Workflows for processing microbial amplicon sequences

Christian Krohn, PhD, RMIT University

2022-07-14

Contents

1 About this GitBook **5**

2 Getting started **7**

3 Miseq library preps **13**

3.1 Introduction 13

3.2 Workflow 15

4 From BaseSpace to Qiime2 and DADA2 **25**

4.1 Introduction 25

4.2 Workflow 29

5 From Qiime2 into R - Initial diagnostics **41**

5.1 Introduction 41

5.2 Workflow 43

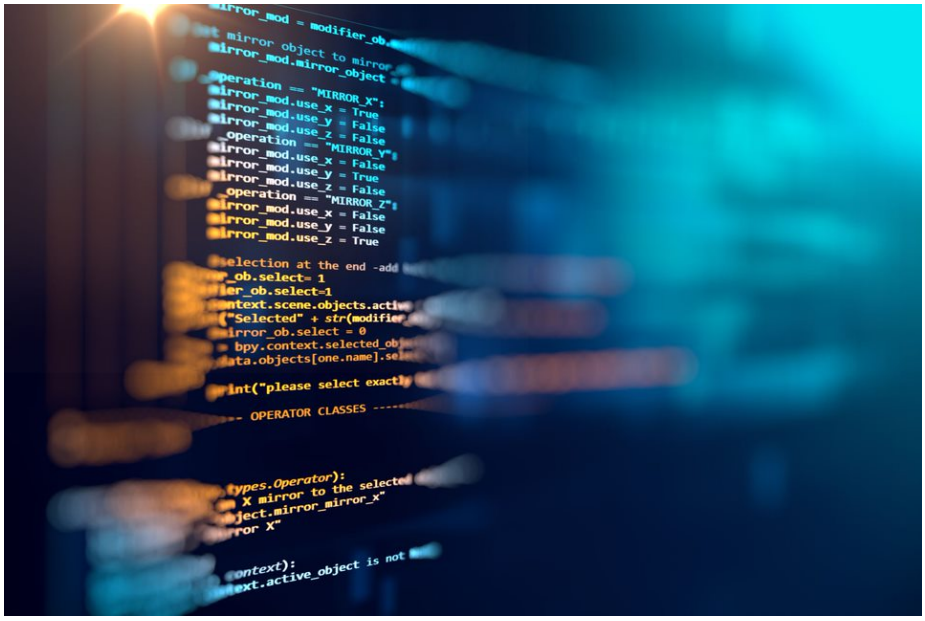
6 Plotting diversity **55**

6.1 Introduction 55

6.2 Workflow 56

Chapter 1

About this GitBook



This GitBook contains workflows for students who want to get started with sequencing microbial amplicons on a Miseq instrument and then process and analyse ASV-based data. It is a compilation of workflows that I have gotten accustomed to during my PhD at La Trobe University with help of many amazing students and colleagues over the years. But it is by no means comprehensive.

One of the biggest hurdles for students that embark on sequencing environmental DNA for the first time, is the effort that is required to understand the various coding languages, file types and formats, packages or platforms that are involved (think Unix, Linux, Slurm, Qiime, R, RMarkdown, python, conda,

ggplot, docker, GitHub, data instances...) before they even can start looking at exploring the data for biological meaning and producing publishable output.

For example, this GitBook is made in Rstudio, using the `rmarkdown` markup language, rendered using the `bookdown` package, and hosted on GitHub. It took me at least five years before I was ready to create and host this content for students. And even then I needed this awesome Bookdown template from Dr. Caspar van Lissa to get started with Bookdown. So it is really helpful to have something to start with.

It was not easy to get over the initial hurdles for me too, so I thought I'd try to compile my workflows. This may help students to get a head start but also helps to scrutinise my approach in the hope that others can point me to better ways of doing things.

Browse to your level

Everyone is at a different level in the journey to exploring microbial diversity. Some have absolutely no idea where to start and others have created their first phylogentic trees with ggtree or similar. The most important prerequisites are patience and persistence... :). Just explore the different chapters and see where it takes you. In fact you will be amazed how much you can achieve with mininal coding knowledge.

Get in touch

We work at the Andy Ball lab, RMIT University, Bundoora, Melbourne. Email me or comment on the discussion section of the GitHub repository for this GitBook. You will need to get a GitHub account to join the discussion. Its free.

Chapter 2

Getting started

Prerequisites will be listed in each sub chapter separately. Following a few general comments:

Access to computational resources

To process data you will need a personal computer or online computational resources with sufficient number of CPUs, working memory and storage. If you want to run most of the processing on a personal computer then I'd recommend a minimum of **4 CPUs, 16 Gb RAM and 500 Gb storage**. However, please note that some alignment tasks require more RAM, hence some of the steps (e.g. taxonomic classifications) may require a more powerful resource. The advantage of using a personal computer is that you can frequently update the latest packages without relying on a university administrator to do that for you. Furthermore, with a personal computer you wont need to share with others, making it easier to run the packages without a workload manager.

Some university may provide their own access to a high performance computer (HPC) for staff and students and likely have certain packages pre-installed. In such cases a batch workload manger such as Slurm may be used, to manage multiple users trying to run computationally heavy jobs. This works fine too but its a bit of a hassle and I would use such an HPC only if really necessary, i.e. smaller jobs on my personal computer and bigger jobs on the HPC.

I am not aware that RMIT University offers that type of service yet.

In Australia there is also the ARDC Nectar Research Cloud, which provides free resources to Universities. This is basically like providing access to an online Linux computer with complete freedom to install any packages as if it was your own. If you are a student you can create a trial account with your student email address (Using your ID only; without the student.edu.au) and check out how it all works, albeit with limited resources. The Nectar service also provides tuturials for starters. However, to get serious with Nectar, you have to request more

resources, which is a fairly straight-forward process. Check out the eligibilities and chat to your supervisor if you are interested.

There are also commercial options, which probably offer very similar type of services to Nectar.

In addition, there is the awesome Galaxy Australia platform, which is an open web-based resource that also contains many tutorials and workflows specific to bioinformatics. Definitely worth checking that out too.

This workbook will be based on using the Nectar Research cloud.

File storage

Working with sequencing data may require you do handle large amounts of data. From experience, a run of a 600µl pool (6-10pM) of amplicons in 2x301 Miseq cycles may produce around 10-20 Gb of FASTQ files. Much more can be expected for NextSeq runs or for long-read sequencing.

Once a Miseq run has finished it is typically available to download from the Illumina Basespace account. After a few runs, the data can accumulate quickly. Hence, **persistent storage space** is required. As long as you have the Basespace account you can keep a copy of the FASTQ file online. However, Basespace may delete your data if the account has been inactive for more than 6 months. Alternatively, data can be upload onto a public repository for persistent storage, such as the Sequence Reads Archives. If you have access to a research cloud or commercial computational resources then you may store data on those. Personally, I prefer a 1TB solid state harddrive.

Command Line

Many of the tools used are managed through command line. Cloud servers are also accessed through command line. That means there is no graphical user interphase and all commands, including installation and running of packages are done through lines of codes on a simple window interphase. This window is called Terminal (Mac, Linux) or Command Prompt application (Windows). There are lots of online sources to help you get started with command line. One example here: https://towardsdatascience.com/a-quick-guide-to-using-command-line-terminal-96815b97b955

Data storage browser

To upload, download, move, rename files on any cloud computer or high performance computer (HPC) at your institution, you will also require a cloud storage browser, such as Cyberduck (Mac) or PuTTY (Windows).

R and R studio

R has become an essential research language. If you want to progress in research it is almost inevitable for you to learn. Just do it. :). Perhaps start here if you dont know where else to start: https://education.rstudio.com/learn/beginner/.

It lists some great step-by-step tutorials on how to install R and Rstudio and then explains the basics.

If you are like me, then you simply copy and paste code from other sites and see what happens. This book should provide you with the required information to enable you to follow the workflows. In case you get stuck, you can either chat to me directly or write on the GitHub Discussion Page for this GitBook (chrismitbiz/ABlab-workflows/discussions/).

At the same time you will want to do short courses (eg. from edX Data Science: R Basics) that delve a little deeper into different data structures such as data frames, matrices, lists and the syntax to arrange stuff. Over time you understand what different lines of code mean and can trouble shoot when things don't work. And things often don't work. You will get used to troubleshooting code :).

Environment managers

Package and environment managers are extremely useful for your workflows. They help to install and run software packages such as qiime2, into individual 'environments' independent of your operating system. The environment manager is installed once and from there you use it to install individual packages. The most commonly used environment manager is Conda. Learn more about it here: https://docs.conda.io/projects/conda/en/latest/user-guide/getting-started.html

Conda can be installed by either installing Anaconda or Miniconda. Boths works the same way. Anaconda requires 5 Gb of discspace and installs everything you can possibly need, while Miniconda is just the raw bones and lets you install things one by one.

Most of the commands to manage environments are done in command line using a terminal. There is also a graphical installer that makes handling the environments a little more visual. That would be my go to. Learn more here: www.anaconda.com.

Once conda installed you can always check what environments are installed with the command `conda env list`. In my case the output looks like this:

```
(base) ubuntu@metagenomics:/pvol$ conda env list
# conda environments:
#
base                  *  /pvol/anaconda3
SqueezeMeta              /pvol/anaconda3/envs/SqueezeMeta
anvio-7.1                /pvol/anaconda3/envs/anvio-7.1
cutadaptenv              /pvol/anaconda3/envs/cutadaptenv
qiime2-2021.8            /pvol/anaconda3/envs/qiime2-2021.8
qiime2-2022.2            /pvol/anaconda3/envs/qiime2-2022.2
```

There are several other environment managers. We are also using Docker in our work but the learning curve is a bit steeper so we wont get into it here.

Explore the help functions

Packages such as qiime2 have great resources to help you understand how to run any of the available commands. For example if you dont know what input parameters are available for the qiime command
`qiime feature-classifier classify-sklearn`, you can simply enter this command into your Terminal (with the qiime environment activated, if it is installed with Conda) and add a `--help` at the end: `qiime feature-classifier classify-sklearn --help`.

Or you just want to know what other commands qiime has in its repertoire, you can run `qiime --help`.

For example, this

```
conda activate qiime2-2022.2
qiime --help
```

Gives

```
Usage: qiime [OPTIONS] COMMAND [ARGS]...

  QIIME 2 command-line interface (q2cli)
  --------------------------------------

  To get help with QIIME 2, visit https://qiime2.org.

  To enable tab completion in Bash, run the following command or add it to
  your .bashrc/.bash_profile:

      source tab-qiime

  To enable tab completion in ZSH, run the following commands or add them to
  your .zshrc:

      autoload -Uz compinit && compinit
      autoload bashcompinit && bashcompinit
      source tab-qiime

Options:
  --version   Show the version and exit.
  --help      Show this message and exit.

Commands:
  info                Display information about current deployment.
  tools               Tools for working with QIIME 2 files.
  dev                 Utilities for developers and advanced users.
  alignment           Plugin for generating and manipulating alignments.
  composition         Plugin for compositional data analysis.
  cutadapt            Plugin for removing adapter sequences, primers, and
                      other unwanted sequence from sequence data.

  dada2               Plugin for sequence quality control with DADA2.
  deblur              Plugin for sequence quality control with Deblur.
  demux               Plugin for demultiplexing & viewing sequence quality.
  diversity           Plugin for exploring community diversity.
  diversity-lib       Plugin for computing community diversity.
  emperor             Plugin for ordination plotting with Emperor.
  feature-classifier  Plugin for taxonomic classification.
  feature-table       Plugin for working with sample by feature tables.
  fragment-insertion  Plugin for extending phylogenies.
  gneiss              Plugin for building compositional models.
  longitudinal        Plugin for paired sample and time series analyses.
  metadata            Plugin for working with Metadata.
  phylogeny           Plugin for generating and manipulating phylogenies.
  quality-control     Plugin for quality control of feature and sequence data.
  quality-filter      Plugin for PHRED-based filtering and trimming.
  sample-classifier   Plugin for machine learning prediction of sample
                      metadata.

  taxa                Plugin for working with feature taxonomy annotations.
  vsearch             Plugin for clustering and dereplicating with vsearch.
(qiime2-2022.2) ubuntu@metagenomics:/pvol$
```

GitHub account

This GitBook, including all its files, is hosted on one of my GitHub repositories (https://github.com/chrismitbiz/ABlab-workflows). If you have any comments you can ask a question on the Discussion Page of this repository. You require a GitHub account to do that. It is free.

Workflows from other lab groups

Searching for "amplicon-sequencing" under "Topics" on GitHub gave 37 results.

- nf-core/ampliseq
  On the top of that list is the nf-core/ampliseq pipeline developed by the nf-core community. It is based on a software called Nextflow which allows to put different processes into a pipeline. This is great for doing things a little more reproducible but it requires you to be fairly knowledgeable with Linux, container software, config files etc.. Not great to learn stuff for beginners.

- Tools-Microbiome-Analysis
  Tools-Microbiome-Analysis is a websites containing a comprehensive list of R packages and, more importantly, tutorials related to analysis of microbial amplicons and ecology. Really good reference to go back to every now and then. **Highly recommended**.

- grimmlab/MicrobiomeBestPracticeReview
  Essential paper to read (Current challenges and best-practice protocols for microbiome analysis, 2021) and a great workflow resource on GitHub. **Essential read**. The Grimm lab in Munich published this paper as well as developed a python and R-based workflow that assists with the recommended best practices (amplicon as well as metagenomic workflows).

- KasperSkytte/ampvis2
  Ampviz2 is an R-package to visualise and analyse 16S rRNA amplicon data. It is always more convenient to have packages that have the details and optics worked out for you. Like phyloseq, the ampviz2 package combines different tables from data (E.g. otu table, taxonomic table, phylogenetic tree, sample data etc) and then provides different functions to apply to that combined object to visualise the data.

Chapter 3

Miseq library preps

3.1 Introduction

Some info on getting your run ready

In our lab we usually prepare a 6-10 pM pool of libraries following the 16S Metagenomic Sequencing Library Preparation protocol from Illumina with 10% Phix. **Download it, study it and refer to it throughout your labwork**. This protocol is not limited to 16S amplicons. You can use it for fungal (e.g. ITS) amplicons too. In the last section there are also valuable tips for pre-PCR and post-PCR lab procedures. There is also some additional Illumina guidance here: https://sapac.support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html

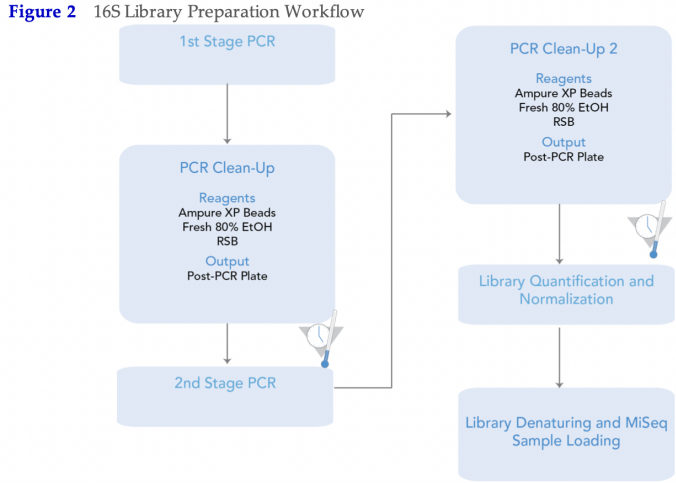
3.1.1 Overview

The protocol encompasses the following steps (copied out of the protocol):

16S Library Preparation Workflow

The following diagram illustrates the workflow using the 16S Library Preparation Protocol. Safe stopping points are marked between steps.

Figure 2 16S Library Preparation Workflow



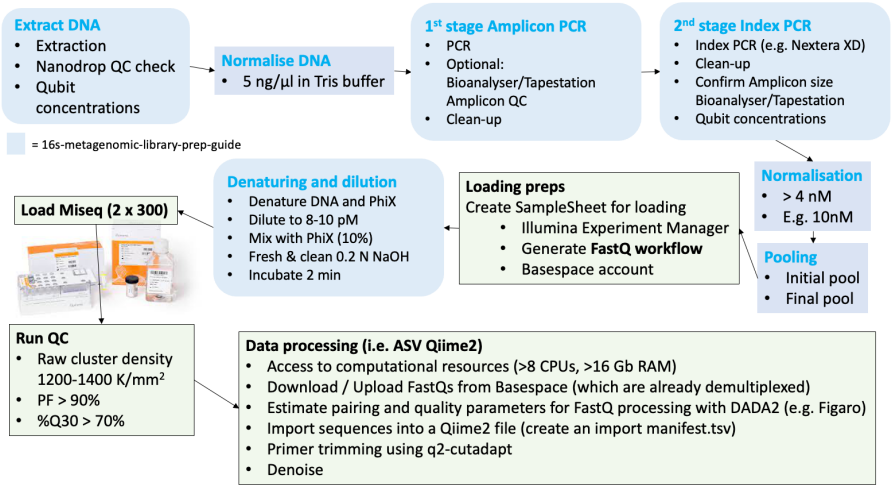
Please note that only **one amplicon target** should be sequenced in each run. That means it is not recommended to sequence for example two different 16S amplicons or one 16S amplicon plus an ITS amplicon. Sequencing different amplicon lengths will result in reduced quality as the the Miseq flowcell will preferentially cluster shorter fragments. However, it is not impossible. We have tried to run four different amplicons on one flowcell before, although with mixed results.

Labwork may take four weeks from DNA extraction to Miseq loading. Double that if you have not done it before. You probably realised that there is quite a bit of labwork required before you finally get FASTQ files and subsequent ASV/taxonomy tables to analyse - the ultimate goal here. The more samples are planned to be included into the pool, the more work and consumables are needed. Illumina recommends no more than 96 samples on one flowcell to get enough depths (i.e. $>$ 100,000 reads per sample). But we have sequenced around 196 samples on one flowcell before, which captured sufficient reads and diversity. This depends on how much detail and rare taxa you want to capture. The less samples you include, the more data and depth for each sample.

Also note that you will require over 20 boxes of tips (mostly 200 µl) for each 96-well plate, so plan your consumables ahead of time. Most of the tips will be used for the PCR clean ups.

After all steps in the above protocol are completed, the Miseq will be loaded

with a 600 µl pool, followed by 56 hours of run-time for 2x301 cycles before FASTQ files are available for download from BaseSpace.



Throughout the 56 hours run you can check how things are tracking either on the BaseSpace website or the BaseSpace phone app. Understanding different QC metrics will help identifying potential problems with the pool or the instrument. It is important to note that this is a bit of a nerve-wrecking moment in the process, as it remains unclear until the metrics come through after around the 21st cycle, whether or not the run will be successful.

3.2 Workflow

3.2.1 DNA extraction

We commonly extract DNA from soils or from wastewater sludges using the *DNeasy Powersoil Pro Kit* (Qiagen) for both, soil and sludge. It results in high quality DNA and includes a bead beating step to make sure the gram positive cells are sufficiently broken up. Although, I don't have evidence for that. In fact, please let us know on the GitHub Discussion Page for this GitBook if you have any comments on that as we are always looking for ways to improve things.

Soil

- For soils, the protocol of the extraction kit is followed, weighing in 0.25 g of soil and making sure to record the exact weights used.

Figure 3.1: "DNA extracts"

- Also make sure to measure the water content of the soil to be able to report DNA concentrations per dry weight of soil.

Sludge

- For DNA extraction of sludge, 0.5 ml is added into an empty 1.7 ml extraction tube (without extraction beads), then centrifuged at 5000 rpm for 5 minutes before removing the supernatent.
- Extraction beads are then added and the Powersoil protocol followed.
- Measure Total Solids (TS) of the sludge to report DNA concentrations per gram of TS.

3.2.2 DNA quality and concentration

Once you have extracted DNA it is necessary to confirm its quality and concentration.

- Quality is assessed using a Nanodrop Spectrophotometer. You are aiming for **260/280 Ratio of 1.8 and a 260/230 Ratio of 2-2.2**. This is explained in the T042-TECHNICAL BULLETIN. The Nanopore also estimates DNA concentrations but it is not recommended to use these concentrations measurements to normalise DNA. It is more accurate and consistent with a Qubit Fluorometer.

- DNA concentrations are best measured using fluorescent dye-based methods such as the **Qubit Fluorometer**. It is easist to use the '*1X*' assays such as the *Qubit 1X dsDNA BR Assay Kit* or the *Qubit 1X dsDNA HS Assay Kit*
  - Broad range (BR) Kit for quantifying extracted environmental DNA and afte the 2nd stage PCR.
  - High Sensitivity (HS) Kit for quantifying the pool at the end.

3.2.3 DNA normalisation

The next step is to normalise all DNA extracts to 5 ng/µl.

- Dilute DNA to equal concentrations, either in a fresh batch of PCR-grade 10mM Tris buffer (pH 8.5) or in PCR-grade water

- Set a pre-define DNA volume, say 2µl for all extracts, and add Tris or water as calculated with

$$C_1V_1 = V_2C_2$$

3.2.4 First and second PCR - Amplicon PCR and Indexing PCR

Follow the 16S Metagenomic Sequencing Library Preparation Guide. Include a negative control and sequence that as well. This allows you to assess background and cross contamination. The whole workflow is best done in 96-well plates, using multi-channel pipettes to avoid indexing mistakes across wells.

Sample list and plate layout

Ensure that all sample IDs, plate-layout and each unique Illumina index carefully planned and printed out before you start your PCRs. Perhaps, stick the layout onto the Bio-safety cabinet where you do most of the pipetting and keep checking you are on the right well and sample. Label all plates even if you discard them afterwards. Labwork can be stressful and you need to trace all your steps across each well of the plate because otherwise you risk that you have indexed the wrong samples and it will become impossible to know which sample you have sequenced.

Primers

Figure 3.2: "The ready-mix working solution for Qubit, which includes dye and buffer. No preparation/premixing required."

Any primers you order for the Miseq will have to include the **Illumina overhang adapter**. The sequence of the overhang adapter is added **in front** of the gene-specific primer sequences. There are two overhang adapters; one for the forward primers and one for the reverse primer. Check out the relevant information on primers from page 3 the 16S Metagenomic Sequencing Library Preparation Guide.

Clean-ups

It is worth doing the clean-ups in a separate deep-well plate ("MIDI" plate, see consumables below) to avoid any blow-out of dried beads at the end of the clean-up. This process is explained in the above mentioned guide too. Tip: Handle the beads extremely carefully after the ethanol has dried. They blow out easily.

Index Kits

The index kits are expensive. Try to handle them with care and replace lids with the provided replacement lids, after every use. Multipipettes are absolutely essential to avoid loosing track on indexes used across the different wells on the plate. The use of the Illumina Plate Fixture comes in handy (see consumables below).

Amplicon QC

After the indexing-PCR products are cleaned it is recommended to run a gel on all samples to confirm that amplification was successful with enough DNA present and that it is the correct amplicon length. It also helps to identify any primer dimers, which may have to be cleaned out.

If you have access to a **Tapestation or a Bioanalyzer** in your lab, perhaps run a random subset of the indexed samples on one of these instruments, just to accurately confirm the average length of the amplicon. This average basepair length of the amplicon is important, as it is the basis for normalising your DNA to nM. In any case, after the index samples are normalised and pooled, it is highly recommended to run the pool on a Bioanalyser too. Otherwise it is too much guesswork in getting the concentrations right. This is further explained below.
Repeat any samples that failed, i.e. that do not show any bands, and ensure you use the same indices.

3.2.5 Normalisation of indexed amplicons

After you finished all PCR work it is time to do another round of Qubitting and normalisation of all samples into a new 96-well plate. The final pool should be 4 nM but we prefer to normalise DNA to a higher concentration first, say 10

nM, then pool and then dilute the pool to 4 nM afterwards.
We are using a very handy excel sheet, which was kindly developed and provided by Sarah Knowler at La Trobe University to prepare and assist in normalisations, converting from ng/µl to nM using the formula:

$$\frac{concentrations\ (ng/\text{ţ}l)}{660g/mol\ x\ average\ library\ size}$$

Contact us if you would like to get a copy of this very handy library preparation excel sheet.

Some more information here: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-10.pdf. There's also a pooling calculator here https://support.illumina.com/help/pooling-calculator/pooling-calculator.htm which may be helpful with your calculations.

3.2.6 Pooling

- Pipette equal volumes of the normalised DNA into one 1.5 ml tube.

- Measure its concentration using the High Sensitivity (HS) Qubit 1X dsDNA HS Assay Kit.

- Assuming the pool was created wtih normalised concentration that are higher than 4 nM, then you will need to dilute the pool down to 4nM. In order to do so you will need to accurately measure the average basepair length of your pool. Do not rely on the ladder on a gel. Always get an accurate read. See next point.

- Measure the average base pair lengths of the pool with a **Bioanalyzer or Tapestation**. Use at least three replicates. The average basepair length is then used to calculate the final concentration required in ng/µl to get a 4 nM pool.

- Dilute the pool to the concentration in ng/µl that reflect 4 nM.

3.2.7 Denaturing of pool and loading of the Miseq

Now you are almost ready load the Miseq. The last steps are done right before you plan to load the instrument. First the pool DNA needs to be denatured, mixed with PhiX and then diluted further to your preferred loading concentration (6 - 10 pM). The final concentrations impact on cluster densities. That means you get more data with higher concentrations but that is at the expense of potential errors during sequencing at high cluster densities. I am still not quite sure what the ideal concentration is. Please let me know if you know more.

The key for the final steps are that..

- ..the reagent cartridge is properly thawed. Perhaps thaw in fridge 3 hours before starting the denaturing steps. Then put into icy water to ensure final and even thawing across the cartridge. Mix cartrige well at the end and ensure that no bubbles are visible before loading.

- .. the 0.2 N sodium hydroxide (NaOH) is made fresh with biological grade NaOH. For example use the *BioUltra, for molecular biology, 10 M NaOH in H2O* from Sigma Aldrich to make a 0.2 N solution (N = Normality but it is equal to Molarity here because there is only one OH in NaOH). I have used analytical-grade NaOH pellets as well, which worked but the RNAse free solutions are safer in terms of inhibitors etc..

- ..that the pool is cooled down immediately in ice or icy water right after the 2 minutes in the 96 degrees C heatblock and pretty much immediately loaded into the Miseq. The Miseq has to be prepped beforehand (I.e. maintenance wash and sample sheet loaded).

- a sample sheet (or so-called manifest file) is created beforehand using the Illumina Experiment Manager Software. This software assist in producing a `.csv` file that is then loaded into the Miseq instrument via a USB and ensures that the indexes you have used are linked to the relevant samples in your pool. Basically a list of sample IDs and index IDs. An example of a sample sheet from one of our runs is shown below.
  IMPORTANT: Make sure to specify the Workflow as **GenerateFastQ**.

- you have a Illumina BaseSpace account.

[Header]												
IEMFileVersion	5											
Investigator Name	AndyBall Lab											
Experiment Name	PrimerTrial											
Date	6/22/2022											
Workflow	GenerateFASTQ											
Application	Assembly											
Instrument Type	MiSeq											
Assay	Nextera XT											
Index Adapters	Nextera XT v2 Index Kit											
Description	threedigestersfourprimersplusPMA											
Chemistry	Amplicon											
[Reads]												
301												
301												
[Settings]												
ReverseComplement	0											
kmer	31											
Adapter	CTGTCTCTTATACACATCT											
[Data]												
Sample_ID	Sample_Name	Sample_Plat	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	GenomeFold	Sample_Proj	Description		
PT_01	PT_01	1	A01	N701-C	TAAGGCGA	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_01V4_UNoPMA		
PT_02	PT_02	1	A02	N702-C	CGTACTAG	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_02V4_UNoPMA		
PT_03	PT_03	1	A03	N703-C	AGGCAGAA	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_03V4_UNoPMA		
PT_04	PT_04	1	A04	N704-C	TCCTGAGC	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_04V4_UNoPMA		
PT_05	PT_05	1	A05	N705-C	GGACTCCT	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_05V4_UNoPMA		
PT_06	PT_06	1	A06	N706-C	TAGGCATG	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_01V4_UPMA		
PT_07	PT_07	1	A07	N707-C	CTCTCTAC	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_02V4_UPMA		
PT_08	PT_08	1	A08	N710-C	CGAGGCTG	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_03V4_UPMA		
PT_09	PT_09	1	A09	N711-C	AAGAGGCA	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_04V4_UPMA		
PT_10	PT_10	1	A10	N712-C	GTAGAGGA	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_05V4_UPMA		
PT_11	PT_11	1	A11	N714-C	GCTCATGA	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_01V4_UPMAx		
PT_12	PT_12	1	A12	N715-C	ATCTCAGG	S513-C	TCGACTAG		PrimerTrial	ARCSR_02_Meso_02V4_UPMAx		
PT_13	PT_13	1	B01	N701-C	TAAGGCGA	S515-C	TTCTAGCT		PrimerTrial	ARCSR_02_Meso_03V4_UPMAx		
PT_14	PT_14	1	B02	N702-C	CGTACTAG	S515-C	TTCTAGCT		PrimerTrial	ARCSR_02_Meso_04V4_UPMAx		
PT_15	PT_15	1	B03	N703-C	AGGCAGAA	S515-C	TTCTAGCT		PrimerTrial	ARCSR_02_Meso_05V4_UPMAx		
PT_16	PT_16	1	B04	N704-C	TCCTGAGC	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_01V4_UNoPMA		
PT_17	PT_17	1	B05	N705-C	GGACTCCT	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_02V4_UNoPMA		
PT_18	PT_18	1	B06	N706-C	TAGGCATG	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_03V4_UNoPMA		
PT_19	PT_19	1	B07	N707-C	CTCTCTAC	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_04V4_UNoPMA		
PT_20	PT_20	1	B08	N710-C	CGAGGCTG	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_05V4_UNoPMA		
PT_21	PT_21	1	B09	N711-C	AAGAGGCA	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_01V4_UPMA		
PT_22	PT_22	1	B10	N712-C	GTAGAGGA	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_02V4_UPMA		
PT_23	PT_23	1	B11	N714-C	GCTCATGA	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_03V4_UPMA		
PT_24	PT_24	1	B12	N715-C	ATCTCAGG	S515-C	TTCTAGCT		PrimerTrial	ARCSR_03_Meso_04V4_UPMA		
PT_25	PT_25	1	C01	N701-C	TAAGGCGA	S516-C	CCTAGAGT		PrimerTrial	ARCSR_03_Meso_05V4_UPMA		
PT_26	PT_26	1	C02	N702-C	CGTACTAG	S516-C	CCTAGAGT		PrimerTrial	ARCSR_03_Meso_01V4_UPMAx		
PT_27	PT_27	1	C03	N703-C	AGGCAGAA	S516-C	CCTAGAGT		PrimerTrial	ARCSR_03_Meso_02V4_UPMAx		
PT_28	PT_28	1	C04	N704-C	TCCTGAGC	S516-C	CCTAGAGT		PrimerTrial	ARCSR_03_Meso_03V4_UPMAx		

Figure 3.3: “Example of a manifest file”

Table 3.1: Example of a consumables list we typically include for a Miseq run

Equipment.consumables	Item.number	Supplier	Comment
Qiagen Powersoil Pro test kit, 250	#47016	Qiagen	
Primers	Various	E.g. Sigma	
Magnetic beads, AMPure XP / JetSeq Clean	BIO-68031	Meridian Life Sciences	For clean-up
KAPA HiFi Hotstart ready mix	KK2602, 6.25ml	Millennium Science	For PCR
MiSeq Reagent Kit v3	MS-102-3003	Illumina	Needs to be a
Nextera XT Index Kit v2, Set A, B, C or D	FC-131-2003	Illumina	One set (A, B
PhiX control	FC-110-3001	Illumina	
Qubit 1X dsDNA BR Assay Kit	Q33266	Thermofisher	DNA and PC
Qubit 1X dsDNA HS Assay Kit	Q33230	Thermofisher	Pool quantific
Bioanalyzer Kit (HighSens or 1000 Kit)	5067-4626 or 5067-1504	Integrated Science Tech	
Tween 20	H5151	Promega	For Miseq inst
96-well, 0.2ml plates, non-skirted	AB0600	Thermofisher	For PCR
96 Well 0.8mL Polypropylene Deepwel Plate	AB0859	Thermofisher	For clean-up
Optical Plate Seals	V7840	Promega	For PCR
0.5mL thin-walled PCR tubes	E4942	Promega	For Qubit

3.2.8 Consumables

All consumables and equipment required for the 16S Miseq library prep protocol are available from page 21 of the above mentioned 16S Metagenomic Sequencing Library Preparation Guide.
The **magnetic stand** for 96-well plates and the TruSeq Index **Plate Fixture Kit** are essential.

Here is what we typically include into our purchasing list. The combined total of the below list costs over AUD$ 10,000.

Chapter 4

From BaseSpace to Qiime2 and DADA2

4.1 Introduction

In this chapter you will learn how to process the FASTQ files that are generated from an amplicon sequencing run on a Miseq instrument. We are using Qiime2. Please also refer to the extensive documentation and tutorials available at https://qiime2.org. You can learn almost everything there.

4.1.1 Prerequisites and required files

- FASTQ files available to download on BaseSpace. Two for each sample. One for read 1 (R1) and another for read 2 (R2). For example, `PT-01_S1_L001_R1_001.fastq.gz` and `PT-01_S1_L001_R2_001.fastq.gz` contains all sequences for sample PT-01.
- Personal computer or cloud computational resources with > 4 CPUS, > 16 RAM, > 100 GB storage recommended.
- Cloud storage browser installed (example Cyberduck or PuTTYm see Chapter 2).
- Qiime2 installed. Always use the latest version.
- All required files in correct format and file extension (see below).
- The BaseSpace Downloader software (available on the BaseSpace page. You will be prompted to download once you go to your files).

Table 4.1: Example of a qiime import manifest.tsv file

sample-id	forward-absolute-filepath	rev
PT-01	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-01_S1_L001_R1_001.fastq.gz	/pv
PT-02	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-02_S2_L001_R1_001.fastq.gz	/pv
PT-03	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-03_S3_L001_R1_001.fastq.gz	/pv
PT-04	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-04_S4_L001_R1_001.fastq.gz	/pv
PT-05	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-05_S5_L001_R1_001.fastq.gz	/pv
PT-06	/pvol/Sequences/20220622_PrimerTrial_RMIT/PT-06_S6_L001_R1_001.fastq.gz	/pv

Table 4.2: Example of a samplesheet.tsv file

#SampleID	#Sludge	WWTP	SludgeType	PMA_treat	Rep_bio	Sludgecollection
#q2:types	categorical	categorical	categorical	categorical	numeric	categorical
PT-01	1	ETP5	Meso	NoPMA	1	9/5/2022
PT-02	1	ETP5	Meso	NoPMA	2	9/5/2022
PT-03	1	ETP5	Meso	NoPMA	3	9/5/2022
PT-04	1	ETP5	Meso	NoPMA	4	9/5/2022
PT-05	1	ETP5	Meso	NoPMA	5	9/5/2022

`Qiimeimportmanifest.tsv` - A tab separated file with three columns where the first column is the sample ID (The exact same ID that was used in the manifest file for loading the Miseq). The second and third columns are the absolute paths to the forward and reverse FASTQ files, respectively. If you are working on a cloud computer then the FASTQ files should be uploaded to that, and the path needs to contain the full path to the relevant folder, including the full name of the file itself. If you process those files on your local drive then the paths need to change to your local folder containing the FASTQ files. In the below example table, the files are located on a Linux cloud computer (Nectar Research Cloud).

It might take some time to compile the import manifest. Each file path has to exact. However, in case you made a mistake in the file names etc, Qiime is generally pretty good at reporting where the error occured.

`samplesheet.tsv` - file containing sample metadata. This file should contain all other environmental measurements you may have done on each sample (pH, EC...), which you want to relate to microbial taxonomy and composition. The same file and format will be used to import metadata into the R package Phyloseq later too.

IMPORTANT: Create the `.tsv` file as a tab separated `.csv` file in Excel first and then manually change the file name from `.csv` to `.tsv`. Not sure how else to easily build the `.tsv` file.

`scriptfile.txt` - A text file that contains your qiime2 scripts. Example below. Code is compiled here and then only the file itself is executed to run any commands in that text. The hash (#) in front of any line of code stops it from being executed. That means all code is normally 'hashed'. If one wants to run the code the lines are 'un-hashed' and the text file is executed in the command line to run the 'un-hashed' code. On the Linux cloud computer the command would be: `sh scriptfile.txt`.

Note that a script file is meant to make life easy and the process replicable. But you could run each of the processes directly by adding the commands into the terminal without a script file.

Briefly on script files for cloud-based high performance computers (HPC) that are using Slurm to manage multiple user runs. If the HPC at your institution is using Slurm, then the script file has to be called with `sbatch scriptfile.txt`. The script files for HPCs using slurm contains a header that determins the allocations (number of CPUs, RAM etc..) for your task. The first lines of the .txt file typically look similar to this:

```
#!/bin/bash
#SBATCH --cpus-per-task=6
#SBATCH --mem-per-cpu=6000
#SBATCH --time=24:00:00
#SBATCH --partition=compute
#SBATCH --job-name=Q_fun
```

Commands, such as this:

```
echo "Starting at: $(date)"
echo "Finished at: $(date)"
```

.. are less important but help to report the start and end times of the executed commands from beginning (where the `echo` command is placed at the beginning of the script) to end (`echo`... placed at the end of the script) into the Terminal Windows. They dont affect the qiime commands in any way.

Slurm and `sbatch` are not needed if you manage your own computer or your 'own' cloud computer. In the example below we are using a Linux cloud computer, provided by the Nectar Research Cloud so we dont need Slurm.

```
sce#!/bin/bash
#SBATCH --cpus-per-task=8
#SBATCH --mem-per-cpu=6000
#SBATCH --time=96:00:00
#SBATCH --partition=compute
#SBATCH --job-name=Q_bac

start=$(date +'%Y-%m-%d %H:%M:%S')
echo "start time $start"


### 24/06/2022 Experiment 1

### USE THIS FILE TO IMPORT FASTQ FILES INTO A qiime2 file called 'demux-paired-end.qza' or demux-sgle.qza" (or
similar)
### Place this script into the same folder that contains the fastq files
### Output files go into another folder for further processing
### Sequences compiled in /pvol/Sequences/20220622_PrimerTrial_RMIT/


# Load qiime
eval "$(conda shell.bash hook)"
conda activate qiime2-2022.2


##### ----- V4_U

###--------------------------- PAIRED END MANIFEST IMPORT (Step 1)

#qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' \
#       --input-path /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/Importmanifest.tsv \
#       --output-path /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/demux-paired-end.qza \
#       --input-format PairedEndFastqManifestPhred33V2

#qiime demux summarize \
#       --i-data /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/demux-paired-end.qza \
#       --o-visualization /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/demux-paired-
end.qzv


###----------------------------- SINGLE ENDED MANIFEST IMPORT (Step 1)

##### ----- V4_U
# qiime tools import --type 'SampleData[SequencesWithQuality]' \
#       --input-path /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/
Importmanifest_single.tsv \
#       --output-path /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/demux_single.qza \
#       --input-format SingleEndFastqManifestPhred33V2

# qiime demux summarize \
#       --i-data /pvol/workdirs/Project1C/202206_PrimerTrial/PreProcessingFastQ/V4_U/demux_single.qza \


conda deactivate
stop=$(date +'%Y-%m-%d %H:%M:%S')
```

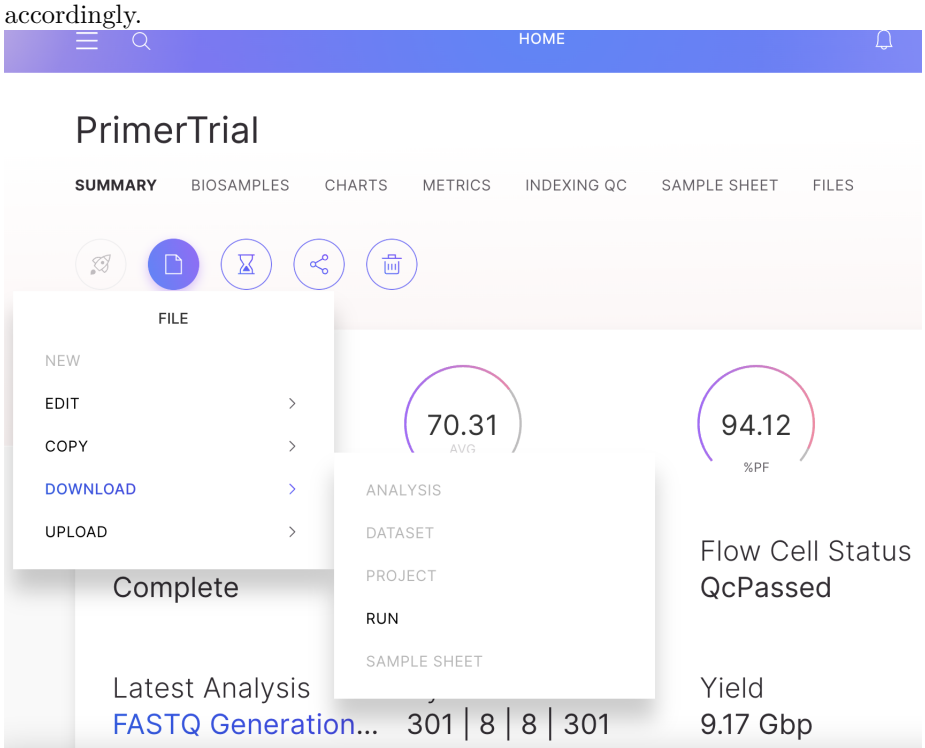
Figure 4.1: “Example of a scriptfile.txt”

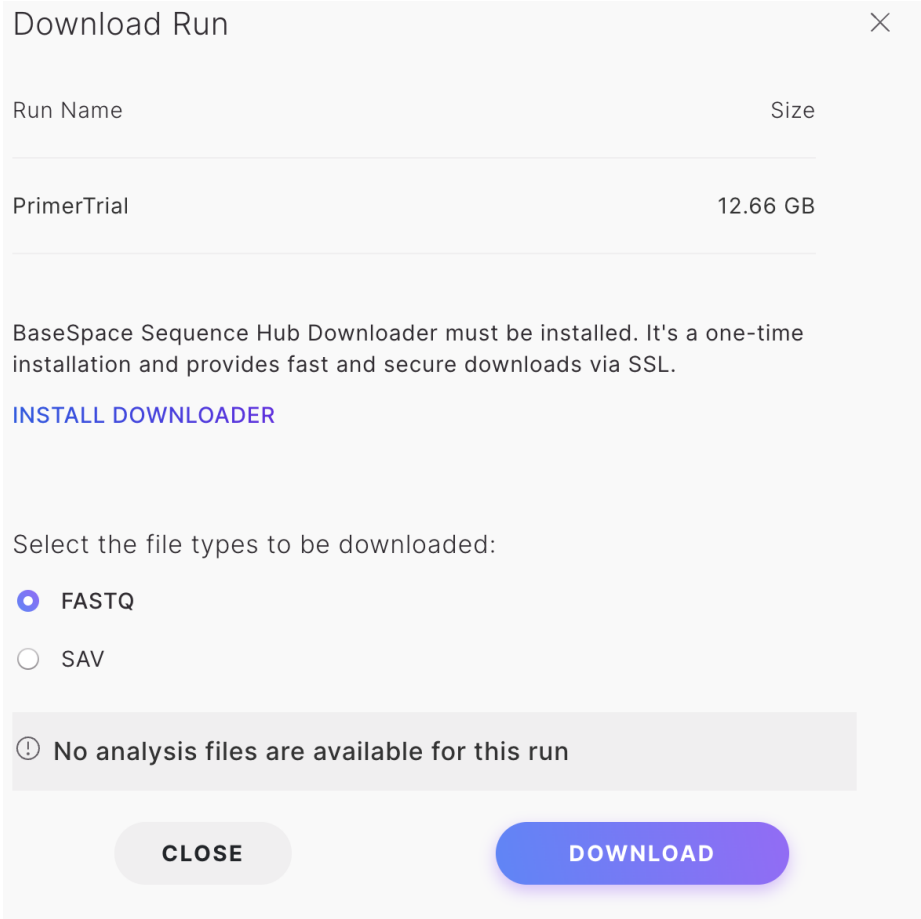
4.2 Workflow

4.2.1 Download FASTQ files

In this example, Fastq files were produced from sequencing 16SrRNA marker genes on an Illumina Miseq instrument. If you ran your own library on your own BaseSpace account, the files should be available to download from that BaseSpace account. If the run was created by another user, that person can invite you to the run, which gives you access to download the files too.

Log in to BaseSpace and into your project. Go to files, run and download FastQ. You will be asked to install a downloader software. Follow instructions accordingly.





The FASTQ files will come in separate folders for each sample. It is necessary to move all files into just ONE folder. That folder will then be used to import the sequences into a qiime object (described below).

4.2.2 Paired end manifest import (Step 1)

Go to the folder that contains all the FASTQ files (all files in one folder). I.e. in Unix/Linux use `cd path/to/folder/` to navigate to it. The scriptfile.txt should also be located here. The following qiime command that will look for the files and then create a qiime object using the sequences found in each FASTQ file. The `--output-path` instructs qiime where to place and how to name the output file, here demux-paired-end.qza is placed into the same directory.
Remove any # in front of your code (already removed in the below example) and

execute the scriptfile.txt with `sh scriptfile.txt` (or `sbatch scriptfile.txt` in case you are using Slurm).

```
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path Qiimeimportmanifest.tsv \
  --output-path demux-paired-end.qza \
  --input-format PairedEndFastqManifestPhred33V2
```

4.2.3 Cutadapt (Step 2)

If all goes well you should now have a demux-paired-end.qza file in the same directory as your FASTQ files , which contains all your sequences. You could move demux-paired-end.qza to another folder and continue all other steps there if you wish. The FASTQ files are not needed anymore. Dont forget to move your script file to where it is most convenient for executing the commands. If you moved the demux-paired-end.qza to another folder it is probably easiest to also move the script file that the same directory (or create a second script file containing the remaining commands).

Go to the folder containing demux-paired-end.qza and use cutadapt to trim out the primer sequences. Unhash relevant lines in your script file, which should be located in the same folder. Run `sh scriptfile.txt`.

In this example we used the V4 primers. Change primer sequence to the exact primers that you used in the amplicon PCR.

```
qiime cutadapt trim-paired \
  --i-demultiplexed-sequences demux-paired-end.qza \
  --p-front-f GTGYCAGCMGCCGCGGTAA \
  --p-front-r GGACTACNVGGGTWTCTAAT \
  --o-trimmed-sequences trimmed_demux-paired-end.qza
```

4.2.4 Read quality assessment

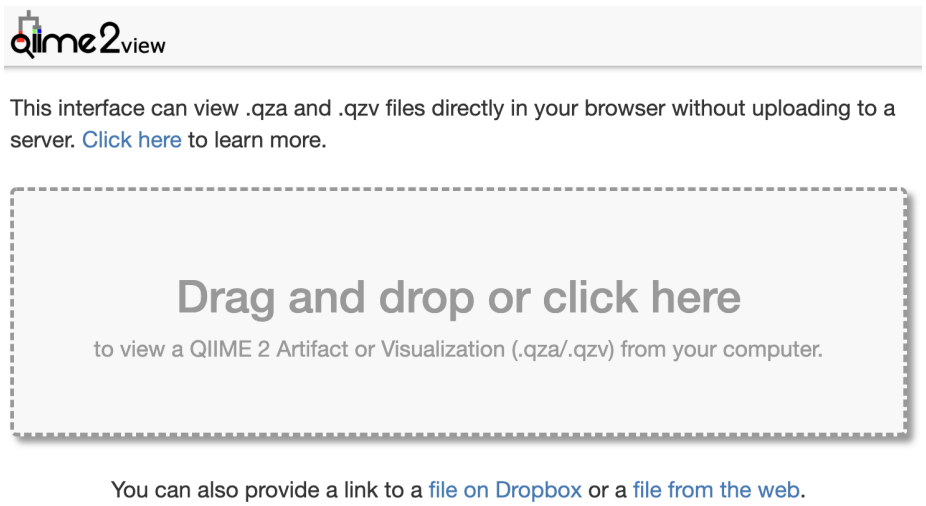
Visualise the output with a .qzv file

The .qza files can be visualised by 'converting' them into .qzv files.
Here we take the trimmed_demux-paired-end.qza and create a trimmed_demux-paired-end.qzv.

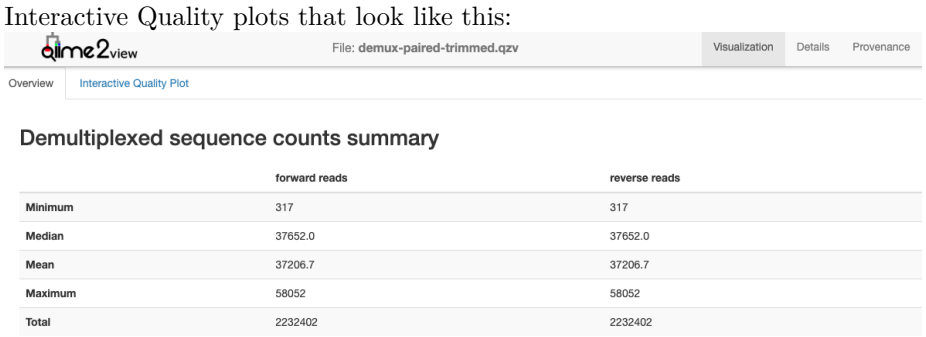
NOTE: View any .qzv file on https://view.qiime2.org. Drag and drop the qzv file into the browser window and inspect the results.

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

Go to https://view.qiime2.org and drag and drop to visualise your .qzv file in the browser.



Now click on the 'Interactive Quality Plot' tab. You will see exactly that: Interactive Quality plots that look like this:



	forward reads	reverse reads
Minimum	317	317
Median	37652.0	37652.0
Mean	37206.7	37206.7
Maximum	58052	58052
Total	2232402	2232402

Note the total number of sequences. You want to retain a high percentage of these after quality filtering with DADA2.

STOP HERE. Inspect the visualisation and decide on location and maximum expected error.
From the output decide where to truncate the forward and reverse reads with

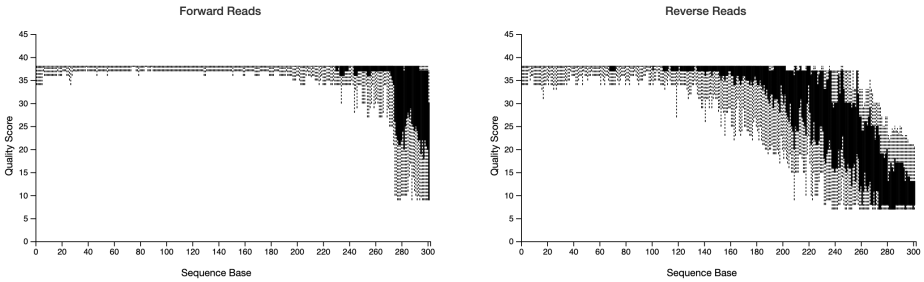


Figure 4.2: "Example of a demux.qzv visualisation"

`p-trunc-len-f`, `p-trunc-len-r`,`--p-max-ee-f` and `--p-max-ee-r` in DADA2 below.

It can take some trial and error to get these setting right. We are using a docker-based package called FIGARO to help us estimate those parameters (Not shown here). But essentially you want to capture high quality reads and be confident about the Amplicon Sequence Variants (ASV), while also capture sufficent depth of ASVs and reads without unnecessarily filtering out too much. In the below example, we have trimmed the forward reads at 272 base pairs with a maximum expected error (max-ee) of 2 (which is the default) and the reverse reads at 151 with a max-ee of 3. I think it is o.k. to relax the max-ee for the reverse reads (which are ALWAYS lower in quality) as I feel more confident about the fact that the reverse reads are paired with the forward reads. Pairing in itself provides increased confidence that the reads do in fact represent a biological relevant sequence. As always please comment on our GitHub discussion page if you have any suggestions here. Thanks!

A minimum overlap between the forward and reverse primer of 20 base pairs is recommended. To overlap can be calculated as following:

length forward read + length reverse read - length amplicon - truncated basepairs forward read - truncated basepairs reverse reads = overlap

So, for example, if we picked `--p-trunc-len-f 272` and `--p-trunc-len-r 151`, we get

$$301 + 301 - 292 - 29 - 150 = 131 \; bp \; overlap$$

In cases where the quality of the reverse reads is very poor, or the amplicon is too long for pairing to work, it is also acceptable to import, trim and denoise only the forward reads. The V4 primer of this example, is nice and short at 292 basepairs, so is great for pairing even at lower reverse-read qualities.

4.2.5 Denoise paired end sequences with dada2 (Step 3)

Once the trimming and max-ee parameters are decided, it is time to run the DADA2 function. This may take a while, depending on total number of samples.

The output will be a feature_table.qza and sample_rep_seqs.qza, containing the ASV abundances and their sequences respectively.

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs trimmed_demux-paired-end.qza \
  --o-table feature_table.qza \
  --o-representative-sequences sample_rep_seqs.qza \
  --p-trim-left-f 0 --p-trim-left-r 0 \
  --p-trunc-len-f 271 \
  --p-trunc-len-r 151 \
    --p-max-ee-f 2 \
    --p-max-ee-r 3 \
  --output-dir dada2 \
  --verbose
```

Summarise and visualise the ASV abundances (feature_table.qza) in a .qzv file.

```
qiime feature-table summarize \
  --i-table feature_table.qza \
  --o-visualization feature_table.qzv \
  --m-sample-metadata-file samplesheet.tsv
```

Note: Look at the feature_table.qzv and record median reads per sample. Compare the total frequency after denoising with the total sequence count from the trimmed_demux-paired-end.qza. You hope to retain a high percentage of total sequences after denoising with DADA2. It is prudent to report that percentage in your methods.

4.2.6 Taxonomic classifier and assignment (Step 4)

The next step is to assign taxonomies to the sequences in the denoised sample sequences. Here, we use a pre-trained classifier that is based on the Silva database. This pre-trained classifier is available on the data resource page of Qiime (Most current link at the time of writing: https://docs.qiime2.org/2022.2/data-resources).

However, in case you used a different primer you would have to create the classifier yourself. Again there is great resource avaiable on https://docs.qiime2.org.

qiime2view File: feature_table29June-1.qzv

Overview | Interactive Sample Detail | Feature Detail

Table summary

Metric	Sample
Number of samples	67
Number of features	5,591
Total frequency	1,585,492

Frequency per sample

	Frequency
Minimum frequency	17.0
1st quartile	17,987.0
Median frequency	22,377.0
3rd quartile	31,129.0
Maximum frequency	50,243.0
Mean frequency	23,664.059701492537

Frequency per sample detail (csv | html)

nber of samples

Figure 4.3: “feature_table.qzv output”

The process is fairly straight forward but takes computational time. Briefly, extract reference reads from a database (i.e. Silva here) based on the primers used. Then use those extracted sequences and fit or train them onto a taxonomy. Basically, predict which amplicon sequence should be what phylym/class/order/family/genus etc.... This trained file becomes a 'classifier' that is used to assign taxonmies on your sequences.

Here we have a pre-trained classifier, silva-132-99-515-806-nb-classifier.qza: The output is a file called taxonomy_silva.qza and taxonomy_silva.qzv in this case.

```
qiime feature-classifier classify-sklearn \
  --i-classifier silva-132-99-515-806-nb-classifier.qza \
  --p-reads-per-batch 10000 \
  --i-reads sample_rep_seqs.qza \
  --o-classification taxonomy_silva.qza \
  --quiet

# Then summarise and visualise the output into a .qza file

qiime metadata tabulate \
--m-input-file taxonomy_silva.qza \
--o-visualization taxonomy_silva.qzv
```

4.2.7 Build phylogenetic tree (Step 5)

The next step is not essential but really good to have. Creating a phylogenetic tree from the amplicon sequences.

In this case we are using the insertion tree method. See https://library.qiime2.org/plugins/q2-fragment-insertion/16/ for more information in this method.

As not all ASVs will be inserted we will filter the feature_table.qza again to keep only those ASVs that are in the tree. You will need the reference file from silva or greengenes. In this case we are using `sepp-refs-silva-128.qza`.

```
qiime fragment-insertion sepp \
  --i-representative-sequences sample_rep_seqs.qza \
  --i-reference-database sepp-refs-silva-128.qza \
  --o-tree insertion-tree.qza \
  --o-placements insertion-placements.qza

qiime fragment-insertion filter-features \
  --i-table feature_table.qza \
```

```
  --i-tree insertion-tree.qza \
  --o-filtered-table feature_table_insertiontreefiltered.qza \
  --o-removed-table removed_features.qza
```

Done!

Everything else including further quality filtering happens with `phyloseq` in R where we will import the following files: `feature_table_insertiontreefiltered.qza`, `taxonomy_silva.qza` and `insertion-tree.qza`.

This will be covered in the next chapter.

4.2.8 All steps combined

Copy and paste this into your script file if needed.

```
# Manifest Import
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path Qiimeimportmanifest.tsv \
  --output-path demux-paired-end.qza \
  --input-format PairedEndFastqManifestPhred33V2

# Cutadapt
qiime cutadapt trim-paired \
  --i-demultiplexed-sequences demux-paired-end.qza \
  --p-front-f GTGYCAGCMGCCGCGGTAA \
  --p-front-r GGACTACNVGGGTWTCTAAT \
  --o-trimmed-sequences trimmed_demux-paired-end.qza

qiime demux summarize \
  --i-data trimmed_demux-paired-end.qza \
  --o-visualization trimmed_demux-paired-end.qzv

# Denoise
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs trimmed_demux-paired-end.qza \
  --o-table feature_table.qza \
  --o-representative-sequences sample_rep_seqs.qza \
  --p-trim-left-f 0 --p-trim-left-r 0 \
  --p-trunc-len-f 271 \
  --p-trunc-len-r 151 \
    --p-max-ee-f 2 \
    --p-max-ee-r 3 \
  --output-dir dada2 \
  --verbose
```

```
qiime feature-table summarize \
  --i-table feature_table.qza \
  --o-visualization feature_table.qzv \
  --m-sample-metadata-file samplesheet.tsv

# Taxonomic assignment
qiime feature-classifier classify-sklearn \
  --i-classifier silva-132-99-515-806-nb-classifier.qza \
  --p-reads-per-batch 10000 \
  --i-reads sample_rep_seqs.qza \
  --o-classification taxonomy_silva.qza \
  --quiet

# Phylogenetic tree
qiime fragment-insertion sepp \
  --i-representative-sequences sample_rep_seqs.qza \
  --i-reference-database sepp-refs-silva-128.qza \
  --o-tree insertion-tree.qza \
  --o-placements insertion-placements.qza

# Final filtering
qiime fragment-insertion filter-features \
  --i-table feature_table.qza \
  --i-tree insertion-tree.qza \
  --o-filtered-table feature_table_insertiontreefiltered.qza \
  --o-removed-table removed_features.qza
```

Qiime2 reference:
Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu

Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

Chapter 5

From Qiime2 into R - Initial diagnostics

5.1 Introduction

In this chapter you will learn how to import Qiime2-produced ASV tables, taxonomy tables and tree files into R. For this exercise we will us publicly available FastQ files that were generated by a research group in Denmark. The files are available in the Sequence Read Archives (SRA) under accession number PRJNA645373.
Reference: [1] C. Jiang, S.J. McIlroy, R. Qi, F. Petriglieri, E. Yashiro, Z. Kondrotaite, P.H. Nielsen, Identification of microorganisms responsible for foam formation in mesophilic anaerobic digesters treating surplus activated sludge, Water Res. 191 (2021) 116779. https://doi.org/10.1016/j.watres.2020.116779.

Sample IDs
You decide on the samplesIDs before you start your library prep and sequencing. For this chapter we use a different dataset to the previous chatper and its sampleIDs are longer, e.g. "SRR12204258" and "SRR12204269", compared to "PT-01", "PT-02" in the previous chapter. It does not matter how you label each of your indexed samples as long as the IDs are unique. And each unique sample ID in the `samplesheet.csv` needs to match the sample IDs of the `feature_table.qza`. You may recall those IDs originate from your Miseq run and are the same sample IDs that are used on the loading manifest file to prepare the run. They subsequently are part of the FastQ filenames and flow through into any relevant Qiime2 data. If a sample ID does not match or if there is a different number of samples between these two input files, then phyloseq will complain.

Here the first three Sample IDs ("SRR12204258" and "SRR12204269" etc..) in the ASV table from this data set.:

	SRR12204258	SRR12204269	SRR12204280
2a4a6962ccd9e5d76122cf3b3335671a	1354	2778	26
04135cecc84e61a9192cd07cf8201b37	106	512	26
7851df9af5c9f989afcc8c4b71da3c58	114	0	467
72f4c77a923b4a2bc3d0a0eaed7b659c	561	192	203

ASV IDs
The same principle applies to each ID of the ASVs. The ASVs IDs have to match between the `feature_table.qza` and the `taxonomy_silva.qza`.

Following the first three ASV IDs (Feature ID) and related taxon assignments in taxonomy_silva.qza of this dataset:

	Feature.ID	Taxon	Confidence
1	2a4a6962ccd9e5d76122cf3b3335671a	d__Bacteria; p__Chloroflexi; c__Anaerolineae; o__Anaeroline...	1.0000000
2	04135cecc84e61a919 2a4a6962ccd9e5d76122cf3b3335671a	__Bacteroidota; c__Bacteroidia; o__Bacteroidal...	0.9927922
3	7851df9af5c9f989afcc8c4b71da3c58	d__Bacteria; p__Chloroflexi; c__Anaerolineae; o__Anaeroline...	0.9215569
4	72f4c77a923b4a2bc3d0a0eaed7b659c	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Bacteroidal...	0.9994765

NOTE: After the denoising, taxonomic classification and tree alignments in Qiime2, I prefer to do all subsequent analysis in R. However, it is also possible to do much of the diversity analysis in Qiime2. In fact, some interesting plug-ins and functions in Qiime2 are not available as packages in R, hence for some specific tasks you may need to keep using Qiime2. It is up to the goals and preferences of the investigator.

Now let's work in R!

Required files

- `samplesheet.tsv` - same file that was used in the visualisation steps in Qiime2

- `feature_table.qza` - if you created a phylogenetic tree from the sequences as described in the previous chapter then the ASVs in the feature table were filtered to match the ASVs in the tree. Hence the files was named `featuretable-insertiontree-filtered.qza` to differentiate it.

- `taxonomy_silva.qza`

- `insertion-tree.qza`

5.2 Workflow

5.2.1 Packages

First install the required packages. Some packages are stored on 'remote' repositories such as GitHub, hence before you can install them you need helper packages such as `remotes`. Other packages, such as `phyloseq` are managed by BioConductor project. They provide their own package manager package for R called `BiocManager`, which helps to install from the same release.

Learn more about phyloseq here: https://joey711.github.io/phyloseq/

Both, remotes and BiocManager are already installed in the below example. Once the packages are installed, load them into your working space.

```
# install.packages("remotes")
# if (!requireNamespace("BiocManager", quietly = TRUE))
#  install.packages("BiocManager")
# remotes::install_github("jbisanz/qiime2R")
# BiocManager::install("phyloseq")

library(qiime2R)  # to import qiime.qza into an R object
library(phyloseq) # To combine all relevant data objects into one object for easy data managemen
library(tidyverse) # Compilation of packages for data management
library(stringr)  # to change some of the strings in taxonomic names
library(vegan)   # A commonly used package for numerical ecology
```

5.2.2 Import qiime-files and create a phyloseq object

Now that packages are loaded we can import the Qiime2 `.qza` and the `.tsv` files. The aim is to import everything as R objects, which are then combined into one phyloseq object. Once we have a phyloseq object, any filtering and visualisations can be run from the one object.

In the previous chapter, we used a `samplesheet.tsv` to visualise and summarise the `feature_table.qza`. That exact same samplesheet.tsv is imported here and slightly changed to make it phyloseq friendly. In this example, all relevant Qiime2 outputs were copied into a folder named 'qiimefiles'.

```
# Sample sheet
metadata <- read_tsv("qiimefiles/samplesheet.tsv")  # lets call the R object 'metadata
# Inspect the metadata object. The second row is qiime-specific information and has to

metadata2 <- metadata[c(2:nrow(metadata)),] %>%     # remove the top row and convert c
  rownames_to_column("spl")%>% # this just adds the rownames to a column and calles in
  mutate_all(type.convert) %>%
  mutate_if(is.factor, as.character) %>%  # reformatting columns to avoid any problems
  column_to_rownames("spl") %>%  # this puts the column "spl" back into rownames
  as_tibble() %>%
  column_to_rownames("#SampleID")

# str(metadata2)  # use the `str` command to inspect the data

# Then decide which of the columns you want to be factors and in which order these fac
metadata2  <- metadata2 %>%
  mutate(Reactor = factor(Reactor)) %>%  # replace column Reactor with the same values
  mutate(Location = factor(loc))     # create a new column 'Location', which is a facto

# Qiime import
SVs <-  qiime2R::read_qza("qiimefiles/featuretable-insertiontree-filtered.qza")
ASVtable <- as.data.frame(SVs$data)   # extract the data table, this has to be a dataf

# Taxonomy
taxonomy <- qiime2R::read_qza("qiimefiles/taxonomy_silva.qza") # import the qiime obje
taxonomy <- taxonomy$data  # extract the taxonomy data
## re-format the taxonomy file to split the taxonomy into columns
## remove the confidence column
taxtable <- taxonomy %>% as_tibble() %>%
  separate(Taxon, sep=";", c("Kingdom","Phylum","Class","Order","Family","Genus","Speci
  dplyr::select(-Confidence) %>%
  column_to_rownames("Feature.ID") %>%
  as.matrix()

# Tree
tree <- read_qza("qiimefiles/insertion-tree.qza")
tree <- tree$data  #extract data

# Create the phyloseq object
ps <-phyloseq(
  otu_table(ASVtable, taxa_are_rows = T),
  sample_data(metadata2),
  phyloseq::tax_table(taxtable),
  phyloseq::phy_tree(tree)
)
```

```
# ps  (not run)
# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 4218 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 14 sample variables ]
# tax_table()   Taxonomy Table:    [ 4218 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 4218 tips and 4217 internal nodes ]

# remove things out of the R environment you dont need.
rm(metadata, metadata2, SVs, taxonomy, taxtable, tree, ASVtable)

## Remove those annoying short codes in front of taxa names (i.e. p__ etc) as they
## dont look good in visualisation
tax_table(ps)[, "Kingdom"] <- str_replace_all(tax_table(ps)[, "Kingdom"], "d__", "")
tax_table(ps)[, "Phylum"] <- str_replace_all(tax_table(ps)[, "Phylum"], " p__", "")
tax_table(ps)[, "Class"] <- str_replace_all(tax_table(ps)[, "Class"], " c__", "")
tax_table(ps)[, "Order"] <- str_replace_all(tax_table(ps)[, "Order"], " o__", "")
tax_table(ps)[, "Family"] <- str_replace_all(tax_table(ps)[, "Family"], " f__", "")
tax_table(ps)[, "Genus"] <- str_replace_all(tax_table(ps)[, "Genus"], " g__", "")
tax_table(ps)[, "Species"] <- str_replace_all(tax_table(ps)[, "Species"], " s__", "")
```

From the output of the `ps` object we can see that there are 4281 ASVs in 51 samples and associated with 14 variables in the metadata.

If you wish, you can access each individual data set from the ps object with functions `otu_table(ps)@.Data`, `sample_data(ps)` `tax_table(ps)@.Data` or `phy_tree(ps)`. This can become handy if you want to change something, such as adding columns to the sample_data or changing them to a factor etc..

```
ASVs <- otu_table(ps)@.Data
metadata <- data.frame(sample_data(ps))
taxtable <- data.frame(tax_table(ps)@.Data)
tree <- phy_tree(ps)

# Inspect individual objects in your own time.
```

5.2.3 Save the `ps` object as an `.rds` file into your working directory

Optional
This step allows you to save and store the R object as a RDS file and load it back

into any future R environment, should you require it. This would be helpful if your workflow is separated into different Rscripts and you want to just load in the RDS file instead of having to import qiime files and create the phyloseq object every time.

```
saveRDS(ps, file = "ps_ProjectX_2022July")

# to load the file back into your R environment:
# ps <- readRDS(file = "ps_ProjectX_2022July")
```

5.2.4 Initial filtering

Keep the orginal `ps` object unchanged. Create new objects for subsequent filtering steps.

Often we simply filter out any ASVs that have less than x number of reads and are present in less than x number of samples. This helps to reduce noise and the overload of zeros in the dataset, which some differential abundance tests can't deal with. Perhaps, it also helps with removing ASVs that were incorrectly denoised with DADA2 in the first place. However, it also means that rare ASVs with a low prevalence are not considered. So filtering has to be done with consideration to the intended analysis.

Then filter out any phyla that came up as uncharacterized and taxa that came up as mitochondria and chloroplast.

```
# filter as needed. This will be your final otu-table.
# Although you can still filter for specific analysis if needed.
# minimum of reads per feature
ps.flt = prune_taxa(taxa_sums(ps) >= 5, ps) #minimum reads per feature

# ps.flt (not run)
# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 3911 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 14 sample variables ]
# tax_table()   Taxonomy Table:    [ 3911 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 3911 tips and 1694 internal nodes ]

#filter any "NA"-phyla that have not been classified i.e.
# contain nothing in the phylum column of the taxtable (just a <NA>)

ps.flt  = subset_taxa(ps.flt , !is.na(Phylum) & !Phylum %in% c(""))

# ps.flt (not run)
# phyloseq-class experiment-level object
```

```
# otu_table()   OTU Table:         [ 3892 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 14 sample variables ]
# tax_table()   Taxonomy Table:    [ 3892 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 3892 tips and 1687 internal nodes ]

# Filter out non-bacteria, mitochondia and chloroplast taxa
ps.flt <- ps.flt %>%
  subset_taxa(Kingdom == "Bacteria" & Family  != "Mitochondria" & Class   != "Chloroplast")

# ps.flt (not run)
# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 3853 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 14 sample variables ]
# tax_table()   Taxonomy Table:    [ 3853 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 3853 tips and 1673 internal nodes ]

# We can filter more for running beta diversity or differential abundance analysis later. This wi
# For example:
# minimum presence in x% of samples (51 * 0.04 = 2 samples)
# ps.flt = filter_taxa(ps.flt, function(x) sum(x > 0) > (0.04*length(x)), TRUE)
# Would result in
# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 1674 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 14 sample variables ]
# tax_table()   Taxonomy Table:    [ 1674 taxa by 7 taxonomic ranks ]
# phy_tree()    Phylogenetic Tree: [ 1674 tips and 1673 internal nodes ]
```

5.2.5 Proportion of ASVs identified

Check what proportion of ASVs could be identified on phylum and genus-level. This is best done with the handy phyloseq function `ntaxa()`, which gives the number of taxa present in the phyloseq object.

This will indicate how many reads you would be discarding, in case you remove ASVs, which were not assigned a taxon.

It is also an indicator of how well the database you have been using (e.g. Silva here) has performed.You could compare that with other databases if you wish.

```
#  percent of phyla identified
ps.flt  = subset_taxa(ps , !is.na(Phylum) & !Phylum %in% c(""))
```

```
# percent of phyla identified
ntaxa(ps.flt) / ntaxa(ps)
```

```
## [1] 0.9947843
```

```
ps.flt  = subset_taxa(ps , !is.na(Genus) & !Genus %in% c(""))
# percent of phyla identified
ntaxa(ps.flt) / ntaxa(ps)
```

```
## [1] 0.940256
```

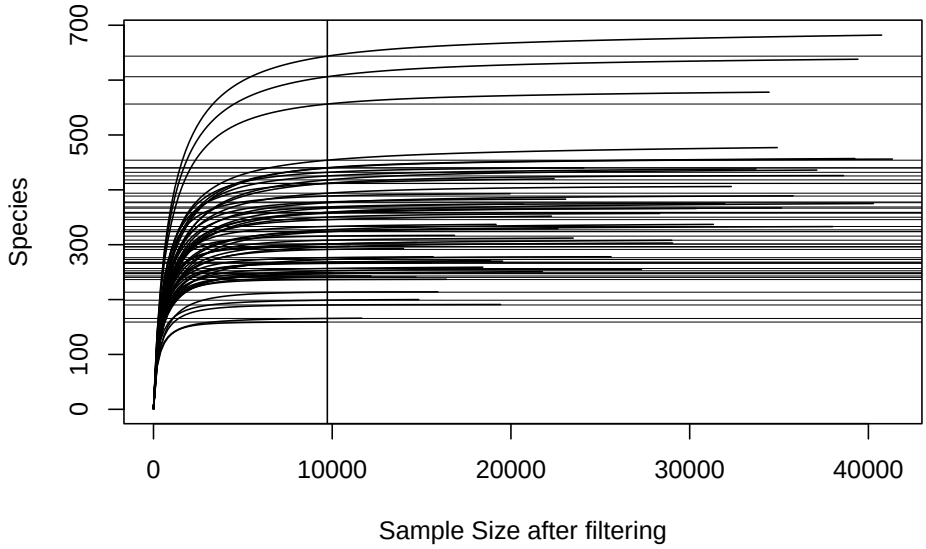
5.2.6 Rarefaction curve

Check if you cover enough depth for diversity analyses. Some samples may have low species richness, including a negative or samples that failed during PCR for example. The rarefaction curve will help assess which samples you may have to exclude from subsequent diversity calculations. It also provides a general view if samples cover enough of the potential richness (Number of ASVs).

The curve shows you the number of 'Species' (i.e. ASVs) on the y axis and the total reads on the x axis for each sample. So basically, it shows how rich each sample is and indicates if the number of reads that are captured in a samples covers enough of the ASVs. If the curve flattens, it indicates that the sample would not get any richer even with more reads...

I typically use the rarefaction curve to check if there are samples that need removing before establishing alpha diversity metrics. In this example, the vertical line helps identify the sample with lowest number of reads. We can check that by running `min(colSums(otu_table(ps.flt)))`, which tells us that they are 9690 reads in the sample with lowest reads. If this sample is kept in and we measure subsequent alpha diversity metrics, where each of the sample-reads are randomly resampled to the lowest read-number (here 9690), then we may not capture the diversity of samples with higher richness. The grey horizontal lines indicate how much richness we may lose if we include the sample with lowest number of reads (the difference between grey horizontal line and the curve of each sample at the highest number of reads on x-axis).

```
vegan::rarecurve(t(otu_table(ps.flt)),
                 step=200, sample = min(colSums(otu_table(ps.flt))),
                 label = FALSE, xlab = "Sample Size after filtering")
```




```
# show the lowest and highest number of sample reads.
max(colSums(otu_table(ps.flt)))
```

```
## [1] 41342
```

```
min(colSums(otu_table(ps.flt)))
```

```
## [1] 9726
```

```
# Filtering options
# filter options (not run), you can filter out specific sample or treatments from the phyloseq ob
# `%notin%` = Negate(`%in%`)  # create a %notin% filter operator for filtering

# Filtering out treatments, e.g. any sample that was labeled "Negative" in the metadata
# ps.flt.flt <-  prune_samples(sample_data(ps.flt)$Treatment %notin% c("Negative"), ps.flt)

# Filtering out any taxa that have zero reads because they were only present in the removed sampl
# ps.flt.flt <- prune_taxa(taxa_sums(ps.flt.flt) != 0, ps.flt.flt)
```

Once you identified the samples you would like to remove, you can filter out specific samples or treatments from the phyloseq objects. For example, with the `prune_samples(sample_data(ps.flt)$Treatment %notin% c("Negative"), ps.flt)` command here we would sample out any samples labeled as "Negative" in the metadata. Afterwards make sure to sample out any taxa that now contain zero reads because they were only present in the removed samples. Perhaps, create new phyloseq objects to differentiate from

your different filtered object `ps.flt.flt <- prune_taxa(taxa_sums(ps.flt) != 0, ps.flt)`.

In this example, we are keeping the sample with lowest reads. No further filtering required at this stage.

5.2.7 Prevalence table

Create a prevalence table to get an overview of phyla with low to high prevalence.

The output is a dataframe showing all phyla in order of total abundance. The mean prevalence defines how often the phylum appears across all samples on average. E.g. a mean prevalence of 10.22 means that ASVs in this phylum were present on average in 10.22 samples. Higher values indicates a 'core' relevance to the sum of samples.

This table is handy to decide if you want to remove certain phyla from some visualisations as they may not contribute to the overall analysis. It also allows for a comparison with filtered and unfiltered data for due diligence. Here it is apparent that phyla that were unclassified (i.e.) contributed 3,522 reads to the total of 1,299,791 reads. That is `3522 / sum(phyloseq::otu_table(ps)) * 100` = 0.27%. Chances are that these NA sequences may just be noise and may not represent biological relevant amplicons. Hence, I would keep them out of any further analysis.
However, in environments where you get higher contributions of unknown amplicons, it may indicate that the taxonomic database does not capture your samples well. In that case it would be wise to keep the NAs in for measuring alpha and beta diversity.

```
## Create a prevalence table to see which phyla have low prevalence
#ps.flt.prev <-  prune_samples(sample_data(physeqB.flt)$Treatment != "Reference", phys
# ps.flt.prevtab <-  prune_samples(sample_data(physeqB.flt.prevtab)$Treatment != "Blan


prevelancedf = apply(X = phyloseq::otu_table(ps.flt),
                     MARGIN = 1,
                     FUN = function(x){sum(x > 0)})

prevelancedf = data.frame(Prevalence = prevelancedf,
                          TotalAbundance = taxa_sums(ps.flt),
                          phyloseq::tax_table(ps.flt))

prevelancedf <- plyr::ddply(prevelancedf, "Phylum", function(df1){
  data.frame(mean_prevalence=mean(df1$Prevalence), total_abundance=sum(df1$TotalAbundan
```

```
  })

prevelancedf %>% arrange(desc(total_abundance))  # print table from highest to lowest abundance
```

```
##                          Phylum mean_prevalence total_abundance
## 1                    Chloroflexi        4.894366          228035
## 2                   Bacteroidota        4.491773          199720
## 3                     Firmicutes        4.298273          179454
## 4               Actinobacteriota        5.101010          161701
## 5                 Proteobacteria        4.111111          151650
## 6               Desulfobacterota        4.895652           57139
## 7                   Synergistota        6.712644           51049
## 8                 Cloacimonadota        6.652174           43842
## 9                Acidobacteriota        3.808989           35158
## 10                  Thermotogota       10.190476           33489
## 11                Spirochaetota        4.475728           24593
## 12             Verrucomicrobiota        2.989899           16246
## 13             Fermentibacterota       35.000000           14812
## 14               Patescibacteria        3.192547           14454
## 15            Caldatribacteriota        5.304348            8934
## 16               Planctomycetota        2.397727            5158
## 17                Armatimonadota        2.611111            4462
## 18 Marinimicrobia_(SAR406_clade)       20.000000            3818
## 19               Hydrogenedentes        3.950000            3411
## 20                  Nitrospirota        8.750000            2853
## 21  SAR324_clade(Marine_group_B)        3.333333            2622
## 22                           WS1        3.625000            2559
## 23              Campilobacterota        3.866667            2465
## 24                Fibrobacterota        2.739130            1874
## 25                   Myxococcota        2.761905            1366
## 26              Bdellovibrionota        2.444444            1133
## 27                 Caldisericota        3.583333            1071
## 28                         WPS-2        2.666667             897
## 29                   Sumerlaeota        3.600000             686
## 30                 Cyanobacteria        9.000000             647
## 31                Fusobacteriota        2.888889             565
## 32          Coprothermobacterota        2.800000             511
## 33               Gemmatimonadota        2.666667             304
## 34               Elusimicrobiota        2.833333             225
## 35                      CK-2C2-2        4.000000             162
## 36             Methylomirabilota        4.000000             159
## 37                  Dependentiae        1.857143              90
## 38              Latescibacterota        2.000000              89
## 39                  Zixibacteria        5.000000              74
## 40                       Sva0485        2.000000              69
```

```
## 41                  Acetothermia        2.000000              59
## 42                           WS4        1.250000              46
## 43                  Hydrothermae        1.000000              26
## 44                  Deinococcota        1.000000              11
## 45                       FCPU426        1.000000               3
## 46              Margulisbacteria        1.000000               3
## 47                        MBNT15        1.000000               2
```

```
## Compare with the original, unfiltered ps object
prevelancedf = apply(X = phyloseq::otu_table(ps),
                     MARGIN = 1,
                     FUN = function(x){sum(x > 0)})

prevelancedf = data.frame(Prevalence = prevelancedf,
                          TotalAbundance = taxa_sums(ps),
                          phyloseq::tax_table(ps))

prevelancedf <- plyr::ddply(prevelancedf, "Phylum", function(df1){
  data.frame(mean_prevalence=mean(df1$Prevalence), total_abundance=sum(df1$TotalAbundan
  })

prevelancedf %>% arrange(desc(total_abundance))  # print table from highest to lowest
```

```
##                          Phylum mean_prevalence total_abundance
## 1                    Chloroflexi        4.865429          228276
## 2                   Bacteroidota        4.451499          202022
## 3                     Firmicutes        4.283019          197782
## 4               Actinobacteriota        5.045131          165791
## 5                 Proteobacteria        4.046709          161897
## 6               Desulfobacterota        4.829060           57187
## 7                   Synergistota        6.533333           51227
## 8                 Cloacimonadota        6.387755           45296
## 9                Acidobacteriota        3.791209           35318
## 10                  Thermotogota       10.190476           33489
## 11                 Spirochaetota        4.443396           25028
## 12             Verrucomicrobiota        2.981308           16649
## 13             Fermentibacterota       35.000000           14812
## 14               Patescibacteria        3.138554           14501
## 15            Caldatribacteriota        5.304348            8934
## 16               Planctomycetota        2.433333            5478
## 17                Armatimonadota        2.611111            4462
## 18 Marinimicrobia_(SAR406_clade)       20.000000            3818
## 19                          <NA>        3.636364            3522
## 20               Hydrogenedentes        3.950000            3411
## 21                  Nitrospirota        8.750000            2853
```

```
## 22              Campilobacterota        4.187500            2710
## 23  SAR324_clade(Marine_group_B)        3.333333            2622
## 24                            WS1        3.625000            2559
## 25                Fibrobacterota        2.750000            1942
## 26                   Myxococcota        2.681818            1369
## 27              Bdellovibrionota        2.444444            1133
## 28                 Caldisericota        3.583333            1071
## 29                         WPS-2        2.666667             897
## 30                   Sumerlaeota        3.600000             686
## 31                 Cyanobacteria        9.000000             647
## 32                Fusobacteriota        2.888889             565
## 33          Coprothermobacterota        2.800000             511
## 34               Gemmatimonadota        2.666667             304
## 35               Elusimicrobiota        2.833333             225
## 36                      CK-2C2-2        4.000000             162
## 37             Methylomirabilota        4.000000             159
## 38                 Dependentiae        1.750000              94
## 39              Latescibacterota        2.000000              89
## 40                  Zixibacteria        5.000000              74
## 41                       Sva0485        2.000000              69
## 42                  Acetothermia        2.000000              59
## 43                           WS4        1.250000              46
## 44                  Hydrothermae        1.000000              26
## 45                  Deinococcota        1.000000              11
## 46                       FCPU426        1.000000               3
## 47              Margulisbacteria        1.000000               3
## 48                        MBNT15        1.000000               2
```

5.2.8 Export ASV and taxa tables to Excel

This might be useful if you would like to share your abundance and taxa tables with someone that does not use R.

```
# extract ASV table from ps object
ASVs <- data.frame(otu_table(ps.flt)) %>%
  rownames_to_column("FeatureID")

# extract taxonomy table from ps object
taxtable <- data.frame(tax_table(ps.flt))
# combine the taxonomy levels into one colum, separated by a ;
taxtable <- taxtable %>%
  rownames_to_column("FeatureID") %>% as_tibble() %>%
  unite(Taxon, sep=";", c("Kingdom","Phylum","Class","Order","Family","Genus","Species"))
# combine ASVs and taxatable into one object
ASVTable <- ASVs %>% left_join(taxtable, by = "FeatureID")
```

```
# write .csv file
# write.csv(ASVTable, "ASVTable.csv") not run
```

Note:
This code is an amalgamation from various sources. Apart from putting it together into this workflow I do not take credit for it.

Chapter 6

Plotting diversity

Some visualisations of alpha and beta diversity.

6.1 Introduction

This chapter introduces you to a basic workflow for plotting alpha and beta diversity using phyloseq and ggplot. It provides examples for loading the phyloseq object from a previously saved .rds file, then how to calculate and include alpha diversity metrics into the sample_data of the phyloseq object, and finally how to plot some of these metrics, individually or combined.

Furthermore, beta diversity, i.e. four different type of ordinations are plotted. The examples, include non-metric dimensional scaling on Bray-Curtis dissimilaries, Principal Component Analysis (Aitchison distance), Redundancy Analysis (Aitchison) and a Principal Coordinate Analysis on unweighted uniFrac distances.

The stats behind the individual metrics are not discussed. If you want to learn more about the various options to analyse your amplicon-derived diversity, perhaps start here: https://sites.google.com/site/mb3gustame/home, the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME). This is one of many other useful links and resources. Some others are listed in chapter 1.

Prior to establishing alpha diversity indices, as well as Bray-Curtis dissimilarities, we will stick to rarefying sequences to a specific library size (library size = sum of sequence counts of a sample) to normalise the data and prevent bias due to library size. Usually, this is done using the smallest library size of all samples. If you want to learn more about rarefying I recommend this paper here: Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. We will follow the recommendations from the authors of this paper to resample without replacement.

Other normalisation methods are also possible and it is up to you, the analyst, to understand and decide on the appropriate methods. I found centred-log ratios (Aitchison distances) also intuitive in interpretation, which are included into the workflow below in a PCA and RDA. Log-ratios are further discussed in the paper Microbiome Datasets Are Compositional: And This Is Not Optional.
In fact, using different types of normalisation, dissimilarities or distances, while providing different types of information, may complement each other and help in interpretation of your data.

6.1.1 Prerequisites

- a phyloseq object; either read in from a pre-saved `.rds` file or created as described in chapter 5.

Time to get started...

6.2 Workflow

6.2.1 Packages

Install the following packages if not already.

```
# BiocManager::install("microbiome") If needed
library(phyloseq)
library(ggpubr)       # a handy helper package for ggplots
library(tidyverse)
theme_set(theme_bw())  # setting the theme for ggplots
library(vegan)
library(microbiome)
```

6.2.2 Load phyloseq object

For detail on how to create a phyloseq object please see chapter 5.

```
# reading in a previously saved phyloseq object
ps <- readRDS('ps_ProjectX_2022July')
ps  # get an overview of number of taxa and samples contained in the phyloseq object
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 4218 taxa and 51 samples ]
## sample_data() Sample Data:       [ 51 samples by 15 sample variables ]
## tax_table()   Taxonomy Table:    [ 4218 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 4218 tips and 4217 internal nodes ]
```

Check the metadata contained in the phyloseq object. This is the original metadata from the public repository where the FastQ files were downloaded from (Sequence Read Archives).

There are 15 columns containing some character (chr), factor or integer (int) variables.

```
metadf <- data.frame(sample_data(ps))
str(metadf)
```

```
## 'data.frame':    51 obs. of  15 variables:
##  $ Assay.Type     : chr  "AMPLICON" "AMPLICON" "AMPLICON" "AMPLICON" ...
##  $ Bases          : int  13254234 24196788 15758554 20048406 37634632 43555904 35107436 138881
##  $ BioProject     : chr  "PRJNA645373" "PRJNA645373" "PRJNA645373" "PRJNA645373" ...
##  $ Bytes          : int  9235913 16404070 10758456 13009806 24207581 27963491 22585685 938886
##  $ Organism       : chr  "anaerobic digester metagenome" "anaerobic digester metagenome" "anae
##  $ collection_date: chr  "30/11/16" "9/3/16" "17/3/16" "5/12/16" ...
##  $ loc            : chr  "Ejby Moelle" "Soeholt" "Esbjerg Vest" "Viborg" ...
##  $ Reactor        : Factor w/ 10 levels "1A","1B","2B",..: 4 4 6 7 7 5 5 5 4 4 ...
##  $ Type           : chr  NA NA NA "DS" ...
##  $ Purpose        : chr  "reactorfoampotentials" "reactorfoampotentials" "reactorfoampotential
##  $ Experiment     : chr  "SRX8715180" "SRX8715169" "SRX8715158" "SRX8715151" ...
##  $ lat_lon        : chr  "55.398077 N 10.415041 E" "56.175302 N 9.582588 E" "55.488235 N 8.430
##  $ ReleaseDate    : chr  "2021-01-01T00:00:00Z" "2021-01-01T00:00:00Z" "2021-01-01T00:00:00Z"
##  $ Sample.Name    : chr  "16SAMP-17236" "16SAMP-17231" "16SAMP-17233" "MQ170915-14" ...
##  $ Location       : Factor w/ 12 levels "Avedoere","Damhusaaen",..: 4 11 5 12 12 12 12 2 11 11
```

6.2.3 Pre-filter, rarefy and calculate diversity

Now the diversity indices can be calculated and added to the metadata of the phyloseq object for plotting. First, the abundances are pre-filtered (as shown in the previous chapter), then the ASV abundances are randomly resampled with `phyloseq::rarefy_even_depth`, such that all samples have the same number

of ASVs.
The diversity indices are then calculated with `phyloseq::estimate_richness` based on these evenly resampled abundances. The dataframe that is created containing the range of indices is then added to the existing metadata of the phyloseq for plotting.

```
# some additional quality filtering - repeating steps from previous chapter.

ps.flt = prune_taxa(taxa_sums(ps) >= 5, ps) #minimum reads per feature
ps.flt  = subset_taxa(ps.flt , !is.na(Phylum) & !Phylum %in% c(""))
ps.flt <- ps.flt %>%
  subset_taxa(Kingdom == "Bacteria" & Family  != "Mitochondria" & Class   != "Chloropla
ps.flt
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 3853 taxa and 51 samples ]
## sample_data() Sample Data:       [ 51 samples by 15 sample variables ]
## tax_table()   Taxonomy Table:    [ 3853 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 3853 tips and 3852 internal nodes ]
```

```
# confirming minimum sample size
min(colSums(otu_table(ps.flt)))
```

```
## [1] 9770
```

```
# quantile summary of sample sizes
quantile(colSums(otu_table(ps.flt)))
```

```
##      0%     25%     50%     75%    100%
##  9770.0 16278.5 23035.0 34173.0 42345.0
```

```
# rarefy - required before measuring alpha diversity.
# This step is to create an abundance table, where the all samples are randomly
# resamples such that all samples have the same number of ASVs.
# It removes any bias but number of re-samples is based on the sample with
# smallest number of ASVs so one has to check and filter beforehand in case
# there are samples with very different number of ASVs. I am generally going
# by a 10x rule. If the lowest number of ASVs is 10x lower than the sample with
# highest number of ASVs, then I keep that sample in for alpha diversity assessments.

ps_rare <- phyloseq::rarefy_even_depth(ps.flt,
    sample.size = min(colSums(otu_table(ps.flt))),
    rngseed = TRUE,
```

```
    replace = FALSE, #without replacement
    trimOTUs = TRUE)

# check rarefied phyloseq object
# ps_rare (not run)

# phyloseq-class experiment-level object
# otu_table()   OTU Table:         [ 3801 taxa and 51 samples ]
# sample_data() Sample Data:       [ 51 samples by 15 sample variables ]
# tax_table()   Taxonomy Table:    [ 3801 taxa by 7 taxonomic ranks ]
 #phy_tree()    Phylogenetic Tree: [ 3801 tips and 3800 internal nodes ]

# Create diversity indices
div.df <- phyloseq::estimate_richness(ps_rare)
# add sample names
div.df$`#SampleID` <- phyloseq::sample_names(ps.flt)
# add diversity indices to phyloseq object
metadf <- data.frame(sample_data(ps.flt))  # export metadata first
metadf <- metadf %>%
  rownames_to_column("#SampleID") %>%
  left_join(div.df, by = "#SampleID") %>%
  column_to_rownames("#SampleID") # add the diversity indices (columns) to the metadata
# add pilou evenness (see calculation in "Numerical Ecology with R - Bocard et al")
metadf$pielou_ev <- div.df$Shannon/log(div.df$Observed)
# replace metadata with temp meta.df which includes the diversity indices
ps.flt@sam_data <- sample_data(metadf)
ps_rare@sam_data <- sample_data(metadf)
# this replaced the old metadata with new metadata that includes the diversity
# indices for this phyloseq object
str(data.frame(sample_data(ps.flt)))
```

```
## 'data.frame':    51 obs. of  25 variables:
##  $ Assay.Type     : chr  "AMPLICON" "AMPLICON" "AMPLICON" "AMPLICON" ...
##  $ Bases          : int  13254234 24196788 15758554 20048406 37634632 43555904 35107436 138881
##  $ BioProject     : chr  "PRJNA645373" "PRJNA645373" "PRJNA645373" "PRJNA645373" ...
##  $ Bytes          : int  9235913 16404070 10758456 13009806 24207581 27963491 22585685 9388862
##  $ Organism       : chr  "anaerobic digester metagenome" "anaerobic digester metagenome" "anae
##  $ collection_date: chr  "30/11/16" "9/3/16" "17/3/16" "5/12/16" ...
##  $ loc            : chr  "Ejby Moelle" "Soeholt" "Esbjerg Vest" "Viborg" ...
##  $ Reactor        : Factor w/ 10 levels "1A","1B","2B",..: 4 4 6 7 7 5 5 5 4 4 ...
##  $ Type           : chr  NA NA NA "DS" ...
##  $ Purpose        : chr  "reactorfoampotentials" "reactorfoampotentials" "reactorfoampotential
##  $ Experiment     : chr  "SRX8715180" "SRX8715169" "SRX8715158" "SRX8715151" ...
##  $ lat_lon        : chr  "55.398077 N 10.415041 E" "56.175302 N 9.582588 E" "55.488235 N 8.430
##  $ ReleaseDate    : chr  "2021-01-01T00:00:00Z" "2021-01-01T00:00:00Z" "2021-01-01T00:00:00Z"
```

```
##  $ Sample.Name    : chr  "16SAMP-17236" "16SAMP-17231" "16SAMP-17233" "MQ170915-14"
##  $ Location       : Factor w/ 12 levels "Avedoere","Damhusaaen",..: 4 11 5 12 12 12
##  $ Observed       : num  161 354 247 265 389 454 406 304 278 371 ...
##  $ Chao1          : num  161 356 247 266 395 ...
##  $ se.chao1       : num  0 2.2689 0.0832 1.0272 3.914 ...
##  $ ACE            : num  161 357 247 266 395 ...
##  $ se.ACE         : num  4.6 9.28 7.43 7.72 9.83 ...
##  $ Shannon        : num  4.24 4.84 4.74 4.78 5.03 ...
##  $ Simpson        : num  0.966 0.973 0.982 0.983 0.986 ...
##  $ InvSimpson     : num  29.6 37.7 56.8 57.8 70.6 ...
##  $ Fisher         : num  27.4 72 46.1 50.2 81 ...
##  $ pielou_ev      : num  0.834 0.825 0.861 0.857 0.843 ...
```

6.2.4 Colors

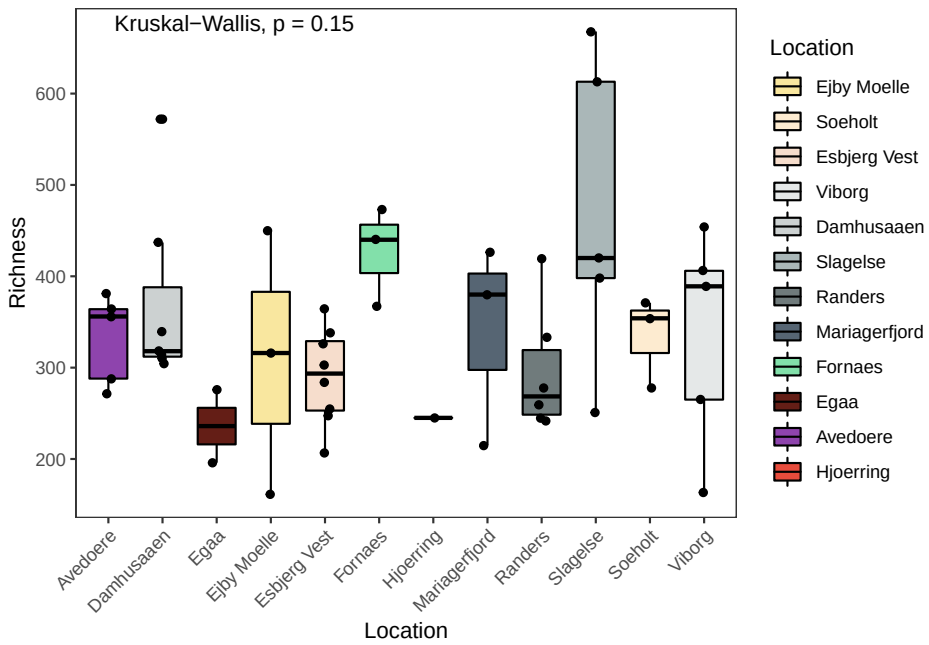
```
# You can check Hex codes with help of https://htmlcolorcodes.com/

cols <- c("#F9E79F", "#FDEBD0", "#F6DDCC",  "#E5E8E8", "#CCD1D1", "#AAB7B8",
          "#707B7C", "#566573","#82E0AA", "#641E16", "#8E44AD", "#E74C3C")

# you can also name each colour according to your treatment.
# This will help ggplot to associated the same colour for each treatment across
# different ggplots. For example, if we want to show different colour by location
# in this dataset:
names(cols)  <- unique((data.frame(sample_data(ps.flt)))$Location)
cols
```

```
##   Ejby Moelle       Soeholt  Esbjerg Vest        Viborg    Damhusaaen
##     "#F9E79F"     "#FDEBD0"     "#F6DDCC"     "#E5E8E8"     "#CCD1D1"
##      Slagelse       Randers Mariagerfjord       Fornaes          Egaa
##     "#AAB7B8"     "#707B7C"     "#566573"     "#82E0AA"     "#641E16"
##      Avedoere     Hjoerring
##     "#8E44AD"     "#E74C3C"
```

6.2.5 Richness

```
p1 <- data.frame(phyloseq::sample_data(ps.flt)) %>%
  ggpubr::ggboxplot(., x = "Location",
                    y = "Observed",
                    fill = "Location",
                    ylab = "Richness",
```

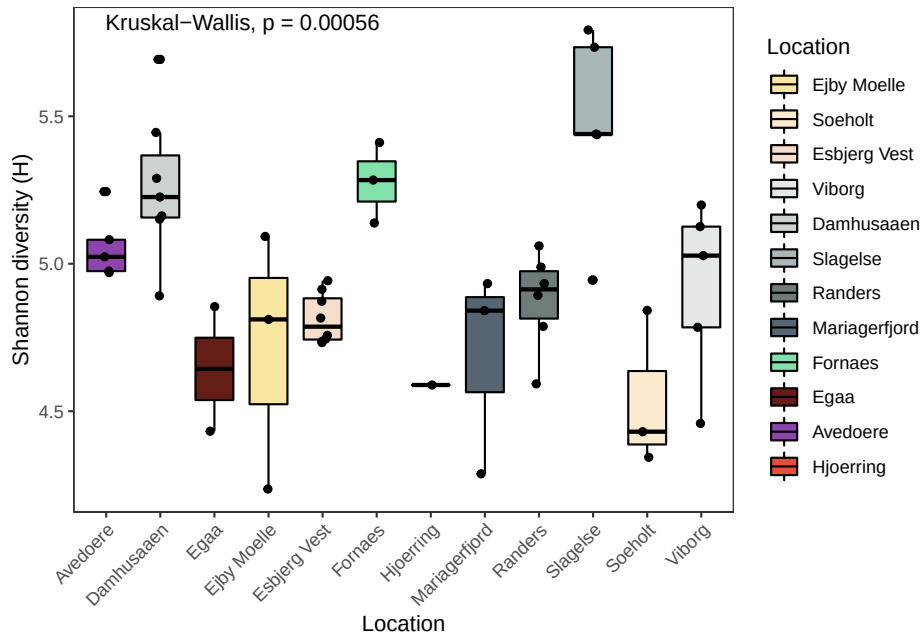
```
                    xlab = "Location",
                    palette = cols) +
    theme_bw() +
    theme( panel.grid.major = element_blank(), panel.grid.minor = element_blank() ) +
    geom_jitter(width = 0.1) +
   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
#    theme(axis.title.x=element_blank(),  # Options
#        axis.text.x=element_blank(),
#        axis.ticks.x=element_blank()) +
  ggpubr::stat_compare_means(method = "kruskal",label.x = 2)
p1
```



6.2.6 Shannon

```
# Shannon
p2 <- data.frame(phyloseq::sample_data(ps.flt)) %>%
  ggpubr::ggboxplot(., x = "Location",
                    y = "Shannon",
                    fill = "Location",
                    ylab = "Shannon diversity (H)",
                    xlab = "Location",
                    palette = cols) +
```

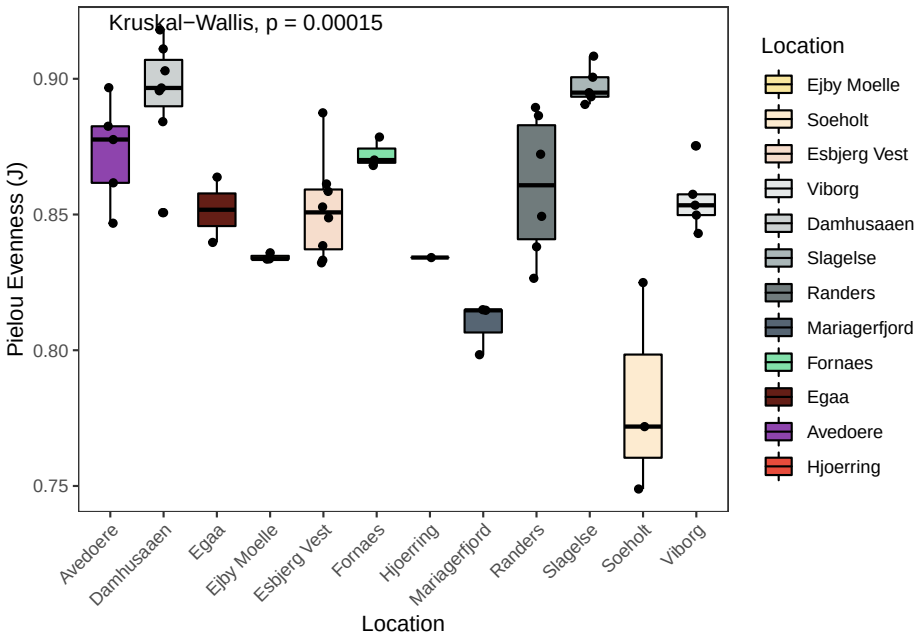
```
    theme_bw() +
    theme( panel.grid.major = element_blank(), panel.grid.minor = element_blank() ) +
    geom_jitter(width = 0.1) +
   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
#    theme(axis.title.x=element_blank(),  # Options
#        axis.text.x=element_blank(),
#        axis.ticks.x=element_blank()) +
  ggpubr::stat_compare_means(method = "kruskal",label.x = 2)
p2
```



6.2.7 Pielou's evenness

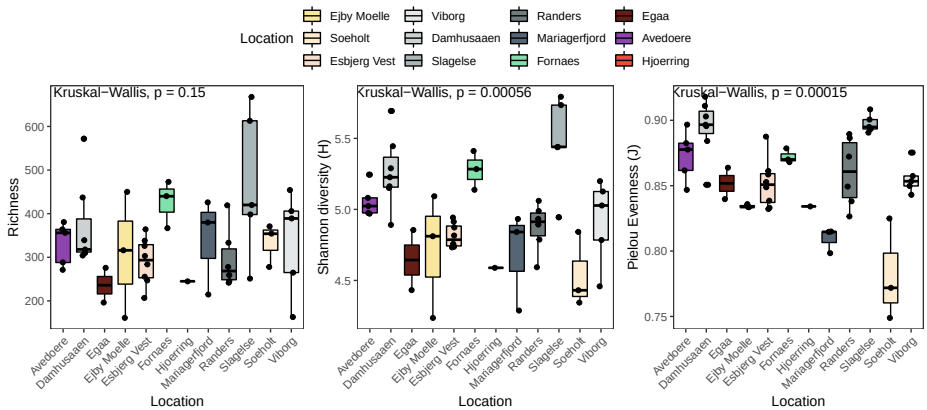
```
# Pielou
p3 <- data.frame(phyloseq::sample_data(ps.flt)) %>%
  ggpubr::ggboxplot(., x = "Location",
                    y = "pielou_ev",
                    fill = "Location",
                    ylab = "Pielou Evenness (J)",
                    xlab = "Location",
                    palette = cols) +
    theme_bw() +
    theme( panel.grid.major = element_blank(), panel.grid.minor = element_blank() ) +
```

```
    geom_jitter(width = 0.1) +
   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
#    theme(axis.title.x=element_blank(),  # Options
#        axis.text.x=element_blank(),
#        axis.ticks.x=element_blank()) +
  ggpubr::stat_compare_means(method = "kruskal",label.x = 2)
p3
```



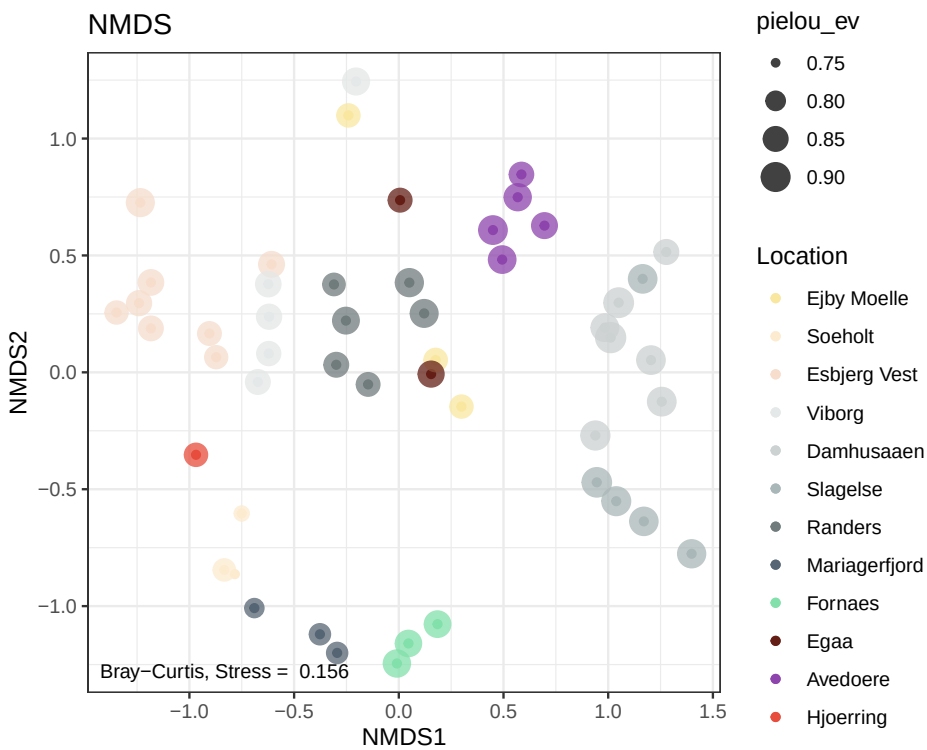
6.2.8 Combined plotting

```
#pdf(NULL)
g1 <- ggpubr::ggarrange(p1, p2, p3, common.legend = TRUE, nrow = 1, legend = 'top')
#x = dev.off()
g1
```



6.2.9 NMDS

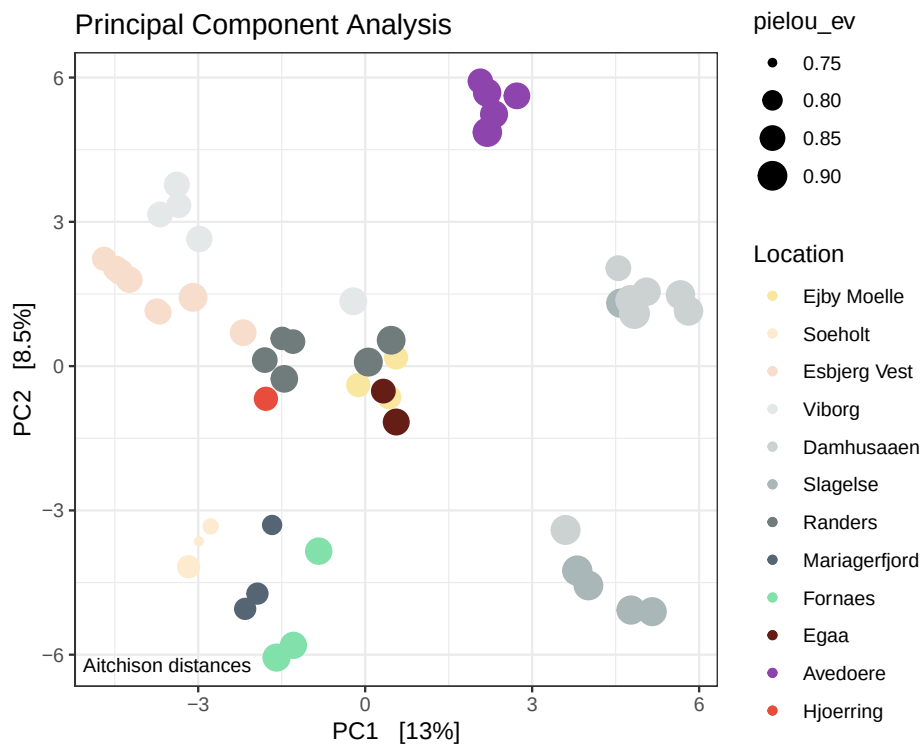
```
physeqNMDS <-  ps_rare  # We are using the rarefied abundances as explained in the int

#ordination
ordination <- phyloseq::ordinate(physeqNMDS  , "NMDS", distance = 'bray') # using Bray

label <- round(ordination$stress,3) # stress label
#plotting ordination
NMDSplot <- phyloseq::plot_ordination(physeqNMDS,
                                      ordination ,
                                      color = "Location",
                                      #shape = "Reactor",
                                      # option to add shapes as data symbols
                                      title = "NMDS") +
     geom_point(aes(size = pielou_ev), alpha = 0.75) +   # shape size in proportion to
     annotate(geom = 'text', label = paste("Bray-Curtis, Stress = ", label),
             x = -Inf, y = -Inf, hjust = -0.05, vjust = -1, size = 3)  +
     scale_color_manual(values = cols)

NMDSplot
```



6.2.10 PCA

```
physeqPCA <-  ps.flt  # just to keep it separate.

# normalisation and transform using the microbiome package
# I chose a centred log-ratio transform and a principal component analysis in a euclidean space.
physeqPCA <- microbiome::transform(physeqPCA, "clr")

#ordination
ordination <- phyloseq::ordinate(physeqPCA , "RDA") # useing the centred-log transformed abundanc

#plotting ordination
PCAplot <- phyloseq::plot_ordination(physeqPCA,
                                     ordination ,
                                     color = "Location",
                                     #shape = "Reactor",
                                    # option to add shapes as data symbols
                                     title = "Principal Component Analysis") +
     geom_point(aes(size = pielou_ev)) +    # # shape size in proportion to Pielou Evenness
```

```
    annotate(geom = 'text', label = paste("Aitchison distances"),
             x = -Inf, y = -Inf, hjust = -0.05, vjust = -1, size = 3)  +
    scale_color_manual(values = cols)

PCAplot
```

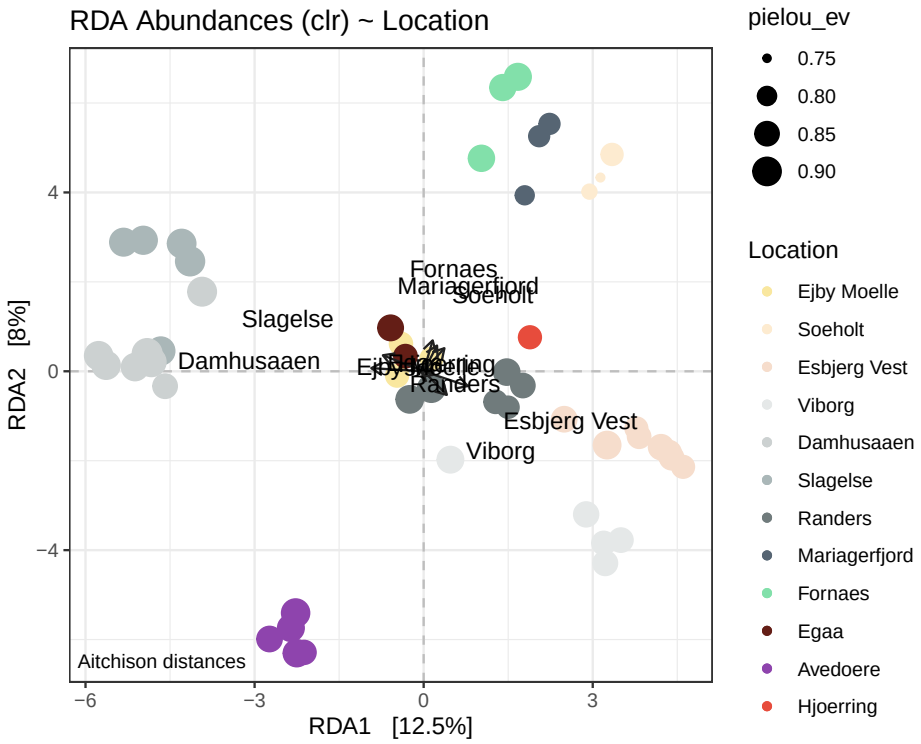


6.2.11 RDA

```
# PERCENT FILTERVALUE FOR ALL RDAs
pcnt <- 0.25
# FORMULA FOR ALL RDAs
form <- formula(~ Location)
# labels for arrows
# pos for arrows
xarw <- 5
yarw <- 5
```

```
# arrow length map
larw <- 1.5
# arrow length map for chromosol rda
larwch <- 1
# fixed width of x axis
xaxis = c(-6.75, 6.5)
# position of annotation
textposx = c(-4.4)
# arrow color
arrowcolor <- "grey13"


physeqPCA <- ps.flt # excluding Heat, H2O2, reference and blank

# normalisation and transform
physeqPCA <- microbiome::transform(physeqPCA, "clr")

#ordination
ordination <- phyloseq::ordinate(physeqPCA, "RDA", "bray", formula = form )

RDAplot <- phyloseq::plot_ordination(physeqPCA, ordination ,
                                color = "Location")  +
  geom_hline(yintercept=0, linetype="dashed", color = "gray") +
  geom_vline(xintercept=0, linetype="dashed", color = "gray") +    # shape size reflective of Pie
     annotate(geom = 'text', label = paste("Aitchison distances"),
             x = -Inf, y = -Inf, hjust = -0.05, vjust = -1, size = 3) +
  geom_point(aes(size = pielou_ev)) +        # shape size in proportion to Pielou Evenness
  #scale_colour_gradient(low = colourlow, high = colourhigh) +
  ggtitle("RDA Abundances (clr) ~ Location") +
  scale_color_manual(values = cols)


# Now add the environmental variables as arrows to either of these p1 or p2
arrowmat <- vegan::scores(ordination, display = "bp")
arrowmat <- data.frame(arrowmat)
#rownames(arrowmat) <- arrowlabel
# Add labels, make a data.frame
arrowdf <- data.frame(labels = rownames(arrowmat), arrowmat)
arrowdf$labels <- arrowdf$labels %>% gsub("Location", "", .) # remove the string "Location" from
# Define the arrow aesthetic mapping
arrow_map <- aes(xend = larw * RDA1,
                 yend = larw * RDA2,
                 x = 0,
                 y = 0,
                 shape = NULL,
```

```
                  color = NULL,
                  label = labels)

label_map <- aes(x = xarw * RDA1,
                  y = yarw * RDA2,
                  shape = NULL,
                  color = NULL,
                  label = labels)

arrowhead = arrow(length = unit(0.02, "npc"))

RDAplot <-RDAplot +
  geom_segment(
    mapping = arrow_map,
    size = .5,
    data = arrowdf,
    color = arrowcolor,
    arrow = arrowhead)  +
  geom_text(
   mapping = label_map,
  size = 4,
  data = arrowdf,
  show.legend = FALSE)


RDAplot
```



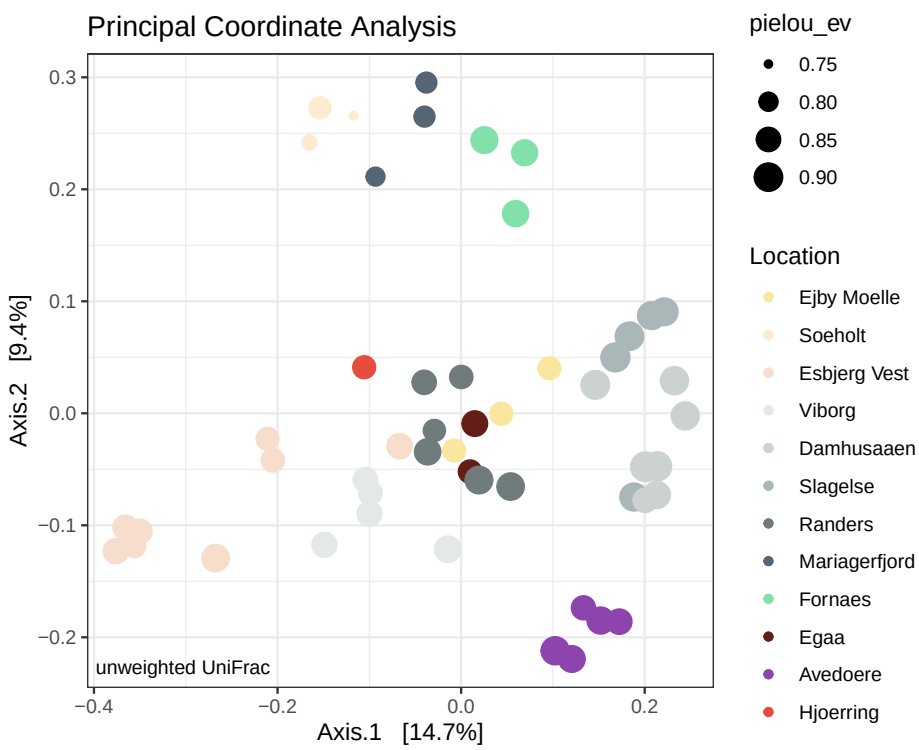

```
# save figure for publication
# ggsave("RDA.png", height=6, width=7.5, units='in', dpi=600)
#knitr::include_graphics("./RDA.png")
```

6.2.12 Unweighted uniFrac

UniFrac distances are derived from phylogenetic distances between ASVs, which means that a phylogenetic tree needs to be included into the phyloseq object.

```
physeqPcOA <-  ps.flt  # just to keep it separate.

#ordination
ordination <- phyloseq::ordinate(physeqPcOA  ,
                                 "PCoA",
                                 distance = 'unifrac') # using just presence and absence (unweigh

#plotting ordination
```

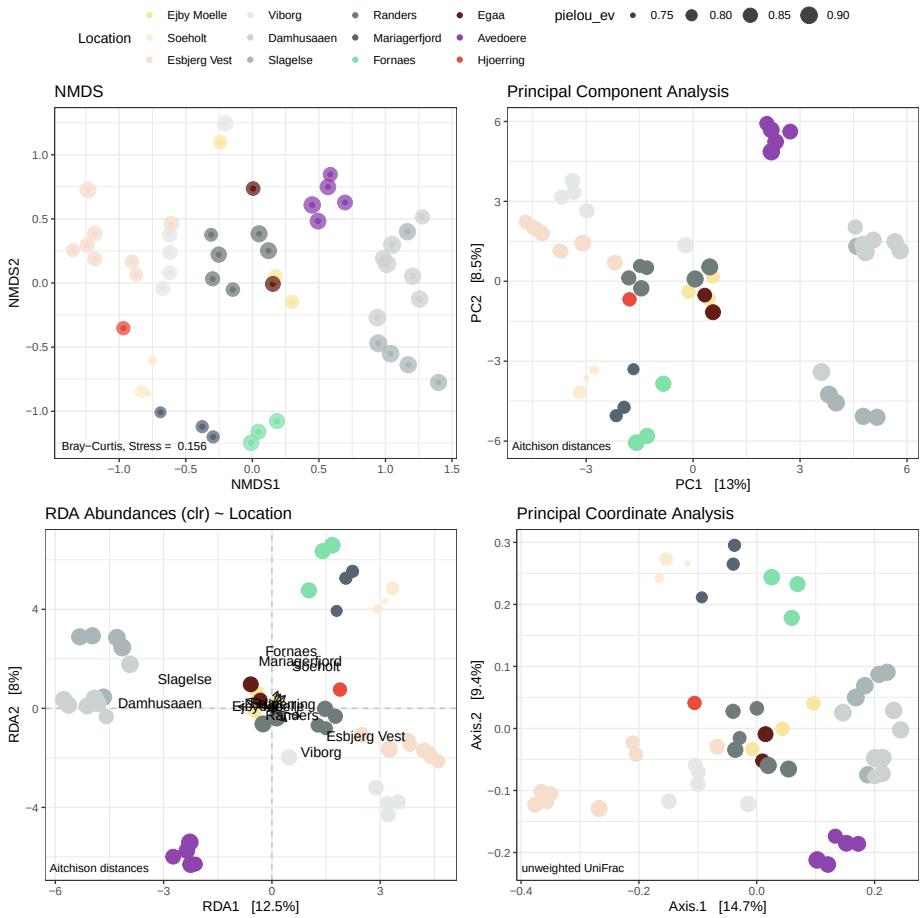
```
PCOAplot <- phyloseq::plot_ordination(physeqPcOA,
                                      ordination ,
                                      color = "Location",
                                      #shape = "Reactor",
                                      # option to add shapes to data symbols
                                      title = "Principal Coordinate Analysis") +
     geom_point(aes(size = pielou_ev)) +    # shape size reflective of Pielou
     annotate(geom = 'text', label = paste("unweighted UniFrac"),
             x = -Inf, y = -Inf, hjust = -0.05, vjust = -1, size = 3)  +
     scale_color_manual(values = cols)

PCOAplot
```



6.2.13 Combined plotting

```
#pdf(NULL)
g2 <- ggpubr::ggarrange(NMDSplot, PCAplot, RDAplot, PCOAplot,
                        common.legend = TRUE, nrow = 2, ncol = 2, legend = 'top')
```

```
#x = dev.off()
g2
```



6.2.14 Save plots to png and pdf

```
# either as png or pdf
ggsave("alphadiversity.png", plot = (g1), height=4, width=10, units='in', dpi=600)
ggsave("alphadiversity.pdf", plot = (g1), height=4, width=10, units='in', dpi=600)

ggsave("betadiversity.png", plot = (g2), height=10, width=10, units='in', dpi=600)
ggsave("betadiversity.pdf", plot = (g2), height=10, width=10, units='in', dpi=600)
```

Note: This code is an amalgamation from various sources. Apart from putting it together into a pipeline I do not take credit for it.