Microbiome-Scale Mixture Use Demonstration

Nate Olson

2017-11-20

1 Introduction

Metagenomics, sequencing microbial community DNA, has greatly advanced our understanding of the microbial world. Targeted sequencing of the 16S rRNA gene, 16S metagenomics, is a commonly used method for sequencing a microbial community. 16S metagenomics is a complex measurement process comprised of multiple molecular laboratory and computational steps (Julia K Goodrich et al. 2014; Kim et al. 2017). There are numerous sources of error and bias in the measurement process, for both the molecular (e.g. PCR and sequencing) and computational steps (e.g. sequence clustering) (D'Amore et al. 2016; Julia K. Goodrich et al. 2014; Brooks et al. 2015). Appropriate datasets and methods are needed to evaluate the 16S measurement process inorder to characterize how these sources of bias and error impact the measurement results and determine where to focus efforts for improving the measurement process.

In order to characterize the accuracy of a measurement process you need a sample or dataset with an expected value to benchmark against. There have been a number of studies characterizing and evaluating different steps in the 16S rRNA metagenomics measurement process which use mock communities, simulated data, or environmental samples. Mock communities consisting of mixtures of cells or DNA from individual organisms have an expected value but are not representative of the complexity of environmental samples in terms of the of number or abundance distributions of organisms (Bokulich et al. 2016). Similar to mock communities simulated data have an expected value that can be used for benchmarking. However, the sequencing error profile is not completely understood and therefore simulated sequencing data does not recapitulate the complexity of sequencing data generated from an environmental sample. While simulated data and mock communities are usefull in evaluating and benchmarking new methods one needs to consider that methods optimized for mock communities and simulated data are not necessarily optimized to handle the biases, noise, and diversity present in real samples. Data generated from environmental samples, which include the biases, error, and diversity of real samples, are often used to benchmark new molecular laboratory and computational methods. However, without an expected value to compare to only measurement precision, similarity of results to thoes generated using a different method, can be evaluated. An alternative to these types of data is sequencing data generated from mixtures of environmental samples. By mixing environmental samples at known proportions you can use information obtained from the unmixed samples and how they were mixed to obtain an expected value for use in assessing the measurement process. Mixtures of environmental samples have previously been used to evaluate gene expression measurements microarrays and RNAseq (Parsons et al. 2015; Pine, Rosenzweig, and Thompson 2011; Thompson et al. 2005)

Mock communities are most commonly used to assess the qualitative characteristics of a dataset. As the number of organisms in the mock community is known the total number of features can be compared to the expected number. The number of observed features in a mock community is often significantly higher than the expected number of organism **REF**. A noteable exception to this is mock community benchmarking studies evaluating sequencing inference method, such as DADA2 (Callahan et al. 2016). The higher than expected number of features is often attributed to sequencing and PCR artifacts as well as reagent contaminants **REF**.

The quantitative characteristics of 16S metagenomic measurement process is normally assessed using both mock communities and simulated data. Mock communities of equimolar and staggered concentration are used to assess the quantiative accuracy of the relative abundance estimates **REF**. Results from relative abundance estimates using mock communities generated from mixtures of DNA have shown taxonomic specific effects where individual taxa are under or over represented in a sample. These taxonomic specific effects have been attributed to primer mismatches **REF**. To assess differential abundance, simulated datasets are used where specific taxa are artifically over represented in one set of samples compared to another **REF**.

Using simulated data to assess log fold-change estimates only evaluate computational steps of the measurement process.

In the present study we developed a mixture dataset of extracted DNA from human stool samples for assessing the 16S metagenomic measurement process. The mixture datasets was processed using three bioinformatic pipelines. We developed metrics for qualitative and quantitative assessment of the bioinformatic pipeline results. The quantitative results were similar across pipelines but the qualitative results varied across pipelines. Additionally, the dataset and metrics developed in this study can be used to evaluate new bioinformatic pipelines.

2 Methods

2.1 Two-Sample Titration Design

Samples from a vaccine trial were used to generate a two-sample titration dataset for assessing 16S metagenomic computational methods (Harro et al. 2011). Pre- and post-exposure sample from five trial participants were selected based on the following criteria no *Escherichia coli* detected in stool samples using qPCR and 16S metagenomic sequencing before exposure (pre-exposure) to Enterotoxigenic *Escherichia coli* (ETEC)) and timepoints with the highest concentration of *E. coli* after exposure (post-exposure) (Pop et al. 2016, Fig. 1A). For the two-sample titration post-exposure samples were titrated into pre-exposure samples with log_2 changes in pre to post sample proportions (Fig. 1B). Unmixed samples were diluted to 12.5 $ng/\mu L$ in tris-EDTA buffer prior to making the two-sample titrations. Initial DNA concentration was measured using NanoDrop ND-1000 (Thermo Fisher Scientific Inc. Waltham, MA USA).

By using a two-sample titration mixture design the expected relative abundance of a feature can be determined using the following equation (1). Where θ_i, is the proportion of post-exposure DNA in titration i, C_{ij} is the relative abundance of feature j in titration i, and C_{post_j} and C_{pre_j} are the relative abundance of feature j in the unmixed pre- and post-exposure samples.

$$C_{ij} = \theta_i C_{post_j} + (1 - \theta_i) C_{pre_j} \tag{1}$$

2.2 Titration Validation

qPCR was used to validate the volumetric mixing of the unmixed samples and check of differences in the proportion of prokaryotic DNA across titrations. To ensure that the two-sample titrations were volumetrically mixed according to the mixture design independent ERCC plasmids were spiked into the unmixed pre- and post-exposure samples (Baker et al. 2005) (NIST SRM SRM 2374) (Table **??**). The ERCC plasmids were resuspendended in 100 $ng/\mu L$ tris-EDTA buffer and 2 $ng/\mu L$ was spiked into the approporiate unmixed sample. Plasmid abundance was quantified using TaqMan gene expression assays (FAM-MGB) (Catalog # 4448892, ThermoFisher) specific to each ERCC plasmids using the TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA). To check for differences in the proportion of bacterial DNA in the pre- and post-exposure samples, bacterial DNA concentration in the titrations was quantified using the Femto Bacterial DNA quantification kit (Zymo Research, Irvine CA). All samples were run in triplicate along with a standard curve. An in-house standard curve consisting of log_{10} dilutions of *E. coli* DNA was used as the standard curve. All qPCR assays were performed using the QuantStudio Real-Time qPCR (ThermoFisher). The amplification data and Ct values were exported from the QuantStudio™ Design and Analysis Software v1.4.1 as tsv files for statistical analysis. Statistical analysis was performed using the R programming language.

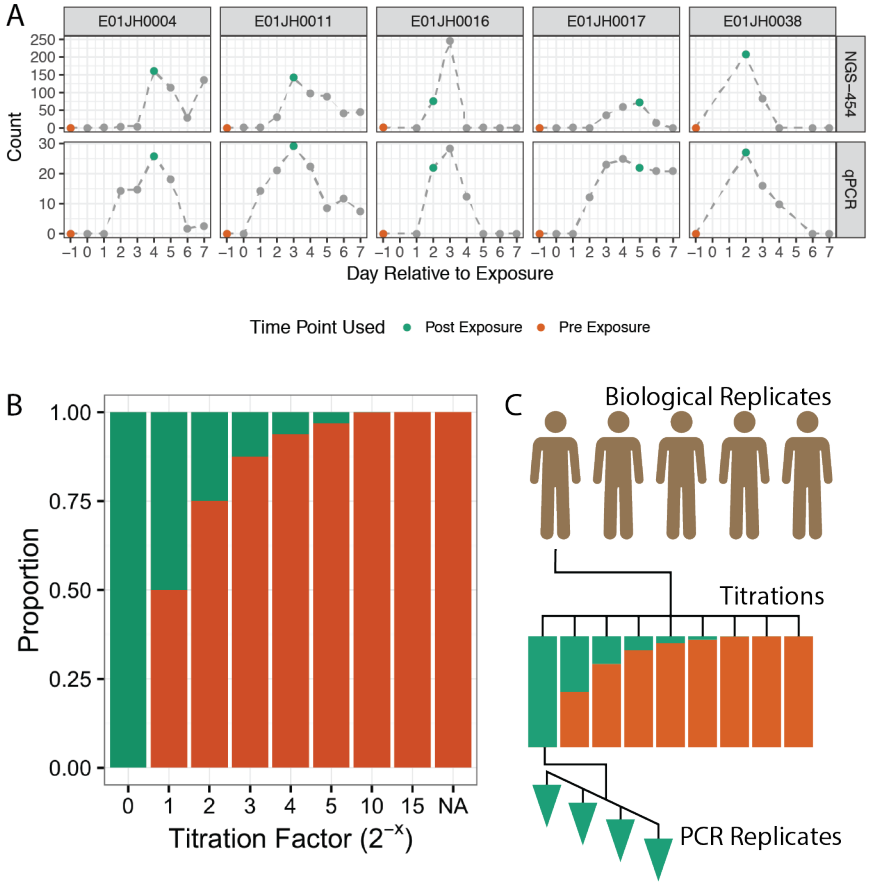

Figure 1: Sample selection and experimental design for two-sample titration 16S rRNA metagenomic sequencing assessment dataset. A) Pre- and post-exposure samples from five participants in a vaccine trial (Harro et al. 2011) were selected based on *Escherichia coli* abundance measured using qPCR and 454 16S rRNA metagenomics sequencing (454-NGS), data from Pop et al. (2016). Pre- and post-exposure samples are indicated with orange and green data points, respectively. Grey points indicates other samples from the vaccine trial time series. B) The pre-exposure samples were titrated into post-exposure samples following a log_2 dilution series. The NA titration factor represents the unmixed pre-exposure sample. C) Pre- and post-exposure samples from the five vaccine trial participants were used to generate independent two-sample titration series. The result was a total of 45 samples, 7 titrations + 2 unmixed samples times 5 biological replicates. Four replicate PCRs were performed for each of the 45 samples resulting in 190 PCRs.

2.3 Sequencing

The 45 samples (seven titrations and two unmixed samples for the five biological replicates) were processed using a standard 16S rRNA amplicon sequencing workflow based on the Illumina 16S library protocol (16S Metagenomic Sequencing Library Preparation, posted date 11/27/2013, dowloaded from https://support.illumina.com). The protocol consisted of an initial 16S rRNA PCR followed by a separate sample indexing PCR prior to normalization and pooling.

A total of 192 PCRs were run including four PCR replicates per sample and 12 no template controls. The 16S PCR targeted the V3-V5 region, Bakt_341F and Bakt_806R (Klindworth et al. 2012). The V3-V5 target region is 464 bp, with forward and reverse reads overlaping by 136 bp, assuming 300 bp paired-end reads (Yang, Wang, and Qian 2016) (http://probebase.csb.univie.ac.at). The primer sequences include additional overhang adapter sequences to facilitate library preparation (forward primer 5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG - 3' and reverse primer 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC - 3'). The 16S targeted PCR was performed according to the Illumina protocol using the KAPA HiFi HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA). The resulting PCR product was verified using agarose gel electrophoresis. Quality control DNA concentration measurements were made after the initial 16S rRNA PCR, indexing PCR, and normalization. DNA concentration was measured using SpextraMax Accuclear Nano dsDNA Assay Bulk Kit (Part# R8357#, Lot 215737, Molecular Devices LLC. Sunnyvale CA, USA) and flourescent measurements were made with a Molecular Devices SpectraMax M2 spectraflourometer (Molecular Devices LLC. Sunnyvale CA, USA).

The 16S rRNA PCR product was used to generate sequencing libraries. The initial PCR products were purified using AMPure XP beads (Beckman Coulter Genomics, Danvers, MA) following the manufactures protocol. After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). Prior to pooling the purified sample concentration was normalized using SequalPrep Normalization Plate Kit(Catalog n. A10510-01, Invitrogen Corp., Carlsbad, CA), according to the manufactuers protocol. The pooled library concentration was measured using the Qubit dsDNA HS Assay Kit (Part# Q32851, Lot# 1735902, ThermoFisher, Waltham, MA USA). Due to the low concentration of the pooled amplicon library the modified protocol for low concentration libraries was used. The library was run on a Illumina MiSeq and base calls were made using Illumina Real Time Analysis Software version 1.18.54. Sequencing data quality control metrics for the 384 datasets in the study (192 samples with forward and reverse reads) were computed using the bioconductor `Rqc` package (Souza and Carvalho 2017) to calculate the quality metrics used in the following analysis.

2.4 Sequence Processing

Sequence data was processed using three bioinformatic pipelines, a _de-novo_clustering method - Mothur (Schloss et al. 2009), an open-reference clustering method - QIIME (Caporaso et al. 2010), and a sequence inference methods - DADA2 (Callahan et al. 2016), unclustered sequences as a control. Code used to run the bioinformatic pipelines is available at https://github.com/nate-d-olson/mgtst_pipelines. The Mothur (version 1.37, http://www.mothur.org/) pipeline was based on the MiSeq SOP (Schloss et al. 2009; Kozich et al. 2013). As a different 16S rRNA region was sequenced than the region the SOP was developed for the procedure was modified to account for smaller overlap between the forward and reverse reads relative to the amplicons used in the protocol. The Mothur pipeline included an initial pre-processing step where forward and reverse reads were merge using the Needleman-Wunsch algorithm. Low quality reads were identified based on presence of ambiguous bases, reads that failed alignment to the SILVA reference database (V119, https://www.arb-silva.de/) (Quast et al. 2012), and chimeras were filtered from the dataset. Chimera filtering was performed using UChime without a reference database (Edgar et al. 2011). OTU clustering was performed using the OptiClust algorithm with a clustering threshold of 0.97 (Westcott and Schloss 2017). The RDP classifier implemented in mothur was used for taxonomic classification against the mothur provided version of the RDP v9 training set (Wang et al. 2007). The QIIME open-reference clustering pipeline for

paired-end Illumina data was performed according to the online tutorial (http://nbviewer.jupyter.org/github/biocore/qiime/blob/1.9.1/examples/ipynb/illumina_overview_tutorial.ipynb) using QIIME version 1.9.1 (Caporaso et al. 2010). Briefly the QIIME pipeline uses fastq-join to merge paired-end reads (Aronesty 2011) and the Usearch algorithm (Edgar 2010) with Greengenes database version 13.8 with a 97% similarity threshold (DeSantis et al. 2006) was used for open-reference clustering. DADA2 a R native pipeline was also used to process the sequencing data (Callahan et al. 2016). The pipeline includes a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP naive bayesian classifier (Wang et al. 2007) and the SILVA database V123 provided by the DADA2 developers (Quast et al. 2012, https://benjjneb.github.io/dada2/training.html). The unclustered pipeline was based on the mothur *de-novo* clustering pipeline, where the paired-end reads were merged, filtered, and then dereplicated. Reads were aligned to the reference Silva alignment (V119, https://www.arb-silva.de/), and reads failing alignment were excluded from the dataset. Taxonomic classification of the unclustered sequences was performed using the same RDP classifier implmented in mothur used for the *de-novo* pipeline. To limit the size of the dataset the most abundant 40,000 OTUs (comparable to the mothur dataset), across all samples, were used as the unclustered dataset.

2.5 Data Analysis

Prior to qualitative and quantitative assessment relative abundance and log fold-changes were estimated and the proportion of prokaryotic DNA in each titration, θ, was inferred. A negative binomial model was used to calculate the average relative abundance across PCR replicates for individual features. Log fold-changes between all titration pairs and pre- and post-exposure samples were calculated using EdgeR (Robinson, McCarthy, and Smyth 2010; McCarthy et al. 2012). To account for differences in the proportion of bacterial DNA in the pre- and post-exposure samples. A linear model was used to infer θ in equation (2), where $\mathbf{C}$ is a vector of counts for a set of features, $\mathbf{C}_{obs_j}$ observed counts for titration j, with $\mathbf{C}_{pre_j}$ cand $\mathbf{C}_{post_j}$ representing the vector of counts for the same features for the unmixed pre- and post-exposure samples. To summarize counts across PCR replicates and account for differences in sequencing depth, negative binomial relative abundance estimates were used to infer θ. 16S rRNA sequencing count data is know to have a non-normal mean-variance relationship resulting in poor model fit for standard linear regression. Generalized linear models provide an alternative to standard least-squares regression however, the above model is additive and therefore unable to directly infer θ_j 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 bootstraping with 1000 replicates. To limit the impact of uninformative and low abundance features a subset of individual specific features were used to infer θ. Featured were observed in at least 14 of the 28 total titration PCR replicates (4 pcr replicates per titration, 7 titrations), greater than 1 log_2 fold-change between the pre- and post-exposure samples, and present in all four or none of the pre- and post-exposure PCR replicates.

$$C_{obs_j} = \theta_j(C_{post_j} - C_{pre_j}) + C_{pre_j} \tag{2}$$

2.6 Quantitative Assessment

To quantitatively assess the count table values the expected relative abundance and log fold-change values were compared to the relative abundance estimates (*obs*) calculated using a negative binomial model and the EdgeR log fold-change estimates. Equation (1) and the inferred θ values were used to calculate the expected feature relative abundance (*exp*). The error rate metrics for the relative abundance estimates were compared across pipelines and biological replicates. Error rate was defined as $|exp - obs|/exp$. Mixed effects models were used to compare feature-level error rate bais and variance across pipelines accounting for individual effect. Feature-level bias and variance were evaluated using the median error rate and robust coefficient of variation ($RCOV = IRQ/median$) respectively. Large feature-level error rate bias and variance outliers were observed, these outliers were excluded from the mixed effects model to minimize biases in the model due to poor fit for a numer of features and were characterized independently.

To assess differential abundance log fold-change estimates, log fold-change between all titrations were compared to the expected log fold-change values for the pre-specific and pre-dominant features. When assuming the feature is only present in pre-exposure samples the expected log fold-change is independent of the observed counts for the unmixed samples. Expected log fold-change between titrations i and j is calculated using (3), where θ is the proportion of post-exposure bacterial DNA in a titration. Pre-dominant and pre-specific features were defined as features observed in all four pre-exposure PCR replicates and a log fold-change between pre- and post-exposure samples greater than 5. Pre-specific features were not observed in any of the post-exposure PCR replicates and pre-dominant features were observed in one or more of the post-exposure PCR replicates. Only individuals with consistent inferred and estimated θ values were included in the log fold-change analysis, E01JH0004, E01JH0011, and E01JH0016.

$$logFC_{ij} = log_2\left(\frac{1-\theta_i}{1-\theta_j}\right) \tag{3}$$

2.7 Qualitative Assessment

For the qualitative measurement assessment we evaluated features only observed in either the unmixed samples, unmixed-specific features, or the titrations, titration-specific features. Features are unmixed- or titration-specific 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. These features can be considered false positives or negatives.

We tested if sampling alone could explain feature specificity. For unmixed-specific features we used a binomial test and for titration-specific features we used a Bayesian hypothesis test. For both tests p-values were adjusted for multiple comparisons using the Benjamini & Hochberg method (Benjamini and Hochberg 1995). To determine if sampling alone can explain unmixed-specific features the binomial test was used to test the following hypothesis;

H_0 - Given no observed counts and the total abundance for a titration the true proportion of a feature is **equal to** the expected proportion.

H_1 - Given no observed counts and the total abundance for a titration the true proportion of a feature is **less than** the expected proportion.

To test if titration-specific features could be explained by sampling alone we used a Bayesian hypothesis test. Simulation was used to estimate probabilities a feature was observed given the observed counts and sequencing depth. For the simulation we assumed a binomial distribution given the observed total abundance and a uniform distribution of proportions, 0 to the minimum expected proportion. The minimum expected proportion, $\pi_{min_{exp}}$, is calculated using the mixture equation (1) and the minimum observed feature proportion for unmixed pre-exposure, $\pi_{min_{pre}}$, and post-exposure $\pi_{min_{post}}$ samples for each individual and pipeline. For features not present in unmixed samples the assumption is that the feature proportion is less than $\pi_{min_{exp}}$.

We formulated our null and alternative hypothesis for the Bayesian test as follows,

H_0 - Given the total abundance for a sample and minimum expected proportion the true proportion of a feature is **less than** the minimum expected observed proportion.
H_1 - Given the total abundance for a sample and minimum expected proportion the true proportion of a feature is **greater than or equal to** the minimum expected proportion.

The following equations @ref(eq:probPi;eq:probC) were used to calculate the p-value for the Bayesian hypothesis test assuming equal priors, i.e. $P(\pi < \pi_{min_{exp}}) = P(\pi \geq \pi_{min_{exp}})$.

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

$$P(C \geq C_{obs}) = P(C \geq C_{obs} | \pi < \pi_{min_{exp}}) P(\pi < \pi_{min_{exp}}) + P(C \geq C_{obs} | \pi \geq \pi_{min_{exp}}) P(\pi \geq \pi_{min_{exp}}) \quad (5)$$

Table 1: Summary statistics for the different bioinformatic pipeliens. DADA2 is a denoising sequence inference pipeline, QIIME is a 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 feature abundance. Filter rate is the proportion of reads that were removed while processing the sequencing data for each bioinformatic pipeline.

Pipelines	Features	Sparsity	Total Abundance	Drop-out Rate
dada2	3144	0.93	68649 (1661-112058)	0.24 (0.18-0.59)
mothur	38469	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
qiime	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

3 Results

3.1 Dataset characteristics

Quality assessment of sequencing run summarizing number of reads per sample. Two barcoded experimental samples have less than 35,000 reads (Fig. 2A). The rest of the samples with less than 35,000 reads are no template PCR controls (NTC). Excluding the one failed reaction with 2,700 reads and the NTCs, the total range in the observed number of sequences per samples is 3195 to 152267 reads with a median library size of 8.9548×10^4. For the expected overlap region, based on primer positions and read lengths (16S PCR fig), the forward read has consistently higher base quality scores relative to the reverse read with a narrow overlap region with high base quality scores for both forward and reverse reads (Fig. 2B).

The sequencing dataset was processed using four bioinformatic pipelines. The resulting count tables were characterized for number of features, sparsity, and filter rate (Table 1, Fig. 2C). The pipelines evaluated have different approaches for handling low quality reads resulting in the large variability in filter rate (Table 1). QIIME pipeline has the highest filter rate and highest number of features per sample. The targeted amplicon region has a relatively small overlap region, 136 bp for 300 bp paired end reads. The high filtration rate is due to the drop in base calling accuracy at the ends of the reads especially the reverse reads resulting in a high frequency of unsuccessfully merged reads pairs (Fig. 2B). Additionally, to remove potential sequencing artifacts from the dataset QIIME excludes singletons, OTUs only observed once in the dataset. The expectation is that this mixture dataset will be less sparse relative to other datasets due to the redundant nature of the samples where 35 of the samples are derived directly from the other 10 samples and there are four PCR replicates for each sample. Sparsity was lower for *de-novo* clustering (QIIME) than sequence inference (DADA2) even though DADA2 has fewer total features. 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.

3.2 Titration Series Validation

In order to use information from the unmixed samples to obtain expected count values for the titrations we first need to evaluate two assumptions about the mixed samples: 1. The samples were mixed volumetrically in a log_2 dilution series according to the mixture design. 2. The unmixed pre- and post-exposure samples have the same proportion of prokaryotic DNA. Exogenous DNA was spiked into the unmixed samples prior to mixing and quantified using qPCR to validate the samples were volumetrically mixed according to the mixture design. To evaluate the second assumption total prokaryotic DNA in the titrations samples was quantified using a qPCR assay targeting the 16S rRNA gene.

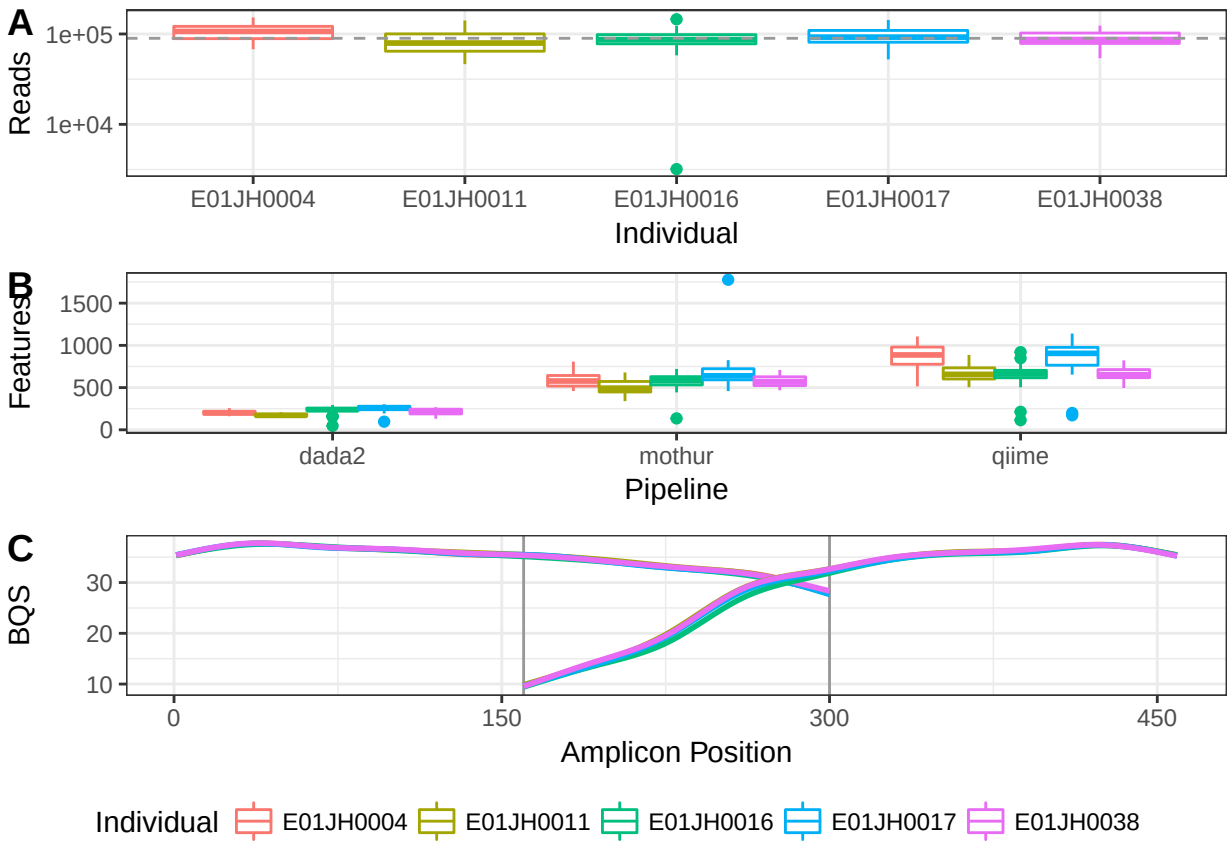

Figure 2: Sequencing dataset summary. (A) Distribution in the number of reads per barcoded sample (Library Size) by individual. Dashed horizontal line indicates overall median library size. (B) Smoothing spline of the base quality score (BQS) by sequencing cycle. Vertical lines indicate approximate overlap region between forward and reverse reads. (C) Distribution of the number of features per sample.

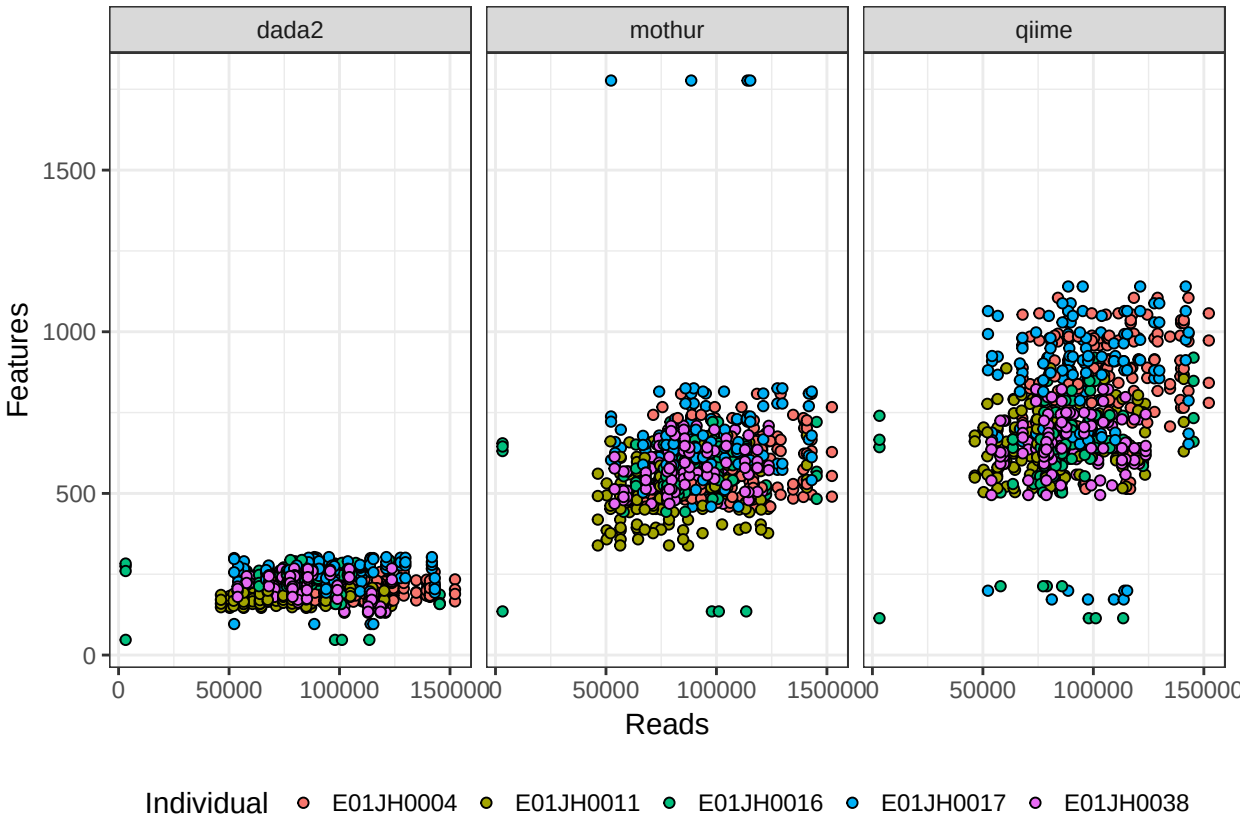

Figure 3: Relationship between the number of reads and features per sample by bioinformatic pipeline.

Table 2: ERCC Spike-in qPCR assay information and summary statistics. ERCC is the ERCC identifier for the ERCC spike-in, Assay is TaqMan assay, and Length and GC are the size and GC contenct 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.

Treatment	Individual	ERCC	Assay	Length	Std. R^2	E	R^2	Slope
Post	E01JH0004	ERCC-00012	Ac03459877-a1	77	0.9996	86.19	0.98	0.92
	E01JH0011	ERCC-00157	Ac03459958-a1	71	0.9995	87.46	0.95	0.90
	E01JH0016	ERCC-00108	Ac03460028-a1	74	0.9991	87.33	0.95	0.84
	E01JH0017	ERCC-00002	Ac03459872-a1	69	0.9968	85.80	0.89	0.93
	E01JH0038	ERCC-00035	Ac03459892-a1	65	0.9984	86.69	0.95	0.94
Pre	E01JH0004	ERCC-00084	Ac03459922-a1	63	0.9972	84.36	0.53	-2.09
	E01JH0011	ERCC-00034	Ac03459987-a1	58	0.9999	87.93	0.52	-1.56
	E01JH0016	ERCC-00057	Ac03460000-a1	78	0.9990	84.22	0.60	-1.95
	E01JH0017	ERCC-00130	Ac03460039-a1	72	0.9979	89.78	0.32	-1.66
	E01JH0038	ERCC-00092	Ac03459925-a1	87	0.9994	84.30	0.21	-1.86

3.2.1 Spike-in qPCR results

Volumetric mixing of the two-sample titration was validated using qPCR to quantify ERCC plasmids spiked into the pre- and post-exposure samples. 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 (Table 2). The qPCR assays targeting the ERCCs spiked into the post-exposure samples had R^2 values and slope estimates close to 1 (Table 2). For a log_2 two-sample-titration mixture design the expected slope is 1, corresponding to a doubling in template DNA every PCR cycle. Slope estimates less than 1 were attributed to the assay standard curve efficiency less than 1 (Table 2). For the pre-exposure ERCCs a regression line was fit to the log_2 pre-exposure sample proportion for titrations 1-4 and the unmixed pre-exposure sample. The change in pre-exposure sample proportion between titrations 5, 10, and 15 (0.97 - 0.99997) is to small for qPCR to detect changes in ERCC spike-in concentration with an expected Ct difference of 0.04 between the titrations 5 and 15. For a regression line was fit to the Ct values and log_2 pre-exposure sample proportion with a -1 expected slope as the spike-in concentration is expected to increase linearly with the proportion of pre-exposure sample and both Ct and pre-exposure sample proportions are on log_2 scales. ERCCs spiked into the pre-exposure samples the R^2 values were low, less than 0.6, with slope estimates between -1.5 and -2.1 (Table 2). Deviation from the expected slope for the pre-exposure ERCC qPCR results is attributed to the small change in spike-in concentration between samples preventing accurate quantification of changes in spike-in concentration between titrations. When considering the quantitative limitations of the qPCR assay these results indicate that the unmixed pre- and post-exposure samples were volumetrically mixed according to the mixture design.

3.2.2 Bacterial DNA Concentration

Prokaryotic DNA concentration changes across titrations indicating the proportion of bacterial DNA from the unmixed pre- and post-exposure samples in a titration is not consistent with the mixture design (Fig. 4). 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-exposure samples the slope of the concentration estimates across the two-sample titration would be 0. For individuals where the proportion of prokaryotic DNA is higher in the pre-exposure samples the

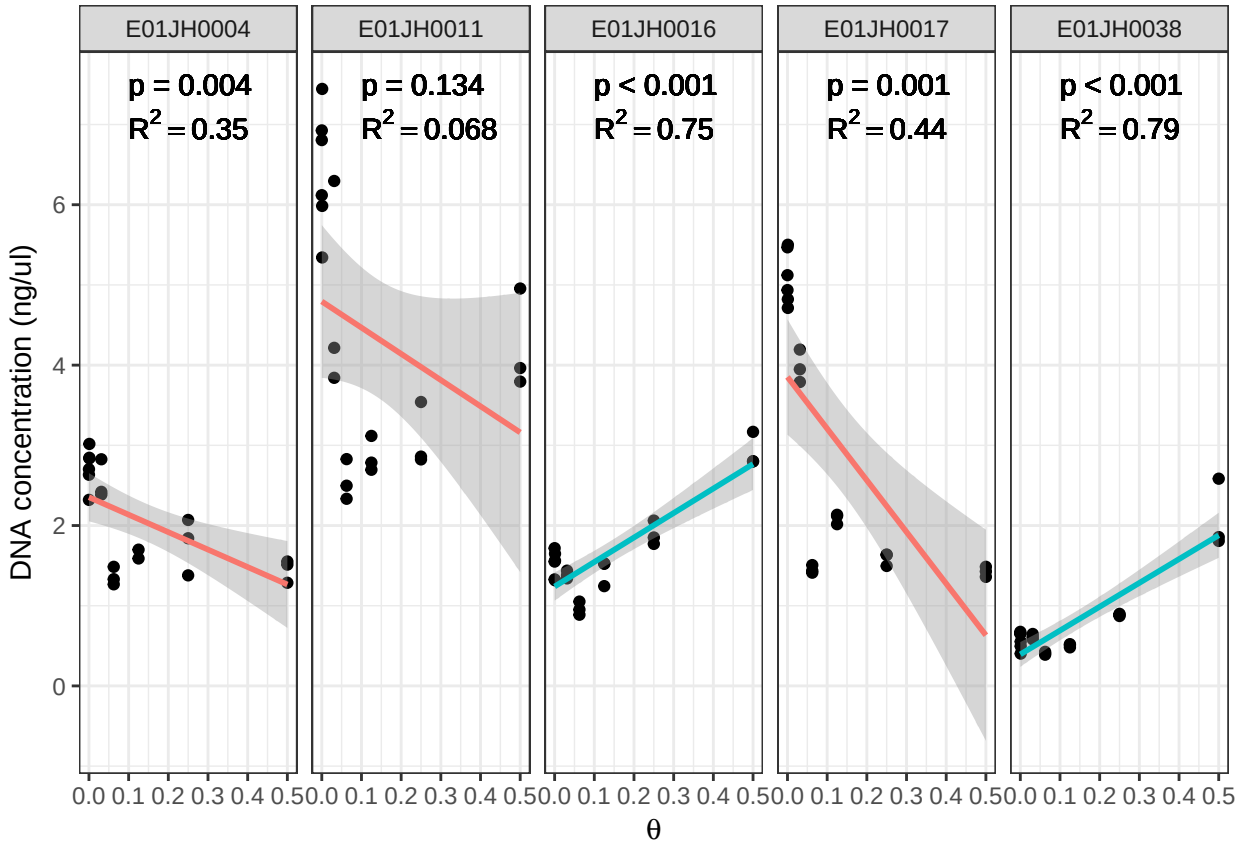

Figure 4: Prokaryotic DNA concentration (ng/ul) across titrations measured using a 16S rRNA qPCR assay. Separate linear models, Prokaryotic DNA concentration versus θ were fit for each individual, and R^2 and p-values were reported. Red lines indicate negative slope estimates and blue lines positive slope estimates. p-value indicates signficant difference from the expected slope of 0. Multiple test correction was performed using the Benjamini-Hochberg method. One of the E01JH0004 PCR replicates for titration 3 ($ heta=0.125$) was identified as an outlier, with a concentration of 0.003, and was excluded from the linear model. The linear model slope was still significantly different from 0 when the outlier was included.

slope will be negative and positive when the proporition is higher for post-exposure samples. The slope estimates are significantly different from 1 for individuals all individuals excluding E01JH0011 (Fig. 4). These results indicate that the proportion of prokaryotic DNA is lower in the post-exposure when compared to the pre-exposure samples for E01JH0004 and E01JH0017 and higher for E01JH0016 and E01JH0038.

Across all titrations median prokaryotic DNA concentration varied by individual, with E01JH0011 having the highest concentration, 3.84 $ng/\mu l$, and E01JH0038 having the lowest concentration, 0.61 $ng/\mu l$. As the DNA concentration for the unmixed samples was normalized to 12.5 $ng/\mu l$ prior to generating the titration series, the proportion of DNA in the samples targeted by 16S sequencing method ranged from 0.31 to 0.05.

3.2.3 Theta Estimates

To account for differences in the proportion of prokaryotic DNA in the pre- and post-exposure (Fig. 4) we inferred the proportion of post-exposure sample prokaryotic DNA in a titration, using the 16S rRNA sequencing data, θ (Fig. 5). Overall the relationship between the inferred and mixture design θ values were consistent across pipelines but not individual whereas the 95% CI varied by both individual and pipeline. For E01JH0004, E01JH0011, and E01JH0016 the inferred and mixture design θ values were in better agreement compared to E01JH0017 and E01JH0038. For E01JH0017 the inferred values were consistently less than the mixture design values, and greater than the mixture design values for E01JH0038. These results were consistent with the qPCR prokaryotic DNA concentration results with E01JH0004 and E01JH0017 having a significantly positive slopes then E01JH0016 and E01JH0038 significantly negative slopes (Fig. 4).

3.3 Measurement Assessment

Next we assessed the qualitative and quanitative nature of 16S rRNA measurement process using our two-sample titration dataset. For the qualitative assessment we analyzed the relative abundance of features only observed in the unmixed samples and titrations. For the quantitative assessment we looked the the relative abundance and differential abundance log fold-change estimates.

3.3.1 Qualitative Assessment

A number of unmixed- and titration-specific features were observed (titration-specific: Fig. 6A, unmixed-specific: Fig. 6B). There were unmixed-specific features with expected counts that could not be explained by sampling alone for all individuals and bioinformatic pipelines (Fig. 6C). However, the proportion of unmixed-specific features that could not be explained by sampling alone varied by bioinformatic pipeline. DADA2 had the highest rate of unmixed-specific features could not be explained by sampling alone whereas QIIME had the lowest rate of unmixed-specific features. Consistent with the distribution of observed counts for titration-specific features more of the DADA2 features could not be explained by sampling alone compared to the other pipelines (Fig. 6D).

3.3.2 Quantitative Assessment

For the relative abundance assessment we evaluated the consistency of the observed and expected relative abundance estimates for a feature across the four titration PCR replicates as well as feature-level bias and variance. The pre- and post-exposure estimated relative abundance and inferred θ values were used to calculate titration and feature level error rates. To prevent over-fitting, unclustered pipeline θ estimates were used to calculate the error rates. Only features observed in all pre- and post-exposure PCR replicates and pre- and post-exposure specific features were included in the analysis (Table 5). Pre- and post-exposure specific features were defined as present in all four PCR replicates of the pre-exposure or post-exposure PCR replicates, respectively, but none of the PCR replicates for the other unmixed sample. There is lower confidence in the relative abundance of a feature in the pre- or post-exposure unmixed samples when the

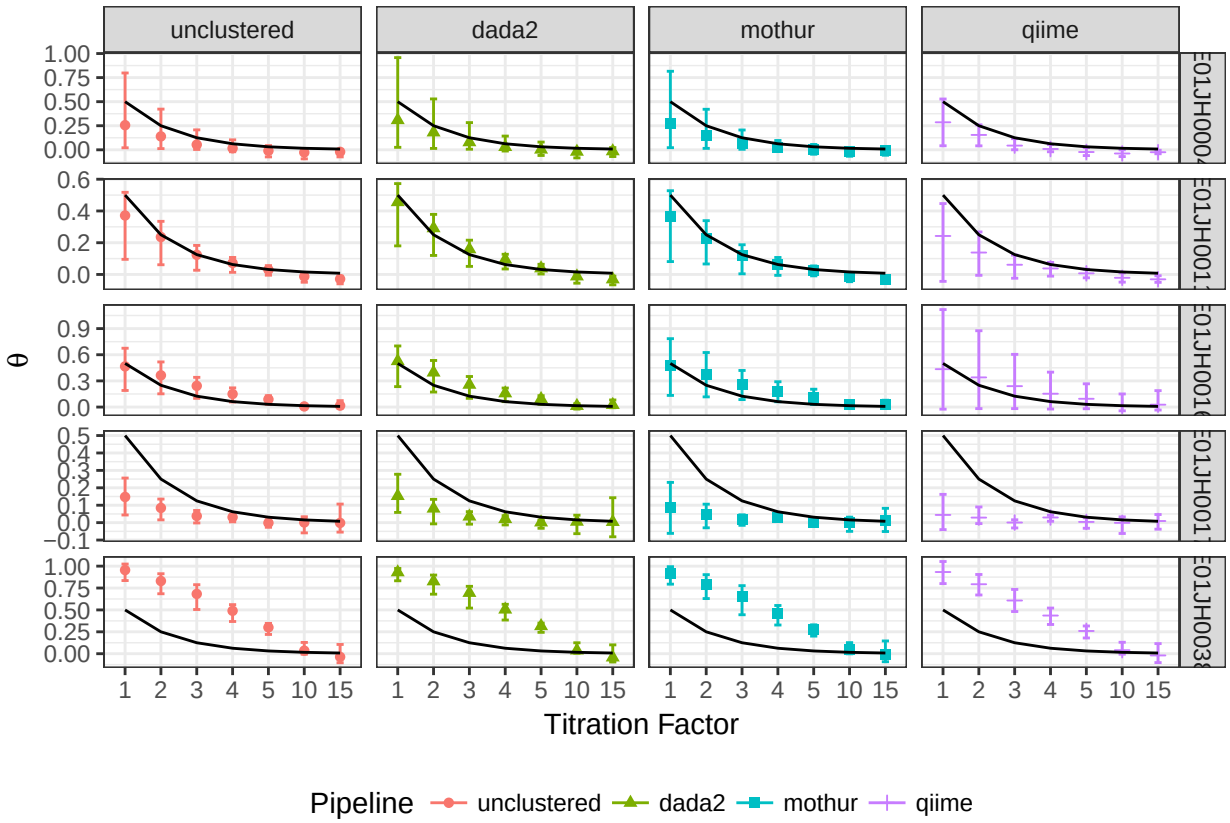

Figure 5: Theta estimates by titration, biological replicate, and bioinformatic pipeline. The points indicate mean estimate of 1000 bootstrap theta estimates and errorbars 95% confidence interval. The black line indicates the expected theta values. Theta estimates below the expected theta indicate that the titrations contains less than expected bacterial DNA from the post-treatment sample. Theta estimates greater than the expected theta indicate the titration contains more bacterial DNA from the pre-treatment sample than expected.

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

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

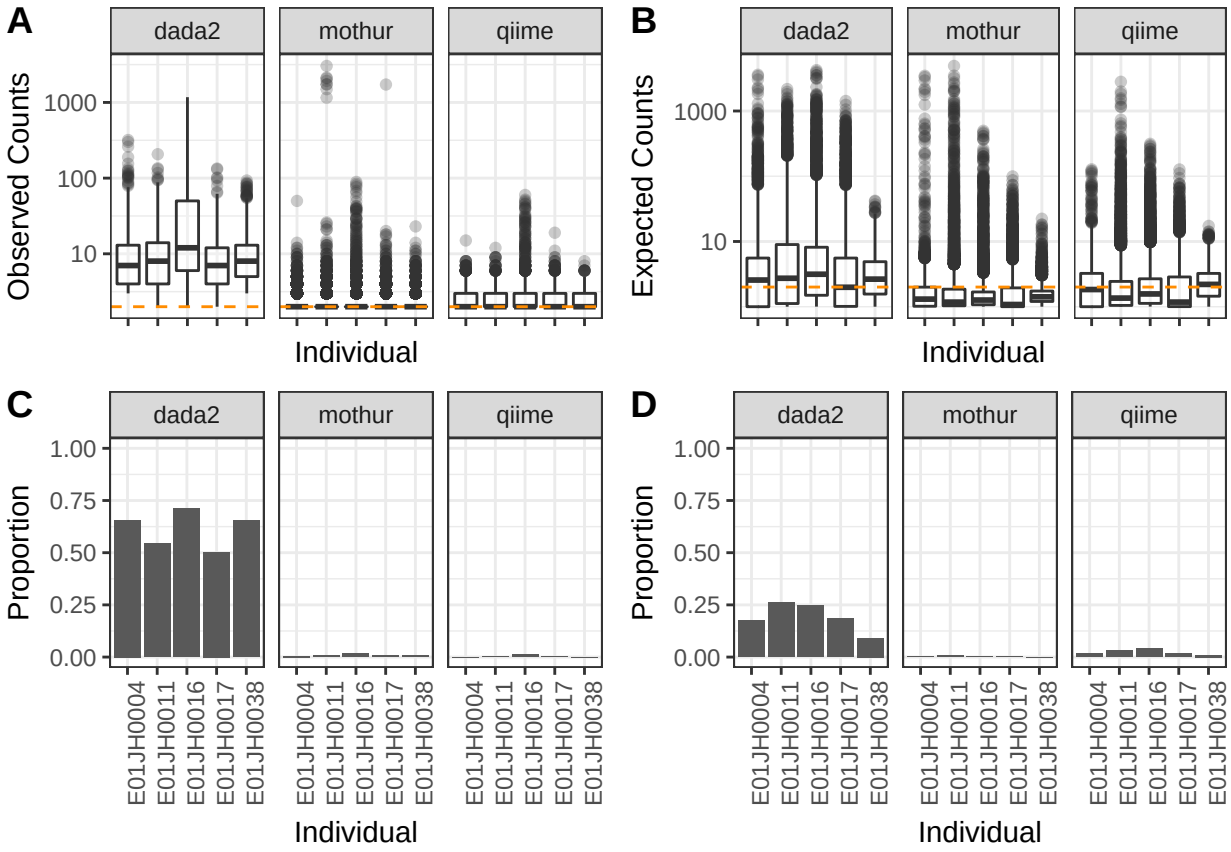

Figure 6: Distribution of (A) observed count values for titration-specific features and (B) expected count values for unmixed-specific features by pipeline and individual. The orange horizontal dashed line indicates a count value of 1. (C) Proportion of unmix-specific features and (D) titration-specific 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 not explained by sampling alone.

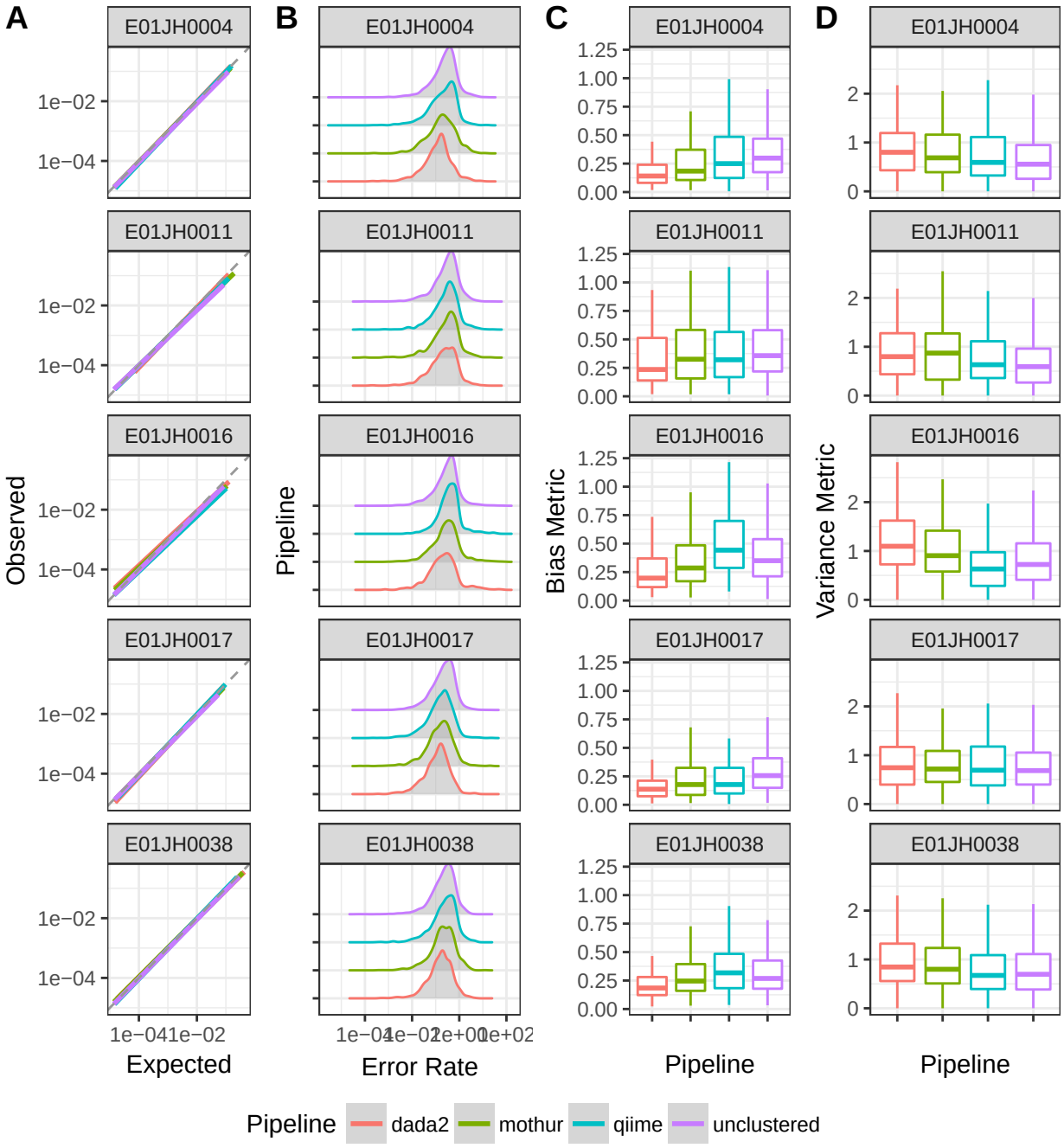

Figure 7: Relative abundance assessment. (A) A linear model of the relationship between the expected and observed relative abundance. The dashed grey line indicates expected 1-to-1 relationship. The plot is split by individual and color is used to indicate the different bioinformatic pipelines. A negative binomial model was used to calculate an average relative abundance estimate across the four PCR replicates. Points with observed and expected relative abundance values less than 1/median library size were excluded from the data used to fit the linear model. (B) Distribution of relative abundance error rate by individual and pipeline. Distribution of feature-level relative abundance (C) bias metric - median error rate and (D) variance - robust coefficient of variation ($RCOV = (IQR)/|median|$) by individual and pipeline. To prevent extreme metric values from obscuring metric value visual comparisons, boxplot outliers, $1.5 imes IQR$ from the median were excluded from the figure.

feature is observed in some of the 4 PCR replicates, therefore these features were not included in the error analysis. Overall aggreement between the inferred and observed relative abundance was high for all individuals and bioinformatic pipelines (Fig. 7A). The error rate distribution was similarly consistent across pipelines, including long tails (Fig. 7B)

To assess the feature-level quantitative accuracy of the relative abundance values we compared the the relative abundance feature-level bias, (median error rate, Fig. 7C), and variance, ($RCOV = (IQR)/|median|$ Fig. 7D). Feature-level bias and variance metrics were compared across pipelines and individuals using a mixed effects model. Large bias and variance values were observed for all pipelines (Table 3). Features with large bias and variance metrics, defined as $1.5 \times IQR$ from the median, were excluded from the analysis to prevent outliers from biasing the comparison. Multiple comparisons test (Tukey) was used to test for significant differences in feature-level bias and variance between pipelines. A one-sided alternative hypothesis to determine which pipelines had a smaller, feature-level error rate. The Mothur, DADA2, and QIIME feature-level bias were all signficantly 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) (7C).
Large variance metric values were observed for all individuals and pipelines (Table 3). The feature-level variance was not significantly different between pipelines, Mothur = 0.83, QIIME = 0.71 and DADA2 = 1 (Fig. 7D). We evalutated whether poor feature-level relative abundance metrics can be attribtuted to specific taxonomic groups or phylogenetic clades (Supplemental Fig. **2017-11-16_feature-phyloSignal.pdf**). While a phylogenetic signal was detected for both the bias and variance metric, we were unable to identify specific taxonomic groups or phylogenetic clades that performed poorly in our assessement.

The agreement between the log-fold change estimates and expected values were individual specific and generally consistent across pipelines (Fig. 8A). The individual specific effect was attributed to the fact that unlike the relative abundance assessment the inferred θ values were not used to calculate the expected values. The inferred θ values were not used to calculate the expected values as we wanted to include all of the titrations and the θ estimates for the higher titrations were not monotomically decreasing and therefore resulted in unrealistic expected log fold-change values, e.g. negative log-fold changes for pre-exposure specific features. The log-fold change estimates and expected values were consistent across pipelines with one noteable exception. For E01JH0011 the Mothur log fold-change estimates were in better agreement wih the expected value compared to the other pipelines. As θ was not corrected for differences in the proportion of prokaryotic DNA between the unmixed pre- and post-exposure samples we are unable to say whether Mothur's performance was better than the other pipelines.

The log fold-change error distribution was consistent across pipelines (Fig. 8B). There was a long tail in the 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 the log-fold change error distribution for log-fold change estimates using different normalization methods (Supplemental Fig. **2017-11-15_norm_comp_logFC.pdf**). The error rate distributions, including the long tails, were consistent across normalization methods. Furthermore as the long tail was observed for the unclustered data as well, the log-fold change estimates contributing to the long tail are likely due to a bias associated with the molecular laboratory portion of the measurement process and not the bioinformatic pipelines. Based on exploratory analysis of the relationship between the log fold-change estimates and expected values for individual features indicated that the long tails were attributed to feature specific performance.

The $1 - slope$ and R^2 values for linear models of the estimated and expected log fold-change for individual features, all titration comparison, were used to characterize the feature-level log fold-change bias (Fig. 8C) and variance across pipelines (Fig. 8D). A bias metric of $1 - slope$ was used, where 0 is the desired value (i.e. log fold-change estimate = log fold-change expected), negative and positive values indicate the log-fold change was consistently under and over estimated, respectively. The linear model R^2 value was used to characterize the feature-level log fold-change variance as it indicates the consistency of the relationship between the log fold-change estimates and expected values are across titration comparisons. Similar to the relative abundance assessment we used a mixed-effects models to account for differences in individuals when comparing bias and variance metrics across pipelines. The log fold-change bias metric and variance metrics were not significantly different between pipelines (Bias: F = 0, 2.51, p = 0.99, 0.08, 8B, (Variance: F = 47.39, 0.23, p = 0, 0.8, Fig. 8C). Next we evaluated whether poor feature-level metrics could be attributed to

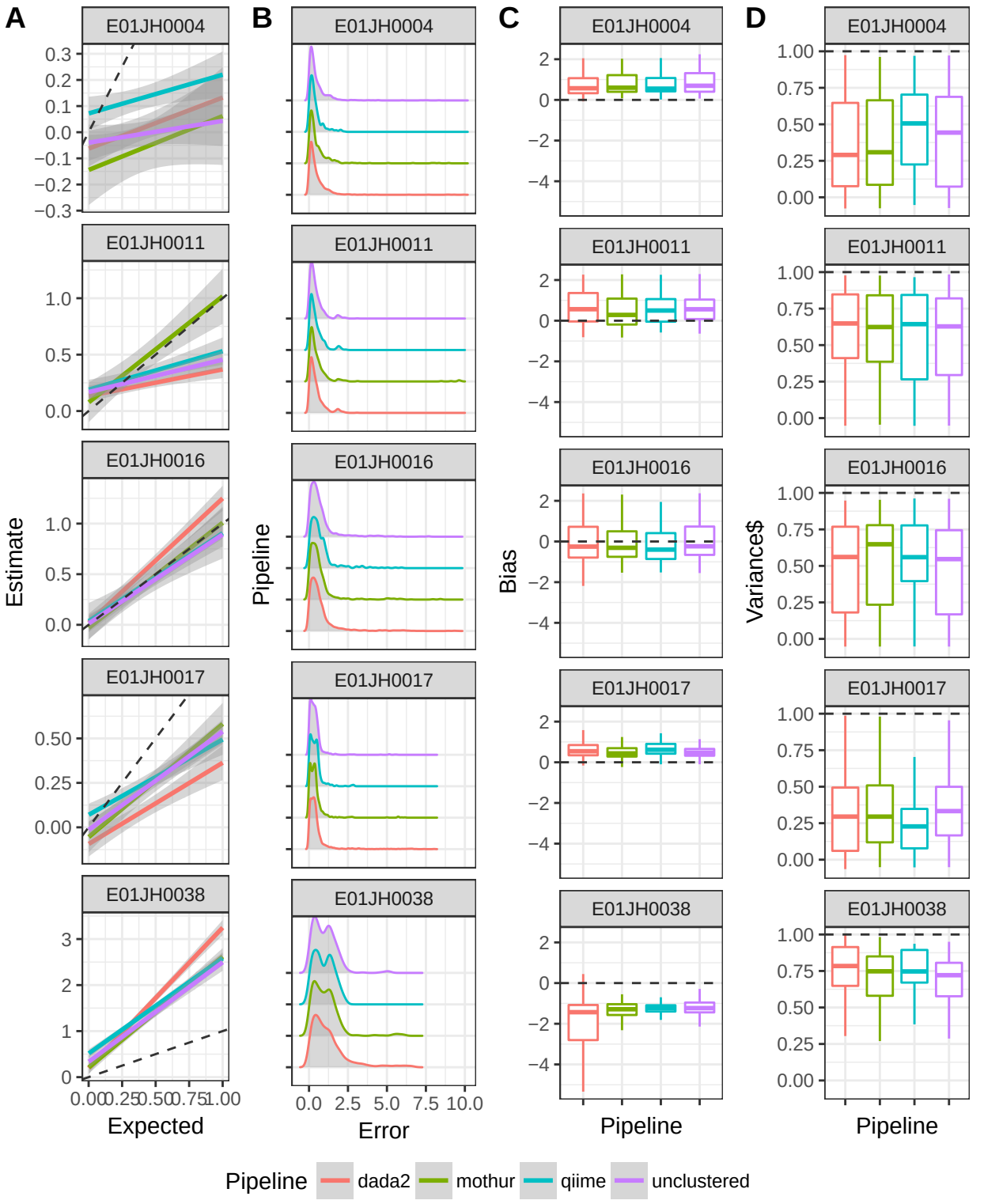

Figure 8: (A) Linear model or the relationship between log fold-change estimates and expeced values for pre-specific and pre-domiant features by pipeline and individual, line color indicates pipelines. Dashed grey line indicates expected 1-to-1 relationship between the estimated and expected log fold-change. (B) Log fold-change error (|exp-est|) distribtion by pipeline and individual. Distribution of feature-level log-fold change error bias (C) and variance (D) metrics by individual 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. To prevent extreme metric values from obscuring metric value visual comparisons, boxplot outliers, $1.5 imes IQR$ from the median were excluded from the figure.

specific clades for taxonomic groups (Supplemental Fig. **2017-11-16_feature-phyloSignal.pdf**). Similar to the relative abundance estimate, while a phylogenetic signal was detected for both the bias and variance metrics, we were unable to identify specific taxonomic groups or phylogenetic clades that performed poorly in our assessement.

4 Discussion

We generated a 16S metagenomic benchmarking dataset with the diversity, relative abundance dynamic range, and sequencing artifacts of a real dataset using mixtures of environmental samples We used the dataset to assess the feature presence/ absence, relative abundance, and log fold-change for count tables generated using four different bioinformatic pipelines

We assessed the performance of three bioinformatic pipelines, open-clustering, *de-novo* clustering, sequencing inference, and unclustered. Running these pipelines on our mixture dataset resulted in a range of total feature abundance and features per sample. Despite the wide range in number of features the pipelines all generated similarly sparse datasets. The sparsity was comparable to previously publised values, 1-3% (Paulson et al. 2013). As the dataset is highly redundant, 180 samples derived from 10 environmental samples lower sparsity was expected. The higher than expected sparsity can be attributed to false positive and false negative or true low abundance features. False positives are sequencing or PCR artifacts that are not appropriately filtered or assigned to an appropriate feature by the bioinformatic pipeline. The 16S region sequenced in the study is larger than the region the de-novo and open clustering pipelines were initially developed for and thus potentailly explaining the higher than expected sparsity (Kozich et al. 2013) **QIIME V4 REF**. The larger region has a smaller overlap between the forward and reverse reads as a result in our study the merging of the forward and reverse reads did not allow for the sequence error correction that occurs when their is greater overlap. However, both the *de-novo* and open-reference clustering pipelines have produced count tables with magnitudes higher than expected features in evaluation studies using defined mixtures of cell and DNA, mock communities **REF**. The sequence inference method produced a count table with significantly fewer features compared to the *de-novo* and open-reference clustering pipelines, yet had comparable sparsity. False negative provide a potential expalination for the higher than expected sparsity. A false negative occurs when a lower abundance sequence representing an organism within the sample is clustered with a higher abundance sequence inflating the dataset sparsity. The qualitative assessment results, specifically the high proportion of titration and unmixed sample specific features that could not be explained by sampling alone, indicates that the sequence inference method had a high false negative feature detection rate. While microbial abundance distributions are known to have long tails, it is likely that the observed sparsity is an artifact of the 16S measurement process based on results from previously mentioned mock community benchmarking studies.

As the qualitative assessment results were pipeline dependent the implications for 16S metagenomic studies vary by pipeline. For *de-novo* and open-reference clustering methods any conclusions made based on low abundance features require additional justification. Specifically, determining whether the feature is a measurement artifact or represents a member of the microbial community. This is especially relevant for studies characterizing the rare biosphere. A study exploring the microbial ecology of the Red-necked stint *Calidris ruficollis*, a migratory shorebird, used a hard filter for low abundance features, but also compared the results with and without the filter ensuring that the study conclusions were not biased by using the artibrary filter or including the low abundant, likely predominantly measurement artifacts, features (Risely et al. 2017). For 16S metagenomic studies using DADA2, missing low abundance features are more likely to impact presence/absence ecological diversity analysis. 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. At the same it is unlikely that the number of features in a sample accurately reflects the true richness of a sample though whether real differences in richness between samples are detectable when a dataset is processed using DADA2 is unknown.

The quantiative assessment results, both relative abundance and log fold-change estimates were individual specific. The individual specific results are a limitation in inferring the proportion of prokaryotic DNA in a titration from post-exposure samples, θ. We were able to use an assay targeting the 16S rRNA gene to detect

changes in the concentation of bacterial DNA across titration but we were unable to estimate the proportion of bacterial DNA in the unmixed samples using the qPCR data. Using the 16S sequencing data we inferred the proportion of bacterial DNA from the post-exposure sample in each titration. However, the uncertainty and accuracy of the inferrence method is not known resulting in an unaccounted for error source. A better method for estimating the proportion of bacterial DNA in the unmixed samples would increase the accuracy of the error metrics.

While the relative abundance bias metric was significantly different between pipelines overall, pipeline had minimal impact on the quantitative assement results when accounting for individual effects. However, large outliers were commonly observed. The outliers could not be attributed to the bioinformatic pipeliens as the outliers were observed for the unclustered datasets as well. Therefore the poor performance for outliers was attributed to the molecular laboratory side of the measurment process. Visual exploration of the results indicates a feature-specific effect. Mismatches in the primer binding regions have been shown to impact PCR efficiency, a potential cause for poor feature-specific performance **REF**. We evaluate the taxonomy and phylogenetic relationship for the relative abundance and log fold-change bias and variance metrics. While a phylogenetic signal was detected for all metrics poor feature-level performance we were unable to attributed to any specific taxonomic group or phylogenetic clade. We were unable to define a set of characteristics (quantitate, taxonomic, or phylogenetic) that can be used to identify poor performing features.

Based on the results of our quantitative assessment community level analysis are likey more accurate than feature-level analyses. Therefore results from individual analysis like differential abundance, which rely on log-fold change estimates, are more susceptible unknown biases and are potentially artifacts of the measurement process. As the outliers were consistently observed across individual, pipeline, normalization method, and log fold-change estimators, the molecular laboratory step in the measurement process is responsible for the measurement artifact. Additional work is needed to further characterize these outliers to increase confidence in the results of feature-level analyses.

5 Conclusions

This two-sample-titration dataset can be used to evaluate and characterize bioinformatic pipelines and clustering methods. The sequence dataset presented in this study can be processed with any 16S bioinformatic pipeline to generate a count table. Our quantitative and qualitative assessment can then be performed on the count table and the results compared to those obtained using the pipelines included in this study. Based on the results of our assessment of four bioinformatic pipelines the pipelines generate sets of features with different characteristics interms of total abundance, features per samples, and total features. The objective of any pipeline is to differentiate true biological sequences from artifacts of the measurement process. Users should consider whether a pipeline minimizes false positives (DADA2) or false negatives (Mothur) is more appropriate for their study objectives. Further more as the feature-specific quantitative assessment results could not be attributed to any feature quantitative, phylogenetic, or taxonomy, feature-level resutlts for any 16S metagenomic study should be interpreted with care. Addressing both of these issues requires advances in both the molecular biology and computational components of the measurement process.

---

6 References

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

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

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

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

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

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods*. Nature Publishing Group.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5). Nature Publishing Group:335–36.

DeSantis, Todd Z, Philip Hugenholtz, Neils Larsen, Mark Rojas, Eoin L Brodie, Keith Keller, Thomas Huber, Daniel Dalevi, Ping Hu, and Gary L Andersen. 2006. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with Arb." *Applied and Environmental Microbiology* 72 (7). Am Soc Microbiol:5069–72.

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

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

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

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

Goodrich, Julia K., Sara C. Di Rienzi, Angela C. Poole, Omry Koren, William A. Walters, J. Gregory Caporaso, Rob Knight, and Ruth E Ley. 2014. "Conducting a Microbiome Study." *Cell* 158 (2). Elsevier Inc.:250–62. https://doi.org/10.1016/j.cell.2014.06.037.

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

Kim, Dorothy, Casey E Hofstaedter, Chunyu Zhao, Lisa Mattei, Ceylan Tanes, Erik Clarke, Abigail Lauder, et al. 2017. "Optimizing Methods and Dodging Pitfalls in Microbiome Research." *Microbiome* 5 (1). BioMed Central:52.

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

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

McCarthy, Davis J., Chen, Yunshun, Smyth, and Gordon K. 2012. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10):–9.

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

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12). Nature Research:1200–1202.

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

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

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

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

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

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23). Am Soc Microbiol:7537–41.

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

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

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

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

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

---

7 Supplemental

8 Theta Inferrence

Table 4: Number of features used to estimate theta by biological replicate and pipeline.

pipe	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
dada2	90	90	144	136	130
mothur	114	104	178	149	177
qiime	145	146	106	155	204
unclustered	346	396	466	343	472

Table 5: Number of features by pipeline and individual used in the relative abundance error rate analysis.

pipe	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
dada2	116	103	155	162	119
mothur	141	129	206	205	184
qiime	226	208	132	231	280
unclustered	715	779	860	683	708

9 Relative Abundance

9.1 Feature Characterization

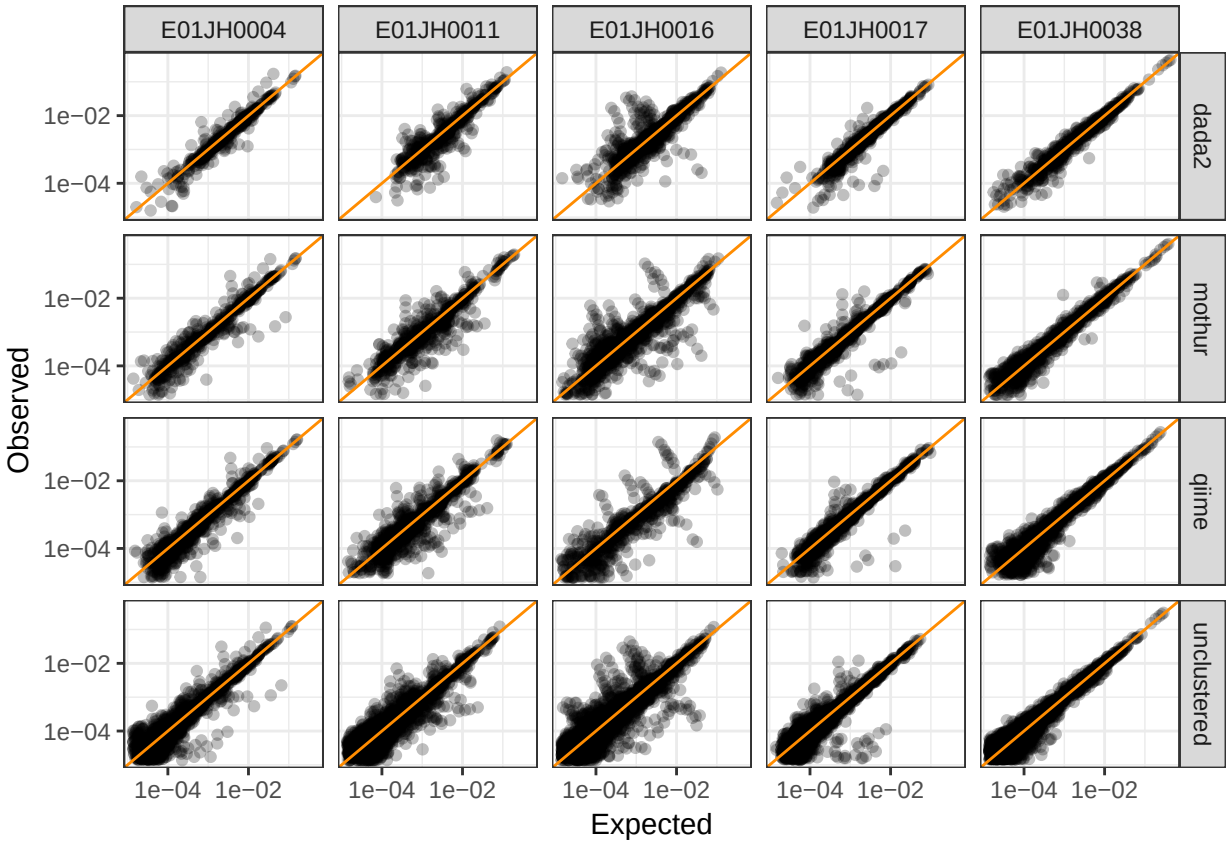

Figure 9: Relationship between expected and observed relative abundance. Dark orange line indicates the expected 1-to-1 relationship.

Table 6: Number of pre-specific and pre-dominant features by individual and pipeline

Individual	Type	dada2	mothur	qiime	unclustered
E01JH0004	dominant	7	11	8	14
E01JH0004	specific	47	11	10	32
E01JH0011	dominant	3	7	6	11
E01JH0011	specific	38	14	11	24
E01JH0016	dominant	4	5	0	7
E01JH0016	specific	84	44	16	65
E01JH0017	dominant	5	11	10	17
E01JH0017	specific	69	19	15	37
E01JH0038	dominant	13	17	11	14
E01JH0038	specific	31	8	3	5

Table 7: Number of pre-specific and pre-dominant features by individual and normalization method for Mothur

Individual	Type	CSS	RAW	RLE	TMM	TSS	UQ
E01JH0004	dominant	11	11	11	11	11	11
E01JH0004	specific	10	12	11	11	12	11
E01JH0011	dominant	8	8	7	7	8	7
E01JH0011	specific	17	17	14	14	18	14
E01JH0016	dominant	4	7	5	5	10	5
E01JH0016	specific	40	46	44	44	53	44
E01JH0017	dominant	9	13	11	11	14	11
E01JH0017	specific	17	21	19	19	25	19
E01JH0038	dominant	7	20	17	17	24	17
E01JH0038	specific	3	10	8	8	12	8

10 log Fold-Change

10.1 Estimator Comparison

10.2 Feature Characterization

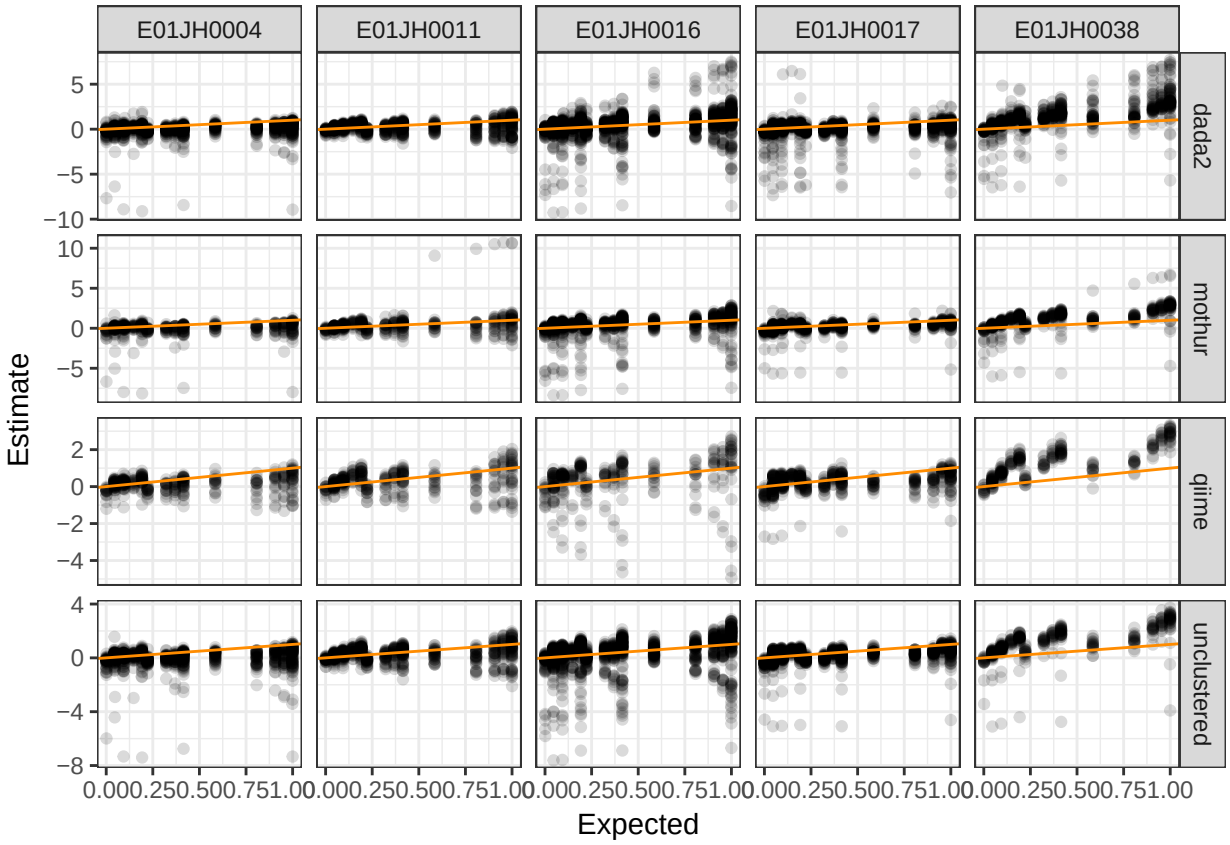

Figure 10: Relationship between log fold-change estimates and expected values. Orange line represents the expected 1-to-1 relationshion.

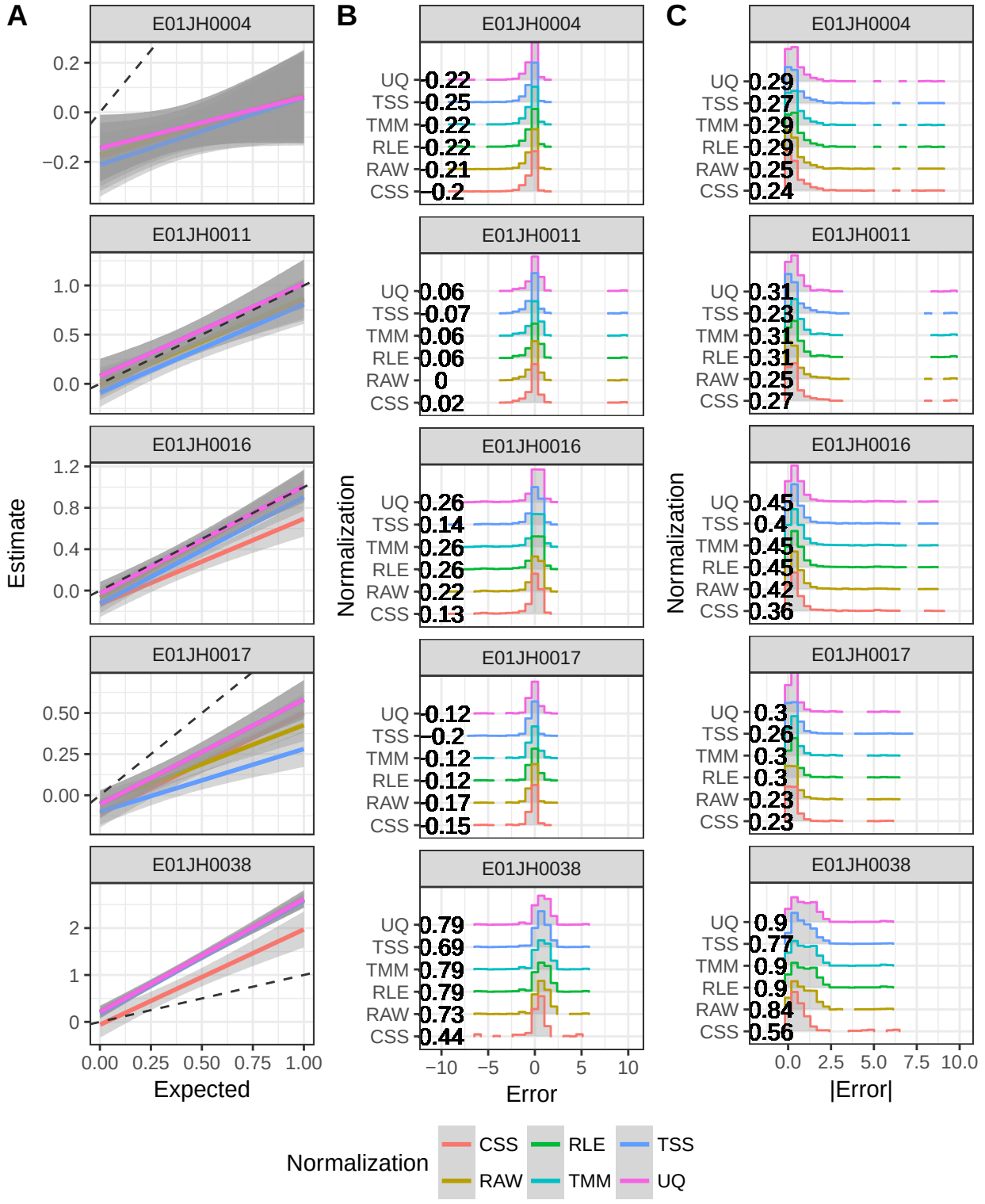

Figure 11: Impact of normalization methods on the agreement between log fold-change estimates and expected values for pre-specific and pre-dominant features. (A) Linear model relating the log fold-change estimates with the expected values by individual and normalization method. (B) Distribution of log fold-change (B) absolute error and (C) error by normalization method and individual.