Pipeline Overview

This Nextflow pipeline contains the following workflow/steps. Note that files with extension ending in `.qza` are `QIIME 2` "artifacts" and can be used as inputs directly with `QIIME 2`. Files ending in `.qzv` can be visualized directly in `QIIME 2 View`. You can also export most `.qza` files and `.qzv` into standard text/formats with `qiime tools export`. E.g. if you export `ccs_reps.qza`, you get a FASTA file for amplicon sequence variants (ASV) generated by `DADA2`

QC

This step uses `seqkit` to generate statistics such as length, quality and GC content of each read from input FASTQ files. In addition, user can use `--filterQ` parameter to use only reads above a certain QV value (default 30). The statistics are used to produce read quality and length distribution in the final HTML report. Output files includes: * `$outdir/filtered_input_FASTQ`: FASTQ files that have filtered to only reads above the threshold quality. * `$outdir/results/reads_QC` * `seqkit.summarised_stats.group_by_samples.tsv`: Read length and quality summary statistics for each sample. The *pretty* version of this file just makes it easy to read on the terminal. * `all_samples_cutadapt_stats.tsv`: Number of reads containing both F27 and R1492 primers. * `all_samples_seqkit.readstats.tsv`: Length, GC content and read quality for each individual read in all samples. * `all_samples_seqkit.summarystats.tsv`: Similar to `seqkit.summarised_stats.group_by_samples.tsv` but with Q1 and Q3 of read qualities.

Primers trimming and importing into QIIME 2 artifact

We use `cutadapt` in this step to trim away the full length F27-R1492 16S primers at both ends of each read. `cutadapt` will also reorientate each reads to the same orientation (important for ASV classification using Naive Bayes classifier in `QIIME 2`). Output files include: * `$outdir/trimmed_primers_FASTQ`: All FASTQ files that have been filtered and trimmed. These are the files used as inputs to `DADA2` * `$outdir/cutadapt_summary`: `cutadapt` report files. For summary of how many reads remained after trimming see `all_samples_cutadapt_stats.tsv` documented in QC section above.

After trimming the primers, the pipeline will load the high quality FASTQ files into `QIIME 2`. This will generate: * `$outdir/import_qiime`: Contains a file named `samples.qza` that contains data from all samples and ready for further analysis within the `QIIME 2` environment, e.g. `DADA2`. * `$outdir/summary_demux`: * `samples.demux.summary.qzv`: An interactive table of the number of reads in each sample that can be opened using `QIIME 2 View`. * `per-sample-fastq-counts.tsv`: Simple count of number of reads in each FASTQ.

DADA2 denoising into ASVs

This pipeline uses `DADA2` implemented within the `QIIME 2` set of plugins (Version 2022-2). However, small modification in the `DADA2` script from the plugin was done in order to allow using reads without trimming primers. This is because many publicly available datasets already have their primers trimmed. By default the plugin will not work with them. With the modification, user can specify `--skip_primer_trim` when running the pipeline to directly use trimmed FASTQ files. In addition, we set the default parameter for minQ in DADA2 filtering step to 0 which is the default DADA2 parameter (but set to 3 by QIIME 2 authors) in favor of the maxEE filter (See discussion `https://github.com/benjjneb/dada2/issues/1125` and `https://github.com/benjjneb/dada2/issues/1216`).

Note that after denoising into ASVs, the pipeline by default will filter to ASVs that exist in a minimum of 2 samples and must have at least 5 supporting reads. These can be changed using `--min_asv_totalfreq` and `--min_asv_sample` parameters. Set both to zeros to disable filtering.

In addition, after denoising and filtering into ASVs, the pipeline automatically calculate a rarefaction depth that includes 90% of the input samples. This is then used to produce rarefaction curve and to calculate core diversity metrics. The automatic calculation works well if majority of the samples have good amount of reads, but can sometimes fail in sequencing run with uneven read distribution (large amount of samples with low reads). The rarefaction depth can be overridden with the `--rarefaction_depth` parameter.

Output from DADA2 includes: * `$outdir/dada2`: * `dada2-ccs_stats.qza`: `DADA2` statistics in such as number of chimeric reads and denoised reads. This is reported in the final HTML report. * `seqtab_nochim.rds`: R object post-DADA2 analysis. This can be loaded into R for further analysis using R. * `dada2-ccs_rep_filtered.qza`: `DADA2` ASV sequences after filtering for minimum samples and frequency (see above). This is also exported into a standard FASTA file as `dada2_ASV.fasta` in this folder (and in the `results` folder). * `dada2-ccs_table_filtered.qza`: `DADA2` ASV frequencies table (i.e. How many reads belong to each ASV). * `$outdir/results`: * `rarefaction_depth_suggested.txt`: Sampling depth covering >90% of the input samples. * `alpha-rarefaction-curves.qzv`: Rarefaction curve that can be opened in `QIIME 2 View` to understand features saturation.

Taxonomy classification and barplots

ASVs generated from DADA2 are classified using VSEARCH and Naive Bayes classifier (See FAQ here). Output includes: * `$outdir/results`: * `best_tax_merged_freq_tax.tsv`: (For Naive Bayes classifier) Table containing ASV ID, sequences, taxonomy, assignment confidence and the counts of the ASVs in each sample. * `feature-table-tax.biom`: (For Naive Bayes classifier) BIOM format containing ASVs count and taxonomy information. Can be imported into popular packages such as `phyloseq` for downstream processing and visualization. * `best_taxonomy_withDB.tsv`: (For Naive Bayes classifier) Taxonomy assigned to each ASV. This file contains the information on which database was used to assign a specific ASV and can be useful for understanding differences in assignments. * `vsearch_merged_freq_tax.tsv`: (For VSEARCH classifier) Table containing ASV ID, sequences, taxonomy, assignment confidence and the counts of the ASVs in each sample. * `feature-table-tax_vsearch.biom`: (For VSEARCH classifier) BIOM format containing ASVs count and taxonomy information. Can be imported into popular packages such as `phyloseq` for downstream processing and visualization. * `taxonomy_barplot_vsearch.qzv`: (For VSEARCH classifier) Interactive taxonomy barplot that can be visualized with `QIIME 2 View`. * `taxanomy_barplot_nb.qzv`: (For Naive Bayes classifier) Interactive taxonomy barplot that can be visualized with `QIIME 2 View`. * `$ourdir/nb_tax` contains classification results using Naive-Bayes approach (`assignTaxonomy` in `DADA2`): * The individual `tsv` file represents classification with a single database. E.g. `gtdb_nb.tsv` refers to classification with just GTDB.

Phylogenetic tree and diversity metrics

This pipeline generates simple phylogenetic tree and diversity metrics using the `qiime phylogeny align-to-tree-mafft-fasttree` and `diversity core-metrics-phylogenetic` plugins. Note that default parameters are used and the pipeline makes no attempt to optimize the steps here, but merely to provide quick results. Please refer to plugins documentation for more information on the methods and outputs. Output includes: * `$outdir/results/phylogeny_diversity`: * `mafft_alignment.qza` and `mafft_alignment_masked.qza`: MAFFT alignment (masked or not) between ASVs from DADA2. * `phylotree_mafft.qza` and `phylotree_mafft_unrooted.qza`: Unrooted and rooted (midpoint-rooting) tree from FASTTREE. * `phylotree_mafft.qza` can be visualized directly using iToL. After uploading to iToL, you can also upload `$outdir/results/taxonomy.vsearch.qza` or `$outdir/results/best_tax.qza` to annotate taxonomies on the phylogenetic tree. * `core-metrics-diversity`: Diversity metrics including Bray-Curtis, Jaccard index and Unifrac distances. The various "Emperor" `.qzv` file can be viewed using QIIME 2 View for interactive visualization of the diversity metrics. The distance matrices are used for the MDS plot in the final HTML report.

Report and visualization

At the end of the pipeline, a report will be generated to report the statistics from the important steps. In addition, "Krona" plots will be generated to allow visualization of the communities in each sample. All HTML files should open in modern internet browser such as Chrome/Safari. * `$outdir/results/visualize_biom.html`: Main HTML report containing many useful statistics and tables of top taxonomies. * `$outdir/results/krona.qzv`: Krona plots that can be opened using `QIIME 2 View`. There is also a separate folder named "krona_html" containing standalone HTMLs if you prefer to open directly in browser without using `QIIME 2 View`. By default this uses the VSEARCH taxonomy output.