Noname manuscript No.
(will be inserted by the editor)

Assessing 16S marker gene survey data analysis methods using mixtures of human stool sample DNA extracts.

ND Olson · MS Kumar · S Hao · W Timp · ML Salit · OC Stine · H Corrada Bravo

Received: date / Accepted: date

Abstract 16S rRNA marker-gene surveys use targeted sequencing to characterize prokaryotic microbial communities. Analysis of these studies is confronted with numerous bioinformatic pipelines and downstream analysis methods, with limited guidance on how to decide between appropriate methods from simulation studies or limited complexity benchmark studies. Appropriate data sets and statistics for assessing these methods are needed. A mixture of environmental samples is one approach for generating assessment data sets with the real data complexity while providing an expected value.We developed a mixture dataset for assessing 16S rRNA bioinformatic pipelines and downstream analysis methods using samples collected from participants in a Enterotoxigenic *Escherichia coli* (ETEC) vaccine trial participants. A two-sample titration mixture design was used where DNA from stool samples prior to ETEC exposure was titrated into stools samples collected after exposure, in effect diluting the amount of ETEC in the mixed sample. The sequencing data were processed using multiple bioinformatic pipelines, DADA2 a sequence inference method, Mothur a *de novo* clustering method, and QIIME with open-reference

ND Olson · ML Salit
Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
E-mail: nolson@nist.gov

ND Olson · MS Kumar · H Corrada Bravo
Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742, USA

S Hao · W Timp
Department of Biomedical Engineering, Johns Hopkins University

OC Stine
School of Medicine, University of Maryland

H Corrada Bravo
Department of Computer Science, University of Maryland, College Park, MD 20742, USA

clustering. The pipelines varied in the number of features and proportion of reads passing quality control but had similar sparsity. The mixture dataset was used to qualitatively and quantitatively assess the count tables generated using the pipelines. Statistical tests were used to determine if features only present in unmixed samples and titrations, *unmixed*- and *titration*-specific features, were had abundance value that could be explained by sampling alone. For Mothur and QIIME less than 5% of *unmixed*- and *titration*-specific feature abundance could not be explained by sampling alone where as for DADA2 greater than 50% of *unmixed*-specific features and 10% of *titration*- specific features could not be explained by sampling alone. The quantitative assessment evaluated pipeline performance by comparing observed to expected relative and differential abundance values. Expected relative abundance and differential abundance values were calculated using information from the unmixed samples and mixture design. Overall the observed relative abundance and differential abundance values were consistent with the expected values. We developed feature-level bias and variance metric to further characterize relative abudance and differential abundance quantitative performance.Relative abundance feature-level bias metric was significantly different across the three platforms with DADA2 having the lowest bias, followed by Mothur, and QIIME.The relative abundance feature-level variance metric and both the differential abundance feature-level bias and variance metrics did not differ significantly across the three pipelines.The dataset and methods developed for this study will serve as a valuable community resource for assessing 16S rRNA marker-gene survey bioinformatic methods.

Keywords 16S rRNA · assessment · bioinformatic pipeline · normalization · differential abundance ·

1 Introduction

Targeted sequencing of the 16S rRNA gene, 16S rRNA marker-gene-surveys, is a commonly used method for characterizing microbial communities, microbiomes. The 16S rRNA marker-gene-survey measurement process includes molecular (e.g. PCR and sequencing) and computational steps (e.g., sequence clustering) (Goodrich et al. 2014). Molecular steps are used to selectively target and sequence the 16S rRNA gene from prokaryotic organisms within a sample. The computational steps convert the raw sequence data into a matrix with feature (e.g., operational taxonomic units) relative abundance values for each sample (Goodrich et al. 2014). Both molecular and computational measurement process steps contribute to the overall measurement bias and dispersion (D'Amore et al. 2016; Goodrich et al. 2014; Brooks et al. 2015). Proper measurement method evaluation allows for the characterization of how individual steps impact the measurement processes as a whole and determine where to focus efforts for improving the measurement process. Appropriate

datasets and methods are needed to evaluate the 16S rRNA marker-gene-survey measurement process. A sample or dataset with "ground truth" is needed to characterize measurement process accuracy. Numerous studies have evaluated quantitative and qualitative characteristics of the 16S rRNA measurement process using mock communities, simulated data, and environmental samples.

To assess the 16S rRNA sequencing measurement process qualitative characteristics of a mock communities are commonly used (Bokulich et al. 2016). As the number of organisms in the mock community is known, the total number of features can be compared to the expected value. The number of observed features in a mock community is often significantly higher than the expected number of organism (Kopylova et al. 2014). The higher than expected number of features is often attributed to sequencing and PCR artifacts as well as reagent contaminants (Brooks et al. 2015; Huse et al. 2010). A notable exception to this is mock community benchmarking studies evaluating sequencing inference method, such as DADA2 (B. J. Callahan et al. 2016). Sequence inference methods aim to reduce the number of sequence artifacts features. While mock communities have a known value, they lack the feature diversity and relative abundance dynamic range of real samples (Bokulich et al. 2016).

The quantitative characteristics of 16S rRNA sequence data are normally assessed using mock communities and simulated data. Mock communities of equimolar and staggered concentration are used to assess relative abundance estimate quantitative accuracy (Kopylova et al. 2014). Results from relative abundance estimates using mock communities generated from mixtures of DNA have shown taxonomic specific effects where individual taxa are under or over represented in a sample. Taxonomic specific biases due to DNA extraction have been shown with Gram negatives having higher extraction efficiency compared to Gram negatives (Costea et al. 2017, @Olson2012). Mismatches in the primer binding sites are also responsible for taxonomic specific effects (Brooks et al. 2015; Klindworth et al. 2012; Gohl et al. 2016). Additionally, sequence template properties such as GC content, sequence secondary structure, and gene flanking regions have been attributed to taxon specific biases (Pinto and Raskin 2012; Hansen et al. 1998; Gohl et al. 2016). Simulated count tables have been used to assess differential abundance method, where specific taxa are artificially overrepresented in one set of samples compared to another (McMurdie and Holmes 2014). Using simulated data to assess log fold-change estimates only evaluates computational steps of the measurement process.

Quantitative and qualitative assessment can also be performed using sequence data generated from mixtures of environmental samples. While simulated data and mock communities are useful in evaluating and benchmarking new methods one needs to consider that methods optimized for mock communities and simulated data are not necessarily optimized for the sequencing error profile and feature diversity of real samples. Data from environmental samples, which are real samples, are often used to benchmark new molecular laboratory and computational methods. However, without an expected value to compare to, only measurement precision can be evaluated. By mixing environmental

samples, an expected value can be calculated using information from the unmixed samples and how they were mixed. Mixtures of environmental samples have previously been used to evaluate gene expression measurements microarrays and RNAseq data(Parsons et al. 2015; Pine, Rosenzweig, and Thompson 2011; Thompson et al. 2005).

In the present study, we developed a mixture dataset of extracted DNA from human stool samples for assessing 16S rRNA sequencing. The mixture datasets were processed using three bioinformatic pipelines. We developed metrics for qualitative and quantitative assessment of the bioinformatic pipeline results. The quantitative results were similar across pipelines but the qualitative results varied across pipelines. We have made both the dataset and metrics developed in this study publically available for evaluating new bioinformatic pipelines.

2 Methods

2.0.1 Two-Sample Titration Design

Samples collected at multiple timepoints during a Enterotoxigenic *E. coli* (ETEC) vaccine trial (Harro et al. 2011) were used to generate a two-sample titration dataset for assessing the 16S rRNA marker-gene survey measurement process. Samples from five trial participants were selected for our two-sample titration dataset. Trial participants (subjects) and sampling timepoints were selected based on *E. coli* abundance data collected using qPCR and 16S rRNA sequencing from Pop et al. (2016). Only individuals with no *E. coli* detected in samples collected from trial participants prior to ETEC exposure twere used for our two-samples titrations. Post ETEC exposure (POST) samples were identified as the timepoint after exposure to ETEC with the highest *E. coli* concentration for each subject (Fig. 1A). Due to limited sample availability, the timepoint with the second highest concentrations for E01JH0016 was used as the POST sample. Independent titration series were generated for each subject, where POST samples were titrated into PRE samples with POST proportions of 1/2, 1/4, 1/8, 1/16, 1/32, 1/1,024, and 1/32,768 (Fig. 1B). Unmixed (PRE and POST) sample DNA concentration was measured using NanoDrop ND-1000 (Thermo Fisher Scientific Inc. Waltham, MA USA). Unmixed samples were diluted to 12.5 $ng/\mu L$ in tris-EDTA buffer before making the two-sample titrations.

For our two-sample titration mixture design, the expected feature relative abundance can be calculated using equation (1), where θ_i, is the proportion of POST DNA in titration i, q_{ij} is the relative abundance of feature j in titration i, and the relative abundance of feature j in the unmixed PRE and POST samples is $q_{pre,j}$ and $q_{post,j}$.

$$q_{ij} = \theta_i q_{post,j} + (1 - \theta_i) q_{pre,j} \tag{1}$$

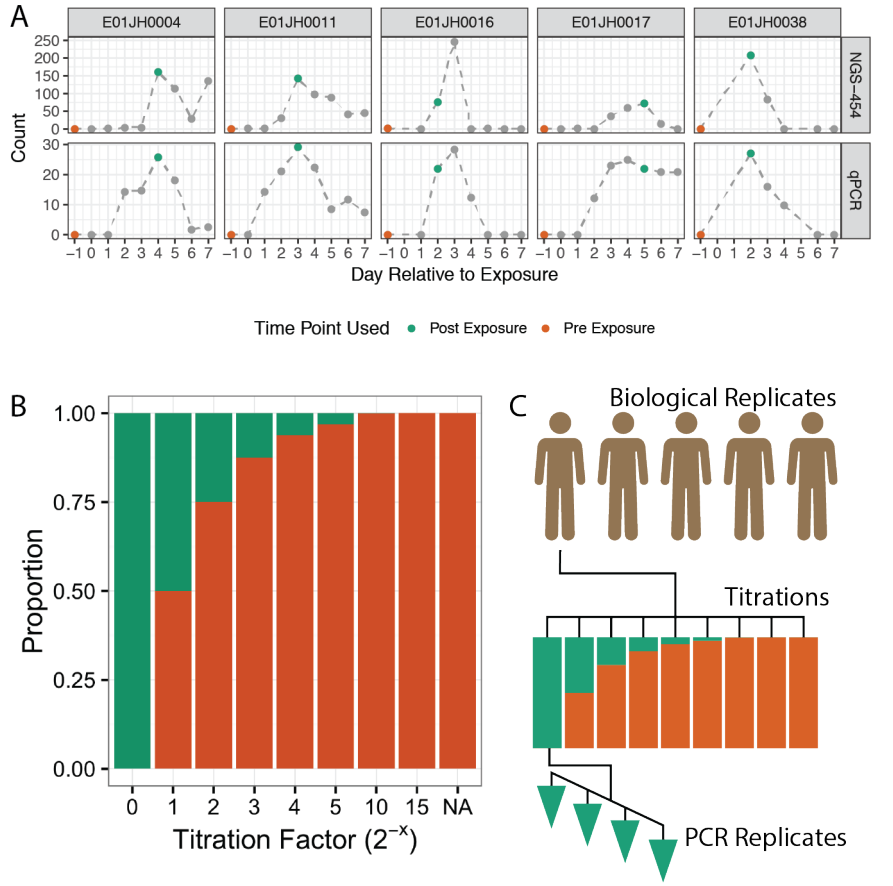

Fig. 1 Sample selection and experimental design for the two-sample titration 16S rRNA marker-gene-survey assessment dataset. A) Pre- and post-exposure (PRE and POST) samples from five vaccine trial participants were selected based on *Escherichia coli* abundance measured using qPCR and 454 16S rRNA sequencing (454-NGS), data from Pop et al. (2016). PRE and POST samples are indicated with orange and green data points, respectively. Grey points are other samples from the vaccine trial time series. B) Proportion of DNA from PRE and POST samples in titration series samples. PRE samples were titrated into POST samples following a log_2 dilution series. The NA titration factor represents the unmixed PRE sample. C) PRE and POST samples from the five vaccine trial participants, subjects, were used to generate independent two-sample titration series. The result was a total of 45 samples, 7 titrations + 2 unmixed samples times 5 subjects. Four replicate PCRs were performed for each of the 45 samples resulting in 190 PCRs.

2.0.2 Titration Validation

qPCR was used to validate volumetric mixing and check for differences in the proportion of prokaryotic DNA across titrations. To ensure that the two-sample titrations were volumetrically mixed according to the mixture design, independent ERCC plasmids were spiked into the unmixed PRE and POST

samples (Baker et al. 2005) (NIST SRM SRM 2374) (Table 2). The ERCC plasmids were resuspended in 100 $ng/\mu L$ tris-EDTA buffer and 2 $ng/\mu L$ was spiked into the appropriate unmixed sample. Plasmids were spiked into unmixed samples after unmixed sample concentration was normalized to 12.5 $ng/\mu L$. POST sample ERCC plasmid abundance was quantified using TaqMan gene expression assays (FAM-MGB) (Catalog # 4448892, ThermoFisher) specific to each ERCC plasmid using the TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA). To check for differences in the proportion of bacterial DNA in the PRE and POST samples, titration bacterial DNA concentration was quantified using the Femto Bacterial DNA quantification kit (Zymo Research, Irvine CA). All samples were run in triplicate along with an in-house *E. coli* DNA log_{10} dilution standard curve. qPCR assays were performed using the QuantStudio Real-Time qPCR (ThermoFisher). Amplification data and Ct values were exported as tsv files using QuantStudio Design and Analysis Software v1.4.1. Statistical analysis was performed on the exported data using custom scripts in R (R Core Team 2018, `https://github.com/nate-d-olson/mgtst_pub`).

2.0.3 Sequencing

The 45 samples (seven titrations and two unmixed samples for each of five subjects) were processed using the Illumina 16S library protocol (16S Metagenomic Sequencing Library Preparation, posted date 11/27/2013, downloaded from `https://support.illumina.com`). This protocol specifies an initial 16S rRNA PCR followed by a sample indexing PCR, followed by normalization and sequencing.

A total of 192 16S rRNA PCR assays were run including four replicates per sample and 12 no-template controls, using Kapa HiFi HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA). The initial PCR assay targeted the V3-V5 region of the 16S rRNA gene, Bakt_341F and Bakt_806R (Klindworth et al. 2012). The V3-V5 region is 464 base pairs (bp) long, with forward and reverse reads overlapping by 136 bp, using 2 X 300 bp paired-end sequencing (Yang, Wang, and Qian 2016) (`http://probebase.csb.univie.ac.at`). Primer sequences include overhang adapter sequences for library preparation (forward primer 5'- TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CCT ACG GGN GGC WGC AG - 3' and reverse primer 5'- GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CHV GGG TAT CTA ATC C - 3'). For quality control, the PCR product was verified using agarose gel electrophoresis to check for appropriate size bands, and concentration measurements were made after the initial 16S rRNA PCR, the indexing PCR, and normalization steps. DNA concentration was measured using SpextraMax Accuclear Nano dsDNA Assay Bulk Kit (Part# R8357#, Lot 215737, Molecular Devices LLC. Sunnyvale CA, USA) and fluorescent measurements were made with a Molecular Devices SpectraMax M2 spectraflourometer (Molecular Devices LLC. Sunnyvale CA, USA).

Initial PCR products were purified using AMPure XP beads (Beckman Coulter Genomics, Danvers, MA) following the manufacturer's protocol. After purification, the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). Prior to pooling purified sample concentration was normalized using SequalPrep Normalization Plate Kit (Catalog n. A10510-01, Invitrogen Corp., Carlsbad, CA), according to the manufacturer's protocol. Pooled library concentration was checked using the Qubit dsDNA HS Assay Kit (Part# Q32851, Lot# 1735902, ThermoFisher, Waltham, MA USA). Due to the low pooled amplicon library DNA concentration, a modified protocol for low concentration libraries was used. The library was run on an Illumina MiSeq, and base calls were made using Illumina Real Time Analysis Software version 1.18.54. Sequencing data quality control metrics for the 384 fastq sequence files (192 samples with forward and reverse reads) were computed using the Bioconductor `Rqc` package (Souza and Carvalho 2017; Huber et al. 2015).

2.0.4 Sequence Processing

Sequence data were processed using four bioinformatic pipelines: a *de-novo* clustering method - Mothur (Schloss et al. 2009), an open-reference clustering method - QIIME (Caporaso et al. 2010), and a sequence inference methods - DADA2 (B. J. Callahan et al. 2016), and unclustered sequences as a control. The code used to run the bioinformatic pipelines is available at `https://github.com/nate-d-olson/mgtst_pipelines`.

The Mothur pipeline follows the developers MiSeq SOP (Schloss et al. 2009; Kozich et al. 2013). The pipeline was run using Mothur version 1.37 (`http://www.mothur.org/`) As we sequenced a larger 16S rRNA region, with smaller overlap between the forward and reverse reads, than the 16S rRNA region the SOP was designed. Pipeline parameters were modified to account for the difference in overlap are noted for individual steps below. The Makefile and scripts used to run the mothur pipeline are available `https://github.com/nate-d-olson/mgtst_pipelines/blob/master/code/mothur`. The Mothur pipeline included an initial preprocessing step where the forward and reverse reads are trimmed and filtered using base quality scores merged into contigs. The following parameters were used for the initial contig filtering, no ambiguous bases, max contig length of 500 bp, and max homopolymer length of 8 bases. For the initial read filtering and merging step, low-quality reads were identified and filtered from the dataset based on the presence of ambiguous bases, failure to align to the SILVA reference database (V119, `https://www.arb-silva.de/`) (Quast et al. 2012), and identification as chimeras. Prior to alignment, the SILVA reference multiple sequence alignment was trimmed to the V3-V5 region, positions 6,388 and 25,316. Chimera filtering was performed using UChime (version v4.2.40) without a reference database (Edgar et al. 2011). OTU clustering was performed using the OptiClust algorithm with a clustering threshold of 0.97 (Westcott and Schloss 2017). The RDP classi-

fier implemented in mothur was used for taxonomic classification against the mothur provided version of the RDP v9 training set (Wang et al. 2007).

The QIIME open-reference clustering pipeline for paired-end Illumina data was performed according to the online tutorial (Illumina Overview Tutorial (an IPython Notebook): open reference OTU picking and core diversity analyses, `http://qiime.org/tutorials/`) using QIIME version 1.9.1 (Caporaso et al. 2010). Briefly, the QIIME pipeline uses fastq-join (version 1.3.1) to merge paired-end reads (Aronesty 2011) and the Usearch algorithm (Edgar 2010) with Greengenes database version 13.8 with a 97% similarity threshold (DeSantis et al. 2006) was used for open-reference clustering.

DADA2, an R native pipeline was also used to process the sequencing data (B. J. Callahan et al. 2016). The pipeline includes a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP nave Bayesian classifier (Wang et al. 2007) and the SILVA database V123 provided by the DADA2 developers (Quast et al. 2012, `https://benjjneb.github.io/dada2/training.html`).

The unclustered pipeline was based on the mothur *de-novo* clustering pipeline, where the paired-end reads were merged, filtered, and then dereplicated. Reads were aligned to the reference Silva alignment (V119, `https://www.arb-silva.de/`), and reads failing alignment were excluded from the dataset. Taxonomic classification of the unclustered sequences was performed using the same RDP classifier implemented in mothur used for the *de-novo* pipeline. To limit the size of the dataset the most abundant 40,000 OTUs (comparable to the mothur dataset), across all samples, were used as the unclustered dataset.

2.0.5 Titration Proportion Estimates

The following linear model (2) was used to infer the proportion of prokaryotic DNA in each titration, θ. Where $\mathbf{Q}_i$ is a vector of titration i feature relative abundance estimates and $\mathbf{Q}_{pre}$ and $\mathbf{Q}_{post}$ are vectors of feature relative abundance estimates for the unmixed PRE and POST samples. Average PCR replicate relative abundance values were calculated using a negative binomial model.

$$\mathbf{Q}_i = \theta_i(\mathbf{Q}_{post} - \mathbf{Q}_{pre}) + \mathbf{Q}_{pre} \tag{2}$$

To fit the model to prevent uninformative and low abundance features from biasing θ estimates only informative features meeting the following criteria were used Features included in the model were observed in at least 14 of the 28 total titration PCR replicates (4 replicates per 7 titrations), demonstrated greater than 2-fold difference in relative abundance between the PRE and POST samples, and were present in either all four or none of the PRE and POST PCR replicates.

16S rRNA sequencing count data is known to have a non-normal mean-variance relationship resulting in poor model fit for standard linear regression

(McMurdie and Holmes 2014). Generalized linear models provide an alternative to standard least-squares regression. The above model is additive and therefore unable to directly infer θ_i in log-space. To address this issue, we fit the model using a standard least-squares regression then obtained non-parametric 95 % confidence intervals for the θ estimates by bootstrapping with 1000 replicates.

2.0.6 Qualitative Assessment

Our qualitative measurement assessment evaluated features only observed in unmixed samples (PRE or POST), *unmixed-specific*, or titrations,*titration-specific*. *Unmixed-* or *titration-specific* features are due to differences in sampling depth (number of sequences) between the unmixed samples and titrations, artifacts of the feature inference process, or PCR/sequencing artifacts. Measurement process artifacts should be considered false positives or negatives. Hypothesis tests were used to determine if differences in sampling depth could account for *unmixed-specific* and *titration-specific* features. p-values were adjusted for multiple comparisons using the Benjamini & Hochberg method (Benjamini and Hochberg 1995). For *unmixed-specific* features, the binomial test was used to evaluate if true feature relative abundance is less than the expected relative abundance. A binomial test could not be used to evaluate *titration-specific* features, as the hypothesis would be formulated as such. Given observed counts and the titration total feature abundance, the true feature relative abundance is equal to 0. As non-zero counts were observed the true feature proportion is non-zero, and the test always fails. Therefore, we formulated a Bayesian hypothesis test for *titration-specific* features.

A Bayesian hypothesis test was used to evaluate if the true feature proportion is less than the minimum detected proportion. The Bayesian hypothesis test was formulated using equation (3). Which when assuming equal priors, $P(\pi < \pi_{min}) = P(\pi \geq \pi_{min})$, reduces to (4). For equations (3) and (4) π is the true feature proportion, π_{min} is the minimum detected proportion, C is the expected feature counts, and C_{obs} is the observed feature counts. Simulation was used to generate possible values of C, assuming C has a binomial distribution given the observed sample total feature abundance, and a uniform probability distribution for π between 0 and 1. π_{min} was calculated using the mixture equation (1) where $q_{pre,j}$ and $q_{post,j}$ are $min(\mathbf{Q}_{pre})$ and $min(\mathbf{Q}_{post})$ across all features for a subject and pipeline. Our assumption is that π is less than π_{min} for features not observed in unmixed samples due to random sampling.

$$
\begin{aligned}
p &= P(\pi < \pi_{min} | C \geq C_{obs}) \\
&= \frac{P(C \geq C_{obs} | \pi < \pi_{min}) P(\pi < \pi_{min})}{P(C \geq C_{obs} | \pi < \pi_{exp}) P(\pi < \pi_{min}) + P(C \geq C_{obs} | \pi \geq \pi_{min}) P(\pi \geq \pi_{min})}
\end{aligned} \tag{3}
$$

$$p = \frac{P(C \geq C_{obs} | \pi < \pi_{min})}{P(C \geq C_{obs})} \tag{4}$$

2.0.7 Quantitative Assessment

Quantitative assessment compared observed relative abundance and log fold-changes to expected values derived from the titration experimental design. Feature average relative abundance across PCR replicates was calculated using a negative binomial model, and used as observed relative abundance values (*obs*) for the relative abundance assessment. Average relative abundance values were used to reduce PCR replicate outliers from biasing the assessment results. Equation (1) and inferred θ values were used to calculate the expected relative abundance values (*exp*). Relative abundance error rate is defined as $|exp - obs|/exp$.

We developed bias and variance metrics to assess feature performance. The feature-level bias and variance metrics were defined as the median error rate and robust coefficient of variation ($RCOV = IQR/median$) respectively. Mixed-effects models were used to compare feature-level error rate bias and variance metrics across pipelines with subject as a random effect. Extreme feature-level error rate bias and variance metric outliers were observed, these outliers were excluded from the mixed effects model to minimize biases due to poor model fit and were characterized independently.

Log fold-change between samples in the titration series including PRE and POST were compared to the expected log fold-change values to assess differential abundance log fold-change estimates. Log fold-change estimates were calculated using EdgeR (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Expected log fold-change for feature j between titrations l and m is calculated using equation (5), where θ is the proportion of POST bacterial DNA in a titration, and q is feature relative abundance. For features only present in PRE samples the expected log fold-change is independent of the observed counts for the unmixed samples and is calculated using (6). Due to a limited number of *PRE-specific* features, both *PRE-specific* and *PRE-dominant* features were used in the differential abundance assessment. *PRE-specific* features were defined as features observed in all four PRE PCR replicates and not observed in any of the POST PCR replicates and *PRE-dominant* features were also observed in all four PRE PCR replicates and observed in one or more of the POST PCR replicates with a log fold-change between PRE and POST samples greater than 5.

$$logFC_{lm,j} = \log_2 \left(\frac{\theta_l q_{post,j} + (1 - \theta_l) q_{pre,i}}{\theta_m q_{post,j} + (1 - \theta_m) q_{pre,j}} \right) \tag{5}$$

$$logFC_{lm,i} = log_2 \left(\frac{1 - \theta_l}{1 - \theta_m} \right) \tag{6}$$

Table 1 Summary statistics for the different bioinformatic pipelines. DADA2 is a denoising sequence inference pipeline, QIIME is an open-reference clustering pipeline, and mothur is a de-novo clustering pipeline. No template controls were excluded from summary statistics. Sparsity is the proportion of 0's in the count table. Features is the total number of OTUs (QIIME and mothur) or SVs (DADA2) in the count. Sample coverage is the median and range (minimum-maximum) per sample total abundance. Drop-out rate is the proportion of reads removed while processing the sequencing data for each bioinformatic pipeline.

Pipelines	Features	Sparsity	Total Abundance	Drop-out Rate
DADA2	3144	0.93	68649 (1661-112058)	0.24 (0.18-0.59)
Mothur	38469	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
QIIME	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

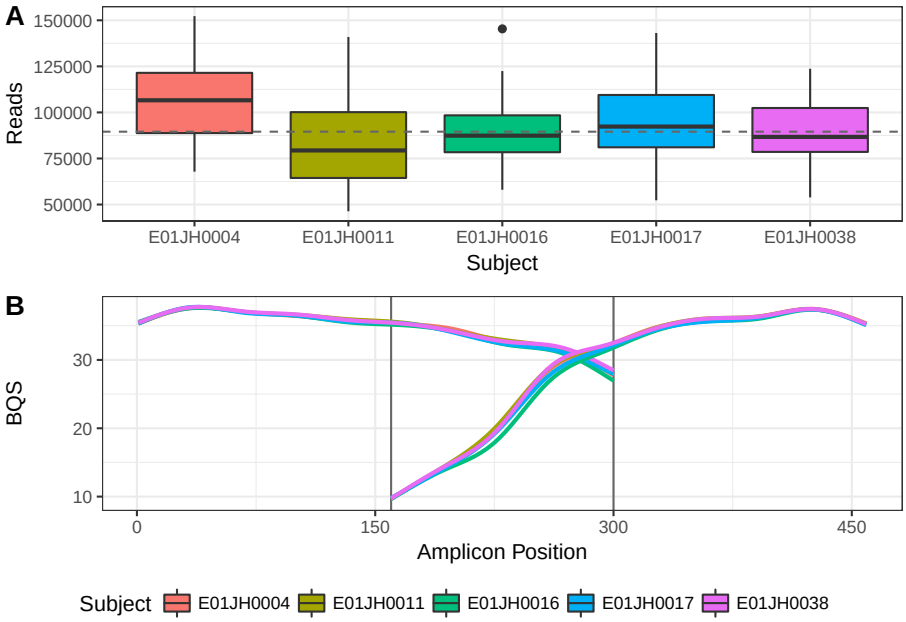

Fig. 2 Sequence dataset characteristics. (A) Distribution in the number of reads per barcoded sample (Library Size) by individual. The dashed horizontal line indicates overall median library size. Excluding one PCR replicate from subject E01JH0016 titration 5 that had only 3,195 reads. (B) Smoothing spline of the base quality score (BQS) across the amplicon by subject. Vertical lines indicate approximate overlap region between forward and reverse reads. Forward reads go from position 0 to 300 and reverse reads from 464 to 164.

3 Results

3.1 Dataset characteristics

We first characterize the number of reads per sample and base quality score distribution. The number of reads per sample and distribution of base quality scores by position was consistent across subjects (Fig. 2). Two barcoded experimental samples had less than 35,000 reads. The rest of the samples

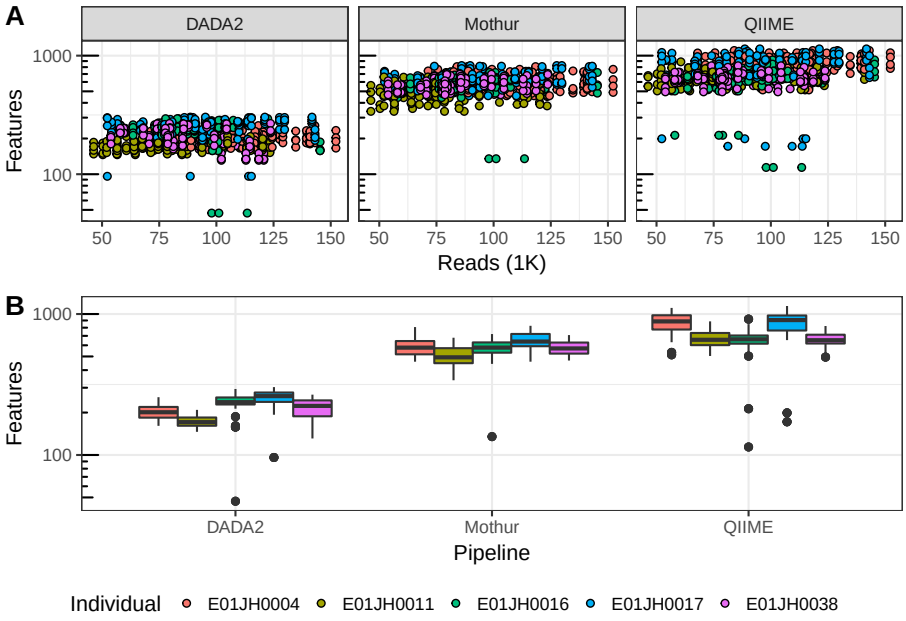

Fig. 3 Relationship between the number of reads and features per sample by bioinformatic pipeline. (A) Scatter plot of observed features versus the number of reads per sample. (B) Observed feature distribution by pipeline and individual. Excluding one PCR replicate from subject E01JH0016 titration 5 with only 3,195 reads, and the Mothur E01JH0017 titration 4 (all four PCR replicates), with 1,777 observed features.

with less than 35,000 reads were no template PCR controls (NTC). Excluding the one failed reaction with 2,700 reads and NTCs, there were 8.9548×10^4 (3195-152267, median and range) sequnces per sample. The forward read has consistently higher base quality scores relative to the reverse read with a narrow overlap region with high base quality scores for both forward and reverse reads (Fig. 2B).

The resulting count tables generated using the four bioinformatic pipelines were characterized for number of features, sparsity, and filter rate(Table 1, Figs. 3B). The pipelines evaluated employ different approaches for handling low quality reads resulting in the large differences in drop-out rate and the fraction of raw sequences not included in the count table (Table 1). QIIME pipeline has the highest drop-out rate and number of features per sample but fewer total features than Mothur. The targeted amplicon region has a relatively small overlap region, 136 bp for 300 bp paired-end reads, compared to other commonly used amplicons (Kozich et al. 2013; Walters et al. 2016). The high drop-off rate is due to low basecall accuracy at the ends of the reads especially the reverse reads resulting in a high proportion of unsuccessfully merged reads pairs (Fig. 2B). Furthermore increasing the drop-out rate, QIIME excludes singletons, OTUs only observed once in the dataset, to remove potential sequencing artifacts from the dataset. QIIME and DADA2 pipelines

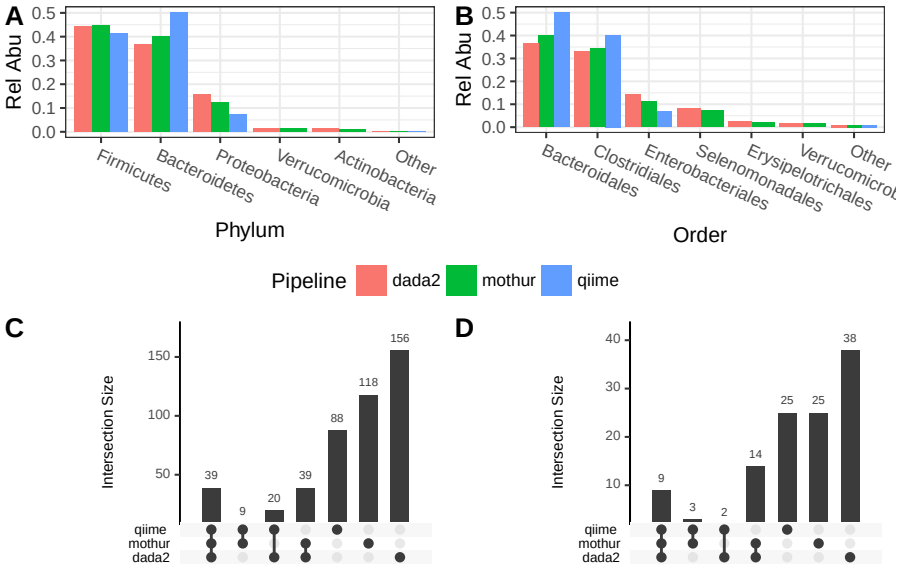

Fig. 4 Comparison of dataset taxonomic composition across pipelines. Phylum (A) and Order (B) relative abundance by pipeline. Taxonomic groups with less than 1% total relative abundance were grouped together and indicated as other. Pipeline genus-level taxonomic assignment set overlap for the all features (C) and the upper quartile genera by relative abundance for each pipeline (D).

were similarly sparse (the fraction of zero values in count tables) despite differences in the number of features and drop-out rate. The expectation is that this mixture dataset will be less sparse relative to other datasets due to the redundant nature of the samples where 35 of the samples are derived directly from the other 10 samples, and four PCR replicates for each sample. With sparsity greater than 0.9 for the three pipelines it is unlikely that any of the pipelines successfully filtered out a majority of the sequencing artifacts.

The dataset taxonomic assignments also varied by pipeline (Fig. 4). Phylum and order relative abundance is similar across pipelines (Fig. 4A & B). Differences are attributed to different taxonomic classification methods and databases. The DADA2 and QIIME pipelines differed from Mothur and QIIME for Proteobacteria and Bacteriodetes. Regardless of threshold, for genus sets most genera were unique to individual pipelines (Fig. 4C & D). Sets with QIIME had the fewest genera, excluding the DADA2-QIIME set. QIIME pipeline was the only one to use the open-reference clustering and the Greengenes database. Mothur and DADA2 both used the SILVA dataset. The Mothur and DADA2 pipeline use different implmentations of the RDP nave Bayesian classifier, which may be partially responsible for the mothur, unclustered, and DADA2 differences.

Table 2 ERCC Spike-in qPCR assay information and summary statistics. ERCC is the ERCC identifier for the ERCC spike-in, Assay is TaqMan assay, and Length and GC are the size and GC content of the qPCR amplicon. The Std. R^2 and Efficiency (E) statistics were computed for the standard curves. R^2 and slope for titration qPCR results for the titration series.

Subject	ERCC	Assay	Length	Std. R^2	E	R^2	Slope
E01JH0004	012	Ac03459877-a1	77	0.9996	86.19	0.98	0.92
E01JH0011	157	Ac03459958-a1	71	0.9995	87.46	0.95	0.90
E01JH0016	108	Ac03460028-a1	74	0.9991	87.33	0.95	0.84
E01JH0017	002	Ac03459872-a1	69	0.9968	85.80	0.89	0.93
E01JH0038	035	Ac03459892-a1	65	0.9984	86.69	0.95	0.94

3.2 Titration Series Validation

To validate the two-sample titration dataset for use in abundance assessment we evaluated two assumptions about the titrations: 1. The samples were mixed volumetrically in a log_2 dilution series according to the mixture design. 2. The unmixed PRE and POST samples have the same proportion of prokaryotic DNA. To validate the sample volumetric mixing exogenous DNA was spiked into the unmixed samples before mixing and quantified using qPCR . To evaluate if the PRE and POST samples had the same proportion of prokaryotic DNA total prokaryotic DNA in the titrations samples was quantified using a qPCR assay targeting the 16S rRNA gene.

3.2.1 Spike-in qPCR results

Titration series volumetric mixing was validated by quantify ERCC plasmids spiked into the POST samples using qPCR. The qPCR assay standard curves had a high level of precision with R^2 values close to 1 and amplification efficiencies between 0.84 and 0.9 for all standard curves indicating the assays were suitable for validating the titration series volumetric mixing (Table 2). For our log_2 two-sample-titration mixture design the expected slope of the regression line between titration factor and Ct is 1, corresponding to a doubling in template DNA every PCR cycle. The qPCR assays targeting the ERCCs spiked into the POST samples had R^2 values and slope estimates close to 1 (Table 2). Slope estimates less than one were attributed to assay standard curve efficiency less than 1 (Table 2). ERCCs spiked into PRE samples were not used to validate volumetric mixing as PRE sample proportion differences were too small for qPCR quantification. The expected C_t difference for the entire range of PRE concentrations in only 1. When considering the quantitative limitations of the qPCR assay these results confirm that the unmixed samples were volumetrically mixed according to the design.

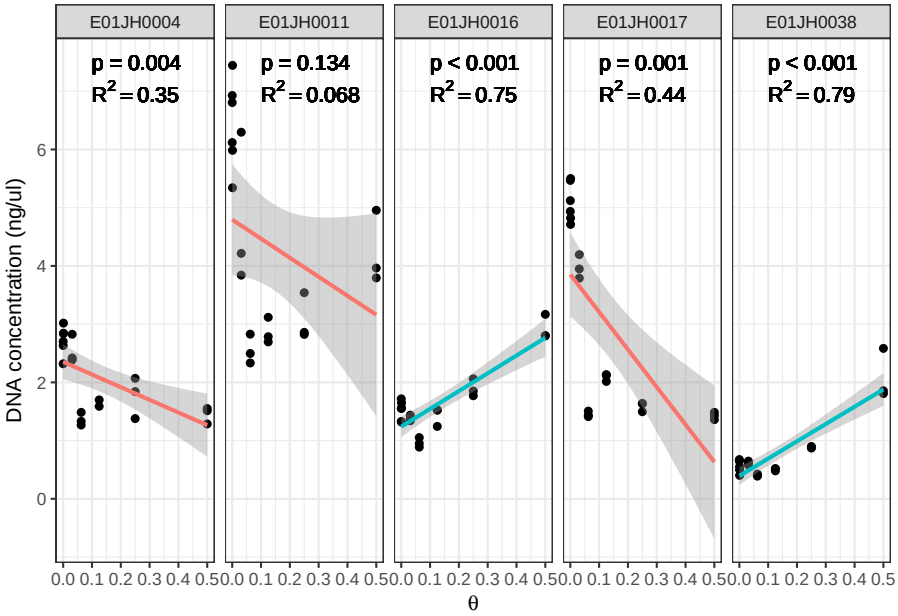

Fig. 5 Prokaryotic DNA concentration (ng/ul) across titrations measured using a 16S rRNA qPCR assay. Separate linear models, Prokaryotic DNA concentration versus θ were fit for each individual, and R^2 and p-values were reported. Red lines indicate negative slope estimates and blue lines positive slope estimates. p-value indicates significant difference from the expected slope of 0. Multiple test correction was performed using the Benjamini-Hochberg method. One of the E01JH0004 PCR replicates for titration 3 ($\theta = 0.125$) was identified as an outlier, with a concentration of 0.003, and was excluded from the linear model. The linear model slope was still significantly different from 0 when the outlier was included.

3.2.2 Bacterial DNA Concentration

The observed changes in prokaryotic DNA concentration across titrations indicate the proportion of bacterial DNA from the unmixed PRE and POST samples in a titration is inconsistent with the mixture design (Fig. 5). A qPCR assay targeting the 16S rRNA gene was used to quantify the concentration of prokaryotic DNA in the titrations. An in-house standard curve with concentrations of 20 ng/ul, 2ng/ul, and 0.2 ng/ul was used, with efficiency 91.49, and R^2 0.999. If the proportion of prokaryotic DNA is the same between PRE and POST samples the slope of the concentration estimates across the two-sample titration would be 0. For subjects where the proportion of prokaryotic DNA is higher in the PRE samples, the slope will be negative and positive when the proportion is higher for POST samples. The slope estimates are significantly different from 0 for all subjects excluding E01JH0011 (Fig. 5). These results indicate that the proportion of prokaryotic DNA is lower in POST when compared to the PRE samples for E01JH0004 and E01JH0017 and higher for E01JH0016 and E01JH0038.

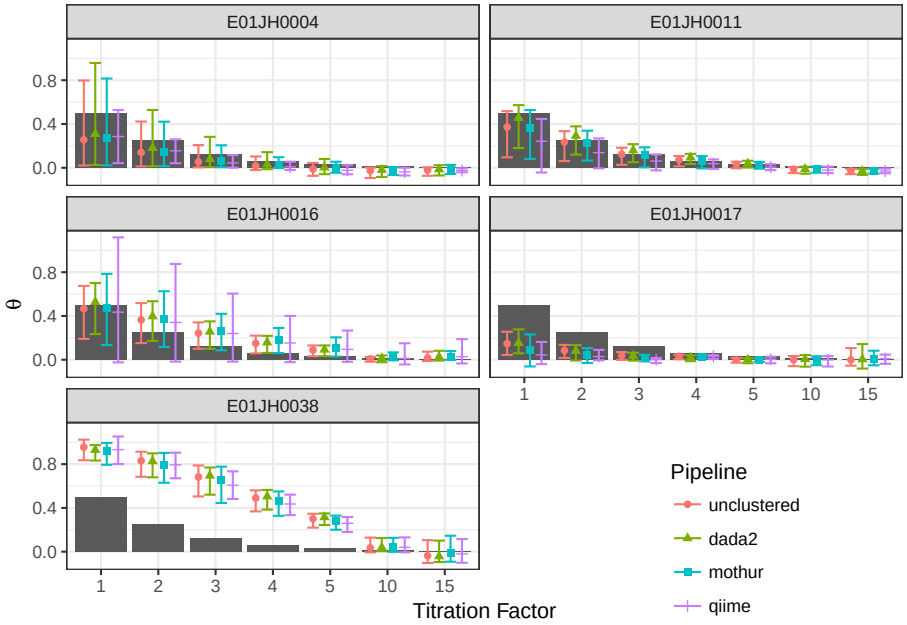

Fig. 6 Theta estimates by titration, biological replicate, and bioinformatic pipeline. The points indicates mean estimate of 1000 bootstrap theta estimates and errorbars 95% confidence interval. The black bar indicate expected theta values. Theta estimates below the expected theta indicate that the titrations contain less than expected bacterial DNA from the POST sample. Theta estimates greater than the expected theta indicate the titration contains more bacterial DNA from the PRE sample than expected.

3.2.3 Theta Estimates

To account for differences in the proportion of prokaryotic DNA in PRE and POST samples (Fig. 5) we inferred the proportion of POST sample prokaryotic DNA in a titration, θ, using the 16S rRNA sequencing data (Fig. 6). Overall the relationship between the inferred and mixture design θ values were consistent across pipelines but not subject whereas the 95% CI varied by both subject and pipeline. For study subjects E01JH0004, E01JH0011, and E01JH0016 the inferred and mixture design θ values were in agreement, in contrast, to study subjects E01JH0017 and E01JH0038. For E01JH0017 the inferred values were consistently less than the mixture design values. Whereas for E01JH0038 the inferred values were consistently greater than the mixture design values. These results were consistent with the qPCR prokaryotic DNA concentration results with significantly positive slopes for E01JH0004 and E01JH0016 and a significantly negative slope for E01JH0038 (Fig. 5).

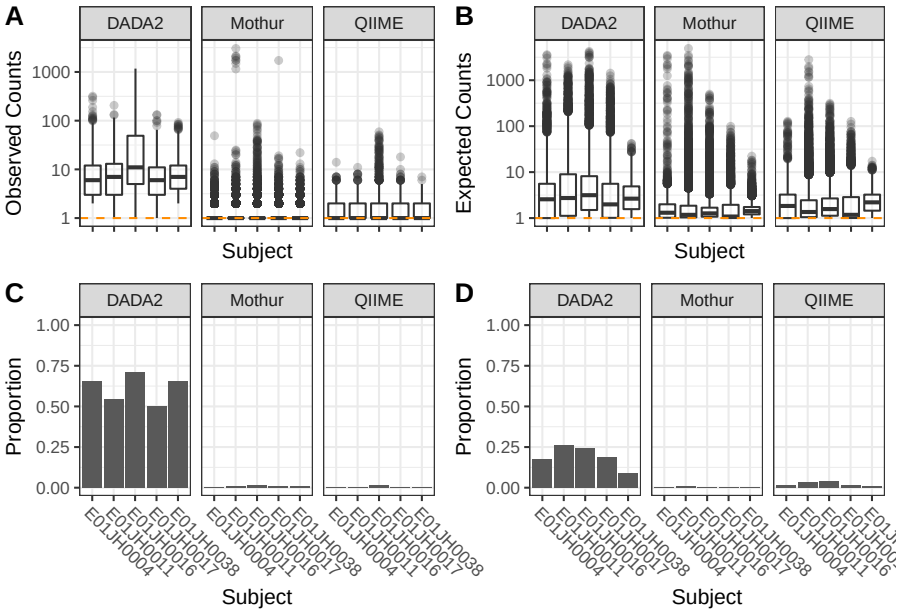

Fig. 7 Distribution of (A) observed count values for titration-specific features and (B) expected count values for unmixed-specific features by pipeline and individual. The orange horizontal dashed line indicates a count value of 1. (C) Proportion of unmix-specific features and (D) titration-specific features with an adjusted p-value < 0.05 for the Bayesian hypothesis test and binomial test respectively. We failed to accept the null hypothesis when the p-value < 0.05, indicating that the discrepancy between the feature only being observed in the titrations or unmixed samples cannot be explained by sampling alone.

3.3 Measurement Assessment

Next, we assessed the qualitative and quantitative nature of 16S rRNA measurement process using our two-sample titration dataset. For the qualitative assessment, we analyzed the relative abundance of features only observed in the unmixed samples or titrations which are not expected given the titration experimental design. The quantitative assessment evaluated relative and differential abundance estimates.

3.3.1 Qualitative Assessment

Unmixed- and titration-specific features were observed for all pipelines (titration-specific: Fig. 7A, unmixed-specific: Fig. 7B). For mixture datasets the low abundance features present only in the unmixed samples and mixtures are expected due to random sampling. For our two-sample titration dataset there were unmixed-specific features with expected counts that could not be explained by sampling alone for all individuals and bioinformatic pipelines (Fig. 7C). However, the proportion of unmixed-specific features that could not be explained by sampling alone varied by bioinformatic pipeline. DADA2 had the

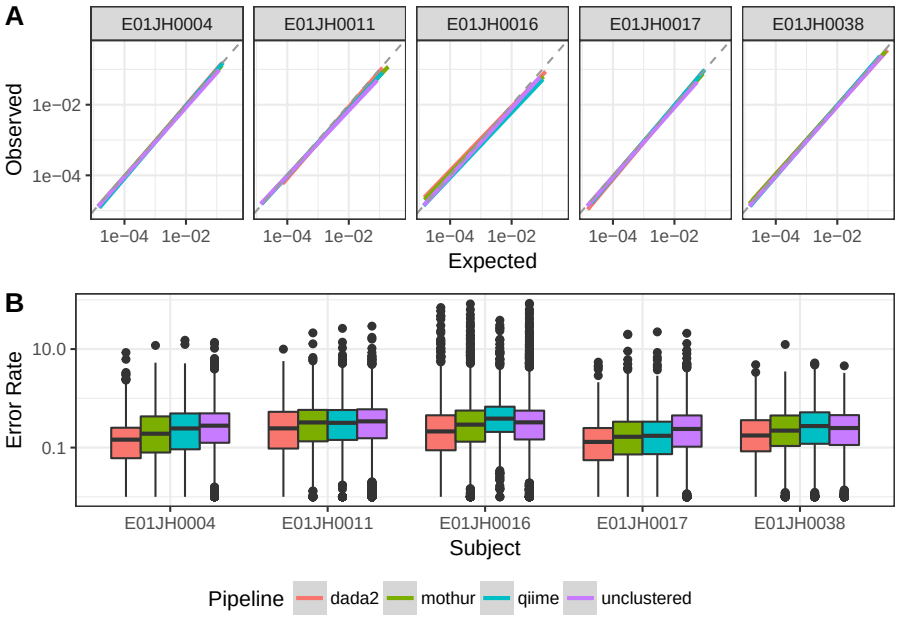

Fig. 8 Relative abundance assessment. (A) A linear model of the relationship between the expected and observed relative abundance. The dashed grey line indicates expected 1-to-1 relationship. The plot is split by individual and color is used to indicate the different bioinformatic pipelines. A negative binomial model was used to calculate an average relative abundance estimate across the four PCR replicates. Points with observed and expected relative abundance values less than 1/median library size were excluded from the data used to fit the linear model. (B) Relative abundance error rate distribution by individual and pipeline.

highest rate of unmixed-specific features not explained by sampling whereas QIIME had the lowest rate. Consistent with the distribution of observed counts for titration-specific features more of the DADA2 features could not be explained by sampling alone compared to the other pipelines (Fig. 7D). Overall, DADA2 resulted in the largest number of observed features inconsistent with the titration experiment design, while the same phenomenon is significantly reduced in the other pipelines.

3.3.2 Quantitative Assessment

For the relative abundance assessment, I evaluated the consistency of the observed and expected relative abundance estimates for a feature and titration as well as feature-level bias and variance. The PRE and POST estimated relative abundance and inferred θ values were used to calculate titration and feature level error rates. Unclustered pipeline θ estimates were used to calculate the error rates for all pipelines to prevent over-fitting. Only features observed in all PRE and POST PCR replicates and PRE and POST specific features were included in the analysis (Table 3). PRE and POST specific features were defined

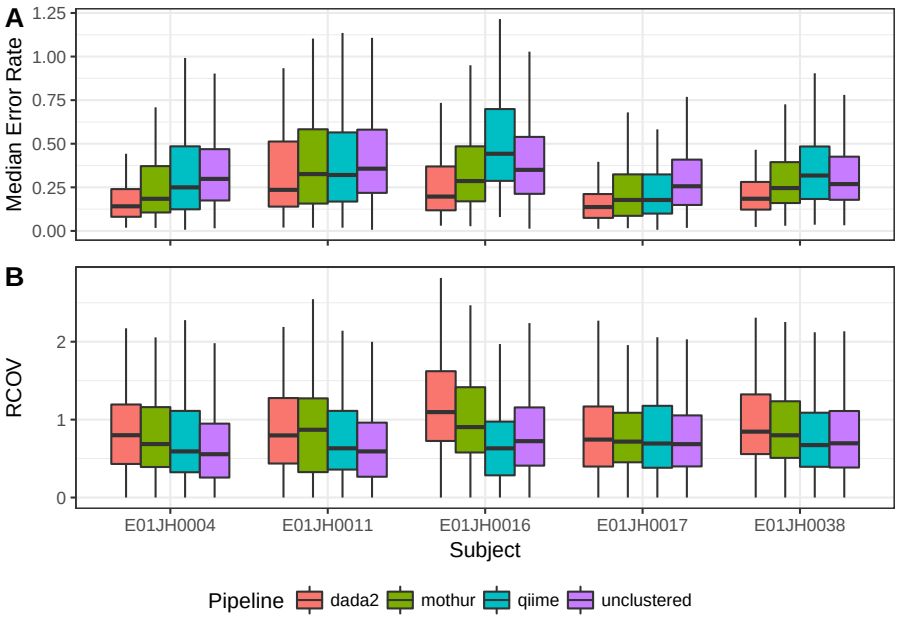

Fig. 9 Comparison of pipeline relative abundance assessment feature-level error metrics. Distribution of feature-level relative abundance (A) bias metric - median error rate and (B) variance - robust coefficient of variation ($RCOV = (IQR)/|median|$) by individual and pipeline. Boxplot outliers, $1.5 \times IQR$ from the median were excluded from the figure to prevent extreme metric values from obscuring metric value visual comparisons.

Table 3 Maximum feature-level error rate bias (median error rate) and variance (robust COV) by pipeline and individual.

Metric	Pipeline	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
Bias	dada2	2.37	2.55	17.03	4.34	0.66
	mothur	5.30	6.76	19.24	4.15	1.93
	qiime	3.99	6.43	8.83	4.80	1.09
	unclustered	6.45	7.24	16.85	4.37	1.91
Variance	dada2	4.60	8.96	7.36	5.91	6.71
	mothur	4.71	7.35	3.71	5.70	8.01
	qiime	4.40	22.57	4.46	17.10	7.91
	unclustered	7.06	10.30	16.94	8.07	6.00

as present in all four PCR replicates of the PRE or POST PCR replicates, respectively, but none of the PCR replicates for the other unmixed samples. There is lower confidence in PRE or POST feature relative abundance when the feature is not observed in some of the 4 PCR replicates, therefore these features were not included in the error analysis. Overall, agreement between the inferred and observed relative abundance was high for all individuals and

bioinformatic pipelines (Fig. 8A). The error rate distribution was similarly consistent across pipelines, including long tails (Fig. 8B)

To assess quantitative accuracy I compared the feature-level relative abundance error rate bias (median error rate, Fig. 9A) and variance ($RCOV = (IQR)/|median|$ Fig. 9B) across pipelines and individuals using mixed effects models. Large bias and variance values were observed for all pipelines (Table 3). Features with large bias and variance metrics (outliers), defined as $1.5 \times IQR$ from the median. To prevent the outliers from biasing the comparison they were not included in the dataset used to fit the mixed effects model. Multiple comparisons test (Tukey) was used to test for significant differences in feature-level bias and variance between pipelines. A one-sided alternative hypothesis was used to determine which pipelines had a smaller, feature-level error rate. The Mothur, DADA2, and QIIME feature-level bias were all significantly different from each other ($p < 1 \times 10^{-8}$). DADA2 had the lowest mean feature-level bias (0.2), followed by Mothur (0.28), with QIIME having the highest bias (0.33) (9B). Large variance metric values were observed for all individuals and pipelines (Table 3). The feature-level variance was not significantly different between pipelines, Mothur = 0.83, QIIME = 0.71 and DADA2 = 1 (Fig. 9B). I evaluated whether poor feature-level relative abundance metrics can be attributed to specific taxonomic groups or phylogenetic clades. While a significant overall phylogenetic signal was detected for both the bias and variance metric, I was unable to identify specific taxonomic groups or phylogenetic clades exceedingly poor performance in our assessment.

The agreement between the log-fold change estimates and expected values were individual specific and consistent across pipelines (Fig. 10A). The individual specific effect was attributed to the fact that unlike the relative abundance assessment the inferred θ values were not used to calculate the expected values. The inferred θ values were not used to calculate the expected values as I wanted to include all of the titrations and the θ estimates for the higher titrations were not monotonically decreasing and therefore resulted in unrealistic expected log fold-change values, e.g., negative log-fold changes for PRE specific features. The log-fold change estimates and expected values were consistent across pipelines with one notable exception. For E01JH0011 the Mothur log fold-change estimates were in better agreement with the expected value compared to the other pipelines. However, as θ was not corrected for differences in the proportion of prokaryotic DNA between the unmixed PRE and POST samples I am unable to say whether Mothur's performance was better than the other pipelines.

The log fold-change error distribution was consistent across pipelines (Fig. 10B). There was a long tail of high error features in the error distribution for all pipelines and individuals. The log fold-change estimates responsible for the long tail could not be attributed to specific titration comparisons. Additionally, I compared the log-fold change error distribution for log-fold change estimates using different normalization methods. The error rate distributions, including the long tails, were consistent across normalization methods. Furthermore, as the long tail was observed for the unclustered data as well, the log-fold change

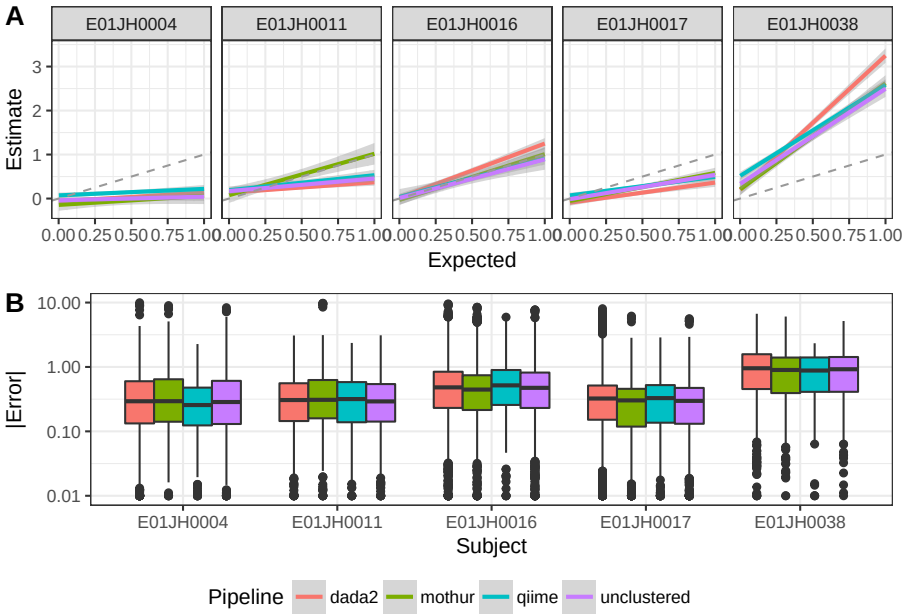

Fig. 10 (A) Linear model or the relationship between log fold-change estimates and expected values for PRE-specific and PRE-dominant features by pipeline and individual, line color indicates pipelines. Dashed grey line indicates expected 1-to-1 relationship between the estimated and expected log fold-change. (B) Log fold-change error (|exp-est|) distribution by pipeline and individual.

estimates contributing to the long tail are likely due to a bias associated with the molecular laboratory portion of the measurement process and not the bioinformatic pipelines. Based on exploratory analysis of the relationship between the log fold-change estimates and expected values for individual features indicated that the long tails were attributed to feature specific performance.

Feature-level log fold-change bias and variance metrics were used to compare pipeline performance (Fig. 10). Feature-level bias and variance metrics are defined as the $1 - slope$ and R^2 for linear models of the estimated and expected log fold-change for individual features and all titration comparisons. For the bias metric, $1 - slope$, the desired value is 0 (i.e., log fold-change estimate = log fold-change expected), with negative values indicating the log-fold change was consistently underestimated and positive values consistently overestimated. The linear model R^2 value was used to characterize the feature-level log fold-change variance as it indicates how consistent the relationship between log fold-change estimates and expected values is across titration comparisons. To compare bias and variance metrics across pipelines mixed-effects models were used. The log fold-change bias and variance metrics were not significantly different between pipelines (Bias: F = 0, 2.51, p = 0.99, 0.08, 10B, Variance: F = 47.39, 0.23, p = 0, 0.8, Fig. 10C). Next, I evaluated whether poor feature-level metrics could be attributed to specific clades for taxonomic

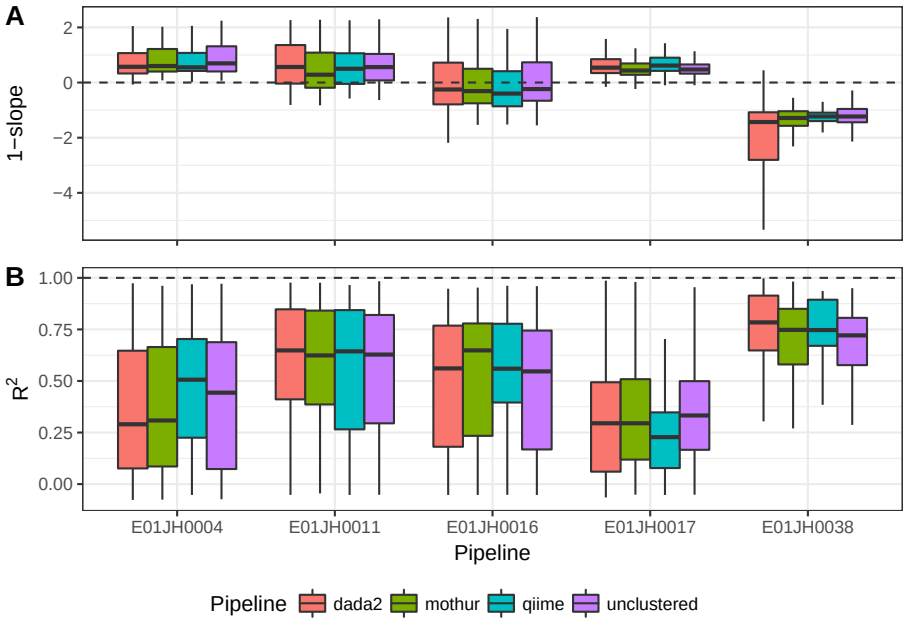

Fig. 11 Feature-level log-fold change error bias (A) and variance (B) metric distribution by subject and pipeline. The bias ($1 - slope$) and variance (R^2) metrics are derived from the linear model fit to the estimated and expected log fold-change values for individual features. Boxplot outliers, $1.5 \times IQR$ from the median were excluded from the figure to prevent extreme metric values from obscuring metric value visual comparisons.

groups. Similar to the relative abundance estimate, while a phylogenetic signal was detected for both the bias and variance metrics, I was unable to identify specific taxonomic groups or phylogenetic clades that performed poorly in our assessment.

4 Discussion

We assessed the quantitative and qualitative characteristics of count tables generated using different bioinformatic pipelines and 16S rRNA marker-gene survey mixture dataset. The mixture dataset followed a two-sample titration mixture design, where DNA collected before and after exposure to pathogenic *Escherichia coli* from five vaccine trial participants (subjects) were mixed following a log_2 dilution series (Fig. 1). Qualitative count table characteristics were assessed using relative abundance information for features observed only in titrations and unmixed samples. We quantitatively assed count tables by comparing feature relative and differential abundance to expected values.

4.0.1 Count Table Assessment Demonstration

We demonstrated our novel assessment approach by evaluating count tables generated using different bioinformatic pipelines, QIIME, Mothur, and DADA2. The Mothur pipeline uses *de novo* clustering for feature inference (Westcott and Schloss 2017; Schloss et al. 2009). Pairwise distances used in clustering are calculated using a multiple sequence alignment. The quality filtered paired-end reads are merged into contigs. The pipeline the aligns contigs to a reference multiple sequence alignment and removes uninformative positions in the multiple sequence alignment. The QIIME pipeline uses open-reference clustering where merged paired-end reads are first assigned to reference cluster centers (Rideout et al. 2014; Caporaso et al. 2010). Next QIIME clusters unassigned reads *de novo*. Unlike Mothur, the QIIME clustering method uses pairwise sequence distances calculated from pairwise sequence alignments. As a result, the QIIME pairwise distances are calculated using the full ˜436 bp sequences whereas Mothur pairwise distances were calculated using a 270 bp multiple sequence alignment. The DADA2 pipeline uses a probability model and maximization expectation algorithm for feature inference (B. J. Callahan et al. 2016). Unlike distance-based clustering methods employed by the Mothur and QIIME pipelines, DADA2 parameters determine if low abundance sequences are grouped with a higher abundance sequence. As a control, we compared our quantitative assessment results for the three pipelines to a count table of unclustered features. The unclustered features were generated using the Mothur pipeline preprocessing methods.

Quantitative Assesssment While the relative abundance bias metric was significantly different between pipelines overall, pipeline choice had minimal impact on the quantitative assessment results when accounting for subject-specific effects. Outlier features, those with extreme quantitative analysis bias and variance metrics, were observed for all pipelines and both relative and differential abundance assessments. Outlier features could not be attributed to bioinformatic pipelines and are likely due to biases in the molecular biology part of the measurement process. Outlier features are not likely a pipeline artifact as they were observed in count tables generated using the unclustered pipeline as well as standard bioinformatic pipelines. We were unable to attribute outlier features to relative abundance values, log fold-change between unmixed samples, and sequence GC content. Features with extreme metric values were not limited to any specific taxonomic group or phylogenetic clade. PCR amplification is a well-known bias molecular biology component of the measurement process. Mismatches in the primer binding regions impact PCR efficiency and are a potential cause for poor feature-specific performance (Wright et al. 2014). Additional research is needed before outlier features are attributed to mismatches in the primer binding regions.

Qualitative Assessment The qualitative assessment evaluated whether features only observed in unmixed samples or titrations could be explained by

sampling alone. Features present only in the titrations or unmixed samples not due to random sampling are bioinformatic pipeline artifacts. These artifacts can be categorized as false negative or false positive features. A false negative occurs when a lower abundance sequence representing an organism within the sample is clustered with a higher abundance sequence from a different organism. False positives are sequencing or PCR artifacts not appropriately filtered or assigned to an appropriate feature by the bioinformatic pipeline.

Count table sparsity, the proportion of zero-valued cells, provides additional insight into the qualitative assessment results. A high rate of false negative features is a potential explanation for the DADA2 count table's poor performance in the qualitative assessment and comparable sparsity to the other pipelines despite having significantly fewer features (Fig. 7, 1). The DADA2 feature inference algorithm may be aggressively grouping lower abundance true sequences with higher abundance sequences. As a result, the low abundance sequences are not present in samples leading to increased sparsity and higher abundance unmixed- and titration-specific features. Adjusting the DADA2 parameters, specifically the `OMEGA_A` parameter in `setDadaOpt`. Along these lines, the DADA2 documentation states that the default setting for `OMEGA_A` is conservative to prevent false positives at the cost of increasing false negatives (B. J. Callahan et al. 2016).

False positive features provide an explanation for Mothur and QIIME pipelines having lower proportion of unmixed- and titration-specific features not explained by sampling but high sparsity (Fig. 7, 1). The statistical tests used to determine if the specific features could be explained by sampling only considers feature abundance. Therefore, the statistical test is not able to distinguish between true low abundance unmixed- and titration-specific features and low abundance sequence artifacts. Mothur and QIIME count tables have ten times and three times more features compared to DADA2, respectively (Table 1). While microbial abundance distributions are known to have long tails, it is likely that the observed sparsity is an artifact of the 16S rRNA sequencing measurement process. Similarly, significantly more features than expected are commonly observed for mock community benchmarking studies evaluating the QIIME and Mothur pipelines (Kozich et al. 2013).

False positive features can be reduced, but not eliminated, using smaller amplicon and prevalence filtering. The 16S rRNA region sequenced in the study is larger than the region the *de-novo*, and open clustering pipelines were initially developed for, potentially explaining the higher than expected sparsity (Kozich et al. 2013). Kozich et al. (2013) were reduced the sequence error rate from 0.29% to 0.06% when using paired-end reads that completely overlap. The larger region has a smaller overlap between the forward and reverse reads. As a result merging of the forward and reverse reads did not allow for the sequence error correction that occurs when a smaller amplicon is used. However, even when targeting smaller regions of the 16S rRNA gene both the *de-novo* (Mothur) and open-reference clustering (QIIME) pipelines produced count tables with significantly more features than expected in evaluation studies using mock communities. Prevalence filtering is used to exclude low abundance fea-

tures, likely predominantly measurement artifacts (B. Callahan et al. 2016). For example, a study exploring the microbial ecology of the Red-necked stint *Calidris ruficollis*, a migratory shorebird, used a hard filter to validate their study conclusions are not biases by false positive features. The study authors compared results with and without prevalence filter ensuring that the study conclusions were not biased by using the arbitrary filter or including the low abundant features (Risely et al. 2017).

4.0.2 Using Mixtures to Assess 16S rRNA Sequencing

Mixtures of environmental samples have previously been used to assess RNAseq and microarray gene expression measurements. However, this is the first time mixtures have been used to assess microbiome measurement methods. Our mixture dataset allowed us to develop novel methods for assessing marker-gene-survey computational methods. Our quantitative assessment allowed for the characterization of relative abundance values using a dataset with a larger number of features and dynamic range compared to assessments using mock communities. As a result, we were able to identify previously unknown feature specific biases. Based on our study results additional experiments can be performed to identify the cause of these biases and develop appropriate methods to account for them. Based on our subject-specific results observation, we recommend that studies based on stool samples seeking inferences in a longitudinal series of multiple subjects carefully estimate bacterial DNA proportions and adjust inferences accordingly. Additionally, our qualitative assessment results, when combined with sparsity information provide a new method for evaluating how well bioinformatic pipelines account for sequencing artifacts without loss of true biological sequences.

There were also limitations using our mixture dataset. These limitations included: Lack of agreement between the proportion of unmixed samples titrations and the mixture design. The number of features used in the different analysis. These limitations are described below along with recommendations for addressing them in future studies.

Differences in the proportion of prokaryotic DNA in the samples used to generate the two-sample titrations series results in differences between the true mixture proportions and mixture design. We attempted to account for differences in mixture proportion from mixture design by estimating mixture proportions using sequence data. Similar to how the proportion of mRNA in RNA samples used in a previous mixture study. We were able to use an assay targeting the 16S rRNA gene to detect changes in the concentration of bacterial DNA across titration, but unable to quantify the proportion of bacterial DNA in the unmixed samples using qPCR data. Using the 16S sequencing data we inferred the proportion of bacterial DNA from the POST sample in each titration. However, the uncertainty and accuracy of the inference method are not known resulting in an unaccounted for error source.

A better method for quantifying sample bacterial DNA proportion or using samples with consistent proportions would increase the expected value

and in-turn error metric accuracy. Limitations in the prokaryotic DNA qPCR concentration assay precision limit the suitability for use in mixture studies. Digital PCR provides a more precise alternative to qPCR and is, therefore, a more appropriate method. Alternatively using samples where the majority of the DNA is prokaryotic would minimize this issue. Mixtures of environmental samples can also be used to assess shotgun metagenomic methods as well. As shotgun metagenomics is not a targeted approach, differences in the proportion of bacterial DNA in a sample would not impact the assessment results in the same way as 16S rRNA marker-gene-surveys.

Using samples from a vaccine trial allowed for the use of a specific marker with an expected response, *E. coli*, during methods development. However, the high level of similarity between the unmixed samples resulted in a limited number of features that could be used in the quantitative assessment results. Using more diverse samples to generate mixtures would address this issue.

4.1 Conclusions

This two-sample-titration dataset can be used to evaluate and characterize bioinformatic pipelines and clustering methods. The sequence dataset presented in this study can be processed with any 16S bioinformatic pipeline. Our quantitative and qualitative assessment can then be performed on the count table and the results compared to those obtained using the pipelines included in this study. The threee pipelines we evaluated produced sets of features varying in total feature abundance, number of features per samples, and total features. The objective of any pipeline is to differentiate true biological sequences from artifacts of the measurement process. In general based on our evaluation results we recommend using for DADA2 for feature-level abundance analysis, e.g. differential abundance testing. While DADA2 performed poorly in our qualitative assessment, the pipeline had performed better in the quantitative assessment compared to the other pipelines. Additionally, the DADA2 poor qualitative assessment results due to false-negative features are unlikely to negatively impact feature-level abundance analysis, though additional research is needed to validate this claim. When determining which pipeline to use for a study, users should consider whether minimizing false positives (DADA2) or false negatives (Mothur) is more appropriate for their study objectives. When a sequencing dataset is processed using DADA2, the user can be more confident that an observed feature represents a member of the microbial community and not a measurement artifact. Pipeline parameter optimization could address DADA2 false-negative issue. For the Mothur and QIIME pipelines, prevalence filtering will reduce the number of false-positive features. Feature-level results for any 16S rRNA marker-gene survey should be interpreted with care, as the biases responsible for poor quantitative assessment are unknown. Addressing both of these issues requires advances in both the molecular biology and computational components of the measurement process.

References

Aronesty, Erik. 2011. "Ea-Utils: Command-Line Tools for Processing Biological Sequencing Data." *Expression Analysis, Durham, NC*.

Baker, Shawn C, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, et al. 2005. "The External Rna Controls Consortium: A Progress Report." *Nature Methods* 2 (10). Nature Publishing Group: 731–34.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 289–300.

Bokulich, Nicholas A, Jai Ram Rideout, William G Mercurio, Arron Shiffer, Benjamin Wolfe, Corinne F Maurice, Rachel J Dutton, Peter J Turnbaugh, Rob Knight, and J Gregory Caporaso. 2016. "Mockrobiota: A Public Resource for Microbiome Bioinformatics Benchmarking." *mSystems* 1 (5). Am Soc Microbiol: e00062–16.

Brooks, J Paul, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, et al. 2015. "The Truth About Metagenomics: Quantifying and Counteracting Bias in 16S rRNA Studies." *BMC Microbiology* 15 (1). BioMed Central: 66.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13: 581–83. https://doi.org/10.1038/nmeth.3869.

Callahan, BJ, K Sankaran, JA Fukuyama, PJ McMurdie, and SP Holmes. 2016. "Bioconductor Workflow for Microbiome Data Analysis: From Raw Reads to Community Analyses [Version 2; Referees: 3 Approved]." *F1000Research* 5 (1492). https://doi.org/10.12688/f1000research.8986.2.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (April). Nature Publishing Group SN -: 335. http://dx.doi.org/10.1038/nmeth.f.303.

Costea, Paul I, Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, et al. 2017. "Towards Standards for Human Fecal Sample Processing in Metagenomic Studies." *Nat. Biotechnol.* 35 (October). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 1069.

D'Amore, Rosalinda, Umer Zeeshan Ijaz, Melanie Schirmer, John G Kenny, Richard Gregory, Alistair C Darby, Christopher Quince, and Neil Hall. 2016. "A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling." *BMC Genomics* 17. BMC Genomics: 1–40. https://doi.org/10.1186/s12864-015-2194-9.

DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. 2006. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and

Workbench Compatible with Arb." *Applied and Environmental Microbiology* 72 (7). Am Soc Microbiol: 5069–72.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than Blast." *Bioinformatics* 26 (19). Oxford University Press: 2460–1.

Edgar, Robert C, Brian J Haas, Jose C Clemente, Christopher Quince, and Rob Knight. 2011. "UCHIME Improves Sensitivity and Speed of Chimera Detection." *Bioinformatics* 27 (16). Oxford Univ Press: 2194–2200.

Gohl, Daryl M, Pajau Vangay, John Garbe, Allison MacLean, Adam Hauge, Aaron Becker, Trevor J Gould, et al. 2016. "Systematic Improvement of Amplicon Marker Gene Methods for Increased Accuracy in Microbiome Studies." *Nat. Biotechnol.*, July.

Goodrich, Julia K, Sara C Di Rienzi, Angela C Poole, Omry Koren, William A Walters, J Gregory Caporaso, Rob Knight, and Ruth E Ley. 2014. "Conducting a Microbiome Study." *Cell* 158 (2). Elsevier: 250–62.

Hansen, Martin Christian, Tim Tolker-Nielsen, Michael Givskov, and Sren Molin. 1998. "Biased 16S rDNA PCR Amplification Caused by Interference from DNA Flanking the Template Region." *FEMS Microbiol. Ecol.* 26 (2). Oxford University Press: 141–49.

Harro, Clayton, Subhra Chakraborty, Andrea Feller, Barbara DeNearing, Alicia Cage, Malathi Ram, Anna Lundgren, et al. 2011. "Refinement of a Human Challenge Model for Evaluation of Enterotoxigenic Escherichia Coli Vaccines." *Clinical and Vaccine Immunology* 18 (10). Am Soc Microbiol: 1719–27.

Huber, W., Carey, V. J., Gentleman, R., Anders, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`.

Huse, Susan M, David Mark Welch, Hilary G Morrison, and Mitchell L Sogin. 2010. "Ironing out the wrinkles in the rare biosphere through improved OTU clustering." *Environmental Microbiology* 12 (7): 1889–98. `https://doi.org/10.1111/j.1462-2920.2010.02193.x`.

Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jrg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glckner. 2012. "Evaluation of General 16S Ribosomal Rna Gene Pcr Primers for Classical and Next-Generation Sequencing-Based Diversity Studies." *Nucleic Acids Research*. Oxford Univ Press, gks808.

Kopylova, Evguenia, Jose A Navas-molina, Cline Mercier, and Zech Xu. 2014. "Open-Source Sequence Clustering Methods Improve the State Of the Art." *mSystems* 1 (1): 1–16. `https://doi.org/10.1128/mSystems.00003-15.Editor`.

Kozich, James J, Sarah L Westcott, Nielson T Baxter, Sarah K Highlander, and Patrick D Schloss. 2013. "Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the Miseq Illumina Sequencing Platform." *Applied and Environmental Microbiology* 79 (17). Am Soc Microbiol: 5112–20.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Res.* 40 (10): 4288–97.

McMurdie, Paul J, and Susan Holmes. 2014. "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." *PLoS Comput. Biol.* 10 (4): e1003531.

Olson, Nathan D, and Jayne B Morrow. 2012. "DNA Extract Characterization Process for Microbial Detection Methods Development and Validation." *BMC Res. Notes* 5 (December): 668.

Parsons, Jerod, Sarah Munro, P Scott Pine, Jennifer McDaniel, Michele Mehaffey, and Marc Salit. 2015. "Using Mixtures of Biological Samples as Process Controls for Rna-Sequencing Experiments." *BMC Genomics* 16 (1). BioMed Central: 708.

Pine, P Scott, Barry A Rosenzweig, and Karol L Thompson. 2011. "An Adaptable Method Using Human Mixed Tissue Ratiometric Controls for Benchmarking Performance on Gene Expression Microarrays in Clinical Laboratories." *BMC Biotechnology* 11 (1). BioMed Central: 38.

Pinto, Ameet J, and Lutgarde Raskin. 2012. "PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets." *PLoS One* 7 (8): e43093.

Pop, Mihai, Joseph N Paulson, Subhra Chakraborty, Irina Astrovskaya, Brianna R Lindsay, Shan Li, Hctor Corrada Bravo, et al. 2016. "Individual-Specific Changes in the Human Gut Microbiota After Challenge with Enterotoxigenic Escherichia Coli and Subsequent Ciprofloxacin Treatment." *BMC Genomics* 17 (1). BioMed Central: 1.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jrg Peplies, and Frank Oliver Glckner. 2012. "The Silva Ribosomal Rna Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1). Oxford University Press: D590–D596.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Rideout, Jai Ram, Yan He, Jose A Navas-Molina, William A Walters, Luke K Ursell, Sean M Gibbons, John Chase, et al. 2014. "Subsampled Open-Reference Clustering Creates Consistent, Comprehensive OTU Definitions and Scales to Billions of Sequences." *PeerJ* 2 (August): e545.

Risely, Alice, David Waite, Beata Ujvari, Marcel Klaassen, and Bethany Hoye. 2017. "Gut Microbiota of a Long-Distance Migrant Demonstrates Resistance Against Environmental Microbe Incursions." *Molecular Ecology*. Wiley Online Library.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported

Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23). Am Soc Microbiol: 7537–41.

Souza, Welliton, and Benilton Carvalho. 2017. *Rqc: Quality Control Tool for High-Throughput Sequencing Data*. `https://github.com/labbcb/Rqc`.

Thompson, Karol L, Barry A Rosenzweig, P Scott Pine, Jacques Retief, Yaron Turpaz, Cynthia A Afshari, Hisham K Hamadeh, et al. 2005. "Use of a Mixed Tissue Rna Design for Performance Assessments on Multiple Microarray Formats." *Nucleic Acids Research* 33 (22). Oxford University Press: e187–e187.

Walters, William, Embriette R Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A Gilbert, et al. 2016. "Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys." *mSystems* 1 (1).

Wang, Qiong, George M Garrity, James M Tiedje, and James R Cole. 2007. "Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16). Am Soc Microbiol: 5261–7.

Westcott, Sarah L, and Patrick D Schloss. 2017. "OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units." *mSphere* 2 (2).

Wright, Erik S, L Safak Yilmaz, Sri Ram, Jeremy M Gasser, Gregory W Harrington, and Daniel R Noguera. 2014. "Exploiting Extension Bias in Polymerase Chain Reaction to Improve Primer Specificity in Ensembles of Nearly Identical Dna Templates." *Environmental Microbiology* 16 (5). Wiley Online Library: 1354–65.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16S rRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17 (1). BioMed Central: 1.