16S rRNA microbial community composition analysis in Qiime2

Input data: FASTQ files.
16S rRNA amplicons
V3-V4 domain (~450bp).

Workflow details (presented on schematics below):

1. **Fastqc** provides basic statistics on each FASTQ file (e.g. average length, GC content, quality values, etc). MultiQC summarizes the fastqc results into one HTML report file (website visualization).

2. **Import** of data to Qiime2 (further named q2) and converting to qiime2 artifact. The procedure requires metafile – a TSV (tab delimited table) file with numeric and/or categorical information characterizing the samples that allows to group and compare samples of similar characteristic.

Example of metafile:

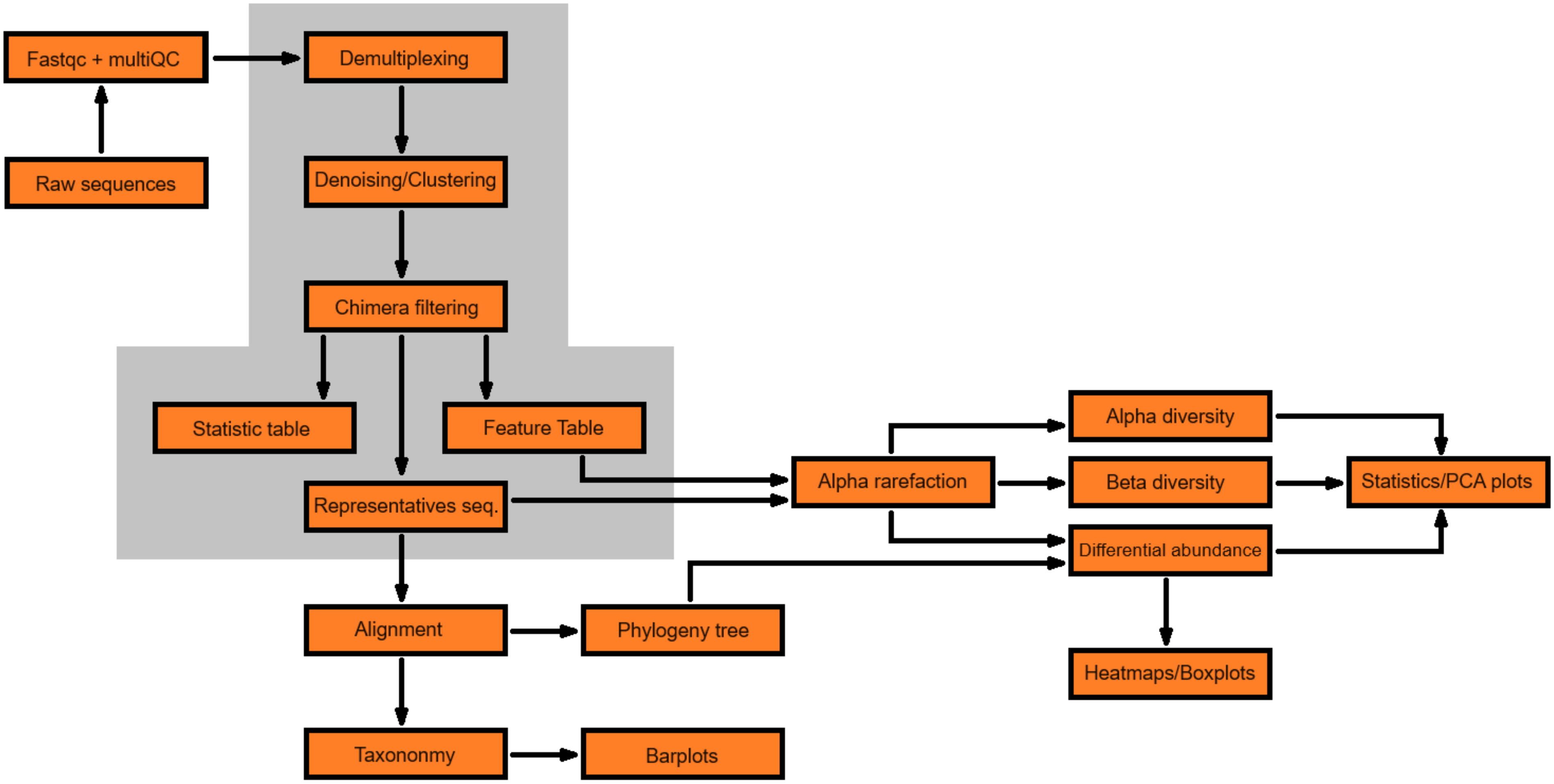
#SampleID	type	sex	age	city	province	dilution	CFU (x10^6)
A1	control	male	11	Johannesburg	Gauteng	1x	1.4
A2	control	male	12	Pretoria	Gauteng	2x	2.2
A3	case	female	14	Durban	KZN	1.5x	10
B1	control	female	14	Cape Town	WC	10x	3.1
B2	case	male	12	Pretoria	Gauteng	2x	7.1
B3	control	female	12	Kimberley	NC	2x	2.9
C1	case	male	13	Stellenbosch	WC	1.3x	1.6

dada2 analysis – gray area on the schematics:

3. **Demultiplexing** – This step looks for barcode sequences at the beginning of your reads (5’ -end) with a certain error tolerance, removes them, and returns sequence data separated by each sample.

4. **Denoising/Clustering** – The denoising step is removing non-biological sequences (e.g. primers, sequencing adapters, PCR spacers, etc). Next clusters sequences into amplicon sequence variants (ASVs) or operational taxonomic units (OTUs) to collapse similar sequences (e.g., those that are ≥ 97% similar to each other) into single replicate sequences also known as OTU picking.

5. **Chimera filtering** – removing of chimeric reads that may be caused by sequencing a chromosomal aberration or by technical issues during sample preparation.
dada2 output are 3 artifacts:
   a) **statistics table** – summarize data reduction on each step and final percent of sequences that survived each step of filtration, trimming and merging.
   b) **representative sequences** – OTU/ASV FASTA file with consensus sequences for grouped reads based on default distance metric. This includes so called singletons – ASVs derived from only one pair of reads – which usually highly uncertain and therefore are removed from further analysis.
   c) **feature table** – essential for further analysis steps table containing correlation between samples, metafile information and ASVs.



Workflow schematics – gray area indicates dada2 analysis step

6. **Alignment** - diversity analyses rely on the phylogenetic similarity between individual features that can then be used by other downstream analyses. Each ASV is compared and clustered with closes “relative” sequence which allows to draw phylogenetic relationship tree between all ASVs.

7. **Taxonomy** - comparing query sequences (i.e. ASVs) to a reference database of sequences with known taxonomic composition by simply finding the closest alignment to a *taxonomy classifier* to determine the closest taxonomic affiliation with some degree of confidence or consensus, based on alignment, k-mer frequencies, etc. The reference classifiers can be build base on few available databases and programs: e.g. greengenes or silva (for 16S, 18S and other rRNAs), BLAST or Kraken2 (for rRNA and other markers). This step allows to classify ASV sequences to different taxa units (order, phyla, kindom, genra, family, class and species). Outcome of this step is a barplot representing taxonomic diversity between and within samples.

8. **Alpha rarefaction** - is a method for normalization via sub-sampling without replacement and is commonly used as a workaround for the issue of uneven sequencing depth. Rarefaction occurs in two steps: first, samples which are below the rarefaction depth are filtered out of the feature table. Then, all remaining samples are subsampled without replacement to get to the specified sequencing depth. It’s both important and sometimes challenging to select a rarefaction depth for diversity analyses.

9. **Alpha diversity** – within sample diversity, consists of few metrics:
    - Shannon’s diversity index (a quantitative measure of community richness)
    - Observed Features (a qualitative measure of community richness)
    - Faith’s Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features)
    - Evenness (or Pielou’s Evenness; a measure of community evenness)

10. **Beta diversity** - between samples diversity, consists of few metrics:
    - Jaccard distance (a qualitative measure of community dissimilarity)
    - Bray-Curtis distance (a quantitative measure of community dissimilarity)
    - unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
    - weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)

11. **Differential abundance** - Compositionality refers to the issue of dealing with proportions in compositions between samples or features. To account for differences caused by sequencing depth, microbial abundances are typically interpreted as proportions (e.g. relative abundance). The data will be analyzed based on 3 types of correlations:
    a) Correlation clustering - if we don’t have relevant prior information about how to cluster together organisms, we can group them together based on how often they co-occur with each other.
    b) Gradient clustering - uses a metafile categories to cluster taxa found in similar sample types. For example evaluating if pH is a driving factor of taxa changes observed when pH values change.
    c) Phylogenetic analysis – utilizes phylogenetic tree (e.g. rooted-tree from previous step) to infer relationships between taxa.

Time frame table

Some of the steps strongly depend on amount of sequences in each sample and diversity between them. 16S amplicon in bacteria is around 450bp long which also extends the time of few steps in the analysis. The below proposed time frame is standard for this type of analysis, nevertheless can change depending on the initial data quality and quantity of sequences. Based on the data provided by the client, the time frame for execution of the project will be **120 working hours**.

Additional remarks

In case only one side of each sequence (e.g. R1/Forward) is available due to technical issues during sequencing, the analysis is still possible. The results will not represent the true diversity of the samples but rather approximated, general version. The amount of ASV might be higher, than when both sides for each read are used, due to shorter length of the analyzed sequences (half the length ~225bp) and therefore lesser similarity during clustering step. This may lead to higher alpha diversity and over-representation of some species.