Metabarcoding of bacteria in the short-beaked echidna *Tachyglossus aculeatus* samples

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 v2021.11 by Ashley Dungan.

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).

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

qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path ~/data/Echidna/data/16S/ \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path ~/data/Echidna/demux/v4demuxed.qza
```

Remove primers

For this experiment, amplicons were amplified following the Earths Microbiome protocol with 515F (Caporaso)–806R (Caporaso) primers targeting the v4 region of the 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/Echidna/demux/v4demuxed.qza \
--p-front-f GTGCCAGCMGCCGCGGTAA \
--p-front-r GGACTACHVGGGTWTCTAAT \
--p-discard-untrimmed \
--p-error-rate 0.15 \
--output-dir ~/data/Echidna/trim \
--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/Echidna/trim/trimmed_sequences.qza \
--o-visualization  ~/data/Echidna/trim/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.

Median quality score for raw reads dropped below 35 at 253 and 208 bp for the forward and reverse reads, respectively. However, I find that being conservative and truncating significantly less than these values provides higher quality data with more reads retained.

In [ ]:
```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs ~/data/Echidna/trim/trimmed_sequences.qza \
--p-trunc-len-f 213 \
--p-trunc-len-r 168 \
--p-n-threads 0 \
--output-dir ~/data/Echidna/trim/dada2out \
--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/Echidna/trim/dada2out/representative_sequences.qza \
--o-visualization ~/data/Echidna/trim/dada2out/16s_rep_seqs.qzv \
--verbose

qiime metadata tabulate \
--m-input-file ~/data/Echidna/trim/dada2out/denoising_stats.qza \
--o-visualization ~/data/Echidna/trim/dada2out/16s_denoising_stats.qzv \
--verbose
```

Train Silva v138 classifier for the 16S rRNA V4 (515-806) region

The newest version of the Silva database (v138) was trained to classify bacterial 16S rRNA reads for V4 region. Reference files silva-138-99-seqs.qza and silva-138-99-tax.qza were downloaded from https://www.arb-silva.de/download/archive/.

16S rRNA reads for the V4 region were extracted.

In [ ]:
```
qiime feature-classifier extract-reads \
--i-sequences ~/data/silva-138-99-seqs.qza \
--p-f-primer GTGCCAGCMGCCGCGGTAA \
--p-r-primer GGACTACHVGGGTWTCTAAT \
--o-reads ~/data/silva_138_16s_515-806.qza \
--verbose
```

The classifier was then trained using a naive Bayes algorithm.

In [ ]:
```
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads ~/data/silva_138_16s_515-806.qza \
--i-reference-taxonomy ~/data/silva-138-99-tax.qza \
--o-classifier ~/data/silva_138_16s_515-806_classifier.qza \
--verbose
```

Assign taxonomy

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

qiime feature-classifier classify-sklearn \
--i-classifier ~/data/silva_138_16s_515-806_classifier.qza \
--i-reads ~/data/Echidna/trim/dada2out/representative_sequences.qza \
--p-n-jobs 1 \
--output-dir ~/data/Echidna/taxonomy/ \
--verbose

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

Build a phylogenetic tree

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/Echidna/tree

qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences ~/data/Echidna/trim/dada2out/representative_sequences.qza \
--o-alignment ~/data/Echidna/tree/aligned_16s_rep_seqs.qza \
--o-masked-alignment ~/data/Echidna/tree/masked_aligned_16s_rep_seqs.qza \
--o-tree ~/data/Echidna/tree/16s_unrooted_tree.qza \
--o-rooted-tree ~/data/Echidna/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.

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 QIIME 2 pipeline.

In [ ]:
```
qiime taxa filter-table \
--i-table ~/data/Echidna/trim/dada2out/table.qza \
--i-taxonomy ~/data/Echidna/taxonomy/classification.qza \
--p-exclude Mitochondria,Chloroplast \
--o-filtered-table ~/data/Echidna/16S_table.qza \
--verbose

qiime feature-table summarize \
--i-table ~/data/Echidna/16S_table.qza \
--m-sample-metadata-file ~/data/Echidna/metadata.tsv \
--o-visualization ~/data/Echidna/16s_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/Echidna/R

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

qiime tools export \
  --input-path ~/data/Echidna/16S_table.qza \
  --output-path ~/data/Echidna/R

biom convert \
-i ~/data/Echidna/R/feature-table.biom \
-o ~/data/Echidna/R/asv-table.tsv \
--to-tsv

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

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. The max depth setting was set to 14926 based on the median number of reads per sequence in the 16s_table.qzv file.

mkdir ~/data/Echidna/downstream

qiime diversity alpha-rarefaction \ --i-table ~/data/Echidna/16S_table.qza \ --i-phylogeny ~/data/Echidna/tree/16s_rooted_tree.qza \ --p-max-depth 14926 \ --m-metadata-file ~/data/Echidna/metadata.tsv \ --o-visualization ~/data/Echidna/downstream/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.

qiime taxa barplot \ --i-table ~/data/Echidna/16S_table.qza \ --i-taxonomy ~/data/Echidna/taxonomy/classification.qza \ --m-metadata-file ~/data/Echidna/metadata.tsv \ --o-visualization ~/data/Echidna/downstream/barchart.qzv \ --verbose