RESEARCH

A framework for assessing 16S marker gene survey data analysis methods using mixtures.

Nathan D. Olson[1,2,3*], M. Senthil Kumar[2,3], Shan Li[4], Domenick J. Braccia[2,3], Stephanie Hao[5], Winston Timp[5], Marc L. Salit[6], O. Colin Stine[4] and Hector Corrada Bravo[2,3,7]

Correspondence: nolson@nist.gov
Biosystems and Biomaterials
Division, National Institute of
tandards and Technology, 100
Bureau Dr., Gaithersburg,
Maryland, 20899 USA
ull list of author information is
vailable at the end of the article

Abstract

Background: Analysis of 16S rRNA marker-gene surveys ~~, used to characterize prokaryotic microbial communities,~~ may be performed by ~~numerous~~ a variety of bioinformatic pipelines and downstream analysis methods. However, appropriate assessment datasets and statistics are needed as there is limited guidance ~~on how~~ to decide between ~~methods, appropriate data sets and statistics for assessing these methods are needed. We developed a mixture dataset with real data complexity and an expected value for assessing~~ available analysis methods. Mixtures of environmental samples are useful for assessment as they provide values calculated from measurements of the unmixed samples and the mixture design that can be compared to values recovered by each bioinformatic method. While experiments mixing complex samples have been used to assess other sequencing methods such as RNAseq, they have yet to be used to assess 16S rRNA sequencing.

Results: We developed an assessment framework for 16S rRNA ~~bioinformatic pipelines and downstream analysis methods . We generate an assessment dataset using~~ sequencing analysis methods based on a two-sample titration mixture ~~design. The sequencing data were processed using multiple bioinformatic pipelines ,~~ dataset and metrics to evaluate OTU count table characteristics. Our qualitative assessment evaluates feature presence/absence exploiting features only present in unmixed samples or titrations by testing if random sampling can explain their observed relative abundance. Our quantitative assessment evaluates how well relative and differential abundance values agree with values expected from the mixture design. We evaluated count tables generated by three commonly used bioinformatic pipelines as demonstration: i) DADA2 a sequence inference method, ii) Mothur a *de novo* clustering method, and iii) QIIME ~~with~~ which uses open-reference clustering. ~~The mixture dataset was used to qualitatively and quantitatively assess count tables generated using the pipelines. The qualitative assessment was used to evalute features only present in unmixed samples and titrations. The abundance~~ Qualitative assessment indicated that the majority of Mothur and QIIME features specific to unmixed samples ~~and~~ or titrations were explained by ~~sampling alone. However, for~~ random sampling alone but not DADA2 ~~over a third of the unmixed sample and titration specific feature abundance could not be explained by sampling alone. The quantitative assessment evaluated pipeline performance by comparing observed to expected relative and differential abundance values. Overall the observed relative abundance~~ features. When combined with assessments of count table sparsity, these results indicate that DADA2 has a higher false negative rate whereas Mothur and QIIME have higher false positive rates. Quantitative assessment indicated that, overall, observed relative abundance and differential abundance values were consistent with ~~the expected values . Though outlier features were observed across all pipelines .~~ expected values for all three pipelines. We also identified subsets of features measured with high error by all pipelines evaluated. We could not identify the source of bias in these poor performing features based on previously studied sources of bias, indicating that further analysis of potentially unknown and unaccounted for biases is warranted.

Conclusions: ~~Using a novel mixture dataset and assessment methods we quantitatively and qualitatively~~ We developed a novel framework for assessing 16S rRNA marker-gene survey analysis methods based on mixture experiments. To demonstrate the assessment framework we evaluated count tables generated using three bioinformatic pipelines. The ~~dataset and methods~~ assessment framework developed for this study will serve as a valuable community resource for assessing 16S rRNA marker-gene survey bioinformatic methods.

Keywords: 16S rRNA gene; assessment; bioinformatic pipeline; normalization; differential abundance

Background

Targeted sequencing of the 16S rRNA gene ~~, commonly known as 16S rRNA marker-gene-surveys, is a commonly used method for characterizing microbial communities, microbiomes~~is commonly used to characterize microbial communities. The 16S rRNA marker-gene-survey measurement process includes molecular ~~(e.g. PCR and sequencing) and computational steps (e.g., sequence clustering) [1]. Molecular steps are used~~ steps to selectively target and sequence the 16S rRNA gene from prokaryotic organisms within a sample ~~. The computational steps~~ and computational steps [1] computational steps convert the raw sequence data into a ~~matrix with feature (e.g., operational taxonomic units)~~ count table of feature relative abundance values ~~, feature abundance relative to all other features, for each sample~~ [1]. Both molecular and computational measurement ~~process steps~~ processes contribute to the overall measurement bias and dispersion [2, 1, 3]. ~~Proper measurement method evaluation allows for~~ The need for datasets characterizing complex microbial communities with some degree "ground truth" has emerged in order to properly characterize the accuracy of 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.~~ marker-gene-survey measurement process.

Diverse bioinformatic pipelines used to generate count tables produce data with diverse characteristics. For example the commonly used QIIME, Mothur, and DADA2 pipelines produce feature sets and count tables with different characteristics. Mothur uses *de novo* clustering for feature inference [4, 5]. Pairwise distances used in clustering are calculated from a multiple sequence alignment. Quality filtered paired-end reads are merged into contigs, then aligned to a reference multiple sequence alignment, followed by the removal of uninformative positions. As a result the feature set representative sequences are shorter than the input amplicons. For the QIIME open-reference clustering pipeline merged paired-end reads are first assigned to reference cluster centers [6, 7]. Next, unassigned reads are clustered *de novo*. Unlike Mothur, the QIIME pipeline clustering method uses pairwise sequence distances calculated from pairwise sequence alignments. As a result, the QIIME pairwise distances are calculated using the full amplicon sequence, whereas Mothur pairwise distances are calculated using multiple sequence alignment with only informative positions. The DADA2 pipeline uses a probability model and maximization expectation algorithm for feature inference [8]. 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.

Numerous studies have evaluated quantitative and qualitative characteristics of the 16S rRNA measurement process using mock communities, simulated data, and environmental samples.

~~To~~ Mock communities are commonly used to assess the qualitative characteristics of the 16S rRNA sequencing measurement process ~~mock communities are commonly used [9]. As the number of organisms in the mock community is known, the total~~

[9]. The use of mock communities in this fashion shows that surveys often result in number of features ~~can be compared to the expected value. The number of observed features in a mock community is significantly higher than the expected number of organism~~ that are significantly higher than the underlying features in the mock community [10]. The higher than expected number of features is often attributed to sequencing and PCR artifacts as well as reagent contaminants [3, 11]. A notable exception ~~to this is mock community benchmarking studies evaluating sequencing inference method~~is count tables generated using feature inference methods, such as DADA2 [8]. Sequence inference methods which aim to reduce the number of ~~sequence artifact features . While mock communities have an expected number of features and composition~~features from sequence artifacts by using statistical models to group sequences by both similarity and abundance. Nonetheless, while mock communities are useful in this type of assessment , they lack the ~~feature diversity and relative abundance~~ diversity and dynamic range of feature present in real samples [9].

~~The quantitative characteristics~~ Quantitative assessment 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 [10]~~is informative but provides an incomplete characterization of the measurement process. Results from relative abundance estimates using mock communities generated from mixtures of single organism's DNA have shown taxonomic specific effects where individual taxa are under or over represented in a sample. For example, Gram-negative bacteria have higher extraction efficiency compared to Gram-positive bacteria, and are thus likely over represented in count tables[12, 13]. Mismatches in the primer binding sites are also responsible for taxonomic specific biases [3, 14, 15]. Additionally, taxon specific biases due to sequence template properties such as GC content, secondary structure, and gene flanking regions have been observed [16, 17, 15]. ~~Simulated count tables~~ However, due to limited community complexity the applicability of mock community assessment results to more complex environmental samples is unknown. Environmental sample complexity can be modeled using simulated and have been used to assess differential abundance ~~method (fold change differences in relative abundance)~~methods, where specific taxa are artificially ~~overrepresented~~ over represented in one set of samples compared to another [18]. ~~Using~~ However, using simulated data to assess log fold-change estimates only evaluates the 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~~ real environmental samples are often used to benchmark new molecular laboratory and computational methods. However, without ~~an expected value to compare to~~expected values for use in assessment, only measurement precision ~~,~~ or agreement with other methods ~~,~~ can be evaluated. By mixing environmental samples, expected values are calculated using information from the unmixed samples and mixture design. Mixtures of environmental samples were previously used to evaluate gene expression measurements [19, 20, 21].

~~In the present study, we developed a mixture dataset of extracted DNA from human stool samples for assessing~~ Here we present a framework for assessing computational methods used to analyze 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 by pipeline. We have made both~~ marker-gene-survey data. The framework is comprised of a 16S rRNA two-sample titration dataset, generated using mixtures of human stool sample DNA extracts, along with metrics to assess the quantitative and qualitative characteristics of count tables generated using marker-gene-survey computational methods. To demonstrate usage of this assessment framework, we evaluated three bioinformatic pipelines. Both the dataset and metrics developed in this study ~~publicly available for evaluating~~ are publicly available and can be used to evaluate and optimize new and existing bioinformatic pipelines.

Results

~~Two-Sample Titration Design~~Assessment Framework

~~Samples collected at multiple timepoints during a Enterotoxigenic *E. coli* (ETEC) vaccine trial [22] were used to generate a two-sample titration dataset for assessing the~~ Our framework assesses the qualitative and quantitative characteristics of the 16S rRNA ~~marker-gene survey measurement process . Samples from five trial participants were selected for our~~ measurement process (Fig. 1). The framework evaluates count tables generated by bioinformatic pipelines from a dataset developed specifically for use in this framework. The qualitative assessment provides insight into how much confidence a user can have in feature presence/absence. The quantitative assessment evaluates the bias and variance of relative and differential abundance estimates.

Assessment Dataset - Mixture Design

Using mixtures of environmental samples we generated a dataset with expected values for use in our assessment framework. For mixture datasets, expected values can be obtained using information from unmixed samples and the mixture design. Our mixture dataset uses a 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. [23]. Only individuals with no *E. coli* detected in samples collected from trial participants prior to ETEC exposure (PRE) were used for our two-samples titrations. Post ETEC exposure (POST) samples were identified as the timepoint~~ titration mixture design, where DNA collected from five vaccine trial participants before and after exposure to ~~ETEC with the highest~~ pathogenic ~~*E.*~~ *Escherichia coli*~~*concentration for each subject (Fig. 2A). Due to limited sample availability, for E01JH0016 the timepoint with the second highest* E. coli *concentration 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*~~ was mixed following a *log*$_2$ dilution series (Fig. 2~~B). 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 mixing.~~

). Each sample was sequenced in quadruplicate. For our two-sample titration mixture design, ~~the~~ expected feature relative abundance ~~can be~~ is 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 $q_{post,j}$. Throughout the rest of the manuscript, samples collected prior to and after *E. coli* exposure are referred to as PRE and POST respectively.

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

~~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. [23]. Counts represent normalized relative abundance values for 454-NGS and copies of the heat-labile toxin gene per μL, a marker gene for ETEC, for qPCR. 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.~~

Qualitative Assessment

The qualitative assessment shows how well pipelines differentiate true biological sequences from measurement process artifacts. Inadequate processing of artifacts results in false positive and false negative features where false positives are features in a count table that are not present in the sequenced sample and false negative features are biological sequences in a sample not represented in the count table. Our qualitative assessment methods characterize the artifactual feature proportion (the frequency of artifactual features in a count table) by estimating the proportion of *titration*- and *unmixed-specific* features (Fig. 1B) that cannot be explained by sampling alone. We combine the artifactual feature proportion assessment results with sparsity estimates to hypothesize whether the artifactual features are primarily false positives or negatives. Sparsity is defined as the fraction of 0 valued cells in the count table (Fig. 1C).

~~Dataset characteristics~~

Quantitative Assessment

To evaluate count table abundance values, our quantitative assessment uses error, bias, and variance metrics (Fig. 1C). Error metrics measure agreement between observed and expected abundance values. The bias and variance metrics summarise feature-level performance. Bias metrics summarise the overall agreement with expected values and the variance metric characterizes the distribution of the agreement. Overall, pipeline performance is evaluated by comparing count table metric distributions. Additionally, feature-level metrics are indicators of feature-specific biases.

Assessment Dataset Characterization and Validation

To assure the mixture dataset is suitable for use in our assessment framework, we first validated the titration series and raw sequence data. The mixture dataset had sufficient sample coverage, reads per sample, and read quality for use in our assessment framework. The number of reads per sample and distribution of base quality scores by position was consistent across subjects (Fig. S5). There were 8.9548×10^4 (152,267 - 3,195) sequences per sample, median and range. Average base quality score was greater than 30 over the length of the amplicon when considering both forward and reverse reads (Fig. S5B).

Additionally, we characterized subject specific differences to inform the interpretation of our assessment results. No subject specific differences in base quality score were observed (Fig. S5). However, average read depth was greater for E01JH004 compared to the other individuals (Fig. S5). Community composition differences between PRE and POST samples and individuals was characterized using alpha and beta diversity (Fig. S6). Overall alpha diversity was higher for POST except for E01JH0011, though differences in diversity between PRE and POST varied by individual. Based on the beta diversity the community composition within individuals differed between the PRE and POST samples. Note that assessment metrics defined above and results reported below are based on within subject comparisons.

To validate the two-sample titration assessment dataset, 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 (ERCC plasmids) were spiked into the unmixed samples before mixing and quantified using qPCR (Fig. S1B). The stool samples used to generate the mixtures have both eukaryotic (primarily human) DNA and prokaryotic DNA. If the proportion of prokaryotic DNA differs between the unmixed samples, then the amount of DNA from the unmixed samples in a titration targeted by 16S rRNA gene sequencing is not consistent with the mixture design. We quantified the proportion of prokaryotic DNA in the unmixed samples using a qPCR assay targeting the 16S rRNA gene (Fig. S1C).

Our assessment dataset validation results indicated that the samples were volumetrically mixed according to the mixture design (Table S1) but prokaryotic DNA proportion varied across the titration series (Fig. S2). To account for

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	38358	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
QIIME	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

deviations from the mixture design due to differences in the proportion of prokaryotic DNA in the unmixed samples, we estimated the proportion of POST in each titration using the 16S rRNA sequencing data (Fig. S3) and the estimated POST proportions were used in our assessment metric calculations. See Supplemental Material for the assessment dataset validation methods and results.

Count Table Assessment Demonstration

Next, we demonstrate the utility of our assessment framework on count tables generated using three different bioinformatic pipelines; DADA2, Mothur and QIIME. First, we provide high level summary statistics for initial insight into how the count tables differ. Next, we compare the assessment framework results for the three count tables.

Count Table Characteristics ~~Sequence dataset characteristics. (A) Distribution in the number of reads per barcoded sample (Library Size) by individual. Boxplots summarize data distribution with horizontal bar as median, boxes indicating interquartile range, whiskers ±1.5 × IQR, and black points outliers. 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.~~

~~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.~~

~~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).~~

~~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. ??). Two barcoded experimental samples had less than 35,000 reads. The rest of the samples with less than 35,000~~

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

~~The resulting~~ The count tables generated using the ~~four bioinformatic pipelines were characterized for~~ three bioinformatic pipelines vary in pre-processing and feature inference methods. These differences are reflected in the count table number of features, ~~sparsity, and filter~~ total abundance, and drop-out rate (Table 1, ~~Figs. ??B~~Fig. S7B). The pipelines evaluated employ different approaches for handling low quality reads resulting in large differences in the 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 [24, 25]. The high drop-out 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. ~~??B). Furthermore, increasing the drop-out~~ S5B). Further increasing the filter rate, QIIME excludes singletons ~~,~~ (features only observed once in the dataset~~, to remove potential sequencing artifacts from the dataset. QIIME and DADA2 pipelines 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. This is due to the redundant nature of the samples where the 35 titration samples are derived directly from the 10 unmixed samples, along with 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.~~).

~~Dataset taxonomic assignments~~ Feature taxonomic composition also varied by pipeline (Fig. ~~??).~~ S8). The three pipelines generated unique feature sets in terms of sequence length and amplicon position (see pipeline description). Therefore, we used feature taxonomic assignments for cross-pipeline community composition comparison. Phylum and order relative abundance is similar across pipelines (Fig. ~~??A~~ S8A & B). The observed differences are attributed to different taxonomic classification methods and databases used by the pipelines. ~~DADA2 and QIIME pipelines differed from Mothur and QIIME for Proteobacteria and Bacteriodetes.~~ Regardless of the relative abundance threshold, ~~for genus sets~~ most genera were unique to individual pipelines (Fig. ~~??C~~ S8C & D). Sets ~~,~~ (shared taxa between pipelines~~,~~) with QIIME had the fewest genera, excluding the DADA2-QIIME set. QIIME was the only pipeline to use open-reference clustering and the Greengenes database. Mothur and DADA2 both used the SILVA dataset. The Mothur and DADA2 pipeline use different ~~implmentations~~ implementations of the RDP naïve Bayesian classifier, which may be partially responsible for the Mothur, unclustered, and DADA2 differences.

~~Titration Series Validation~~

Qualitative Assessment

~~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. The stool samples used to generate the mixtures have both eukaryotic (primarily human) DNA and prokaryotic DNA. If the proportion of prokaryotic DNA differs between the unmixed samples, then the amount of DNA from the unmixed samples in a titration targeted by 16S rRNA gene sequencing is not consistent with the mixture design. 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.~~

~~*Spike-in qPCR results*~~

~~Titration series volumetric mixing was validated by quantifying 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 **??**). 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 **??**). Slope estimates less than one were attributed to assay standard curve efficiency less than 1 (Table **??**). 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 two-sample titration mixture design.~~

~~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 SlopeE01JH0004 012 Ac03459877-a1 77 0.9996 86.19 0.98 0.92E01JH0011 157 Ac03459958-a1 71 0.9995 87.46 0.95 0.90E01JH0016 108 Ac03460028-a1 74 0.9991 87.33 0.95 0.84E01JH0017 002 Ac03459872-a1 69 0.9968 85.80 0.89 0.93E01JH0038 035 Ac03459892-a1 65 0.9984 86.69 0.95 0.94~~

~~*Prokaryotic DNA Concentration*~~

~~Observed changes in prokaryotic DNA concentration across titrations indicate the proportion of prokaryotic DNA from the unmixed PRE and POST samples in a titration is inconsistent with the mixture design (Fig. **??**). 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. **??**). 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.~~

~~*Theta Estimates*~~

~~Human stool sample DNA extracts vary in the proportion of eukaryotic (primarily human) and prokaryotic DNA in the sample. To account for differences in the proportion of prokaryotic DNA in PRE and POST samples (Fig. **??**) we inferred the proportion of POST sample prokaryotic DNA in a titration, θ, using the 16S rRNA sequencing data (Fig. **??**). Overall the relationship between the inferred and mixture design θ values were consistent across pipelines but not subject whereas the θ estimate 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 significantly negative slope for E01JH0038 (Fig. **??**).~~

~~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.~~

~~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~~ To evaluate feature presence-absence, the framework's qualitative assessment measures artifactual feature proportion and count table sparsity. Low abundance features present only in unmixed samples or ~~titrations. These features are not expected given the titration experimental design. The quantitative assessment evaluated relative and differential abundance estimates.~~

~~*Qualitative Assessment*~~

~~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.~~

~~Unmixed- and titration-specific~~ titration samples are expected due to random sampling. *Unmixed-* and *titration-specific* features were observed for all pipelines (~~titration-specific~~*titration-specific*: Fig. 3A, ~~unmixed-specific~~*unmixed-specific*: Fig. 3B). ~~For mixture datasets low abundance features present only in unmixed samples and mixtures are expected due to random sampling~~ Overall, the DADA2 count table had the largest number of artifactual features (Table S3). A summary of the *titration-specific* artifactual features is provided in the supplementary material.

We next assessed the proportion of these artifactual features that could be explained by sampling effects alone. For our two-sample titration dataset~~there were unmixed-specific~~ , there were *unmixed-specific* features with expected counts not which could not be explained by sampling alone for all individuals and bioinformatic pipelines (Fig. 3C). However, the proportion of ~~unmixed-specific~~ *unmixed-specific* features that could not be explained by sampling alone varied by bioinformatic pipeline. DADA2 had the highest proportion of ~~unmixed-specific features not explained by sampling~~ *unmixed-specific* artifactual features whereas QIIME had the lowest proportion ~~. Consistent~~ which is 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~~ *titration-specific* feature observed counts (Fig. 3D). ~~Overall, the~~

We expected this mixture dataset to be less sparse relative to other datasets due to the redundant nature of the samples where the 35 titration samples are derived directly from the 10 unmixed samples, along with four PCR replicates for each sample. We observed overall sparsity of 0.93 and 0.94 for DADA2 ~~count table had the largest number of observed features inconsistent with the titration experiment design, while the same phenomenon is significantly reduced in the other pipelines~~ and QIIME respectively, and a higher value of 0.98 for Mothur 1).

To account for differences in microbial community composition across the five individuals we also measured sparsity at the individual level (Table S2). Sparsity at the individual-level is lower than overall sparsity for all three pipelines. In this case, average sparsity across individuals for 0.70 and 0.76 for DADA2 and Mothur, while QIIME had a lower average sparsity across individuals of 0.56. Differences in alpha and beta diversity for the five individual unmixed samples are consistent with individual level sparsity and therefore reflects differences in individual microbial community composition.

Based on the artifactual feature proportions and count table sparsity, DADA2 artifactual features are likely due to false negative features, whereas the Mothur and QIIME high sparsity values were attributed to false positive features. Based on the observed sparsity levels it is unlikely that any of the pipelines successfully filtered out a majority of the sequencing artifacts. Both unmixed- and titration-specific features that can and cannot be explained by sampling alone contribute to sparsity and the differences in the artifactual feature proportion and sparsity provide insight into how the pipelines treat sequencing artifacts.

Quantitative Assessment

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

~~For the relative abundance assessment~~

Relative Abundance Assessment To assess count table feature relative abundance values, we 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 relative abundance error rates. Relative abundance error rate is defined as $|exp - obs|/exp$, where exp and obs is the expected and observed relative abundance. To control for biases in feature inference, the three pipelines were compared to an unclustered dataset. The unclustered count table was generated using the 40,000 most abundant features from Mothur's initial preprocessing (see Methods for details). 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 ~~??). PRE and POST specific features were defined as present in all four 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 all 4 PCR replicates, therefore these features were not included in the analysis.~~ S3). Overall, agreement between inferred and observed relative abundance was high for all individuals and bioinformatic pipelines (Fig. 4A). The error rate distribution was similarly consistent across pipelines, including long tails (Fig. 4B).

To assess quantitative accuracy across pipelines, we compared the feature-level relative abundance error rate bias ~~(median error rate, Fig. 5A) and variance $(RCOV = (IQR)/|median|$ Fig. 5B) across pipelines and individuals~~ and variance using mixed effects models. To control for subject specific differences, subject was included in the model as a random effect. Large bias and variance metric values were observed for all pipelines (Table ~~??). Features with large bias and variance metrics, $1.5 \times IQR$ from the median , were deemed outliers. To prevent these outlier features from biasing the comparison they were not 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 smaller feature-level error rate~~S3). Feature-level relative abundance error rate bias (median error rate, Fig. 5A) was significantly different between pipeline, but no statistically significant differences were observed for the variance metric, $(RCOV = (IQR)/|median|$, Fig. 5B) across pipeline. The Mothur, DADA2, and QIIME feature-level ~~bias~~ biases 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) (5B). Large variance metric values were observed for all individuals and pipelines (Table ~~??~~S3). The feature-level variance was not significantly different between pipelines~~, ~~: Mothur = 0.83, QIIME = 0.71 and DADA2 = 1 (Fig. 5B). ~~We 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, no specific taxonomic groups or phylogenetic clades were identified with exceptionally poor performance in our assessment.~~

The agreement between log-fold change estimates and expected values were individual specific and consistent across pipelines (Fig. 6A). The individual specific effect was attributed to the fact that unlike relative abundance assessment, the inferred θ values were not used to calculate expected values. Inferred θ values were not used to calculate the expected values because all of the titrations and the θ estimates for the higher titrations were ~~included and they were~~ not monotonically decreasing~~and therefore~~. Using the inferred θ 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~~: for subject E01JH0011, the Mothur log fold-change estimates were more consistent with expected values than the other pipelines. However, as θ was not corrected for differences in the proportion of prokaryotic DNA between the unmixed PRE and POST samples, it cannot be said whether Mothur's performance was better than the other pipelines.

The log fold-change error distribution was consistent across pipelines (Fig. 6B). 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, we compared ~~log-fold change~~ error distributions for log-fold change estimates using different normalization methods. Error rate distributions, including the long tails, were consistent across normalization methods. ~~Furthermore,~~ Seeing as the long tail was observed for the unclustered data as well, the log-fold change 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~~computational portion. 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. 6). Similar to relative abundance, 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 consistency between log fold-change estimates and expected values 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, 6B, Variance: F = 47.39, 0.23, p = 0, 0.8, Fig. 6C). ~~We also evaluated whether poor feature-level metrics could be attributed to specific clades for taxonomic groups. Similar to the relative abundance estimate, while a phylogenetic signal was detected for both the bias and variance metrics, no specific taxonomic groups or phylogenetic clades that performed poorly were identified.~~

Discussion

Mixtures of environmental samples have been used to assess RNAseq and microarray gene expression measurements [19, 20, 21]. However, this is the first time mixtures have been used to assess microbiome measurement methods. We developed a novel assessment framework utilizing a mixture dataset for evaluating marker-gene-survey computational methods (Fig. 1).

~~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~~ Using mixtures of environmental samples, expected values for use in assessment can be obtained using information from unmixed samples and how the samples were mixed. Our assessment dataset follows a two-sample titration mixture design, where DNA collected from five vaccine trial participants before and after exposure to pathogenic *Escherichia coli* ~~from five vaccine trial participants (subjects) were~~ was mixed following a log_2 dilution series (Fig. 2). ~~Qualitative count table~~ Count table qualitative characteristics were assessed using relative abundance information for features observed only in titrations (titration-specific) and unmixed samples ~~. We quantitatively assed count tables by comparing feature relative and differential abundance~~ (unmixed-specific) (Fig. 1B). Statistical tests were used to determine if the absence of unmixed-specific features from titrations or absence of titration-specific features from unmixed samples could be explained by random sampling. Count tables were quantitatively assessed by comparing observed feature relative abundance and feature differential abundance estimates to expected values. Quantitative performance was characterized using error rate, along with feature-level bias variance metrics we developed (Fig. 1C).

Count Table Assessment Demonstration

We demonstrated our ~~novel assessment approach by evaluating~~ assessment framework on count tables generated ~~using different~~ by three commonly used bioinformatic pipelines, QIIME, Mothur, and DADA2. The ~~Mothur pipeline uses *de novo* clustering for feature inference [4, 5]. Pairwise distances used in clustering are calculated using a multiple sequence alignment. The quality filtered paired-end reads are merged into contigs. The pipeline then 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 [6, 7]. 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~~ objective of any pipeline is to differentiate true biological sequences from measurement process artifacts along with accurate abundance estimates. Our qualitative assessment results, when combined with sparsity information provides a new method for evaluating how well bioinformatic pipelines account for sequencing artifacts without loss of true biological sequences. Additionally, our quantitative assessment results identified previously unknown feature specific biases in abundance estimates.

The qualitative assessment evaluates if titration- and unmixed-specific features can be explained by random sampling alone (Fig. 1B). Titration- and unmixed-specific features not explained by sampling are artifacts of the measurement process. These artifacts can be viewed as false-positives, not representative of actual sequences in a sample, or false-negatives, actual sequences in a sample not represented by count table features. Artifacts can be PCR errors such as chimeras, reads with high sequencing error rates, or cross sample contamination [26][27][28]. Count table sparsity information (the proportion of zero-valued cells) provides additional insight into the qualitative assessment results.

A high false negative rate provides an explanation for DADA2~~pipeline uses a probability model and maximization expectation algorithm for feature inference [8]. Unlike distance-based clustering methods employed by the Mothur and QIIME pipelines ,~~ 's high proportion of artifact titration- and unmixed-specific features and count table having comparable sparsity to the other pipelines despite having significantly fewer features (Fig. S5 and Table 1). The DADA2 ~~parameters determine if low abundance~~ feature inference algorithm may be aggressively grouping lower abundance true sequences with higher 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~~ result, the low abundance sequences are not present in samples leading to increased sparsity and high abundance unmixed- and titration-specific features. This aggressive grouping of sequences is a design choice made by the algorithm developers. The DADA2 documentation states that the default setting for `OMEGA_A` is conservative to prevent false positives at the cost of increasing false negatives [8]. Using the qualitative assessment methods described here, a user can adjust the `OMEGA_A` parameter to obtain a false-negative rate appropriate for their study.

~~*Quantitative Assessment*~~ 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~~deviations in the proportion of prokaryotic DNA from PRE and POST samples in a titration from the mixture design. 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 are not likely a pipeline artifact~~

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 unlikely pipeline artifacts as they were observed in count tables generated using the unclustered pipeline as well as standard bioinformatic pipelines. ~~We~~ Additionally, we were unable to attribute outlier features to relative abundance values, log fold-change between unmixed samples, and sequence GC content. ~~Features~~ Furthermore, features with extreme metric values were not limited to any specific taxonomic group or phylogenetic clade. ~~Outlier features could not be attributed to bioinformatic pipelines and are likely due to biases in the molecular biology part of the measurement process. PCR amplification is~~ PCR amplification bias (a well-known source of bias in the molecular biology part of the measurement process) is one possible explanation for the outlier features [29]. Mismatches in the primer binding regions impact PCR efficiency and are a potential cause for poor feature-specific performance [30]. Additional research is needed before outlier features ~~are~~ can be 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 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 assessmentresults. A high rate of false negative features is a potential explanation for~~ Based on our assessment results, we suggest using DADA2 ~~count table's poor performance in the qualitative assessment and comparable sparsity~~ for feature-level abundance analysis, e.g. differential abundance testing. While DADA2 performed poorly in our qualitative assessment, the pipeline performed better in the quantitative assessment compared to the other pipelines~~despite having significantly fewer features (Fig. 3 and Table 1). The~~ . Additionally, 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~~ poor qualitative assessment results due to false-negative features are unlikely to negatively impact feature-level abundance analysis. When determining which pipeline to use for a study, users should consider whether minimizing false positives (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 [8].~~) or false negatives (Mothur) is more appropriate for their study objectives. Based on our findings we find that users of DADA2 can be more confident that an observed feature represents a member of the microbial community and not a measurement artifact, but careful examination of sequences assigned to features of interest should still be performed.

~~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. 3 and Table 1). The statistical tests used to determine if the specific features could be explained by sampling alone 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. Mothurand 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 [24].~~

~~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 developed for, potentially explaining the higher than expected sparsity [24]. Kozich et al. [24] reduced the sequence error rate from 0.29% to 0.06% by using paired-end reads that completely overlap. The larger region used in this study has a smaller overlap between the forward and reverse reads. As a result, merging the forward and reverse reads did not allow for 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 features, predominantly measurement artifacts [31]. 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 [32].~~

~~*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. Using our mixture dataset we developed 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 mock community assessments. As a result, we identified previously unknown feature specific biases. Based on our subject-specific results observation, we recommend that studies using 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~~

Using Mixtures to Assess 16S rRNA Sequencing - Lessons Learned

There are limitations using our ~~mixture dataset. These limitations included:~~ assessment dataset, these include: (1) Lack of agreement between the proportion of ~~unmixed samples~~ prokaryotic DNA from the unmixed samples in the titrations and the mixture design. ~~The~~ (2) The mixture design resulted in a limited number of features ~~used in the different analysis~~and range of expected log-fold changes. 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 resulted 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~~ to estimate mixture proportions similar to how mRNA proportions in RNA samples ~~was~~ were used in a previous mixture study [19]. We used an assay targeting the 16S rRNA gene to detect changes in the concentration of prokaryotic DNA across titrations, but were unable to quantify the proportion of prokaryotic DNA in the unmixed samples using qPCR data. Using the 16S ~~sequencing data~~ rRNA sequencing data, we inferred the proportion of prokaryotic 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~~ source of error.

A better method for quantifying sample prokaryotic DNA proportion or using samples with consistent proportions would increase confidence in the expected value and, in-turn, error metric accuracy. Limitations in the prokaryotic DNA qPCR assay's concentration precision limits the assay’~~ssuitability~~s 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 prokaryotic 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 PRE and POST 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. Alternatively, instead of mixing PRE and POST samples from the same individual, mixing PRE and POST samples from different individuals would have resulted in additional features for use in our quantitative assessment. While unmixed sample similarity impacts the number of features that can be used in the quantitative assessment, the qualitative assessment is not impacted by unmixed sample similarity. Finally, a symmetric mixture design, for example one with unmixed PRE and POST ratios of 1:4, 1:2, 1:1, 2:1, and 4:1, would provide a larger dynamic range of abundance values for assessing both PRE and POST specific features.

Conclusions

~~Our two-sample-titration dataset and assessment methods~~ Our assessment framework can be used to evaluate and characterize ~~bioinformatic pipelines and clustering methods. The sequence dataset presented in this study can be processed with~~ 16S rRNA marker-gene survey analysis methods, in particular count tables produced by any 16S rRNA 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 presented here. The three pipelines we evaluated produced sets of features varying in total feature abundance, number of featuresper samples, and total features. The objective of any pipeline is to differentiate true biological sequences from measurement process artifacts. In general, based on our evaluation results we suggest using~~ We demonstrated our assessment framework with three commonly used bioinformatic pipelines. Our qualitative assessment results indicated that the QIIME and Mothur pipelines produced count table with more false-positive features whereas the DADA2 ~~for~~ count table had more false-negative features. Overall the three pipelines performed well in our quantitative assessment. However, feature-level ~~abundance analysis, e.g. differential abundance testing. While DADA2 performed poorly in our qualitative assessment, the pipeline performed better in the quantitative assessmentcompared to the other pipelines. Additionally,~~ analysis identified poorly performing features and the ~~DADA2 poor qualitative assessment results due to false-negative features are unlikely to negatively impact~~ sources of bias responsible for this poor feature-level ~~abundance analysis . 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~~ quantitative performance are unknown. Therefore, 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.

Methods

~~*Titration Validation*~~

~~qPCR was used to validate volumetric mixing and check for differences in the proportion of prokaryotic DNA across titrations. To ensure the~~

Assessment Framework

To assess the qualitative and quantitative performance of marker-gene survey analysis methods we developed a framework utilizing our two-sample ~~titrations were volumetrically mixed according to the mixture design, independent ERCC plasmids were spiked into the unmixed PRE and POST samples [33] (NIST SRM SRM 2374) (Table ??) . The ERCC plasmids were resuspended in 100 μL tris-EDTA buffer and~~

~~2 μL of resuspended plasmids 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 and TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA).~~ titration dataset(Fig. 1). Qualitative assessment evaluates feature presence-absence. The quantitative assessment evaluates the relative and differential abundance estimates.

~~To check for differences in the proportion of bacterial DNA in the~~

Assessment Dataset - Mixture Design

To provide a dataset with real-world complexity and expected values for qualitative and quantiative assessment we used mixtures of environmental samples. Samples collected at multiple timepoints during a Enterotoxigenic *E. coli* (ETEC) vaccine trial [22] were used to generate a two-sample titration dataset (Fig. 2). 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. [23]. Only individuals with no *E. coli* detected in samples collected from trial participants prior to ETEC exposure (PRE) were 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. 2A). Due to limited sample availability, for E01JH0016 the timepoint with the second highest *E. coli* concentration was used as the POST sample. Independent titration series were generated for each subject. 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. 2B). Unmixed (PRE and POST~~samples, titration bacterial~~) sample 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^{TM} Design and Analysis Software v1.4.1. Statistical analysis was performed on the exported data using custom scripts in R [34]. The qPCR data and scripts used to analyze the data are available at .~~

~~*Sequencing*~~

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 mixing. The resulting titration series was composed of 45 samples~~(~~, seven titrations and two unmixed samples for each of ~~five subjects)~~ the five subjects.

The 45 samples 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 PCR of the 16S rRNA gene~~PCR~~, followed by a sample indexing PCR, sample concentration normalization, and sequencing.

A total of 192 16S rRNA PCR assays were ~~run including four~~ sequenced across two 96-well plates including four PCR 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 [14]. 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 [35] (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~~ Kapa HiFi HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA) was used to PCR the 16S rRNA gene. The PCR product amplicon size was verified using agarose gel electrophoresis~~to check amplicon size~~. Concentration measurements were made after the initial 16S rRNA PCR, the indexing PCR, and normalization steps. DNA concentration was measured using the QuantIT Picogreen dsDNA Kit (Cat # P7589, ThermoFisher Scientific) and fluorescent measurements were made with a Synergy2 Multi-Detection MicroPlate Reader (BioTek Instruments, Inc, Winooski, VT).

Initial PCR products were purified using 0.8X 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) and then purified using 1.12X AMPure XP beads. 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. The sequence data was deposited in the NCBI SRA archive under Bioproject PRJNA480312. Individual SRA run accession numbers and metadata in Supplemental Table. 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 [36, 37]. ~~The sequence data was deposited in the NCBI SRA archive under Bioproject PRJNA480312. Individual SRA run accession numbers and metadata in Supplemental Table.~~

~~*Sequence Processing*~~

Sequence data were processed using four bioinformatic pipelines: a *de-novo* clustering method - Mothur [5], an open-reference clustering method - QIIME [7], and a sequence inference method - DADA2 [8], 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 developer's MiSeq SOP [5, 24]. The pipeline was run using Mothur version 1.37 (http://www.mothur.org/). 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 modified to account for 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 includes an initial preprocessing step where the forward and reverse reads are trimmed and filtered using base quality scores and were merged into single contigs for each read pair. 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/) [38], 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 [26]. OTU clustering was performed using the OptiClust algorithm with a clustering threshold of 0.97 [4]. The RDP classifier implemented in Mothur was used for taxonomic classification against the Mothur provided version of the RDP v9 training set [39].

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 [7]. Briefly, the QIIME pipeline uses fastq-join (version 1.3.1) to merge paired-end reads [40] and the Usearch algorithm [41] with Greengenes database version 13.8 with a 97% similarity threshold [42] was used for open-reference clustering.

DADA2, an R native pipeline was also used to process the sequencing data [8]. The pipeline includes a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP naïve Bayesian classifier [39] and the SILVA database V123 provided by the DADA2 developers [38, 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.

~~*Titration Proportion Estimates*~~

~~The following linear model was used to infer the proportion of prokaryotic DNA, θ, in each titration. 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. Feature relative abundance estimates were calculated using a negative binomial model.~~

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

~~To fit the model and prevent uninformative and low abundance features from biasing θ estimates, only 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 [18]. Generalized linear models provide an alternative to standard least-squares regression. The above model is additive and therefore θ_i cannot be directly inferred 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.~~

Qualitative Assessment

~~Our qualitative measurement~~

Artifactual Feature Proportion Our qualitative assessment evaluated features only observed in unmixed samples (PRE or POST) ~~, *unmixed-specific*, or titrations, *titration-specific*. *Unmixed-*~~ or only in titrations. The former we will refer to as unmixed-specific features and the latter we will refer to as titration-specific features (Fig. 1B). *Unmixed-* and *titration-specific* features ~~are~~ can arise from errors in the PCR/sequencing, feature inference processes, or 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.~~. To provide context for the artifactual feature proportion results count table sparsity was used (Fig. 1C). Sparsity is defined as the proportion of 0 valued cells in a matrix.

Hypothesis tests were used to determine if ~~differences in sampling depth~~ random sampling alone, here sequencing depth, could account for ~~*unmixed-specific*~~*unmixed-* and *titration-specific* features. p-values were adjusted for multiple comparisons using the Benjamini & Hochberg method [43]. For *unmixed-specific* features, ~~the~~ a 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~~ The binomial test was infeasible for *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~~. Because the count table abundance values for these features was 0 in the unmixed samples, their estimated probability of occurrence π_{min} is equal to ~~0. As non-zero counts were observed the true feature proportion is non-zero, and the test always~~ 0, and thus, the binomial test fails. Therefore, we formulated a Bayesian hypothesis test for *titration-specific* features ~~.~~

~~A Bayesian hypothesis test~~ detailed by equation (2). This Bayesian approach was used to evaluate if the true feature proportion is less than the minimum detected proportion. ~~The Bayesian hypothesis test was formulated using equation . Which~~ Note that when assuming equal priors, ~~$P(\pi < \pi_{min}) = P(\pi \geq \pi_{min})$,~~ $P(\pi < \pi_{min}) = P(\pi > \pi_{min})$, (2) reduces to (3). ~~For equations and π is~~ We define π as the true feature proportion, ~~π_{min} is~~ π_{min} the minimum detected proportion, ~~C is~~ C the expected feature counts, and ~~C_{obs} is~~ C_{obs} 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}~~ Count values for C were simulated using a beta prior (with varying alpha and beta values) for $\pi > \pi_{min}$ and a uniform distribution for $\pi < \pi_{min}$. Higher values of alpha and beta will skew the prior right and left respectively. Our Bayesian hypothesis tests (Eg. (3)) results were largely unaffected by beta distribution parameterization (Fig. S4). π_{min} was calculated using the mixture equation (1) where ~~$q_{pre,j}$ and $q_{post,j}$ are $min(\mathbf{Q}_{pre})$ and $min(\mathbf{Q}_{post})$~~ $q_{pre,j}$ 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~~ π is less than ~~π_{min}~~ π_{min} for features not observed in unmixed samples~~due to random sampling.~~. Artifacts not explained by sequencing alone are likely errors in the sequence measurement and inference processes, and thus, false positives or negatives.

$$\begin{aligned} p &= P(\pi < \pi_{min} | C \geq C_{obs}) \\ &= \frac{P(C \geq C_{obs} | \pi < \pi_{min}) P(\pi < \pi_{min})}{P(C \geq C_{obs} | \pi < \pi_{exp}) P(\pi < \pi_{min}) + P(C \geq C_{obs} | \pi \geq \pi_{min}) P(\pi \geq \pi_{min})} \end{aligned} \tag{2}$$

$$p = \frac{P(C \geq C_{obs} | \pi < \pi_{min})}{P(C \geq C_{obs})} \tag{3}$$

Quantitative Assessment

For quantitative assessment, we 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 [44, 45]. Expected log fold-change for feature j between titrations l and m is calculated using equation (4), 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 (5). Features only observed in POST samples, *POST-specific*, expected log fold-change values can be calculated in a similar manner. However, *POST-specific* features were rarely observed in more than one titration and therefore were not suitable for use in our assessment. 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.

$$logFC_{lm,j} = \log_2\left(\frac{\theta_l q_{post,j} + (1-\theta_l)q_{pre,i}}{\theta_m q_{post,j} + (1-\theta_m)q_{pre,j}}\right) \tag{4}$$

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

Count Table Assessment Demonstration

Demonstrate framework by comparing the qualitative and quantitative assessment results across the three pipelines. We first characterized overall differences in the count tables produced by the three pipelines. This characterization included calculating the number of features, total abundance by sample, dropout-rate, and taxonomic composition.

Qualitative Assessment

For the qualitative assessment we compare the proportion of artifactual features. The artifactual feature proportion was defined as the proportion of *unmixed-* and *titration-specific* features with abundance values that could not be explained by sampling alone. These are PCR replicates with p-values less than 0.05 after multiple hypothesis test correction for the binomial and bayesian hypothesis tests described in the assessment framework methods section. We additionally used the count table sparsity values to draw conclusions regarding the mechanism responsible for different artifactual feature proportions.

Quantitative Assessment

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 excluded from this analysis to minimize biases due to poor model fit. Features with large bias and variance metrics, $1.5 \times IQR$ from the median, were deemed outliers. These outlier features were characterized independently in a separate analysis.

We fit the following mixed effect model to test for differences in measurement bias across pipelines

$$e_{ijk} = b + b_i + z_j + \epsilon_{ijk}$$

where e_{ijk} is the observed error across features and tritations k for pipeline i on individual j. b_i is a fixed term modeling the pipeline effect, z_j is a random effect (normally distributed with mean 0) capturing overall bias differences across individuals. We fit a similar model for differences in error variance across pipelines.

We used estimated terms $\hat{b}_i$ from the mixed effects model to test for pair-wise differences across pipelines. These multiple comparisons were performed with Tukey's HSD test. A one-sided alternative hypothesis was used to determine which pipelines had smaller feature-level error rate.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Sequence data was deposited in the NCBI SRA archive under Bioproject PRJNA480312. Individual SRA run accession numbers and metadata in Supplemental Table. The code used to run the bioinformatic pipelines is available at `https://github.com/nate-d-olson/mgtst_pipelines`. Scripts used to analyze the data are available at `https://github.com/nate-d-olson/mgtst_pub`.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially supported by National Institutes of Health (NIH) [NIH R01HG005220 to H.C.B.]

Authors' contributions

NDO, HCB, OCS, MS, and WT designed the experiment, SL and SH performed the laboratory work. NDO, HCB~~and MS~~, MS, and DJB analyzed the data. NDO, DJB, and HCB wrote the manuscript. All authors provided feedback on manuscript drafts and approved the final manuscript.

Acknowledgements

The authors would like to thank the two anonymous reviewers, Mihai Pop, Scott Pine, Scott Jackson, Justin Zook, Nathan Swenson, and Prachi Kulkarni for feedback on manuscript drafts. Joseph Paulson and Justin Wagner provided helpful insight during the development of the project. Opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendations or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

Author details

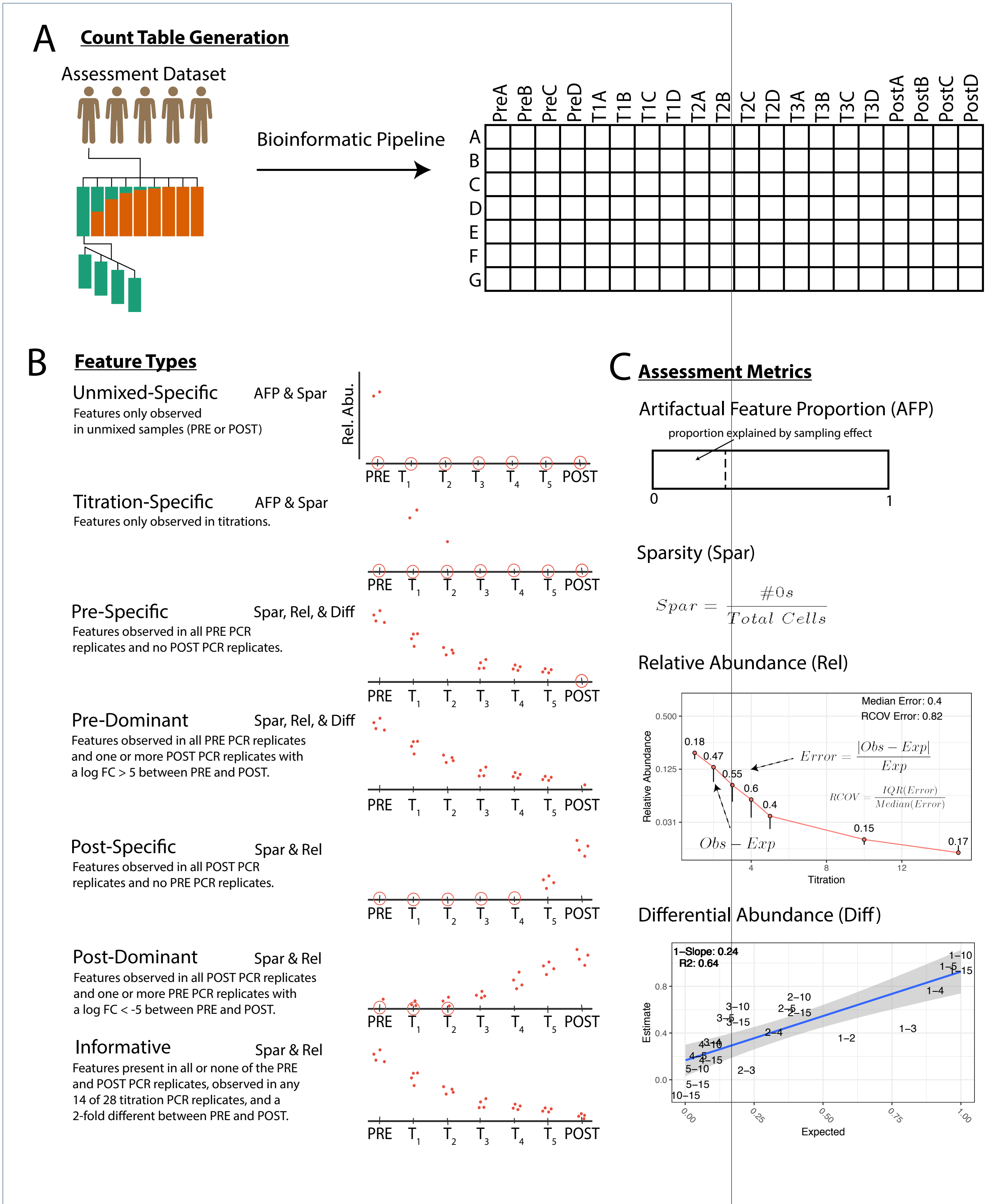
[1]Biosystems and Biomaterials Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, Maryland, 20899 USA. [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, 8314 Paint Branch Dr. College Park, Maryland, 20742 USA. [3]University of Maryland Institute of Advanced Computer Studies, University of Maryland, College Park, 8223 Paint Branch Dr. College Park, Maryland, 20742 USA. [4]Department of Epidemiology and Public Health, University of Maryland School of Medicine, 655 W. Baltimore Street, Baltimore, Maryland, 21201 USA. [5]Department of Biomedical Engineering, Johns Hopkins University, 720 Rutland Ave., Baltimore, Maryland, 21205 USA. [6]Joint Initiative for Metrology in Biology, 443 Via Ortega, Stanford, CA, 94305 USA. [7]Department of Computer Science, University of Maryland, College Park, 8223 Paint Branch Dr. College Park, Maryland, 20742 USA.

References

1. Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R., Ley, R.E.: Conducting a microbiome study. Cell **158**(2), 250–262 (2014)
2. D'Amore, R., Ijaz, U.Z., Schirmer, M., Kenny, J.G., Gregory, R., Darby, A.C., Quince, C., Hall, N.: A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics **17**, 1–40 (2016). doi:10.1186/s12864-015-2194-9
3. Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G., Reris, R.A., Sheth, N.U., Huang, B., Girerd, P., *et al.*: The truth about metagenomics: quantifying and counteracting bias in 16s rrna studies. BMC microbiology **15**(1), 66 (2015)
4. Westcott, S.L., Schloss, P.D.: Opticlust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere **2**(2) (2017)
5. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., *et al.*: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology **75**(23), 7537–7541 (2009)
6. Rideout, J.R., He, Y., Navas-Molina, J.A., Walters, W.A., Ursell, L.K., Gibbons, S.M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J.C., Gilbert, J.A., Huse, S.M., Zhou, H.-W., Knight, R., Caporaso, J.G.: Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ **2**, 545 (2014)
7. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: Qiime allows analysis of high-throughput community sequencing data. Nature Methods **7**, 335 (2010). Correspondence
8. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature Methods **13**, 581–583 (2016). doi:10.1038/nmeth.3869
9. Bokulich, N.A., Rideout, J.R., Mercurio, W.G., Shiffer, A., Wolfe, B., Maurice, C.F., Dutton, R.J., Turnbaugh, P.J., Knight, R., Caporaso, J.G.: mockrobiota: a public resource for microbiome bioinformatics benchmarking. mSystems **1**(5), 00062–16 (2016)
10. Kopylova, E., Navas-molina, J.A., Mercier, C., Xu, Z.: Open-Source Sequence Clustering Methods Improve the State Of the Art. mSystems **1**(1), 1–16 (2014). doi:10.1128/mSystems.00003-15.Editor
11. Huse, S.M., Welch, D.M., Morrison, H.G., Sogin, M.L.: Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental microbiology **12**(7), 1889–98 (2010). doi:10.1111/j.1462-2920.2010.02193.x
12. Costea, P.I., Zeller, G., Sunagawa, S., Pelletier, E., Alberti, A., Levenez, F., Tramontano, M., Driessen, M., Hercog, R., Jung, F.-E., Kultima, J.R., Hayward, M.R., Coelho, L.P., Allen-Vercoe, E., Bertrand, L., Blaut, M., Brown, J.R.M., Carton, T., Cools-Portier, S., Daigneault, M., Derrien, M., Druesne, A., de Vos, W.M., Finlay, B.B., Flint, H.J., Guarner, F., Hattori, M., Heilig, H., Luna, R.A., van Hylckama Vlieg, J., Junick, J., Klymiuk, I., Langella, P., Le Chatelier, E., Mai, V., Manichanh, C., Martin, J.C., Mery, C., Morita, H., O'Toole, P.W., Orvain, C., Patil, K.R., Penders, J., Persson, S., Pons, N., Popova, M., Salonen, A., Saulnier, D., Scott, K.P., Singh, B., Slezak, K., Veiga, P., Versalovic, J., Zhao, L., Zoetendal, E.G., Ehrlich, S.D., Dore, J., Bork, P.: Towards standards for human fecal sample processing in metagenomic studies. Nat. Biotechnol. **35**, 1069 (2017)
13. Olson, N.D., Morrow, J.B.: DNA extract characterization process for microbial detection methods development and validation. BMC Res. Notes **5**, 668 (2012)
14. Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glöckner, F.O.: Evaluation of general 16s ribosomal rna gene pcr primers for classical and next-generation sequencing-based diversity studies. Nucleic acids research, 808 (2012)
15. Gohl, D.M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T.J., Clayton, J.B., Johnson, T.J., Hunter, R., Knights, D., Beckman, K.B.: Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat. Biotechnol. (2016)
16. Pinto, A.J., Raskin, L.: PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS One **7**(8), 43093 (2012)
17. Hansen, M.C., Tolker-Nielsen, T., Givskov, M., Molin, S.: Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. FEMS Microbiol. Ecol. **26**(2), 141–149 (1998)
18. McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput. Biol. **10**(4), 1003531 (2014)
19. Parsons, J., Munro, S., Pine, P.S., McDaniel, J., Mehaffey, M., Salit, M.: Using mixtures of biological samples as process controls for rna-sequencing experiments. BMC genomics **16**(1), 708 (2015)
20. Pine, P.S., Rosenzweig, B.A., Thompson, K.L.: An adaptable method using human mixed tissue ratiometric

controls for benchmarking performance on gene expression microarrays in clinical laboratories. BMC biotechnology **11**(1), 38 (2011)

21. Thompson, K.L., Rosenzweig, B.A., Pine, P.S., Retief, J., Turpaz, Y., Afshari, C.A., Hamadeh, H.K., Damore, M.A., Boedigheimer, M., Blomme, E., *et al.*: Use of a mixed tissue rna design for performance assessments on multiple microarray formats. Nucleic acids research **33**(22), 187–187 (2005)
22. Harro, C., Chakraborty, S., Feller, A., DeNearing, B., Cage, A., Ram, M., Lundgren, A., Svennerholm, A.-M., Bourgeois, A.L., Walker, R.I., *et al.*: Refinement of a human challenge model for evaluation of enterotoxigenic escherichia coli vaccines. Clinical and Vaccine Immunology **18**(10), 1719–1727 (2011)
23. Pop, M., Paulson, J.N., Chakraborty, S., Astrovskaya, I., Lindsay, B.R., Li, S., Bravo, H.C., Harro, C., Parkhill, J., Walker, A.W., *et al.*: Individual-specific changes in the human gut microbiota after challenge with enterotoxigenic escherichia coli and subsequent ciprofloxacin treatment. BMC genomics **17**(1), 1 (2016)
24. Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D.: 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), 5112–5120 (2013)
25. Walters, W., Hyde, E.R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J.A., Jansson, J.K., Caporaso, J.G., Fuhrman, J.A., Apprill, A., Knight, R.: Improved bacterial 16S rRNA gene (v4 and v4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. mSystems **1**(1) (2016)
26. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R.: Uchime improves sensitivity and speed of chimera detection. Bioinformatics **27**(16), 2194–2200 (2011)
27. Edgar, R.C.: UNCROSS2: identification of cross-talk in 16S rRNA OTU tables (2018)
28. D'Amore, R., Ijaz, U.Z., Schirmer, M., Kenny, J.G., Gregory, R., Darby, A.C., Shakya, M., Podar, M., Quince, C., Hall, N.: A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics **17**, 55 (2016)
29. Sze, M.A., Schloss, P.D.: The impact of dna polymerase and number of rounds of amplification in pcr on 16s rrna gene sequence data. bioRxiv (2019). doi:10.1101/565598. https://www.biorxiv.org/content/early/2019/03/04/565598.full.pdf
30. Wright, E.S., Yilmaz, L.S., Ram, S., Gasser, J.M., Harrington, G.W., Noguera, D.R.: Exploiting extension bias in polymerase chain reaction to improve primer specificity in ensembles of nearly identical dna templates. Environmental microbiology **16**(5), 1354–1365 (2014)
31. Callahan, B., Sankaran, K., Fukuyama, J., McMurdie, P., Holmes, S.: Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 2; referees: 3 approved]. F1000Research **5**(1492) (2016). doi:10.12688/f1000research.8986.2
32. Risely, A., Waite, D., Ujvari, B., Klaassen, M., Hoye, B.: Gut microbiota of a long-distance migrant demonstrates resistance against environmental microbe incursions. Molecular ecology (2017)
33. Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., *et al.*: The external rna controls consortium: a progress report. Nature methods **2**(10), 731–734 (2005)
34. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/
35. Yang, B., Wang, Y., Qian, P.-Y.: Sensitivity and correlation of hypervariable regions in 16s rrna genes in phylogenetic analysis. BMC bioinformatics **17**(1), 1 (2016)
36. Souza, W., Carvalho, B.: Rqc: Quality Control Tool for High-Throughput Sequencing Data. (2017). R package version 1.10.2. `https://github.com/labbcb/Rqc`
37. Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., Bravo, C., H., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, D., K., Irizarry, A., R., Lawrence, M., Love, I., M., MacDonald, J., Obenchain, V., Ole's, K., A., Pag'es, H., Reyes, A., Shannon, P., Smyth, K., G., Tenenbaum, D., Waldron, L., Morgan, M.: Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods **12**(2), 115–121 (2015)
38. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic acids research **41**(D1), 590–596 (2012)
39. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and environmental microbiology **73**(16), 5261–5267 (2007)
40. Aronesty, E.: ea-utils: Command-line tools for processing biological sequencing data. Expression Analysis, Durham, NC (2011)
41. Edgar, R.C.: Search and clustering orders of magnitude faster than blast. Bioinformatics **26**(19), 2460–2461 (2010)
42. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L.: Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. Applied and environmental microbiology **72**(7), 5069–5072 (2006)
43. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289–300 (1995)
44. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
45. McCarthy, D.J., Chen, Y., Smyth, G.K.: Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. **40**(10), 4288–4297 (2012)



A **Count Table Generation**

Assessment Dataset

Bioinformatic Pipeline

	PreA	PreB	PreC	PreD	T1A	T1B	T1C	T1D	T2A	T2B	T2C	T2D	T3A	T3B	T3C	T3D	PostA	PostB	PostC	PostD
A																				
B																				
C																				
D																				
E																				
F																				
G																				

B **Feature Types**

Unmixed-Specific — AFP & Spar
Features only observed in unmixed samples (PRE or POST)

Titration-Specific — AFP & Spar
Features only observed in titrations.

Pre-Specific — Spar, Rel, & Diff
Features observed in all PRE PCR replicates and no POST PCR replicates.

Pre-Dominant — Spar, Rel, & Diff
Features observed in all PRE PCR replicates and one or more POST PCR replicates with a log FC > 5 between PRE and POST.

Post-Specific — Spar & Rel
Features observed in all POST PCR replicates and no PRE PCR replicates.

Post-Dominant — Spar & Rel
Features observed in all POST PCR replicates and one or more PRE PCR replicates with a log FC < -5 between PRE and POST.

Informative — Spar & Rel
Features present in all or none of the PRE and POST PCR replicates, observed in any 14 of 28 titration PCR replicates, and a 2-fold different between PRE and POST.

C **Assessment Metrics**

Artifactual Feature Proportion (AFP)

Sparsity (Spar)

$$Spar = \frac{\#0s}{Total\ Cells}$$

Relative Abundance (Rel)

$$Error = \frac{|Obs - Exp|}{Exp}$$

$$RCOV = \frac{IQR(Error)}{Median(Error)}$$

$$Obs - Exp$$

Differential Abundance (Diff)

Figure 1 Assessment Framework. A) Count tables evaluated by the assessment framework are generated from the assessment dataset using marker-gene survey bioinformatic pipelines. Count table rows are features identified by the bioinformatic pipeline and column are samples, four PCR replicates (labeled A-D) were sampled for PRE and POST and titrations, to simplify the diagram only three titrations are shown. B) Pictorial depiction of abundance values of the seven feature types observed and used in the assessment framework. C) Qualitative and quantitative assessment metrics used in the assessment framework. The artifactual feature proportion metric (AFP) is a qualitative assessment of feature presence/absence based on unmixed-specific or titration-specific artifactual features. Sparsity (SPAR) is a qualitative assessment of the proportion of observed

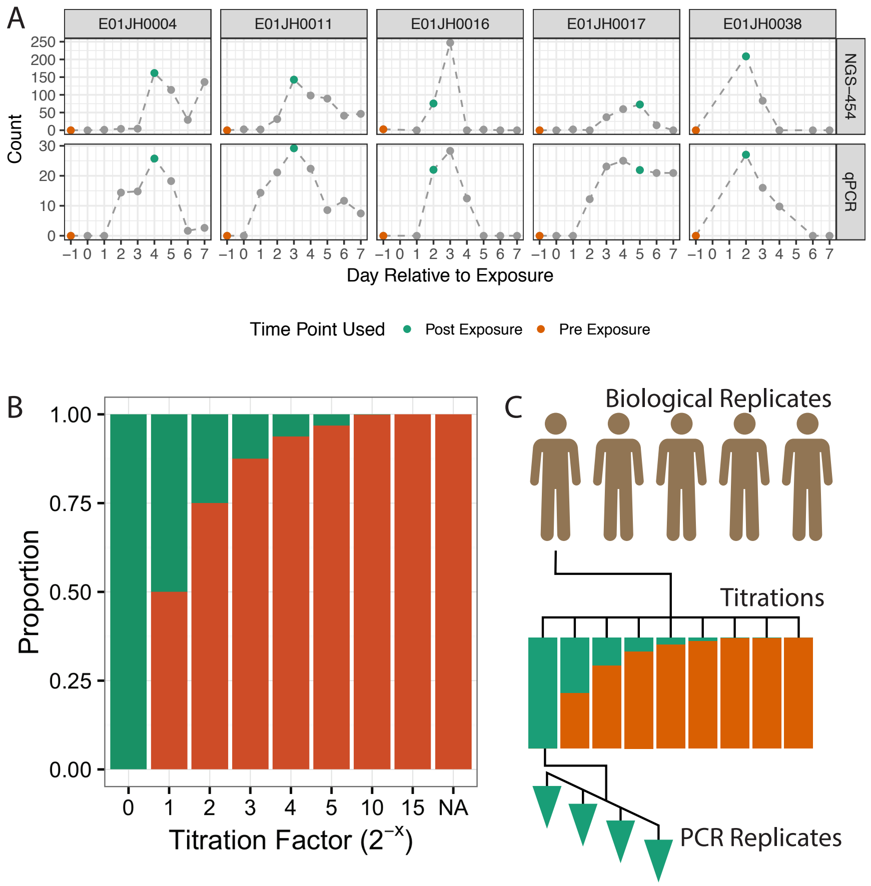


Figure 2 Sample selection and experimental design for the two-sample titration 16S rRNA marker-gene-survey assessment dataset. A) Pre- and post-exposure (PRE and POST) samples from five vaccine trial participants were selected based on *Escherichia coli* abundance measured using qPCR and 454 16S rRNA sequencing (454-NGS), data from Pop et al. [23]. Counts represent normalized relative abundance values for 454-NGS and copies of the heat-labile toxin gene per μL, a marker gene for ETEC, for qPCR. 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.

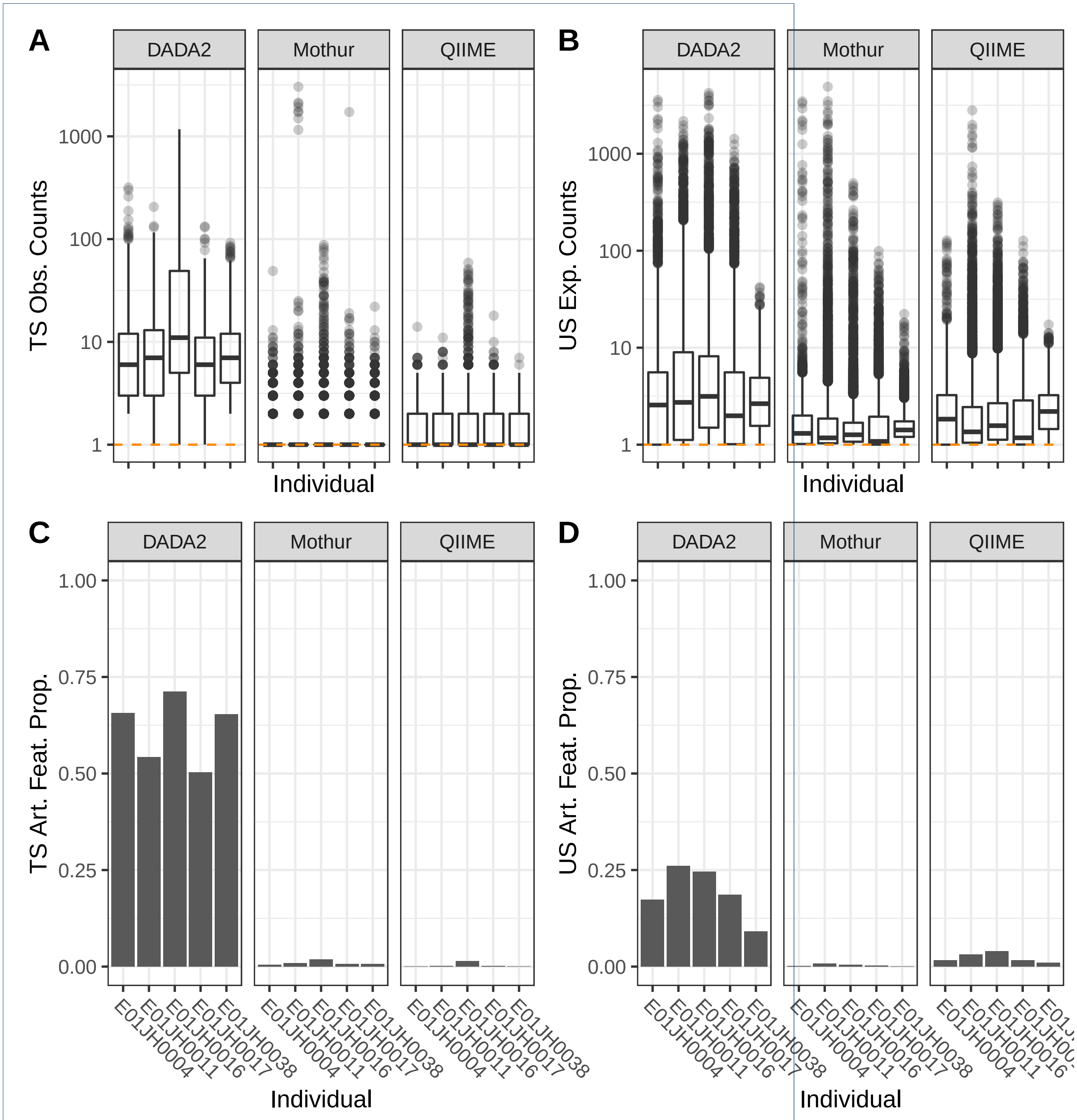


Figure 3 ~~Prokaryotic DNA concentration~~ Distribution of (~~ng/ul~~A) ~~across titrations measured using a 16S rRNA qPCR assay. Separate linear models, Prokaryotic DNA concentration versus θ were fit~~ observed count values for ~~each individual,~~ titration-specific features and ~~R^2~~ (B) expected count values for unmixed-specific features by pipeline and ~~p-values were reported~~individual. ~~Red lines indicate negative slope estimates and blue lines positive slope estimates. p-value~~ The orange horizontal dashed line indicates ~~significant difference from the expected slope~~ a count value of ~~0. The grey regions indicate the linear model 95% confidence interval. Multiple test correction was performed using the Benjamini-Hochberg method. One of the E01JH0004 PCR replicates for titration 3~~ 1. (~~$\theta = 0.125$~~C) ~~was identified as an outlier, with a concentration~~ Proportion of ~~0.003,~~ titration-specific features and ~~was excluded from~~ (D) unmixed-specific features with an adjusted p-value $<$ 0.05 for the ~~linear model~~Bayesian hypothesis test and binomial test respectively. ~~The linear model slope was still significantly different from 0~~ We failed to accept the null hypothesis when the ~~outlier was included~~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.

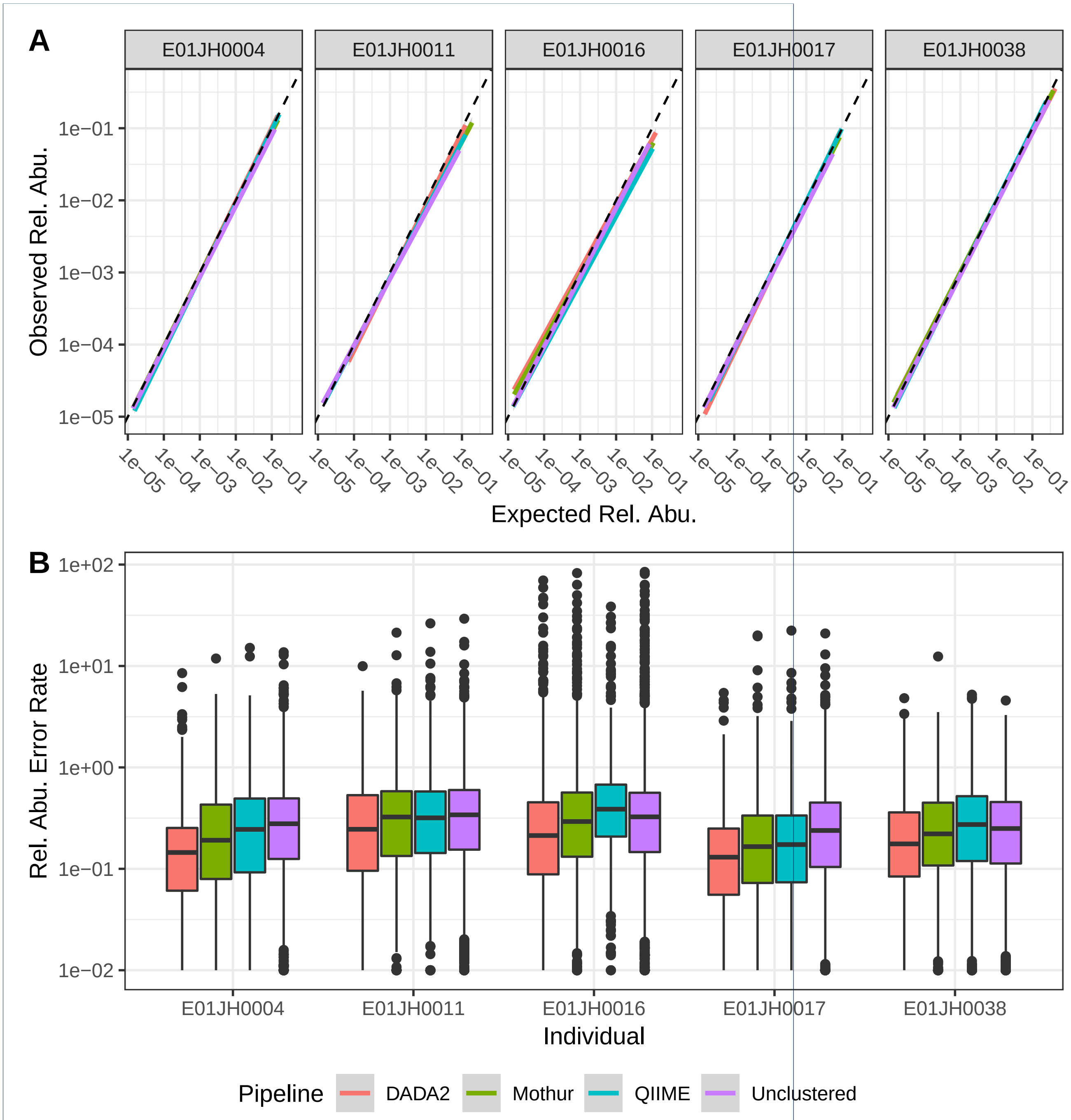


Figure 4 Relative abundance assessment. (A) A linear model of the relationship between the expected and observed relative abundance. The dashed grey line indicates expected 1-to-1 relationship. The plot is split by individual and ~~color is used to indicate the different~~ bioinformatic ~~pipelines~~pipeline indicated by line color. A negative binomial model was used to calculate an average relative abundance estimate across ~~the four~~ PCR replicates. ~~Points~~ To highlight quantitative performance for higher abundance features, points with observed and expected relative abundance values less than 1/median~~library size~~ (total abundance) were excluded from the ~~data used to fit the linear model~~plot. (B) Relative abundance error rate ($|expected - observed|/expected$) distribution by individual and pipeline.

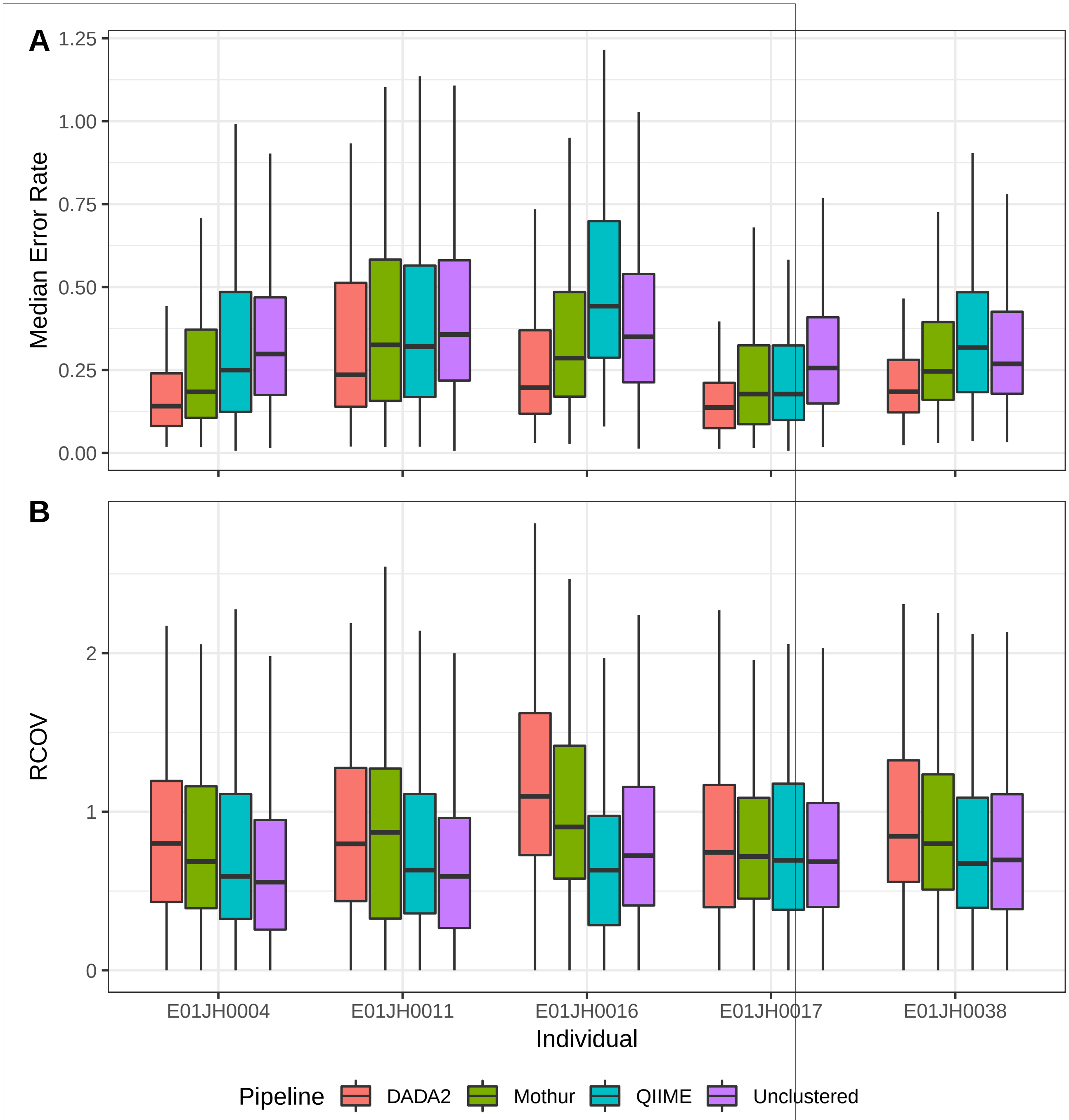


Figure 5 Comparison of pipeline relative abundance assessment feature-level error metrics. Distribution of feature-level relative abundance (A) bias metric - median error rate and (B) variance - robust coefficient of variation (~~$RCOV = (IQR)/|median|$~~$RCOV = IQR/|medianerrorrate|$) by individual and pipeline. For both the bias and variance metrics lower values are better. 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.

Differential Abundance Assessment

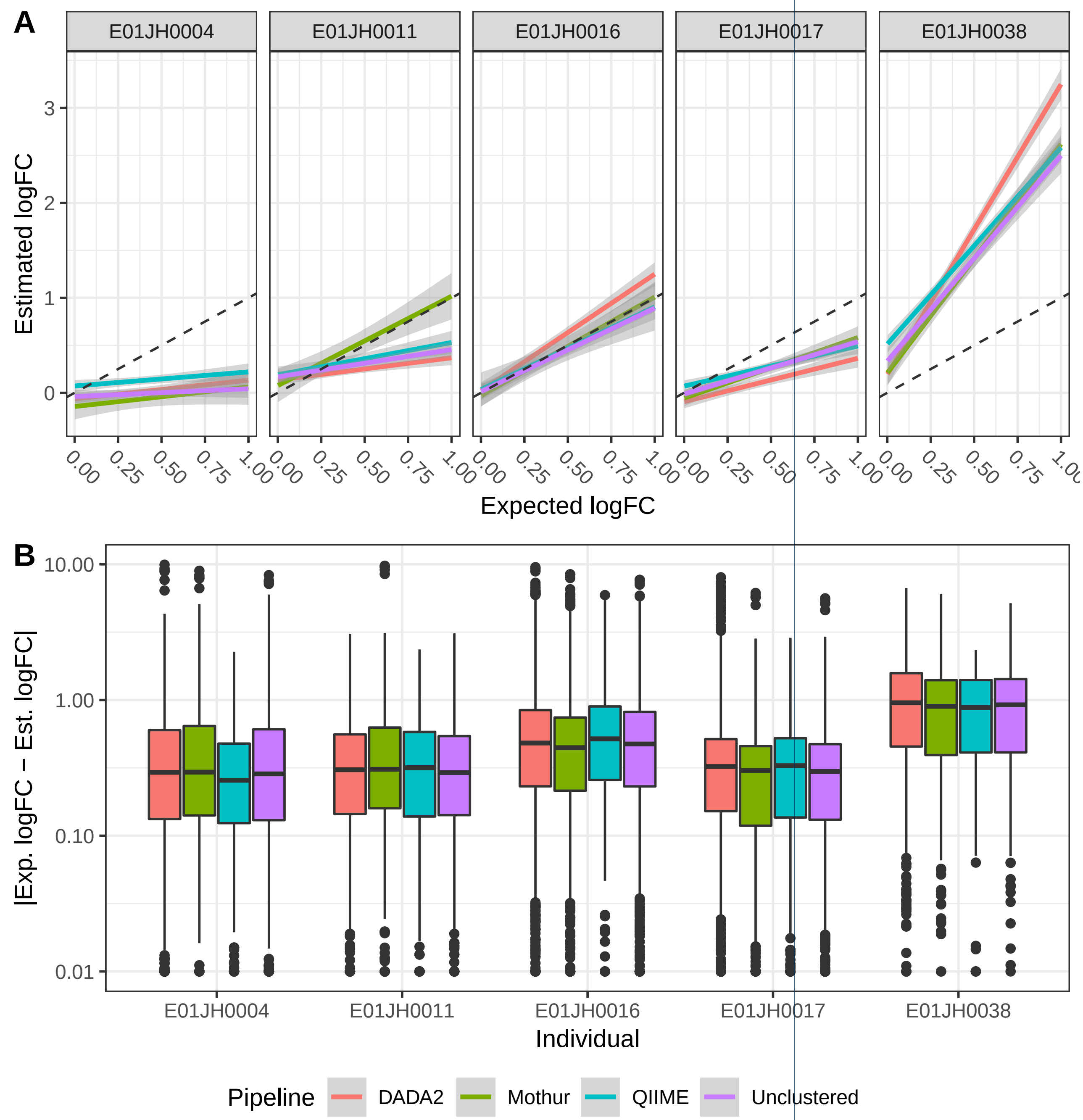


Figure 6 Differential abundance quantitative assessment. (A) Linear model ~~or~~ of the relationship between ~~log fold-change estimates~~ estimated and expected ~~values~~ log fold-change relative abundance between titrations for PRE-specific and PRE-dominant features by pipeline and individual, line 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.

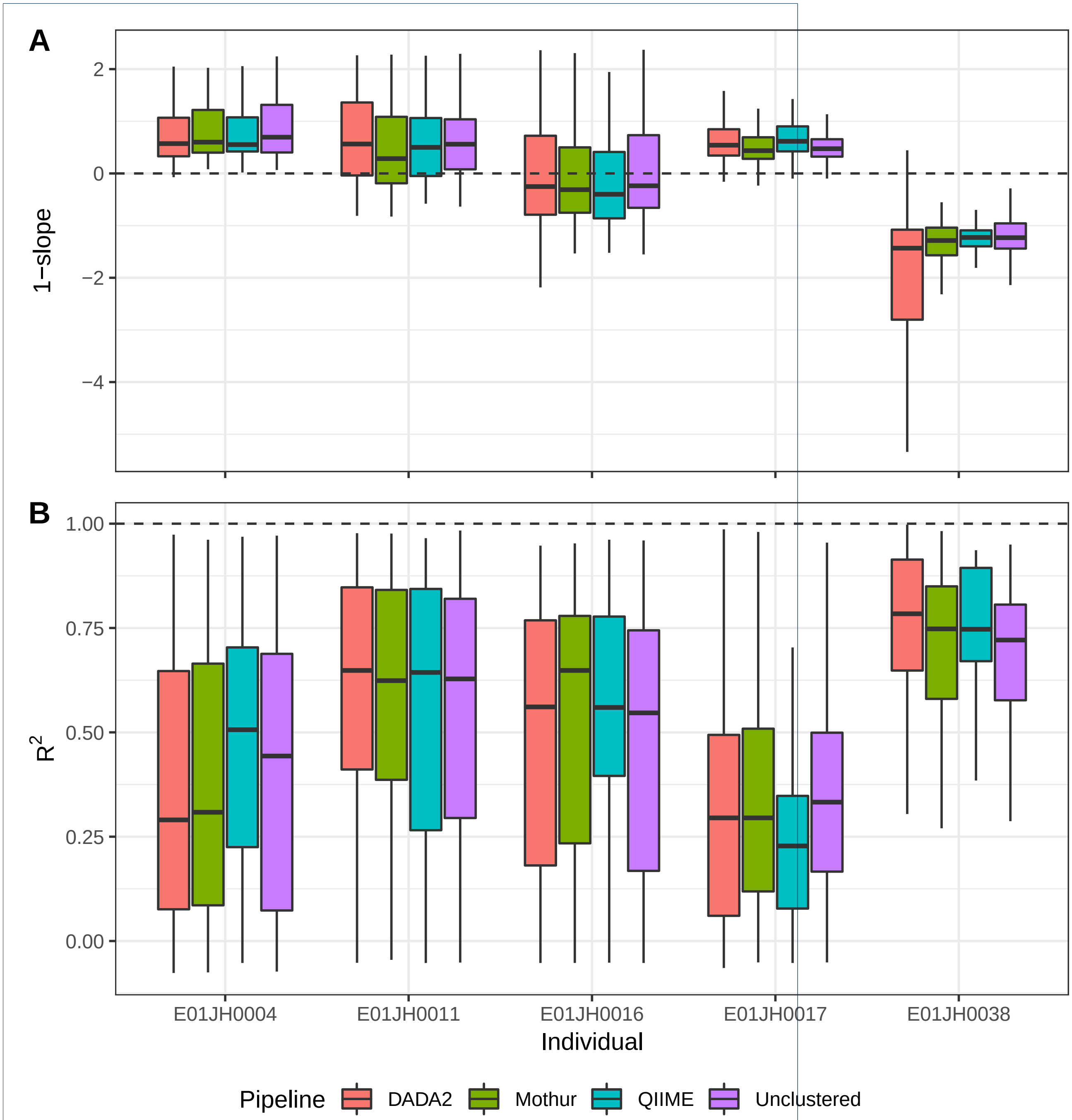


Figure 7 Feature-level ~~log-fold~~ differential abundance assessment.Log-fold change error bias (A) and variance (B) metric distribution by subject and pipeline. The bias $(1 - slope)$ and variance (R^2) metrics are derived from the linear model fit to the estimated and expected log fold-change values 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.