Microbiome-Scale Mixture Use Demonstration

Nate Olson

2017-11-09

1 Introduction

Metagenomics, sequencing the DNA from a microibal community, 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, as the targeted approach allows for a more in-depth exploration of a microbial community taxonomic compositions compared to shotgun metagenomics where whole genomes are sequenced. 16S metagenomics is a complex measurement process comprised of multiple molecular laboratory and computation 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 laboratory (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 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 communities efforts for improving the measurement process.

A key step in the measurement process is clustering, the grouping of sequences into biologically relevant units, or operational taxonomic units (OTUs). There are a number of different clustering methods.
The two most commonly used clustering methods are *de-novo* clustering and open-reference clustering. *de-novo* clustering algorithms group sequences based on a defined similarity threshold (Westcott and Schloss 2015). Open-reference clustering matches sequences to a set of previously clustered reference sequences (*de-novo*) then perform *de-novo* clustering on the sequences in the dataset that do not match to sequences in the reference dataset with the desired similarity threshold (He et al. 2015). A third methods for clustering, sequence inference, uses statistical models or algorithms to differentiate true biological sequences within a dataset from sequencing errors (Callahan et al. 2016; Amir et al. 2017; Eren et al. 2015).

Further challenging the measurement process is the compositional nature of the 16S data, that is the proporition of an organism within a sample is being measured and not the absolute abundance (Tsilimigras and Fodor 2016). Sequencing data only provide information regarding the relative abundance of organisms within a samples to other organisms within the same sample. When comparing the relative abundance of an organism across samples you are comparing organismal abundance relative to the rest of the organisms within the sample. As a result an organism can have the same absolute abundance in two samples but due to differences in either the microbial community composition or for targeted assays such as 16S metagenomics differences in the proportion of human DNA in the sample.

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 all of which use mock communities, simulated data, or environmental samples. Mock communities consisting of mixtures of cells or DNA from individual organisms and simulated data have been previously used to evaluate different aspects of the measurement process (Bokulich et al. 2016). Mock communities 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. 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 are not necessarily optimized to handle the additional biases, noise, and diversity present in real samples. Data generated from environmental samples, which include the biase, 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 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)

- Application to 16S
  - We generated a data set using mixtures of extracted DNA from human stool samples for assessing the 16S metagenomic measurement process.
  - Processed the resulting dataset with three bioinformatic pipelines and performed a quantitative and qualitative assessment of the resulting count tables.
  - Results indicate that ....

2 Methods

2.1 Two-Sample Titration Design

Samples from a vaccine trial were selected for use in the study (Harro et al. 2011). 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 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 the 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 (**TODO** Table ERCC) (Baker et al. 2005) (NIST SRM SRM 2374). 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

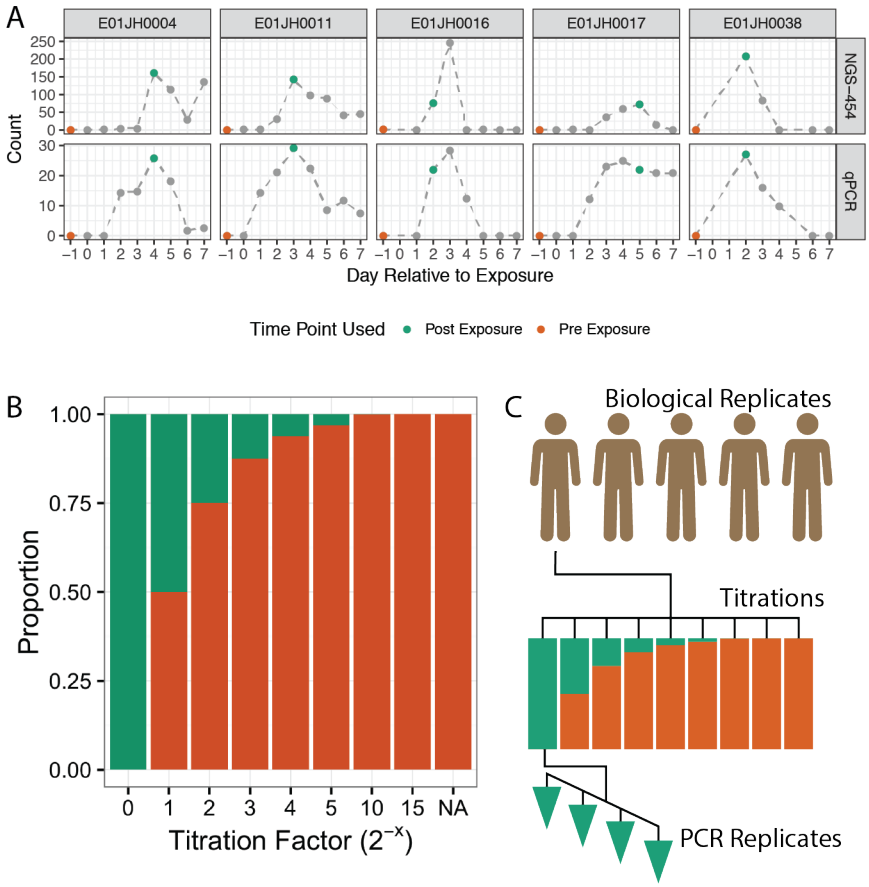

Figure 1: Sample selection and experimental design for two-sample titration 16S rRNA metagenomic sequencing assessment dataset. A) Pre- and post-exposure samples from five participants in a vaccine trial (Harro et al. 2011) were selected based on *Escherichia coli* abundance measured using qPCR and 454 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. Grey 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.

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.

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 (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'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGC-CTACGGGNGGCWGCAG - 3' and reverse primer 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGA-GACAGGACTACHVGGGTATCTAATCC - 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.

2.4 Sequence Processing

Sequence data was processed using four bioinformatic pipelines, Mothur (Schloss et al. 2009), QIIME (Caporaso et al. 2010), DADA2 (Callahan et al. 2016), and 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 used 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), the Usearch algorithm (Edgar 2010) and 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 and filtered using the `make.contigs` command 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. 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. Taxonomic classification of the unclustered sequences was performed using the same RDP classifier implmented in mothur used for the *de-novo* pipeline.

2.5 Data Analysis

- To generate summaries of QA metrics for the 384 datasets in the study (192 samples with forward and reverse reads) used the bioconductor `Rqc` package (REF) to calculate the quality metrics used in the following analysis.
- negative binomial model was used to calculate the average relative abundance across PCR replicates.
- log changes between titrations and pre- and post-exposure samples were calculated using EdgeR (Robinson, McCarthy, and Smyth 2010; McCarthy et al. 2012).

2.5.1 Theta Inference

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 features were used to infer θ. Features used were individual specific. To use a feature was 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 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. The error rate bias and variance 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 COV, $IRQ/median$, 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 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 or an artifact of the feature inference process.

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 Monte-Carlo simulation and 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 Monte-Carlo simulation and a Bayesian hypothesis test. 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. $P(\pi < \pi_{min_{exp}}) = P(\pi \geq \pi_{min_{exp}})$.

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

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

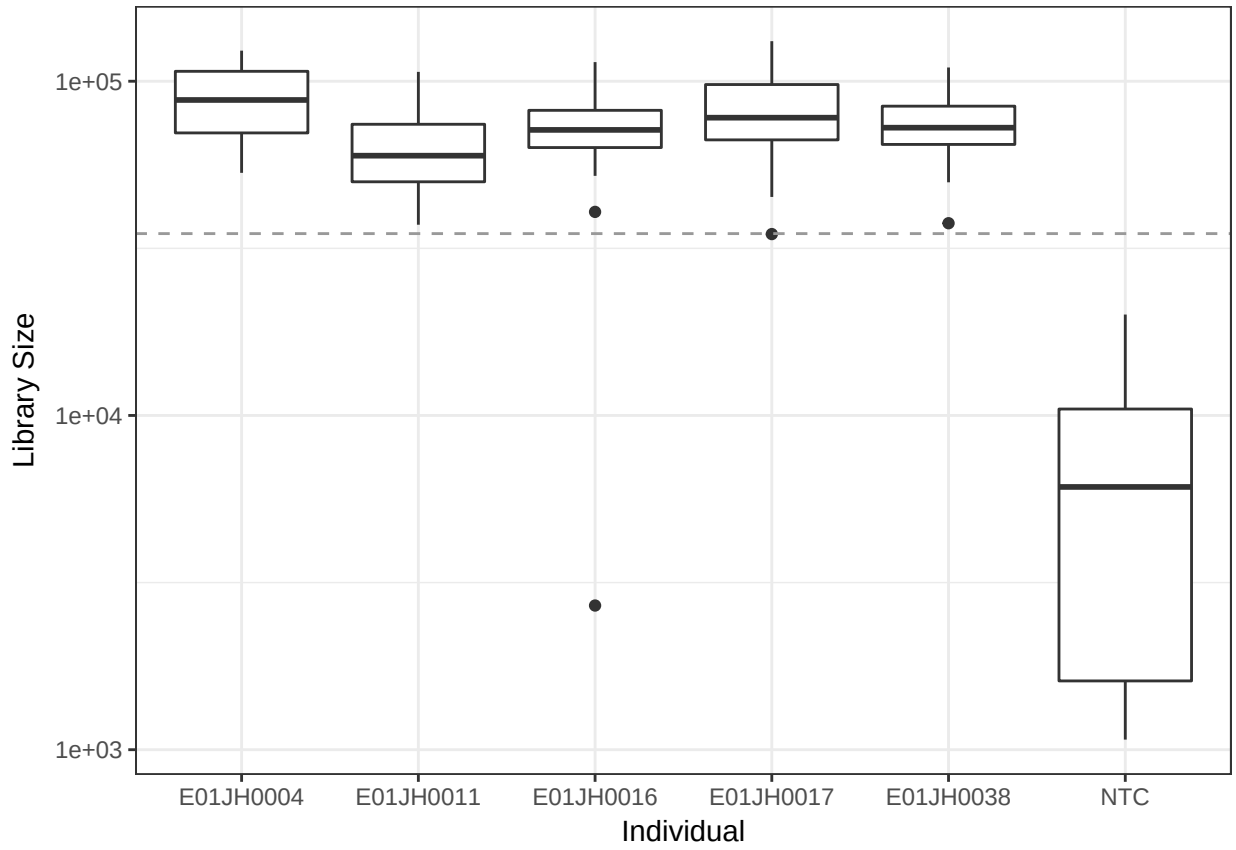

Figure 2: Distribution in the number of reads per barcoded sample (Library Size) by individual. Dashed horizontal line indicates 35,000 reads per barcoded sample.

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 2. 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 34889 to 131760 reads with a median library size of 73571. 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. 3).

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). Different pipelines 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 while the 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 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. 3). 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

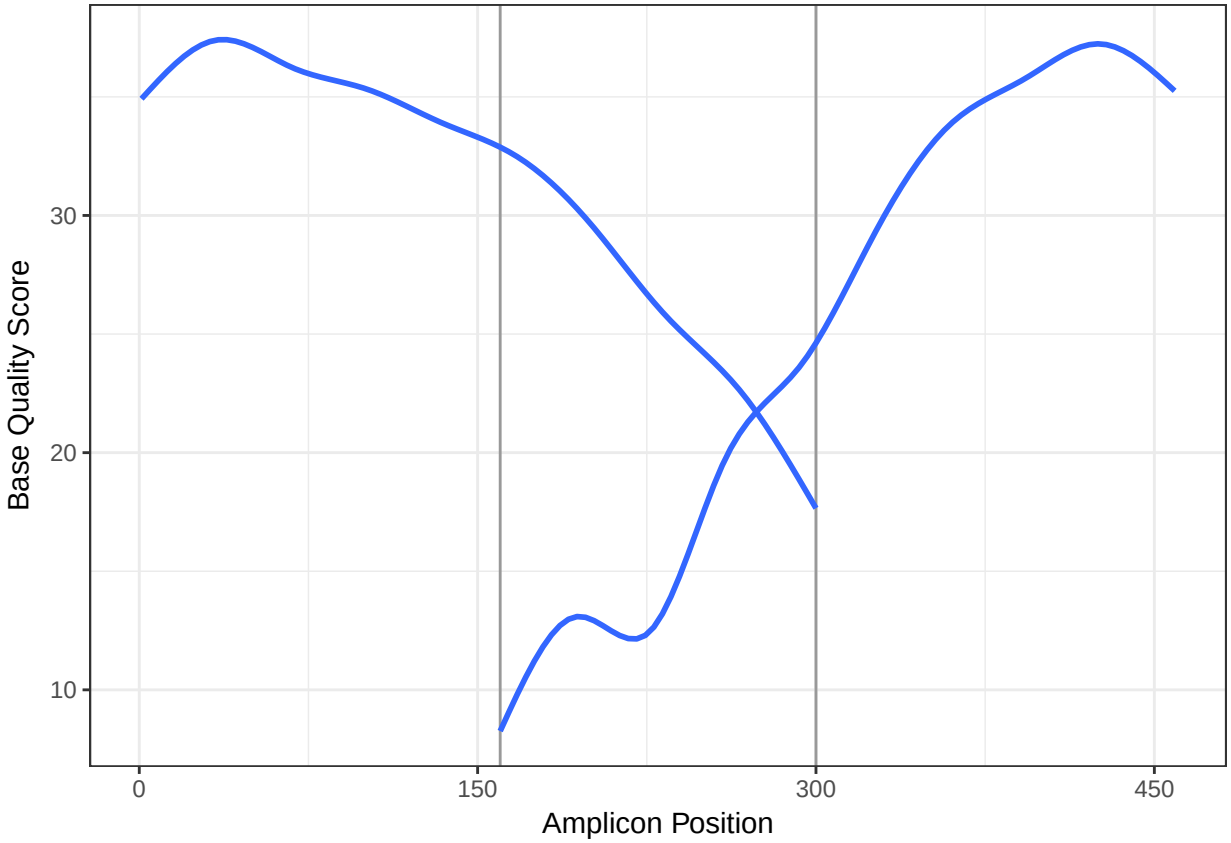

Figure 3: Smoothing spline of the base quality score by sequencing cycle. Vertical lines indicate approximate overlap region between forward and reverse reads.

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	Sample Coverage	Filter Rate
dada2	3144	0.93	68649 (1661-112058)	0.08 (-0.16-0.39)
mothur	38469	0.98	53775 (1265-87806)	0.28 (0.12-0.53)
qiime	11385	0.94	25254 (517-46897)	0.65 (0.49-0.96)

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.

The number of features per sample varied by bioinformatic pipeline (Fig. 4). The number of observed features by sample was more correlated between the QIIME and Mothur pipelines compared to the DADA2 pipeline (Fig. 4 A-C). Of the four samples with low numbers of features for the QIIME pipeline, only one of the samples had low number of observed features for the other two pipelines as well.

3.2 Titration Series Validation

In order to use information from the unmixed samples to obtain expected count values for the titrations we need to evaluate two assumptions about the mixed samples: 1. The samples were mixed volumetrically in a log_2 dilution series. 2. The unmixed pre- and post-exposure samples have the same proportion of bacterial 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 expectations. Total bacterial DNA in the unmixed samples was quantified using a qPCR assay targeting the 16S rRNA gene.

3.2.1 Spike-in qPCR results

The 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). Because of the log_2 two-sample-titration mixture design the expected slope is 1, for a doubling in template DNA every PCR cycle. Slope estimates less than 1 were attributed to the assay standard curve amplification factors being less than 2 (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 the 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) when 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. The 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 the accurate quantification of spike-in concentration. When taking into consideration 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.

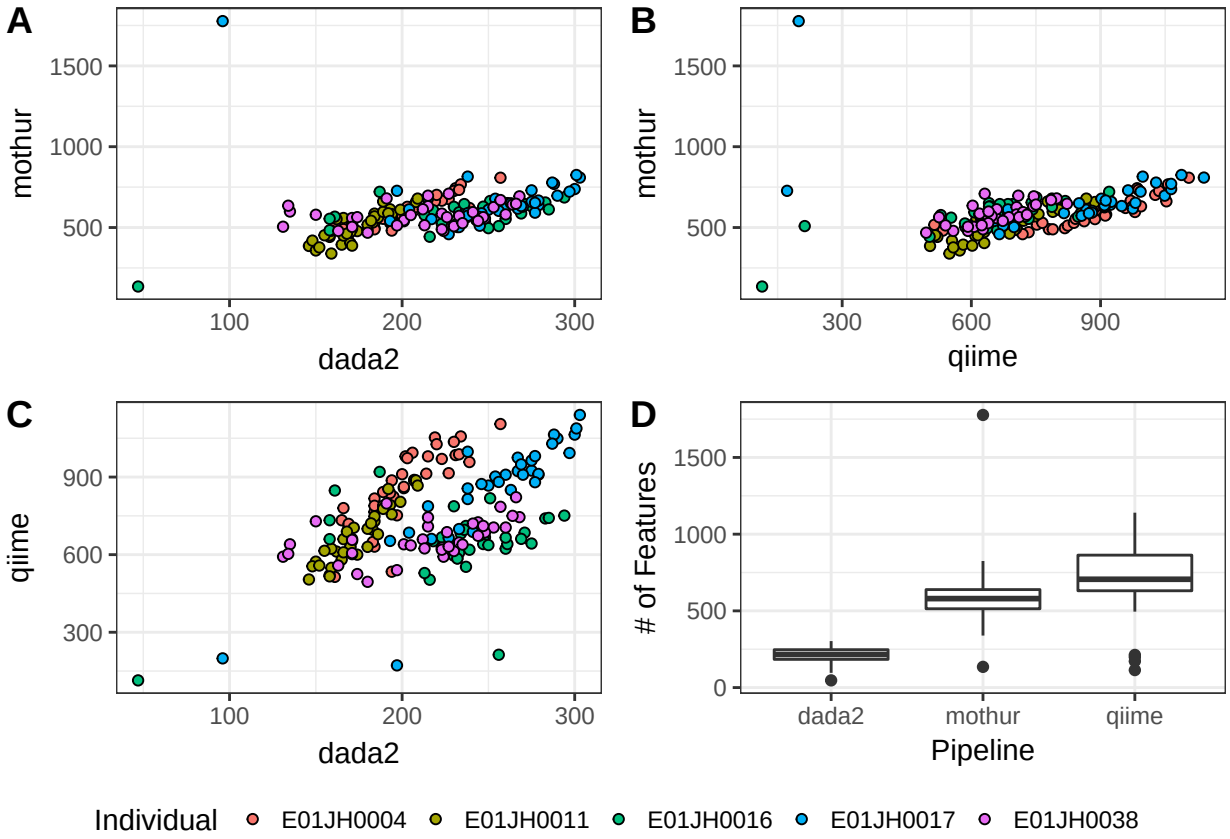

Figure 4: Comparison of the number of observed features per sample by bioinformatic pipeline. (A) Mothur v. DADA2, (B) Mothur v. QIIME, and (C) QIIME v. DADA2

Table 2: ERCC Spike-in qPCR summary statistics. R^2, Efficiency (E), and amplification factor (AF) for standard curves. R^2 and slope for titration qPCR results for the titration series.

Treatment	Individual	Std. R^2	E	AF	R^2	Slope
Post	E01JH0004	0.9996	86.19	1.86	0.98	0.92
	E01JH0011	0.9995	87.46	1.87	0.95	0.90
	E01JH0016	0.9991	87.33	1.87	0.95	0.84
	E01JH0017	0.9968	85.80	1.86	0.89	0.93
	E01JH0038	0.9984	86.69	1.87	0.95	0.94
Pre	E01JH0004	0.9972	84.36	1.84	0.53	-2.09
	E01JH0011	0.9999	87.93	1.88	0.52	-1.56
	E01JH0016	0.9990	84.22	1.84	0.60	-1.95
	E01JH0017	0.9979	89.78	1.90	0.32	-1.66
	E01JH0038	0.9994	84.30	1.84	0.21	-1.86

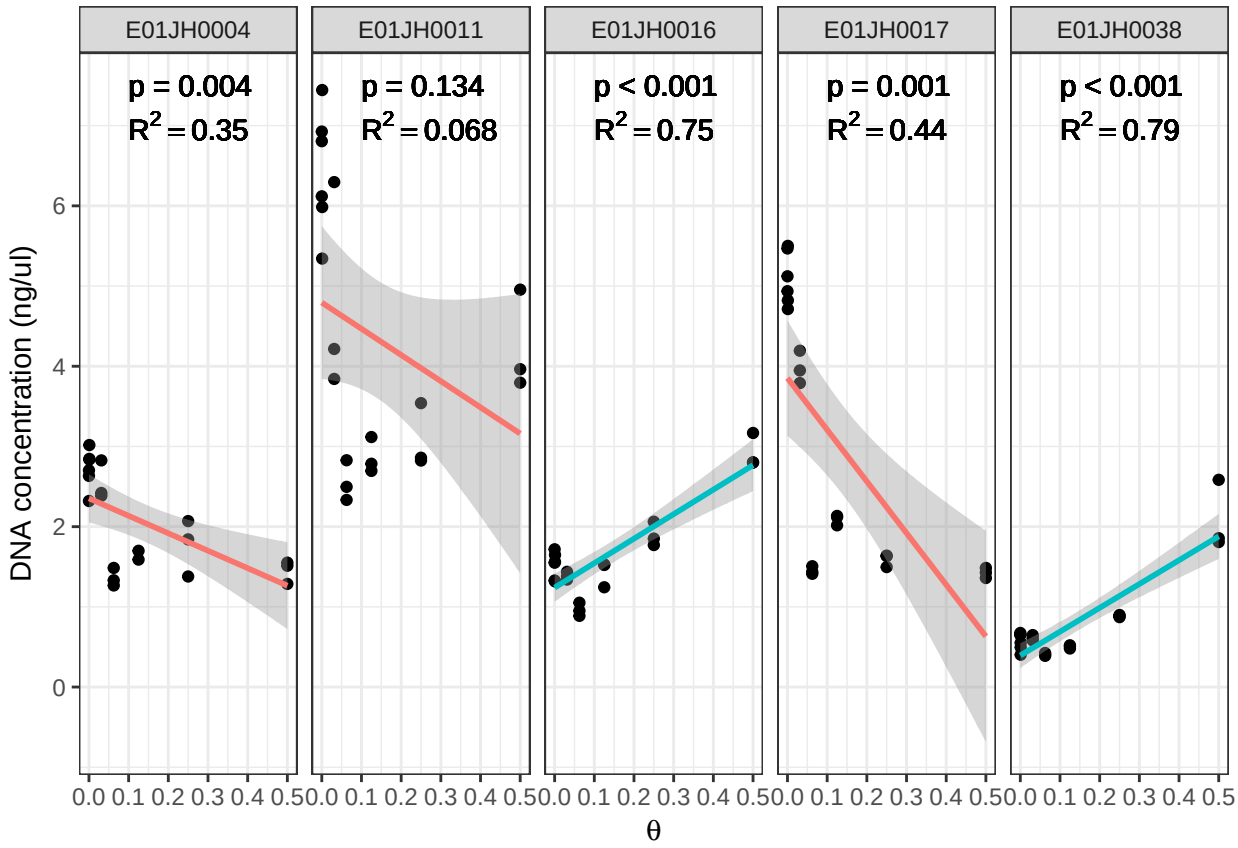

Figure 5: Prokaryotic DNA concentration (ng/ul) across titrations measured using a 16S rRNA qPCR assay. Separate linear models, [DNA]~θ were fit for each individual, R^2 and p-values were reported for each model. Red lines indicate negative slope estimates and blue lines positive slope estimates. p-value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

3.2.2 Bacterial DNA Concentration

Prokaryotic DNA concentration changes across titrations (Fig. 5) indicating the proportion of bacterial DNA from the unmixed pre- and post-exposure samples in a titration is not consistent with the mixture design. 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 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. 5). 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.

Table 3: 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

3.2.3 Theta Estimates

To account for differences in the proportion of prokaryotic DNA in the pre- and post-exposure (Fig. 5) we inferred θ, proportion of post-exposure sample prokaryotic DNA in a titration, using the 16S rRNA sequencing data (Fig. 6). Overall the relationship between the inferred and mixture design θ values were consistent across pipelines but not individual whereas the 95% CI varied by both individual and pipeline. For E01JH0004, 11, and 16 the inferred and mixture design θ values were in better agreement compared to E01JH0017 and E01JH0038. For E01JH0017 and E01JH00038 the inferred values were consistently less than and greater than the mixture design values, respectively. These results were consistent with the qPCR prokaryotic DNA concentration results with E01JH0017 having a significantly positive slope and E01JH0038 a significantly negative slope (Fig. 5).

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

There are a number of unmixed- and titration-specific features with a range of observed (titration-specific, Fig. 7A) and expected (unmix-specific, Fig. 7B) counts. There were unmixed-specific features with expected counts that could not be explained by sampling alone for all biological replicates and bioinformatic pipelines (Fig. 7C). However, the proportion of unmixed-specific features that could not be explained by sampling alone varied by bioinformatic pipeline with over half of the DADA2 unmixed-specific features could not be explained by sampling alone whereas QIIME had the lowest rate of unmixed-specific features that could not be explained by sampling alone. Consistent with the distribution of observed counts for titration-specific features more of the DADA2 features could not be explained by sampling alone compared to the other pipelines (Fig. 7D).

3.3.2 Quantitative Assessment

Overall aggreement between the inferred and observed relative abundance was high for all individuals and bioinformatic pipelines (Fig. 8A). 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, θ estimates for the unclustered pipeline were used to calculate the error rates for the other three pipelines.
Only features observed in all pre- and post-exposure PCR replicates and pre- and post-exposure specific features were included in the analysis (Table **??**). 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 feature is observed in some of the 4 PCR

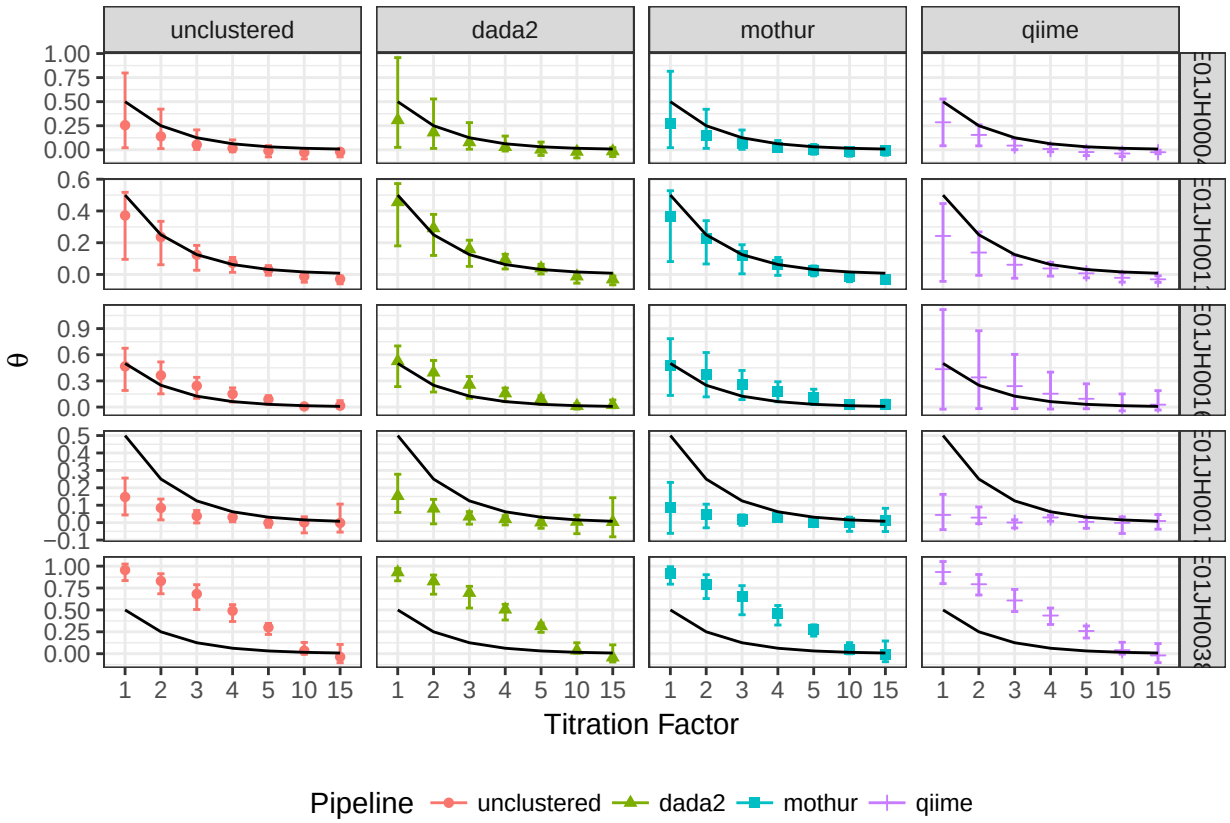

Figure 6: Theta estimates by titration, biological replicate, and bioinformatic pipeline. The points indicate mean estimate of 1000 bootstrap theta estimates and errorbars 95% confidence interval. The black line indicates the expected theta values. Theta estimates below the expected theta indicate that 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 4: Maximum feature-level error rate bias (median error rate) and variance (robust COV) by pipeline and individual.

Metric	Pipeline	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
Median	dada2	2.37	2.55	17.03	4.34	0.54
	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
RCOV	dada2	151.12	165.34	277.86	48.45	25.34
	mothur	57.99	29.56	43.74	340.26	1991.54
	qiime	470.91	409.21	575.73	144.63	153.58
	unclustered	3442.72	4638.81	313.60	109.16	311.81

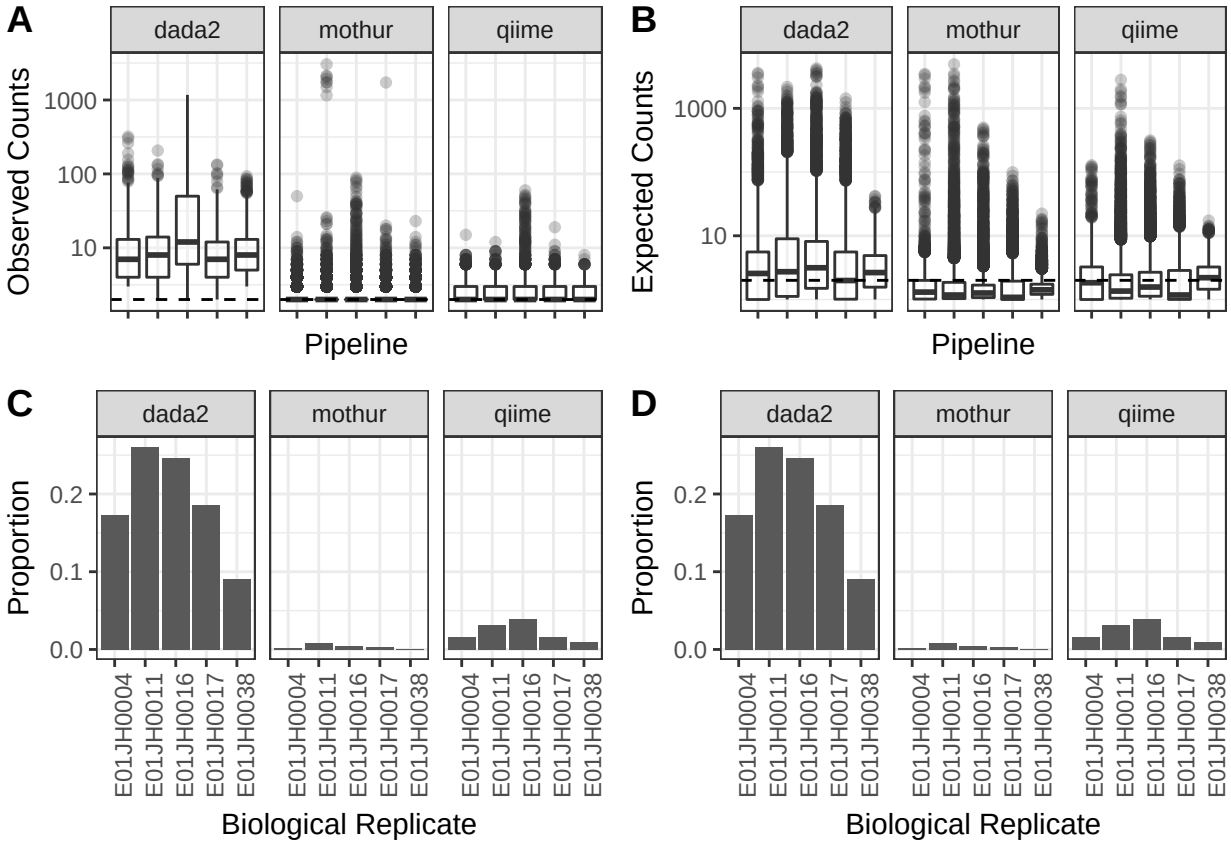

Figure 7: Distribution of (A) observed count values for titration-specific features and (B) expected count values for unmixed-specific features by pipeline and individual. The horizontal dashed line indicates a count value of 1. (C) Proportion of unmix-specific features and (D) titration-specific features with an adjusted p-value < 0.05 for the bayesian hypothesis test and binomial test respectively. We fail to accept the null hypothesis when the p-value < 0.05, indicating that for these features the discrepancy between the feature not being observed in the titration and present in the unmixed samples is not explained by sampling alone.

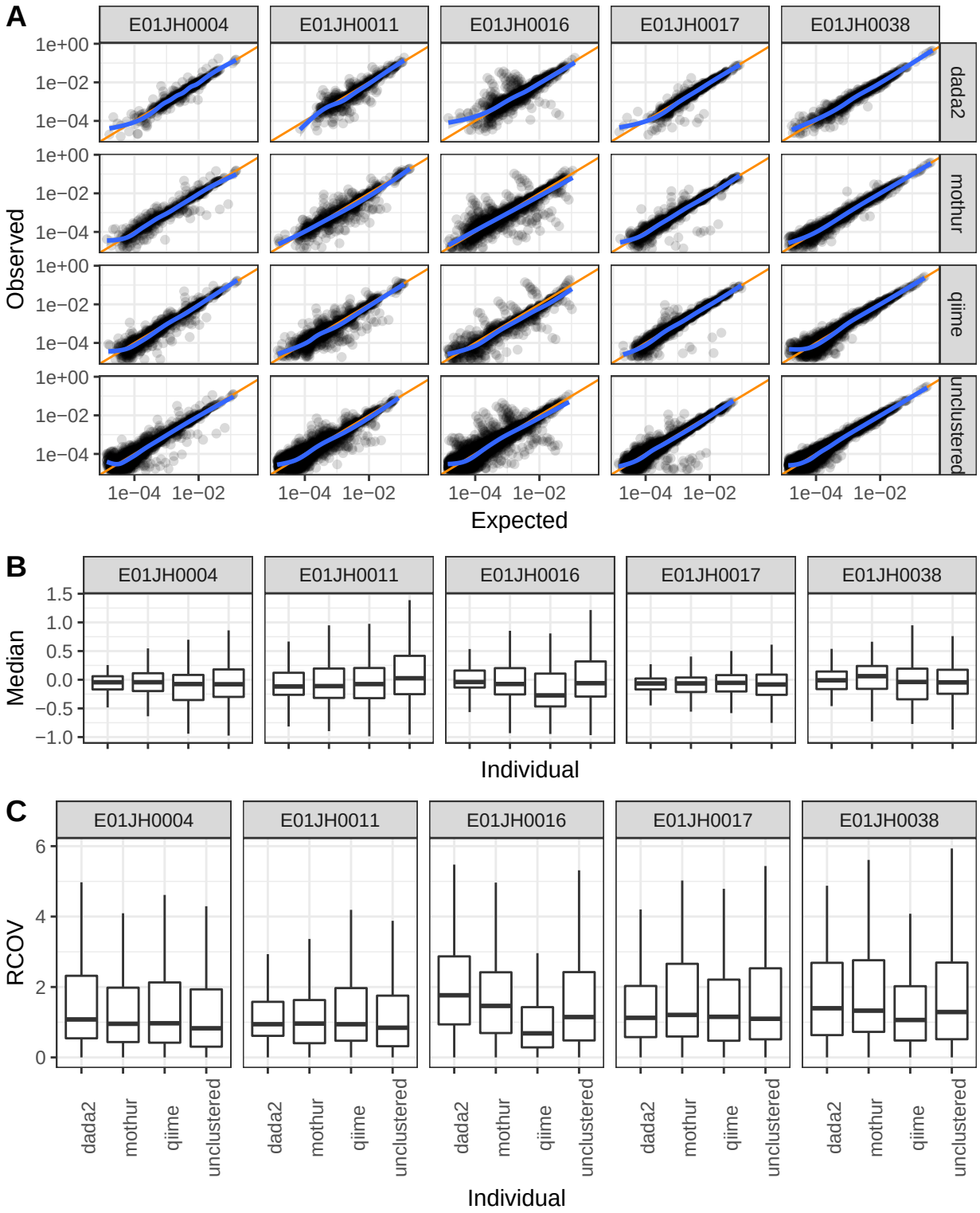

Figure 8: (A) Expected and observed count relationship. Orange line indicates expected 1-to-1 relationship. Blue line a smoothed regression line of the observed and expected value relationship. Distribution of feature-level relative abundance (B) median error rates and (C) robust coefficient of variation (RCOV) by individual and pipeline.

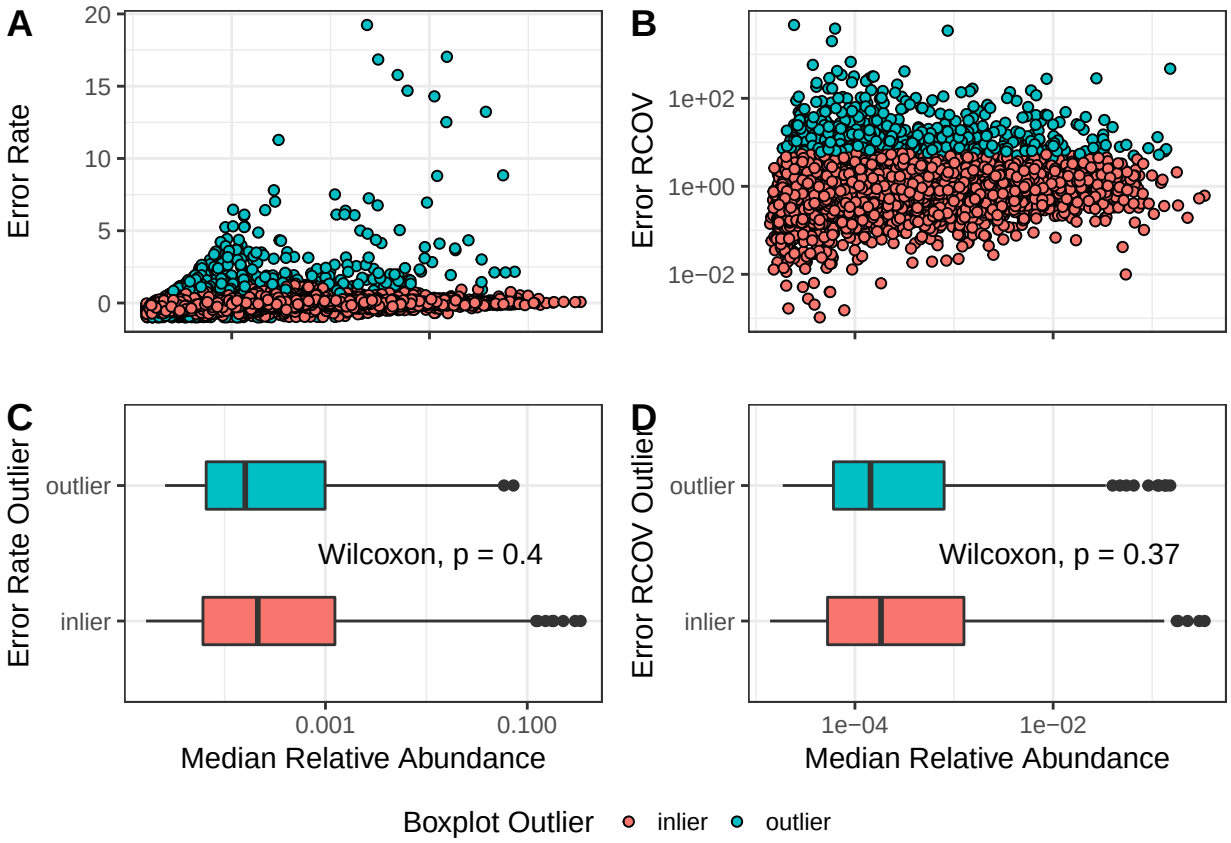

Figure 9: Feature-level median relative abundance is lower for error rate outlier features but not for RCOV outlier features. (A) Relationship between median relative abundance to the feature-level (A) median error rate and (B) error RCOV. Boxplots summarizing the feature-level median relative abundance between (C) error rate and (D) RCOV outlier and non-outlier features for features. Only features with median error rates greater than 0 included in the error rate boxplot.

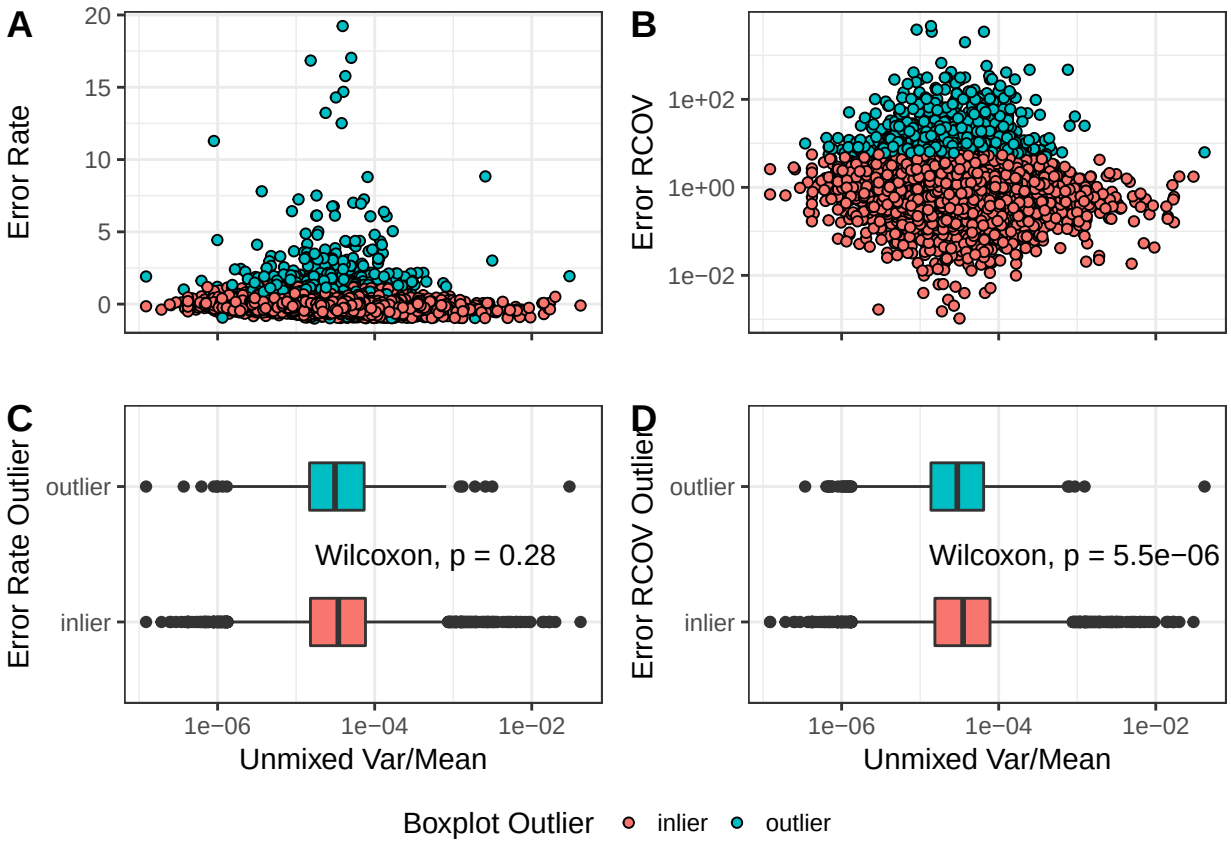

Figure 10: The variance-mean relationship for the unmixed sample relative abundance values does not explain outlier median and RCOV features. (A) Relationship between unmixed sample relative abundance variance/mean to the feature-level (A) median error rate and (B) error RCOV. Boxplots summarizing the unmixed sample relative abundance variance/mean between (C) error rate and (D) RCOV outlier and non-outlier features for features. Only features with median error rates greater than 0 included in the error rate boxplot.

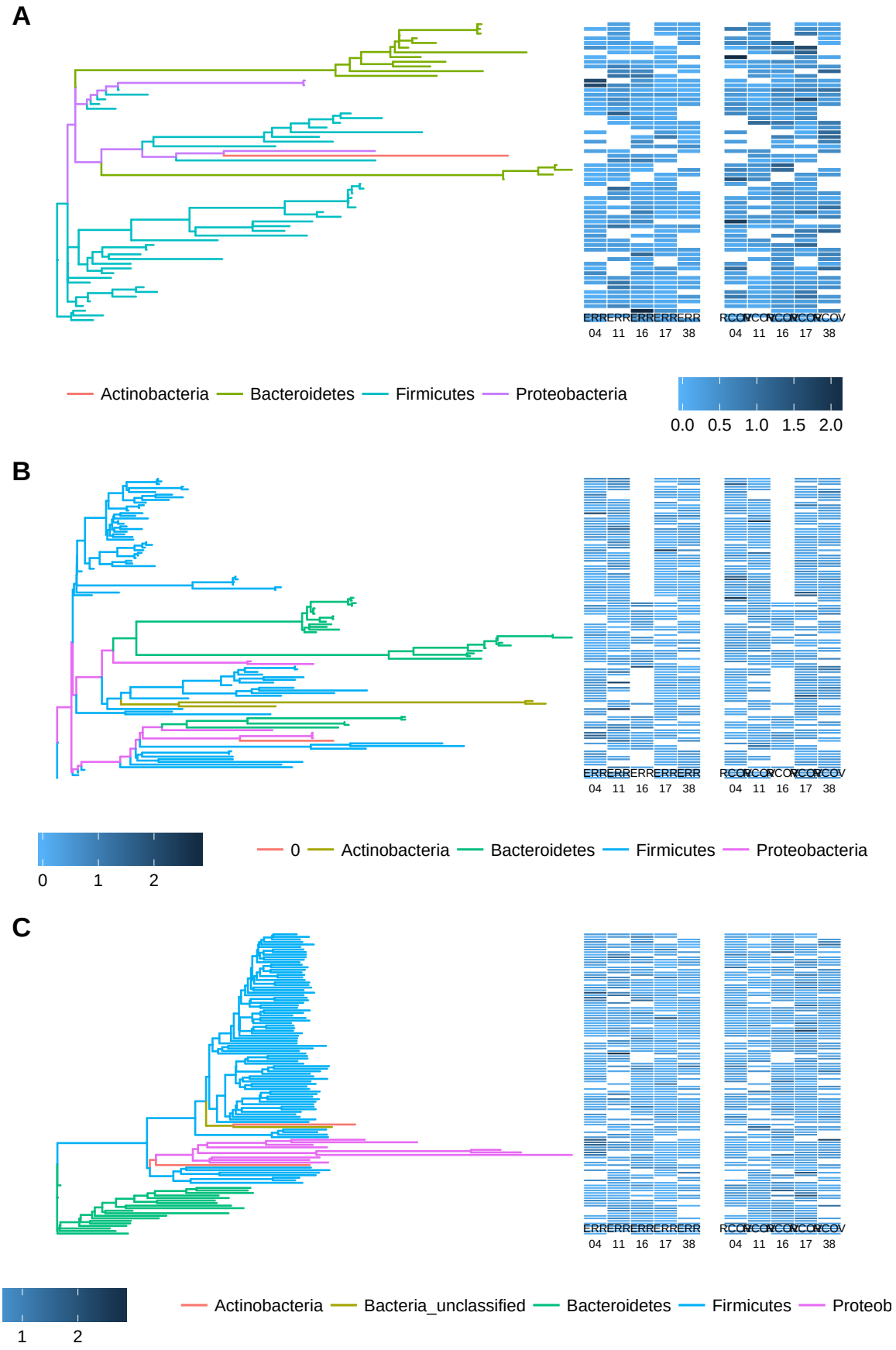

Figure 11: Phylogenetic analysis of feature-level relative abundance error metrics, as heatmaps across individuals. Subplots for individual pipelines (A) DADA2, (B) QIIME, (C) Mothur. Feature phylum assignment is indicated by branch color.

Table 5: 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

replicates, therefore these features were not included in the error analysis. The deviation from the expected value on the low end varies by biological replicate and pipeline.

Next we evaluated the feature-level quantitative accuracy of the relative abundance values by comparing the distribution of the median error and robust coefficient of variation ($RCOV = (IQR)/|median|$)for the relative abundance error rate (Fig. 8). Feature-level median error rates and RCOV were compared across pipelines and individuals using a mixed effects model. Large error rates and RCOV values were observed for all pipelines (Table 4). Features with large error rates, defined as $1.5 \times IQR$ from the median, were excluded from the analysis to prevent outliers from biasing the comparison. The mean feature-level error rate by pipeline was negative for all three pipelines (DADA2 -0.05, Mothur -0.06, and QIIME -0.11). Multiple comparisons test (Tukey) was used to test for significant differences in feature-level error between pipelines. A one-sided alternative hypothesis to determine which pipelines had a smaller, closer to zero, feature-level error rate. The Mothur and DADA2 mean feature-level error rates were closer to zero and signficantly different from the QIIME pipeline, (Mothur v. QIIME t = -3.7, p = 0; DADA2 v. QIIME t = -3.98, p = 0). Though the Mothur and DADA2 mean feature-level error rates were not significantly different from eachother, (t = -0.55, p = 0.57). Unlike feature-level error rates, large RCOV was observed for all individuals and pipelines (Table 4). Outlier values were also excluded from the RCOV analysis. The feature-level RCOV was not significantly different between pipelines, Mothur = 1.31, QIIME = 1.08 and DADA2 = 1 (Fig. 8C).

In an attempt to identify feature characteristics that could be attributed to poor performance the feature-level error-rate and RCOV were compared to the unmixed sample relative abundance and relative abundance across the titrations. Additionally, feature-level relative abundance error metrics were compared across individuals to try and identify any relationship between poor performance and phylogeny. Outlier feature-level error-rates had lower median relative abundance, but not RCOV (Fig. 9). The median relative abundance was significantly lower for features identified as outliers based on the feature-level error rate, but only when considering features with positive error rates. This is likely due to the more extreme error rates all being positive. For the RCOV the feature-level median relative abundance values were not significantly different between the outlier and non-outlier features.
The error rate is dependent on the accuracy of the relative abundance estimates for the unmixed pre- and post-exposure samples. The feature-level median error-rate and RCOV was compared to the the unmixed sample variance/mean relative abundance to determine if extreme error-rate and RCOV values could be attributed to variability in relative abundance between PCR replicates for the unmixed samples. The variance/mean for the unmixed samples was lower for the outliers compared to the non-outliers for both the feature-level error rate and RCOV. Investigation of relationship between phylogeny and feature-level relative abundance error metrics (Fig. 11). Only features included in the relative abundance error analysis for at least three of the five individuals were included in the figure. No clear relationship between feature-level error rate or RCOV and phylogeny of representative sequences.

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 log fold-change bias and variance across pipelines. A error 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

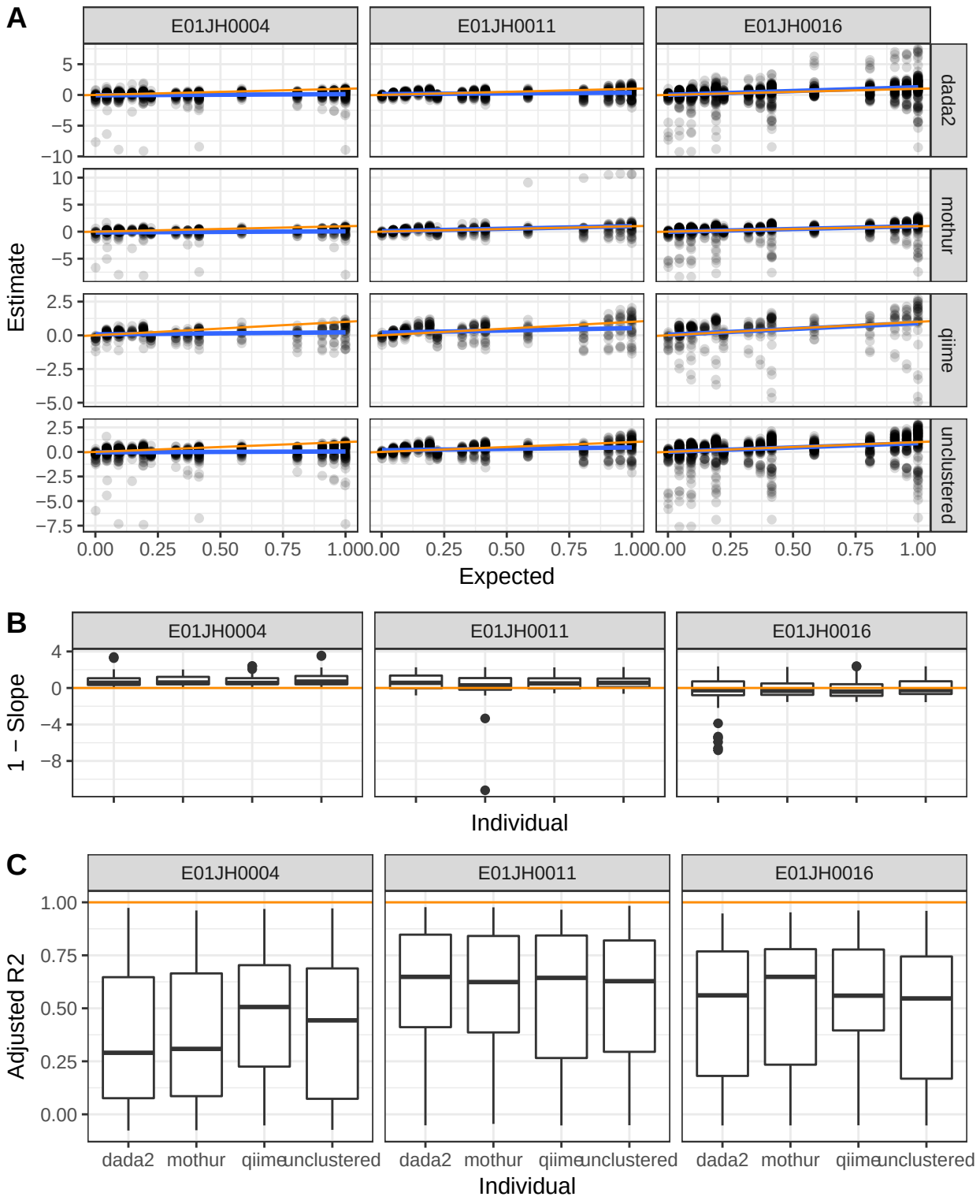

Figure 12: Relationship between the observed and expected logFC for pre-specific and pre-domiant features by pipeline and individual for all titration pair comparisons. Orange line indicates expected 1-to-1 relationship between the estimated and expected log fold-change. Blue line is a linear model was fit to the data and grey area is the models uncertainty estimate. Distribution of feature-level log-fold change error by individual and pipeline as the (B) 1 - slope and (C) R^2 for a linear model fit to the estimated and expected log fold-change values

fold-change variance as it indicates how consistent 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 take into account differences in individuals when comparing the log fold-change error rates between pipeline. The log fold-change bias metric was not significantly different between pipelines (F = 2.03, 0.36, p = 0.15, 0.7, DADA2 0.39, Mothur 0.34, QIIME 0.46, 12B). The log fold-change variance metric was not significantly different between pipelines (F = 80.19, 0.71, p = 0, 0.49, DADA2 0.47, Mothur 0.49, QIIME 0.51, Fig. 12C).

While some of the poor agreement can be attributed to low abundance (noisy data), or smaller differences in the log fold-change between pre- and post-exposure samples. An additional mixed-effects model was used to determine feature characteristics that are correlated with logFC error rate. Increased estimated logFC and logCPM were significantly related to lower error rates. Analysis of the log fold-change error results indicated a feature specific performance effect.

4 Discussion

By using mixtures of environmental samples we were able to generate a 16S metagenomic benchmarking dataset with the diversity, relative abundance dynamic range, and sequencing artifacts of a real dataset. 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 There were two primary limitations of the study that were a product of the experimental design. Only features that were differentially abundant between the pre- and post-exposure were used in the assessment. Using samples from the vaccine trial provided a specific features, *E. coli* that could be used during method development. However, only a limited number of features were differentially abundant between the pre- and post-exposure samples resulting in a smaller set of features that could be used in our assessment. Generating mixtures of samples with less similarity would increase the number of features used in the assessment. Additionally, using samples from other environments would increase the taxonomic diversity of features used in the assessment and potentially allowing for a more rigorus evaluation of the relationship between the assessment metrics and phylogeny. The second limitation of the experimental design was the difference in the proportion of bacterial DNA between the pre- and post-exposure samples. 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.

Evaluated the performance of four different bioinformatic pipelines, open-clustering, de-novo clustering, sequencing inference, and unclustered. Running these pipelines on the same 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 datasets with similar levels of sparsity. As the dataset is highly redundant, 180 samples derived from 10 envirnomental samples lower sparsity was expected. The qualitative assessment results indicate that the sequence inference method had a high rate of false negative features. This high false negative rate likey resulted in higher sparsity. For the other pipelines the high sparsity is attributed to features that were artifacts of the sequencing process, false positives. The high rate of false positive features have been observed in benchmarking studies using mock communities. The 16S region sequenced in the study is larger than the region the de-novo and open clustering pipelines were initially developed for. 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.

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, how do you know 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 **SOME BIRD** used a hard filter for low abundance features, but also compared the results with and without the filter ensuring that any conclusions were not biases by using the artibrary filter or including the low abundance features that are likely predominantly measurement artifacts **REF**. For 16S metagenomic studies using DADA2, missing low abundance features are more likely to impact presence/absence diversity analysis. Though a user can be more confident that an observed feature represents a member of the microbial community and not a measurement artifact. It is unlikely that the number of features in a sample accurately reflects the true richness of a sample though how well the results datasets are able to detect real differences in richness between samples 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. Pipeline had minimal effect on the quantiative results when accounting for individual effects. However, large outliers were commonly observed. Visual exploration of the results indicates a feature-specific effect. Experimental factors such as feature-level relative abundance of the titrations or unmixed samples could not account for the outliers. Additionally, outliers were not restricted to specific taxonomic groups. What dependency means for 16S gene surveys, when does bioinformatic pipeline matter and when does it not matter? Unable to define a set of characteristics that define can be used to identify poor performing features. In general low abundance features are noisier but high abundance features are not necessarily accurate?

Outliers as a product of artificial merging or splitting of features

Compositionality nature of 16S metagenomic data is well known within the field. The proportion of prokaryotic DNA in a sample is rarely taken into consideration in 16S studies. Our error metrics decrease when using our inferred θ values to correct for difference in prokaryotic DNA proportions. The extent to which the impact of differences in the proportion of prokaryotic DNA between sets of samples compared in a 16S studies is not well known. However, our results indicate that accounting for differences in proportion of prokaryotic DNA can improve the accuracy of the relative abundance comparisons.

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 numbers of features. The objective of any pipeline is to differentiate true biological sequences in a dataset from artifacts of the measurement process. Users should consider whether a pipeline is minimizes false positives (DADA2) or false negatives (Mothur). Continued effort to improve the 16S metagenomic measurement process to better understand and account for differences in the proportion of prokaryotic DNA in the samples and sequence artifacts. Addressing both of these issues requires advances in both the molecular biology and computational components of the measurement process.

6 Session information

6.1 Git repo commit information

The current git commit of this file is 719e5bae295bdac12d65d8038d7bb1867d52c4e7, which is on the master branch and was made by Nate Olson on 2017-11-09 11:14:50. The current commit message is fleshed out discussion. The repository is online at https://github.com/nate-d-olson/mgtst-pub

6.2 Platform Information

```
##  setting  value
##  version  R version 3.4.2 (2017-09-28)
##  system   x86_64, darwin15.6.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-11-09
```

6.3 Package Versions

package	version	date	source
ape	5.0	2017-10-30	CRAN (R 3.4.2)
base	3.4.2	2017-10-04	local
bindrcpp	0.2	2017-06-17	CRAN (R 3.4.0)
Biobase	2.36.2	2017-05-04	Bioconductor
BiocGenerics	0.22.1	2017-10-07	Bioconductor
BiocParallel	1.10.1	2017-05-03	Bioconductor
Biostrings	2.44.2	2017-07-21	Bioconductor
broom	0.4.2	2017-02-13	CRAN (R 3.4.0)
datasets	3.4.2	2017-10-04	local
DelayedArray	0.2.7	2017-06-03	Bioconductor
dplyr	0.7.4	2017-09-28	CRAN (R 3.4.2)
forcats	0.2.0	2017-01-23	CRAN (R 3.4.0)
foreach	1.4.3	2015-10-13	CRAN (R 3.4.0)
GenomeInfoDb	1.12.3	2017-10-05	Bioconductor
GenomicAlignments	1.12.2	2017-08-19	Bioconductor
GenomicRanges	1.28.6	2017-10-04	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.4.0)
ggpubr	0.1.5	2017-08-22	CRAN (R 3.4.1)
ggtree	1.8.2	2017-08-19	Bioconductor
git2r	0.19.0	2017-07-19	CRAN (R 3.4.1)
glmnet	2.0-13	2017-09-22	CRAN (R 3.4.2)
graphics	3.4.2	2017-10-04	local
grDevices	3.4.2	2017-10-04	local
IRanges	2.10.5	2017-10-08	Bioconductor
kableExtra	0.6.1	2017-11-01	CRAN (R 3.4.2)
knitr	1.17	2017-08-10	CRAN (R 3.4.1)
limma	3.32.10	2017-10-13	Bioconductor

magrittr	1.5	2014-11-22	CRAN (R 3.4.0)
MASS	7.3-47	2017-02-26	CRAN (R 3.4.0)
Matrix	1.2-11	2017-08-16	CRAN (R 3.4.1)
matrixStats	0.52.2	2017-04-14	CRAN (R 3.4.0)
metagenomeSeq	1.18.0	2017-04-25	Bioconductor
methods	3.4.2	2017-10-04	local
modelr	0.1.1	2017-07-24	CRAN (R 3.4.1)
multcomp	1.4-7	2017-09-07	CRAN (R 3.4.1)
mvtnorm	1.0-6	2017-03-02	CRAN (R 3.4.0)
nlme	3.1-131	2017-02-06	CRAN (R 3.4.0)
parallel	3.4.2	2017-10-04	local
ProjectTemplate	0.8	2017-08-09	CRAN (R 3.4.1)
purrr	0.2.4	2017-10-18	CRAN (R 3.4.2)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.4.0)
readr	1.1.1	2017-05-16	CRAN (R 3.4.0)
readxl	1.0.0	2017-04-18	CRAN (R 3.4.0)
Rqc	1.10.2	2017-07-13	Bioconductor
Rsamtools	1.28.0	2017-04-25	Bioconductor
S4Vectors	0.14.7	2017-10-08	Bioconductor
ShortRead	1.34.2	2017-10-08	Bioconductor
stats	3.4.2	2017-10-04	local
stats4	3.4.2	2017-10-04	local
stringr	1.2.0	2017-02-18	CRAN (R 3.4.0)
SummarizedExperiment	1.6.5	2017-09-29	Bioconductor
survival	2.41-3	2017-04-04	CRAN (R 3.4.0)
TH.data	1.0-8	2017-01-23	CRAN (R 3.4.0)
tibble	1.3.4	2017-08-22	CRAN (R 3.4.1)
tidyr	0.7.2	2017-10-16	CRAN (R 3.4.2)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.4.0)
treeio	1.0.2	2017-05-01	Bioconductor
utils	3.4.2	2017-10-04	local
XVector	0.16.0	2017-04-25	Bioconductor

References

Amir, Amnon, Daniel McDonald, Jose A Navas-Molina, Evguenia Kopylova, James T Morton, Zhenjiang Zech Xu, Eric P Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." *mSystems* 2 (2). Am Soc Microbiol: e00191–16.

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. doi: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.

Eren, A Murat, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin. 2015. "Minimum Entropy Decomposition: Unsupervised Oligotyping for Sensitive Partitioning of High-Throughput Marker Gene Sequences." *The ISME Journal* 9 (4). Nature Publishing Group: 968–79.

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. doi: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.

He, Yan, J Gregory Caporaso, Xiao-Tao Jiang, Hua-Fang Sheng, Susan M Huse, Jai Ram Rideout, Robert C Edgar, et al. 2015. "Stability of Operational Taxonomic Units: An Important but Neglected Property for Analyzing Microbial Diversity." *Microbiome* 3 (1). BioMed Central: 20.

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.

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.

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.

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.

Tsilimigras, Matthew CB, and Anthony A Fodor. 2016. "Compositional Data Analysis of the Microbiome: Fundamentals, Tools, and Challenges." *Annals of Epidemiology* 26 (5). Elsevier: 330–35.

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. 2015. "De Novo Clustering Methods Outperform Reference-Based Methods for Assigning 16S rRNA Gene Sequences to Operational Taxonomic Units." *PeerJ* 3. PeerJ Inc.: e1487.

———. 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.