**Comparison of Bioinformatics Tools and Procedures for Food Processing Surface Microbiota Characterization Using 16S RNA Sequences**

**QIIME 2 Workflow Document-created by Sarah E.Daly**

**Relevant files are available in NCBI # . 16S rRNA sequencing data collected from July 2022 to May 2023. Sequenced by Novogene. See Materials & Methods for more information. Commands to type or paste in the terminal are in** Courier New.

# QIIME 2 Workflow

# Starting QIIME 2

1. Machine: 96 cores; 512GB RAM. I would suggest using the VNC interface rather than PuTTY for easier viewing and file transfer.
2. Cornell BioHPC users only: you need to set up conda environment first and then deactivate when done: <https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=388#c>

export LC\_ALL=en\_US.utf-8

export LANG=en\_US.utf-8

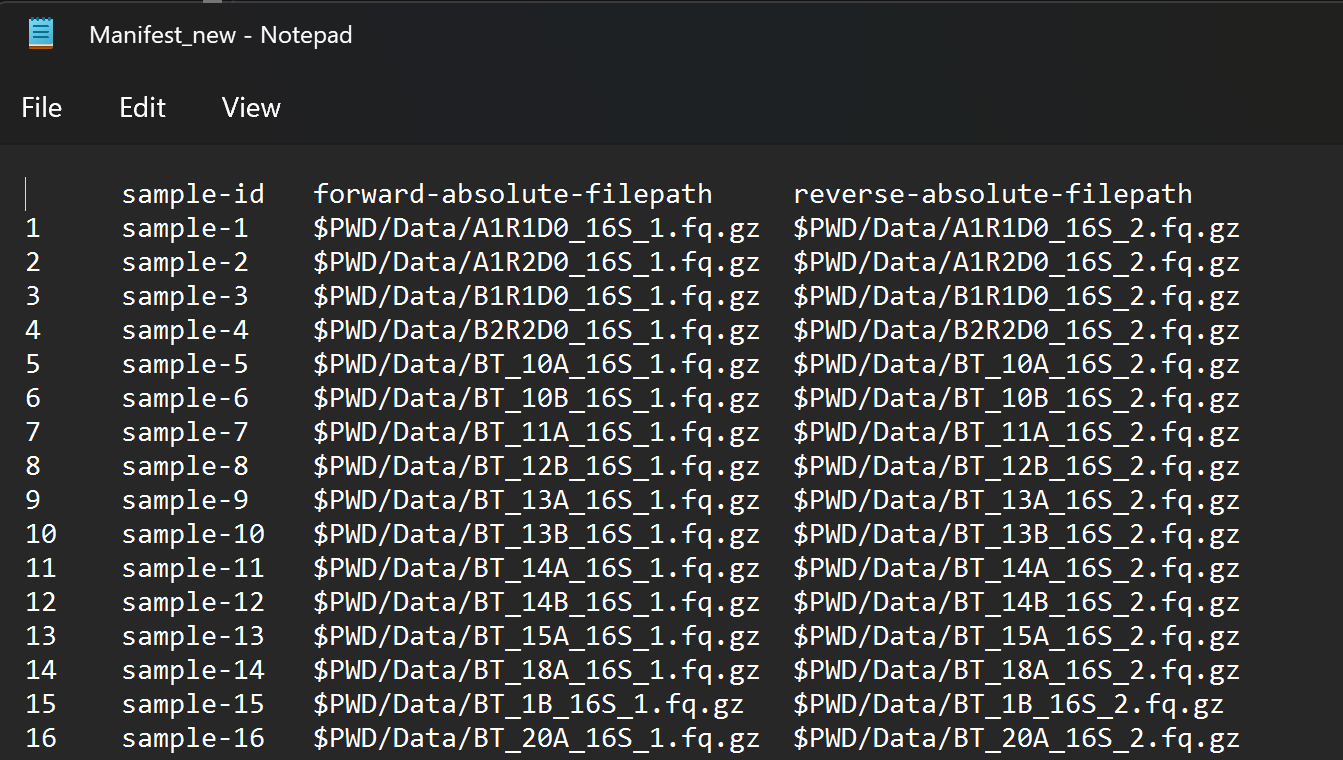
source /programs/miniconda3/bin/activate qiime2-2022.2

1. Then type “qiime” into command line to open QIIME 2. (I did this in qiime2022.2)
2. Type “source tab-qiime” so you can tab complete file names
3. Type in “qiime --help" to see the list of plug-ins (similar to “packages” in R) installed (dada2 is the most important we are using)

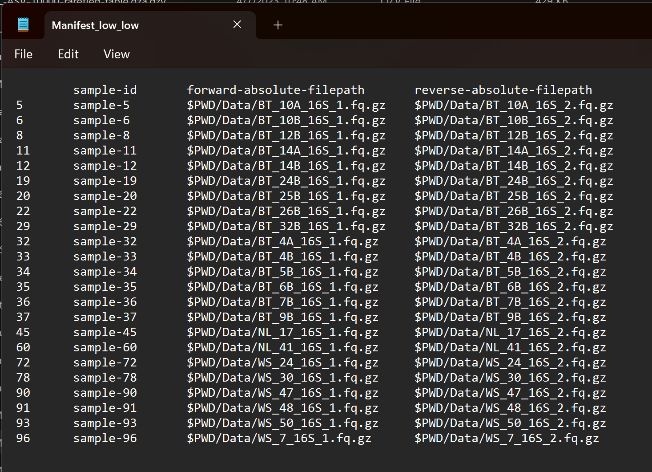
# Importing Sequence Files to QIIME 2

1. Download or copy the 16S sequence folder from NCBI into your personal directory.
2. Filter for the demultiplexed gz files (Search in windows for files that end in : \*\_1.fq.gz & \*\_2.fq.gz) and then drag them in your folder. Transfer these files into a new folder marked “Data” (there should be no subfolders).
3. Transfer “Data” to your home directory in QIIME 2. Check that all files have the same extension (.fq.gz) or QIIME will not import it.
4. In Qiime, change your working directory (cd ‘directory’) to “Data” and check the compression of each file (should be all Oks). 
   1. **for** f **in** \*.fq.gz; **do** gunzip -t -v $f; **done**
5. Go back to your working directory (cd ..)
6. The file names do not match the basic QIIME 2 defaults. Therefore, we need to create a Manifest File with the name and path of all sequence files. Go to R Studio on your home computer and open Manifest File Generator.Rmd.
7. Read the directions and create a Manifest file. Export it as a .csv and open. Save it as a .txt. Note the file must be in this format.

***Troubleshooting:*** Find and replace $PWD with the path to the file if you have issues with QIIME 2 opening it.

****

1. In QIIME 2, convert the sequence files to a .qza artifiact & view basic sequence and sample info. It will generate an .html that you can view in a web browser. You should perform this command in your working directory.

**

qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path Manifest\_SFWS.txt --output-path paired-end-demux-SFWS.qza --input-format PairedEndFastqManifestPhred33V2

qiime demux summarize --i-data paired-end-demux-SFWS.qza --o-visualization paired-end-demux-SFWS.qza

qiime tools view paired-end-demux-SFWS.qza.qzv

qiime tools validate paired-end-demux-SFWS.qza

Result paired-end-demux.qza appears to be valid at level=max.

\*\*Forward and Reverse read “Total” must match. If not, check each sample and remove samples that don’t match\*\*

# Importing Metadata in QIIME 2

1. Got to the .csv version of your manifest file.
2. Keep sample-id column (first column must be this); you can delete other columns as you add more
3. If you need barcodes, they can be found in SampleSeq\_info in the original sequence files
4. Keep cells blank for missing data
5. Data must be categorical or numeric
6. Use simple headings; use (-) between words if you need a heading with more than 2 words
7. Copy and paste in additional metadata categories you want.
8. Extract Facility, Sampling Season, and original sample ID from this file info in the metadata file and input into basic metadata sheet.
9. Save as a .txt file and transfer to your QIIME 2 working directory
10. Visualize the metadata in your working directory.

qiime metadata tabulate --m-input-file Metadata\_SFWS.txt --o-visualization tabulated-sample-metadata.qzv

qiime tools view tabulated-sample-metadata.qzv

# Truncating and Denoising Reads

1. We need to determine at what quality ends can be trimmed and reads can still be merged without losing info. This can be somewhat subjective.
2. *Denoise* to remove and/or correct noisy reads, correct errors in marginal sequences (DADA2), remove chimeric sequences, remove singletons, join denoised paired-end
3. According to the report we have, the minimum base length of forward reads is 226 and reverse reads is 223. Generally, the sequence quality looks good.
4. There are multiple criteria for trimming
   1. **Manual trimming**: Trim if median quality score < 20

\*\*I suggest using a machine with 96 cores and letting the process run with all the cores (add the --p-n-threads 0 flag to the command). This seems to save considerable time—see below.

**ASV Method**

**Manual Trimming by Quality Score (took about half an hour to run)**

qiime dada2 denoise-paired --i-demultiplexed-seqs paired-end-demux-SFWS.qza --p-trunc-len-f 0 --p-trunc-len-r 0 --p-trunc-q 20 --p-n-threads 0 --o-table table-SFWS.qza --o-representative-sequences ASV-rep-seqs-SFWS.qza --o-denoising-stats denoising-stats-SFWS.qza --verbose

**OTU Method**

***De novo* clustering of a feature table can be performed as follows. In this example, clustering is performed at 99% identity to create 99% OTUs.**

qiime vsearch cluster-features-de-novo **\**

--i-table table-SFWS.qza **\**

--i-sequences ASV-rep-seqs-SFWS.qza **\**

--p-perc-identity 0.99 **\**

--o-clustered-table otu-table-dn-99-SFWS.qza **\**

--o-clustered-sequences out-rep-seqs-dn-99-SFWS.qza

**View output from quality score trimming:**

**#Denoising Stats**

qiime metadata tabulate --m-input-file denoising-stats-SFWS.qza --o-visualization 16S-denoising-stats-FWS.qzv

qiime tools view 16S-denoising-stats-FWS.qzv

**#Sequence length**

**#Can be used to calculate genome size**

qiime feature-table tabulate-seqs --i-data ASV-rep-seqs-SFWS.qza --o-visualization 16S-rep-seqs-asv-SFWS.qzv

qiime tools view 16S-rep-seqs-asv-SFWS.qzv

qiime feature-table tabulate-seqs --i-data out-rep-seqs-dn-99-SFWS.qza --o-visualization 16S-rep-seqs-otu-SFWS.qzv

qiime tools view 16S-rep-seqs-otu-SFWS.qzv

**#Summarize feature count**

qiime feature-table summarize --i-table table-SFWS.qza --o-visualization table\_summary-FWS.qzv

qiime tools view table\_summary-SFWS.qzv

**#Click on “Interactive” which can give insight into impact of rarefaction on sequences**

**“Feature” data = data on representative sequences**

# Classifying Sequence Taxonomy

Machine-learning-based classification methods are available through classify-sklearn. These classifiers must be *trained*, e.g., to learn which features best distinguish each taxonomic group, adding an additional step to the classification process. [Classifier training](https://docs.qiime2.org/2022.8/tutorials/feature-classifier/) is **reference database- and marker-gene-specific** and only needs to happen once per marker-gene/reference database combination; that classifier may then be re-used as many times as you like without needing to re-train!

We will train the [Naive Bayes](http://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes) classifier using [Greengenes](http://qiime.org/home_static/dataFiles.html) reference sequences and classify the representative sequences. We select classify-sklearn with a Naive Bayes classifier because it can slightly outperform other methods for the 16S rRNA gene (<https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0470-z>)

List of taxonomy databases:

<http://qiime.org/home_static/dataFiles.html>

Download a pre-made classifier here:

* <https://docs.qiime2.org/2022.11/data-resources/>
* Select [Greengenes 13\_8 99% OTUs full-length sequences](https://data.qiime2.org/2022.11/common/gg-13-8-99-nb-classifier.qza)  [in Box gg-13-8-99-nb-classifier.qza]

Greengenes

<https://greengenes.secondgenome.com/?prefix=downloads/greengenes_database/gg_13_5/>

**#Classifying sequences from the ASV table**

qiime feature-classifier classify-sklearn \

--i-classifier gg-13-8-99-nb-classifier.qza \

--i-reads ASV-rep-seqs-SFWS.qza \

--o-classification taxonomy-asv-SFWS.qza

*\*May take at least 10 minutes to run.*

**#Classifying sequences from the OUT table**

qiime feature-classifier classify-sklearn \

--i-classifier gg-13-8-99-nb-classifier.qza \

--i-reads out-rep-seqs-dn-99-SFWS.qza \

--o-classification taxonomy-otu—99-SFWS.qza

**#create a phylogeny tree**

**ASV method**

qiime phylogeny align-to-tree-mafft-fasttree \

--i-sequences ASV-rep-seqs-SFWS.qza \

--output-dir phylogeny-align-to-tree-mafft-fasttree-SFWS-asv

**OTU method**

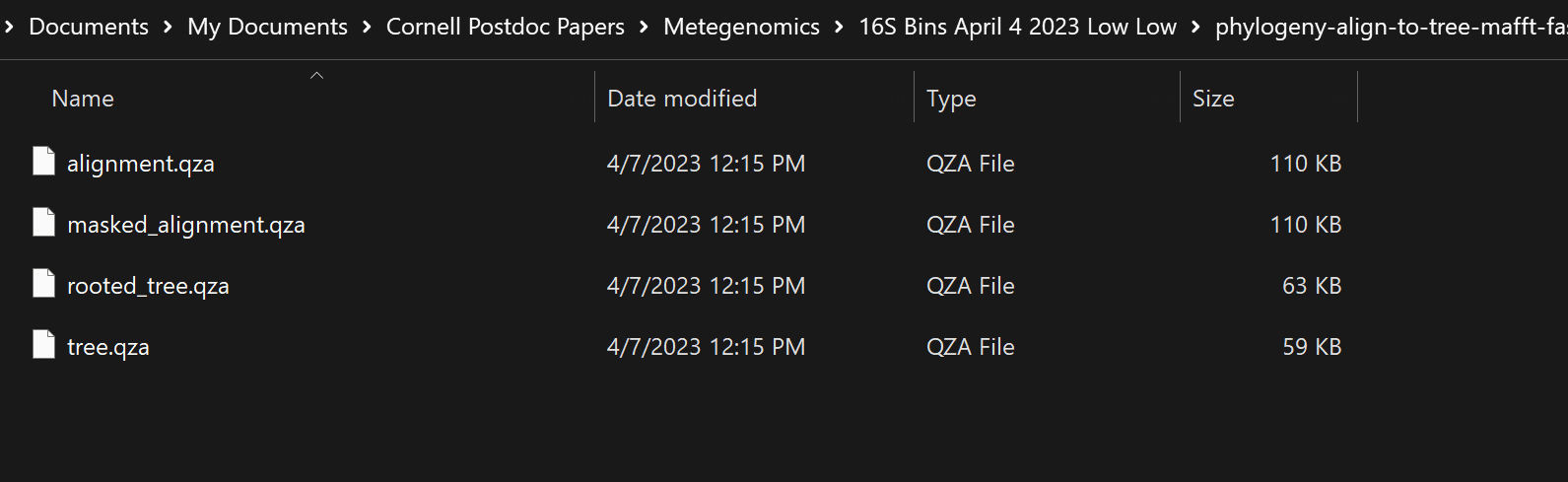
**#create a phylogeny tree**

qiime phylogeny align-to-tree-mafft-fasttree \

--i-sequences out-rep-seqs-dn-99-SFWS.qza \

--output-dir phylogeny-align-to-tree-mafft-fasttree-SFWS-otu

**“Rooted\_tree” is located in this directory:**



**\*No visual output from tree, but it can be used in R phyloseq.**

**Visualize taxonomy results in Barplots**

qiime taxa barplot --i-table table-SFWS.qza --i-taxonomy taxonomy-asv-  
SFWS.qza --m-metadata-file Metadata\_SFWS.txt--o-visualization barplotbbb-asv-SFWS.qzv

qiime taxa barplot --i-table otu-table-dn-99-SFWS.qza --i-taxonomy taxonomy-out-99-SFWS.qza --m-metadata-file Metadata\_SFWS.txt --o-visualization barplotbbb-otu-SFWS.qzv

**#Make barplot to view**

qiime tools view barplotbbb-asv-FWS.qzv

qiime tools view barplotbbb-otu-FWS.qzv

**#Remove low frequency counts & view new barplots**

qiime feature-table filter-features --i-table table-SFWS.qza --p-min-samples 2 --p-min-frequency 1500 --o-filtered-table filtered-table-SFWS-asv.qza

qiime taxa barplot --i-table filtered-table-SFWS-asv.qza --i-taxonomy taxonomy-asv-SFWS.qza --m-metadata-file Metadata\_SFWS.txt --o-visualization barplotbbb-asv-SFWS-filetered.qzv

qiime feature-table filter-features --i-table otu-table-dn-99-SFWS.qza --p-min-samples 2 --p-min-frequency 1500 --o-filtered-table filtered-table-SFWS-otu.qza

qiime taxa barplot --i-table filtered-table-SFWS-otu.qza --i-taxonomy taxonomy-otu-99-SFWS.qza --m-metadata-file Metadata\_SFWS.txt --o-visualization barplotbbb-otu-SFWS-filtered.qzv

Cornell BioHPC: Deactivate the qiime environment after you are done:

conda deactivate

Make sure to copy files back to your home directory(/home/netid)!

Cancel your reservation in BioHPC if you are done!

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*END \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

# R PHYLOSEQ ANALYSIS

Import files into Rstudio for each combination (Medium ASV, Medium OTU, ect)

-otu or asv table

-phylogenic tree

-taxonomic table

-metadata file

#See Code for individual analyses