PLOS ONE

RESEARCH ARTICLE

Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing

Phillip R. Myer[1]*, Tara G. McDaneld[2], Larry A. Kuehn[2], Keith D. Dedonder[3¤a], Michael D. Apley[3], Sarah F. Capik[3¤b], Brian V. Lubbers[3], Gregory P. Harhay[2], Dayna M. Harhay[2], John W. Keele[2], Madison T. Henniger[1], Brooke A. Clemmons[1], Timothy P. L. Smith[2]

1 Department of Animal Science, University of Tennessee Institute of Agriculture, University of Tennessee, Knoxville, TN, United States of America, **2** USDA-ARS, U.S. Meat Animal Research Center, Clay Center, NE, United States of America, **3** College of Veterinary Medicine, Kansas State University, Manhattan, KS, United States of America

¤a Current address: Veterinary and Biomedical Research Center, Inc, Manhattan, KS, United States of America
¤b Current address: College of Veterinary Medicine & Biomedical Sciences, College Station, TX, United States of America
* pmyer@utk.edu

OPEN ACCESS

Citation: Myer PR, McDaneld TG, Kuehn LA, Dedonder KD, Apley MD, Capik SF, et al. (2020) Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. PLoS ONE 15(7): e0235498. https://doi.org/10.1371/journal.pone.0235498

Editor: Bernd Schierwater, Tierarztliche Hochschule Hannover, GERMANY

Received: July 30, 2019

Accepted: June 17, 2020

Published: July 13, 2020

Data Availability Statement: These sequence data are available from the NCBI Sequence Read Archive (SRA Accession PRJNA548468).

Funding: The authors received no funding for this work.

Competing interests: The co-author KD Dedonder is currently employed by Veterinary and Biomedical Research Center, Inc, Manhattan. However, the past work was completed at Kansas State University during their education program. This is reflected in the affiliation statement. The current

Abstract

Surveys of microbial populations in environmental niches of interest often utilize sequence variation in the gene encoding the ribosomal small subunit (the 16S rRNA gene). Generally, these surveys target the 16S genes using semi-degenerate primers to amplify portions of a subset of bacterial species, sequence the amplicons in bulk, and assign to putative taxonomic categories by comparison to databases purporting to connect specific sequences in the main variable regions of the gene to specific organisms. Due to sequence length constraints of the most popular bulk sequencing platforms, the primers selected amplify one to three of the nine variable regions, and taxonomic assignment is based on relatively short stretches of sequence (150–500 bases). We demonstrate that taxonomic assignment is improved through reduced unassigned reads by including a survey of near-full-length sequences specific to the target environment, using a niche of interest represented by the upper respiratory tract (URT) of cattle. We created a custom Bovine URT database from these longer sequences for assignment of shorter, less expensive reads in comparisons of the upper respiratory tract among individual animals. This process improves the ability to detect changes in the microbial populations of a given environment, and the accuracy of defining the content of that environment at increasingly higher taxonomic resolution.

Introduction

A critical step in 16S-based microbiome research is the assignment of microbial taxonomy. Currently, most research utilizes public 16S reference databases, such as Greengenes, SILVA, RDP, or GenBank, to assign taxonomy to sequence data [1–4]. However, these databases may

company provided no support in the form of salary for author KDD, and did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. There are no relevant patents or products to declare. This does not affect our adherence to PLOS ONE policies on data or materials sharing.

pose limitations to taxonomic assignment due to the large amount and length of sequences available as well as the variable quality of data. For example, 43% of full-length 16S small-subunit rRNA gene sequences available in the GenBank database were identified as chimeras and unable to be classified, resulting in low-quality data that can ultimately alter taxonomic assignment [1]. Unlike GenBank, Greengenes and SILVA utilize a chimera-checking and alignment algorithm for quality filtering [5, 6]. Even with these quality-checking programs, these public 16S reference databases contain between 0.2% to 2.5% mislabeled sequences, with errors associated with knowledge bases that have misidentified species or misidentified sequence data that result in incorrectly assigned taxa [7]. When these reference databases were compared among each other using taxonomy annotations provided by RDP to assess taxonomic assignment, Greengenes and SILVA significantly differed from each other, with an error rate around 17% [8]. Not only do these public databases contain significant error rates but also thousands of sequences. In August 2007, SILVA had 546,521 sequences available [6]; now, SILVA reports having over 6 million aligned rRNA sequences [9]. The sequences available in these databases are not always full-length sequences as well, and have historically focused on amplicons of hypervariable regions [10]. The classification accuracy differs among hypervariable regions, and can negatively impact taxonomic assignment based on the regions referenced [11]. Greengenes provides full-length and chimera-checked 16S rRNA gene sequences, created from multiple curators who provide their sequences from a single study. However, among many curators, there can be multiple incongruencies in taxonomic nomenclatures at the phylum level (1). Ultimately, this has the potential to lead to errors in assignment when using these public databases. In addition, the extensive number of sequences to process can have a large impact on computational resources [12]. Therefore, to assist with the limitations associated with public 16S reference databases, using custom 16S reference databases developed for specified microbial niches has the potential to improve taxonomic assignment while reducing the computational load that is associated with public 16S reference databases. With increased performance, researchers can more accurately investigate the impact of microbial communities and understand their roles in specific environments, particularly with diseases that are high-impact in certain industries.

Bovine respiratory disease (BRD), or shipping fever, substantially impacts the U.S. beef cattle industry, with economic losses exceeding $1 billion annually [13, 14]. Calf mortality, veterinary and/or treatment costs, and labor are among the primary contributors to the financial burden of BRD on the industry. The manifestation of the disease can result in reduced growth and reproductive performance [15, 16], subsequently affecting long-term herd performance and longevity. A variety of complex factors can be involved with BRD pathogenesis that involve host elements, environmental elements, inadequate management and housing, as well as viral and bacterial pathogens, such as *Mannheimia haemolytica* and *Mycoplasma bovis* [17, 18]. Yet, the presence of these pathogens in the upper respiratory tract (URT) of cattle may not necessarily result in disease [19]. Indeed, microbes such as *M. haemolytica* are opportunistic and result in diseased states when the animal is exposed to environmental stressors such as transport, weaning, or viral pathogens [20]. As BRD is reliant on opportunistic conditions, research is beginning to focus on the association of these bacterial pathogens with the animal's nasopharyngeal microbiome [21–23]. Understanding these associations are key in determining the role of the microbiome on BRD, and how microbial community dysbiosis may impact the incidence of respiratory disease in cattle.

Currently, microbiome research efforts with regard to BRD have focused on characterizing the microbiome of the bovine nasal cavity in normal and diseased states in varying stages of production [23–25]. These community-based studies conducted 16S rRNA analyses utilizing short-read technologies. However, as is similar in other fields of animal agriculture microbiome research, typically a large quantity of sequence reads is left unassigned (approximately 10–20% in BRD research) [26] and/or operational taxonomic units (OTUs) with low sequence counts are removed from analyses. These shortcomings decrease the power of analyses to detect small, potentially important variations in microbial community structure and function that may have phenotypic or disease-related impacts. Greater use of community-based microbial databases has been a more practical option in defining accurate microbiomes [27, 28]. Utilizing niche-specific microbial databases developed from full-length or near-full-length 16S reads for short-read 16S analyses would aid in microbial identification and further limit unassigned taxa [29, 30]. In BRD research this is critical, as optimizing the ability to accurately and comprehensively investigate the nasopharyngeal microbiome permits researchers with greater opportunities to define the complex interactions in this multifactorial disorder. In this report, we describe the development of a 16S rRNA sequence database for the bacterial community present in the URT of cattle, to provide improved assignment of taxa within the bovine URT and as a site-specific tool for examining microbiome dysbiosis involved in BRD.

Materials and methods

Database construction

Animal selection and sampling. This study was approved and carried out in accordance with the recommendations of the Kansas State University and U.S. Meat Animal Research Center Institutional Animal Care and Use Committees. Animal selection and sampling procedures for the "KSU" population were conducted as described in DeDonder and colleagues [31]. The KSU population consisted of a total of 180 cattle with an initial body weight ranging from 164–269 kg, obtained from Athens, Tennessee (*n* = 60), Maryville, Missouri (*n* = 60), and Richmond, Kentucky (*n* = 60). Comingled steers and bulls were acquired from multiple origins and mixed breeds to obtain a broad spectrum of variability for bacterial classification. The cattle were transported to a feeding facility in Kansas and housed in open-air, dirt floor, group housing pens. Cattle were randomly assigned one of two treatments upon arrival at the facility, receiving either mock treatment saline at 2 mL / 49.9 kg or metaphylactic gamithromycin at the dose of 2 mL / 49.9 kg subcutaneously in the neck. Additional information regarding vaccines, growth implants, and diets are further described according to DeDonder and colleagues [31]. Animal selection and sampling procedures for the "USMARC" population were conducted as described by McDaneld et al. (24). Samples were collected in the years 2010–2012 from animals belonging to advanced generations of the U.S. Meat Animal Research Center GPE (Germplasm Evaluation Program; [32]) herd, Clay Center, Nebraska. Approximately 800 animals were produced in multiple-sire matings of crossbred cows to F1 and purebred bulls from various breeds, consisting of variable fractions of 18 breeds: Angus, Hereford, Red Angus, Brahman, Charolais, Gelbvieh, Limousin, Simmental, Brangus, Beefmaster, Shorthorn, Maine Anjou, Santa Gertrudis, Chiangus, Salers, Braunvieh, South Devon, and Tarentaise. As a number of factors affect the microbial communities of the nares in cattle with regard to BRD, such as environment, management, sex, breed, and clinical treatment [23, 33], this optimized strategy and diverse mix of cattle was used to establish the bacterial Bovine URT 16S rDNA database for cattle with BRD.

The KSU population was sampled multiple times, including at the sale barn prior to departure on transport trucks, during unloading upon arrival at the feedlot, and at least twice more during the 28-day feeding period. Some animals had additional sample collected if they were diagnosed as potential BRD cases. Sampling was conducted by deep nasopharyngeal swab inserted through the nares by a trained veterinarian as described [17]. Briefly, the double guarded, sterile uterine swab was inserted into the nasal cavity to the point of resistance against the swab, indicating nasopharyngeal tonsilar tissue. The swab was rotated to obtain sample from the pharyngeal tissues. During swab removal, the guarded sleeve protected the swab from contamination. One swab was inserted in each nostril, and kept separate during transport. The swabs were stored in liquid Amies media and sent on ice to the USDA-US Meat Animal Research Center (Clay Center, NE) for same-day processing (approximately 4 hours in transit). The tubes containing media and swabs were vortexed for 30 seconds to displace bacteria still stuck on the swabs, and then the swabs were removed. Aliquots from the tubes of independent swabs from both nares of each animal were taken, and then the remainder of each pair of tubes was combined into a single animal-by-time-point combined sample. Glycerol was added to 20% final solution to each combined sample and independent aliquot, mixed by inversion, allowed to stand for 5 minutes on ice, and then stored at -80˚C.

The USMARC population was sampled using 6-inch nasal swabs of calves at initial vaccination (approximately 40 days of age), preconditioning (approximately 130 days of age), weaning (approximately 150 days of age), and if the calf was diagnosed with BRD during a 1-5-week period after weaning. For sampling, the nose of the animal was wiped cleaned with a single-use towel if fecal material was present. The unguarded 6-inch nasal swab was then gently inserted into the nasal cavity at an approximate depth of 6 inches. The nasal swab was then rotated and removed. After collection of the sample, 6-inch nasal swabs were placed in buffered peptone water with 12% glycerol, drop frozen in liquid nitrogen directly after collection, and stored at -80˚C.

DNA extraction, amplification, and sequencing of database samples. For the KSU samples, a 1mL aliquot of sample was placed into LoBind tubes (Eppendorf, Hamburg, Germany) containing Nuclei Lysis Solution from the Promega Wizard Genomic DNA Purification Kit (Madison, WI, USA) and an isopropanol precipitation was performed on ice using a total of 600μL of 100% isopropanol. Samples were then centrifuged at 14,500 x *g* for 2 min. at 4˚C. A 600μL volume of cold Nuclei Lysis buffer was added to the pellet, vortexed, and incubated at 80˚C for 7 min. Samples were then transferred to QiaShredder (Qiagen, Hilden, Germany) tubes and centrifuged at 14,500 x *g* for 5 min. at 4˚C. Samples were then vortexed and incubated at 80˚C for 5 min. Following incubation at room temperature for 10 min, 3μL of RNAse solution (Promega Wizard Genomic DNA Purification Kit) was added to the sample, mixed via inversion, and incubated at 37˚C in a water bath for 30 min. Following incubation at room temperature for 10 min, 200μL of Protein Precipitation Solution (Promega Wizard Genomic DNA Purification Kit) was added to the sample, vortexed, and placed on ice for 5min. Samples were centrifuged at 14,500 x *g* for 3 min. at 4˚C, supernatant transferred to a new tube, and 600μL of 100% isopropanol added to the sample and mixed via inversion. Another round of centrifugation at 14,500 x *g* for 2 min. at 4˚C was completed, supernatant removed, and 600μL of 70% ethanol was added to the sample. A final centrifugation occurred at 14,500 x *g* for 2 min. and the 70% ethanol was removed via aspiration. The nucleic acid pellet was dried at room temperature for 15 min. and re-suspended in 50μL of DNA rehydration solution (Promega Wizard Genomic DNA Purification Kit). DNA quality was checked via electrophoresis in a 2% (w/v) agarose gel in TBE buffer (1X), and concentration was determined using a Nanodrop 1000 spectrophotometer (ThermoScientific, Wilmington, USA).

For the USMARC samples, total DNA was extracted from each swab using a commercial kit (PowerSoil DNA kit; Qiagen, Germantown, MD) as directed by the manufacturer, and initial DNA quantity was evaluated with a DNA spectrophotometer (DeNovix DS-11 FX Series; Wilmington, DE). Equal amounts of DNA from each swab were then pooled within collection year and time point of sampling (initial vaccination, preconditioning, weaning, after weaning) and were chosen for amplification of the 16S ribosomal RNA (rRNA) gene. There were 114 pools sequenced with long reads representing over 800 animals.

DNA library preparation was performed by polymerase chain reaction (PCR) amplification similar to Myer and colleagues [29]. PCR amplification and DNA library preparation of the V1–V8 hypervariable regions of the bacterial 16S rRNA gene were performed using universal primers 27F (`5'- AGAGTTTGATCCTGGCTCAG`) and 1392R (`5'-GACGGGCGGTGTGTAC`). Amplification consisted of 20 cycles with an annealing temperature of 58˚C and extension time of 90 seconds. Products were purified using AmPure® bead purification (Agencourt, Beverly MA), converted to libraries using Template Preparation Kit v1.0 (Pacific Biosciences), and quantified by the PicoGreen® dsDNA quantitation kit (Invitrogen, Carlsbad, CA) and by real-time PCR on the LightCycler 480 system (Roche, Mannheim, Germany). Libraries were sequenced on the Pacific Biosciences RSII instrument (Pacific Biosciences, Menlo Park, CA).

Bovine URT database sequence data analyses, development, and phylogenetic analyses. All sequencing data were curated using mothur (v1.39.5) [34]. The consensus fastq files were parsed so that scores of zero were interpreted as corresponding to an ambiguous base call. Also, because the consensus sequence can be generated in the forward and reverse complement orientations, proper orientation was verified. Sequences were then filtered for quality ($\geq$ Q30) using a rolling window approach (window = 30), homopolymers $>$ 8 were discarded, any sequences with ambiguous bases were removed, and sequences that contained read lengths shorter than 1350bp or larger than 1400 were removed. Alignment was conducted against the SEED SILVA database (v128) [2]. Pre-clustering of aligned sequences was conducted at diffs = 14. De novo detection of chimeras was conducted using UCHIME [35]. Classification of denoised reads was performed against greengenes (v13_8_99) [1] using a naïve Bayesian classifier at a cutoff of 80. All reads were clustered with VSEARCH at a 97% similarity threshold. Following clustering, a consensus classification of OTUs using a 50% consensus clustering threshold was performed. Then, a taxonomic database consisting of the most abundant sequence of each OTU was constructed. Rarefaction curves examining sequencing depth were constructed in the Quantitative Insights Into Microbial Ecology (QIIME) bioinformatics pipeline, version 1.9.1 [36]. Using the whole database alignment, Newick-formatted, approximately maximum-likelihood phylogenetic trees were calculated and built in QIIME using FastTree [37]. Trees were viewed using Interactive Tree of Life (iTOL) [38]. These sequences are available from the NCBI Sequence Read Archive (SRA Accession PRJNA548468).

Validation analyses

DNA amplification and sequencing. The 16S ribosomal RNA (rRNA) gene variable regions were amplified from additional pooled samples of the USMARC population. Three pooled samples were randomly chosen for sequencing and analyses. DNA was extracted as described above. Library preparation was performed using standard PCR (AccuPrime, Invitrogen, Carlsbad, CA) and primers that amplify variable regions 1 through 3 of the 16S rRNA gene [39]. Quality and quantity of the resulting 16S rRNA gene amplification was checked on the Fragment Analyzer (Advanced Analytical, Ankeny, IA) and then sequenced utilizing the MiSeq Illumina Sequencer (Illumina, San Diego, CA) with a MiSeq Reagent Kit v3 to generate 2x300 paired end reads. These sequences are also available from the NCBI Sequence Read Archive (SRA Accession PRJNA548468).

Sequence data analyses. All MiSeq sequencing data for validation analyses were curated using mothur (v1.39.5) [34]. Sequences were then filtered for quality ($\geq$ Q30) using a rolling window approach (window = 30), homopolymers $>$ 8 were discarded, any sequences with ambiguous bases were removed, and sequences that contained read lengths shorter than 300bp were removed. Alignment was conducted against the SEED SILVA database (v128). Pre-clustering of aligned sequences was conducted at diffs = 14. *De novo* detection of chimeras was conducted using UCHIME. Classification of denoised reads was performed against Greengenes (v13_8_99) [1] and SILVA (v128) (2) using a naïve Bayesian classifier at a cutoff of 80. For comparison, classification of denoised reads was also performed against the Bovine URT database. All reads were clustered with VSEARCH at a 97% similarity threshold. Following clustering, a consensus classification of OTUs using a 50% consensus clustering threshold was performed. Sequences and their resultant taxonomic classification that were successfully assigned by the Bovine URT database but unassigned against the Greengenes database were re-analyzed using the Bovine URT database. To accomplish this, unassigned Greengenes sequences were filtered from the FASTA and OTU tables to retain only those sequences that were unassigned. Utilizing the filtered FASTA and OTU tables, the classification of the filtered unassigned sequences was then performed against the Bovine URT database as described previously to ascertain differences in taxonomic assignment using the Bovine URT database. Phylum-level taxa were assessed for normality using the PROC UNIVARIATE procedure in SAS 9.4 (SAS Institute, Cary, NC). Normality was determined based on visual distribution of histograms and a Shapiro-Wilk statistic of ≥ 0.90. Those variables following a normal distribution were analyzed using one-way analysis of variance (ANOVA) for multiple independent groups [40] to compare among Bovine URT, SILVA, and Greengenes datasets. Those variables following a non-normal distribution were first ranked and then analyzed using a one-way ANOVA. Significance was determined at $P \leq 0.05$ and tendencies were considered from $0.05 < P \leq 0.10$ and the Benjamini–Hochberg method used for multiple-testing corrections [41]. Taxonomic composition correlation between the database pipelines was conducted using Pearson correlation, where *P*-values were calculated using a two-sided test of significance using a t-distribution.

Results

Summary of database composition

A summary of database statistics is included in Table 1. A total of 4,364,029 reads were used for Bovine URT database development following pipeline QC, with 170,730 unique sequences totaling 235,106,702 bp with an average length of 1,377 ± 36 bp. A large quantity of the database consists of singletons, totaling 150,723 reads. This is due to the inclusion of unique reads, including ones with only 1bp difference. As a means to evaluate the completeness of detection of bacterial species, a rarefaction curve of annotated species richness was constructed and resulted in adequate read depth as depicted by the asymptote (Fig 1). Clustering the reads using presently known species resulted in a total number of 2,111 observed species (S1 File). Overall, the Bovine URT database contains a small number of abundant species, but a large quantity of rare species.

Table 1. Bovine URT database statistics.

Metric	Value
Number of Sequences	170,730
Average Sequence Length (bp[1])	1,377 ± 36
Mean GC Content (%)	49 ± 4
Number of Species-Level OTU[2]	2,111
Number of Genus-Level OTU[2]	510

[1] BP: Base pairs

[2] OTU: Operational taxonomic units

https://doi.org/10.1371/journal.pone.0235498.t001

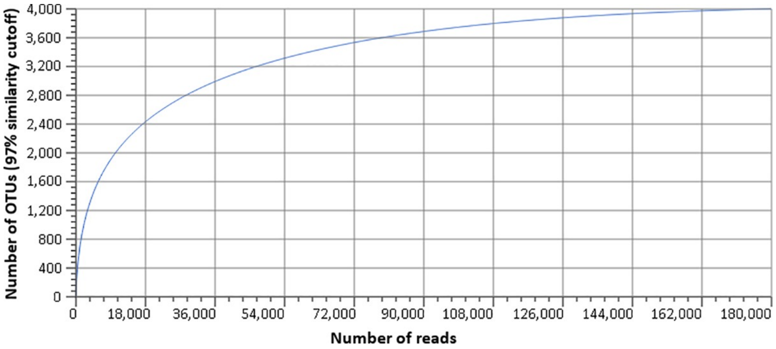

Fig 1. Rarefaction curve of annotated species richness within the Bovine URT bacterial 16S rDNA reference database.

https://doi.org/10.1371/journal.pone.0235498.g001

A detailed phylogenetic tree of the Bovine URT bacterial 16S database at the level of genus is displayed in Fig 2. A detailed taxonomic distribution of the reads at the level of phylum and genus within the database are listed in Fig 3. *Mycoplasma* was the dominant genus, accounting for over 57% of the reads. The only other genus observed at ≥1% included *Moraxella* (3.4%), *Pasteurella* (2.3%), *Ureaplasma* (2.1%), *Terrimonas* (1.6%), *Fusobacterium* (1.4%), *Clostridium* (1.2%), *Ruminococcus* (1.2%), *Microbacterium* (1.1%), and *Streptobacillus* (1.0%). *Mannheimia* was present at 0.7% (S2 File). Only 793 sequences were unable to be identified (0.4%).

Reference database performance

Fig 4 shows the comparison among taxonomic composition at the phylum level when analyzing short read 16S amplicons against the Greengenes and SILVA database in a de novo pipeline and against the Bovine URT database. The top phyla from the Greengenes analysis at ≥1% included Actinobacteria (27.6%), Firmicutes (23.7%), Bacteroidetes (20.7%), Tenericutes (13.8%), and Proteobacteria (6.1%) (S3 File). The top phyla from SILVA analysis at >1% included Actinobacteria (27.9%), Firmicutes (23.6%), Bacteroidetes (20.8%), Tenericutes (13.7%), and Proteobacteria (6.0%) (S3 File). The top phyla from the Bovine URT database analysis at ≥1% included Actinobacteria (27.9%), Firmicutes (23.8%), Bacteroidetes (21.0%), Tenericutes (14.2%), and Proteobacteria (6.1%) (S3 File). Unassigned taxa in the Bovine URT database analysis was 0.2% contrasted to the Greengenes and SILVA analyses, at 4.2% and 4.0% respectively (S3 File). The unassigned taxa were the only group at the phylum level that differed among the Greengenes (4.2 ± 0.21%), SILVA (4.1 ± 0.24%), or Bovine URT (0.2 ± 0.09%) databases ($P = 0.0025$). When contrasting taxonomic composition, specifically at the genus level, the Bovine URT compositional comparisons against the Greengenes and SILVA taxonomic compositions indicated statistically significant correlations ($r = 0.226$; $P < 0.001$ and $r = 0.154$; $P < 0.001$, respectively). All alpha diversity indices among the three analysis methods were not different (Fig 5). Many sequences and their resultant taxa were successfully assigned by the Bovine URT database but unassigned against the Greengenes database and included those of the order Clostridiales, the family Microbacteriaceae, and genera *Mycoplasma* and *Moraxella*, to name a few (S4 File).

Discussion

Over the past decade, high-throughput sequencing has dominated microbial ecology and permitted deep investigation into microbial community composition. As a result, numerous databases containing annotated microbial genomes and 16S amplicon reads of numerous

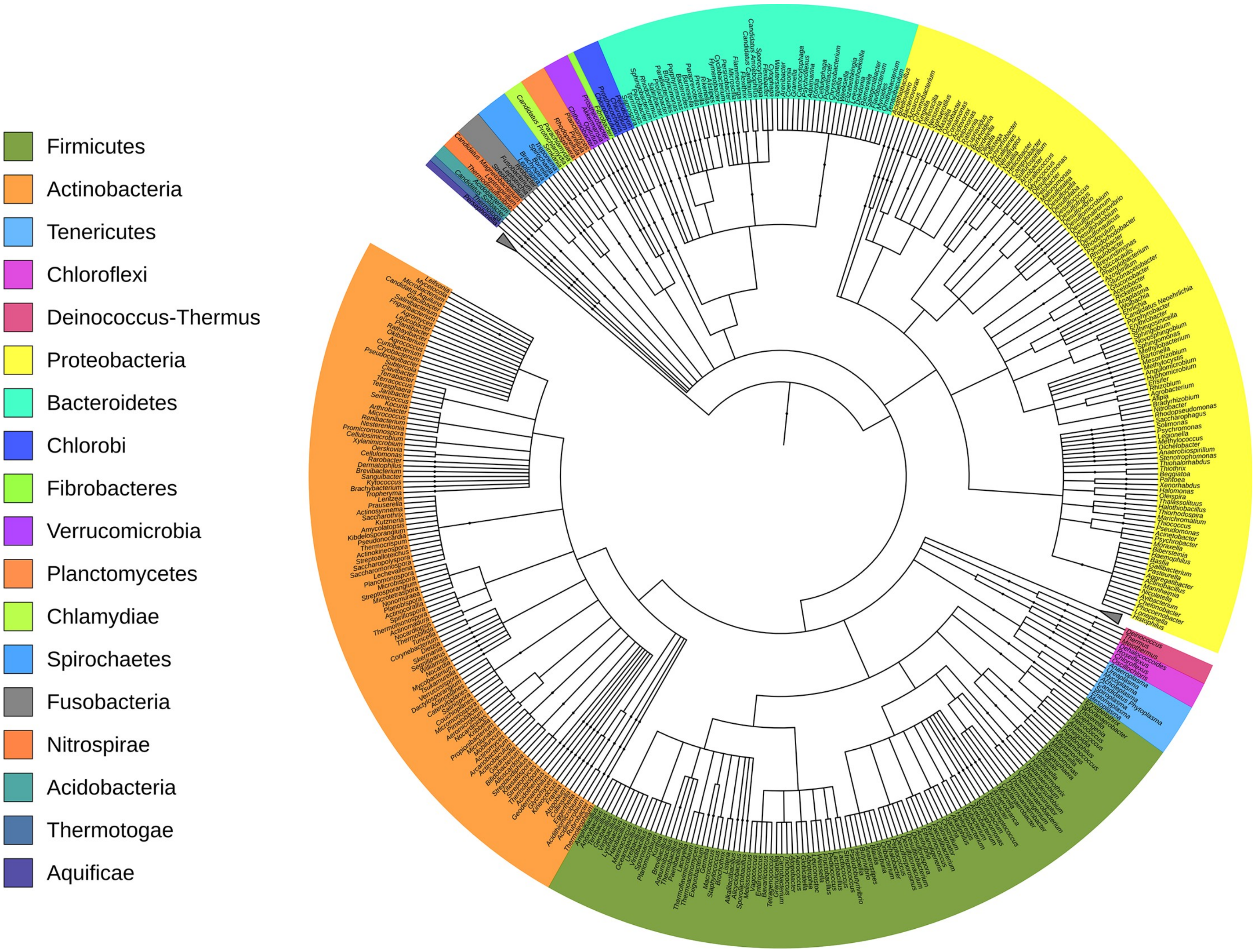

Fig 2. Circular maximum likelihood phylogenetic tree at the level of genus of the Bovine URT 16S rDNA reference database. Tree image was generated using ITOL. Genera are color-coded by phylum.

https://doi.org/10.1371/journal.pone.0235498.g002

microbial ecosystem origins have been constructed, including but not limited to RDP [3], Greengenes [1], SILVA [2], and GenBank [4]. In the typical 16S analysis pipeline, OTU analyses are utilized first, as they are powerful tools in that they do not require a reference database, and function by identifying similar sequences based on sequence identity. Using these databases, taxonomic-dependent methods are then used to assign taxonomic annotations to OTUs based on the OTU analysis. Consequently, the 16S analysis pipeline is limited by the reference database utilized in analyses. Taxonomic annotation of discovered and binned sequences from 16S rDNA read data is a critical step in the canonical 16S amplicon sequencing pipeline. Yet, the use of public databases containing microbial genomic data from numerous microbial ecosystems may not be specific or of high enough quality to properly analyze niche-specific microbial communities. Public 16S databases are popular, but are of variable quality [42, 43] due to their rapid growth and sequence types [44]. The issue is also greater for identification of short reads, commonly used in 16S analyses due to the great depth achieved. With short-read analysis, gaining information beyond the genus level and avoiding mismatches due to the variety of

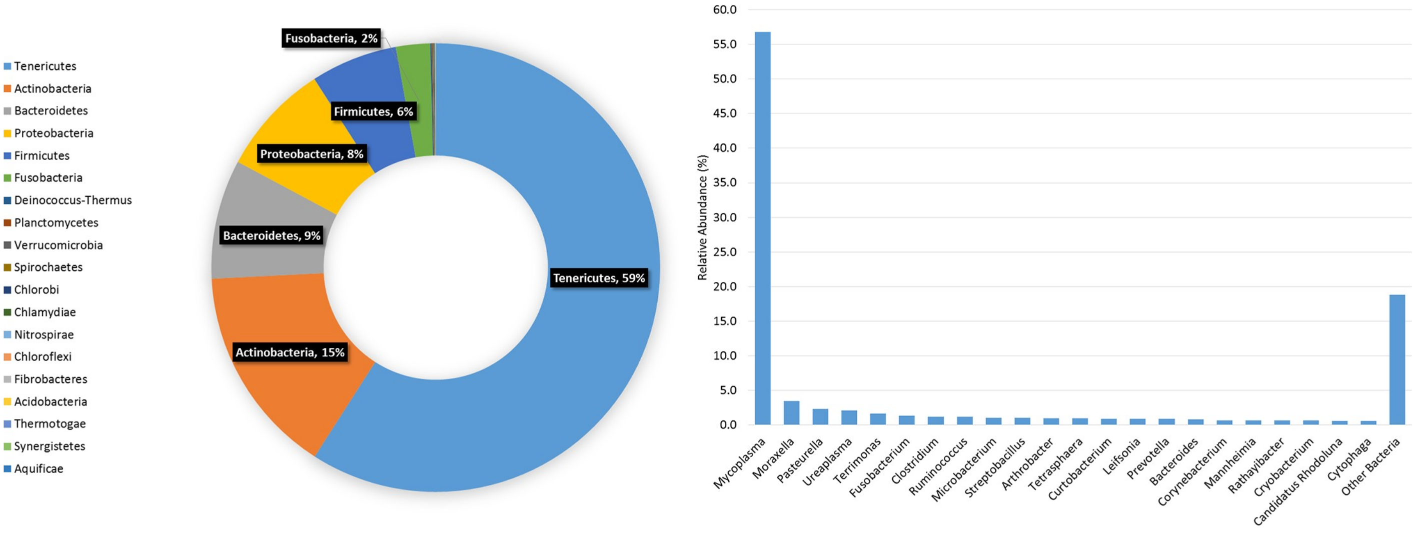

Fig 3. Bacterial taxonomic profile of total 16S rDNA gene sequences in the Bovine URT reference database at the phylum level (donut chart; left) and genus level (bar chart; right). Genus abundances classified by representation at ≥0.6% of total sequences.

https://doi.org/10.1371/journal.pone.0235498.g003

comparative lengths of sequence in the database is often problematic. It is often necessary to increase the taxonomic resolution of microbial community analyses, especially when determining pathology or examining welfare issues, as species-level function and dynamics are essential [45]. The objective of the study was to evaluate the impact of creating a niche-specific, near-full-length, 16S rRNA reference database on the ability to assign taxonomy from shorter partial sequence covering only one to three variable regions. This was completed by analyzing the URT of cattle, a niche with substantial sample to sample variability and relatively high community diversity. Importantly, a niche-specific custom database to be used as a reference also has the disadvantage of being circularly self-referenced and fail to identity taxa that were not present at its setup. These failures could be infrequent but, nevertheless, may be important for the specific habitat. To overcome this consideration for this evaluation, we had available a broad breadth of samples taken at varying locations, environments, breeds, US regions, and time points collected as part of studies on bovine respiratory disease.

The cattle utilized for the development of the Bovine URT database originated from numerous locations and were of varying health status. This variety provided a great resource for broad sampling of URT microbial diversity, as demonstrated in the database phylogenetic tree. Sequences from the phylum Tenericutes and its genus *Mycoplasma* dominated the dataset, which may be due in part to effects of antibiotic treatment of a portion of the animals, or to the stress of transport that places the animal at increased risk for disease (35). Consistent with this supposition, the *Mannheimia* genus was relatively abundant (0.7%). A variety of organisms commonly found in numerous other microbial ecosystems were also identified, expanding the scope of the Bovine URT database for BRD microbial analysis. Organisms in the database that are found in various microbial ecosystems include, but are not limited to, *Prevotella*, commonly found in gastrointestinal tracts [46], *Streptobacillus*, commonly found in mammalian disease [47], and *Clostridium*, whose species are found in a wide variety of hosts and have various functions and pathogenicity [46, 48–50]. In all, there were a total of 2,111 observed species, reflecting encounters with diverse environments by the cattle URT as they move from farm to feedlot. Indeed, this diversity reflects the broad sampling of animals across breeds and environments, an important factor in developing a comprehensive, custom database for the cattle

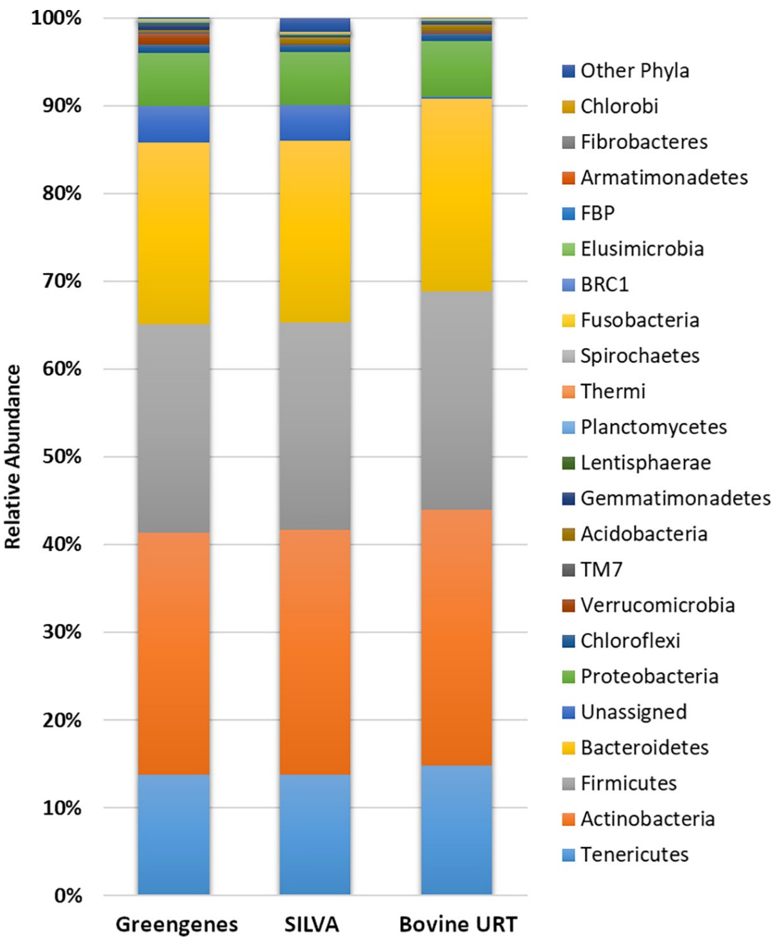

Fig 4. Relative abundance of bacterial phyla in the multiple diverse samples identified using the Greengenes, SILVA, and Bovine URT databases on small read 16S amplicon sequencing from the bovine URT.

https://doi.org/10.1371/journal.pone.0235498.g004

URT. The rarefaction curve of the sequences utilized in the Bovine URT database indicates that the level of sequencing depth was adequate to detect the majority of the microbial communities across the dataset. Taking into account the use of near full length 16S reads, deep sequencing depth, variable health statuses, and a large sample population of cattle, the constructed Bovine URT 16S rDNA database represents one of the more comprehensive microbial datasets to date with reference to BRD.

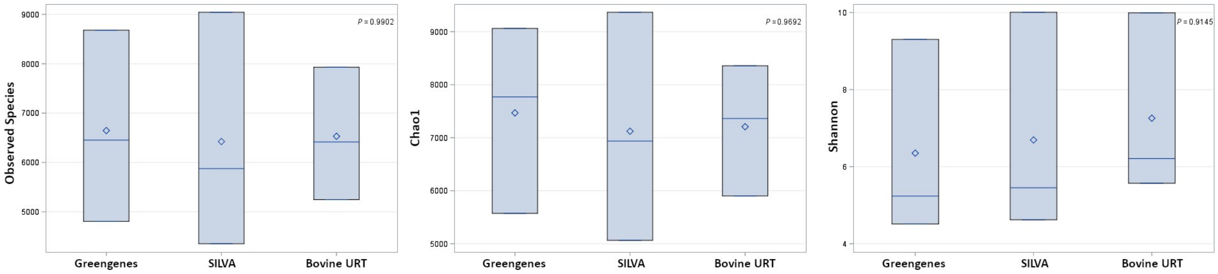

Fig 5. Boxplots of alpha diversity indices among Greengenes, SILVA, and Bovine URT database microbial profiling methods. Alpha diversity indices include A) observed species, B) Chao1, and C) Shannon diversity index.

https://doi.org/10.1371/journal.pone.0235498.g005

Numerous studies have suggested the use of niche-specific databases for 16S analysis of microbial profiling [12, 29, 51, 52]. In support of this concept, validation of the Bovine URT database was completed using short read amplicons from samples of the USMARC population against the long-read Bovine URT database, as well as against the SILVA and Greengenes databases. Interestingly, the profiling methods referencing the Bovine URT, SILVA, and Greengenes databases produced considerably similar microbial community profiles. Even given the observation of animal-to-animal variation within the long read data, the phylum and genus distribution did not vary greatly between methods, substantiating the utility of the Bovine URT database based on its similarity to the two public databases. However, there was a significant difference in the number of unassigned taxa, which was likely a result of greater classification accuracy. Indeed, these re-assigned taxa using the Bovine URT database likely accounted for some of the small differences observed among the analyses when using the Bovine URT database, such as the phyla Tenericutes, Firmicutes, and Bacteroidetes (S4 File). Nucleotide variation along the 16S rRNA gene is considerable, and short reads can only capture a fraction of this variability. Even when full length reads are analyzed, studies have determined that reads may be taxonomically similar at the genus level. Against the RDP database, one study found as many as 5.5% full-length reads similar at the genus level [53]. Given the variability of sequence length and variability of well-characterized species in public databases, short reads are likely matched to multiple fragments along longer reads of the 16S sequence, resulting in false positives or misclassifications cofounding analyses and exponentially augmenting species estimates, phylogenetic diversity, and taxonomic assignments [54, 55]. Mock community analyses with *de novo* OTU picking strategies have demonstrated inflated OTU counts and diversity [55, 56], and many of the spurious OTUs detected are present at low abundances. This concept was supported by the great discrepancy among analyses in the current study pertaining to the difference in unassigned taxa, as well as the number of genera identified among the three analyses, not to be confused with taxonomic composition and phylogenetic diversity. The Bovine URT analysis identified 550 genera, whereas the number of genera in the SILVA and Greengenes analyses were far greater at 981 and 748 respectively (S3 File). This discprepency is expected to be the result of overestimation due to misclassification to reference sets or low quality assignment [52]. Importantly, during development of the Bovine URT database, the taxonomy was assigned to the long-read sequences using Greengenes. However, this does not imply similar database functionality between the two databases. Once created, the Bovine URT database was unique from the Greengenes database. When short-read samples were mapped to both databases for validation, a difference in results would be expected, as there are different sequences between the databases which can result in the aforementioned differences identified in the current study.

There are common drawbacks and pitfalls inherent to many microbial community surveys. Amplification biases due to unequal matching performance of primers across different species are also considerations of additional sources of error. This can lead to major artifacts in the reported taxa proportions within the community. These primer biases have been reported in our previous work and the work of others [29, 30, 57]. However, in this study, this concern is limited with regard to Bovine URT database validation. We utilized a diverse set of microbial samples for database construction coupled with the same primer pairs among all replicates for validation analyses between two distinct databases. Bovine URT database validation using test sequences among different variable regions may output slightly different results [57], but is beyond the scope and intent of this study. Ultimately, the near-full length read database represents a resource to better interpret the data regardless of primers and hypervariable regions.

Longer reads offer greater potential to identify higher quality OTUs [29], and using these longer reads as a standardized reference is promising to capture accurate microbial composition and diversity. The abundance of unassigned taxa among the methods further supports this theory, and encourages the use of reference databases. The significantly reduced unassigned taxa alludes to a more thorough community composition analysis when assigning short reads to a long read reference database. Moreover, assigning short reads to taxa from a database containing classified long reads that are analyzed and included specifically for the ecosystem studied allowed for a reduction in redundancy and reduced the likelihood that taxa went unassigned, as only 793 of over 170,000 sequences were unable to be identified (0.4%) in the Bovine URT database.

Conclusions

We were able to demonstrate greater bacterial assignment and reduced unassigned reads using the Bovine URT database developed from near-full-length 16S reads for short-read 16S analyses of URT bacteria in cattle. The analyses utilizing the Bovine URT reference database were also able to identify bacteria previously reported in bovine BRD research demonstrating its effective practice [23–25, 33]. Accurate taxonomic annotation of sequencing reads is imperative for profiling the composition of microbiomes. This is also critical in application, as identifying microbiome dysbioses or pathogens permits researchers with greater opportunities to define the complex interactions in multifactorial disorders, such as BRD. The development of this Bovine URT reference database provides researchers a tool with considerable improvements in performance for the analysis of bacterial communities, and in this study, those of the URT relating to cattle diagnosed with BRD.

Supporting information

S1 File. Species in the reference Bovine URT database.
(XLSX)

S2 File. Genera in the reference Bovine URT database.
(XLSX)

S3 File. Phyla and genera compositions from the Greengenes, SILVA, and Bovine URT analyses.
(XLSX)

S4 File. Taxa that could be assigned by the Bovine URT but not the Greengenes database.
(XLSX)

Acknowledgments

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. USDA is an equal opportunity provider and employer.

Author Contributions

Conceptualization: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Timothy P. L. Smith.

Data curation: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Madison T. Henniger, Brooke A. Clemmons, Timothy P. L. Smith.

Formal analysis: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Brooke A. Clemmons, Timothy P. L. Smith.

Investigation: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Timothy P. L. Smith.

Methodology: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Madison T. Henniger, Timothy P. L. Smith.

Project administration: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Timothy P. L. Smith.

Resources: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Timothy P. L. Smith.

Supervision: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Timothy P. L. Smith.

Validation: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Madison T. Henniger, Timothy P. L. Smith.

Visualization: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, Timothy P. L. Smith.

Writing – original draft: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Michael D. Apley, Madison T. Henniger, Brooke A. Clemmons, Timothy P. L. Smith.

Writing – review & editing: Phillip R. Myer, Tara G. McDaneld, Larry A. Kuehn, Keith D. Dedonder, Michael D. Apley, Sarah F. Capik, Brian V. Lubbers, Gregory P. Harhay, Dayna M. Harhay, John W. Keele, Madison T. Henniger, Brooke A. Clemmons, Timothy P. L. Smith.

References

1. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology. 2006; 72(7):5069–72. https://doi.org/10.1128/AEM.03006-05 PMID: 16820507
2. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2012; 41 (D1):D590–D6.
3. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research. 2013; 42(D1):D633–D42.
4. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, Pruitt KD, et al. GenBank. Nucleic acids research. 2017.
5. Balvočiūtė M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? BMC genomics. 2017; 18(2):114.
6. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic acids research. 2007; 35(21):7188–96. https://doi.org/10.1093/nar/gkm864 PMID: 17947321

7. Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic acids research. 2016; 44(11):5022–33. https://doi.org/10.1093/nar/gkw396 PMID: 27166378

8. Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases. PeerJ. 2018; 6:e5030. https://doi.org/10.7717/peerj.5030 PMID: 29910992

9. Silva. SILVA 132 Release Information [Available from: https://www.arb-silva.de/documentation/release-132/.

10. Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. PloS one. 2015; 10(2):e0116106. https://doi.org/10.1371/journal.pone.0116106 PMID: 25646627

11. Vilo C, Dong Q. Evaluation of the RDP classifier accuracy using 16S rRNA gene variable regions. Metagenomics. 2012; 1:1–5.

12. Ritari J, Salojärvi J, Lahti L, de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC genomics. 2015; 16(1):1056.

13. Griffin D, Chengappa M, Kuszak J, McVey DS. Bacterial pathogens of the bovine respiratory disease complex. Veterinary Clinics: Food Animal Practice. 2010; 26(2):381–94.

14. USDA. Types and Costs of Respiratory Disease Treatments in U.S. Feedlots: U.S. Department of Agriculture, Animal and Plant Health Inspection Service, Veterinary Services, Centers for Epidemiology and Animal Health; 2013.

15. Virtala A-M, Mechor G, Gröhn Y, Erb H. The effect of calfhood diseases on growth of female dairy calves during the first 3 months of life in New York State. Journal of dairy science. 1996; 79(6):1040–9. https://doi.org/10.3168/jds.S0022-0302(96)76457-3 PMID: 8827469

16. Waltner-Toews D, Martin S, Meek A. The effect of early calfhood health status on survivorship and age at first calving. Canadian Journal of Veterinary Research. 1986; 50(3):314. PMID: 3742366

17. Taylor J, Doyle D, Blackall P, Confer A. Use of REP-PCR and 16S rRNA gene sequencing for comparison of Mannheimia haemolytica isolates obtained from fatal cases of bovine respiratory disease in the USA and Australia. Australian veterinary journal. 2014; 92(1–2):15–23. https://doi.org/10.1111/avj.12137 PMID: 24471877

18. Cornelissen JB, Bree FM, Wal FJ, Kooi EA, Koene MG, Bossers A, et al. Mycoplasma detection by triplex real-time PCR in bronchoalveolar lavage fluid from bovine respiratory disease complex cases. BMC veterinary research. 2017; 13(1):97. https://doi.org/10.1186/s12917-017-1023-6 PMID: 28390431

19. Allen J, Viel L, Bateman K, Rosendal S, Shewen P, Physick-Sheard P. The microbial flora of the respiratory tract in feedlot calves: associations between nasopharyngeal and bronchoalveolar lavage cultures. Canadian journal of veterinary research. 1991; 55(4):341. PMID: 1790489

20. Rice J, Carrasco-Medina L, Hodgins D, Shewen P. Mannheimia haemolytica and bovine respiratory disease. Animal Health Research Reviews. 2007; 8(2):117–28. https://doi.org/10.1017/S1466252307001375 PMID: 18218156

21. Taylor J, Holland B, Step D, Payton M, Confer A. Nasal isolation of Mannheimia haemolytica and Pasteurella multocida as predictors of respiratory disease in shipped calves. Research in veterinary science. 2015; 99:41–5. https://doi.org/10.1016/j.rvsc.2014.12.015 PMID: 25599936

22. Johnston D, Earley B, Cormican P, Murray G, Kenny DA, Waters SM, et al. Illumina MiSeq 16S amplicon sequence analysis of bovine respiratory disease associated bacteria in lung and mediastinal lymph node tissue. BMC veterinary research. 2017; 13(1):118. https://doi.org/10.1186/s12917-017-1035-2 PMID: 28464950

23. Holman DB, Timsit E, Amat S, Abbott DW, Buret AG, Alexander TW. The nasopharyngeal microbiota of beef cattle before and after transport to a feedlot. BMC microbiology. 2017; 17(1):70. https://doi.org/10.1186/s12866-017-0978-6 PMID: 28330466

24. McDaneld TG, Kuehn LA, Keele JW. Evaluating the microbiome of two sampling locations in the nasal cavity of cattle with bovine respiratory disease complex (BRDC)1. Journal of Animal Science. 2018; 96 (4):1281–7.

25. McMullen C, Orsel K, Alexander TW, van der Meer F, Plastow G, Timsit E. Evolution of the nasopharyngeal bacterial microbiota of beef calves from spring processing to 40 days after feedlot arrival. Veterinary microbiology. 2018; 225:139–48. https://doi.org/10.1016/j.vetmic.2018.09.019 PMID: 30322526

26. McDaneld TG, Kuehn LA, Keele JW. Microbiome of the upper nasal cavity of beef calves prior to weaning12. Journal of Animal Science. 2019; 97(6):2368–75. https://doi.org/10.1093/jas/skz119 PMID: 31144716

27. Ross EM, Moate PJ, Bath CR, Davidson SE, Sawbridge TI, Guthridge KM, et al. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. BMC genetics. 2012; 13 (1):53.

28. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database. 2010; 2010.

29. Myer PR, Kim M, Freetly HC, Smith TP. Evaluation of 16S rRNA amplicon sequencing using two next-generation sequencing technologies for phylogenetic analysis of the rumen bacterial community in steers. Journal of microbiological methods. 2016; 127:132–40. https://doi.org/10.1016/j.mimet.2016.06.004 PMID: 27282101

30. Myer PR, Kim M, Freetly HC, Smith TP. Metagenomic and near full-length 16S rRNA sequence data in support of the phylogenetic analysis of the rumen bacterial community in steers. Data in brief. 2016; 8:1048–53. https://doi.org/10.1016/j.dib.2016.07.027 PMID: 27508263

31. DeDonder K, Apley M, Li M, Gehring R, Harhay D, Lubbers B, et al. Pharmacokinetics and pharmacodynamics of gamithromycin in pulmonary epithelial lining fluid in naturally occurring bovine respiratory disease in multisource commingled feedlot cattle. Journal of veterinary pharmacology and therapeutics. 2016; 39(2):157–66. https://doi.org/10.1111/jvp.12267 PMID: 26441021

32. Schiermiester LN, Thallman RM, Kuehn LA, Kachman SD, Spangler ML. Estimation of breed-specific heterosis effects for birth, weaning, and yearling weight in cattle1. Journal of Animal Science. 2015; 93 (1):46–52. https://doi.org/10.2527/jas.2014-8493 PMID: 25568356

33. Holman DB, Timsit E, Alexander TW. The nasopharyngeal microbiota of feedlot cattle. Scientific reports. 2015; 5:15557. https://doi.org/10.1038/srep15557 PMID: 26497574

34. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology. 2009; 75(23):7537–41. https://doi.org/10.1128/AEM.01541-09 PMID: 19801464

35. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011; 27(16):2194–200. https://doi.org/10.1093/bioinformatics/btr381 PMID: 21700674

36. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods. 2010; 7(5):335. https://doi.org/10.1038/nmeth.f.303 PMID: 20383131

37. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one. 2010; 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

38. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research. 2016; 44(W1):W242–W5. https://doi.org/10.1093/nar/gkw290 PMID: 27095192

39. Myer PR, Smith TPL, Wells JE, Kuehn LA, Freetly HC. Rumen Microbiome from Steers Differing in Feed Efficiency. PLOS ONE. 2015; 10(6):e0129174. https://doi.org/10.1371/journal.pone.0129174 PMID: 26030887

40. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017; 5(1):27. https://doi.org/10.1186/s40168-017-0237-y PMID: 28253908

41. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995:289–300.

42. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Applied and environmental microbiology. 2005; 71(12):7724–36. https://doi.org/10.1128/AEM.71.12.7724-7736.2005 PMID: 16332745

43. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Applied and environmental microbiology. 2006; 72(9):5734–41. https://doi.org/10.1128/AEM.00556-06 PMID: 16957188

44. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2015; 44(D1):D733–D45. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

45. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017; 550(7674):61. https://doi.org/10.1038/nature23889 PMID: 28953883

46. Myer P, Freetly H, Wells J, Smith T, Kuehn L. Analysis of the gut bacterial communities in beef cattle and their association with feed intake, growth, and efficiency. Journal of animal science. 2017; 95 (7):3215–24. https://doi.org/10.2527/jas.2016.1059 PMID: 28727105

47. Fornefett J, Krause J, Klose K, Fingas F, Hassert R, Eisenberg T, et al. Comparative analysis of clinics, pathologies and immune responses in BALB/c and C57BL/6 mice infected with Streptobacillus moniliformis. Microbes and infection. 2018; 20(2):101–10. https://doi.org/10.1016/j.micinf.2017.10.001 PMID: 29024796

48. Tariq R, Singh S, Gupta A, Pardi DS, Khanna S. Association of gastric acid suppression with recurrent Clostridium difficile infection: a systematic review and meta-analysis. JAMA internal medicine. 2017; 177(6):784–91. https://doi.org/10.1001/jamainternmed.2017.0212 PMID: 28346595

49. Brüggemann H, Bäumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, et al. The genome sequence of Clostridium tetani, the causative agent of tetanus disease. Proceedings of the National Academy of Sciences. 2003; 100(3):1316–21.

50. Weimer PJ, Stevenson DM. Isolation, characterization, and quantification of Clostridium kluyveri from the bovine rumen. Applied microbiology and biotechnology. 2012; 94(2):461–6. https://doi.org/10.1007/s00253-011-3751-z PMID: 22159841

51. Yoon S-H, Ha S-M, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International journal of systematic and evolutionary microbiology. 2017; 67(5):1613–7. https://doi.org/10.1099/ijsem.0.001755 PMID: 28005526

52. Franzén O, Hu J, Bao X, Itzkowitz SH, Peter I, Bashir A. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. Microbiome. 2015; 3(1):43.

53. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology. 2007; 73 (16):5261–7. https://doi.org/10.1128/AEM.00062-07 PMID: 17586664

54. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018; 6(1):90. https://doi.org/10.1186/s40168-018-0470-z PMID: 29773078

55. Edgar RC. Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. PeerJ. 2017; 5:e3889. https://doi.org/10.7717/peerj.3889 PMID: 29018622

56. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. BMC microbiology. 2016; 16(1):123. https://doi.org/10.1186/s12866-016-0738-z PMID: 27342980

57. Rosselli R, Romoli O, Vitulo N, Vezzi A, Campanaro S, de Pascale F, et al. Direct 16S rRNA-seq from bacterial communities: a PCR-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. Scientific Reports. 2016; 6:32165. https://doi.org/10.1038/srep32165 PMID: 27577787