Pipeline Characterization

Nate Olson

2017-04-06

The sequencing dataset was processed using three bioinformatic pipelines. The following analysis provides and overview of the resulting dataset; count table characteristics, sample coverage, overall similarity between pre- and post-treatment samples.

Count Table Characteristics

Total number of features for all samples and count table sparsity for the bioinformatic pipelines. The expectation is that this dataset will be more sparse relative to other datasets due to the redundant nature of the samples where 35 of the 45 samples are derived directly from the other 10 samples and that there are four PCR replicates for each sample. Sparsity 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. See individual pipeline reports for which steps reads are excluded from the datasets. QIIME pipeline has the highest filter rate while the highest number of features per sample.

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	38358	0.98	53775 (1265-87806)	0.4 (0.35-0.62)
qiime	11385	0.94	25254 (517-46897)	0.7 (0.62-0.97)

Sample Coverage

Rarefaction curves are commonly used to demonstrate how well a community has been sampled (**REF**). The slope of the curve decreases as sample coverage increases. Rarefaction curve for sequence inference has flattened out indicating the community has been fully sampled. Whereas the curves have not reached their asymptotes for *de novo* and open reference clustering indicating that the community has not been fully sampled. Comparison of the individual sample rarefaction curves to the rarefaction curves after aggregating counts for the four replicates indicates that the rarefaction curves of the individual for the *de novo* and open-reference clustering more representative of the diversity in the four replicates combined than the features generated using the sequence inference method.

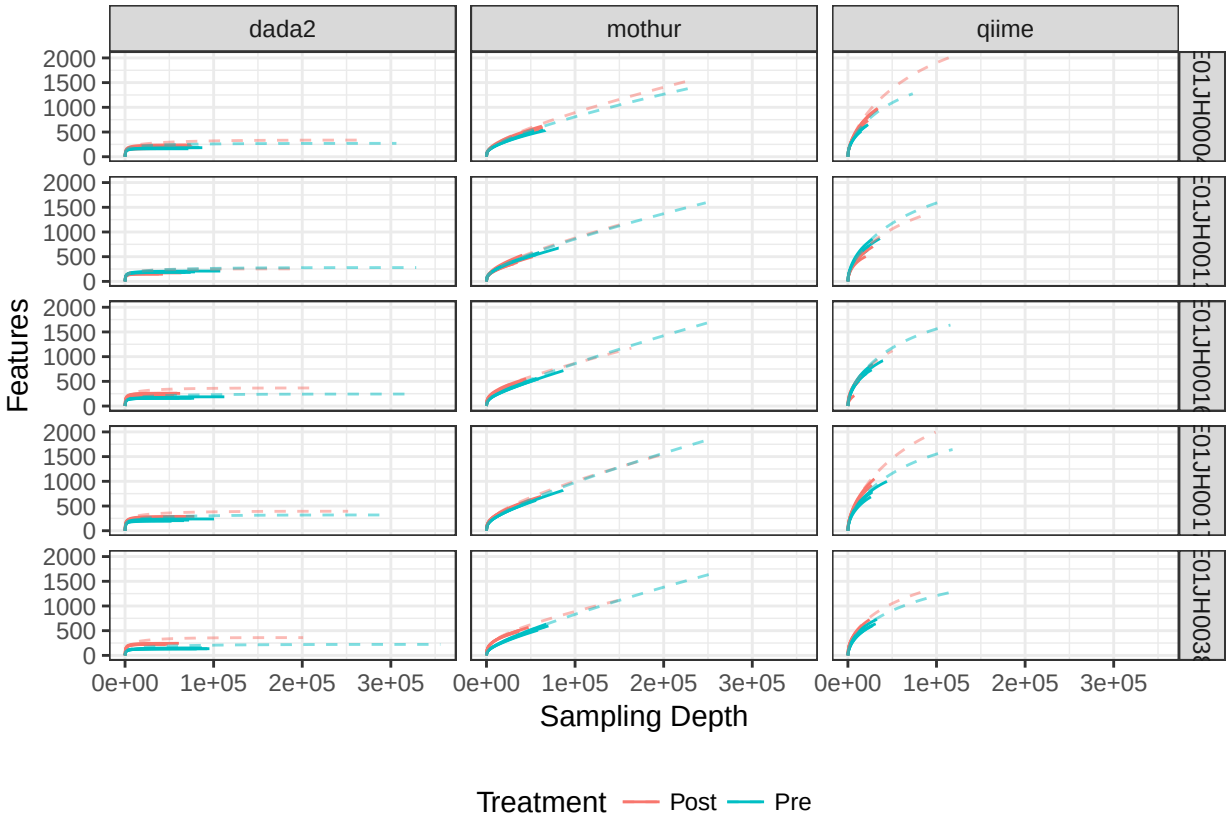
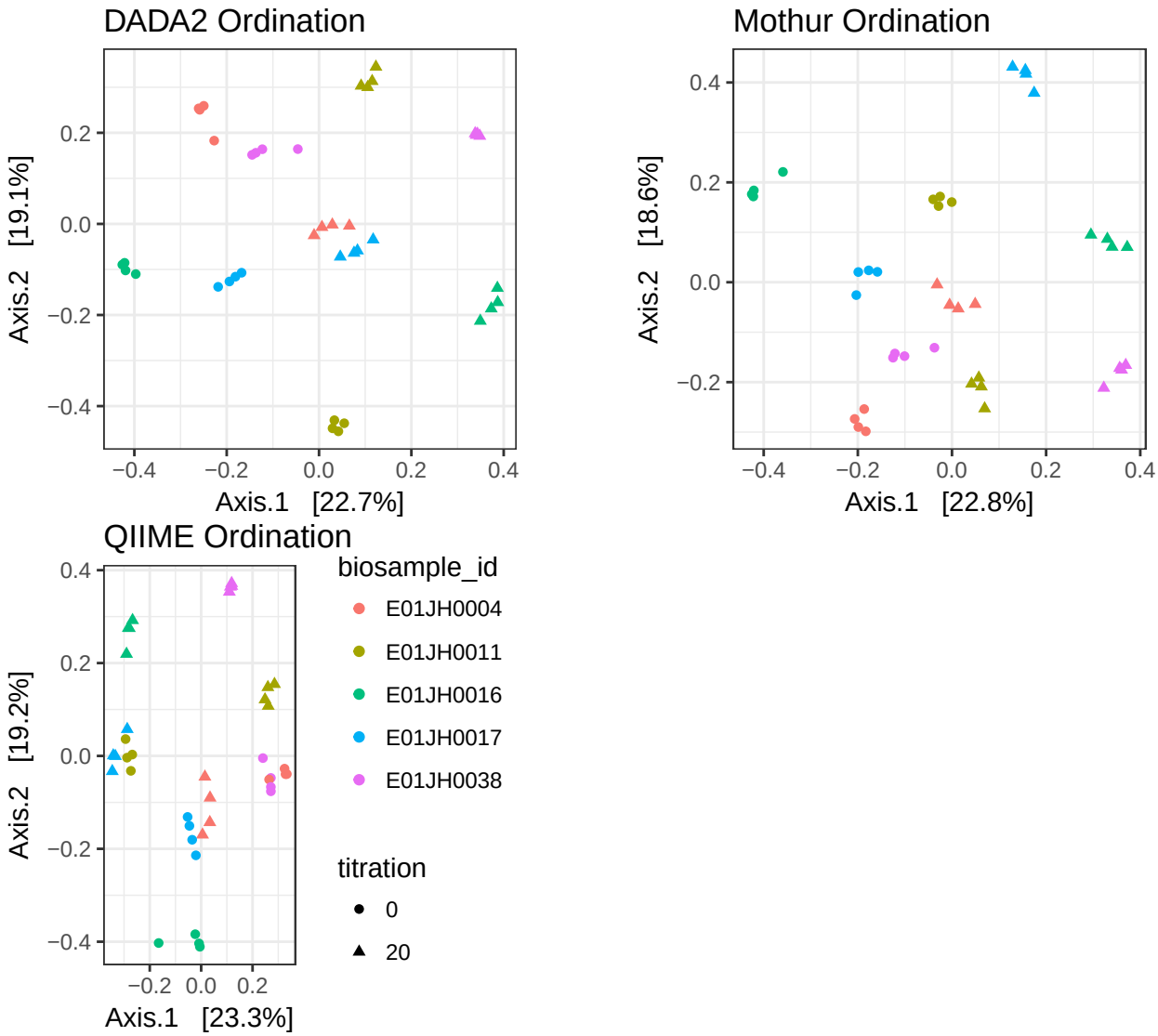


Figure 1: Rarefaction curves for the unmixed pre- and post-treatment samples. The solid lines represent individual PCR replicates and the dashed lines the pooled replicates.

Similarity Between Pre- and Post-Treatment Samples

Replicates group together.
Overall samples group together more by treatment status than biological sample @ref(fig:ordPlots). **TODO** Clean-up formatting



Conclusion

Session information

```
##  setting  value
##  version  R version 3.3.3 (2017-03-06)
##  system   x86_64, darwin15.6.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
```

```
##  tz       America/New_York
##  date     2017-04-06
```

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	1.10.3	2017-03-28	Bioconductor
GenomicAlignments	1.10.1	2017-03-28	Bioconductor
GenomicRanges	1.26.4	2017-03-28	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	2.8.2	2017-03-28	Bioconductor
knitr	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.3)
limma	3.30.13	2017-03-28	Bioconductor
locfit	1.5-9.1	2013-04-20	CRAN (R 3.3.1)
Matrix	1.2-8	2017-01-20	CRAN (R 3.3.3)
metagenomeSeq	1.16.0	2016-11-07	Bioconductor
modelr	0.1.0	2016-08-31	cran (@0.1.0)
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	1.1.0	2017-03-22	CRAN (R 3.3.2)
readxl	0.1.1	2016-03-28	cran (@0.1.1)
Rqc	1.8.0	2016-11-07	Bioconductor
Rsamtools	1.26.1	2016-11-07	Bioconductor
S4Vectors	0.12.2	2017-03-28	Bioconductor
sads	0.3.1	2016-05-13	CRAN (R 3.3.2)
savR	1.12.0	2016-11-07	Bioconductor
ShortRead	1.32.1	2017-03-28	Bioconductor
stringr	1.2.0	2017-02-18	CRAN (R 3.3.2)
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	0.14.1	2017-03-28	Bioconductor