A

Parameters in config.yaml	Output to evaluate
snakemake dada2_pe_denoise # use manifest file to import fastq.gz sequence files manifest_pe: 00-data/manifest_pe.csv # use pre-imported reference database refseqs_qza: 01-imported/refseqs.qza reftax_qza: 01-imported/reftax.qza # choose dada2 parameters based on fastq error profiles dada2pe_trunc_len_f: 240 dada2pe_trunc_len_r: 190	fastq_summary.qzv • median Q-score <30 occurs at fwd. position 267 and rev. position 233 -> trimming at 240 and 190 is acceptable for this amplicon (<300 bp) • 16 samples (fwd. & rev.) all have 1000 reads per sample (test dataset) repseqs.qzv & repseqs_lengths.tsv • of 301 repseqs, most are 253 bp and max is 255 bp except two that are much longer (416 bp, 417 bp) -> filter by length max 260 bp table.qzv • of 16 samples, lowest count per sample is 511 -> set core sampling depth (rarefaction) to 500 (check again after filtering)
snakemake dada2_pe_taxonomy_unfiltered # choose taxonomic classification method (*) classify_method: consensus-vsearch	taxonomy.qzv • 10 repseqs are "Unassigned" and 2 repseqs are "d__Eukaryota" -> filter by keywords "unassigned,eukaryota" taxa_barplot.qzv • the contribution of Unassigned and Eukaryota groups is <10%; still want to filter them
snakemake dada2_pe_diversity_unfiltered # choose MSA parameters (*) alignment_method: muscle alignment_muscle_maxiters: 2 alignment_muscle_diags: -diags # choose outlier detection parameters (*) odseq_distance_metric: linear odseq_bootstrap_replicates: 100 odseq_threshold: 0.025 # choose subsampling (rarefaction) level core_sampling_depth: 500 alpha_max_depth: 500 # choose beta group significance parameters (*) beta_group_column: region beta_group_method: permanova beta_group_pairwise: --p-pairwise	rooted_tree.qzv • feature metadata coloring confirms we should filter Unassigned and Eukaryota repseqs_properties.pdf • confirms we should filter Unassigned and Eukaryota and sequences longer than 260 bp; don't need to filter all outliers alpha_rarefaction.qzv • observed features plateaus at ~450–500 sequences per sample observed_features_group_significance.qzv • difference between regions (Open Water vs. Western Boundary) is not significant by Kruskal–Wallis, but filter size is significant unweighted_unifrac_emperor.qzv • separation by region (axis 2) and filter size (axes 1 & 2) beta_group_significance.qzv • distance based on region is significant
snakemake dada2_pe_report_unfiltered # choose theme for html report report_theme: github	report_dada2-pe_unfiltered.html • a summary of the results and metadata and links to output files are presented in this HTML report
snakemake dada2_pe_report_filtered (**) # choose terms to filter from taxonomy exclude_terms: unassigned,eukaryota # choose repseq length limits repseq_min_length: 0 repseq_max_length: 260	table.qzv • of 16 samples, lowest count per sample is 507 -> it was ok to leave subsampling (rarefaction) depth at 500 rooted_tree.qzv • feature metadata coloring confirms Unassigned and Eukaryota were removed, and tree topology is more homogeneous repseqs_properties.pdf • confirms long sequences and Unassigned and Eukaryota were removed, resulting in fewer gaps in the multiple sequence alignment report_dada2-pe_filtered.html • a summary of the results and metadata and links to output files are presented in this HTML report

(*) these steps can be defined before starting the workflow, as they do not depend on the output of previous steps

(**) all steps are being run at once by using the report command