A Sample Mixture Experiment to Assess 16S rRNA Metagenomic Normalization and Differential Abundance Methods.

Nate Olson

2017-03-07

Chapter 1

Authors

Nathan D. Olson
* National Institute of Standards and Technology
* University of Maryland

Stephanie Hao
* Johns Hopkins University

Winston Timp
* Johns Hopkins University

Marc Salit
* National Institute of Standards and Technology
* Stanford University

O. Colin Stine
* School of Medicine, University of Maryland

Hector Corrada Bravo
* University of Maryland

Chapter 2

Abstract

Chapter 3

Background

- General Introduction
  - Microbiome and 16S metagenomics
  - Differential abundance and normalization
    - Role of normalization in addressing biases, e.g. differences in sequencing depth and sample coverage
  - Connection between 16S and RNAseq
- Use of mixtures for RNAseq validation
  - Complexity of real samples
  - Provides a level of truth
- 16S metagenomic measurement process - specifically sequence processing
- Assessing normalization and differential abundance
  - Current approaches, and limitations
    * how do spike-ins fit into this work
  - Bias-variance trade off
- Study objectives
  - Demonstrate how mixtures can be used for 16S
  - Generate a dataset for use in evaluating 16S
  - Develop methods for assessing normalization and differential abundance
  - Identification of feature characteristics responsible for low performance

Chapter 4

Methods

Dataset of environmental sample mixtures was generated and used to evalute the abundance values and log-fold change values for count tables generated using three different bioinformatic pipelines.

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 as those with no *Eshechichia coli* detected in stool samples before exposure to Enterotoxigenic *Escherichia coli* (ETEC)) and timepoints with the highest concentration of *E. coli* after exposure (Pop et al. 2016) @ref(fig:experimenta_design) (Panel A). For the two-sample titration post-treatment samples (stool samples collected after exposure to *E. coli* ETEC) were titrated into pre-treatment samples (stool samples collected *before* exposure to *E. coli* ETEC) with log_2 changes in pre to post sample proportions @ref(fig:experimenta_design) (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).

Titration Validation

TODO Supplemental Table with ERCC plasmids, qPCR assay, and experimental design.
Table design 1. Sample, Treatment, ERCC plasmid, qPCR assay, cat number 2. sample, PCR plate, well position, qPCR assay, CT value 3. plate, well position, absorbance

To ensure that the two-sample titrations were correctly mixed independent ERCC plasmids were spiked into the unmixed pre- and post-treatment samples (Supplemental Table ERCC). The ERCC plasmids were resuspendended in 100 $ng/\mu L$ tris-EDTA buffer and 2 $ng/\mu L$ was spiked into each 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-treatment samples, the amount of bacterial DNA was quantified using the Femto Bacterial DNA quantification kit (Zymo Research, Irvine CA). All two-sample titrations and unmixed samples were run in triplicate along with a standard curve. An in-house standard curve consisting of log10 dilutions of *E. coli* DNA was used as the standard curve. (**TODO** Supplemental Material justification for using in-house instead of manufacturer provided 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.

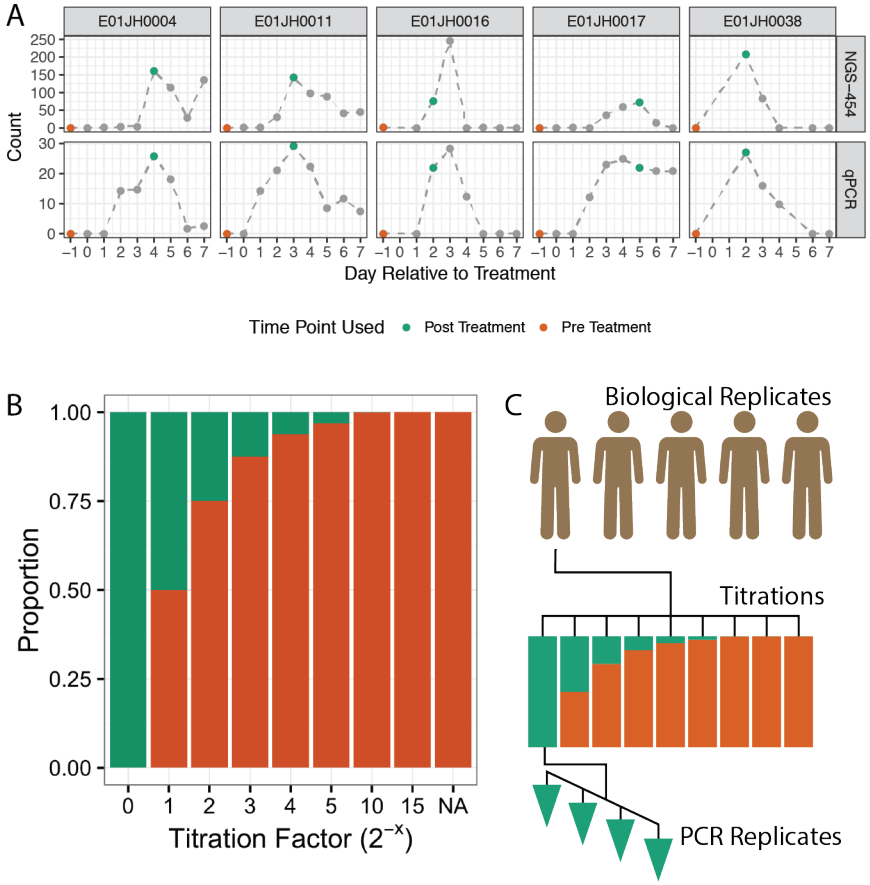


Figure 4.1: Sample selection and experimental design for two-sample titration 16S rRNA metagenomic sequencing assessment dataset. A) Pre- and post-treatment samples from five participants in a vaccine trial (Harro et al. 2011) were selected based on Escherichia coli abundance measured using qPCR and 454 16S rRNA metagenomics sequencing (454-NGS), data from Pop et al. 2016. Pre- and post-treatment samples are indicated with orange and green data points. Grey indicates other samples from the vaccine trial time series. B) The pre-treatment samples were titrated into post-treatment samples following a log_2 dilution series. The NA titration factor represents the unmixed pre-treatment sample. C) The five vaccine trial participants are biological replicates and independent sets of two-sample titrations were mixed for each. 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.

Sample Processing Workflow

The resulting in 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 (REF). The protocol consisted of an initial 16S rRNA PCR followed by a separate sample indexing PCR. A total of 192 PCRs were run including four PCR replicates per sample and 12 no template controls (**TODO** Supplemental PCR plate figure). After the initial PCR and clean-up the 192 PCRs were split into technical replicates and sent to two laboratories for library preparation and sequencing. The concentration of the resulting indexed 16S PCR products were normalized using the SequalPrep Normalization Plate Kit(Catalog n. A10510-01, Invitrogen Corp., Carlsbad, CA). The normalized indexed samples were pooled and sequenced on the Illumina MiSeq (Illumina Inc., San Diego, CA).

Library Preparation and Sequencing

The 16S PCR targeted the V3-V4 region, Bakt_341F and Bakt_806R (Klindworth et al. 2012). The V3-V4 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) (**TODO** Supplemental Figure Amplicon Region). The primer sequences include additional overhang adapter sequences to facilitate library preparation (5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG - 3') and reverse primers (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC).

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, Mocledular Devices LLC. Sunnyvale CA, USA) flourescent measurements were made with a Molecular Devices SpectraMax M2 spectraflourometer (Mocledular Devices LLC. Sunnyvale CA, USA). After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (**Check Barcodes** Illumina Inc., San Diego CA). The purified sample concentration was normalized using SequalPrep (ThermoFisher, Waltham MA), according to the manufactuers protocol and pooled prior to sequencing. The pooled library concentration was measured using the Qubit dsDNA HS Assay Kit (Part# Q32851, Lot# 1735902, ThermoFisher, Waltham, MA USA). Due to 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.

Sequence Processing

Sequence data was processed using a number of bioinformatic pipelines, Mothur (Schloss et al. 2009), QIIME (Caporaso et al. 2010), DADA2 (Callahan et al. 2016), and an in-house pipeline with *de-novo* clustering and phylogenetic placement. 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, see the Makefile in the project github repository (**TODO** add website). 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). Average neighbor clustering for OTU clustering using pairwise sequences distances calculated from the reference based multiple sequence alignment. The RDP classifier implemented in mothur was used for taxonomic classification against the mothur provided

version of the RDP v9 training set (Q. Wang et al. 2007). The QIIME 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). The methods included open reference clustering **TODO: ADD MORE TO THE DESCRIPTION** (Caporaso et al. 2010). DADA2 a R native pipeline was also used to process the sequencing data (Callahan et al. 2016). The pipeline included a sequence inference step and taxonomic classification using the DADA2 implementation of the RDP naive bayesian classifier. The in-house pipeline used Sickle for read trimming (Joshi NA 2011), Pandaseq (Masella et al. 2012) for merging paired-end reads, DNAclust for OTU assignment using a 0.99 similarity threshold (Ghodsi, Liu, and Pop 2011), and a phylogenetic placement based method TIPP was used for taxonomic assignment (Nguyen et al. 2014).

Data Analysis

All data analysis was performed using the statistical programming language R (REF) and the RStudio IDE (REF). Initial quality assessment of the sequence files (fastq) was performed using the Bioconductor package Rqc (REF).

Measurement Assessment

- Measurement assessment of the count table values sample level feature abundance

- Assessment count values with and without normalization and transformation
- Variance - variability in observed counts between the four PCR replicates
  - Coefficient of variation was calculated sd/mu for each feature, pipeline, and normalization method
  - **TODO** Method used to compare CV values
- Bias - how well the observed count values agree with expected values
  - The expectd count values were calculated as follows, with $p = 2^{-}t$ and t is the titration factor. C_{exp} is the expected count value and C_{post} is the observed count value for the unmixed post-treatment sample and C_{pre} is the unmixed pre-treatment sample. To account for potential within and between plate effects the observed counts for the pre- and post-treatment replicate in the same plate and side of the replicate 96 well plate as the samples being assessed were used to calculate the expected count value.

$$C_{exp} = [C_{post} \times p] + [C_{pre} \times (1 - p)]$$

* The following error metric was used to summarize measurement bias

$$E = \frac{\sqrt{1/n \sum_i^n (C_{obs} - C_{exp})^2}}{\sqrt{C_{exp}^2}}$$

- Normalization Methods
  - None
  - TSS - total sum scaling
  - CSS - cumulative sum scaling
  - Senthil's method
  - rareify
  - square root transformation - AIST spike-ins
  - Holmes arcsin transformation

log-fold Change Assessment

- Differential Abundance Methods
  - metagenomeFeatures
  - DESeq
  - EdgeR
  - Limma?
- Variance estimates provided by differential abundance methods were used to compare the logFC variance between bioinformatic pipelines, normalization methods, and differential abundance methods (**Q** Not sure can directly compare differential abundance methods as differences in model assumptions determine the variability)
- Bias
  - expectation calculation

Chapter 5

Results

Sample Selection

Five biological replicates from the ETEC vaccine trial were selected based on the absence of detectable *E. coli* in the pre-treatment sample and the post-treatment sample with the highest abundance of *E. coli* measured using qPCR and 16S rRNA metagenomics (Pop et al. 2016) @ref(fig:experimenta_design). Due to limited material availability for biological replicate E01JH0016, post-treatment day 1 was used instead of day 2. For biological replicate E01JH0017, there was a discrepancy between the maximum abundance post-treatment time point between the qPCR and 454 data. The post-treatment time point with the maximum qPCR abundance value was used in this study.

Sequence QA

Potentially Move To Supplemental

16S PCR validation

The initial 16S PCR product was first verified for amplification and amplicon size using gel electrophoresis. The concentration of the PCR product was assessed using pico green, samples with negative measured concentration values (excluding no template controls (NTC)) were also check using Qubit. Only one of the 180 samples did not successfully amplify E01JH0016 dilution 5 (position F9, plate 1). All but one of the no template control concentration measurements was less than 1 **ng per ul**, when the one sample was checked with Qubit the concentration was too low to measure (Supplemental Material).

Normalization

Library Pooling

Seq Data QA

- Variability in library size
- seq error rate - especially for reverse reads and for overlap region

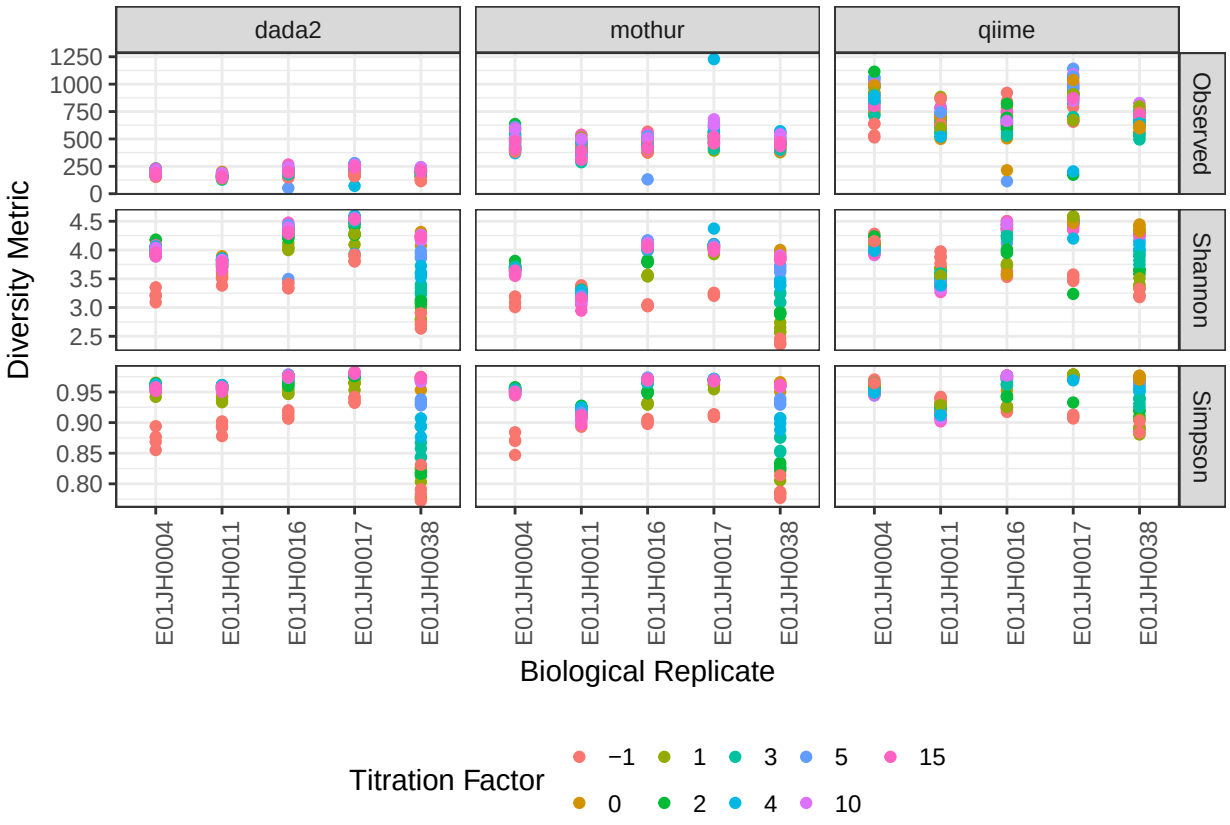


Figure 5.1: Alpha diversity metrics calculated for the different sequence processing pipelines.

Sequence Processing

Pipeline characteristics

Section objectives * make non-quantitative statements * capturing differences in quality across samples * Statements/ Figures showing how datasets behave

MOVE TO ARTIFACTS

- Characterization of different pipelines
  - total number of features
    - DADA2 3691
    - mothur 31948
    - qiime 11381
  - different taxonomic assignments
  - number of assigned vs. non-assigned

Alpha Diversity (richness comparison)

Comparison of feature richness between bioinformatic pipelines (@ref(fig:alpha_div)).

Rarefaction Curves

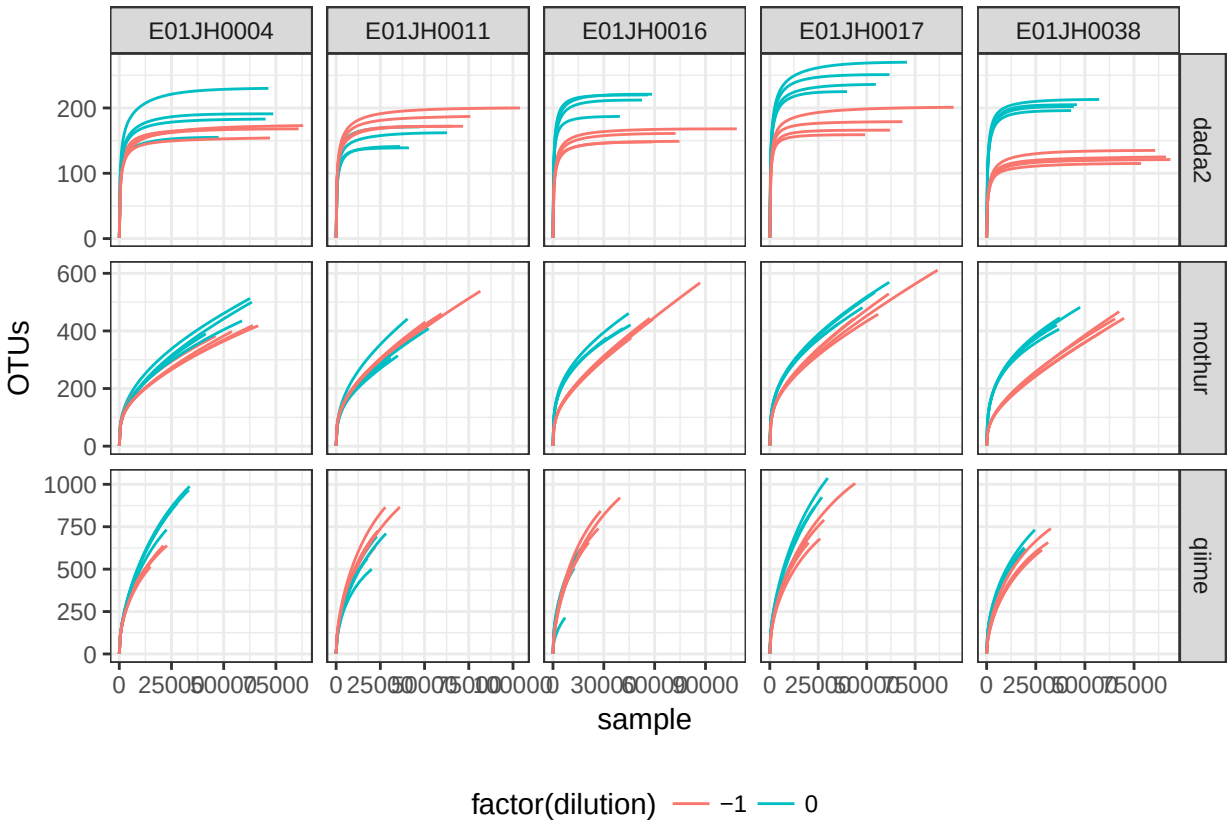
- Use to show coverage

- Unmixed only (include pooled replicates)

```
curve_df <- mrexp %>% map_df(get_curve_df, .id = "pipe")
```

```
## Joining, by = "samID"
## Joining, by = "samID"
## Joining, by = "samID"
```

```
ggplot(curve_df) +
    geom_path(aes(x = sample, y = OTUs, color = factor(dilution), group = samID)) +
    facet_grid(pipe~sampleID, scales = "free") +
    theme_bw() + theme(legend.position = "bottom")
```



Characterization of Differences in Feature Number

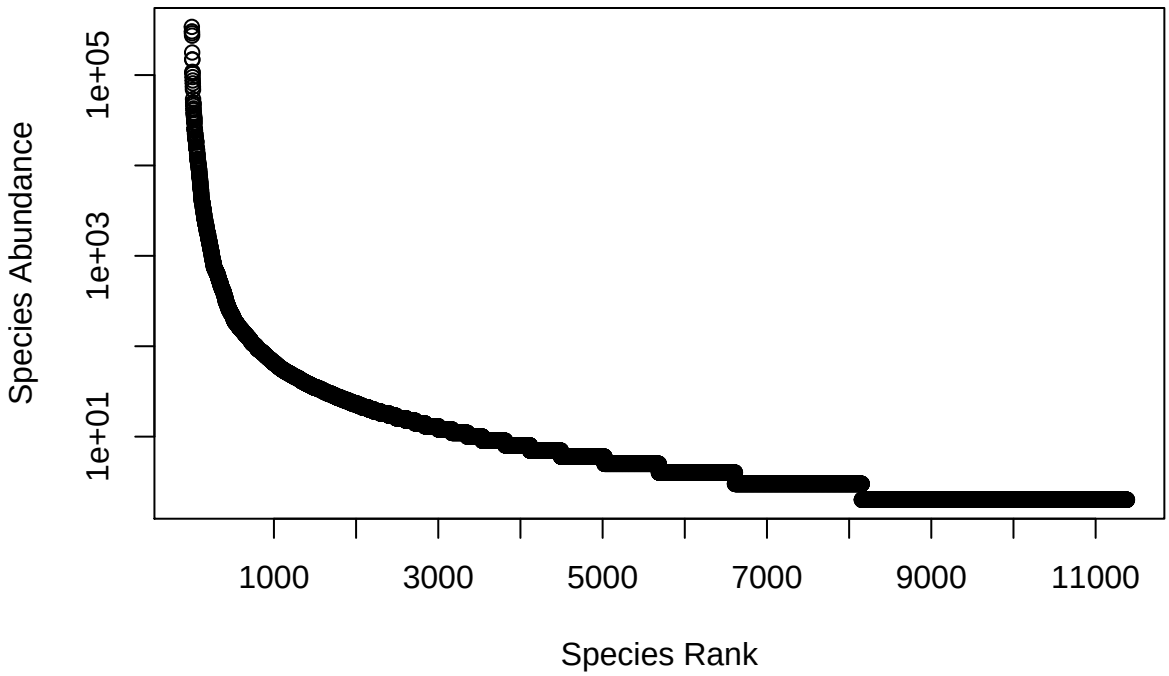
- Exploration of DADA2 features - specifically low total number of features
- QIIME lower number of sequences?

Beta Diversity - Overall sample similarity

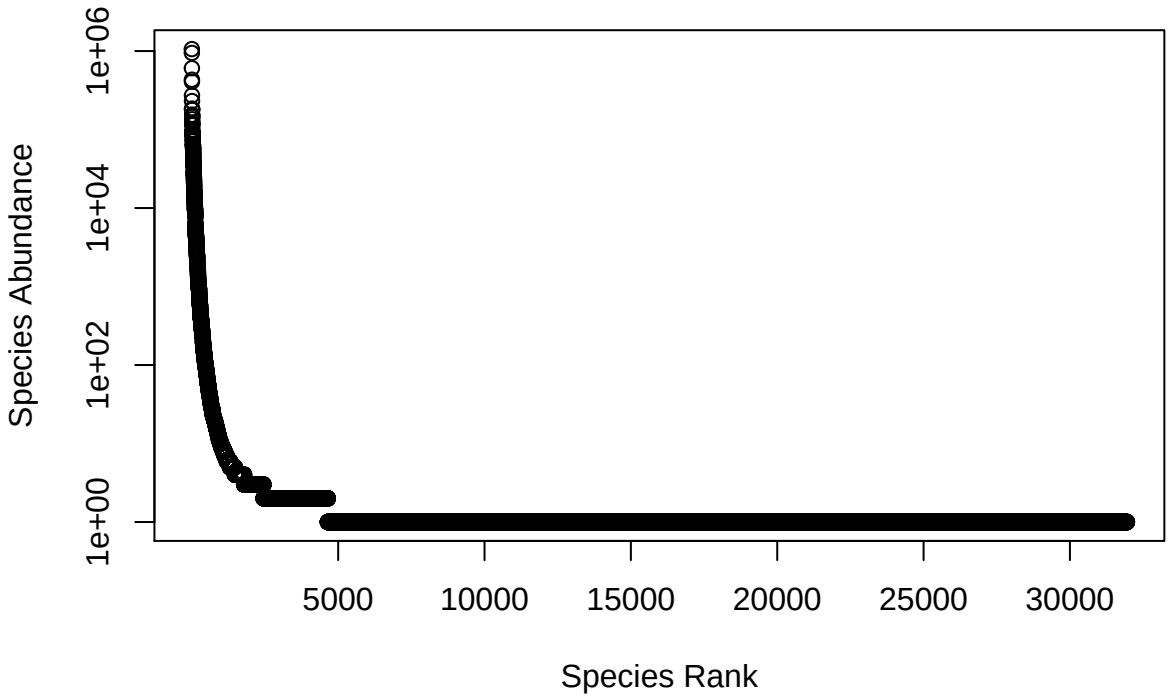
Species abundance distributions

- Characterize differences between clustering methods

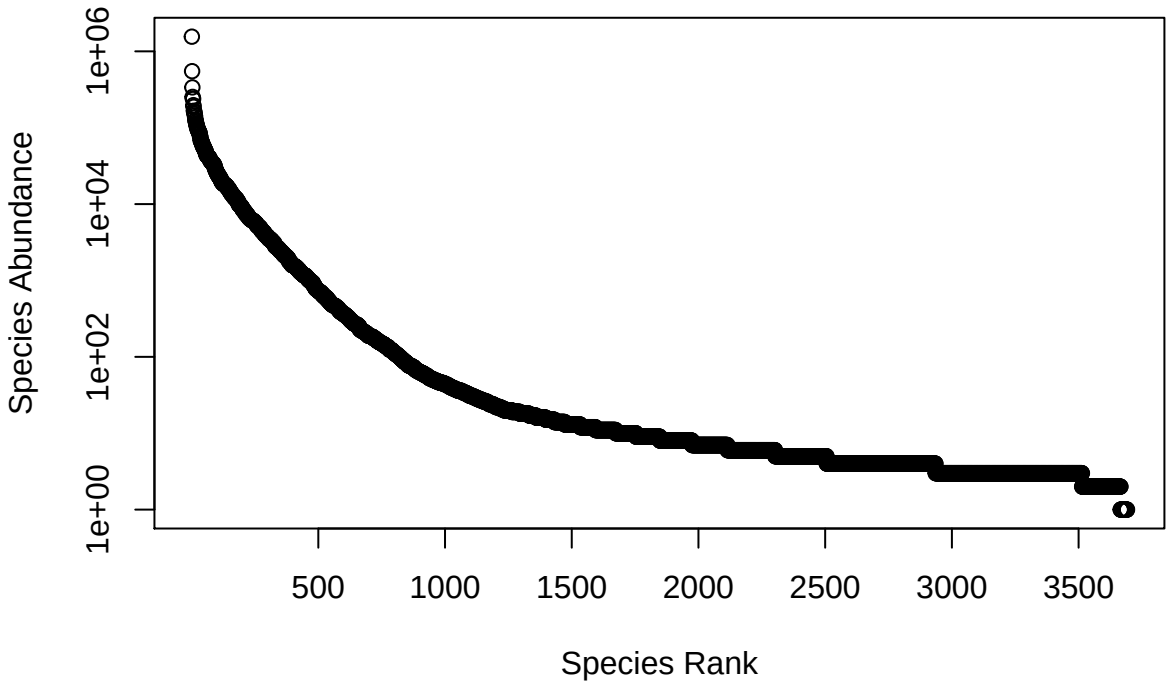
```
mrexp$qiime@assayData$counts %>% rowSums() %>% sads::rad() %>% plot()
```




```
mrexp$mothur@assayData$counts %>% rowSums() %>% sads::rad() %>% plot()
```




```
mrexp$dada2@assayData$counts %>% rowSums() %>% sads::rad() %>% plot()
```



MA plot to demonstrate observed fold changes

```
sample_ids <- pData(mrexp$dada2)$sampleID %>% unique()
sample_ids <- sample_ids[sample_ids != "NTC"]

ma_df <- mrexp %>% map_df(get_ma_df_by_sample, sample_ids, .id = "pipe")
```

```
# __TODO__ Filter low quality samples
ggplot(ma_df) + geom_point(aes(x = A, y = logFC, group = otu, color = sampleID), alpha = 0.5) +
      geom_hline(aes(yintercept = -1), linetype = 2) +
      geom_hline(aes(yintercept = 1), linetype = 2) +
      theme_bw() +
      scale_x_log10() +
      facet_wrap(~pipe)
```

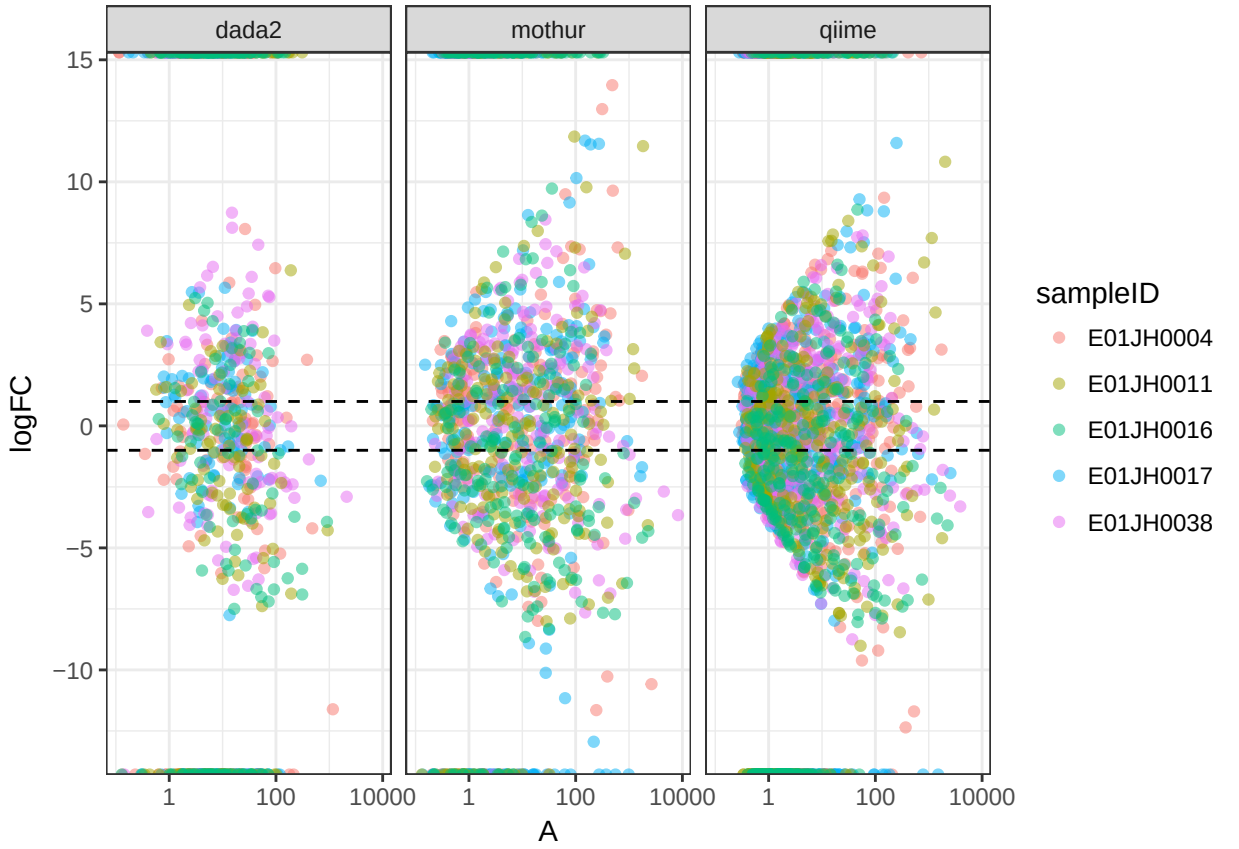


Figure 5.2: MA plot for different bioinformatic pipelines. Points at top and bottom of plots are OTUs only present in pre-treatment and post-treatment samples respectively.

Chapter 6

Count Table Value Analysis

See `artifacts/count_values.pdf`

- Overall Bias and Variance Summary
- features only in Pre and post unmixed
- effectiveness of high and low variance replicates **Not sure what I meant here**
- Correlating factors with feature level normalization performance
  - factors: well position, primer mismatches, GC, high level taxonomic group (gram positive vs. gram negative)
  - Potentially test for a phylogenetic signal?

Chapter 7

Log Fold Change Analysis

- Bias variance, pre and post only
- Correlate with factors

Chapter 8

Discussion

Chapter 9

Session information

```
s_info <- devtools::session_info()
print(s_info$platform)
```

```
##  setting  value
##  version  R version 3.3.2 (2016-10-31)
##  system   x86_64, darwin15.6.0
##  ui       RStudio (1.0.136)
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-03-07
```

```
s_info$packages %>% filter(`*` == "*") %>% select(-`*`) %>%
      knitr::kable()
```

package	version	date	source
bbmle	1.0.18	2016-02-11	CRAN (R 3.3.2)
Biobase	2.34.0	2016-11-07	Bioconductor
BiocGenerics	0.20.0	2016-11-07	Bioconductor
BiocParallel	1.8.1	2016-11-07	Bioconductor
Biostrings	2.42.1	2016-12-19	Bioconductor
DESeq	1.26.0	2016-11-28	Bioconductor
DESeq2	1.15.28	2017-02-02	bioc (readonly/DESeq2@125913)
dplyr	0.5.0	2016-06-24	CRAN (R 3.3.2)
edgeR	3.16.5	2017-02-02	Bioconductor
forcats	0.2.0	2017-01-23	CRAN (R 3.3.2)
foreach	1.4.3	2015-10-13	CRAN (R 3.3.1)
GenomeInfoDb	1.10.2	2017-01-04	Bioconductor
GenomicAlignments	1.10.0	2016-11-07	Bioconductor
GenomicRanges	1.26.2	2017-01-04	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.3.2)
glmnet	2.0-5	2016-03-17	CRAN (R 3.3.1)
IRanges	2.8.1	2016-11-18	Bioconductor
knitr	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.2)
limma	3.30.9	2017-02-02	Bioconductor
locfit	1.5-9.1	2013-04-20	CRAN (R 3.3.1)
Matrix	1.2-8	2017-01-20	CRAN (R 3.3.2)

package	version	date	source
metagenomeSeq	1.16.0	2016-11-07	Bioconductor
modelr	0.1.0	2016-08-31	cran ((**???**))
permute	0.9-4	2016-09-09	CRAN (R 3.3.1)
phyloseq	1.19.1	2017-01-04	Bioconductor
ProjectTemplate	0.7	2016-08-11	CRAN (R 3.3.1)
purrr	0.2.2	2016-06-18	CRAN (R 3.3.1)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.1)
readr	1.0.0	2016-08-03	CRAN (R 3.3.1)
readxl	0.1.1	2016-03-28	cran ((**???**))
Rqc	1.8.0	2016-11-07	Bioconductor
Rsamtools	1.26.1	2016-11-07	Bioconductor
S4Vectors	0.12.1	2016-12-19	Bioconductor
sads	0.3.1	2016-05-13	CRAN (R 3.3.2)
savR	1.12.0	2016-11-07	Bioconductor
ShortRead	1.32.0	2016-11-07	Bioconductor
stringr	1.1.0	2016-08-19	CRAN (R 3.3.1)
SummarizedExperiment	1.4.0	2016-11-07	Bioconductor
tibble	1.2	2016-08-26	CRAN (R 3.3.1)
tidyr	0.6.1	2017-01-10	CRAN (R 3.3.2)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.3.2)
vegan	2.4-2	2017-01-17	CRAN (R 3.3.2)
XVector	0.14.0	2016-11-07	Bioconductor

Chapter 10

References

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.

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.

Ghodsi, Mohammadreza, Bo Liu, and Mihai Pop. 2011. "DNACLUST: Accurate and Efficient Clustering of Phylogenetic Marker Genes." *BMC Bioinformatics* 12 (1). BioMed Central: 1.

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.

Joshi NA, Fass JN. 2011. "Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for Fastq Files (Version 1.33)." https://github.com/najoshi/sickle.

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.

Masella, Andre P, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown, and Josh D Neufeld. 2012. "PANDAseq: Paired-End Assembler for Illumina Sequences." *BMC Bioinformatics* 13 (1). BioMed Central: 31.

Nguyen, Nam-phuong, Siavash Mirarab, Bo Liu, Mihai Pop, and Tandy Warnow. 2014. "TIPP: Taxonomic Identification and Phylogenetic Profiling." *Bioinformatics* 30 (24). Oxford Univ Press: 3548–55.

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.

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.

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.