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

Phillip R. Myer [a,*], MinSeok Kim [b], Harvey C. Freetly [b], Timothy P.L. Smith [b]

[a] Department of Animal Science, University of Tennesse Institute of Agriculture, University of Tennessee, Knoxville, TN 37996
[b] USDA-ARS, U.S. Meat Animal Research Center, Clay Center, NE 68933 [1]

ARTICLE INFO

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.

© 2016 Elsevier B.V. All rights reserved.

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 and Knight, 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 and Raskin, 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., 2004; 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

---

☆ Mention 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.
* Corresponding author.
E-mail addresses: pmyer@utk.edu (P.R. Myer), mkim2276@korea.kr (M. Kim), harvey.freetly@ars.usda.gov (H.C. Freetly), tim.smith@ars.usda.gov (T.P.L. Smith).
[1] USDA is an equal opportunity provider and employer.

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 et al., 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, 2015b, 2015c, 2016a).

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., 2016a 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'- GACGGCGGTGTGTAC) 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 (≥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 (≥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 phylogenic 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