Amplicon Data Analysis Using QIIME2

You will do:

- Quality filtering.
- Denoising.
- Picking of features and representative sequences.
- Assign taxonomy to features.
- Phylogenetic tree treation.
- Calculate alpha and beta diversity.

The tutorial is based on Belén Carbonetto's exellent tutorial "16S based microbial diversity analysis" from Elixir Excelerate Workshop on Marine Metagenomics, Oeiras May 2018, and the official Qiime2 tutorials.

This exercise consists of six parts:

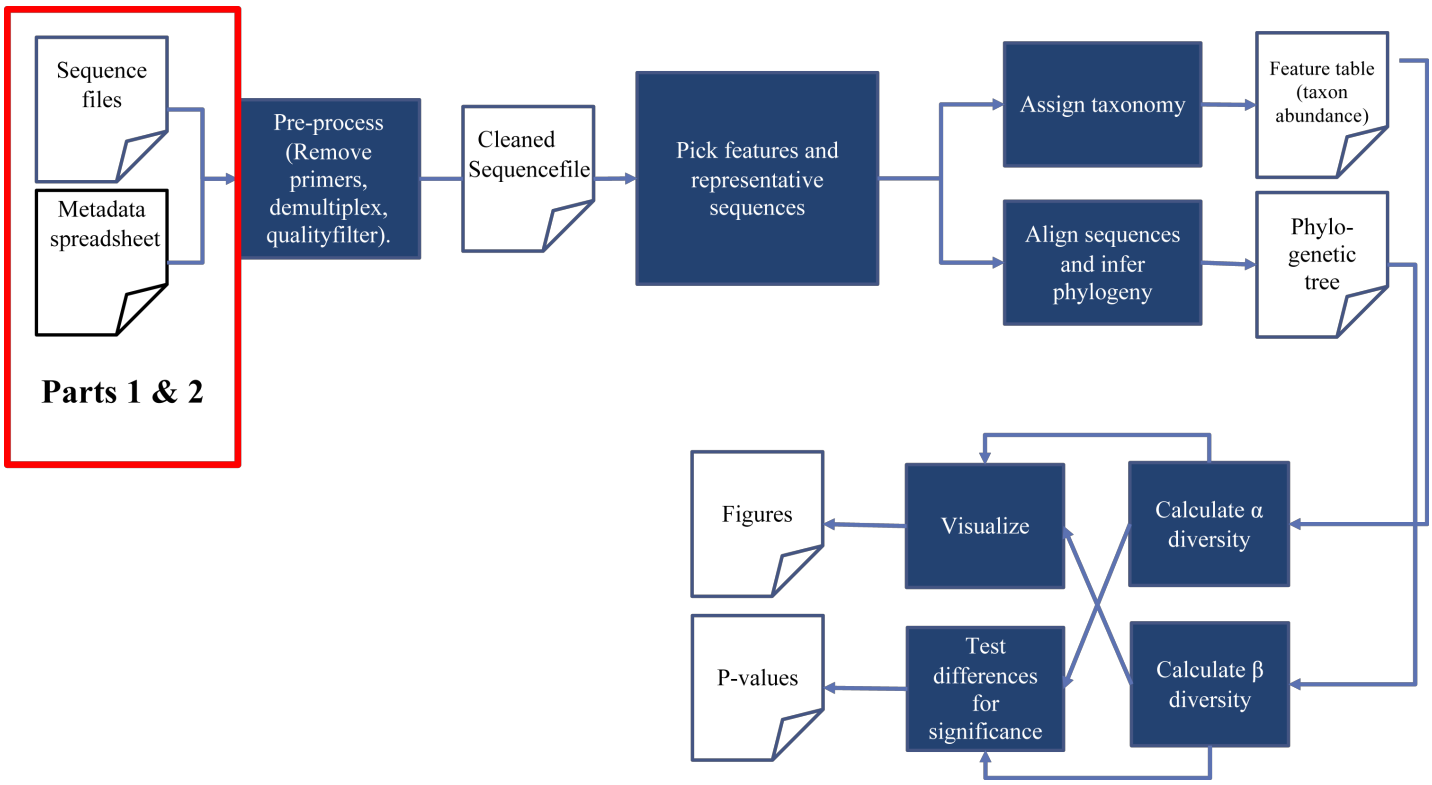
1. Data set and input files
2. Import files in Qiime2
3. Quality filtering, denoising and picking of features and representative sequences
4. Assign taxonomy to features
5. Align sequences and infer phylogeny
6. Calculate alpha- and beta diversity

Please note:

- Every task that you will perfom is printed in blue and marked with roman numerals e.g I].
- Questions for you to answer will be marked with ❓
- You can make tabbing work with Qiime2 if you run this command:

```
source tab-qiime
```

Part 1. Data set and input files



I] Locate the files containing the data you analyse with Qiime2 in this tutorial. You'll find them in the folder named `fastq` in `/home/practical/1/`.

```
cd /home/practical/1/fastq/

ls
```

The dataset we are going to use is originally from Nguyen & Landfald (2015). Four soft bottom sediment samples were collected at different geographical locations on the oceanic polar front in the Barents Sea at different depth and temperatures. Reads are amplicons of the V3–V4 region of the 16S rRNA gene. They are Illumina paired-end reads, already demultiplexed from the sequencing platform (two files/sample), with primers but no barcodes.

Besides the .fastq files you will find **metadata.tsv** in the `/home/practical/1/` folder.

This is a tab separated value table that should include **metadata needed to process sequences and test your hypothesesis**. The easiest way to make a mapping file is with a spreadsheet. Google Docs is prefered, because Excel will corrupt gene symbols, interpret numbers as dates, etc, changes that are not reversible.

It should look something like this:

#SampleID	Sample_Name	Station	Latitude	DepthGr	Depth	TempGr	Temp	TOCgr	TOC
#q2:types	categorical	categorical	categorical	categorical	numeric	categorical	numeric	categorical	numeric
37.S37	SampleA	St1	73	Deep	474	Over2.5	2.7	1	0.73
38.S38	SampleB	St2	73	Deep	460	Over2.5	2.8	3	2.24
44.S44	SampleC	St8	77	Shallow	189	Under1.5	1.1	1	1.21
45.S45	SampleD	St9	77	Shallow	194	Under1.5	1.2	1	1.21

The column labels are always in the first row, # indicates that the line is not going to be read as a sample. Sample IDs must be in the first column, the rest of the columns include metadata.

- Sample IDs needs to be unique.
- It should be 36 characters long or less.
- It should contain only ASCII alphanumeric characters ([a-z], [A-Z], or [0-9]), the period (.) character, or the dash (-) character.

The Sample IDs is the only metadata *required* by Qiime2. However you can, and should, add as many additional metadata as possible to improve your analysis. Metadata columns can contain *categorical* or *numerical* values. They must also include labels. You can read more about Qiime2 metadata files here

▶ **Note**

▶ **Note to QIIME 1 Users**

Progress tracker

Part 1 of 6 completed

Part 2. Import files in Qiime2

I] First thing first: you should create a working directory where you keep the output files of this session:

```
mkdir working_dir
```

II] Move into the working directory.

```
cd working_dir
```

All data that is used as input to QIIME2 should be in form of QIIME2 artifacts, which contain information about the type of data and the format of the data. So, the first thing you need to do is import the sequence data files into a QIIME2 artifact. The semantic type of our data is

`SampleData[SequencesWithQuality]`, i.e Sequences with quality scores (.fastq files), where each set of sequences is associated with a sample identifier (i.e. demultiplexed sequences).

In order to create an artifact we need a **manifest file** (a comma-separated .csv text file) to "guide" QIIME2 to the fastq files to import. This file maps sample identifiers to fastq.gz or fastq absolute filepaths that contain sequence and quality data for the sample. Normally you will have to create this file in the initial steps of your analysis. **We have provided it for you in this tutorial. It is in the `/home/practical/1/` folder. You just need to make some small changes to it**.

III] Open the manifest file (manifest.csv) in a text editor (Whatever makes you comfortable). It looks something like this:

```
#PairedEndFastqManifestPhred33_qiime2_session
sample-id,absolute-filepath,direction
37.S37,/PATH/fastq/37_S37_L001_R1_001.fastq,forward
37.S37,/PATH/fastq/37_S38_L001_R2_001.fastq,reverse
38.S38,/PATH/fastq/38_S38_L001_R1_001.fastq,forward
38.S38,/PATH/fastq/38_S38_L001_R2_001.fastq,reverse
44.S44,/PATH/fastq/44_S44_L001_R1_001.fastq,forward
44.S44,/PATH/fastq/44_S44_L001_R2_001.fastq,reverse
45.S45,/PATH/fastq/45_S45_L001_R1_001.fastq,forward
45.S45,/PATH/fastq/45_S45_L001_R2_001.fastq,reverse
```

IV] Please complete the file with the correct path for each fastq file in your computer (replace "PATH" for the real path to your fastq files). **Hint: the complete path starts with /home/**

Remember! Full/complete/absolute path!!!!

▶ **More about "Fastq manifest" formats**

V] You must activate the Qiime environment in the Terminal for Qiime2 to work.

```
conda activate qiime2
```

VI] You are now ready to create a QIIME2 artifact with your data:

```
qiime tools import /
    --type 'SampleData[PairedEndSequencesWithQuality]' /
    --input-path [/absolute/full/path/to/your/]manifest.csv /
    --output-path paired-end-demux.qza /
    --input-format PairedEndFastqManifestPhred33
```

NB! Backslash (/) is line break (line continuation) used to break up a command onto multiple lines. Leave them out and type the commands as one run-on line.

▶ **What did I just do?**

▶ **Note on the structure of arguments**

We guess you would like to know how your import worked? The .qza artifact does not directly show you this. You can view the descriptive statistics of the sample sizes (i.e. min, max, median, mean, histogram) and samples quality based on Quality Score per base by creating a QIIME2 **visualization artifact (.qzv)** from the data artifact **paired-end-demux.qza**.

All QIIME2 visualizers (i.e., commands that take a `–o-visualization` parameter) will generate a .qzv file.

```
    qiime demux summarize /
        --i-data paired-end-demux.qza /
        --o-visualization paired-end-demux.qzv
```

You can view these files with qiime tools view.

```
qiime tools view paired-end-demux.qzv
```

After you are done viewing type `q` in the terminal window to quit. This doesn't close web page (but the page becomes unreliable).

VII] Explore data

❓ What sizes are our samples? (how many reads/sequence count?).

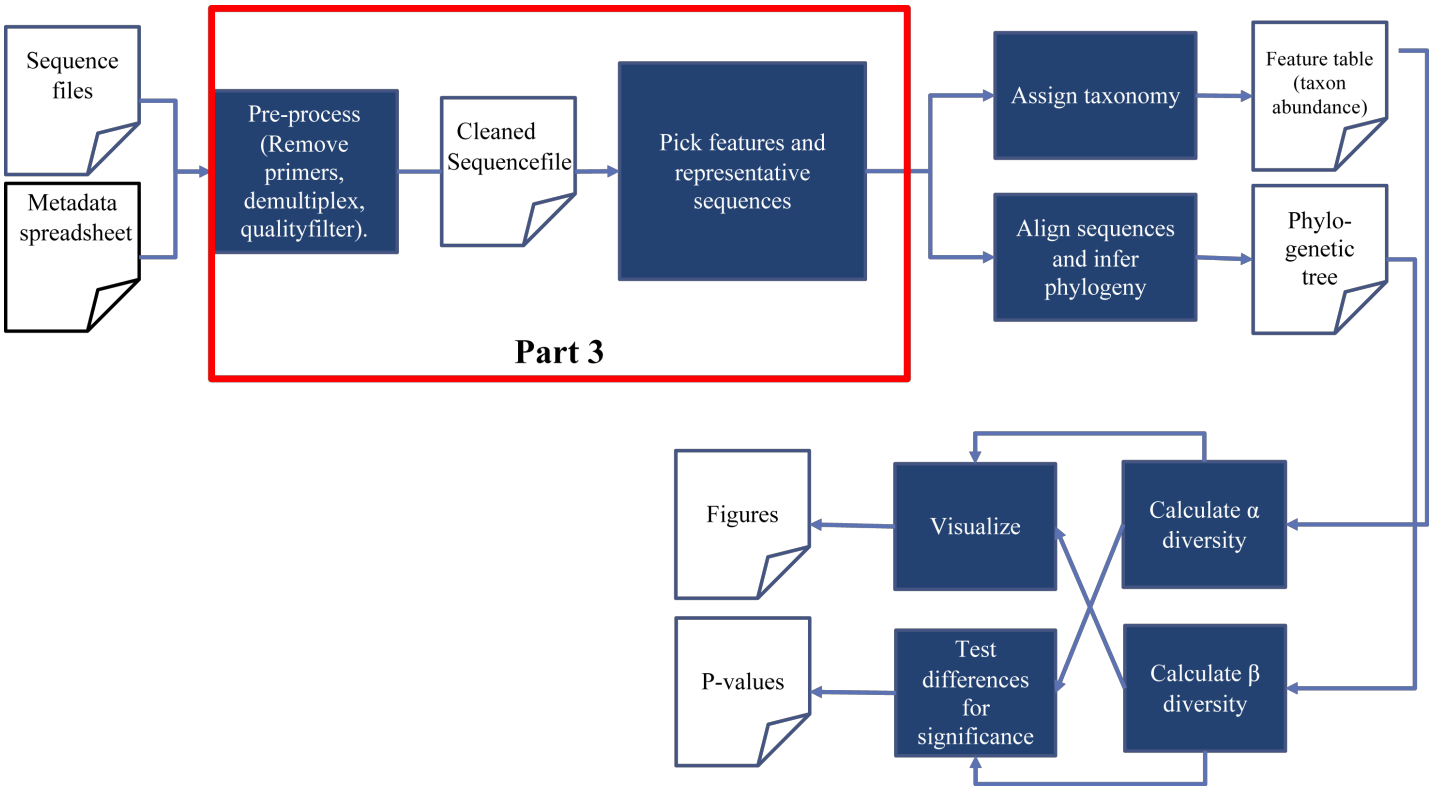
▶ **Solution** - Click to expand

▶ **Note on visualization**
▶ 💡 **ProTip .qza and .qzv**
▶ **Qiime2 tracks Artifact Provenance**

Progress tracker

Part 2 of 6 completed

Part 3. Quality filtering, denoising and picking of features and representative sequences



We are now ready to perform quality control and feature picking. QIIME 2 plugins are available for several quality control methods, including DADA2, Deblur, and basic quality-score-based filtering. In this tutorial we use DADA2. This method is based on correcting (where possible) Illumina sequencing mistakes. As implemented in the `q2-dada2` plugin, this quality control process will filter, trim and dereplicate data, merge paired reads and remove chimeric sequences. It will filter any phiX reads (commonly present in marker gene Illumina sequence data) that are identified in the sequencing data. Features (also called amplicon sequence variants or ASVs) are inferred by a *de novo* process in which biological (true) sequences are discriminated from errors on the basis of the expectation that biological sequences are more likely to be repeatedly observed than are error-containing sequences.

I] Reexamine the sequence quality from the **Interactive Quality Plot** tab in the **paired-end-demux.qzv** file you made. There are two plots per sample. The plot on the left presents the quality scores for the forward reads, and the plot on the right presents the quality scores for the reverse reads. You'll use these plots to determine what trimming parameters to use for denoising with DADA2, and then denoise the reads using `dada2 denoise-paired`.

The `dada2 denoise-paired` method requires two parameters that are used in quality filtering:

- `--p-trunc-len-f` : The position at which forward read sequences should be truncated due to decrease in quality.
- `--p-trunc-len-r` : Position at which reverse read sequences should be truncated due to decrease in quality.

❓ Based on the plots you see in paired-end-demux.qzv, what values would you choose for `--p-trunc-len` and `--p-trim-left` in this case?

▶ **Solution** - Click to expand

You also have the original amplicon primers in the dataset. The primers 341F and 785R are specific for the V3-V4 region. These are concidered non-biological nucleotides by DADA2 and should be removed. You can do this in Qiime2 with the `--p-trim-left-f` and `--p-trim-left-r` parameters which trims off the first m bases of fw and rv sequences. 342F is 17 nts long and 785R is 21 nts. We pass the values 17 and 21 to `--p-trim-left-f` and `--p-trim-left-r` respectively.

II] Run the following command in order to apply DADA2 on **paired-end-demux.qza**:

```
qiime dada2 denoise-paired /
    --i-demultiplexed-seqs paired-end-demux.qza /
    --p-trunc-len-f 290 /
    --p-trunc-len-r 220 /
    --p-trim-left-f 17 /
    --p-trim-left-r 21 /
    --output-dir DADA2 /
    --p-n-threads 8
```

This command may take up to 30 minutes to run (in the computers set up for the session, in your own computer this will depend on the processor and RAM features)

III] While you wait, open a new terminal, activate the QIIME2 environment and read the specifications of the parameters we have just used:

```
source activate qiime2-2018.8

qiime dada2 denoise-paired --help
```

❓ Are we removing chimeras? If so, which method are we using? In case you wonder: Chimeras??.

▶ **Solution**

❓ What are you expecting as output files?

▶ **Solution**

IV] You can now create a summary of the results and visualize it using `qiime tools view`

▶ **In case your run "ne-e-e-ver" finishes!**

The `feature-table summarize` command will give you information on how many sequences are associated with each sample and with each feature, histograms of those distributions, and some related summary statistics.

```
qiime feature-table summarize /
    --i-table DADA2/table.qza /
    --o-visualization DADA2/table-dada2.qzv /
    --m-sample-metadata-file metadata.tsv
```

V] Explore results:

❓ How many features has DADA2 resolved?

▶ **Solution**

❓ How many reads remain after quality filtering?

▶ **Solution**

❓ Which sample has the minimum frequency of reads? and the maximum?

▶ **Solution**

If you would like to visualize **representative sequences**, the `feature-table tabulate-seqs` command will provide a mapping of feature IDs to sequences, and provide links to easily BLAST each sequence against the NCBI nt database.

VI] Convert it to a visualization:

```
qiime feature-table tabulate-seqs /
    --i-data DADA2/representative_sequences.qza /
    --o-visualization DADA2/representative_sequences.qzv
```

VII] Explore results

And if you want to look at the statistics from the filteringsteps, they are in the **denoising_stats.qza** artifact.

VIII] Convert it to a visualization:

```
qiime metadata tabulate /
    --m-input-file DADA2/denoising_stats.qza /
    --o-visualization DADA2/denoising_stats.qzv
```

IX] Explore denoising stats.

You now have a feature table: **table.qza**, and you have assigned a representative sequence for each feature: **representative_sequences.qza**. Before you continue with the analysis. It is possible to export the feature table so you can explore data with softwares other than Qiime2 if needed.

X] In order to export **table.qza** run:

```
cd DADDA2/

qiime tools export /
    --input-path table.qza /
    --output-path exported_table

cd exported_table

ls
```

You will get a .biom format file.

XI] The .biome file can further be exported as a textfile:

```
biom convert -i feature-table.biom -o feature-table.txt --to-tsv

ls

cd ../../
```

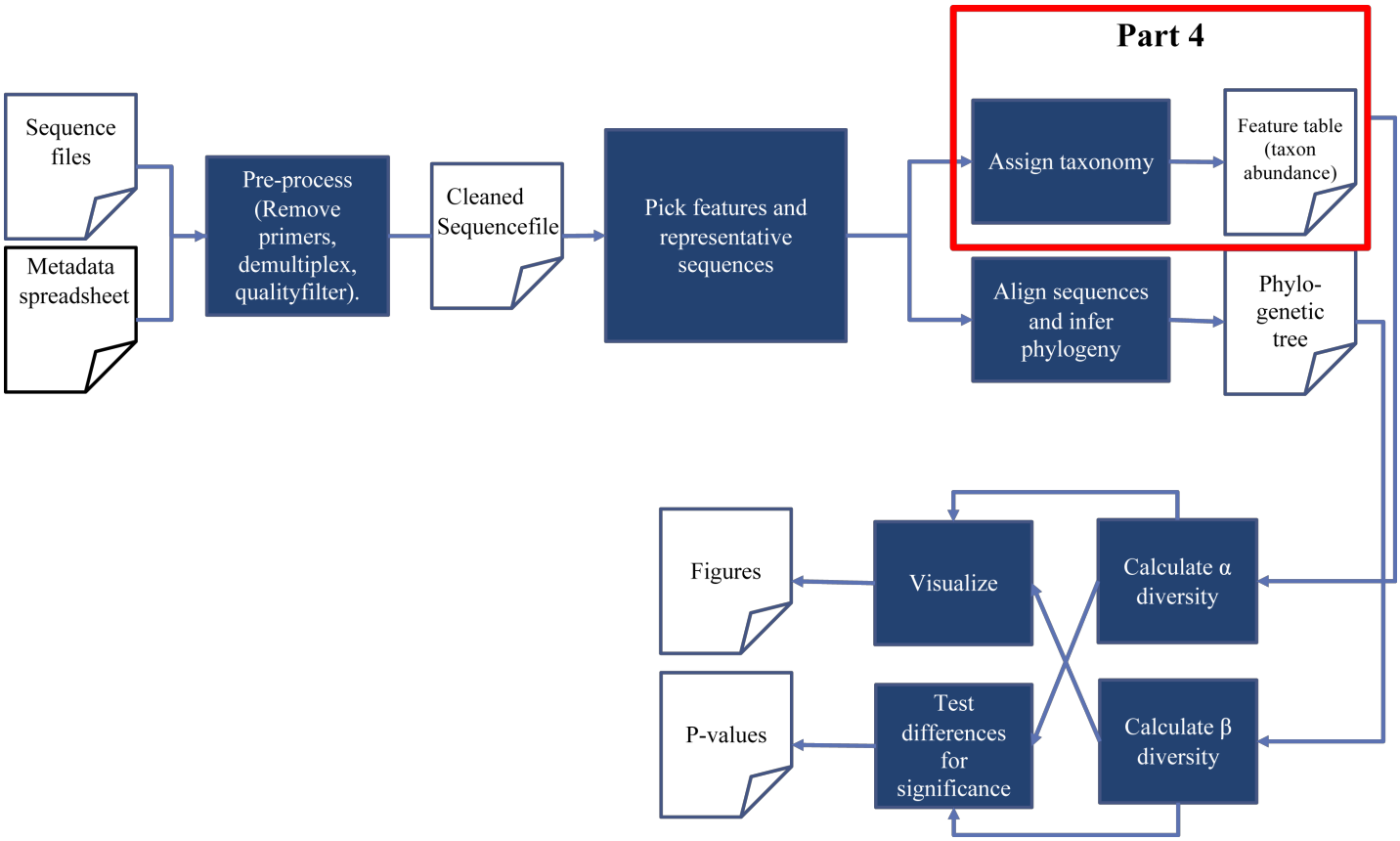
You have just converted the feature table **table.qza** into **feature-table.biom**, which is useful for analysis with tools like as microbiomeanalyst and a **feature-table.txt** which can be visualized in any spreadsheet app.

You are now ready to assign taxonomy to your samples 

Progress tracker

Part 3 of 6 completed

Part 4. Assign taxonomy to features



Sequence variants are of limited usefulness by themselves. As scientists we often want more than just the diversity of the sample. We would like to know the identify the organisms that are present in a sample. That requires us to identify the taxonomy of (= assign taxonomy to) each sequence feature.

In order to do this, we are going to use a **pre-trained classifier**. Generally it is best to train such a classifier on the exact region of the 16S that was sequenced (or 18S or ITS etc that you work on in the future). Here you use the SILVA database (release 132) (we don't have SILVAmar ready for QIIME2 yet). For this tutorial the classifier was trained on the Silva 99% OTUs full-16S rRNA gene sequences trimmed to the V3-V4 region with our specific amplicon primers.

▶ **Which database should I use?**

▶ **In case you wonder how we did it**

❓ Take a look at Qiime2's `feature-classifier` plugin. All taxonomy assignment methods in Qiime2 are here. (Hint: use `--help`). What other methods are there in Qiime2 for assigning taxonomy?

▶ **Solution**

I] You are now ready to run the classification and assign taxonomy to the representative sequences from the DADA2 feature picking using the trained classifier:

```
qiime feature-classifier classify-sklearn /
    --i-classifier classifier_NB_V3-V4_SILVA_99.qza /
    --i-reads DADA2/representative_sequences.qza /
    --o-classification taxonomy.qza
```

II] Now you create visualization files to explore the results:

▶ **In case your run takes "forever"!**

```
qiime metadata tabulate /
    --m-input-file taxonomy.qza /
    --o-visualization taxonomy.qzv
```

This will output a table with each feature ID; its classification and the confidence level for the taxonomy assignment. Note that you can export a .tsv with these results.

III] Look at the the **taxonomy.qzv** file. We can see the data presented at different taxonomic levels and grouped by different experimental factors.

IV] You can also create a visualization file based on bar plots using the following command

```
qiime taxa barplot \
    --i-table DADA2/table.qza \
    --i-taxonomy taxonomy.qza \
    --m-metadata-file metadata.tsv \
    --o-visualization taxa-bar-plots.qzv
```

V] Explore the results and take into account:

Level 1 = Kingdom (e.g Bacteria)
Level 2 = Phylum (e.g Proteobacteria)
Level 3 = Class (e.g Gammaproteobacteria)
Level 4 = Order (e.g Vibrionales)
Level 5 = Family (e.g Vibrionaceae)
Level 6 = Genus (e.g Aliivibrio)
Level 7 = Species (e.g A. fischeri)

Note that you can sort samples by metadata and change colors and taxonomic levels. Morevoer, you can download figures and data in .csv format.

VI] Look for mitochondrial or chloroplast contamination. It should not be that many.

In Qiime2 you can filter your feature table and representative sequences to remove known contaminants or

non-target groups, e.g., host DNA including mitochondrial or chloroplast sequences. The following command will remove all features that contain either mitochondria or chloroplast in their taxonomic annotation.

```
qiime taxa filter-table \
    --i-table DADA2/table.qza \
    --i-taxonomy taxonomy.qza \
    --p-exclude mitochondria,chloroplast \
    --o-filtered-table table-no-mitochondria-no-chloroplast.qza
```

There are of course other filtering methods you can use.

Alternative] **If you have time!**: If you made a taxa-bar-plot artifact of **table-no-mitochondria-no-chloroplast.qza** and explored it.

❓ Did the contamination go away?

VII] Say you want to look more closely at the archeae. These can be filtered to its own table with the `--p-include` argument like this

```
qiime taxa filter-table /
    --i-table DADA2/table.qza /
    --i-taxonomy taxonomy.qza /
    --p-include archaea /
    --o-filtered-table table-archaea.qza
```

VIII] Make a taxa-bar-plots artifact of **table-archaea.qza** and explore it. Sort samples by **DepthGr**.

Say you want to know if the differences observed in the plots for deep and shallow waters are statisticaly significant (the statistical material is small for this example but we'll do it just to see how it works). In this case we can use ANCOM (Analysis of Composition of Microbiomes). This method accounts for the compositional nature of 16S data. If you want to get a deeper insight into the problem of compositional data please read this.

In brief ANCOM adds a **pseudocount** to each count value and then compares the log ratio of the abundance of each taxon to the abundance of all the remaining taxa one at a time. (Transformations after Aitchison 1986). As with any bioinformatics method, you should be aware of the assumptions and limitations of ANCOM before using it. ANCOM assumes that few (less than about 25%) of the features are changing between groups. This is not always the case! If you expect that more features are changing between your groups, you should not use ANCOM as it will be more error-prone (an increase in both Type I and Type II errors is possible). If you want further detalis on how ANCOM works please read this.

IX] You will test for differential abundances at the family level. In order to do that, we need first to get the

table of relative abundances of reads for each family for each sample:

```
qiime taxa collapse /
    --i-table table-archaea.qza /
    --i-taxonomy taxonomy.qza /
    --p-level 5 /
    --o-collapsed-table table-archaea-l5.qza
```

ANCOM operates on a `FeatureTable[Composition]` QIIME 2 artifact, which is based on frequencies of features on a per-sample basis, but cannot tolerate frequencies of zero. To build the composition artifact, a `FeatureTable[Frequency]` artifact must be provided to add-pseudocount (an imputation method), which will produce the `FeatureTable[Composition]` artifact.

X] Add pseudocount:

```
qiime composition add-pseudocount /
    --i-table table-archaea-l5.qza /
    --o-composition-table comp-table-archaea-l5.qza
```

XI] You can now run differential abundance test

```
qiime composition ancom /
    --i-table comp-table-archaea-l5.qza /
    --m-metadata-file metadata.tsv /
    --m-metadata-column DepthGr /
    --o-visualization l5-ancom-archaea-Depth.qzv
```

XII] Let's take a look at the results:

❓ Which families differ in abundance between samples from deep and shallow waters? In which group is each family more abundant?

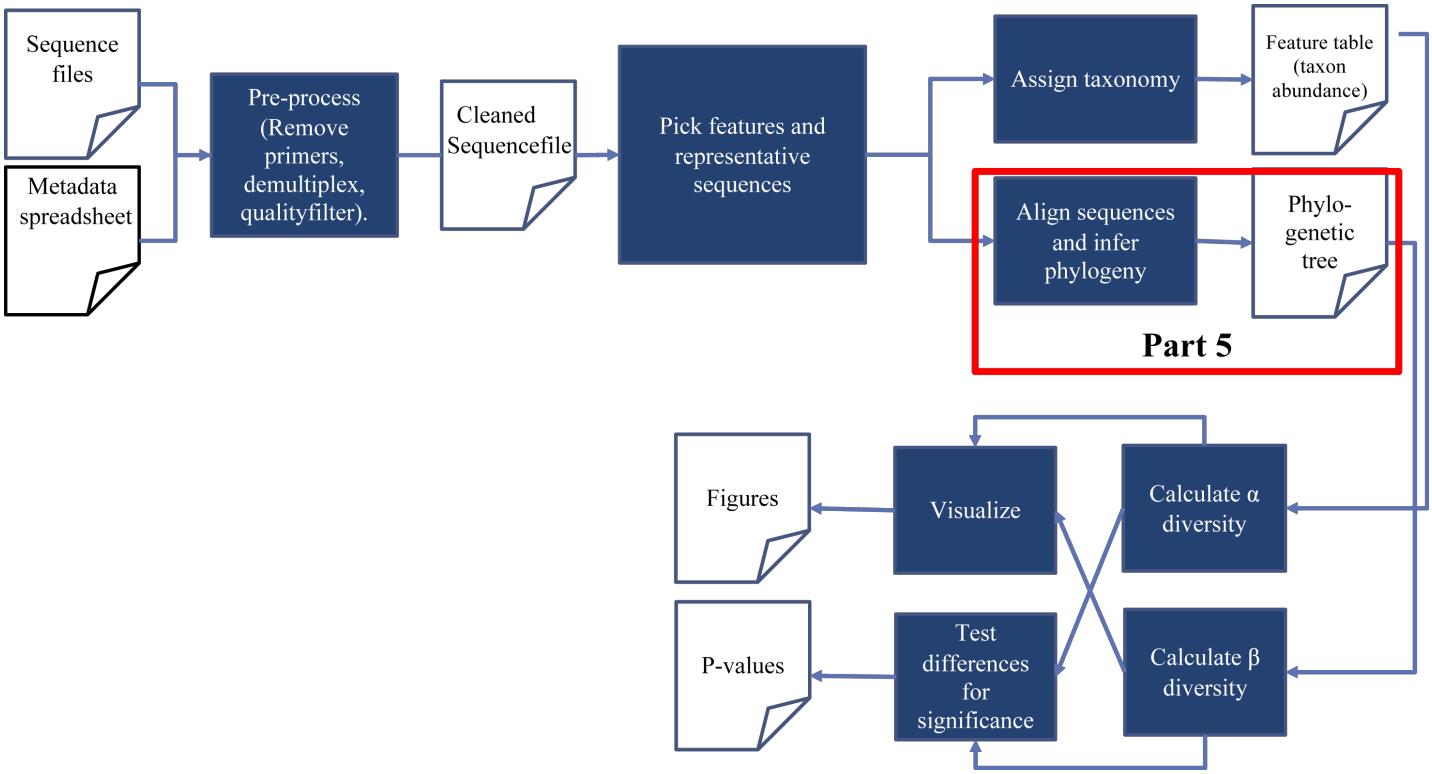
▶ **Solution**

Note: ANCOM does not report p-values but a table with information on the rejection (or not) of H0. They also provide the statistic "W" values and information of the distribution of data in percentiles for each tested group. You can test differeces in taxa abundances using ANCOM for all taxa levels. It is not recommended to use taxa classification below level 6 (i.e. Genus).

Progress tracker

Part 4 of 6 completed

Part 5. Align sequences and infer phylogeny



Later we are going to calculate some diversity metrics and distances that will include phylogenetic information of the features. QIIME2 supports several phylogenetic diversity metrics, including *Faith's Phylogenetic Diversity* and *weighted and unweighted UniFrac*. In addition to counts of features per sample (i.e., the data in the `FeatureTable[Frequency]` QIIME2 artifact), these metrics require a rooted phylogenetic tree relating the features to one another. This information will be stored in a `Phylogeny[Rooted]` QIIME 2 artifact. To generate a phylogenetic tree we will use `align-to-tree-mafft-fasttree` pipeline from the `q2-phylogeny` plugin.

First, the pipeline uses the **mafft** program to perform a multiple sequence alignment of the sequences in our `FeatureData[Sequence]` to create a `FeatureData[AlignedSequence]` QIIME 2 artifact.
Next, the pipeline masks (or filters) the alignment to remove positions that are highly variable. These positions are generally considered to add noise to a resulting phylogenetic tree.
Following that, the pipeline applies **FastTree** to generate a phylogenetic tree from the masked alignment. The FastTree program creates an unrooted tree.
In the final step midpoint rooting is applied to place the root of the tree at the midpoint of the longest tip-to-tip distance in the unrooted tree.
In previous versions of Qiime2 you would need 4 steps to do this!

I] Align sequences and create a tree

```
qiime phylogeny align-to-tree-mafft-fasttree /
    --i-sequences DADA2/representative_sequences.qza /
    --o-alignment aligned_rep_seqs.qza /
    --o-masked-alignment masked_aligned_rep_seqs.qza /
    --o-tree unrooted_tree.qza /
    --o-rooted-tree rooted_tree.qza
```

❓ Which output-files did you get?

▶ **Solution**

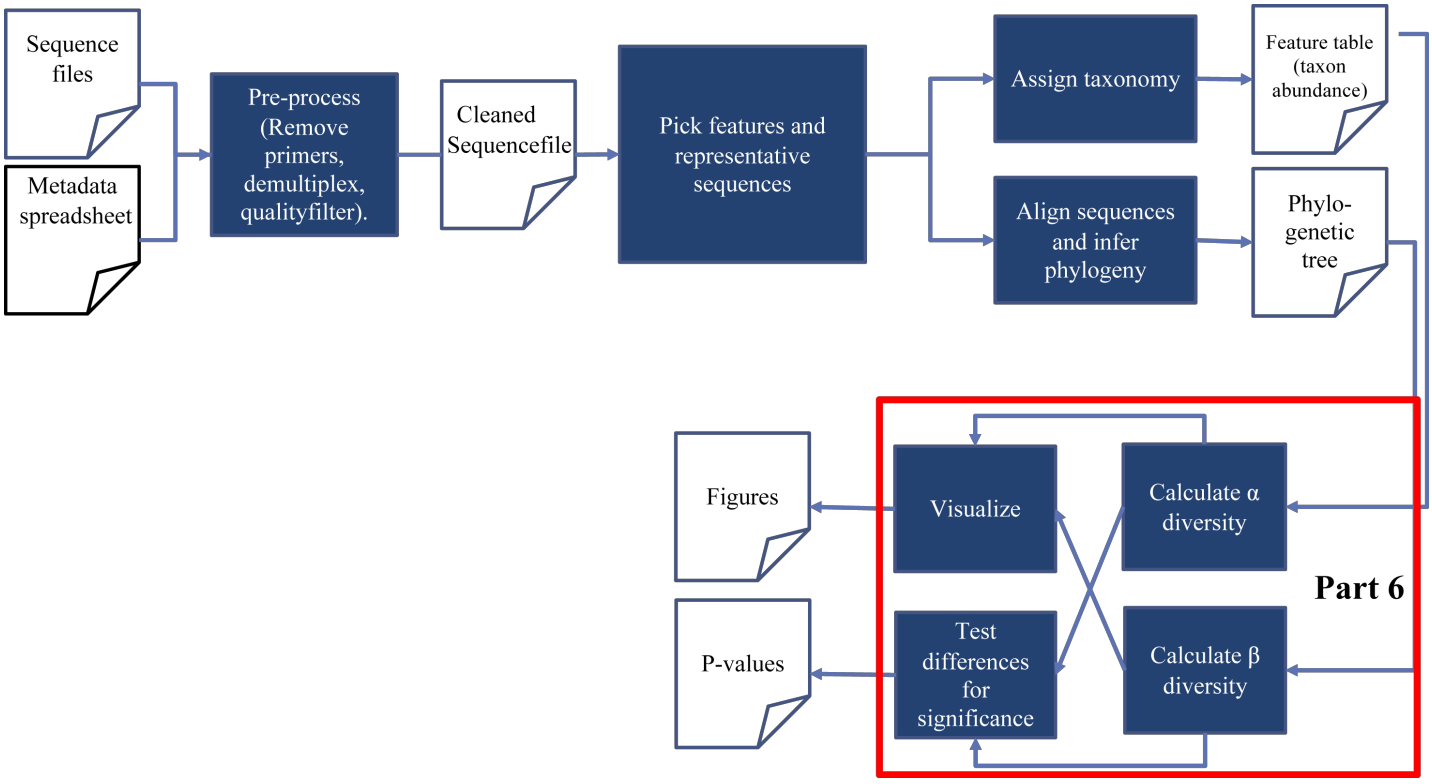
Note: No visualizations will be produced.

Now that we have the phylogenetic tree we are ready to calculate diversity.

Progress tracker

Part 5 of 6 completed

Part 6. Calculate alpha- and beta diversity



In microbiome experiments, investigators frequently wonder about things like:

- How many different species/OTUs/ASVs are present in my samples?
- How much phylogenetic diversity is present in each sample?

- How similar/different are individual samples and groups of samples?
- What factors (e.g., pH, elevation, blood pressure, body site, or host species just to name a few examples) are associated with differences in microbial composition and biodiversity?

Questions like this can be answered by alpha- and beta-diversity analyses. **Alpha diversity** measures the level of diversity **within individual samples**. **Beta diversity** measures the level of diversity or dissimilarity **between samples**. We can then use this information to statistically test whether alpha diversity is different between groups of samples (indicating, e.g., that those groups have more/less species richness) and whether beta diversity is greater between groups (indicating, e.g., that samples within a group are more similar to each other than those in another group, suggesting that membership within these groups is shaping the microbial composition of those samples).

QIIME 2's diversity analyses are available through the `diversity` plugin, which supports computing alpha- and beta- diversity metrics, applying related statistical tests, and generating interactive visualizations. You'll first apply the `core-metrics-phylogenetic` pipeline, which **rarefies** a `FeatureTable[Frequency]` to a user-specified depth, computes several alpha and beta diversity metrics, and generates principle coordinates analysis (PCoA) plots using Emperor for each of the beta diversity metrics. Note: Rarefying data is a matter of debate[1] [2] and is not to be confused with rarefaction

An important parameter that needs to be provided to this script is `--p-sampling-depth`, which is the even sampling depth. Because most diversity metrics are sensitive to different sampling depths across different samples, this script will randomly subsample the counts from each sample to the value provided for this parameter. For example, if you provide `--p-sampling-depth 500`, this step will subsample the counts in each sample without replacement so that each sample in the resulting table has a total count of 500. If the total count for any sample(s) are smaller than this value, those samples will be dropped from the diversity analysis. Choosing this value is tricky. We recommend making your choice by reviewing the information presented in the **table.qzv** file that was created above and choosing a value that is as high as possible (so you retain more sequences per sample) while excluding as few samples as possible.

▶ **Note**

I] View the table.qzv QIIME 2 artifact, and in particular the **Interactive Sample Detail tab** in that visualization.

?

1. What value would you choose to pass for `--p-sampling-depth`?
2. How many samples will be excluded from your analysis based on this choice?
3. How many total sequences will you be analyzing in the core-metrics-phylogenetic command?

▶ **Solution**

II] When you have your `--p-sampling-depth`; run
`qiime diversity core-metrics-phylogenetic`

```
qiime diversity core-metrics-phylogenetic /
    --i-phylogeny DADA2/rooted_tree.qza /
    --i-table DADA2/table.qza /
    --p-sampling-depth xxxx /
    --m-metadata-file ../mapping_file.tsv /
    --output-dir core-metrics-results
```

III] Inspect results in the output dir.

❓ What metrics do you get?

▶ **Solution**

After computing diversity metrics, you can begin to explore the microbial composition of the samples in the context of the sample metadata. This information is present in the sample metadata file **metadata.tsv**. You can do this in the next tutorial on MicrobiomeAnalyst so we'll just show you a few examples on what you can do in Qiime2.

Several visualization files were created, you will have one **.qzv** file for each beta diveristy metric that was calculated.

IV] Explore results using **qiime tools view**

V] Take a look at the results for all beta diveristy distances. (Take advantage of the interactive functionalities)

❓ Can you observe any pattern or grouping of samples?

▶ **Solution**

VI] We can now test for the significance of these grouping by using the following command:

```
qiime diversity beta-group-significance /
    --i-distance-matrix unweighted_unifrac_distance_matrix.qza /
    --m-metadata-file ../metadata.tsv /
    --m-metadata-column DepthGr /
    --o-visualization unweighted_unifrac_DepthGr_significance.qzv
```

VII] Inspect the visualization file:
❓ Does the test confirm what we have observed in the PCoA plot?

▶ **Solution**

Remember you can do the same analysis for every distance matrix, just change the `--i-distance-matrix` parameter and the name of the output file.

Now lets look at alpha diversity results.

VIII] A good way to explore this is by making comparisons between groups of samples:

```
qiime diversity alpha-group-significance \
    --i-alpha-diversity observed_otus_vector.qza \
    --m-metadata-file metadata.tsv \
    --o-visualization observed_otus_vector-group-significance.qzv
```

❓ Is richnness different between deep and and shallow water samples (DetphGr)?

▶ **Solution**

Remember you can do the same analysis for every alpha diveristy metric calculated just change the `--i-alpha-diversity` parameter.

Progress tracker

Part 6 of 6 completed

🏆Tutorial Completed. Great j😀b 🏆🤝👏

1. McMurdie et al. (2014) states that rarefying is a large waste of data and statistical power and that other normalization methods should be used. ↩

2. Weiss et al. (2017)says that also these other methods are inappropriate for microbiome data. ↩