Using A Two Sample Titration Mixture Design to Evaluate Normalization and Differential Abundance Methods.

Nate Olson

2017-01-24

Chapter 1

Abstract

Chapter 2

Background

- General Introduction
  - Microbiome and 16S metagenomics
  - Differential abundance and normalization
  - Connection between 16S and RNAseq
- Use of mixtures for RNAseq validation
  - Complexity of real samples
  - Provides a level of truth
- 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

Chapter 3

Methods

Experimental Design

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 thoes 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). 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 **TODO**.

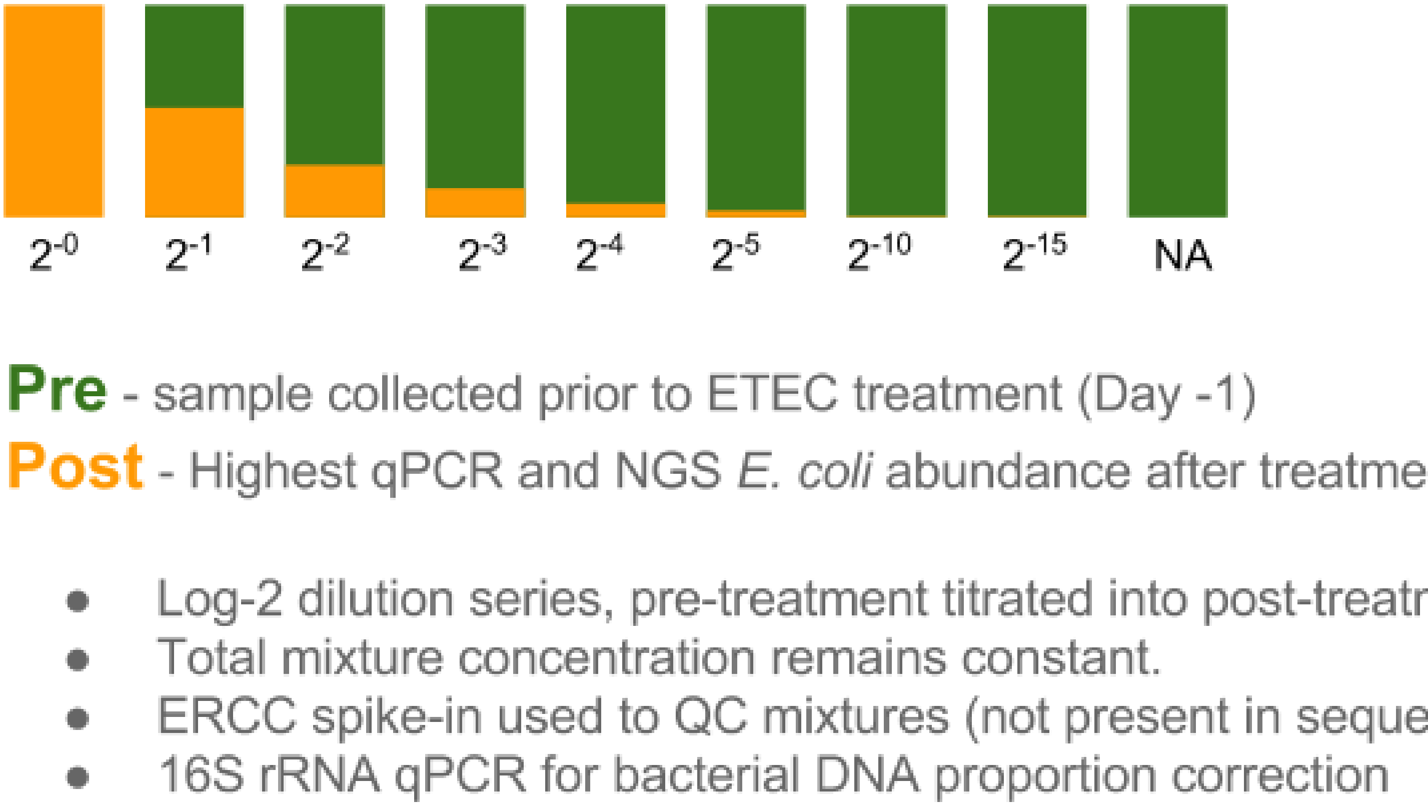
Titration Validation

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 first 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) for each of the 10 ERCC plasmids with the TaqMan Universal MasterMix II (Catalog # 4440040, ThermoFisher Waltham, MA USA).

Inorder to account for differences in the proportion of bacterial DNA in the pre and post-treatment samples used to generate the two-sample titrations, 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 tripplicate 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.



Pre - sample collected prior to ETEC treatment (Day -1)

Post - Highest qPCR and NGS *E. coli* abundance after treatme

- Log-2 dilution series, pre-treatment titrated into post-treatr
- Total mixture concentration remains constant.
- ERCC spike-in used to QC mixtures (not present in seque
- 16S rRNA qPCR for bacterial DNA proportion correction

Figure 3.1: Schematic of two sample titration process with the log_2 titrations used in the study.

Sample Processing Workflow

The resulting 45 two-sample titrations 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 seperate sample indexing PCR. A total of 192 PCRs were run including four PCR replicates of each of the 45 mixtures and 12 no template controls @ref(fig:experimental_design). After the initial PCR and clean-up the 192 PCRs were split into technical replicates and sent to two laboratories for the remainder of the library prepartion 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) then pooled and sequenced on a single Illumina MiSeq (Illumina Inc., San Diego, CA) run.

Library Preparation and Sequencing

The 16S PCR targeted the V34 region, Bakt_341F and Bakt_806R (Klindworth et al. 2012). The V34 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 included additional overhang adapter seuqences to facilitate library preparation (5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGC-CTACGGGNGGCWGCAG - 3') and reverse primers (GTCTCGTGGGCTCGGAGATGTGTATAAGA-GACAGGACTACHVGGGTATCTAATCC).

The reaction was performed according to the Illumina protocol using the KAPA HiFI HotStart ReadyMix reagents (KAPA Biosystems, Inc. Wilmington, MA). The resulting PCR product was then verified using gel electrophoresis. Next the PCR product was purified and concentration was assessed using picogreen (**Reagent Info, plate reader info**). After purification the 192 samples were indexed using the Illumina Nextera XT index kits A and D (Illumina Inc., San Diego CA). The indexed PCR products were again purified and concentration was assessed using picogreen (**Reagent Info**). The purified sample concentration was normalized using SequalPrep (ThermoFisher, Waltham MA), according to the manufactuers protocol and pooled prior to sequencing. Due to low concentration of the pooled amplicon library the follow method was used **TODO**. The library was run on a Illumina MiSeq and base calls were made using **_TODO**.

Sequence Processing

Sequence data was processed using a number of commonly used pipelines, Mothuur (Schloss et al. 2009), QIIME (Caporaso et al. 2010), DADA2 (Callahan et al. 2016), and an in-house pipeline with de-novo clustering 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)). Modifications to the SOP as a different 16S rRNA region was sequenced than the region the SOP 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 (REF). 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. 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 and *de novo* clustering with and without chimera removal (Caporaso et al. 2010). DADA2 an 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 naieve bayesian classifier. The in-house pipeline used Sickle for read trimming (Joshi NA 2011), Pandaseq (REF) 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 (REF).

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).

Normalization and Differential Abundance

- Normalization Methods
  - None
  - TSS - total sum scaling
  - CSS - cumulative sum scaling
  - Senthil's method
  - rarefaction
- Differential Abundance Methods
  - metagenomeFeatures
  - DESeq
  - EdgeR
  - Limma?

Chapter 4

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 availabliliy 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 timepoint between the qPCR and 454 data. The post-treatment timepoint with the maximum qPCR abundance value was used in this study.

Sequence QA

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 also assessed using pico green, samples with negative measured concentation values (excluding no template controls (NTC)) were also check using Qubit. Only one of the 180 samples did not successfully amplify E01JH0016 dilution 5 (postion 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

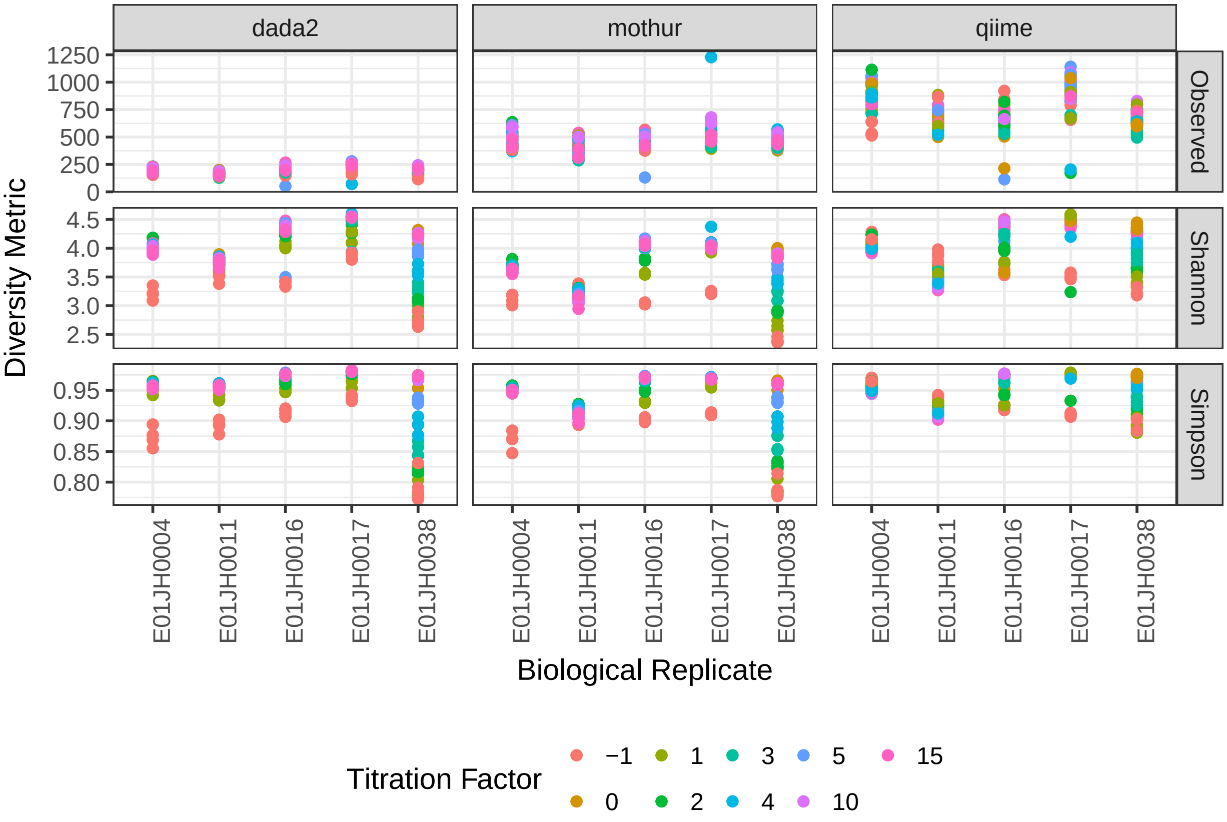


Figure 4.1: Alpha diversity metrics calculated for the different sequence processing pipelines.

Sequence Processing

Pipeline characteristics

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

- Characterization of different pipelines
  - number of clusters
    - 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)).

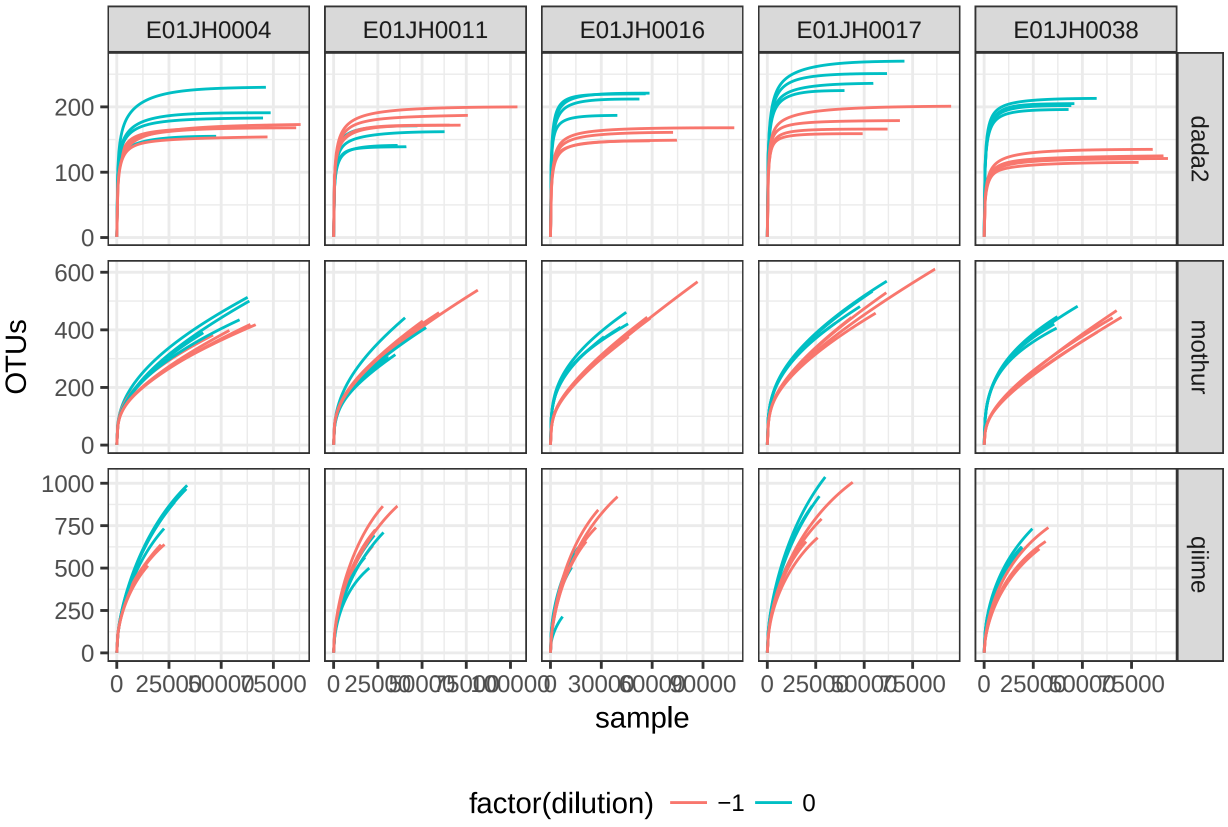
Rarefaction Curves

- 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")
```

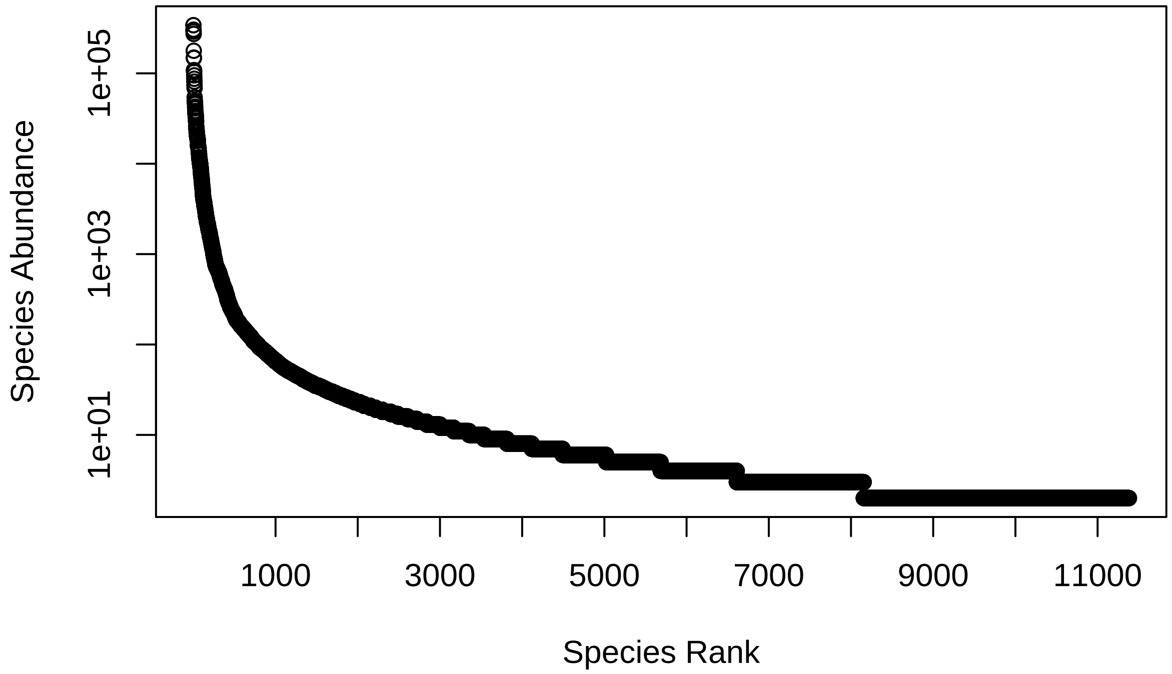


Beta Diversity - Overall sample similarity

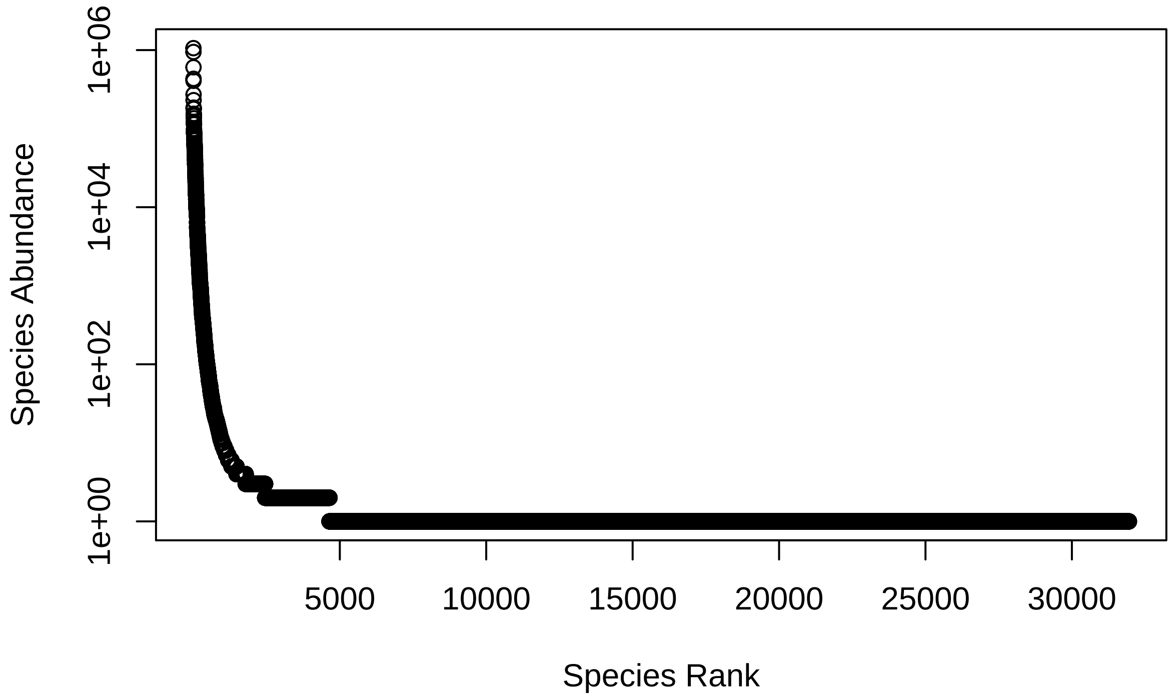
Species abundance distributions

- Characterize differences between clustering methods

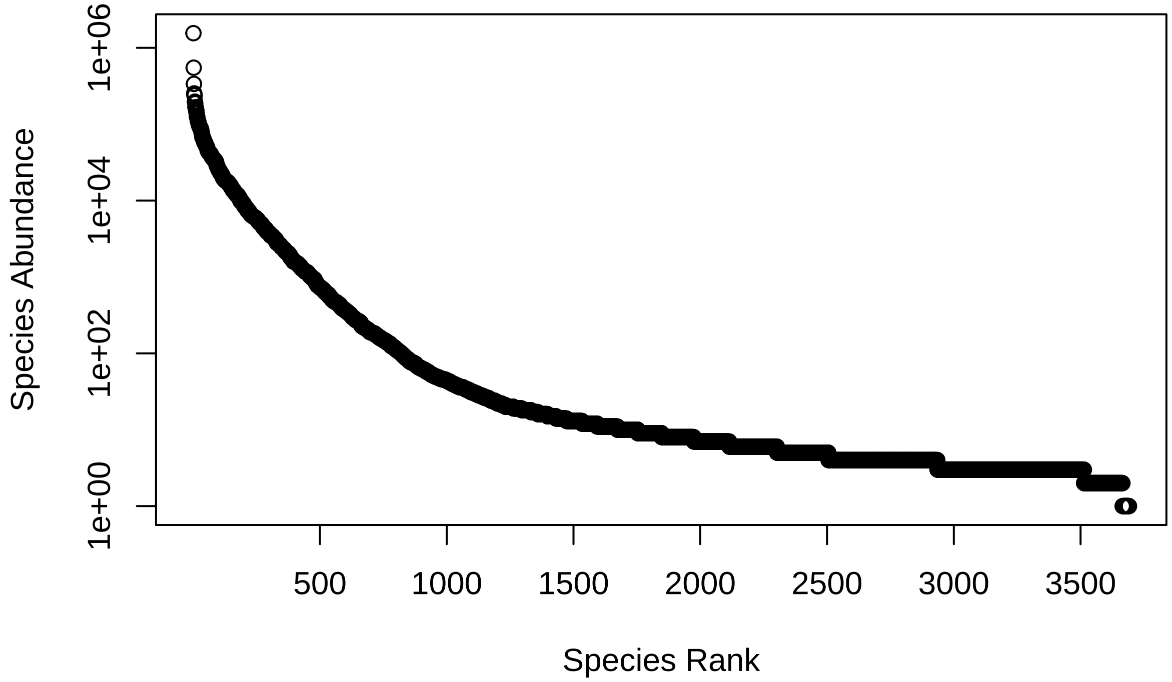
```
mrexp$qiime@assayData$counts %>% rowSums() %>% sads::rad() %>% plot()
```



```
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)
```

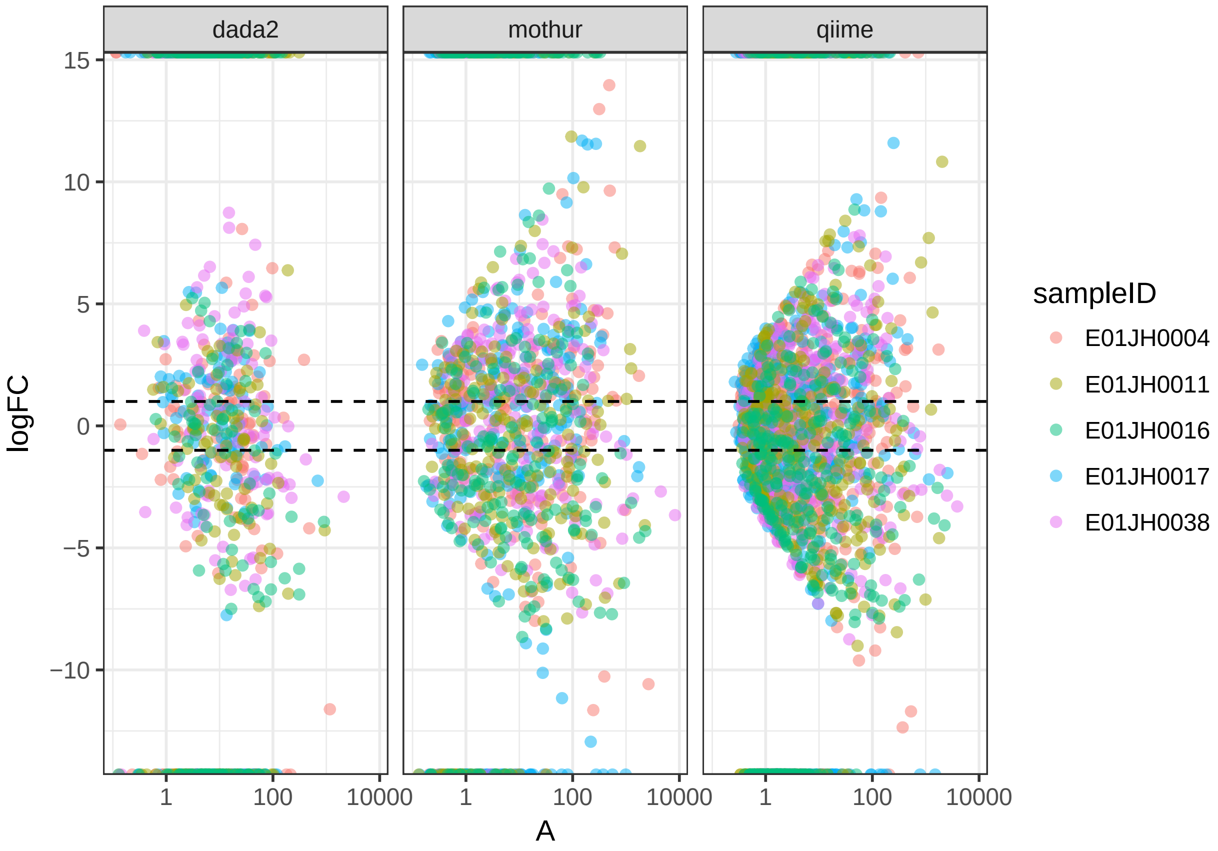


Figure 4.2: MA plot for different bioinformatic pipelines. Points at top and bottom of plots are OTUs only present in pre-treatment and post-treatment samples respectively.

Chapter 5

Normalization

- Overall Bias and Variance Summary
- features only in Pre and post unmixed
- effectiveness of high and low variance replicates
- 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 6

Differential Abundance

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

Chapter 7

Discussion

Chapter 8

Session information

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

```
##  setting  value
##  version  R version 3.3.2 (2016-10-31)
##  system   x86_64, darwin15.6.0
##  ui       RStudio (1.0.44)
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-01-24
```

```
knitr::kable(s_info$packages)
```

package	*	version	date	source
acepack		1.4.1	2016-10-29	CRAN (R 3.3.1)
ade4		1.7-5	2016-12-13	CRAN (R 3.3.2)
annotate		1.52.1	2017-01-04	Bioconductor
AnnotationDbi		1.36.0	2016-11-07	Bioconductor
AnnotationHub		2.6.4	2016-11-28	Bioconductor
ape		4.0	2016-12-01	CRAN (R 3.3.2)
assertthat		0.1	2013-12-06	CRAN (R 3.3.1)
backports		1.0.5	2017-01-18	CRAN (R 3.3.2)
base64enc		0.1-3	2015-07-28	CRAN (R 3.3.1)
bbmle	*	1.0.18	2016-02-11	CRAN (R 3.3.2)
bindr		0.1	2016-11-13	cran ((**???**))
bindrcpp	*	0.1	2016-12-11	cran ((**???**))
Biobase	*	2.34.0	2016-11-07	Bioconductor
BiocGenerics	*	0.20.0	2016-11-07	Bioconductor
BiocInstaller		1.24.0	2016-11-07	Bioconductor
BiocParallel	*	1.8.1	2016-11-07	Bioconductor
BiocStyle		2.2.1	2016-11-28	Bioconductor
biomaRt		2.30.0	2016-11-07	Bioconductor
biomformat		1.2.0	2016-11-07	Bioconductor
Biostrings	*	2.42.1	2016-12-19	Bioconductor
biovizBase		1.22.0	2016-11-07	Bioconductor
bitops		1.0-6	2013-08-17	CRAN (R 3.3.1)

package	*	version	date	source
bookdown		0.3	2016-11-28	CRAN (R 3.3.2)
broom		0.4.1	2016-06-24	CRAN (R 3.3.1)
BSgenome		1.42.0	2016-11-07	Bioconductor
caTools		1.17.1	2014-09-10	CRAN (R 3.3.1)
checkmate		1.8.2	2016-11-02	CRAN (R 3.3.2)
cluster		2.0.5	2016-10-08	CRAN (R 3.3.2)
codetools		0.2-15	2016-10-05	CRAN (R 3.3.2)
colorspace		1.3-2	2016-12-14	CRAN (R 3.3.2)
dada2	*	1.2.1	2017-01-04	Bioconductor
data.table		1.10.0	2016-12-03	CRAN (R 3.3.2)
DBI		0.5-1	2016-09-10	CRAN (R 3.3.1)
DESeq	*	1.26.0	2016-11-28	Bioconductor
devtools	*	1.12.0	2016-06-24	CRAN (R 3.3.1)
dichromat		2.0-0	2013-01-24	CRAN (R 3.3.1)
digest		0.6.11	2017-01-03	CRAN (R 3.3.2)
dplyr	*	0.5.0.9000	2017-01-17	Github (hadley/dplyr@165b
DT	*	0.2	2016-08-09	CRAN (R 3.3.1)
ensembldb		1.6.2	2016-11-18	Bioconductor
evaluate		0.10	2016-10-11	CRAN (R 3.3.1)
forcats	*	0.2.0	2017-01-23	CRAN (R 3.3.2)
foreach	*	1.4.3	2015-10-13	CRAN (R 3.3.1)
foreign		0.8-67	2016-09-13	CRAN (R 3.3.2)
Formula		1.2-1	2015-04-07	CRAN (R 3.3.1)
gdata		2.17.0	2015-07-04	CRAN (R 3.3.1)
genefilter		1.56.0	2016-11-07	Bioconductor
geneplotter		1.52.0	2016-11-07	Bioconductor
GenomeInfoDb	*	1.10.2	2017-01-04	Bioconductor
GenomicAlignments	*	1.10.0	2016-11-07	Bioconductor
GenomicFeatures		1.26.2	2016-12-19	Bioconductor
GenomicFiles		1.10.3	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)
gplots		3.0.1	2016-03-30	CRAN (R 3.3.1)
gridExtra		2.2.1	2016-02-29	CRAN (R 3.3.1)
gtable		0.2.0	2016-02-26	CRAN (R 3.3.1)
gtools		3.5.0	2015-05-29	CRAN (R 3.3.1)
GUILDS		1.3	2016-09-26	CRAN (R 3.3.2)
haven		1.0.0	2016-09-23	CRAN (R 3.3.1)
highr		0.6	2016-05-09	CRAN (R 3.3.1)
Hmisc		4.0-2	2016-12-31	CRAN (R 3.3.2)
hms		0.3	2016-11-22	CRAN (R 3.3.2)
htmlTable		1.8	2017-01-03	CRAN (R 3.3.2)
htmltools		0.3.5	2016-03-21	CRAN (R 3.3.1)
htmlwidgets		0.8	2016-11-09	CRAN (R 3.3.2)
httpuv		1.3.3	2015-08-04	CRAN (R 3.3.1)
httr		1.2.1	2016-07-03	CRAN (R 3.3.1)
hwriter		1.3.2	2014-09-10	CRAN (R 3.3.1)
igraph		1.0.1	2015-06-26	CRAN (R 3.3.1)
interactiveDisplayBase		1.12.0	2016-11-07	Bioconductor
IRanges	*	2.8.1	2016-11-18	Bioconductor
iterators		1.0.8	2015-10-13	CRAN (R 3.3.1)

package	*	version	date	source
jsonlite		1.2	2016-12-31	CRAN (R 3.3.2)
KernSmooth		2.23-15	2015-06-29	CRAN (R 3.3.2)
knitr	*	1.15.1	2016-11-22	CRAN (R 3.3.2)
labeling		0.3	2014-08-23	CRAN (R 3.3.1)
lattice	*	0.20-34	2016-09-06	CRAN (R 3.3.2)
latticeExtra		0.6-28	2016-02-09	CRAN (R 3.3.1)
lazyeval		0.2.0	2016-06-12	CRAN (R 3.3.1)
limma	*	3.30.7	2016-12-19	Bioconductor
locfit	*	1.5-9.1	2013-04-20	CRAN (R 3.3.1)
lubridate		1.6.0	2016-09-13	CRAN (R 3.3.1)
magrittr	*	1.5	2014-11-22	CRAN (R 3.3.1)
markdown		0.7.7	2015-04-22	CRAN (R 3.3.1)
MASS		7.3-45	2016-04-21	CRAN (R 3.3.2)
Matrix	*	1.2-8	2017-01-20	CRAN (R 3.3.2)
matrixStats		0.51.0	2016-10-09	CRAN (R 3.3.1)
memoise		1.0.0	2016-01-29	CRAN (R 3.3.1)
metagenomeSeq	*	1.16.0	2016-11-07	Bioconductor
mgcv		1.8-16	2016-11-07	CRAN (R 3.3.2)
mgtst	*	0.1.0	2016-09-02	local
mime		0.5	2016-07-07	CRAN (R 3.3.1)
mnormt		1.5-5	2016-10-15	CRAN (R 3.3.1)
modelr	*	0.1.0	2016-08-31	cran ((**???**))
multtest		2.30.0	2016-11-07	Bioconductor
munsell		0.4.3	2016-02-13	CRAN (R 3.3.1)
nlme		3.1-130	2017-01-24	CRAN (R 3.3.2)
nnet		7.3-12	2016-02-02	CRAN (R 3.3.2)
numDeriv		2016.8-1	2016-08-27	CRAN (R 3.3.1)
permute	*	0.9-4	2016-09-09	CRAN (R 3.3.1)
phyloseq	*	1.19.1	2017-01-04	Bioconductor
plyr		1.8.4	2016-06-08	CRAN (R 3.3.1)
png		0.1-7	2013-12-03	cran ((**???**))
poilog		0.4	2016-11-15	local
ProjectTemplate	*	0.7	2016-08-11	CRAN (R 3.3.1)
psych		1.6.12	2017-01-08	CRAN (R 3.3.2)
purrr	*	0.2.2	2016-06-18	CRAN (R 3.3.1)
R6		2.2.0	2016-10-05	CRAN (R 3.3.1)
RColorBrewer	*	1.1-2	2014-12-07	CRAN (R 3.3.1)
Rcpp	*	0.12.9	2017-01-14	cran ((**???**))
RcppParallel		4.3.20	2016-08-16	CRAN (R 3.3.2)
RCurl		1.95-4.8	2016-03-01	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 ((**???**))
reshape2		1.4.2	2016-10-22	CRAN (R 3.3.1)
rhdf5		2.18.0	2016-11-07	Bioconductor
rmarkdown		1.3	2016-12-21	CRAN (R 3.3.2)
rpart		4.1-10	2015-06-29	CRAN (R 3.3.2)
rprojroot		1.2	2017-01-16	CRAN (R 3.3.2)
Rqc	*	1.8.0	2016-11-07	Bioconductor
Rsamtools	*	1.26.1	2016-11-07	Bioconductor
RSQLite		1.1-2	2017-01-08	CRAN (R 3.3.2)
rstudioapi		0.6	2016-06-27	CRAN (R 3.3.1)
rtracklayer		1.34.1	2016-11-07	Bioconductor

package	*	version	date	source
rvest		0.3.2	2016-06-17	cran ((**???**))
S4Vectors	*	0.12.1	2016-12-19	Bioconductor
sads	*	0.3.1	2016-05-13	CRAN (R 3.3.2)
sapkotaUtils	*	0.0.0.9000	2016-09-19	Github (nate-d-olson/sapko
savR	*	1.12.0	2016-11-07	Bioconductor
scales		0.4.1	2016-11-09	CRAN (R 3.3.2)
shiny		1.0.0	2017-01-12	CRAN (R 3.3.2)
ShortRead	*	1.32.0	2016-11-07	Bioconductor
stringi		1.1.2	2016-10-01	CRAN (R 3.3.1)
stringr	*	1.1.0	2016-08-19	CRAN (R 3.3.1)
SummarizedExperiment	*	1.4.0	2016-11-07	Bioconductor
survival		2.40-1	2016-10-30	CRAN (R 3.3.1)
tibble	*	1.2	2016-08-26	CRAN (R 3.3.1)
tidyr	*	0.6.1	2017-01-10	CRAN (R 3.3.2)
tidyverse	*	1.1.0	2017-01-20	CRAN (R 3.3.2)
VariantAnnotation		1.20.2	2016-12-19	Bioconductor
vegan	*	2.4-2	2017-01-17	CRAN (R 3.3.2)
VGAM		1.0-3	2017-01-11	CRAN (R 3.3.2)
withr		1.0.2	2016-06-20	CRAN (R 3.3.1)
XML		3.98-1.5	2016-11-10	CRAN (R 3.3.2)
xml2		1.1.0	2017-01-07	CRAN (R 3.3.2)
xtable		1.8-2	2016-02-05	CRAN (R 3.3.1)
XVector	*	0.14.0	2016-11-07	Bioconductor
yaml		2.1.14	2016-11-12	CRAN (R 3.3.2)
zlibbioc		1.20.0	2016-11-07	Bioconductor

Chapter 9

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.

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.

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.

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.