# Illumina sequence analysis pipeline (written by: Timothy Lim, version 30/06/2025)

#Changelog: added decontamination pipeline

## Un-tar raw reads file

#Set up a new folder, change directory to that folder (for un-tar), and un-tar the file to that folder

cd ~/folder\_for\_untar

tar -xvf "~/sequences.tar"

## Rename fastq files to CASAVA format

#Move all the files without CASAVA naming format (see if there is lane number *L* in the name) to a folder first, then change directory and execute script

cd ~/to\_be\_renamed\_CASAVA

source ~/rename\_to\_casava.sh

#Move the renamed files from casava\_renamed folder to raw\_reads\_combined folder, then copy the remaining files (already in CASAVA naming format) from their respective folders to into the combined folder

## Fastp pipeline

#launch fastp module

module load fastp/0.20.0

#Change directory which contains the raw reads (I made backup raw\_reads\_combined folder)

cd ~/raw\_reads\_combined\_fastp

#Run fastp

source ~/Scripts/FASTP.sh

#Delete files without prefix “trimmed” in that folder, then remove prefix-trimmed after fastp

source ~/Scripts/RemoveTrimmed.sh

#Sort by size in ascending order, and delete the fastq files with 1kb.

## FIGARO pipeline

#Activate FIGARO

source ~/miniconda3/bin/activate

conda activate figaro

#Set QT\_QPA\_PLATFORM to avoid error, and use FIGARO

export QT\_QPA\_PLATFORM='offscreen'

python ~/miniconda3/envs/figaro/figaro-master/figaro/figaro.py -i ~/raw\_reads\_combined\_fastp -o ~/FIGARO\_output -f 20 -r 20 -a 490 -F illumina -m 20

#View FIGARO results

cd ~/FIGARO\_output

less trimParameters.json

## QIIME2 pipeline

#Requested for 367GB memory and 48 cores (adjust the memory and core numbers as you see fit) in genomics partition, using the following command:

smux new-session --cpuspertask=48 --mem=367G --time=0-4:00:00 --partition=genomics --qos=genomics

#If you do not have access to genomics partition, you can request for a normal partition, using the following command:

smux new-session --cpuspertask=36 --mem=367G --time=7-

#Attach to assigned session using the following command (based on JOB ID using *show\_job* command, in this case I was assigned JOB ID 12345678):

smux attach-session 12354678

#Load QIIME2 version 2023.9-amplicon (latest available version on MASSIVE platform):

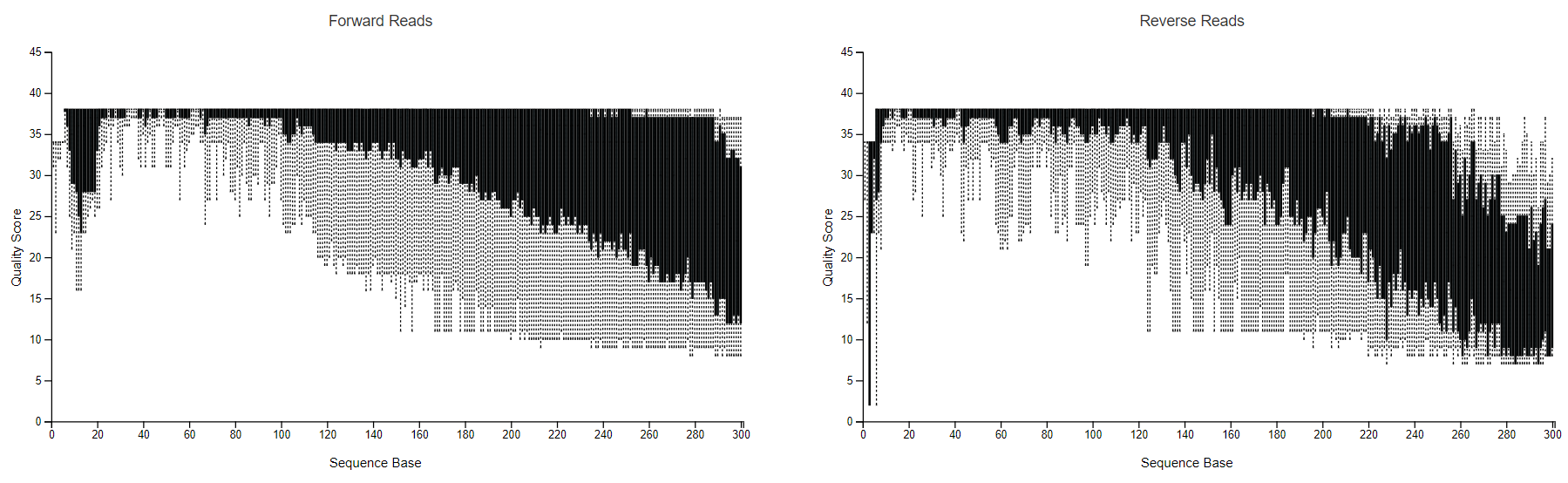
module load qiime2/2023.9-amplicon

#Import 2 single ended reads and making it QIIME format:

qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-format CasavaOneEightSingleLanePerSampleDirFmt --input-path ~/raw\_reads\_combined\_fastp --output-path ~/Import/ImportedReads.qza

#Demultiplexing data and to generate summary of demultiplexing (example of demux data included)

qiime demux summarize --i-data ~/Import/ImportedReads.qza --o-visualization ~/Import/ImportedReads.qzv



#Perform DADA2 denoise with parameters suggested by FIGARO (see *FIGARO and FASTP setup docx* file for more information): forward trimming length = 20; reverse trimming length = 20; forward truncation length = 296; reverse truncation length = 254; max error forward and reverse = 5

qiime dada2 denoise-paired --i-demultiplexed-seqs ~/Import/ImportedReads.qza --p-trunc-len-f 296 --p-trunc-len-r 254 --p-trim-left-f 20 --p-trim-left-r 20 --p-n-threads 32 --p-no-hashed-feature-ids --p-max-ee-f 5 --p-max-ee-r 5 --output-dir ~/dada2 --verbose

#Tabulate metadata to provide state on what got excluded/removed in denoising process

qiime metadata tabulate --m-input-file ~/dada2/denoising\_stats.qza --o-visualization ~/dada2/denoising\_stats.qzv

qiime feature-table tabulate-seqs --i-data ~/dada2/representative\_sequences.qza --o-visualization ~/dada2/representative\_sequences.qzv

#Filter out feature present in less than 1% of samples

qiime feature-table filter-features --i-table ~/dada2/table.qza --p-min-samples 2 --o-filtered-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza

#summarise feature table 🡪 provides what happened, where sequences went, how many crap sequences found; table shows the final number of sequences per sample

qiime feature-table summarize --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza --o-visualization ~/dada2/filtered\_tables/sample-contingency-filtered-table.qzv

#**IF** you want to filter out some non-relevant samples, and summarise as qzv file, then follow the command below. For this guideline it is assumed that there are some non-relevant samples to be filtered off.

qiime feature-table filter-samples --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza --m-metadata-file ~/Diversity/WhatYouWantTobeRetainedInMetadataPool.txt --o-filtered-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza

qiime feature-table summarize --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza --o-visualization ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qzv

After doing these, I then went to create mapping file, which contained the following parameters:

1. #SampleID

2. Env (Environment, or sample type)

3. SourceSink

4. SequencingMethod

#Perform decontamination via *decontam*, using prevalent method (i.e., method extraction blanks (MEB) as control)

qiime quality-control decontam-identify --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza --m-metadata-file ~/Diversity/MappingFile.txt --p-method prevalence --p-prev-control-column Env --p-prev-control-indicator Meb --o-decontam-scores ~/Decontam/decontam\_results.qza --verbose

qiime quality-control decontam-score-viz --i-decontam-scores ~/Decontam/decontam\_results.qza --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza --o-visualization ~/Decontam/decontam\_results.qzv

qiime quality-control decontam-remove --i-decontam-scores ~/Decontam/decontam\_results.qza --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table.qza --o-filtered-table ~/Decontam/sample-contingency-filtered-table-decontaminated.qza

qiime feature-table summarize --i-table ~/Decontam/sample-contingency-filtered-table-decontaminated.qza --o-visualization ~/Decontam/sample-contingency-filtered-table-decontaminated.qzv

#Create phylogenetic tree, using mafft system. Note that the *--p-n-threads* parameter needs to be lower than the allocated core for your session.

qiime phylogeny align-to-tree-mafft-fasttree --i-sequences ~/dada2/representative\_sequences.qza --o-alignment ~/dada2/aligned\_representative\_sequences.qza --o-masked-alignment ~/dada2/masked\_aligned\_representative\_sequences.qza --o-tree ~/dada2/unrooted\_tree.qza --o-rooted-tree ~/dada2/rooted\_tree.qza --p-n-threads 35 --verbose

#Use alpha rarefaction to identify the rarefaction depth; using max depth of 12345 (3rd quartile sequences; it will be different in all cases), 10000, and 20000

qiime diversity alpha-rarefaction --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza --i-phylogeny ~/dada2/rooted\_tree.qza --p-max-depth 12345 --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/alpha-rarefaction.qzv

qiime diversity alpha-rarefaction --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza --i-phylogeny ~/dada2/rooted\_tree.qza --p-max-depth 10000 --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/alpha-rarefaction\_10000.qzv

qiime diversity alpha-rarefaction --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza --i-phylogeny ~/dada2/rooted\_tree.qza --p-max-depth 20000 --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/alpha-rarefaction\_20000.qzv

#Find sequencing depth from alpha rarefaction plotting, then go with diversity core-metrics (and rarefaction)

qiime diversity core-metrics-phylogenetic --i-phylogeny ~/dada2/rooted\_tree.qza --i-table ~/dada2/filtered\_tables/sample-contingency-filtered-table-v1.qza --p-sampling-depth 10000 --m-metadata-file ~/Diversity/MappingFile.txt --output-dir ~/Phylo/core-metrics-results

#Since rarefaction depth is set at 10000, remove SampleIDs that has reads below 10000

#Alpha diversity analysis

qiime diversity alpha-group-significance --i-alpha-diversity ~/Phylo/core-metrics-results/faith\_pd\_vector.qza --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/core-metrics-results/Alpha-group-significance/faith-pd-group-significance.qzv

qiime diversity alpha-group-significance --i-alpha-diversity ~/Phylo/core-metrics-results/observed\_features\_vector.qza --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/core-metrics-results/Alpha-group-significance/observed\_features\_vector\_visual.qzv

qiime diversity alpha-group-significance --i-alpha-diversity ~/Phylo/core-metrics-results/evenness\_vector.qza --m-metadata-file ~/Diversity/MappingFile.txt --o-visualization ~/Phylo/core-metrics-results/Alpha-group-significance/evenness-group-significance.qzv

#Beta diversity analysis (Create another mapping file just for beta diversity analysis. You can duplicate from previously created ones, but remember to not leave any group with only one unique sample. In this case since we previously set sampling depth at 10000, therefore remove those samples under 10000 in mapping file/metadata file)

qiime diversity beta-group-significance --i-distance-matrix ~/Phylo/core-metrics-results/unweighted\_unifrac\_distance\_matrix.qza --m-metadata-file ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt --m-metadata-column Env --o-visualization ~/Phylo/core-metrics-results/Beta-group-significance/unweighted-unifrac-body-site-significance.qzv --p-pairwise

qiime diversity beta-group-significance --i-distance-matrix ~/Phylo/core-metrics-results/weighted\_unifrac\_distance\_matrix.qza --m-metadata-file ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt --m-metadata-column Env --o-visualization ~/Phylo/core-metrics-results/Beta-group-significance/weighted-unifrac-body-site-significance.qzv --p-pairwise

qiime diversity beta-group-significance --i-distance-matrix ~/Phylo/core-metrics-results/jaccard\_distance\_matrix.qza --m-metadata-file ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt --m-metadata-column Env --o-visualization ~/Phylo/core-metrics-results/Beta-group-significance/jaccard\_distance\_matrix\_visual.qzv --p-pairwise

qiime diversity beta-group-significance --i-distance-matrix ~/Phylo/core-metrics-results/bray\_curtis\_distance\_matrix.qza --m-metadata-file ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt --m-metadata-column Env --o-visualization ~/Phylo/core-metrics-results/Beta-group-significance/bray\_curtis\_distance\_matrix\_visual.qzv --p-pairwise

**Taxonomy**

wget all .qza files from: <https://www.dropbox.com/sh/ibpy9j0clw8dzwm/AAAIVuYnqUzAOxlg2fijePQna/ver_0.02?dl=0&subfolder_nav_tracking=1> (Remember to change directory cd; Wget “link of the files”; remove ?dl=0 at the end of each file)

qiime feature-classifier extract-reads --i-sequences ~/Taxonomy/SILVA-138-SSURef-Full-Seqs.qza --p-f-primer GCCTACGGGNGGCWGCAG --p-r-primer GGACTACHVGGGTATCTAATCC --p-trunc-len 500 --o-reads ~/Taxonomy/SILVA-138-SSURef-Full-Seqs\_trained.qza

qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads ~/Taxonomy/SILVA-138-SSURef-Full-Seqs\_trained.qza --i-reference-taxonomy ~/Taxonomy/Silva-v138-full-length-seq-taxonomy.qza --o-classifier ~/Taxonomy/v138\_trained\_classifier.qza

qiime feature-classifier classify-sklearn --i-classifier ~/Taxonomy/v138\_trained\_classifier.qza --i-reads ~/dada2/representative\_sequences.qza --o-classification ~/Taxonomy/v138\_taxonomy.qza --verbose (**IF** you encounter error in this, most likely it is due to not enough memory)

qiime metadata tabulate --m-input-file ~/Taxonomy/v138\_taxonomy.qza --o-visualization ~/Taxonomy/v138\_taxonomy.qzv

#Export script from qza to tsv format

qiime tools export --input-path ~/Taxonomy/v138\_taxonomy.qza --output-path ~/Taxonomy/v138\_taxonomy

#Based on this result – I used the resulting csv to remove taxonomic allocations of less than 95% confidence. Saved document as .tsv and did the following (Open in Excel, sort and remove; remember to put #SampleID):

qiime tools import --type 'FeatureData[Taxonomy]' --input-path ~/Taxonomy/v138\_taxonomy/filtered\_taxonomy.txt --output-path ~/Taxonomy/v138\_taxonomy/filtered\_taxonomy.qza

#Filter original table before proceeding to plot bar plots (use rarefied table instead of normal table)

qiime feature-table filter-features --i-table ~/Phylo/core-metrics-results/rarefied\_table.qza --m-metadata-file ~/Taxonomy/v138\_taxonomy/filtered\_taxonomy.qza --o-filtered-table ~/Taxonomy/table-filtered-95-tax.qza

#Construct bar plots

qiime taxa barplot --i-table ~/Taxonomy/table-filtered-95-tax.qza --i-taxonomy ~/Taxonomy/v138\_taxonomy.qza --m-metadata-file ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt --o-visualization ~/Taxonomy/taxa\_bar\_plot.qzv

#Taxonomy bar plots were made for both full and 95% confidence allocations

## SourceTracker

#export the filtered feature table (rarefied table) as a .biom file, create a .tsv from .biom file for input to SourceTracker: [https://forum.qiime2.org/t/exporting-and-modifying-biom-tables-e-g-adding-taxonomy-annotations/3630](https://forum.qiime2.org/t/exporting-and-modifying-biom-tables-e-g-adding-taxonomy-annotations/3630%20%20%20%20)

qiime tools export --input-path ~/Phylo/core-metrics-results/rarefied\_table.qza --output-path ~/Phylo/core-metrics-results

biom convert -i ~/Phylo/core-metrics-results/feature-table.biom -o ~/Phylo/core-metrics-results/feature-table.tsv --to-tsv

biom head -i ~/Phylo/core-metrics-results/feature-table.tsv

#QIIME2 unload and SourceTracker load

*module unload qiime2/2023.9-amplicon*

*module load sourcetracker/2.0.1*

#SourceTracker run, run following command 5 times, change output folder name from *stout1* to *stout2*… to *stout5*

sourcetracker2 gibbs -i ~/Phylo/core-metrics-results/feature-table.tsv -m ~/Diversity/MappingFile\_RemoveBelow10000\_Betadiversity.txt -o ~/SourceTracker/Exports/stout5 --burnin 100 --source\_rarefaction\_depth 10000 --sink\_rarefaction\_depth 10000

#Notes/suggestions on SourceTracker:

1. To allow improved sensitivity in the detection of faecal contributions, the remaining water samples (except for the water sample which was assigned as sink) from similar environment (e.g., neighbouring beaches) can be included as sources in addition to the faecal source samples. This is to account for background contributions to allow users have better understanding of the unknowns. As the source pool will be different for each sample, users can speed up the process by using the leave one out (LOO) pipeline to generate the SourceTracker commands. Refer to *Bioinformatics-IlluminaMGI/16S/LOO/* folder for more information.