Metabarcoding of bacteria in coral samples treated with or without PMA

16S rRNA gene metabarcoding

This file outlines the steps, beginning with raw data from the sequencing facility, that were followed to produce the 16S rRNA metabarocoding files used in R for data analysis. All intermediate .qza and .qzv files are provided. This data processing was completed in QIIME2 v2022.11 by Ashley Dungan in Nov. 2022.

Import data

Data from the Walter and Eliza Hall Institute (WEHI) came as paired-end, demultiplexed (.fastq) files with primers and overhang sequences still attached still attached. Raw files are stored on Mediaflux, OneDrive, and an external harddrive. File names were adjusted and gzipped to satisfy QIIME2 requirments (+++_L(0-9)(0-9)(0-9)_R(1-2)_001.fastq.gz).

Samples were sequenced across two seperate MiSeq runs - each run was processed separately through dada2 then merged prior to classification.

In [ ]:
```
# Things to do when terminal is opened up, at the start of each session
byobu -S qiime2
source activate qiime2-2022.11

rm -r ~/data/tmp
mkdir ~/data/tmp
export TMPDIR=~/data/tmp
echo $TMPDIR #should print ~/data/tmp
```

In [ ]:
```
mkdir ~/data/bact_viability/demux

qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path ~/data/bact_viability/raw_data/viability \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path ~/data/bact_viability/demux/demuxed.qza

qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path ~/data/bact_viability/raw_data/Laura \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path ~/data/bact_viability/demux/demuxed2.qza
```

Remove primers

For this experiment, amplicons were amplified with 784F-1061R primers targeting the V5V6 region of the bacterial 16S rRNA gene. The reads come back from the sequencer with primers attached, which are removed before denoising using **cutadapt**. With cutadapt, the sequence

specified and all bases prior are trimmed; most sequences were trimmed at ~50 base pairs (bp). An error rate of 0.15 was used to maximize the number of reads that the primers were removed from while excluding nonspecific cutting. Any untrimmed read was discarded.

In [ ]:
```
qiime cutadapt trim-paired \
--i-demultiplexed-sequences ~/data/bact_viability/demux/demuxed.qza \
--p-front-f AGGATTAGATACCCTGGTA \
--p-front-r CRRCACGAGCTGACGAC \
--p-discard-untrimmed \
--p-error-rate 0.15 \
--output-dir ~/data/bact_viability/trim \
--verbose

qiime cutadapt trim-paired \
--i-demultiplexed-sequences ~/data/bact_viability/demux/demuxed2.qza \
--p-front-f AGGATTAGATACCCTGGTA \
--p-front-r CRRCACGAGCTGACGAC \
--p-discard-untrimmed \
--p-error-rate 0.15 \
--output-dir ~/data/bact_viability/trim2 \
--verbose
```

Create and interpret sequence quality data

I created a viewable summary file to evaluate the data quality. The visualization was downloaded and viewed at https://view.qiime2.org.

In [ ]:
```
qiime demux summarize \
--i-data  ~/data/bact_viability/trim/trimmed_sequences.qza \
--o-visualization  ~/data/bact_viability/trim/trimmed_sequences.qzv

qiime demux summarize \
--i-data  ~/data/bact_viability/trim2/trimmed_sequences.qza \
--o-visualization  ~/data/bact_viability/trim2/trimmed_sequences.qzv
```

Quality control of data

Raw, trimmed sequences were quality assessed using the **dada2** plugin within QIIME 2 (Callahan et al., 2016). This plugin utilizes denoising by producing fine-scale resolution through amplicon sequencing variants (**ASVs**), resolving differences of as little as a single nucleotide (Callahan et al., 2016). Its workflow consists of filtering, dereplication, reference-free chimera detection, and paired-end reads merging (Callahan et al., 2016). Using dada2, I performed this error correction and quality filtering to generate a feature table. I used the "pseudo" method used to pool samples for denoising. The pseudo-pooling method is used to approximate pooling of samples. In this approach, samples are denoised independently once, ASVs detected in at least 2 samples are recorded, and samples are denoised independently a second time, but this time with prior knowledge of the recorded ASVs and thus higher sensitivity to those ASVs.

Median quality score for raw reads dropped below 35 at 235/209 and 183/127 bp for the forward and reverse reads, respectively. However, I find that being conservative and truncating

significantly less than these values by 20 bp provides higher quality data with more reads retained.

In [ ]:
```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs ~/data/bact_viability/trim/trimmed_sequences.qza \
--p-trunc-len-f 205 \
--p-trunc-len-r 163 \
--p-n-threads 0 \
--p-pooling-method 'pseudo' \
--output-dir ~/data/bact_viability/trim/dada2out2 \
--verbose

qiime dada2 denoise-paired \
--i-demultiplexed-seqs ~/data/bact_viability/trim2/trimmed_sequences.qza \
--p-trunc-len-f 209 \
--p-trunc-len-r 127 \
--p-n-threads 0 \
--p-pooling-method 'pseudo' \
--output-dir ~/data/bact_viability/trim2/dada2out2 \
--verbose
```

Generate summary files

The metadata file was verified using the plugin for Google Sheets, keemei. All summary files were downloaded and viewed at https://view.qiime2.org. Where appropriate, csv files were downloaded from view.qiime2.org for further data exploration. A .fasta file with all representative sequences was downloaded.

In [ ]:
```
qiime feature-table tabulate-seqs \
--i-data ~/data/bact_viability/trim/dada2out2/representative_sequences.qza \
--o-visualization ~/data/bact_viability/trim/dada2out2/16s_rep_seqs.qzv \
--verbose

qiime metadata tabulate \
--m-input-file ~/data/bact_viability/trim/dada2out2/denoising_stats.qza \
--o-visualization ~/data/bact_viability/trim/dada2out2/16s_denoising_stats.qzv \
--verbose

qiime feature-table tabulate-seqs \
--i-data ~/data/bact_viability/trim2/dada2out2/representative_sequences.qza \
--o-visualization ~/data/bact_viability/trim2/dada2out2/16s_rep_seqs.qzv \
--verbose

qiime metadata tabulate \
--m-input-file ~/data/bact_viability/trim2/dada2out2/denoising_stats.qza \
--o-visualization ~/data/bact_viability/trim2/dada2out2/16s_denoising_stats.qzv \
--verbose
```

For Run 1, all 317 samples sent for 16S rRNA gene metabarcoding were successfully sequenced. Prior to quality control there were 10.2M sequences (min=5, max=239,452, average=32,332). An average of 63.46% of reads (6.5M, min=1, max=146,367, average=20,519) were kept after filtering, denoising, and chimera removals. 95 samples contained fewer than 2800 reads - coral samples in this group (n=68) were sequenced again on Run 2.

For Run 2, all samples had reads except Mu5.Sept22 (there were no files for this one).

Merge datasets

DADA2 is the last step in this analysis that needs to be run on a per sequencing run basis. Now I'll merge the artifacts generated by those two commands.

In [ ]:
```
qiime feature-table merge \
--i-tables ~/data/bact_viability/trim/dada2out2/table.qza \
--i-tables ~/data/bact_viability/trim2/dada2out2/table.qza \
--o-merged-table ~/data/bact_viability/merge_table.qza

qiime feature-table merge-seqs \
--i-data ~/data/bact_viability/trim/dada2out2/representative_sequences.qza \
--i-data ~/data/bact_viability/trim2/dada2out2/representative_sequences.qza \
--o-merged-data ~/data/bact_viability/merged_representative_sequences.qza
```

Identifier-based filtering

I then used identifier-based filtering to retain only those samples associated with this experiment. A TRUE/FALSE column was added to the metadata file, which was used to select the samples to continue processing through the QIIME2 pipeline.

In [ ]:
```
qiime feature-table filter-samples \
--i-table ~/data/bact_viability/merge_table.qza \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--p-where "[Ashley]='TRUE'" \
--o-filtered-table ~/data/bact_viability/ashley_table.qza

qiime feature-table filter-samples \
--i-table ~/data/bact_viability/merge_table.qza \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--p-where "[Laura]='TRUE'" \
--o-filtered-table ~/data/bact_viability/laura_table.qza

qiime feature-table summarize \
--i-table ~/data/bact_viability/ashley_table.qza \
--m-sample-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/ashley_table.qzv \
--verbose

qiime feature-table summarize \
--i-table ~/data/bact_viability/laura_table.qza \
--m-sample-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/laura_table.qzv \
--verbose
```

Assign taxonomy

I first trained the Silva database to classify bacterial 16S rRNA reads for the variable 5 and 6 (V5V6) regions. After training the classifier, each ASV was classified to the highest resolution based on this classifier. I then generated a viewable summary files of the taxonomic assignments, which was downloaded and viewed at https://view.qiime2.org.

n_jobs = 1 This script was run using all available cores

In [ ]:
```
qiime feature-classifier classify-sklearn \
--i-classifier ~/data/silva_138_16s_v5v6_classifier.qza \
--i-reads ~/data/bact_viability/merged_representative_sequences.qza \
--p-n-jobs 1 \
--output-dir ~/data/bact_viability/taxonomy/ \
--verbose

qiime metadata tabulate \
--m-input-file ~/data/bact_viability/taxonomy/classification.qza \
--o-visualization ~/data/bact_viability/taxonomy/taxonomy.qzv \
--verbose
```

Build a phylogenetic tree

"A phylogenetic tree was produced in QIIME 2 by aligning ASVs using the PyNAST method (Caporaso et al., 2010) with mid-point rooting."

The next lines of code do the following:

1. Perform an alignment on the representative sequences.
2. Mask highly variable regions of the alignment.
3. Generate a phylogenetic tree.
4. Apply mid-point rooting to the tree.

In [ ]:
```
mkdir ~/data/bact_viability/tree

qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences ~/data/bact_viability/merged_representative_sequences.qza \
--o-alignment ~/data/bact_viability/tree/aligned_16s_rep_seqs.qza \
--o-masked-alignment ~/data/bact_viability/tree/masked_aligned_16s_rep_seqs.qza \
--o-tree ~/data/bact_viability/tree/16s_unrooted_tree.qza \
--o-rooted-tree ~/data/bact_viability/tree/16s_rooted_tree.qza \
--p-n-threads 1 \
--verbose
```

Filter

We filter out reads classifed as mitochondria and chloroplast. Unassigned ASVs were retained. We then generated a viewable summary file of the new table to see the effect of filtering. According to QIIME developer Nicholas Bokulich, low abundance filtering (i.e. removing ASVs containing very few sequences) is not necessary under the ASV model.

In [ ]:
```
qiime taxa filter-table \
--i-table ~/data/bact_viability/ashley_table.qza \
--i-taxonomy ~/data/bact_viability/taxonomy/classification.qza \
--p-exclude Mitochondria,Chloroplast \
--o-filtered-table ~/data/bact_viability/16S_ashley_table.qza \
--verbose

qiime taxa filter-table \
--i-table ~/data/bact_viability/laura_table.qza \
```

```
--i-taxonomy ~/data/bact_viability/taxonomy/classification.qza \
--p-exclude Mitochondria,Chloroplast \
--o-filtered-table ~/data/bact_viability/16S_laura_table.qza \
--verbose

qiime feature-table summarize \
--i-table ~/data/bact_viability/16S_ashley_table.qza \
--m-sample-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/16S_ashley_table.qzv \
--verbose

qiime feature-table summarize \
--i-table ~/data/bact_viability/16S_laura_table.qza \
--m-sample-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/16S_laura_table.qzv \
--verbose
```

Exporting data for analysis in R

"ASV, taxonomy, metadata and phylogenetic tree files were imported into R and combined into a phyloseq object (McMurdie and Holmes, 2013)."

You need to export your ASV table, taxonomy table, and tree file for analyses in R. Many file formats can be accepted.

Export unrooted tree as .nwk format as required for the R package 'phyloseq.'

Create a BIOM table with taxonomy annotations. A FeatureTable[Frequency] artifact will be exported as a BIOM v2.1.0 formatted file. Then export BIOM as TSV

Export Taxonomy as TSV

In [ ]:

```
mkdir ~/data/bact_viability/R_ashley

qiime tools export \
  --input-path ~/data/bact_viability/tree/16s_unrooted_tree.qza \
  --output-path ~/data/bact_viability/R_ashley

qiime tools export \
  --input-path ~/data/bact_viability/16S_ashley_table.qza \
  --output-path ~/data/bact_viability/R_ashley

biom convert \
-i ~/data/bact_viability/R_ashley/feature-table.biom \
-o ~/data/bact_viability/R_ashley/asv-table.tsv \
--to-tsv

qiime tools export \
--input-path ~/data/bact_viability/taxonomy/classification.qza \
--output-path ~/data/bact_viability/R_ashley

mkdir ~/data/bact_viability/R_laura

qiime tools export \
  --input-path ~/data/bact_viability/tree/16s_unrooted_tree.qza \
  --output-path ~/data/bact_viability/R_laura

qiime tools export \
  --input-path ~/data/bact_viability/16S_laura_table.qza \
```

```
  --output-path ~/data/bact_viability/R_laura

biom convert \
-i ~/data/bact_viability/R_laura/feature-table.biom \
-o ~/data/bact_viability/R_laura/asv-table.tsv \
--to-tsv

qiime tools export \
--input-path ~/data/bact_viability/taxonomy/classification.qza \
--output-path ~/data/bact_viability/R_laura
```

TSV files were opened in excel, headers adjusted for analysis in R, and the data ordered consistently, i.e. the order of the ASVs in the taxonomy table rows was the same order of ASVs in the columns of the ASV table. The taxonomy file was cleaned up by leaving blank cells where the level of classification was not completely resolved.

Downstream Analyses on QIIME 2

Rarefaction curves

I generated rarefaction curves to determine whether the samples have been sequenced deeply enough to capture all the community members.

In [ ]:

```
mkdir ~/data/bact_viability/downstream_ashley

qiime diversity alpha-rarefaction \
--i-table ~/data/bact_viability/16S_ashley_table.qza \
--p-max-depth 20000 \
--p-min-depth 500 \
--p-steps 40 \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/downstream_ashley/16s_alpha_rarefaction_500.
--verbose

mkdir ~/data/bact_viability/downstream_laura

qiime diversity alpha-rarefaction \
--i-table ~/data/bact_viability/16S_laura_table.qza \
--i-phylogeny ~/data/bact_viability/tree/16s_rooted_tree.qza \
--p-max-depth 15000 \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/downstream_laura/16s_alpha_rarefaction.qzv \
--verbose
```

Barchart

Create bar charts to compare the relative abundance of ASVs across samples. You can interactively view the barplot on view.qiime2.org.

In [ ]:

```
qiime taxa barplot \
--i-table ~/data/bact_viability/16S_ashley_table.qza \
--i-taxonomy ~/data/bact_viability/taxonomy/classification.qza \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/downstream_ashley/barchart.qzv \
--verbose
```

```
qiime taxa barplot \
--i-table ~/data/bact_viability/16S_laura_table.qza \
--i-taxonomy ~/data/bact_viability/taxonomy/classification.qza \
--m-metadata-file ~/data/bact_viability/metadata.tsv \
--o-visualization ~/data/bact_viability/downstream_laura/barchart.qzv \
--verbose
```

Basic visualizations and statistics

Alphadiversity

The following is taken directly from the Moving Pictures tutorial and adapted for this data set. QIIME 2's diversity analyses are available through the **q2-diversity** plugin, which supports computing alpha and beta diversity metrics, applying related statistical tests, and generating interactive visualizations. We'll first apply the core-metrics-phylogenetic method, which rarefies a FeatureTable[Frequency] to a user-specified depth, computes several alpha and beta diversity metrics, and generates principle coordinates analysis (PCoA) plots using Emperor for each of the beta diversity metrics.

The metrics computed by default are:

- Alpha diversity
- Shannon's diversity index (a quantitative measure of community richness)
- Observed OTUs (a qualitative measure of community richness)
- Faith's Phylogenetic Diversity (a qualitiative measure of community richness that incorporates phylogenetic relationships between the features)
- Evenness (or Pielou's Evenness; a measure of community evenness)
- Beta diversity
- Jaccard distance (a qualitative measure of community dissimilarity)
- Bray-Curtis distance (a quantitative measure of community dissimilarity)
- unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
- weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)

An important parameter that needs to be provided to this script is --p-sampling-depth, which is the even sampling (i.e. rarefaction) depth that you determined above. Because most diversity metrics are sensitive to different sampling depths across different samples, this script will randomly subsample the counts from each sample to the value provided for this parameter. For example, if you provide --p-sampling-depth 500, this step will subsample the counts in each sample without replacement so that each sample in the resulting table has a total count of 500. If the total count for any sample(s) are smaller than this value, those samples will be dropped from the diversity analysis. Choosing this value is tricky. To pick this value, you should review the 16s_alpha_rarefaction.qzv and 16s_table.qzv files created above. I selected a value of 2500 in order to retain as many sequences per sample while excluding as few samples as possible. Retained 570,000 (8.28%) features in 228 (71.92%) samples at the specifed sampling depth.

In [ ]:

```
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny ~/data/bact_viability/tree/16s_rooted_tree.qza \
```

```
  --i-table ~/data/bact_viability/16S_ashley_table.qza \
  --p-sampling-depth 1467 \
  --m-metadata-file ~/data/bact_viability/metadata.tsv \
  --output-dir ~/data/bact_viability/adiversity_ashley

qiime diversity core-metrics-phylogenetic \
  --i-phylogeny ~/data/bact_viability/tree/16s_rooted_tree.qza \
  --i-table ~/data/bact_viability/16S_laura_table.qza \
  --p-sampling-depth 1400 \
  --m-metadata-file ~/data/bact_viability/metadata.tsv \
  --output-dir ~/data/bact_viability/adiversity_laura
```