Microbial biomarkers as indicators of sperm quality in an insect

QIIME2 Code for 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 v2021.8 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 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). All 16S rRNA gene raw files (515F/806R) from this run were processed together through dada2 to maximize QC and filtering. Afterwhich, ASVs were classified using the most recent version of the Silva database (v138).

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

Reference files silva-138-99-seqs.qza and silva-138-99-tax.qza were downloaded from https://www.arb-silva.de/download/archive/.

Using the original Earths Microbiome primers (515F/806R), 16S rRNA reads for the V4-5 region were extracted.

The primers used for this study (with WEHI overhang) are: -515f WEHI 5' - GTGACCTATGAACTCAGGAGTCGTGCCAGCMGCCGCGGTAA (Caporaso et al., 2011) -806r WEHI 5' - CTGAGACTTGCACATCGCAGCGGACTACHVGGGTWTCTAAT (Caporaso et al., 2011)

In [ ]:
```
# Things to do when terminal is opened up, at the start of each session
byobu -S qiime2
source activate qiime2-2021.8

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

qiime feature-classifier extract-reads \
--i-sequences silva-138-99-seqs.qza \
--p-f-primer GTGYCAGCMGCCGCGGTAA \
--p-r-primer GACTACHVGGGTATCTAATCC \
--o-reads silva_138_16s_341-805.qza \
--verbose
```

The classifier was then trained using a naive Bayes algorithm.

In [ ]:
```
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads silva_138_16s_341-805.qza \
--i-reference-taxonomy silva-138-99-tax.qza \
--o-classifier silva_138_16s_341-805_classifier.qza \
--verbose
```

Import raw data into QIIME2

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

qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path ~/data/cricket/rawdata/ \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--output-path ~/data/cricket/demux/v5v6demuxed.qza
```

Remove primers

These sequences had the primers attached, which were 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/cricket/demux/v5v6demuxed.qza \
--p-front-f GTGYCAGCMGCCGCGGTAA \
--p-front-r GACTACHVGGGTATCTAATCC \
--p-discard-untrimmed \
--p-error-rate 0.15 \
--output-dir ~/data/cricket/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/cricket/trim/trimmed_sequences.qza \
--o-visualization  ~/data/cricket/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 224 and 175 bp for the forward and reverse reads, respectively. I find that truncating at 20 bp less than these values provides higher quality data with more reads retained.

In [ ]:
```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs ~/data/cricket/trim/trimmed_sequences.qza \
--p-trunc-len-f 204 \
--p-trunc-len-r 155 \
--p-n-threads 0 \
--output-dir ~/data/cricket/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/cricket/trim/dada2out/representative_sequences.qza \
--o-visualization ~/data/cricket/trim/dada2out/16s_rep_seqs.qzv \
--verbose

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

qiime feature-table summarize \
--i-table ~/data/cricket/trim/dada2out/table.qza \
--m-sample-metadata-file ~/data/cricket/metadata.tsv \
--o-visualization ~/data/cricket/trim/dada2out/16s_table.qzv \
--verbose
```

After QC, 13 cricket samples (+ 6/7 extraction blanks and 2/3 NTPCRs) had fewer than 1000 reads: 128D_F17 39SV_F13 79_F13 92_F17 19D_F12 128SV_F17 128_F17 86_F17 28SV_F12 19AG_F11 82AG_F13 75SV_F13 73SV_F13

The 7th extraction blank had 1074 reads and the 3rd NTPCR had 2009 reads. Overall this data set and the controls are of a very high quality.

An average of 61.5% of reads were kept after dada2. Data for the output from dada2 can be found in the "denoising_stats_cricket_data" excel spreadsheet.

All ASVs sequences can be found in the file "sequences.fasta". The 50 most abundant ASVs were run through BLASTn to identify a closest match. That information can be found in the excel spreadsheet "Most_abundant_ASVs." In total, 2342 ASVs were identified, most are 251 bp.

In the 158 samples, the mean number of reads was 9,034 (min=0, median=8,326, max=30,689).

Assign taxonomy

The newest version of the Silva database (v138) was trained to classify bacterial 16S rRNA reads for the variable 4 and 5 (V4V5) regions. Then, 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_341-805_classifier.qza \
--i-reads ~/data/cricket/trim/dada2out/representative_sequences.qza \
--p-n-jobs 1 \
--output-dir ~/data/cricket/taxonomy/ \
--verbose
```

```
qiime metadata tabulate \
--m-input-file ~/data/cricket/taxonomy/classification.qza \
--o-visualization ~/data/cricket/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/cricket/tree

qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences ~/data/cricket/trim/dada2out/representative_sequences.qza \
--o-alignment ~/data/cricket/tree/aligned_16s_rep_seqs.qza \
--o-masked-alignment ~/data/cricket/tree/masked_aligned_16s_rep_seqs.qza \
--o-tree ~/data/cricket/tree/16s_unrooted_tree.qza \
--o-rooted-tree ~/data/cricket/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/cricket/trim/dada2out/table.qza \
--i-taxonomy ~/data/cricket/taxonomy/classification.qza \
--p-exclude mitochondria,chloroplast \
--o-filtered-table ~/data/cricket/trim/dada2out/16s_table_filtered.qza \
--verbose

qiime feature-table summarize \
--i-table ~/data/cricket/trim/dada2out/16s_table_filtered.qza \
--m-sample-metadata-file ~/data/cricket/metadata.tsv \
--o-visualization ~/data/cricket/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).

In [ ]:
```
mkdir ~/data/cricket/R

qiime tools export \
```

```
  --input-path ~/data/cricket/tree/16s_unrooted_tree.qza \
  --output-path ~/data/cricket/R

qiime tools export \
  --input-path ~/data/cricket/trim/dada2out/16s_table_filtered.qza \
  --output-path ~/data/cricket/R

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

qiime tools export \
--input-path ~/data/cricket/taxonomy/classification.qza \
--output-path ~/data/cricket/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 20000.

mkdir ~/data/cricket/downstream

qiime diversity alpha-rarefaction \ --i-table ~/data/cricket/trim/dada2out/16s_table_filtered.qza \ --i-phylogeny ~/data/cricket/tree/16s_rooted_tree.qza \ --p-max-depth 20000 \ --p-min-depth 500 \ --p-steps 40 \ --m-metadata-file ~/data/cricket/metadata.tsv \ --o-visualization ~/data/cricket/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/cricket/trim/dada2out/16s_table_filtered.qza \ --i-taxonomy ~/data/cricket/taxonomy/classification.qza \ --m-metadata-file ~/data/cricket/metadata.tsv \ --o-visualization ~/data/cricket/downstream/barchart.qzv \ --verbose