**QIIME2 2024.10 (performed 2025)**

For using the new version of qiime2 in the institute terminal

module load miniconda3/latest

conda activate /$HOME/vast/qiime2-amp-2024.10

conda activate qiime2-amp-2024.10

**Processing of Vaginal samples**

These are contained in Runs 8 (Bacterial sequences region V4 of the 16S rRNA) and 17 (Bacterial and fungal ITS sequences).

1. ***Importing sequence data – This is with multiplexed samples***

qiime tools import --type EMPPairedEndSequences --input-path /$HOME/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Raw\_data/ENDIA\_Raw\_data\_16SrRNA\_sequencing/ENDIA\_Raw\_data\_16S\_Run8/SeqRun8/ --output-path /$HOME/vast/QRun8/QRun8.qza

qiime tools import --type EMPPairedEndSequences --input-path /$HOME/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Raw\_data/ENDIA\_Raw\_data\_16SrRNA\_sequencing/ENDIA\_Raw\_data\_16S\_Run17/SeqRun17/ --output-path /$HOME/vast/QRun17/QRun17.qza

For this you need to have the forward sequences with name forward.fastq.gz and reverse with name reverse.fastq.gz and the barcodes.fastq.gz, compressed in the directory and nothing more. I temporarily changed the names of the sequence files but saved the original name in a text file called NameR8.

This step is quick and doesn’t need to be done in a jobscript.

Note: Every qiime2 step was done for each set of sequencing runs, 8 and 17, but sometimes I show below only how it was run with run 8.

1. ***Demultiplexing sequences***

qiime demux emp-paired --i-seqs QRun8.qza --m-barcodes-file /$HOME/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_16SrRNA\_sequencing/ENDIA\_16S\_Run8/ENDIA\_Processed\_data\_16S\_Run8/Step1/Mapping\_file\_Seq\_8-Run.csv --m-barcodes-column BarcodeSequence --o-per-sample-sequences demux --o-error-correction-details demux-details.qza --p-no-golay-error-correction

qiime demux emp-paired --i-seqs QRun17.qza --m-barcodes-file /$HOME/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_16SrRNA\_sequencing/ENDIA\_16S\_Run17/ENDIA\_Processed\_data\_16S\_Run8/Step1/Mapping\_file\_Run\_17.csv --m-barcodes-column BarcodeSequence --o-per-sample-sequences demux --o-error-correction-details demux-details.qza --p-no-golay-error-correction

***Summary of the demultiplexing results:***

qiime demux summarize --i-data demux.qza --o-visualization demux.qzv

1. ***Trimming of PCR primers***

qiime cutadapt trim-paired --i-demultiplexed-sequences ./demuxQR8.qza --p-front-f GGACTACNVGGGTWTCTAAT --p-front-r GTGYCAGCMGCCGCGGTAA --o-trimmed-sequences trimmed\_demuxQR8.qza -- verbose

1. ***Sequence quality control and feature table construction using DADA2***

qiime dada2 denoise-paired --i-demultiplexed-seqs /$HOME/vast/QRun8/trimmed\_demuxQR8.qza --p-trunc-len-f 200 --p-trunc-len-r 160 --o-representative-sequences /$HOME/vast/QRun8.qza --o-table /$HOME/vast/QRun8/table-dada2\_QR8.qza --o-denoising-stats /$HOME/vast/QRun8/stats-dada2\_QR8.qza --p-max-ee-r 3

qiime dada2 denoise-paired --i-demultiplexed-seqs /$HOME/vast/QRun17/trimmed\_demuxQR17.qza --p-trunc-len-f 200 --p-trunc-len-r 160 --o-representative-sequences /$HOME/vast/QRun17/rep-seqs-dada2\_QR17.qza --o-table /$HOME/vast/QRun17/table-dada2\_QR17.qza --o-denoising-stats /$HOME/vast/QRun17/stats-dada2\_QR17.qza --p-max-ee-r 3

***Summary of feature table***

1. qiime feature-table summarize --i-table table-dada2\_QR8.qza --o-visualization table-dada2\_QR8.qzv --m-sample-metadata-file /$HOME/vast/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_16SrRNA\_sequencing/ENDIA\_16S\_Run8/ENDIA\_Processed\_data\_16S\_Run8/Step1/Mapping\_file\_Seq\_8-Run.csv
2. qiime metadata tabulate --m-input-file stats-dada2\_QR8.qza --o-visualization stats-dada2\_QR8.qzv
3. qiime feature-table tabulate-seqs --i-data rep-seqs-dada2\_QR8.qza --o-visualization rep-seqs-dada2\_QR8.qzv
4. ***Generate a tree for phylogenetic diversity analyses***

Just use one command and obtain the final rooted tree.

qiime phylogeny align-to-tree-mafft-fasttree --i-sequences /$HOME/vast/QRun8/rep-seqs-dada2\_QR8.qza --o-alignment /$HOME/vast/QRun8/aligned-rep-seqs-dada2\_QR8.qza --o-masked-alignment /$HOME/vast/QRun8/masked-aligned-rep-seqs-dada2\_QR8.qza --o-tree /$HOME/vast/QRun8/unrooted-tree\_QR8.qza --o-rooted-tree /$HOME/vast/QRun8/rooted-tree\_QR8.qza

1. ***Taxonomic classification***

For this I will used the most current release of the Silva database 138, but I need to train it.

wget -O "silva-138-99-seqs-515-806.qza" <https://data.qiime2.org/2024.10/common/silva-138-99-seqs-515-806.qza>

wget -O "silva-138-99-tax-515-806.qza" <https://data.qiime2.org/2024.10/common/silva-138-99-tax-515-806.qza>

qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads /$HOME/vast/QRun8/Silva\_full/silva-138-99-seqs-515-806.qza --i-reference-taxonomy /$HOME/vast/QRun8/Silva\_full/silva-138-99-tax-515-806.qza --o-classifier /$HOME/vast/QRun8/Silva\_full/silva-138-99-515-806-nb-Q2024.10-classifier.qza

Using an interactive partition Job because it needs more memory:

salloc --partition interactive --job-name "Qiime\_DB" --cpus-per-task 2 --time 08:00:00 --mem 50G

Then need to activate conda again:

conda activate /$HOME/vast/qiime2-amp-2024.10

conda activate qiime2-amp-2024.10

And then the qiime command

qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads /$HOME/vast/QRun8/Silva\_full/silva-138-99-seqs-515-806.qza --i-reference-taxonomy /$HOME/vast/QRun8/Silva\_full/silva-138-99-tax-515-806.qza --o-classifier /$HOME/vast/QRun8/Silva\_full/silva-138-99-515-806-nb-Q2024.10-classifier.qza

Classify sequences:

qiime feature-classifier classify-sklearn --i-classifier /$HOME/vast/QRun8/Silva\_full/silva-138-99-515-806-nb-Q2024.10-classifier.qza --i-reads /$HOME/vast/QRun8/rep-seqs-dada2\_QR8.qza --o-classification /$HOME/vast/QRun8/Taxonomy\_QR8.qza

qiime metadata tabulate --m-input-file /$HOME/vast/QRun8/Taxonomy\_QR8.qza --o-visualization /$HOME/vast/QRun8/Taxonomy\_QR8.qzv

qiime taxa barplot --i-table ./table-dada2\_QR8.qza --i-taxonomy ./Taxonomy\_QR8.qza --m-metadata-file /$HOME/vast/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_16SrRNA\_sequencing/ENDIA\_16S\_Run8/ENDIA\_Processed\_data\_16S\_Run8/Step1/Mapping\_file\_Seq\_8-Run.csv --o-visualization taxa-bar-plots\_QR8.qzv

qiime taxa barplot --i-table ./table-dada2\_QR17.qza --i-taxonomy ./Taxonomy\_QR17.qza --m-metadata-file /$HOME/vast/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_16SrRNA\_sequencing/ENDIA\_16S\_Run17/ENDIA\_Processed\_data\_16S\_Run17/Mapping\_file\_Run\_17.csv --o-visualization taxa-bar-plots\_QR17.qzv

**Formatting samples to input to rStudio for further analysis**

***Vaginal dataset (Bacterial with 2 sequencing runs)***

1. Building a tree using the set of representative sequences from the 2 runs from the same project

🡪 Merge representative sequences from the vaginal project

qiime feature-table merge-seqs --i-data /$HOME/vast/QRun8/rep-seqs-dada2\_QR8.qza /$HOME/vast/QRun17/rep-seqs-dada2\_QR17.qza --o-merged-data /$HOME/vast/rep-seqs-dada2\_QVaginal8\_17.qza

🡪 Build a phylogenetic tree using rep seqs from both runs:

qiime phylogeny align-to-tree-mafft-fasttree --i-sequences /$HOME/vast/rep-seqs-dada2\_QVaginal8\_17.qza --o-alignment /$HOME/vast/aligned-rep-seqs-dada2\_QVaginal8\_17.qza --o-masked-alignment /$HOME/vast/masked-aligned-rep-seqs-dada2\_QVaginal8\_17.qza --o-tree /$HOME/vast/unrooted-tree\_QVaginal8\_17.qza --o-rooted-tree /$HOME/vast/rooted-tree\_QVaginal8\_17.qza

1. Merging feature tables from the 2 runs

Merging tables before converting to biom.

qiime feature-table merge --i-tables ./QRun8/table-dada2\_QR8.qza ./QRun17/table-dada2\_QR17.qza --p-overlap-method sum --o-merged-table ./Vaginal\_merged-tables 8\_17.qza

1. Merging taxonomies from the 2 runs

qiime feature-table merge-taxa --i-data ./QRun8/Taxonomy\_QR8.qza ./QRun17/Taxonomy\_QR17.qza --o-merged-data Vaginal\_merged\_Taxonomy8\_17.qza

1. Use biom to input to pyloseq

I needed to install biom to be able to use it

pip install numpy

pip install biom-format

a.- Unzip file Table to extract biom file

b.- Transform biom into txt: biom convert -i table Vaginal\_merged-tables8\_17B.biom -o Vaginal\_merged-tables8\_17B.txt --to-tsv

c.- Convert again to biom file as follows: biom convert -i Vaginal\_merged-tables8\_17B.txt -o Vaginal\_merged-tables8\_17G.biom --table-type="OTU table" --to-json

d.- Add the taxonomy: biom add-metadata --sc-separated taxonomy --observation-header OTUID,taxonomy --observation-metadata-fp Vaginal\_merged\_Taxonomy8\_17.tsv -i Vaginal\_merged-tables8\_17G.biom -o Vaginal\_merged-tables8\_17G\_wTax.biom

e.- Add the metadata: biom add-metadata -i Vaginal\_merged-tables8\_17G\_wTax.biom -o Vaginal\_merged-tables8\_17G\_wTaxMet.biom --sample-metadata-fp Mapping\_file\_Run\_8\_17.csv

1. Apply abundance and contamination filters to qiime2 features.

We got everything into phyloseq, including representative fasta sequences. Using exactly the same pipeline, we obtain 157 reference sequences/features after filters (abundance and kingdom-unassigned features). Before we had 160 features. When inspecting the feature IDs (which are 100% associated to the sequence itself) we can see that we have exactly the same feature IDs as before, just 2 were excluded.

Also we wrote in the paper “Of the 334 samples, 172 were from women with T1D. Of these, 160 from women with T1D and 150 from women without T1D were randomly selected for analysis of the bacterial microbiome by 16S rRNA”. And before doing rarefaction we have 334 samples in this step. But we will have 310, exactly the same as in the previous analysis.

1. Check taxonomy classification and use BLAST n to improve when needed.

The taxonomic classification from Qiime2 2024.10 did not improve when compared to previous pipeline using an older version of Qiime2 and Blanst n was used to further classify the sequences as before obtaining the same results.

Until this point everything seems to be very similar and we will be performing the statistical analysis in the same way as before.

**Processing of Mycobiome Vaginal samples**

1. Fungal sequences are in run 17. Therefore import was already performed in step A above.
2. ***Demultiplexing sequences***

We included in the Mapping file only barcodes for ITS1 even though in the sequencing run file we have ITS1 and ITS2, because we will preform the whole analysis with ITS1 only.

qiime demux emp-paired --i-seqs QRun17.qza --m-barcodes-file /$HOME/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_Mycobiome\_sequencing/ENDIA\_Mycobiome\_Run17/ENDIA\_Processed\_data\_Mycobiome\_Run17/Mapping\_file\_Vaginal\_ITS1\_sequencing.csv --m-barcodes-column BarcodeSequence --o-per-sample-sequences demux\_ITS1 --o-error-correction-details demux-details.qza --p-no-golay-error-correction

1. ***Trimming of primers***

qiime cutadapt trim-paired --i-demultiplexed-sequences demux\_ITS1.qza --p-front-f CTTGGTCATTTAGAGGAAGTAA --p-front-r GCTGCGTTCTTCATCGATGC --o-trimmed-sequences trimmed\_demux\_ITS1.qza --verbose

Here, we are going to filter only Pair 1. For this, file trimmed\_demux\_ITS1.qza was unzipped and only sample files with front primer (\_1) were kept in a folder. The we imported them as before, but know as single end:

qiime tools import --type 'SampleData[SequencesWithQuality]' --input-path /$HOME/ENDIA/ ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_Mycobiome\_sequencing /demux\_ITS1\_R1/ --input-format CasavaOneEightSingleLanePerSampleDirFmt

--output-path /$HOME/vast/VMycobiome/demux-single-end\_ITS1\_R1.qza

From the qiime manual:

In the Casava 1.8 demultiplexed (single-end) format, there is one fastq.gz file for each sample in the study which contains the *single-end* reads for that sample. The file name includes the sample identifier and should look like L2S357\_15\_L001\_R1\_001.fastq.gz. This is exactly our case because we unzip the demultiplexed file with the ITS1 sequences file per sample, stayed with \_R1 files and applied as above.

1. ***Sequence quality control and feature table construction using DADA2***

qiime dada2 denoise-single --i-demultiplexed-seqs /$HOME/vast/VMycobiome/demux-single-end\_ITS1\_R1.qza --p-trunc-len 255 --o-representative-sequences /$HOME/vast/VMycobiome/rep-seqs\_ITS1\_R1.qza --o-table /$HOME/vast/VMycobiome/table\_ITS1\_R1.qza --o-denoising-stats /$HOME/vast/VMycobiome/stats\_ITS1\_R1.qza --p-max-ee 3

qiime feature-table summarize --i-table /$HOME/vast/VMycobiome/table\_ITS1\_R1.qza --o-visualization /$HOME/vast/VMycobiome/table\_ITS1\_R1.qzv --m-sample-metadata-file /$HOME/vast/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_Mycobiome\_sequencing/ENDIA\_Processed\_data\_Vaginal\_mycobiome\_ITS/Mapping\_file\_Vaginal\_ITS\_sequencing\_S.csv

qiime metadata tabulate --m-input-file stats\_ITS1\_R1.qza --o-visualization stats\_ITS1\_R1.qzv

qiime feature-table tabulate-seqs --i-data rep-seqs\_ITS1\_R1.qza --o-visualization rep-seqs\_ITS1\_R1.qzv

1. ***Taxonomic classification***

We downloaded the ref seq and taxonomy from the unite database webpage and imported into qiime2.

qiime tools import --type 'FeatureData[Sequence]' --input-path sh\_refs\_qiime\_ver10\_99\_04.04.2024.fasta --output-path sh\_refs\_qiime\_ver10\_99\_04.04.2024.qza

qiime tools import --type 'FeatureData[Taxonomy]' --input-format HeaderlessTSVTaxonomyFormat --input-path sh\_taxonomy\_qiime\_ver10\_99\_04.04.2024.txt --output-path sh\_taxonomy\_qiime\_ver10\_99\_04.04.2024.qza

qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads /$HOME/vast/VMycobiome/sh\_refs\_qiime\_ver10\_99\_04.04.2024.qza --i-reference-taxonomy /$HOME/vast/VMycobiome/sh\_taxonomy\_qiime\_ver10\_99\_04.04.2024.qza --o-classifier /$HOME/vast/VMycobiome/sh\_taxonomy\_qiime\_ver10\_99\_04.04.2024-classifier.qza

**Classify sequences:**

qiime feature-classifier classify-sklearn --i-classifier /$HOME/vast/VMycobiome/sh\_taxonomy\_qiime\_ver10\_99\_04.04.2024-classifier.qza --i-reads /$HOME/vast/VMycobiome/rep-seqs\_ITS1\_R1.qza --o-classification /$HOME/vast/VMycobiome/Taxonomy\_ITS1\_R1.qza

qiime metadata tabulate --m-input-file /$HOME/vast/VMycobiome/Taxonomy\_ITS1\_R1.qza --o-visualization /$HOME/vast/VMycobiome/Taxonomy\_ITS1\_R1.qzv

qiime taxa barplot --i-table /$HOME/vast/VMycobiome/table\_ITS1\_R1.qza --i-taxonomy /$HOME/vast/VMycobiome/Taxonomy\_ITS1\_R1.qza --m-metadata-file /$HOME/vast/ENDIA/ENDIA\_Sequencing\_data/ENDIA\_Processed\_data/ENDIA\_Processed\_data\_Mycobiome\_sequencing/ENDIA\_Processed\_data\_Vaginal\_mycobiome\_ITS/Mapping\_file\_Vaginal\_ITS\_sequencing\_S.csv --o-visualization /$HOME/vast/VMycobiome/taxa-bar-plots\_QVMy.qzv

1. ***Generate a tree for phylogenetic diversity analyses***

qiime phylogeny align-to-tree-mafft-fasttree --i-sequences /$HOME/vast/VMycobiome/rep-seqs\_ITS1\_R1.qza --o-alignment /$HOME/vast/VMycobiome/alignment\_ITS1\_R1.qza --o-masked-alignment /$HOME/vast/masked\_ITS1\_R1.qza VMycobiome/ --o-tree /$HOME/vast/VMycobiome/unrooted\_tree\_ITS1\_R1.qza --o-rooted-tree /$HOME/vast/VMycobiome/rooted\_tree\_ITS1\_R1.qza

**Formatting samples to input to rStudio for further analysis**

***Vaginal dataset (Fungal with 1 sequencing run)***

1. Use biom to input to pyloseq

a.- Unzip file Table to extract biom file:

b.- Transform biom into txt: biom convert -i ./table\_ITS1\_R1B.biom -o ./table\_ITS1\_R1.txt --to-tsv

c.- Convert again to biom file as follows: biom convert -i ./table\_ITS1\_R1.txt -o table\_ITS1\_R1G.biom --table-type="OTU table" --to-json

d.- Add the taxonomy: biom add-metadata --sc-separated taxonomy --observation-header OTUID,taxonomy --observation-metadata-fp Taxonomy\_ITS1\_R1.txt -i table\_ITS1\_R1G.biom -o table\_ITS1\_R1\_wTax.biom

e.- Add the metadata: biom add-metadata -i table\_ITS1\_R1\_wTax.biom -o table\_ITS1\_R1\_wTaxMet.biom --sample-metadata-fp ../Mapping\_file\_Vaginal\_ITS\_sequencing\_S.csv

1. Apply abundance and contamination filters to qiime2 features.

We got everything into phyloseq, including representative fasta sequences.

Interesting to note that the taxonomic classification using the new UNITE database V10 now includes something called Species Hypotheses (SH). SH is a concept used by UNITE to group sequences into species-like entities based on sequence similarity thresholds. Here's what it means in more detail:

**Species Hypotheses (SH)**

* **Definition**: SHs are clusters of fungal DNA sequences that are considered to represent a species or species-like unit. These clusters are determined using sequence similarity thresholds applied to marker genes, such as the **ITS region**, which is the primary fungal DNA barcode.
* **Format**: The taxonomy output might include labels like shXXXXX.X. For example, sh1578002.08FU:
  + **sh1578002**: The unique identifier for the Species Hypothesis.
  + **.08**: The version of the UNITE database where this SH was defined.
  + **FU**: Indicates that this SH is from fungi (as opposed to other kingdoms).

**Why Use SHs?**

* SHs are particularly useful in fungal taxonomy because the taxonomy of fungi is not yet fully resolved, and many fungal sequences in databases are from uncultured or undescribed species. The SH system provides a stable way to refer to groups of sequences even if their formal taxonomic names are unknown or incomplete.
* It helps researchers work with operational taxonomic units (OTUs) or amplicon sequence variants (ASVs) that lack detailed taxonomic information.

**Practical Implications in QIIME2:**

* When using the UNITE database with QIIME2, if a sequence can be assigned to a known species, the output may include the full taxonomy (e.g., k\_\_Fungi; p\_\_Ascomycota; c\_\_Eurotiomycetes; ...).
* If a species-level assignment is not available, the sequence may be assigned to an SH (e.g., k\_\_Fungi; p\_\_Ascomycota; c\_\_Dothideomycetes; ...; shXXXXX.X), representing a species hypothesis without a formal name.

After the abundance filter we end up with 593 fungal features and 154 samples (with 147 having more than 5000 sequences per sample). Before we had 144 samples left after all filters.