Impact of DNA sequencing and analysis methods on 16S rRNA gene bacterial community analysis in dairy products
Zhengyao Xue\textsuperscript{a}, Mary E Kable\textsuperscript{a,*}, and Maria L Marco\textsuperscript{a#}

\textsuperscript{a}Department of Food Science & Technology, University of California, Davis

Running title: Dairy microbiota analysis comparisons

#Address correspondence to Maria L. Marco, mmarco@ucdavis.edu.

*Present address: US Department of Agriculture, Western Human Nutrition Research Center, Davis, CA

Word count for the abstract: 242
Word count for the importance: 139
Word count for the text: 4431
Abstract

DNA sequencing and analysis methods were compared for 16S rRNA V4 PCR amplicon and gDNA mock communities encompassing nine bacterial species commonly found in milk and dairy products. The communities were examined using Illumina MiSeq and Ion Torrent PGM DNA sequencing methods followed by the QIIME 1 (UCLUST) and Divisive Amplicon Denoising Algorithm 2 (DADA2) data analysis pipelines including taxonomic comparisons to the Greengenes and Ribosomal Database Project (RDP) databases. Examination of the PCR amplicon mock community with these methods resulted in Operation Taxonomy Units (OTUs) and Amplicon Sequence Variants (ASVs) that ranged from a low of 13 to high of 118 and were dependent on the DNA sequencing method and read assembly step. The elevated numbers of OTUs and ASVs included assignments to spurious taxa as well as sequence variants of the nine species included in the mock community. Comparisons between the gDNA and PCR amplicon mock communities showed that combining gDNA from the different strains prior to PCR resulted in up to 8.9-fold greater numbers of spurious OTUs and ASVs. However, the DNA sequencing method and initial data assembly steps conferred the largest effects on predictions of bacterial diversity, independent of the mock community type (PCR amplicon or gDNA; Bray-Curtis $R^2 = 0.88$ and weighted Unifrac, $R^2 = 0.32$). Overall, DNA sequencing performed with the Ion Torrent PGM and analyzed with DADA2 and the Greengenes database resulted in the most accurate predictions of the mock community phylogeny, taxonomy, and diversity.
Importance

Validated methods are urgently needed to improve DNA-sequence based assessments of complex bacterial communities. In this study, we used 16S rRNA PCR amplicon and gDNA mock community standards, consisting of nine, dairy-associated bacterial species, to evaluate the most commonly applied 16S rRNA marker gene DNA sequencing and analysis platforms used in evaluating dairy and other bacterial habitats. Our results show that bacterial metataxonomic assessments are largely dependent on the DNA sequencing platform and read curation method used. DADA2 improved sequence annotation compared with QIIME 1, and when combined with the Ion Torrent PGM DNA sequencing platform and the Greengenes database for taxonomic assignment, the most accurate representation of the dairy mock community standards was reached. This approach will be useful for validating sample collection and DNA extraction methods and ultimately investigating bacterial population dynamics in milk and dairy-associated environments.
Introduction

Advancements in massively-parallel, DNA sequencing technologies have resulted in a dramatic increase in knowledge of the microorganisms found in natural environments, food systems, and the human body. 16S rRNA gene amplicon sequencing, in particular, has been a cornerstone for investigating bacterial diversity and phylogeny. This approach has enabled the simultaneous identification of the majority of bacteria in complex microbial communities. Although analysis of 16S rRNA gene diversity has provided significant new perspectives on bacterial habitats, there remain challenges to sample preparation, DNA sequencing, and data analysis approaches for ensuring accurate measurements of bacterial populations.

To address these issues, recent studies have compared sample collection methods (1-3) and storage conditions (2, 4-9). These studies generally showed that differences in bacterial composition caused by those methodological alterations are relatively minor compared to inter-sample variation (1, 3, 5-9). DNA extraction methods, on the other hand, can result in major changes to estimates of bacterial proportions between Gram-positive and Gram-negative bacteria which are more or less difficult to lyse (4, 7, 10-15). Moreover, PCR can also introduce bias depending on the DNA polymerase (16), number of cycles (17) and variable region of the 16S rRNA gene being compared (4, 18, 19).

DNA sequencing platforms, including 454 pyrosequencing, Illumina, Ion Torrent and Pacific Biosciences have also been shown to cause variation in the final community assessments (4, 18-21). However, some of studies also employed different post-run strategies (4, 20), rendering the relative influence of DNA sequencing and read assembly and trimming unclear on the outcomes. Moreover, data analysis methods, especially read clustering approaches (i.e. generating representative sequences), are known to have a significant impact on the
interpretation of bacterial composition (21-27). Reference-based sequence clustering tends to underestimate community diversity because it is limited by the sequences included in the reference database (21, 23, 27). *De novo* sequence clustering can recover more sequence variants than reference-based methods, but can also result in unstable Operational Taxonomic Units (OTUs) between projects that are composed of different sequences with each clustering iteration at the 97% nucleotide sequence identity threshold (28, 29). As a result, there is now effort to move away from OTU based methods towards DNA sequences that represent single nucleotide variation (30, 31). One of the amplicon sequence variants (ASVs) clustering methods is DADA2 (Divisive Amplicon Denoising Algorithm 2), which builds a quality-based model for filtering error and identifying variation in 16S rRNA gene sequences (26).

Herein, we sought to compare different DNA sequencing and data analysis strategies for the capacity to accurately detect the composition of a mock bacterial community consisting of nine species commonly found in milk. To eliminate biases introduced by sample type and DNA extraction method, we employed two different mock communities consisting of either organism-specific PCR amplicons or purified genomic DNA (gDNA). This approach allowed us to compare the performance of two popular benchtop DNA sequencers (Illumina MiSeq and Ion Torrent PGM), sequence read assembly, and OTU/ASV analysis methods.

**Results**

Comparison of different DNA sequence analysis methods for 16S rRNA V4 region reads generated with the Illumina MiSeq and Ion Torrent PGM. A mock community was prepared by combining 16S rRNA V4 region PCR amplicons from nine bacterial strains in equimolar quantities prior to DNA sequencing on Ion Torrent PGM and Illumina MiSeq instruments.
Sequences were either assembled (Illumina MiSeq) or maintained as single-end (Illumina MiSeq and Ion Torrent PGM) reads. A low percentage of reads were identified as chimeras (0 to 1.4%) (Table S1) and the remaining reads were analyzed using QIIME 1 (UCLUST) or DADA2 pipelines for OTU or ASV identification using the Greengenes (version 13.8) and RDP (version GOLD for QIIME 1 and version 11.5 for DADA2) reference databases.

**Illumina MiSeq paired-end assemblies.** QIIME 1 analysis of Illumina MiSeq paired-end, assembled reads with recommended parameters (32) resulted in total OTU numbers that were at least 4.2-fold greater than the expected nine OTUs encompassing strains included in the mock community (Table 2). When the Greengenes database was used for OTU alignment, 85 OTUs were identified. The majority of those OTUs (65) were assigned to taxa included in the mock community. Although the numbers of OTUs varied for each taxon, a single OTU representative encompassed the majority (> 60%) of reads for each of the mock community members (Table 2). For example, out of nine *Staphylococcus* OTUs identified with Greengenes, 99% of the reads were represented by one OTU. The remaining 20 OTUs identified with Greengenes were either designated as taxa that were not included in the mock community or were designated only to the order level. When the RDP database was used as the reference database for QIIME 1 analysis, the total OTUs decreased to 65, largely due to the reduction in spurious “other” assignments (Table 2). However, the overall OTU number remained considerably higher than the expected nine OTUs based on the mock community composition.

The numbers of ASVs identified with DADA2 from Illumina paired-end assemblies were lower than the numbers of OTUs assigned with QIIME 1 (Table 2). Because DADA2 assigns ASVs independently from taxonomic reference databases, total ASV numbers were the same using both Greengenes and RDP. The only distinction was that one *Clostridiaceae* ASV and both
Escherichia ASVs identified using Greengenes were designated as Clostridium and Enterobacteriaceae in RDP (Table 2).

**Illumina MiSeq unassembled, single-end reads.** Without read assembly, the QIIME 1 pipeline resulted in 68 and 36 total OTUs with the Greengenes and RDP databases, respectively (Table 3). These OTU numbers were lower compared to the paired-end assemblies (Table 2). DADA2, on the other hand, resulted in a slightly higher number of ASVs (40 ASVs) than the paired-end assemblies (38 ASVs) (Table 2 and Table 3). More reads were regarded as “other” taxa, and this result was most likely due to the shorter lengths of the unassembled, single-end MiSeq reads. Between the two reference databases, Greengenes resulted in more accurate taxonomic assignments with DADA2. Two Bacillales ASVs and one Clostridiales ASV that were included in the “other” ASV category by RDP were assigned as Bacillus and Clostridiaceae with Greengenes. Moreover, the Escherichia/Shigella ASV in RDP was unambiguously allotted to the Escherichia genus by Greengenes (Table 3).

**Ion Torrent PGM reads.** The application of QIIME 1 with the Greengenes database to the Ion Torrent reads resulted in the highest number of OTUs out of any of the methods applied (Table 4). The use of RDP in QIIME 1 also yielded high OTU numbers, comparable to those found for the paired-end Illumina MiSeq assemblies (Table 4). Conversely, DADA2 resulted in only 13 ASVs (Table 4). The 4 additional ASVs compared to expected were created due to errors in the homopolymer regions (Fig. S1) and the 13 ASVs were distributed across the 9 bacterial taxa included in the mock community, with the exception of one ASV with ambiguous taxonomy (Bacillales) identified using RDP, which was identified as Bacillus using Greengenes. Greengenes also improved the assignment of E. coli to the genus level as opposed to the family level in RDP (Table 4).
For each of the three DNA sequencing/read curation methods tested, DADA2 assigned fewer ASVs per taxon and resulted in fewer spurious ASVs than QIIME 1 (UCLUST) assigned OTUs except in Illumina single-end results analyzed with RDP database. DADA2 taxonomic identification was more specific with the Greengenes than the RDP database. Therefore, the combined DADA2/Greengenes approach was used for the subsequent analyses described below.

Assessments of the gDNA mock community were altered depending on DNA sequencing platform. A gDNA mock community was prepared by mixing equal quantities of gDNA from the nine milk-associated bacterial species prior to barcoded 16S rRNA V4 region PCR amplification. The PCR products were then used for sequencing on either the Illumina MiSeq or Ion Torrent PGM followed by analysis with the DADA2/Greengenes method. More chimeras were found for the gDNA mock community (ranging from 0.3 to 4.6 %) than the PCR amplicon mock community (Table S1). However, except for the known variation in platform-dependent read lengths, nucleotide sequences of the most abundant ASVs assigned to each of the nine mock community species were identical between the Illumina MiSeq (single- and paired-ends) and Ion Torrent PGM platforms (Illumina MiSeq paired-end assembly, Fig. S2; Illumina MiSeq single-end reads, Fig. S3; and Ion Torrent PGM reads, Fig. S4). Nucleotide sequence alignments of those ASVs to the corresponding ASVs identified from the PCR amplicon mock community also showed 100% nucleotide sequence conservation (Illumina MiSeq paired-end assembly, Fig. S2; Illumina MiSeq single-end reads, Fig. S3; and Ion Torrent PGM reads, Fig. S4).

For both Illumina MiSeq assembled and unassembled (single-end) reads, the gDNA mock community resulted in high numbers of ASVs (Table 5). These numbers were higher than found for the PCR amplicon mock community (Table 2 and Table 3). This result was primarily
due to the higher quantities of spurious ASVs (e.g. *Anaeroplasma*, *Bacteroides*, *Desulfovibrio* and *Prevotella*) (Table S2) present at low proportions. Conversely, the same number of 13 ASVs were found for the gDNA and PCR amplicon mock communities when the Ion Torrent PGM was used (Table 5).

The alpha diversity indices Shannon Index and Chao1 of the gDNA mock community were elevated compared to the PCR amplicon mock community for both paired-end and single-end Illumina MiSeq results (Fig. 1). Comparisons to *in silico* reconstructions of the mock communities showed that the Shannon Index and Chao1 were also significantly higher than expected for both Illumina methods. For the PCR amplicon mock community, both indices were significantly higher than expected (Fig. 1A and C). However, only Shannon, and not Chao1 reached significance for the gDNA mock community (Fig. 1B and D). On the other hand, when the Ion Torrent PGM was used, the Shannon Index and Chao1 of the PCR amplicon and gDNA mock communities resembled expected values (Fig. 1).

**Ion Torrent PGM sequencing with the DADA2/Greengenes method resulted in more accurate representations of the gDNA and PCR amplicon mock communities.** DNA sequencing approaches were next compared for their capacity to yield the expected beta diversity and proportions of bacterial taxa included in the two mock communities. According to UPGMA hierarchical clustering of Bray-Curtis dissimilarity metrics, results from the three DNA sequencing approaches (Illumina MiSeq paired-end assembly, single-end, and Ion Torrent PGM) formed separate clusters away from the expected bacterial composition, independent of the gDNA or 16S rRNA PCR amplicon community type (Fig. 2A). Conversely, UPGMA of the weighted Unifrac distance metrics, clustered the sequences according to mock community type.
(Fig. 2B). These comparisons showed that the gDNA mock communities sequenced with the Ion Torrent PGM were the most similar to theoretical (expected) proportions. No single method was found best suited for representing the PCR amplicon mock community (Fig. 2B). To assess whether the use of DADA2/Greengenes influenced this outcome, the other data analysis methods were compared and it was found that the DNA sequencing platform used was consistently influential on mock community beta-diversity (e.g. QIIME 1 with the Greengenes database is shown in Fig. S5).

We next compared the relative abundances of individual taxa across the three DNA sequencing approaches. For the 16S rRNA PCR amplicon mock community, the proportions of most bacterial species were mostly not significantly altered compared to expected theoretical values. Exceptions to this finding were the reduced proportions of *Enterococcaceae* and *Enterococcus* found for Illumina paired-end assemblies and *Streptococcus* for both Illumina MiSeq methods as well as the Ion Torrent PGM platform (Fig. 3). For the gDNA mock community, the proportions of *Escherichia* and *Streptococcus* were significantly different from expected for all three DNA sequencing platforms. The proportions of *Pseudomonas* were also significantly lower than expected for the single- and paired-end assemblies from the Illumina MiSeq, and the proportions of *Bacillus* and *Lactococcus* were also altered for the paired-end assemblies. Lastly, there were higher proportions of “other” taxa for both Illumina MiSeq methods (Fig. 3). Overall, even though DNA sequencing with the Ion Torrent PGM combined with DADA2/Greengenes analysis did not completely provide the expected bacterial composition, this approach resulted in the most accurate representations of the bacteria and their proportions in both gDNA and PCR amplicon mock communities.
Discussion

By comparing DNA sequencing methods, analysis algorithms, and reference databases using dairy relevant bacterial DNA (PCR amplicon and gDNA) mock communities, we found that the DADA2/Greengenes data analysis methods with the Ion Torrent PGM yielded the most accurate interpretations of the 16S rRNA V4 variable region of dairy mock communities relative to the other methods (Illumina MiSeq, QIIME 1, RDP) tested.

Application of DADA2 and QIIME 1 (UCLUST) analysis pipelines to the same 16S rRNA gene data showed that DADA2 assigned fewer total and spurious OTUs/ASVs than QIIME 1 even with stringent filtering (32). Because read length was kept consistent within each DNA sequencing and data assembly platforms, platform-specific differences in OTU/ASV numbers were mainly derived from the core algorithms used for filtering and clustering representative sequences. The DADA2 core algorithm includes the error-rate-based denoising, the isBimeraDenovo chimera identification, and the ASV inference (26). In QIIME 1, the core analysis includes the USEARCH chimera identification and OTU picking strategy (33).

Comparison of OTUs and ASVs using the DADA2 and QIIME 1 pipelines, respectively, also showed that the DADA2 analysis pipeline was able to assign ASVs to more specific taxonomic levels (genus) than QIIME 1. This could be the result of the different taxonomy classifiers employed by DADA2 (RDP’s naive Bayesian classifier) and QIIME 1 (UCLUST classifier) (34).

OTU and ASV taxonomy assignments were also compared with consideration to 16S rRNA gene reference databases. Results from the different combinations of analysis methods and reference databases showed that the majority of OTUs/ASVs detected were representatives of bacterial taxa included in the 16S rRNA PCR amplicon and gDNA mock communities. Each bacterial species was represented by at least a single OTU/ASV. Additional OTUs/ASVs were
largely due to low abundance sequence variants. DNA sequences of the predominant
ASVs/OTUs were 100% identical between gDNA and PCR amplicon mock communities, further
supporting the precision of the technique. When QIIME 1 was applied, the RDP gold database
(35) yielded lower numbers of total OTUs than found with Greengenes 13.8, independent of
whether the Illumina MiSeq or Ion Torrent PGM were used to generate the DNA sequence reads.
However, the RDP gold database has not been updated since 2011 (35) and could potentially be
missing many bacterial sequences, leading less differentiation between OTUs. With the DADA2
pipeline, ASVs were inferred prior to taxonomy assignment (26), resulting in the same total ASV
numbers for both RDP 11.5 and Greengenes 13.8 databases. However, assignments of DADA2
ASVs were still influenced by reference database-specific taxonomic nomenclature and DNA
sequences (36) such that Greengenes was provided deeper, more accurate taxonomic
assignments than found with RDP.

The Illumina MiSeq and Ion Torrent PGM methods also clearly impacted the outcomes
of our mock community analyses. The Illumina MiSeq is well-established and known for its low
error rate, high volume read outputs, and low sequencing cost per Gb (37, 38). The Ion Torrent
PGM method is recognized for shorter run times, lower instrument cost, and flexibility in
sequencing scale per run by the use of different sequencing chips (37, 38), and therefore could be
of relevance for dairy and other food products with short shelf-life times and low sample
numbers. For both single-end and paired-end assembled Illumina MiSeq reads, a greater number
of unexpected taxa and OTUs/ASVs were observed compared to the Ion Torrent PGM. This
finding could be the result of substitution errors Illumina MiSeq read accumulated leading to
taxonomy misidentification (39, 40). This issue became particularly apparent with the finding of
unrelated taxa such as *Anaeroplasma*, *Bacteroides*, *Desulfovibrio* and *Prevotella* that were not
included in the gDNA mock community. Use of the Ion Torrent PGM with our read trimming parameters resulted in the lowest numbers of DADA2 assigned ASVs. At 13 ASVs for both mock communities, this number was only slightly greater than the nine predicted. All thirteen ASVs were repeatedly assigned to members of the mock communities, except for one low abundance ASV when RDP was applied. The four additional ASVs were the result of read errors in the homopolymer regions, a common Ion Torrent error model (20, 37, 38). Interestingly, the Ion Torrent PGM reads resulted in the highest numbers of OTUs when QIIME 1 was used to analyze the data. This might have been due to the higher number of erroneous reads that were passed by QIIME 1 filtering, but were identified as sequence chimeras and artifacts by DADA2.

Alpha and beta diversity metrics were impacted by the DNA sequencing method. Calculations of beta diversity from the DADA2/Greengenes pipeline showed that the results from the Illumina MiSeq were similar to Ion Torrent PGM, especially for the PCR amplicon mock community. However, both paired and single end Illumina MiSeq methods resulted in erroneous estimates of Streptococcus and Lactococcus. Because Streptococcus and Lactococcus share similar 16S rRNA gene sequences, variation in their relative abundances could be caused by the accumulation of substitution errors, a common error that occurs with those instruments (39, 40). Alpha-diversity and beta-diversity assements on data collected with the Ion Torrent PGM and analyzed by DAD2/Greengenes most closely resembled the in silico reconstructions of the mock communities. Taxonomic proportions were also less impacted compared to results from the Illumina MiSeq and limited to reduced proportions of Escherichia and increased Streptococcus levels. This finding is similar to previous studies in which Proteobacteria proportions were underestimated and Firmicutes proportions were overestimated due to negative correlations with genomic G+C content (41). To minimize PCR preference of A+T rich bacterial
organisms, methods such as increasing the denaturing time (41) could be used for future dairy
microbiome studies.

By the use of bacterial DNA standards from nine dairy-relevant bacterial species, we
found that DNA sequencing and analysis pipelines contributed significant variations to
OTU/ASV distributions and observed bacterial diversities. Moreover, PCR biases and errors
from multi-template DNA amplifications are not entirely filtered with the Illumina MiSeq
method. Overall, the Ion Torrent PGM DNA sequencer combined with the DADA2/Greengenes
pipeline led to more accurate OTU/ASV assignments and bacterial diversity measurements under
our study conditions. With DADA2 being wrapped in the QIIME 2, we agree with the QIIME 2
developers that new sequencing results should be analyzed using QIIME 2 with standardized
analysis pipeline (e.g: DADA2) instead of QIIME 1 (UCLUST) (42). Further improvements
might be reached by refinements to taxonomy classifiers (34) and updating reference databases
to emphasize bacteria found in different environments such as dairy foods (36). Moreover,
additional studies are required to further establish the relevance of PCR protocols on the bacterial
diversity outcomes, and the analysis methods should be tested on mock communities consisting
of intact bacterial cells exposed to selected DNA extraction and purification methods (43). The
findings here and the continued development of bacterial diversity analysis methods should
result in reliable comparisons within and between microbial habitats.

Materials and Methods

Bacterial strains and culture conditions. Bacterial strains representing species commonly
found in bovine milk were used to construct a mock bacterial community (Table 1). Each
bacterial strain was grown in standard laboratory culture medium with negative controls for that
species and harvested at early stationary phase by centrifugation at 13,000xg for 2 min.

Laboratory culture medium was as follows: *Bacillus subtilis*, *Pseudomonas fluorescens*, and *Escherichia coli* - LB Lennox broth (Thermo Fisher Scientific); *Enterococcus faecalis* and *Streptococcus agalactiae* - Brain Heart Infusion broth (Thermo Fisher Scientific);

*Staphylococcus aureus* - Tryptic Soy broth (Becton Dickinson); *Corynebacterium bovis* - Tryptic Soy broth (Becton Dickinson) with 0.1% tween 80; *Lactococcus lactis* - M17 broth (Becton Dickinson) with 0.5% glucose; and *Clostridium tyrobutyricum* - Reinforced Clostridial broth (Becton Dickinson). All strains were incubated at 37 °C, with the exception of *B. subtilis*, *L. lactis* and *P. fluorescens*, which were incubated at 30 °C. *B. subtilis*, *C. bovis*, *E. faecalis*, *E. coli* and *P. fluorescens* were grown under aeration (250 rpm).

**Genomic DNA extraction and PCR amplification.** Genomic DNA was extracted using the MagMAX Total Nucleic Acid Isolation Kit (Thermo Fisher Scientific, Vilnius, LT) according to the manufacturer’s protocol with the repeat bead beating method on a FastPrep-24 instrument (MP Biomedicals LLC). DNA concentration was measured with the Qubit 3.0 Fluorometer using the Qubit dsDNA HS Assay Kit (Life Technologies, Eugene, OR). PCR amplification was performed using ExTaq DNA polymerase (TaKaRa, Otsu, Japan) and primers F515 and R806 (44) with a random 8 bp barcode on the 5’ end of F515 for sample multiplexing (45, 46). PCR was initiated at 94 °C for 3 min and followed by 35 cycles of 94 °C for 45 sec, 54 °C for 60 sec, and 72 °C for 30 sec with a final extension step at 72 °C for 10 min. Negative controls were run for each barcoded primer. No PCR product for the negative controls was observed on a 1.5% agarose gel. PCR products were pooled and then gel purified with the Wizard SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA).
Preparation of the mock communities. For the gDNA mock community, 100 ng gDNA isolated from each of the strains was pooled in three separate replicates. The proportion of each bacterial strain in the gDNA mock community was determined by taking into account the genome size and 16S rRNA gene copy number (Table 1). To construct the amplicon mock community, gDNA of the nine bacterial strains was amplified in triplicate by using three different barcoded PCR primers. Amplicon concentrations were measured with the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Eugene, OR) prior to pooling at equal molar concentrations.

DNA sequencing. For Illumina sequencing, the KAPA HTP library preparation kit (KK8234, Kapa Biosystems, Pittsburgh, PA) was used for the ligation of NEXTflex™ adapters (Bioo Scientific, Austin, TX) to the 16S rRNA amplicons prior to 250 bp paired-end sequencing (with 7% PhiX control) on an Illumina MiSeq instrument at the University of California, Davis genome center (http://genomecenter.ucdavis.edu/). For Ion Torrent sequencing, non-barcoded Ion A and Ion P1 adapter were ligated to the pooled amplicons followed by templating, enrichment and sequencing on the One-Touch 2, One-Touch ES systems and Ion PGM using the 400 sequencing kit and a 318 v2 chip (Life Technologies, Carlsbad, CA).

16S rRNA gene sequence analysis. An in silico mock community, termed “expected”, was created using the 16S V4 amplicon sequences from published genomes and reference genomes for the specific bacterial species (Table 1). In addition, the expected 16S V4 region copy numbers were normalized based on the genome size and 16S rRNA gene copy numbers.
Illumina MiSeq sequencing outputs were trimmed with the fastx_tools (47) to keep the first 245 and 170 bases for the forward and reverse reads respectively. Ion Torrent sequence output BAM file was converted to FASTQ format using BEDTools (48) and reads shorter than 200bp were also removed. The first 280 bases of the Ion Torrent reads were kept.

The FASTQ files were then analyzed with QIIME version 1.9.1 and DADA2 1.6.0 (26, 49). In QIIME 1, Illumina reads from the two orientations (forward and reverse) were analyzed either with or without assembly where the `join_paired_ends.py` (fastq-join method) (50) script was used with minimum 100 bp overlap and 1% maximum difference between overlapping sequences. Ion Torrent, single-end and paired-end assembled Illumina FASTQ files were then removed of barcode (8 bases) and primer regions (forward primer: 21 bases, reverse primer: 20 bases) and were demultiplexed using the `split_libraries_fastq.py` script with no barcode error and quality filtered at Q30. Chimeric sequences were identified and removed using USEARCH method (33, 35) and the Greengene database version 13.8 (51, 52). Sequences from both Illumina and Ion Torrent as well as the in silico mock community with expected proportions were merged together as one fasta file for operational taxonomic unit (OTU) clustering using the `pick_open_reference_otus.py` script with recommended parameters (32) and the UCLUST method at 97% similarity thresholds. Greengenes version 13.8 (51, 52) and RDP_GOLD (35) databases were used as references for OTU assignments. Archaea, Chloroplast and low abundance (0.005%) OTUs were removed from the OTU tables (32).

In DADA2, for single-end analysis, the truncated Illumina and Ion Torrent FASTQ files after barcode (8 bases) and primer sequences (forward primer: 21 bases, reverse primer: 20 bases) trimming were demultiplexed using `split_libraries_fastq.py` script with no barcode error and no quality filter (-r 999, -n 999, -q 0, -p 0.0001). Since the single-end reads were already quality
trimmed, no additional truncation was performed in DADA2 to be consistent in read length with
QIIME 1 analysis. For paired-end analysis, in order to get matched sequence files, raw Illumina
reads were demultiplexed in pairs using the idemp tool (53) with no barcode error. Barcode,
forward and reverse primer regions were then trimmed with fastx_tools (47). Resulting reads
were truncated in DADA2 to keep the first 196 bases of the forward reads and 121 bases of the
reverse reads, which were then merged with 51bp minimum overlap to be consistent with QIIME
1 method in resulting read length. The error model learning, dereplication and chimera removal
steps were performed in DADA2 with default parameters. Taxonomy was assigned to the
resulting amplicon sequence variants (ASVs) using RDP database version 11.5 (54) and
Greengenes database version 13.8 with the minimum bootstrap confidence at 80 (51, 52).
Ion Torrent, Illumina single-end and paired-end assembled reads were merged together with the in
silico mock community using the phyloseq package in R (55), and singletons and low abundance
(0.005%) ASVs were removed to be consistent with QIIME 1 analysis.

Accession numbers. Joined- and single-end DNA sequences after quality filtering and trimming
were deposited in the Qiita database (https://qiita.ucsd.edu) under study ID 11351 and in the
European Nucleotide Archive (ENA) under accession number ERP104377.

Statistics. OTU/ASV counts were rarefied at 5483 sequences per sample to retain all samples for
downstream analyses. Significant differences in the observed mock community composition
(alpha diversity and taxonomic distribution) were determined by ANOVA with the Bonferroni’s
multiple comparisons test. A p-value of <0.05 indicates significance. The significance of sample
clustering was indicated by permutational multivariate analysis of variance using the adonis function from the vegan package in R (56) with a p-value of <0.05 through 9,999 permutations.

Acknowledgements

The authors would like to thank Yanin Srisengfa for her extensive help with bacterial culture. The authors would also like to thank NIZO food research for providing strain *Lactococcus lactis* IL1403, and Dr. Steven Lindow at University of California, Berkeley for providing strain *Pseudomonas fluorescens* A506.

The California Dairy Research Foundation grant “Rapid methods for microbial detection and quality assessment of milk” funded this study. The funding agency did not participate in study design, data collection, or interpretation of the data.

References

1. Dominianni C, Wu J, Hayes RB, Ahn J. 2014. Comparison of methods for fecal microbiome biospecimen collection. BMC Microbiol 14:6.

2. Flores R, Shi JX, Yu GQ, Ma B, Ravel J, Goedert JJ, Sinha R. 2015. Collection media and delayed freezing effects on microbial composition of human stool. Microbiome 3:11.

3. Hsieh YH, Peterson CM, Raggio A, Keenan MJ, Martin RJ, Ravussin E, Marco ML. 2016. Impact of different fecal processing methods on assessments of bacterial diversity in the human intestine. Front Microbiol 7:1643.

4. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 2016. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol 16:123.
5. Fouhy F, Deane J, Rea MC, O'Sullivan O, Ross RP, O'Callaghan G, lant BJ, Stanton C. 2015. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. PLoS One 10:e0119355.

6. Rubin BER, Gibbons SM, Kennedy S, Hampton-Marcell J, Owens S, Gilbert JA. 2013. Investigating the impact of storage conditions on microbial community composition in soil samples. PLoS One 8:e70460.

7. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, Hwang J, Chen J, Berkowsky R, Nessel L, Li HZ, Bushman FD. 2010. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol 10:206.

8. Carroll IM, Ringel-Kulka T, Siddle JP, Klaenhammer TR, Ringel Y. 2012. Characterization of the fecal microbiota using high-throughput sequencing reveals a stable microbial community during storage. PLoS One 7:e46953.

9. Bai GY, Gajer P, Nandy M, Ma B, Yang HQ, Sakamoto J, Blanchard MH, Ravel J, Brotman RM. 2012. Comparison of storage conditions for human vaginal microbiome studies. PLoS One 7:e36934.

10. Quigley L, O'Sullivan O, Beresford TP, Ross RP, Fitzgerald GF, Cotter PD. 2012. A comparison of methods used to extract bacterial DNA from raw milk and raw milk cheese. J Appl Microbiol 113:96-105.

11. Barratt MJ, Lebrilla C, Shapiro HY, Gordon JI. 2017. The gut microbiota, food science, and human nutrition: a timely marriage. Cell Host Microbe 22:134-141.
12. Mackenzie BW, Waite DW, Taylor MW. 2015. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol 6:130.

13. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 12:87.

14. Yuan SQ, Cohen DB, Ravel J, Abdo Z, Forney LJ. 2012. Evaluation of methods for the extraction and purification of DNA from the human microbiome. PLoS One 7:e33865.

15. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM, Satsangi J, Flint HJ, Parkhill J, Lees CW, Hold GL, UK IBD Genetics Consortium. 2014. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One 9:e88982.

16. Brandariz-Fontes C, Camacho-Sanchez M, Vila C, Vega-Pla JL, Rico C, Leonard JA. 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. Sci Rep 5:8056.

17. Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. PLoS One 7:e29973.

18. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771.
19. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:55.

20. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG. 2014. Performance comparison of Illumina and Ion Torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol 80:7583-7591.

21. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PLoS One 9:e94249.

22. Sinclair L, Osman OA, Bertilsson S, Eiler A. 2015. Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the Illumina platform. PLoS One 10:e0116955.

23. Golob JL, Margolis E, Hoffman NG, Fredricks DN. 2017. Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. BMC Bioinformatics 18:283.

24. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahe F, He Y, Zhou HW, Rognes T, Caporaso JG, Knight R. 2016. Open-source sequence clustering methods improve the state of the art. mSystems 1:e00003-15.

25. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16.
26. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581.

27. Edgar RC. 2017. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. PeerJ 5:e3889.

28. Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3:e1487.

29. He Y, Caporaso JG, Jiang XT, Sheng HF, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou HW. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 3:20.

30. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639.

31. Edgar RC. 2017. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. bioRxiv https://doi.org/10.1101/192211.

32. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10:57-U11.

33. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.

34. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Caporaso JG. 2018. Optimizing taxonomic classification of marker gene amplicon sequences. PeerJ Preprints doi:10.7287/peerj.preprints.3208v2.
35. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194-2200.

36. Balvociute M, Huson DH. 2017. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics 18:114.

37. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.

38. Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333.

39. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17:125.

40. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37.

41. Laursen MF, Dalgaard MD, Bahl MI. 2017. Genomic GC-content affects the accuracy of 16S rRNA gene sequencing based microbial profiling due to PCR bias. Front Microbiol 8:1934.

42. Carporaso G. 2017. A response to “Accuracy of microbial community diversity ...” by R Edgar for QIIME users. https://forum.qiime2.org/t/a-response-to-accuracy-of-microbial-community-diversity-by-r-edgar-for-qiime-users/1456. Accessed November 14, 2017.
43. Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. Appl Environ Microbiol 84:e02627-17.

44. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108:4516-4522.

45. Bokulich NA, Joseph CML, Allen G, Benson AK, Mills DA. 2012. Next-generation sequencing reveals significant bacterial diversity of botrytized wine. PLoS One 7:e36357.

46. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5:235-237.

47. Hannon Lab. 2014. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. http://hannonlab.cshl.edu/fastx_toolkit/download.html. Accessed May 30, 2017.

48. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

49. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttenhower C, 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.

50. Aronesty E. 2013. Comparison of sequencing utility programs. Open Bioinformatics J 7:1.
51. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610-618.
52. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. 2012. Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys. ISME J 6:94-103.
53. Wu Y. 2014. Barcode demultiplex for Illumina I1, R1, R2 fastq.gz files. https://github.com/yhwu/idemp. Accessed November 20, 2017.
54. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res 42:D633-D642.
55. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217.
56. Oksanen J, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHM, Szoecs E, Wagner H. 2018. vegan: community ecology package. https://cran.r-project.org/web/packages/vegan/index.html. Accessed February 21, 2018.
57. Lee J, Jang YS, Han MJ, Kim JY, Lee SY. 2016. Deciphering Clostridium tyrobutyricum metabolism based on the whole-genome sequence and proteome analyses. MBio 7:e00743-16.
58. Schroder J, Glaub A, Schneider J, Trost E, Tauch A. 2012. Draft genome sequence of Corynebacterium bovis DSM 20582, which causes clinical mastitis in dairy cows. J Bacteriol 194:4437-4437.
59. Kim EB, Kopit LM, Harris LJ, Marco ML. 2012. Draft genome sequence of the quality control strain Enterococcus faecalis ATCC 29212. J Bacteriol 194:6006-6007.

60. Bolotin A, Wincker P, Mauger S, Jaillon O, Malarme K, Weissenbach J, Ehrlich SD, Sorokin A. 2001. The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp lactis IL1403. Genome Res 11:731-753.

61. Loper JE, Hassan KA, Mavrodi DV, Davis EW, Lim CK, Shaffer BT, Elbourne LDH, Stockwell VO, Hartney SL, Breakwell K, Henkels MD, Tetu SG, Rangel LI, Kidarsa TA, Wilson NL, de Mortel JEV, Song CX, Blumhagen R, Radune D, Hostetler JB, Brinkac LM, Durkin AS, Kluepfel DA, Wechter WP, Anderson AJ, Kim YC, Pierson LS, Pierson EA, Paulsen IT. 2012. Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genet 8:e1002784.

62. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Cordani JJ, Connerton IF, Cummings NJ, Daniel RA, Denziot F, Devine KM, Dusterhoft A, Ehrlich SD, Emmerson PT, Entian KD, Errington J, Fabret C, Ferrari E, Foulger D, Fritz C, Fujita M, Fujita Y, Fuma S, Galizzi A, Galleron N, Ghim SY, Glaser P, Goffeau A, Golightly EJ, Grandi G, Guiseppi G, Guy BJ, Haga K, et al. 1997. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390:249-56.

63. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C,
Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157 : H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11-22.

Bouchard D, Peton V, Almeida S, Le Marechal C, Miyoshi A, Azevedo V, Berkova N, Rault L, Francois P, Schrenzel J, Even S, Hernandez D, Le Loir Y. 2012. Genome asequence of *Staphylococcus aureus* Newbould 305, a strain associated with mild bovine mastitis. J Bacteriol 194:6292-6293.

Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM. 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A 99:12391-12396.
### Table 1. Bacterial strains and expected relative abundances in the gDNA mock community.

| Strain                                      | 16S rRNA gene copies<sup>a</sup> | % of total<sup>b</sup> | Genomic reference |
|---------------------------------------------|----------------------------------|------------------------|-------------------|
| *Bacillus subtilis* S44                    | 10                               | 16.67                  | NA<sup>c</sup>    |
| *Clostridium tyrobutyricum* ATCC 25755      | 6                                | 2.29                   | (57)              |
| *Corynebacterium bovis* ATCC 7715           | 1                                | 2.78                   | (58)              |
| *Enterococcus faecalis* ATCC 29212          | 4                                | 9.28                   | (59)              |
| *Escherichia coli* ATCC 700728              | 7                                | 8.95                   | NA                |
| *Lactococcus lactis* IL1403                 | 6                                | 17.82                  | (60)              |
| *Pseudomonas fluorescens* A506              | 6                                | 7.08                   | (61)              |
| *Staphylococcus aureus* ATCC 29740          | 5                                | 12.44                  | NA                |
| *Streptococcus agalactiae* ATCC 27956       | 7                                | 22.7                   | NA                |

<sup>a</sup> rRNA gene copies per genome based on genome reference

<sup>b</sup> Percentage of total bacterial 16S rRNA gene in the mock community according to DNA concentration

<sup>c</sup> For strains that lack whole genome sequences, the genome sizes and 16S rRNA gene copy numbers of reference strain were used (62-65).
Table 2. OTU/ASV distribution of the 16S rRNA PCR amplicon mock community following Illumina MiSeq DNA sequencing and paired-end assembly.

| Taxonomy           | Greengenes | RDP     | Greengenes | RDP     |
|--------------------|------------|---------|------------|---------|
| Bacillaceae        | 10 (83%)   | 7 (99%) | 2 (90%)    | 2 (90%) |
| Bacillus           | 1 (100%)   |         | 2 (90%)    | 2 (90%) |
| Clostridiaceae     | 3 (73%)    |         | 1 (100%)   |         |
| Clostridium        | 3 (87%)    | 7 (99%) | 4 (86%)    | 5 (85%) |
| Corynebacterium    | 14 (80%)   | 11 (88%)| 2 (91%)    | 2 (91%) |
| Enterococcaceae    |            | 2 (99%) |            |         |
| Enterococcus       | 4 (99%)    | 3 (85%) | 3 (85%)    |         |
| Enterobacteriaceae | 4 (99%)    | 17 (60%)| 2 (88%)    |         |
| Escherichia        |            |         | 2 (88%)    |         |
| Lactococcus        | 6 (99%)    | 1 (100%)| 3 (89%)    | 3 (89%) |
| Pseudomonas        | 10 (74%)   | 8 (66%) | 3 (85%)    | 3 (85%) |
| Staphylococcus     | 9 (99%)    | 7 (65%) | 3 (89%)    | 3 (89%) |
| Streptococcus      | 2 (80%)    | 2 (99%) | 2 (89%)    | 2 (89%) |
| Other              | 20 (17%)   | 2 (71%) | 13 (20%)   | 13 (20%)|
| **Sum**            | **85**     | **65**  | **38**     | **38**  |

*Each value represents the average number of OTU/ASV (n= 3) and mean percentage (%) of sequence reads assigned to the most abundant OTU/ASV within that taxon.*
Table 3. OTU/ASV distribution of the 16S rRNA PCR amplicon mock community following Illumina MiSeq DNA sequencing without paired-end assembly.

| Taxonomy                | Qiime 1 Greengenes | Qiime 1 RDP | DADA2 Greengenes | DADA2 RDP |
|-------------------------|---------------------|-------------|------------------|-----------|
| **Bacillus**            | 9 (97%)^a           | 2 (99%)     | 3 (67%)          | 1 (100%)  |
| **Clostridaceae**       | 1 (100%)            | 1 (100%)    | 1 (100%)         |           |
| **Clostridium**         | 3 (99%)             | 3 (99%)     | 3 (99%)          | 3 (99%)   |
| **Corynebacterium**     | 9 (96%)             | 7 (95%)     | 2 (75%)          | 2 (75%)   |
| **Enterococccaceae**    | 1 (100%)            |             |                  |           |
| **Enterococcus**        | 2 (99%)             | 2 (99%)     | 2 (60%)          | 2 (60%)   |
| **Enterobacteriaceae**  | 4 (99%)             | 8 (93%)     |                  |           |
| **Escherichia/Shigella**|                    |             | 1 (100%)         |           |
| **Escherichia**         |                    |             |                  | 1 (100%)  |
| **Lactococcus**         | 6 (99%)             | 2 (99%)     | 1 (100%)         | 1 (100%)  |
| **Pseudomonas**         | 11 (91%)            | 4 (99%)     | 2 (65%)          | 2 (65%)   |
| **Staphylococcus**      | 12 (98%)            | 3 (99%)     | 2 (76%)          | 2 (76%)   |
| **Streptococcus**       | 4 (92%)             | 2 (99%)     | 1 (100%)         | 1 (100%)  |
| **Other**               | 6 (28%)             | 2 (91%)     | 22 (12%)         | 25 (89%)  |
| **Sum**                 | 68                  | 36          | 40               | 40        |

^a Each value represents the average number of OTU/ASV (n= 3) and mean percentage (%) of sequence reads assigned to the most abundant OTU/ASV within that taxon.
Table 4. OTU/ASV distribution of the 16S rRNA PCR amplicon mock community following Ion Torrent PGM sequencing.

| Taxonomy           | QIIME 1 Greengenes | QIIME 1 RDP | DADA2 Greengenes | DADA2 RDP |
|--------------------|--------------------|-------------|------------------|-----------|
| Bacillaceae        | 3 (88%)<sup>a</sup> | 8 (95%)     | 2 (95%)          | 1 (100%)  |
| Bacillus           | 21 (75%)           | 8 (95%)     | 1 (100%)         | 1 (100%)  |
| Clostridium        | 5 (70%)            | 7 (99%)     | 1 (100%)         | 1 (100%)  |
| Corynebacterium    | 15 (75%)           | 11 (88%)    | 1 (100%)         | 1 (100%)  |
| Enterococccaceae   | 2 (50%)            |             |                  |           |
| Enterococcus       | 9 (98%)            | 2 (99%)     | 1 (100%)         | 1 (100%)  |
| Enterobacteriaceae | 13 (95%)           | 17 (60%)    |                  | 1 (100%)  |
| Escherichia        | 1 (100%)           |             |                  |           |
| Lactococcus        | 6 (99%)            | 1 (100%)    | 2 (99%)          | 2 (99%)   |
| Pseudomonas        | 13 (71%)           | 8 (66%)     | 2 (99%)          | 2 (99%)   |
| Staphylococcus     | 21 (97%)           | 7 (65%)     | 2 (99%)          | 2 (99%)   |
| Streptococcus      | 6 (83%)            | 2 (99%)     | 1 (100%)         | 1 (100%)  |
| Other              | 6 (40%)            | 2 (71%)     | 0                | 1 (100%)  |
| Sum                | 118                | 67          | 13               | 13        |

<sup>a</sup>Each value represents the average number of OTU/ASV (n= 3) and mean percentage (%) of sequence reads assigned to the most abundant OTU/ASV within that taxon.
Table 5. ASV distribution of the gDNA mock community following different sequencing methods\(^a\).

| Taxonomy          | Illumina paired | Illumina single | Ion Torrent |
|-------------------|-----------------|-----------------|-------------|
| Bacillus          | 4 (86%)\(^b\)   | 2 (96%)         | 2 (95%)     |
| Clostridiaceae    | 3 (92%)         | 1 (100%)        |             |
| Clostridium       | 4 (80%)         | 2 (94%)         | 1 (100%)    |
| Corynebacterium   | 2 (96%)         | 1 (100%)        | 1 (100%)    |
| Enterococcus      | 3 (88%)         | 1 (100%)        | 1 (100%)    |
| Escherichia       | 2 (96%)         | 1 (100%)        | 1 (100%)    |
| Lactococcus       | 3 (88%)         | 3 (67%)         | 2 (99%)     |
| Pseudomonas       | 3 (94%)         | 1 (100%)        | 2 (99%)     |
| Staphylococcus    | 2 (90%)         | 2 (99%)         | 2 (99%)     |
| Streptococcus     | 3 (90%)         | 2 (99%)         | 1 (100%)    |
| Other             | 116 (12%)       | 111 (10%)       | 0           |
| **Sum**           | 145             | 127             | 13          |

\(^a\) Results were based on DADA2 analysis pipeline with the Greengenes database.

\(^b\) Each value represents the average number of OTU/ASV (n= 3) and mean percentage (%) of sequence reads assigned to the most abundant OTU/ASV within that taxon.
FIG 1. Alpha diversity measurements of mock community samples. Shannon index of (A) PCR amplicon mock community and (B) gDNA mock community. Chao1 index of (C) amplicon mock community and (D) gDNA mock community. Results shown were analyzed following DADA2 pipeline and Greengenes database. Each bar represents the mean ± stdev for three replicates. Alpha diversity measurements for each community were compared to expected values using the ANOVA with the Bonferroni’s multiple comparisons test. P values <0.05 were considered to be significant from the expected values and were indicated by the presence of asterisk signs above each bar plot.
FIG 2. Relative proportions of taxa and UPGMA hierarchical clustering of the mock communities. UPGMA hierarchical clustering was based on the (A) Bray Curtis dissimilarity matrix and (B) weighted Unifrac distance matrix. Expected taxa (9 bacterial species) are labelled with the corresponding taxonomic level from the DNA sequencing results. Each bar contains the results from each of the three mock community replicates tested using different DNA sequencing methods. Results shown were analyzed following the DADA2 pipeline with Greengenes database.
FIG 3. Relative abundance of taxa in the 16S rRNA PCR amplicon and gDNA mock communities. Relative abundance of expected taxa are labelled with the corresponding taxonomic level from sequencing results. Amplicon stands for the 16S rRNA PCR amplicon mock community and gDNA stands for the gDNA mock community. Results shown were analyzed following DADA2 pipeline with Greengenes database. Each bar represents the mean ± stdev for three replicates. Proportions for each community were compared to expected proportions using the ANOVA with the Bonferroni’s multiple comparisons test. P values <0.05 were considered to be significant from the expected values and were indicated by the presence of asterisk signs above each bar plot.