April 2025

Joyalea Carson-Austin, MSc Thesis

University of British Columbia Okanagan

Supervisor: Dr. Miranda Hart

Attribution:

This pipeline is based on the publicly available amplicon sequencing analysis protocol:  
Langille Lab's Microbiome Helper SOP – Amplicon-SOP-v2 (QIIME2-2024.5),  
available at:  
<https://github.com/LangilleLab/microbiome_helper/wiki/Amplicon-SOP-v2-(qiime2%E2%80%902024.5)>

This version has been modified to suit the specific requirements of my master's thesis research on arbuscular mycorrhizal fungi (AMF). All modifications are clearly annotated:

|  |
| --- |
| **Bolded headers *==*** IMR protocol  *Italics* == IMR protocol  Blue text ==modifications made to the pipeline to fit my sequence data/research questions  Grey == QIIME 2 commands with annotations  Green ==File outputs, file paths and visual outputs |

Sequences and files used in this modified pipeline are available at:  
<https://github.com/Joya-CA/AMF-sequences-thesis-2025.git>

This modified version is shared in accordance with GitHub and academic fair use conventions. Please refer to the original Microbiome Helper repository for the most up-to-date and supported versions of the protocol.

Comeau, A.M., Langille, M.G.I., & LangilleLab. (2024). *Microbiome Helper Amplicon SOP v2 (QIIME 2-2024.5)*. GitHub repository: <https://github.com/LangilleLab/microbiome_helper/wiki/Amplicon-SOP-v2-(qiime2%E2%80%902024.5)>**Requirements**

*This workflow assumes that you have QIIME 2 (version qiime2-amplicon-2024.5) installed in a conda environment and that you also have*[*FASTQC*](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)*and [MultiQC](https://multiqc.info/) installed in this environment.*

*This workflow also assumes that the input is raw paired-end MiSeq data in demultiplexed FASTQ format located within a folder called raw\_data. The filenames are expected to look like this: 105CHE6WT\_S325\_L001\_R1\_001.fastq.gz, where each field separated by an \_ character corresponds to:*

1. *The sample name (105CHE6WT)*
2. *The sample number on the run (S325)*
3. *The lane number (L001)*
4. *The read number (R1 for forward and R2 for reverse)*
5. *The set number (001)*

*Often you will need to re-name your files to match this syntax. However, QIIME 2 accepts many different formats so if your files are not already in this format (e.g. not demultiplexed) you would just need to use slightly different commands for importing your data.*

Sequences were provided demultiplexed and in the Cassava format

* 1. **First steps – move out of root operating system**

# Move out of root operating systems

root@Jigglerjr:~# su - joya

**1.2 Activate QIIME 2 conda environment**

# Activate conda and Qiime2

(base) joya@Jigglerjr:~$ conda activate qiime2-amplicon-2024.10

* 1. **Format metadata file**

**See QIIME 2 “Metadata”**

File path: "C:\Users\Joyalea\Documents\Bionformatics\_MARCH\METADATA.tsv"

Ubuntu uses linux style file paths, so the file path needs to be converted to the proper format

To convert:

C: → /mnt/c

\ (backslashes) → / (forward slashes)

E.g. "C:\Users\Joyalea\Documents\Bionformatics\_MARCH\METADATA.tsv"

🡪 /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv

**1.4 Set number of cores**

Performed manually within individual tasks

**1.5 Inspect read quality**

Completed in STEP 1.9

**1.6 Import FASTQs as QIIME 2 artifact**

*To standardize QIIME 2 analyses and to keep track of provenance (i.e. a list of what commands were previously run to produce a file) a special format is used for all QIIME 2 input and output files called an "artifact" (with the extension QZA). The first step is to import the raw reads as a QZA file.*

# Convert data to demux.qza using Cassava File Format

joya@Jigglerjr:~$ qiime tools import \

--type 'SampleData[PairedEndSequencesWithQuality]' \

--input-path "/mnt/c/Users/Joyalea/Documents/UBCO/Thesis/Results - thesis/Bionformatics sequences/CarsonAustinCustomAMF" \

--output-path demux.qza \

--input-format CasavaOneEightSingleLanePerSampleDirFmt

*All of the FASTQs are now in the single artifact file, demux.qza. This file format can be a little confusing at first, but it is actually just a zipped folder. You can manipulate and explore these files better with the qiime tools utilities (e.g. peek and view).*

**1.7 Check for primers in the sequences**

Function notes: ‘grep’ searches inside a file, ‘zgrep’ searches inside a zipped file

# Navigate to the directory where the files are

cd "/mnt/c/Users/Joyalea/Documents/UBCO/Thesis/Results - thesis/Bionformatics sequences/CarsonAustinCustomAMF"

# List FASTQ file names

ls \*.fastq.gz

# Search for primers in reads

zgrep -m 5 'CAGCCGCGGTAATTCCAGCT' M23\_S178\_L001\_R1\_001.fastq.gz # WANDA

zgrep -m 5 'GAACCCAAACACTTTGGTTTCC' M23\_S178\_L001\_R2\_001.fastq.gz # AML2

# Some hits

# Check for AML1 sequences

zgrep -m 5 'ATCAACTTTCGATGGTAGGATAGA' M23\_S178\_L001\_R1\_001.fastq.gz # AML1

# AML1 primer sequence not present

**1.8 Trim primers with cutadapt**

*Screen out reads that do not begin with primer sequence and remove primer sequence from reads using the [cutadapt](http://cutadapt.readthedocs.io/en/stable/guide.html) QIIME 2 plugin.*

# Trim primers using cutadapt: https://forum.qiime2.org/t/demultiplexing-and-trimming-

adapters-from-reads-with-q2-cutadapt/2313

joya@Jigglerjr:~$ qiime cutadapt trim-paired \

--i-demultiplexed-sequences demux.qza \

--p-front-f CAGCCGCGGTAATTCCAGCT \

--p-front-r GAACCCAAACACTTTGGTTTCC \

--p-discard-untrimmed \

--p-no-indels \

--o-trimmed-sequences reads\_trimmed.qza

Output files: reads\_trimmed.qza

# Check for presence of primers

# Export trimmed file to FASTQ format

qiime tools export \

--input-path reads\_trimmed.qza \

--output-path trimmed\_fastq

cd trimmed\_fastq

# Check for presence of primers

# Forward primer

zgrep 'CAGCCGCGGTAATTCCAGCT' M23\_S178\_L001\_R1\_001.fastq.gz

# Reverse primer

zgrep ' GAACCCAAACACTTTGGTTTCC ' M23\_S178\_L001\_R1\_002.fastq.gz

# No primers remaining

cd # Return to home directory

Output files: trimmed\_fastq

**1.9 Summarize trimmed FASTQs**

*You can run the demux summarize command after trimming the reads to get a report of the number of reads per sample and quality distribution across the reads. This generates a more basic output compared to FASTQC/MultiQC, but is sufficient for this step.*

# Summarize trimmed FASTQs

qiime demux summarize \

--i-data reads\_trimmed.qza \

--o-visualization reads\_trimmed\_summary.qzv

cp reads\_trimmed.qzv /mnt/c/Users/Joyalea/Documents/

Visual output: reads\_trimmed.qzv /mnt/c/Users/Joyalea/Documents/

# Check how many were removed after trimming

qiime demux summarize \

--i-data reads\_qza.qza \

--o-visualization reads\_summary.qzv

cp reads\_summary.qzv /mnt/c/Users/Joyalea/Documents/

Visual output: reads\_summary.qzv /mnt/c/Users/Joyalea/Documents/

# Lost ~2.7% of reads, move onto DADA2

*Note that we gave the output file above the extension .qzv since this a special type of artifact file - a****visualization****. You can look at the visualization by uploading the file to the*[*QIIME2 view website*](https://view.qiime2.org/)*and clicking on the Interactive Quality Plot tab at the top of the page*

**2. Denoising the reads into amplicon sequence variants**

*At this stage the main 2 pipelines you can use are based on either*[*deblur*](https://github.com/biocore/deblur)*or*[*DADA2*](https://benjjneb.github.io/dada2/tutorial.html)*. Below we will describe the commands for running deblur. See*[*here*](https://github.com/LangilleLab/microbiome_helper/wiki/QIIME2-DADA2-Quick-Reference)*if you are interested in running DADA2.*

**2.1 DADA2 Pipeline**

*There are detailed tutorials for how to run DADA2 with QIIME 2 on the QIIME 2 website. This page is simply a quick reference for convenience. Running the DADA2 pipeline is more straight-forward than deblur, but is slower. It is also more flexible with the truncation lengths, but note that these are specified for forward and reverse reads separately. The important thing to consider here is what lengths to use so that the forward and reverse reads can be stitched together accurately. You may also want to lower the maximum number of expected errors allowed depending on your data quality.*

*Here is a quick reference for running DADA2 in QIIME2 (set # of threads to CPU capacity).*

*ITS and SSU regions are variable in length and should not be set to a specific truncation.*

Note: Leave truncation variable due to variation in SSU region!

# DADA2 Denoising

qiime dada2 denoise-paired \

--i-demultiplexed-seqs reads\_trimmed.qza \

--p-trunc-len-f 0 \ #No truncation due to region variability

--p-trunc-len-r 0 \

--p-max-ee-f 2 \ #Allow up to 2 expected errors per forward read

--p-max-ee-r 3 \ #Allow up to 3 expected errors per reverse read

--p-n-threads 4 \

--output-dir /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output

File output: /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output

**2.2 Summarizing DADA2 output**

*Once a denoising pipeline has been run you can summarize the output table with the below command, which will create a visualization artifact for you to view.*

# Check retained reads

qiime metadata tabulate \

--m-input-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/denoising\_stats.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/denoising\_stats.qzv

Visual output: /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/denoising\_stats.qzv

# Check visual to ensure sufficient reads were retained

*We will use this visualization later to determine the the cut-offs for filtering the table below, but for now you should mainly take a look at the visualization to****ensure that sufficient reads have been retained after running DADA2****. This denoising tool filters out reads that either do match to known noise or that do not match with low similarity to the expected amplicon region. If your samples have very low depth after running deblur (compared to the input read depth) this could be a red flag that either you ran the tool incorrectly, you have a lot of noise in your data, or that deblur is inappropriate for your dataset.*

**3. Create a summary table to check retained ASVs**

# Create summary table to check retained ASVs

qiime feature-table summarize \

--i-table /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/table.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/table\_summary.qzv \

--m-sample-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADA.txt

Visual output: /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/table\_summary.qzv \

# Review summary table for information on # of unique ASVs, # of sequences per sample and mean sequence number. Used to determine rarefication.

**3.1 Cluster to 97% identity and assign taxonomy using MAARJAM database (vsearch blast)**

*We recommend users use a widely used classifier to help ensure there are no unexpected issues with the Naive-Bayes model. We previously maintained primer-specific classifiers, which theoretically can provide more accurate classifications, but we no longer do this due to concerns regarding issues with the trained models that are difficult to catch if only a couple people are running them. No matter what approach you use, it's a good idea to run a few sanity checks on the output to make sure it worked correctly for your data (see below).*

As I amplified the SSU region with AMF specific primers, I BLASTed My sequences against the MAARJAM database. To do this, the database reference sequences need to be imported.

# Make directory to store 97% identity results

mkdir -p /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN

# Copy MAARJAM reference databases (TYPE) to file path

cp /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/maarjam\_ref\_\*.qza \

/mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

File output: /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/maarjam\_ref\_seqs.qza

# Cluster sequences to 97% identity using de novo clustering

qiime vsearch cluster-features-de-novo \

--i-table /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/table.qza \

--i-sequences /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/dada2\_output/representative\_sequences.qza \

--p-perc-identity 0.97 \

--p-threads 4 \

--o-clustered-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table.qza \

--o-clustered-sequences /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza

Output files:

--o-new-reference-sequences /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_all\_refs.qza

Saved FeatureTable[Frequency] to: /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table.qza

Saved FeatureData[Sequence] to: /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza

Saved FeatureData[Sequence] to: /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_all\_refs.qza

# Recreate summary output table

qiime feature-table summarize \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table\_summary.qzv \

--m-sample-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv

Output visualisation: /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table\_summary.qzv

**3.2 Run taxonomic classification (blastn: not Comeau)**

‘blastn’ was used as opposed to ‘consensus blast’, as it allows for query coverage to be set and e values to be set (Comeau used a trained Naïve-Bayes classifier. This is not a well supported option for AMF – poor database coverage and no pre-built classifiers available.

# Alignment with blastn (97% identity threshold, 95% query coverage, 1 e 50 )

# Identity threshold and query coverage are based off Vasar et al. (2017) Supplementary Materials, <https://static-content.springer.com/esm/art%3A10.1007%2Fs00572-017-0791-y/MediaObjects/572_2017_791_MOESM1_ESM.pdf>

#Export representative sequences from QIIME 2 artifact

# Export inside WSL home

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza \

--output-path ~/otu97\_rep\_seqs\_export

# Then copy the results back into the Windows folder

cp -r ~/otu97\_rep\_seqs\_export /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

# Export MAARJAM reference sequences from QIIME 2 artifact

# Export to WSL home directory

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_ref\_seqs.qza \

--output-path ~/maarjam\_ref\_export

# Copy exported folder to target Windows directory

cp -r ~/maarjam\_ref\_export /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

# Build BLAST database from MAARJAM reference FASTA

mkdir -p /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_blast\_db

makeblastdb \

-in /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_ref\_export/dna-sequences.fasta \

-dbtype nucl \

-out /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_blast\_db/maarjam\_ref

# Run BLASTn with 97% identity, 95% coverage, and e-value cutoff of 1e-50

blastn \

-query /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/rep\_seqs\_export/dna-sequences.fasta \

-db /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_blast\_db/maarjam\_ref \

-out /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/blast\_results\_97.tsv \

-evalue 1e-50 \

-perc\_identity 97 \

-qcov\_hsp\_perc 95 \

-outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore"

# Convert to a QIIME2 compatible version

**#Performed in R (see ‘QIIME2Conversion\_TaxonomyAssignment\_Markdown.Rmd’)**

#reimport

qiime tools import \

--type 'FeatureData[Taxonomy]' \

--input-format HeaderlessTSVTaxonomyFormat \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu97.tsv \

--output-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu97.qza

#Filter the feature table to match

qiime feature-table filter-features \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu97.qza \

--o-filtered-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table\_classified\_only.qza

#Create blast table and FILTER OUT UNASSIGNED SEQUENCES

qiime taxa filter-table \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table\_classified\_only.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu97.qza \

--p-exclude Unassigned \

--o-filtered-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/table\_blast\_filtered\_otu97.qza

# BLASTn unassigned features at 90% identity match with 95% query

# Export sequences

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza \

--output-path ~/rep\_seqs\_export\_97

cp -r ~/rep\_seqs\_export\_97 /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

**#Performed in R (see ‘QIIME2Conversion\_TaxonomyAssignment\_Markdown.Rmd’)**

# Filter Unassigned sequences

qiime feature-table filter-seqs \

--i-data /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/unassigned\_ids\_97.tsv \

--o-filtered-data /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/unassigned\_otu\_seqs.qza

#Export unassigned sequences

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/unassigned\_otu\_seqs.qza \

--output-path ~/unassigned\_otu\_seqs\_export

cp -r ~/unassigned\_otu\_seqs\_export /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

# Run BLASTn at 90% identity

blastn \

-query /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/unassigned\_otu\_seqs\_export/dna-sequences.fasta \

-db /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/maarjam\_blast\_db/maarjam\_ref \

-out /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/blast\_results\_90.tsv \

-evalue 1e-50 \

-perc\_identity 90 \

-qcov\_hsp\_perc 95 \

-outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore"

**#Summarized in R (see ‘QIIME2Conversion\_TaxonomyAssignment\_Markdown.Rmd’)**

#Reimport to QIIME2

qiime tools import \

--type 'FeatureData[Taxonomy]' \

--input-format HeaderlessTSVTaxonomyFormat \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu90\_filtered\_97.tsv \

--output-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu90\_filtered\_97.qza

#Merge 90% and 97% assigned sequences

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu97.qza \

--output-path ~/taxonomy\_otu97\_export

cp -r ~/taxonomy\_otu97\_export /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

qiime tools export \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_otu90\_filtered\_97.qza \

--output-path ~/taxonomy\_otu90\_filtered\_export\_97

cp -r ~/taxonomy\_otu90\_filtered\_export\_97 /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/

# Reimport merged taxonomy

qiime tools import \

--type 'FeatureData[Taxonomy]' \

--input-format HeaderlessTSVTaxonomyFormat \

--input-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.tsv \

--output-path /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza

# Filter table to match

qiime feature-table filter-features \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_table.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--o-filtered-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu\_table\_merged.qza

#Create interactive barplot

qiime taxa barplot \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu\_table\_merged.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged\_barplot.qzv

# Recreate summary table\_filtered

qiime feature-table summarize \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu\_table\_merged.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/table\_summary\_merged.qzv \

--m-sample-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv

**3.3 Assess subset of taxonomic assignments with BLAST**

*The performance of the taxonomic classification is difficult to assess without a gold-standard reference, but nonetheless one basic sanity check is to compare the taxonomic assignments with the top BLAST hits for certain ASVs.*

*It is simple to do this with QIIME 2 by running:*

*qiime feature-table tabulate-seqs --i-data deblur\_output/representative\_sequences.qza \*

*--o-visualization deblur\_output/representative\_sequences.qzv*

*The file deblur\_output/representative\_sequences.qzv is a QIIME 2 visualization file that you can open in the*[*QIIME 2 viewer*](https://view.qiime2.org/)*. The format makes it easy to BLAST certain ASVs against the NCBI nt database. By comparing these BLAST hits with the taxonomic assignment of ASVs generated above you can reassure yourself that the taxonomic assignments overall worked correctly. It's a good idea to select ~5 ASVs to BLAST for this validation, which should be from taxonomically different groups, such as different phyla, according to the taxonomic classifier.*

90% matched FASTA sequences were reblasted against the NCBI database and the AMFungal database to compare identity matches between MAARJAM and NCBI nt.

**4. Filtering resultant table**

*Filtering the denoised table is an important step of microbiome data analysis. You can see more details on this process in the*[*QIIME 2 filtering tutorial*](https://docs.qiime2.org/2022.11/tutorials/filtering/)*.*

**4.1 Filter out rare ASVs**

*Based on the summary visualization created in step 3.1 above you can choose a cut-off for how frequent a variant needs to be (and optionally how many samples need to have the variant) for it to be retained. One possible choice would be to remove all ASVs that have a frequency of less than 0.1% of the mean sample depth. This cut-off excludes ASVs that are likely due to MiSeq bleed-through between runs (reported by Illumina to be 0.1% of reads). To calculate this cut-off you would identify the mean sample depth in the visualization created in step 2.4, multiply it by 0.001, and round to the nearest integer.*

*Once you've determined how you would like to filter your table you can do so with this command (X is a placeholder for your choice):*

Filtering singletons was performed to reduce noise that may appear with low abundance reads. While estimating MiSeq bleed – estimated at 0.1% of sequence abundance -- is a common approach in microbiome analysis, AMF are in lower abundance than many other microbial communities (e.g. bacteria). Thus, applying a higher read threshold might disproportionately remove rare AMF taxa. Therefore, singleton removal was chosen as a more conservative filtering strategy for this dataset.

# Filter out singletons

qiime feature-table filter-features \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu\_table\_merged.qza \

--p-min-frequency 2 \

--p-min-samples 1 \

--o-filtered-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza

# Recreate summary table\_filtered

qiime feature-table summarize \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/table\_summary\_merged\_filtered\_97.qzv \

--m-sample-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv

# Create taxa barplot

qiime taxa barplot \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxa\_barplot\_merged\_97.qzv

#Create taxa barplot collapsed to the level of treatment

qiime feature-table summarize \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/table\_summary\_merged\_filtered\_97.qzv \

--m-sample-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv

#Create Sample level barplot

qiime taxa barplot \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxa\_barplot\_merged\_97.qzv

#Create table grouped by treatment

qiime feature-table group \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--p-mode sum \

--p-axis sample \

--o-grouped-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/grouped\_by\_treatment\_table.qza

#Create taxa barplot grouped by sample

qiime taxa barplot \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/grouped\_by\_treatment\_table.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA\_GROUPED.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxa\_barplot\_by\_treatment.qzv

**4.5 Exclude low-depth samples**

*Often certain samples will have quite low depth after these filtering steps, which can be excluded from downstream analyses since they will largely add noise. There is no single cut-off that works best for all datasets, but researchers often use minimum cut-offs within the range of 1000 to 4000 reads. You can also use a cut-off much lower than this if you want to retain all samples except those that failed entirely (e.g. depth < 50 reads). Ideally you would choose this cut-off after visualizing rarefaction curves to determine at what read depth the richness of your samples plateaus and choose a cut-off as close to this plateau as possible while retaining sufficient sample size for your analyses.*

*To perform this rarefaction curve analysis you would first need to summarize the filtered table we produced in the last step:*

Note: Perform rarefication curves on the merged tables (create summary table)

*From this table you need to determine the maximum depth across your samples. You can then generate the rarefaction curves with this command (where X is a placeholder for the max depth across samples).*

# Check rarefication curves

qiime diversity alpha-rarefaction \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--p-max-depth 44842 \

--p-steps 20 \

--p-metrics 'observed\_features' \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/alpha\_rarefaction\_observed\_features\_97.qzv

*Take a look at these curves to help decide on a minimum depth cut-off for retaining samples. Once you decide on a hard cut-off you can exclude samples below this cut-off with this command (where SET\_CUTOFF is a placeholder for the minimum depth you select).* ***Alternatively, if you do not wish to exclude any samples****then you can simply make a copy of the QZA file with the final table filename (i.e. cp* ***filtered\_rep\_seqs\_clean\_FINAL.qza; mine: assigned\_rep\_seqs\_97s),*** *since this is the filename used for the remaining commands below.*

Diversity plateaus before the lowest sampling depth (12494), therefore, all samples were retained.

**5. QIIME 2 de novo tree creation (18S and ITS data)**

*Given the lack of a pre-calculated reference tree for 18S and ITS data (unlike the above 16S tree) to do direct placement, we create the tree de novo here below.*

**5.1 Making multiple-sequence alignment**

*We'll need to make a multiple-sequence alignment of the OTUs before running [FastTree](http://www.microbesonline.org/fasttree/). First, we'll make a folder for the output files.*

# Make the denovo tree (unrooted)

#Create output directory for merged tree

mkdir -p /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged

# Perform multiple sequence alignment with MAFFT

qiime alignment mafft \

--i-sequences /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/otu97\_rep\_seqs.qza \

--p-n-threads 4 \

--o-alignment /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/aligned\_seqs.qza

# Mask variable positions in the alignment

qiime alignment mask \

--i-alignment /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/aligned\_seqs.qza \

--o-masked-alignment /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/masked\_aligned\_seqs\_97.qza

# Construct unrooted tree using FastTree

qiime phylogeny fasttree \

--i-alignment /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/masked\_aligned\_seqs\_97.qza \

--p-n-threads 4 \

--o-tree /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/unrooted\_tree\_97.qza

# Root the tree at the midpoint

qiime phylogeny midpoint-root \

--i-tree /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/unrooted\_tree\_97.qza \

--o-rooted-tree /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/rooted\_tree\_97.qza

**7. Generate stacked barchart of taxa relative abundances**

*A more useful output is the interactive stacked bar-charts of the taxonomic abundances across samples, which can be output with this command:*

# Create taxa barplot

# Create taxonomy barplot with merged data

qiime taxa barplot \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--i-taxonomy /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxonomy\_merged.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/taxa\_barplot\_97\_merged\_FINAL.qzv

**8. Calculating diversity metrics and generating ordination plots**

*Common alpha and beta-diversity metrics can be calculated with a single command in QIIME 2. In addition, ordination plots (such as PCoA plots for weighted UniFrac distances) will be generated automatically as well. This command will also rarefy all samples to the sample sequencing depth before calculating these metrics (X is a placeholder for the lowest reasonable sample depth; samples with depth below this cut-off will be excluded).*

# Run core metrics rarefied to lowest sampling depth

# Run core-metrics-phylogenetic analysis (rarefied to 12434)

# Run core diversity metrics using merged filtered table and tree

qiime diversity core-metrics-phylogenetic \

--i-phylogeny /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/tree\_out\_97\_blastn\_merged/rooted\_tree\_97.qza \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--p-sampling-depth 12434 \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--output-dir /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged \

--p-n-jobs-or-threads 4

# Alpha diversity significance testing

qiime diversity alpha-group-significance \

--i-alpha-diversity /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/shannon\_vector.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/shannon\_group\_significance.qzv

qiime diversity alpha-group-significance \

--i-alpha-diversity /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/observed\_features\_vector.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/richness\_group\_significance.qzv

qiime diversity alpha-group-significance \

--i-alpha-diversity /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/evenness\_vector.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/evenness\_group\_significance.qzv

# Beta diversity PERMANOVA tests

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--p-method permanova \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_permanova.qzv

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--p-method permanova \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_permanova.qzv

# Beta diversity dispersion (PERMDISP) tests

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--p-method permdisp \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_permdisp.qzv

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--p-method permdisp \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_permdisp.qzv

#DATE HARVESTED – BETA DIVERSITY

# Beta diversity PERMANOVA tests

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--p-method permanova \

--p-permutations 999 \

--p-pairwise \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_permanova\_date.qzv

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--p-method permanova \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_permanova\_date.qzv

# Beta diversity dispersion (PERMDISP) tests

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--p-method permdisp \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/bray\_curtis\_permdisp\_date.qzv

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--p-method permdisp \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_permdisp\_date.qzv

qiime diversity beta-group-significance \

--i-distance-matrix /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_distance\_matrix.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--p-method permanova \

--p-pairwise \

--p-permutations 999 \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/core\_metrics\_97\_blastn\_merged/jaccard\_permanova\_pairwise\_date.qzv

**9. Identifying differentially abundant features with ANCOM**

[*ANCOM*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450248/)*is one method to test for differences in the relative abundance of features between sample groupings. It is a compositional approach that makes no assumptions about feature distributions. However, it requires that all features have non-zero abundances so a pseudocount first needs to be added (1 is a typical pseudocount choice):*

QIIME forum on running ANCOMs on fungal ASVs: https://forum.qiime2.org/t/taxa-collapse-shows-too-general-taxonomic-assignations/30526/9

# Add pseudocount to the merged, filtered table

qiime composition add-pseudocount \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97.qza \

--p-pseudocount 1 \

--o-composition-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97\_pseudocount.qza

# Run ANCOM by Treatment

qiime composition ancom \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97\_pseudocount.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column Treatment \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/ancom\_output\_treatment\_97\_merged.qzv

# Run ANCOM by Harvest Date

qiime composition ancom \

--i-table /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/merged\_filtered\_table\_97\_pseudocount.qza \

--m-metadata-file /mnt/c/Users/Joyalea/Documents/Bionformatics\_MARCH/METADATA.tsv \

--m-metadata-column DateHarvested \

--o-visualization /mnt/c/Users/Joyalea/Documents/Bioinformatics\_97\_BLASTN/ancom\_output\_date\_97\_merged.qzv

**Other resources**

*There are many other possible QIIME 2 analyses that we recommend you look into. You may also find these other resources useful:*

* [ALDEx2](https://bioconductor.org/packages/release/bioc/html/ALDEx2.html) for testing for differential relative abundances in R
* [corncob](https://github.com/bryandmartin/corncob) for testing for differential relative abundances in R
* [Building Random Forest Models in R](https://github.com/LangilleLab/microbiome_helper/wiki/Random-Forest-Tutorial)
* [PICRUSt2](https://github.com/picrust/picrust2/wiki) for metagenome prediction based on amplicon sequences
* [STAMP](http://kiwi.cs.dal.ca/Software/STAMP) for straight-forward visualization (see [workflow here](https://github.com/LangilleLab/microbiome_helper/wiki/STAMP-preparation))

**Changelog**

* **v2020.8** (2020/10/30): Changed qiime quality-filter q-score-joined to be qiime quality-filter q-score.
* **v2020.8** (2020/10/30): Updated taxa classifiers, but no longer output classifiers for individual 16S/18S variable regions.
* **v2022.2** (2022/11/09): Changed to state removing unclassified ASVs is not the recommended default when filtering out "contaminants".
* **v2022.11** (2023/01/04): Changed qiime vsearch join-pairs to qiime vsearch merge-pairs, and the option --o-joined-sequences to --o-merged-sequences