Microbiome-Scale Mixture Use Demonstration

Nate Olson

2017-10-16

1 Introduction

Metagenomics, sequencing the DNA from a microibal community, has greater advanced our understanding of the microbial world. Targeted sequencing of the 16S rRNA gene, 16S metagenomics, is a commonly used method to sequence a microbial community as the targeted approach allows for a more in-depth exploration of a microbial communities taxonomic compositions compared to shotgun metagenomics where whole genomes are seqeunced. 16S metagenomics is a complex measurement process comprised of multiple molecular laboratory and computation steps (Julia K Goodrich et al. 2014). 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).

A key step in the measurement process is the grouping of sequences into biologically relevant units, or operational taxonomic units (OTUs), a process commonly referred to as clustering. There are a number of difference clustering methods which generate OTUS with different characteristics. The two most commonly used clustering methods are *de-novo* clustering and open-reference clustering. There are a number of different algorithms for *de-novo* clustering, though they all attempt to group sequences in a dataset within a defined similarity threshold (Westcott and Schloss 2015). Open-reference clustering matches seqeunces to a set of reference sequences that have been previously clustered (*de-novo*) then performing *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).

Further challenging the measurement process is the compositional nature of data, that is the proportion of an organism within a sample is being measured (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 extracted DNA.

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 mehtods optimized for this type of data are not necessarily optimizd to handle the additional biases and noise present in real data. Data generated from environmental 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 Generating Mixtures

Dataset comprised of mixtures of environmental samples was generated and used to assess the count tables generated using three different bioinformatic pipelines.

2.1.1 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 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. 1 Panel A). 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. 1 Panel B). 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

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 account 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.

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.

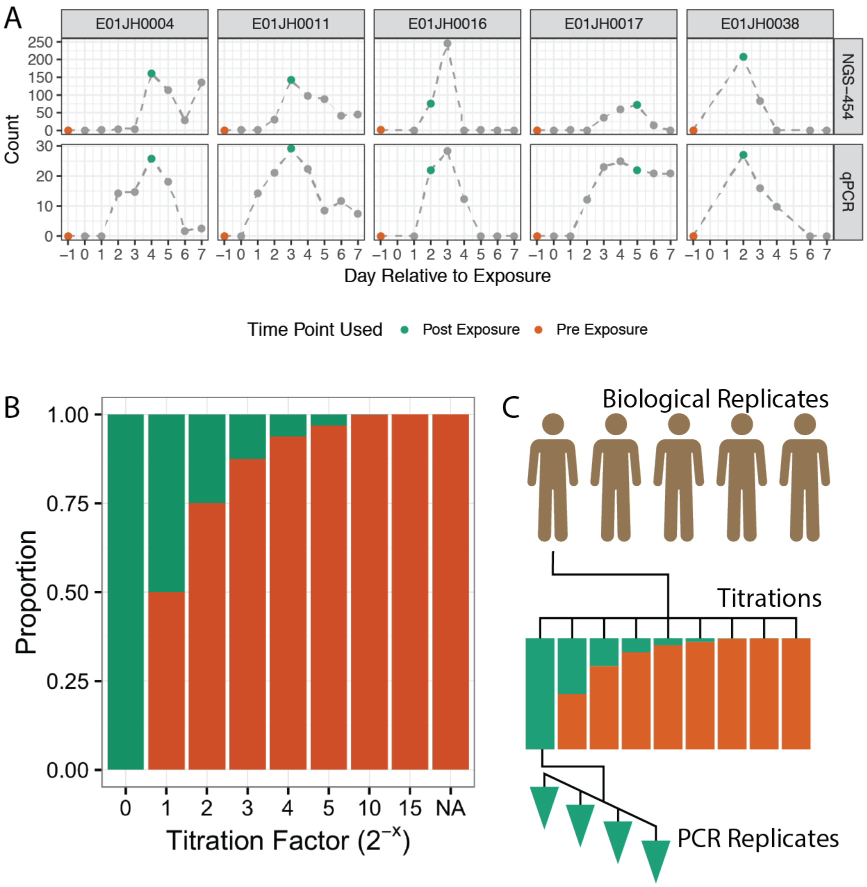

Figure 1: Sample selection and experimental design for two-sample titration 16S rRNA metagenomic sequencing assessment dataset. A) Pre- and post-exposure samples from five participants in a vaccine trial (Harro et al. 2011) were selected based on *Escherichia coli* abundance measured using qPCR and 454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

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 (5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGC-CTACGGGNGGCWGCAG - 3') and reverse primers (GTCTCGTGGGCTCGGAGATGTGTATAAGA-GACAGGACTACHVGGGTATCTAATCC). 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, the indexing PCR, and after 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 then used to generate libraries for sequencing. The initial PCR product was 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 (**Check Barcodes** 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.

2.3.1 Sequencing Data Quality Assessment

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.

2.4 Sequence Processing

Sequence data was processed using three 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 compared to the amplicons the protocol was developed for. The Mothur pipeline included an initial pre-processing step where forward and reverse reads were merge using the Needleman-Wunsch algorithm. Low quality reads, presence of ambiguous bases, reads that failed alignment to the SILVA reference database (https://www.arb-silva.de/), 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). The unclustered was based off of the mothur *de-novo* clustering pipeline, where the sequences were the paired-end reads were merged and filtered using the `make.contigs` command and then dereplicated. Reads were aligned to the reference Silva alignment and reads failing alignment were excluded from the dataset. The most abundant 40,000 OTUs, 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

- 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 (robinson2010, 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 C_{obs_j} observed counts for titration j, counts for unmixed C_{pre_j} and C_{post_j}. **TODO** Revise notation to indicate vector of relative abundance values for feature set. 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 inferθ_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 (Table 4). Features used were individual specific. To be included in the following analysis 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 quanitatively 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.

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. Only individuals with consistent inferred and estimated θ values were included in the log fold-change analysis, E01JH0004, E01JH0011, and E01JH0016. Pre-dominant and pre-specific features were identified based on log fold-changes between pre- and post-exposure samples and number of PCR replicates the feature was observed in for pre- and post-exposure PCR replicates (Table 6). 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. 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.

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

2.7 Qualitative Assessment

For the qualitative measurement assessment we evaluated features only observed in either the unmixed pre- and post-exposure 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 test 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 (Eq. (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 (Eq. (4), (5)) were used to calculate the p-value for the Bayesian hypothesis test assuming equal priors, i.e. $P(\pi < \pi_{min_{exp}}) = P(\pi \geq \pi_{min_{exp}})$.

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

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

NOTE Not sure this is appropriate due to the difference in the range of π used for the null and alternative hypothesis. May also want to consider a different alternative hypothesis π upper limit, potentially using the $\pi_{min_{pre}}$, and post treatment $\pi_{min_{post}}$ to calculate a more realistic upper limit.

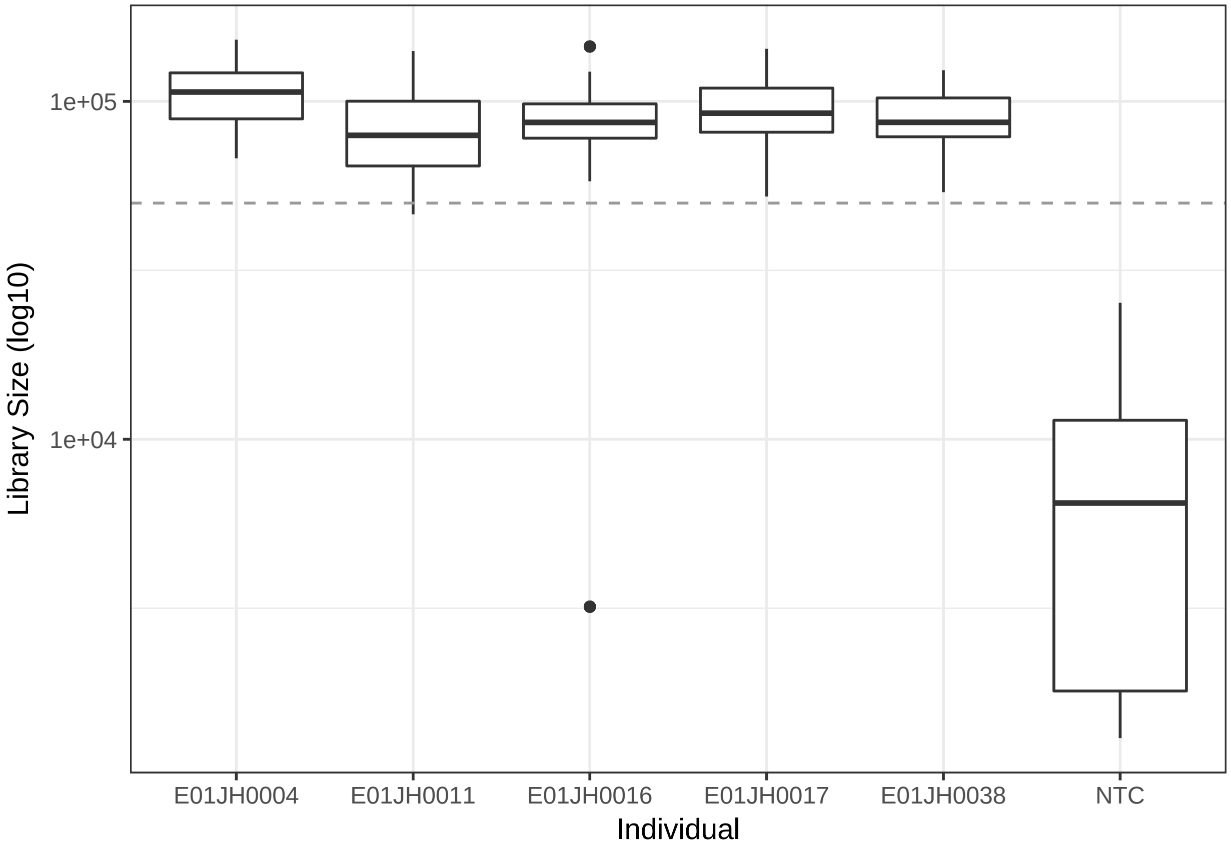

Figure 2: Distribution in the number of reads per barcoded sample (Library Size) by individual. Dashed horizontal line indicates 50,000 reads per barcoded sample.

3 Results

3.1 Dataset characteristics

Quality assessment of sequencing run summarizing number of reads per sample. Two barcoded experimental samples have less than 50,000 reads 2. The rest of the samples with less than 50,000 reads are negative PCR controls (NTC). Excluding the one failed reaction the total range in the observed number of sequences per sample is approximately 40,000 to 150,000 reads.

The sequencing dataset was processed using three bioinformatic pipelines. The resulting count tables were characterized for number of features, sparsity, and filter rate (Table 1). 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. Different pipelines have different approaches for handling low quality reads. QIIME pipeline has the highest filter rate while the highest number of features per sample.

The number of features per sample varied by bioinformatic pipeline (Fig 3). The number of observed features by sample was more correlated between the QIIME and Mothur pipelines compared to the DADA2 pipeline (Fig **??** 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.

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)	0.24 (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)

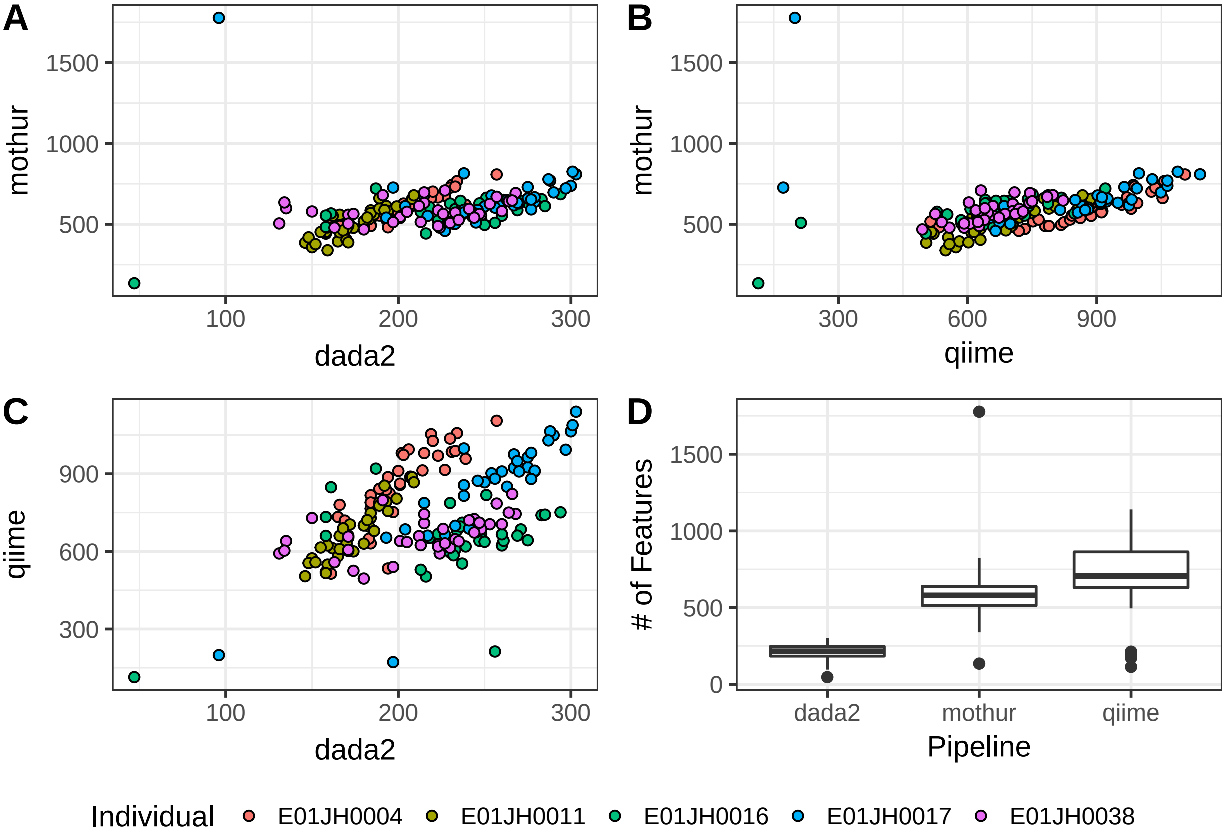

Figure 3: Comparison of the number of observed features per sample by bioinformatic pipeline. (A) Mothur v. DADA2, (B) Mothur v. QIIME, and (C) QIIME v. DADA2

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 were 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). The expected slope is 1, for a doubling every cycle. Slope estimates less than 1 were attributed to the assay standard curve amplification factors being less than 2 (Table 2). The 1-4 titration factor samples had Ct values consistently above the regression line (Figure 4). DNA from the unmixed samples were used to generate titrations 1 and 5 which may account for the offset in the regression line for titrations 1-4. The deviation from the regression line is small and unlikely to significantly impact the quantiative and qualitative assessment. For the pre-exposure ERCCs a regression line was fit to the log2 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 0.0401571 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 (Fig. 4), with a -1 expected slope. 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.

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.

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

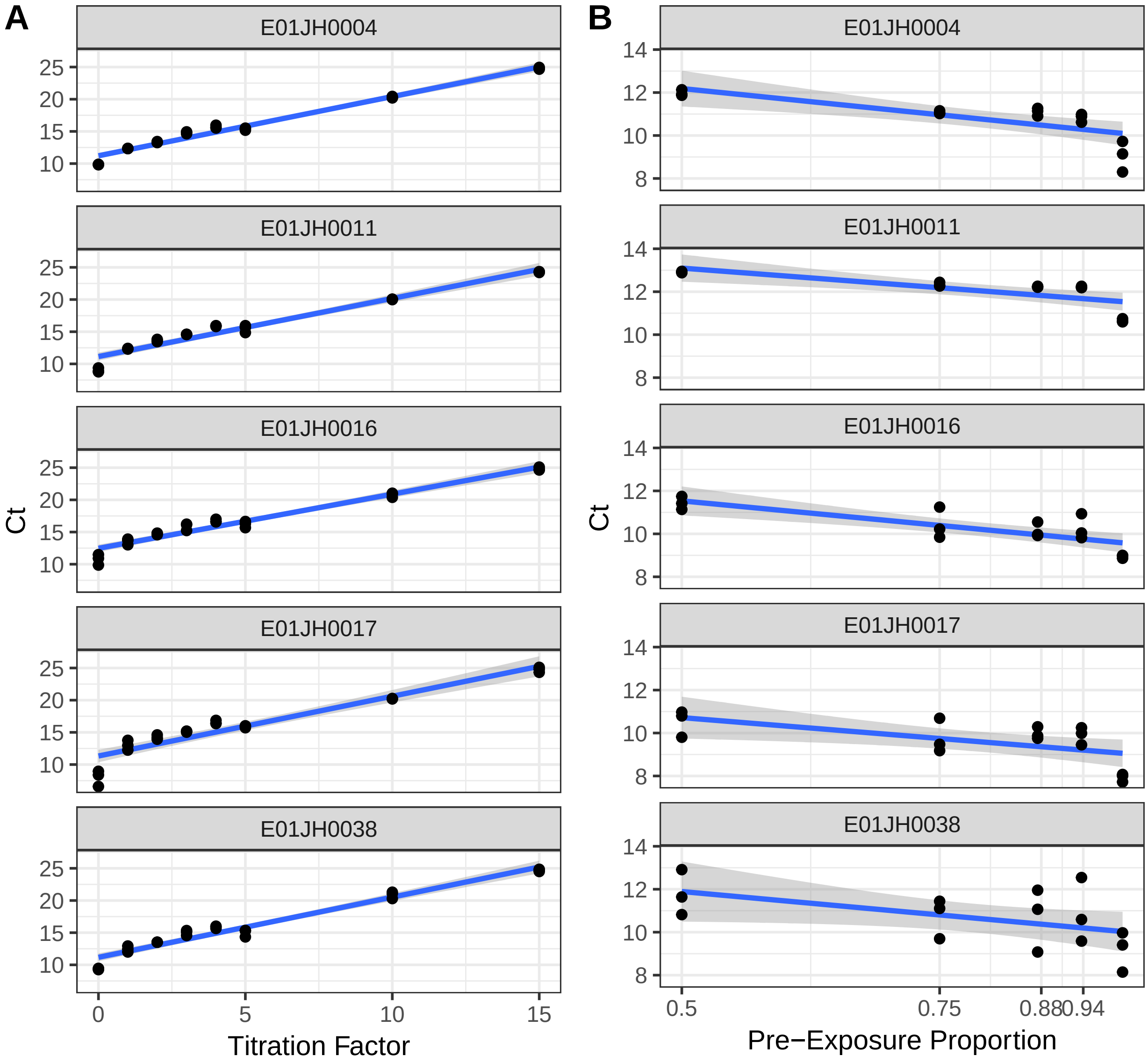

Figure 4: qPCR ERCC spike-in results for ERCC spiked into unmixed (A) Post-exposure samples and (B) Pre-exposure samples (titrations 1-4 only). X-axis is on a log2 scale with expected slope of 1 and -1 for Post-exposure and Pre-exposure spike-ins respectively.

Table 3: Slope estimates for linear model of prokaryotic DNA concentration and titration factor. Separate linear models were fit for each titrations 1-4 for each individual. Multiple test correction was performed using the Benjamini-Hochberg method. p-value indicates signficant difference from the expected slopes of 0.

Individual	Slope	Std. Error	Adj. p-value	R^2
E01JH0004	0.1786	0.1132	0.1735	0.0960
E01JH0011	0.0572	0.2010	0.7806	-0.0703
E01JH0016	-0.4006	0.0739	0.0006	0.6699
E01JH0017	0.4969	0.1310	0.0037	0.4887
E01JH0038	-0.3417	0.0804	0.0024	0.5491

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

3.2.2 Bacterial DNA Concentration

The proportion of prokaryotic DNA 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. 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. Standard curve efficiency 91.49, and R^2 1. 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. For titrations 1-5 the slope estimates are significantly different from 1 for individuals E01JH00016, E01JH0017, and E01JH00038 (Table 3, Fig. 5). These results indicate that the proportion of prokaryotic DNA is higher in the post-exposure sample than the pre-exposure sample for E01JH0016 and E01JH0038, lower in the post-exposure sample for E01JH0017, with no detectable difference for E01JH0004 and E01JH0011.

3.2.3 Theta Estimates

To account for differences in the proportion of prokaryotic DNA in the pre- and post-exposure samples we attempted to infer 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 size 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).

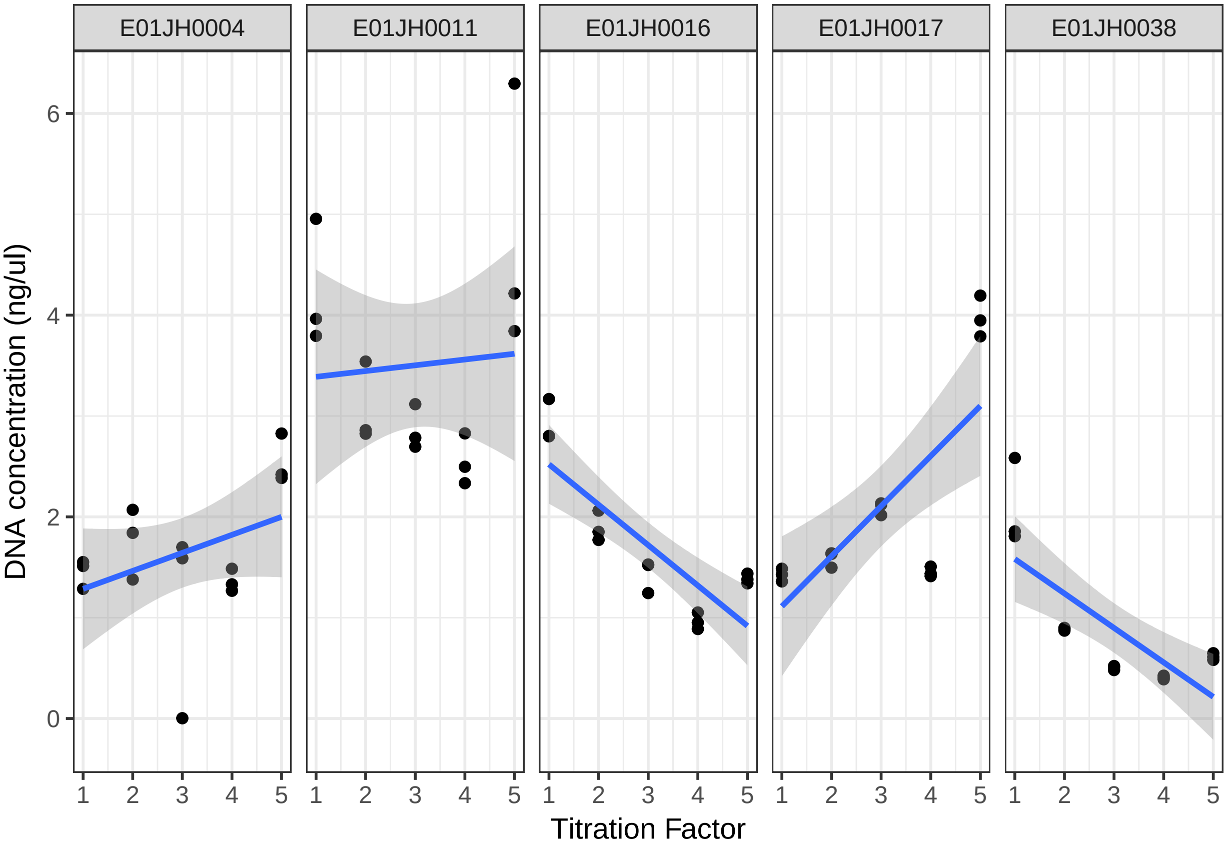

Figure 5: Prokaryotic DNA concentration (ng/ul) across titrations 1-4 measured using a 16S rRNA qPCR assay

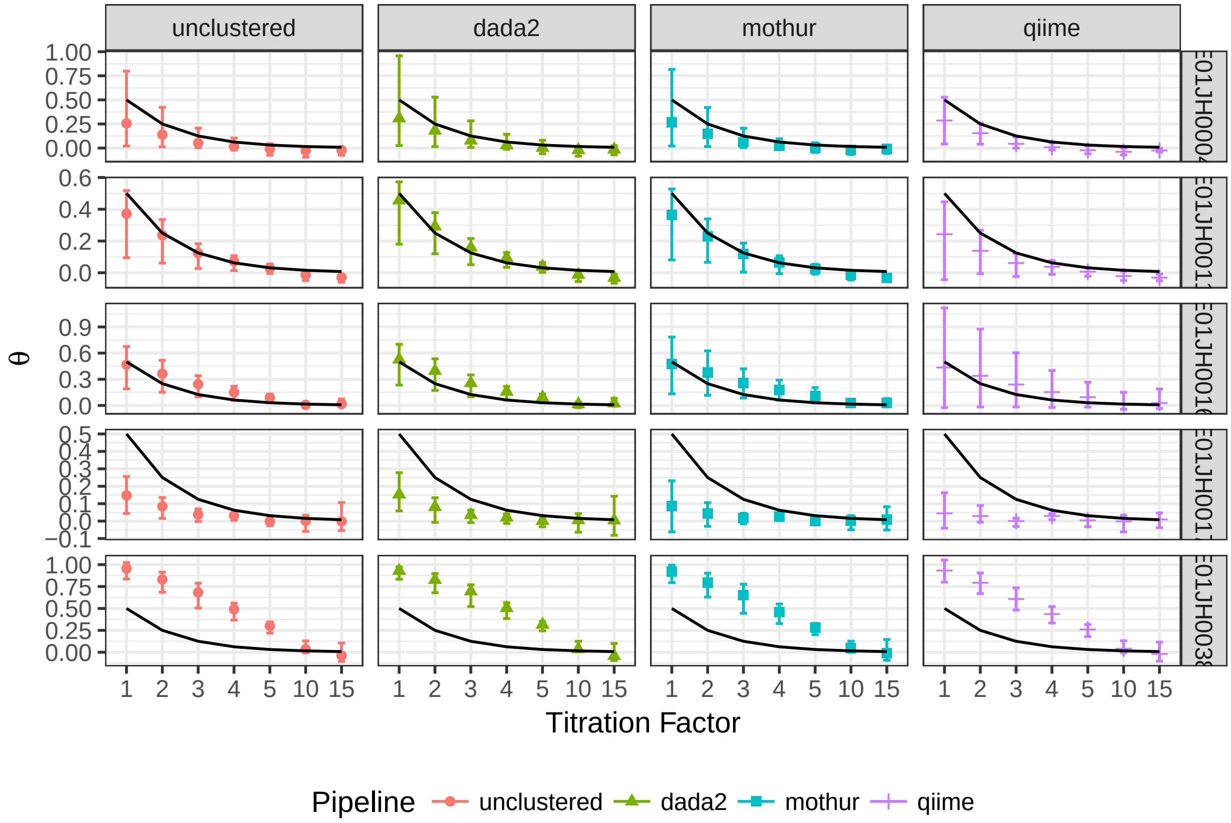

Figure 6: Theta estimates by titration, biological replicate, and bioinformatic pipeline. The points indicate mean estimate of 1000 bootstrap theta estimates and errorbars 95% confidence interval. The black line indicates the expected theta values. Theta estimates below the expected theta indicate that 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.

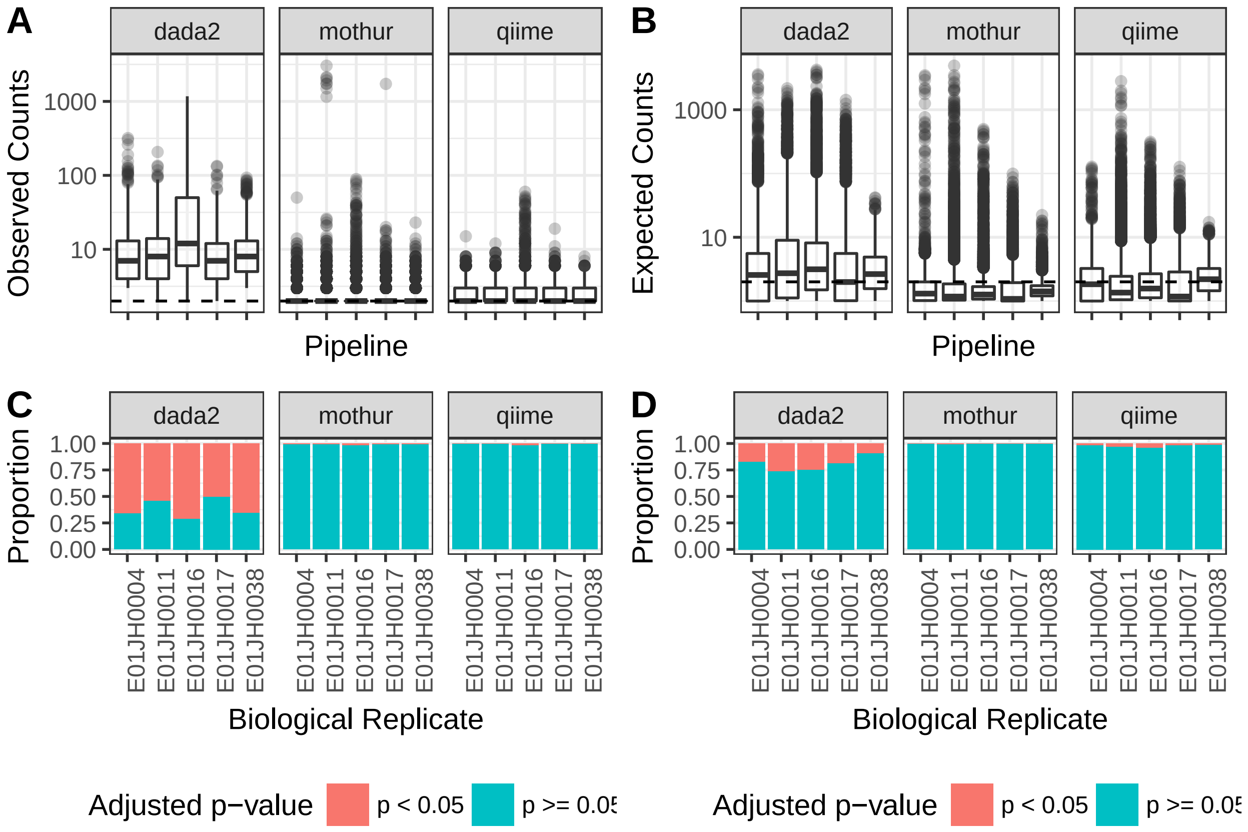

Figure 7: Distribution of (A) observed count values for titration-specific features and (B) expected count values for unmixed-specific features by pipeline and individual. The horizontal dashed line indicates a count value of 1. (C) Proportion of unmix-specific features and (D) titration-specific features 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.

3.3 Measurement Assessment

Next we assessed the 16S rRNA measurement process using our two-sample titration dataset. We assessed the qualitative and quanitative nature of the 16S metagenomics measurement process. For the qualitative assessment we looked 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) and expected (unmix-specific) counts. (Fig. 7A-B). 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 features with 0 observed counts 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).

Table 5: Maximum feature-level error rate bias (median error rate) and variance (robust COV) by pipeline and individual.

Pipeline	Metric	E01JH0004	E01JH0011	E01JH0016	E01JH0017	E01JH0038
dada2	Median	1.66	1.40	17.03	0.88	0.63
mothur	Median	2.39	3.71	19.24	2.09	1.76
qiime	Median	1.44	3.31	8.83	2.39	1.09
unclustered	Median	2.65	4.38	16.85	2.10	1.85
dada2	RCOV	4.50	3.91	7.36	3.95	6.75
mothur	RCOV	68.56	4.37	3.71	4.18	4.77
qiime	RCOV	39.36	5.19	2.40	7.92	17.22
unclustered	RCOV	11.89	14.45	10.48	5.90	8.36

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 theta values were used to calculate titration and feature level 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 **??**). 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 replicates, therefore these features were not included in the error analysis. For all pipelines the expected relative abundance is greater than the observed relative abundance for relative abundance estimates less than 1e-4. The deviation from the expected value on the low end varies by biological replicate and pipeline. Outliers are observed for all pipelines and individuals.

Next we evaluated the quantitative accuracy of the relative abundance values by comparing the distribution of the feature-level median error and feature-level robust coefficient (RCOV=(maximum - minimum)/median) of variation for the relative abundance error rate across pipelines (Fig. 8). Feature-level median error rates and RCOV were compared across pipelines and individuals using a mixed effects model. Large error rates were observed with all pipelines for E01JH0016 (Table 5). 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. When accounting for biological replicate effect DADA2 had a lower feature-level error rate compalred to mothur and qiime, qiime and mothur are not different from each other.
Unlike feature-level error rates, large RCOV was observed for all individuals and pipelines (Table 5). Outlier values were also excluded from the RCOV analysis. The feature-level RCOV was higher for DADA2 compared to Mothur and QIIME though not significantly (Fig. 8C).

Across all pipelines and individuals the estimated and expected logFC estimates were positively correlated (Fig. 9A). However, the slope of the linear relationship between the expected and estimated log fold-change estimates were less than 1 for all pipelines and individuals, with low adjusted R^2 values, with median and minimum-maximum of 0.035 (0.00048-0.093)) and slope estimates (0.075 (0.018-0.12)). Outlier features were the primary driver of the lower than expected slope estimates. **Need to verify** The bias and variance of the log fold-change estimates were compared between pipelines. 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. There was no statistical difference in the log fold-change error feature-level median error rate 9B) or error rate RCOV 9C). 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.

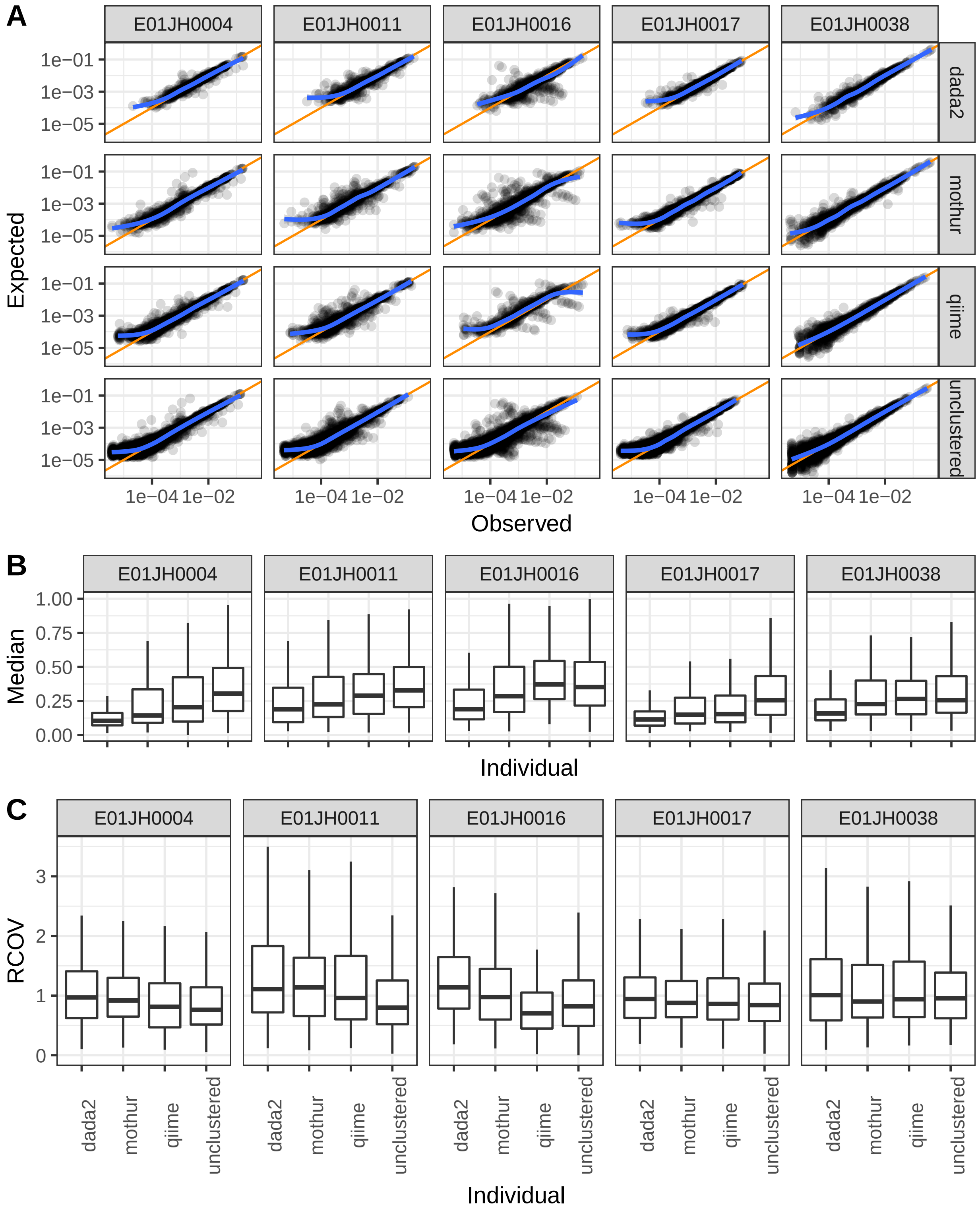

Figure 8: (A) Expected and observed count relationship. Orange line indicates expected 1-to-1 relationship. Blue line a smoothed regression line of the observed and expected value relationship. Distribution of feature-level relative abundance (B) median error rates and (C) robust coefficient of variation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

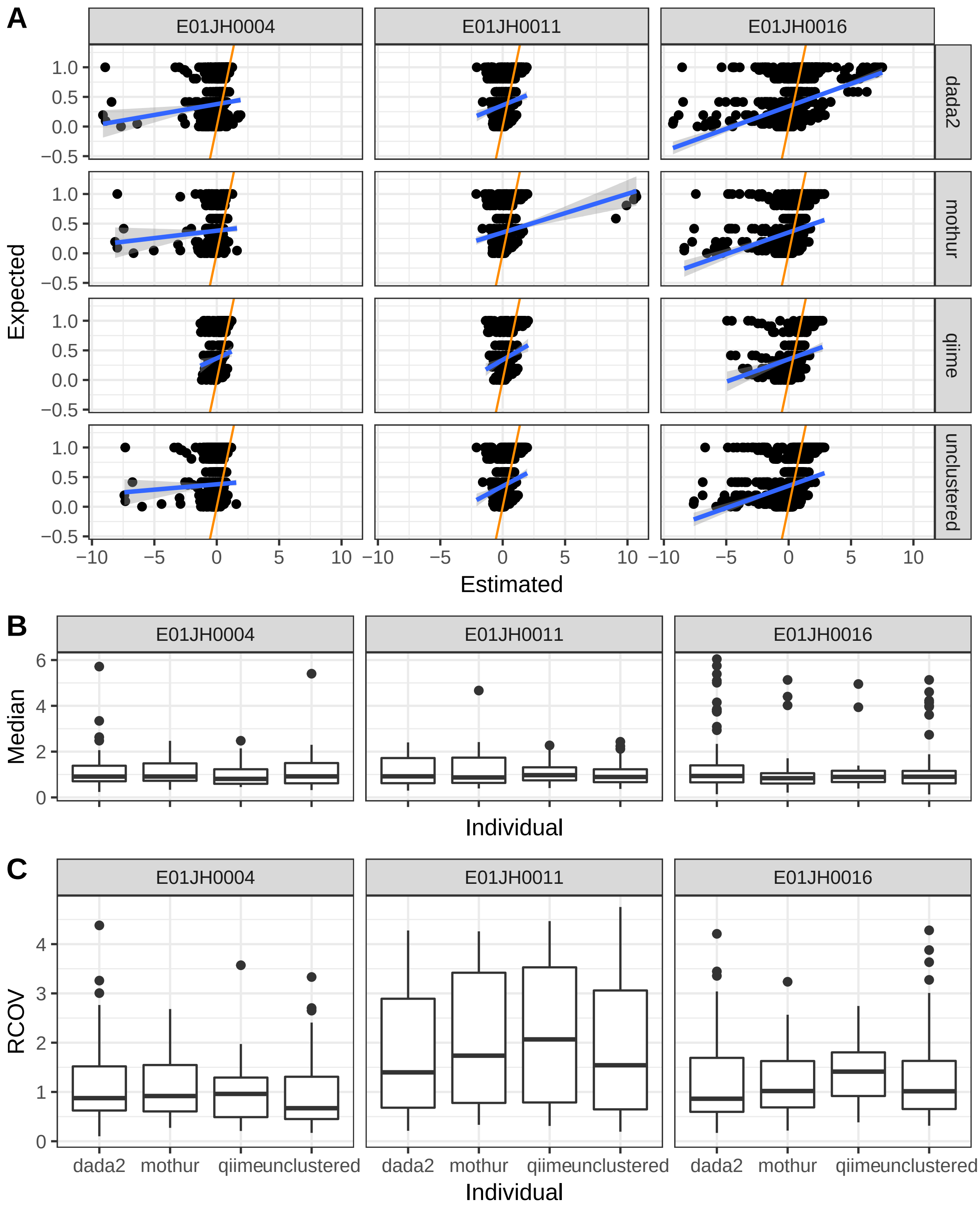

Figure 9: Relationship between the observed and expected logFC for pre-specific and pre-domiant features by pipeline and individual for all titration pair comparisons. Orange line indicates expected 1-to-1 relationship between the estimated and expected logFC. Blue line is a linear model was fit to the data and grey area is the models uncertainty estimate. Distribution of feature-level (B) median error rate and (C) robust coefficient of variation (RCOV) by individual and pipeline

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

4 Discussion

- Sample experimental design
  - Limitation: number of features with differentially abundant between pre- and post-exposure
- Bacterial DNA proportion
  - Limitation: additional uncertainty in expected values
- Pipeline characterization differences
  - Number of features per sample and total abundance
  - DADA2 has higher feature abundance due to fewer features and lower filter rate. Resulting in increased statistical power.
- Why qualitative analysis is pipeline dependent?
  - What dependency means for 16S gene surveys
  - DADA2 spurious OTUs a result of higher counts and therefore increased statistical power.
- Why quantitative analysis is biological replicate dependent?
  - What dependency means for 16S gene surveys, when does bioinformatic pipeline matter and when does it not matter?
  - Differences in proportion of bacterial DNA between the pre- and post-exposure samples drives individual specific results.
  - The proportion of non-prokaryotic DNA in a sample is not taken into considering for nearly all 16S studies.
  - How do differences impact inferences drawn from statistical analyses?
- Relative abundance
  - Outliers
  - Need to summarise across replicates
  - Noisy data
- log fold-change - outlier features
- Relationship between factors impacting quant and qual analysis

5 Conclusions

- How this dataset can be used to evaluate and characterize bioinformatic pipelines and clustering methods.
- Given study results
  - How would you analyze 16S sequencing data assuming current methods?
  - How would you like to analyze 16S sequencing data?
  - What are the limitations of current methods?
  - What would you like to see a clustering method/ pipeline be able to do?
    - What should be improved?

6 Session information

6.1 Git repo commit information

The current git commit of this file is 9ef7acaa8d627098261e03bfe0628a6271c31468, which is on the master branch and was made by nate-d-olson on 2017-10-16 14:56:11. The current commit message is combined figures and tables. 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, darwin16.7.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-10-16
```

6.3 Package Versions

package	version	date	source
base	3.4.2	2017-09-29	local
bindrcpp	0.2	2017-06-17	CRAN (R 3.4.0)
Biobase	2.36.2	2017-06-21	Bioconductor
BiocGenerics	0.22.0	2017-05-04	Bioconductor
BiocParallel	1.10.1	2017-05-04	Bioconductor
Biostrings	2.44.1	2017-06-21	Bioconductor
broom	0.4.2	2017-02-13	CRAN (R 3.4.0)
datasets	3.4.2	2017-09-29	local
DelayedArray	0.2.7	2017-06-21	Bioconductor
dplyr	0.7.2	2017-07-20	CRAN (R 3.4.1)
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	1.12.2	2017-06-21	Bioconductor
GenomicAlignments	1.12.1	2017-06-21	Bioconductor
GenomicRanges	1.28.4	2017-07-19	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.4.0)
ggpubr	0.1.4	2017-06-28	CRAN (R 3.4.1)
git2r	0.19.0	2017-07-19	CRAN (R 3.4.1)
glmnet	2.0-10	2017-05-06	CRAN (R 3.4.0)
graphics	3.4.2	2017-09-29	local
grDevices	3.4.2	2017-09-29	local
IRanges	2.10.2	2017-06-21	Bioconductor
knitr	1.17	2017-08-10	CRAN (R 3.4.1)
limma	3.32.3	2017-07-19	Bioconductor
magrittr	1.5	2014-11-22	CRAN (R 3.4.0)
Matrix	1.2-11	2017-08-21	CRAN (R 3.4.2)
matrixStats	0.52.2	2017-04-14	CRAN (R 3.4.0)
metagenomeSeq	1.18.0	2017-05-04	Bioconductor
methods	3.4.2	2017-09-29	local
modelr	0.1.1	2017-07-24	CRAN (R 3.4.1)
parallel	3.4.2	2017-09-29	local
ProjectTemplate	0.8	2017-08-09	CRAN (R 3.4.1)
purrr	0.2.3	2017-08-02	CRAN (R 3.4.1)
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	2017-07-19	Bioconductor
Rsamtools	1.28.0	2017-05-04	Bioconductor
S4Vectors	0.14.3	2017-06-21	Bioconductor
ShortRead	1.34.0	2017-05-04	Bioconductor
stats	3.4.2	2017-09-29	local
stats4	3.4.2	2017-09-29	local
stringr	1.2.0	2017-02-18	CRAN (R 3.4.0)
SummarizedExperiment	1.6.3	2017-06-21	Bioconductor
tibble	1.3.3	2017-05-28	CRAN (R 3.4.0)
tidyr	0.6.3	2017-05-15	CRAN (R 3.4.0)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.4.0)
utils	3.4.2	2017-09-29	local
XVector	0.16.0	2017-05-04	Bioconductor

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. 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.

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.

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.

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.