While overestimating the abundance of Proteobacteria and underestimating the abundance of Bacteroidetes from conventional estimates of the rumen microbial community (Jami et al., 2014; de Menezes et al., 2011; Pitta et al., 2014), the current data nonetheless indicated that changes in the taxonomic profiles may be less due to primer bias, and more related to the ability to indentify key taxa from the shorter 16S amplicon. This may prove critical in studies of the rumen, as recent DNA and RNA studies have implicated the possible increased role of Proteobacteria in rumen metabolism (Kang et al., 2013). These data also highlight the necessity for further analysis and finer resolution examination of bacterial microbial communities, not relying solely on the results of a single 16S fragment or variable region. It is evident that either applying the datasets synergistically or as controls for accuracy may better resolve microbial community analyses with ribosomal gene fragments. Comparatively, one can arrive at a more comprehensive community, better aiding the understanding of underlying relationships and interactions.

In addition to variation in the relative taxonomic abundance, the study also examined the classification accuracy between the variable regions surveyed. Classification success is an underlying factor in community analysis, where much of the data may be lost, depending on the techniques utilized in the study. The relative taxonomic abundance data in Figure 2 identified the unassigned reads only at that specific taxonomic level; however, the study also aimed to quantify the total accuracy, including cumulative unassigned reads at all levels. Figure 4 illustrates that the proportion of successfully classified 16S rRNA V1-V8 reads was significantly greater at each taxonomic level compared to those of the V1-V3 reads using the Greengenes 16S rRNA Gene Database, which is consistent with current research demonstrating the improved

accuracy of classification utilizing longer 16S amplicons (Nossa et al., 2010). It has also been demonstrated that smaller 16S amplicons generally negatively impact classification success (Liu et al., 2008). As a whole, the abundance of unclassified sequences in both datasets can be related to database and niche. In this case, it has been demonstrated that phylogenetic analysis of organisms known to colonize the rumen have been hindered by limited reference datasets (Jumpstart, 2012). Further development and expansion of the 16S reference sets will likely improve the performance of classification. Additionally, this also urges the use of niche-specific reference databases, which would better resolve community analyses at many levels. The results are also dependent upon the bias of sequences available in a respective database. Generally, amplification loci specific to either end of the 16S gene will have a low number of database matches due to the less often sequenced areas (Baker et al., 2003). The V1-V3 dataset is subject to such bias, adding further support for the use of longer reads for microbiome analysis, whether for direct determination or for classification accuracy and abundance.

In order to compare the degree of similarity among communities established by the amplicon groups rather than among the samples, we utilized UniFrac distance matrices using jackknifed UPGMA clustering as well as PCoA. The UPGMA trees (Figure 5) and PCoA (Figure 6) illustrated that there were differences between the unweighted and weighted UniFrac distances at both branch length and clustering from samples using the amplicon sequence data, and these differences were significant. The results suggest that the partial and near full-length sequence regions of the analyzed amplicons significantly affected the phylogenetic evaluation of microbial communities using UniFrac. Additionally, these data demonstrated differing community analyses among differing amplicon sequence data, adding further support for the careful selection of 16S amplicons in microbial community analysis. The significant differences

in the beta diversity inferred from the two identified 16S rRNA gene fragments verified changes in the two communities that may not be apparent by simple taxonomic or alpha diversity comparisons. The microbial community profiles identified in this study from the partial or near full-length regions depict significantly different, non-comparable community profiles, which must be carefully considered when determining factors and their effect on community composition and structure.

4.1 Conclusions

Amplicon choice in 16S rRNA analyses is a significant stage in conducting 16S-based microbial community research. The current study highlights the differences of the rumen bacterial community in steers as a function of 16S rRNA amplicon. The study demonstrated the increased phylogenetic resolution, differential taxonomic classification, greater taxonomic classification accuracy, as well as differences in community composition, when using near-full length 16S reads in the rumen of steers compared to the shorter V1-V3 reads. Although this may warrant the use of full-length or near full-length reads for microbial community analyses, the aforementioned factors, such as budgetary constraints, may limit the ability to conduct large, in-depth studies using longer reads generated by the Pacific Biosciences RSII instrument. Community-based microbial databases may be a more practical option in determining an accurate microbiome (Ross et al., 2012). In such cases, longer reads generated by the PacBio instrument may be better at determining the niche-specific microbiome, as a reference for greater throughput, short-read, 16S analysis. This would also aid with database accuracy and limiting unassigned taxa. A synergistic approach utilizing both 16S amplicon sizes may be warranted in order to achieve a more comprehensive analysis of bacterial communities.

Research examining ecological niches and bacterial community structure and phylogeny, such as the cattle rumen, has been revolutionized by the use of high throughput platforms and great sequencing depth. The use of next-generation sequencing platforms, such as the MiSeq, has permitted the generation of enormous amounts of data for community analysis. Consequently, its use is also dependent upon shorter read lengths, which may inhibit accurate community analysis depending on the niche examined. This study used two sequencing platforms to determine the efficacy of reporting ruminal microbial communities from the shorter V1-V3 16S rRNA amplicon, with the underlying concept that many studies have, and will continue to use high-throughput platforms to acquire microbiome data. Although such platforms may be optimal in terms of time and quantity of data acquisition, this study supports the use of many methods initially, to ensure subsequent accurate community analyses.

ACKNOWLEDGEMENTS

We thank Bob Lee, Sue Hauver, Kelsey McClure, Renee Godtel, and Brooke Clemmons for technical assistance. This project is partially supported by Agriculture and Food Research Initiative Competitive Grant No. 2011-68004-30214 from the USDA National Institute of Food and Agriculture.

REFERENCES

- Aronesty, E.** 2011. *ea-utils* : Command-line tools for processing biological sequencing data. <http://code.google.com/p/ea-utils>.
- Baker GC, Smith JJ, Cowan DA.** 2003. Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods*. **55**(3):541-555.

Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers, *Journal of Microbiological Methods*. **55**(3):541-555.

Bath C, Morrison M, Ross EM, Hayes B J, Cocks B G. 2013. The symbiotic rumen microbiome and cattle performance: a brief review. *Animal Production Science*. **53**:876–881.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. Qiime allows analysis of high throughput community sequencing data. *Nat. Methods* **7**:335–336. doi:10.1038/nmeth.f.303.

Chakravorty S, Helb D, Burday M, Connel N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiology Methods*. **69**(2):330-339.

Claesson MJ, O'Sullivan O, Wang Q, Nikkila J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW. 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE*. **4**:e6669.

de Menezes AB, Lewis E, O'Donovan M, O'Neill BF, Clipson N, Doyle EM. 2011. Microbiome analysis of dairy cows fed pasture or total mixed ration diets. *FEMS Microbiol. Ecol.* **78**(2):256-265.

DeSantis T Z, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol.* **72**:5069-72.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19):2460-2461.

Fichot EB, Norman SR. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*. **1**:10.

Forney LJ, Zhou X, Brown CJ. 2004. Molecular microbial ecology: land of the one-eyed king. *Curr Opin Microbiol*. **7**:210–220.

Giovannoni SJ, Suzuki MT. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol*. **62**(2):625-30.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ, The Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW. 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*. **21**:494–504.

Hamady M, Knight R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*. **19**:1141–1152.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*. **4**:e1000255.

Ibarbalz FM, Pérez MV, Figuerola ELM, Erijman L. 2014. The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics. *PLoS ONE*. **9**(6):e99722.

Jami E, Mizrahi I. 2012. Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. *PLoS ONE* **7**(3):e33306.

- Jami E, White BA, Mizrahi I.** 2014. Potential Role of the Bovine Rumen Microbiome in Modulating Milk Composition and Feed Efficiency. *PLoS ONE*. **9**(1): e85423.
- Jumpstart Consortium Human Microbiome Project Data Generation Working Group.** 2012. Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research. *PLoS ONE*. **7**(6):e39315.
- Kang SH, Evans P, Morrison M, McSweeney C.** 2013. Identification of metabolically active proteobacterial and archaeal communities in the rumen by DNA- and RNA-derived 16S rRNA gene. *Journal of Applied Microbiology*. **115**:644–653.
- Kim M, Morrison M, Yu Z.** 2011. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microbiol Methods*. **84**(1):81-87.
- Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C , Horn M, Glockner FO.** 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*. **41**(1): e1.
- Lane DJ.** 1991. 16S/23S rRNA sequencing. In: *Nucleic acid techniques in bacterial systematics*. Stackebrandt, E., and Goodfellow, M., eds., John Wiley and Sons, New York, NY, pp. 115-175.
- Li H, Zhang Y, Li DS, Xu H, Chen GX, Zhang CG.** 2009. Comparisons of different hypervariable regions of rrs genes for fingerprinting of microbial communities in paddy soils. *Soil Biology and Biochemistry*. **41**:954-968.
- Liu Z, DeSantis TZ, Andersen GL, Knight R.** 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*. **36**:e120.
- Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*. **35**:120.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmiento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG. 2014. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol.* **16**(9):2659-2671.

Lozupone CA, Knight R. 2005. Unifrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**:8228–8235.

Martin A. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology.* **68**:3673-3682.

McCann JC, Wiley LM, Forbes TD, Rouquette FM Jr, Tedeschi LO. 2014. Relationship between the Rumen Microbiome and Residual Feed Intake-Efficiency of Brahman Bulls Stocked on Bermudagrass Pastures. *PLoS ONE* **9**(3):e91864

Myer PR, Kim M, Freetly HC, Smith TPL. 2016. Analysis of metagenomic and near full-length 16S rRNA sequences to determine methods for phylogenetic analysis of the rumen bacterial community in steers. Data in Brief. *Submitted.*

Myer PR, Smith TPL, Wells JE, Kuehn LA, Freetly HC. 2015a. Rumen microbiome from steers differing in feed efficiency. *PLoS ONE* **10**:e0129174. doi:10.1371/journal.pone.0129174.

Myer PR, Wells JE, Smith TPL, Kuehn LA, Freetly HC. 2015b. Cecum microbial communities from steers differing in feed efficiency. *J. Anim. Sci.* **93**:5327–5340. doi:10.2527/jas.2015-9415.

Myer PR, Wells JE, Smith TPL, Kuehn LA, Freetly HC. 2015c. Microbial community profiles of the colon from steers differing in feed efficiency. *SpringerPlus* **4**(1):1– 13. doi:10.1186/s40064-015-1201-6.

Myer PR, Wells JE, Smith TPL, Kuehn LA, Freetly HC. 2016. Microbial community profiles of the jejunum from steers differing in feed efficiency. *J. Anim. Sci.* **94**:327–338. doi: 10.2527/jas2015-9839

Nossa CW, Oberdorf WE, Yang L, Aas JA, Paster BJ, Desantis TZ, et al. 2010. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol.* **16**:4135–4144.

Pinto AJ, Raskin L. 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One.* **7**:e43093.

Pitta D, Pinchak W, Dowd S, Osterstock J, Gontcharova V. 2010. Rumen bacterial diversity dynamics associated with changing from bermudagrass hay to grazed winter wheat diets. *Microb Ecol.* **59**:511–522.

Pitta DW, Kumar S, Vecchiarelli B, Shirley DJ, Bittinger K, Baker LD, Ferguson JD, Thomsen N. 2014. Temporal dynamics in the ruminal microbiome of dairy cows during the transition period. *J Anim Sci.* **92**(9):4014-4022.

Polz, MF, Cavanaugh, CM. 1998. Bias in Template-to-Product Ratios in Multitemplate PCR. *Appl. Environ. Microbiol.* **64**(10):3724-3730.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE.* **5**:E9490.

Ross EM, Moate PJ, Bath CR, Davidson SE, Sawbridge TI, Guthridge KM, Cocks BG, Hayes BJ. 2012. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. *BMC Genetics* **13**:53.

- Schloss PD, Westcott SL, Ryabin T, et al.** 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol.* **75**(23):7537-41. doi: 10.1128/AEM.01541-09.
- Stevenson DM, Weimer PJ.** 2007. Dominance of *Prevotella* and low abundance of classical ruminal bacterial species in the bovine rumen revealed by relative quantification real-time PCR. *Applied Microbiology and Biotechnology.* **75**(1):165-174.
- Turner S, Pryer KM, Miao VPW, Palmer JD.** 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol.* **46**:327–338.
- Vinje H, Liland KH, Almøy T, Snipen L.** 2015. Comparing K-mer based methods for improved classification of 16S sequences. *BMC bioinformatics*, **16**(1):205.
- von Wintzingerode F, Gobel UB, Stackebrandt E.** 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev.* **21**:213–229.
- Yu Z, Morrison M.** 2004. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *Biotechniques.* **36**(5):808-813.

Amplicons Technology	V1-V8	V1-V3
	PacBio	MiSeq
Number of Normalized Reads	40,000	25,000
Mean Length	1397.67 ± 0.29	483.43 ± 1.97
No. of OTUs	1116 ± 269	1010 ± 446
Chao1	1652 ± 729	1511 ± 655
Shannon Diversity	5.80 ± 0.86	6.13 ± 1.55
Good's Coverage	98.85 ± 1.58	98.47 ± 1.13

Table 1. Diversity statistics among reads from grouped samples ($\geq 97\%$ sequence similarity). Means among the groups were compared using ANOVA and the Tukey's Test ($P < 0.05$).