Pipeline Characterization

Nate Olson

2017-03-08

Introduction

The sequencing dataset was processed using three (**TODO** Fourth pipeline - Popline) bioinformatic pipelines. The following analysis provides and overview of the resulting datasets;

1. based on the count table characteristics (number of features and sparsness)
2. sample coverage (rarefaction curves),
3. variability and range of observed log-fold change values (MA plots beween pre- and post-treatment samples)
4. overall similarity between pre- and post-treatment samples (beta-diversity ordination)
5. Feature specificity - how much information for the titration values is provided by the unmixed samples

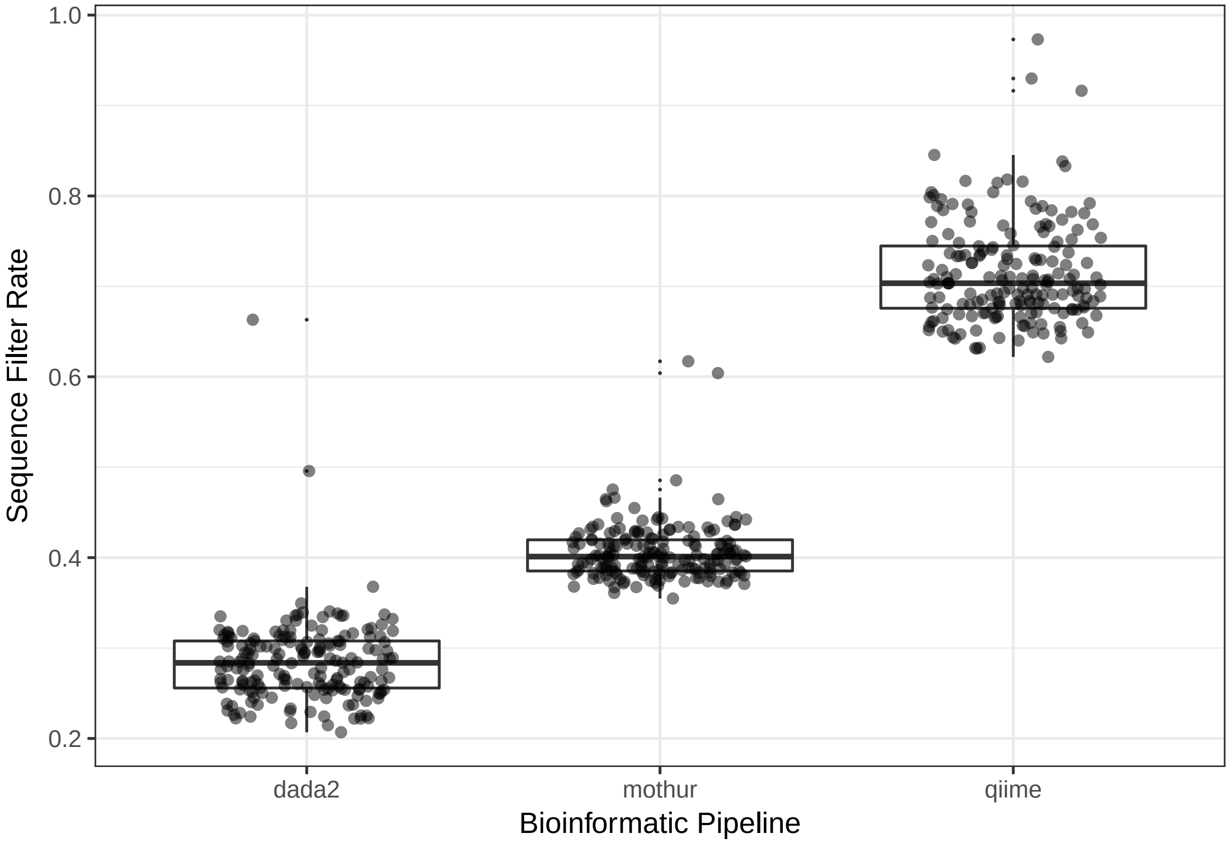
Count Table Characteristics

Total number of features for all samples and count table sparsity for the bioinformatic pipelines.

NOTE Sparsity lower for De-novo clustering (QIIME) than sequence inference (DADA2) even though DADA2 has fewer total features.

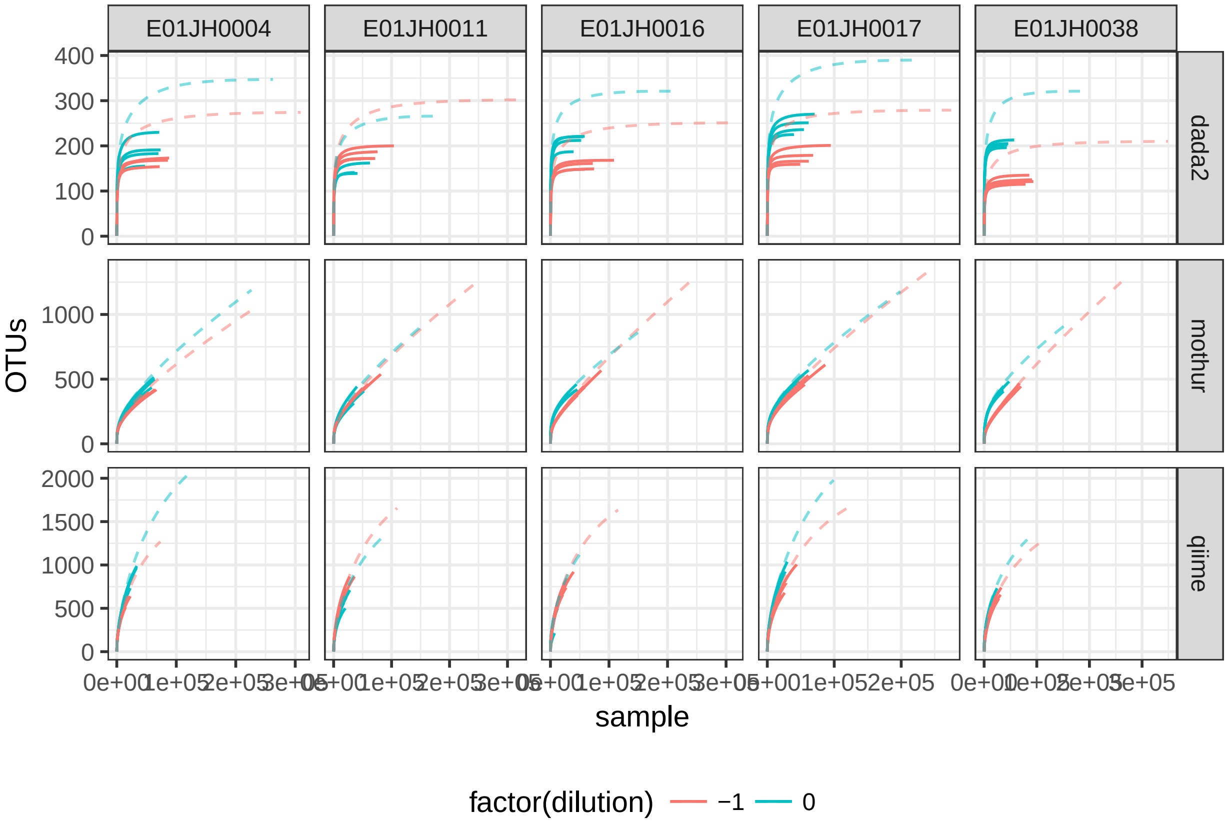
pipe	features	sparsity
dada2	3691	0.9470424
mothur	31948	0.9852194
qiime	11381	0.9349667

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.



Sample Coverage

Sample coverage show using rarefaction curves. 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 flatten out 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.



Range of logFC values

To get a general idea regarding the overall range of logFC values to expect for the titrations provided an MA plot comparing the pre- and post-treatment samples.

TODO Comment on the relationship between the observed logFC values and titrations

Similarity between pre- and post-treatment samples

Beta diversity ordination plot

Feature Specificity

see Mix Specific Feature report see Pre- Post-Specific report

Conclusion

Session information

```
##  setting  value
##  version  R version 3.3.2 (2016-10-31)
##  system   x86_64, darwin15.6.0
##  ui       unknown
```

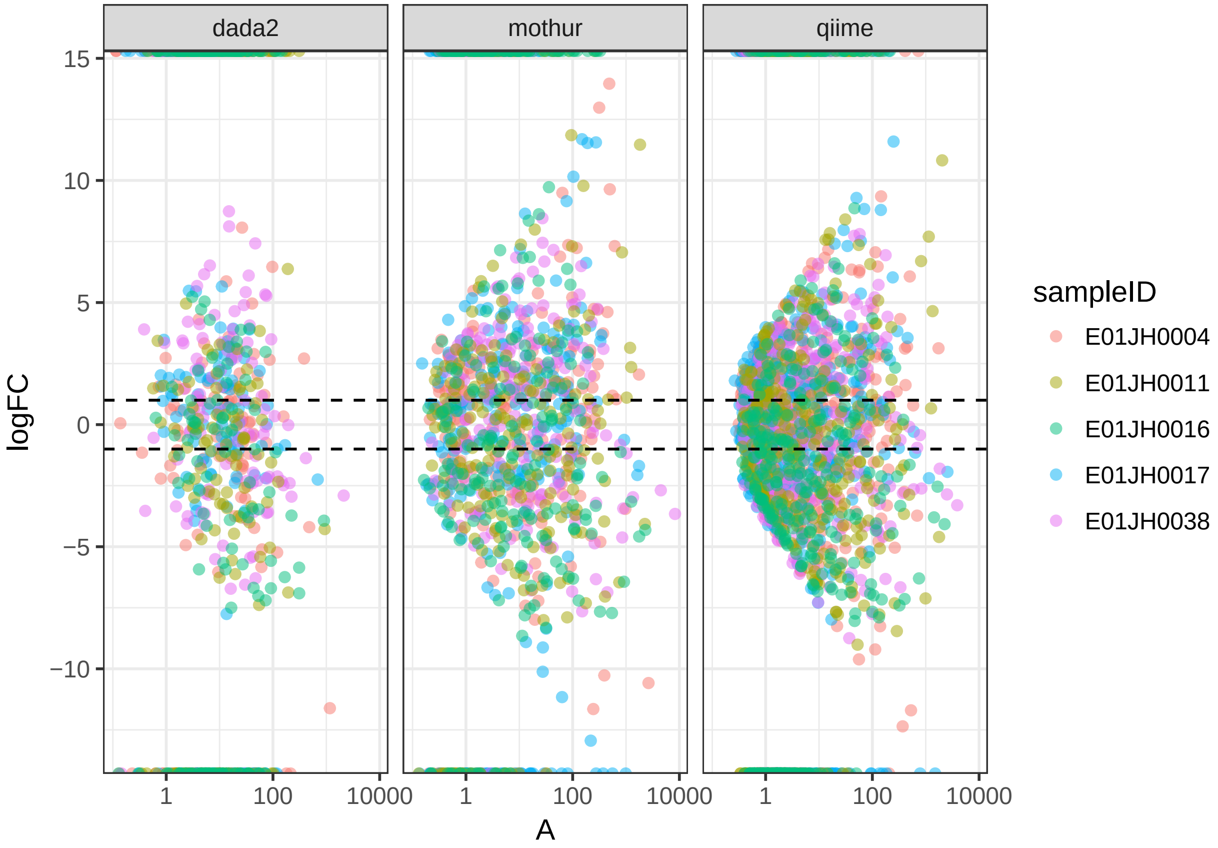


Figure 1: MA plot for different bioinformatic pipelines. Points at top and bottom of plots are OTUs only present in pre-treatment and post-treatment samples respectively.

```
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-03-08
```

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)
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	1.0.0	2016-08-03	CRAN (R 3.3.1)
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	0.12.1	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