{: style="width:350px; padding-right:50px"}     {: style="width:150px"}

# QIIME2

Anticipated workshop duration when delivered to a group of participants is **4 hours**.

For queries relating to this workshop, contact Melbourne Bioinformatics (bioinformatics-training@unimelb.edu.au).

## Overview

### Topic

- [x] Genomics
- [ ] Transcriptomics
- [ ] Proteomics
- [ ] Metabolomics
- [ ] Statistics and visualisation
- [ ] Structural Modelling
- [ ] Basic skills

### Skill level

- [ ] Beginner
- [x] Intermediate
- [ ] Advanced

This workshop is designed for participants with command-line knowledge. You will need to be able to `ssh` into a remote machine, navigate the directory structure and `scp` files from a remote computer to your local computer.

### Description

What is the influence of genotype (intrinsic) and environment (extrinsic) on anemone-associated bacterial communities?

**Data:** Illumina MiSeq v3 paired-end (2 × 300 bp) reads (FASTQ).

**Tools:** QIIME 2

**Pipeline:**

*Section 1:* Importing, cleaning and quality control of the data
*Section 2:* Taxonomic Analysis
*Section 3:* Building a phylogenetic tree
*Section 4:* Basic visualisations and statistics
*Section 5:* Exporting data for further analysis in R
*Section 6:* Extra Information

## Learning Objectives

At the end of this introductory workshop, you will:

- Take raw data from a sequencing facility and end with publication quality graphics and statistics
- Answer the question *What is the influence of genotype (intrinsic) and environment (extrinsic) on anemone-associated bacterial communities?*

# Tutorial layout

- There is a `Table of contents` on the right-hand side which can be used to easily navigate through the tutorial by clicking the relevant section.

```
These grey coloured boxes are code blocks. The rectangular boxes in the top
right hand corner of this code block/grey box can be used to copy the code to
the clipboard.
```

??? example "Coloured boxes like these with > on the far right hand side, can be clicked to reveal the contents." REVEALED!

!!! attention "Attention: Pay attention to the information in these boxes." Important information, hints and tips.

# Requirements and preparation

!!! attention "Important" **Attendees are required to use their own laptop computers.**

```
At least one week before the workshop, if required, participants should install the software below.  T
```

## Mode of Delivery

This workshop will be run on a Nectar Instance. For more information click here.

It is possible to apply for your own Nectar allocation to run this workshop independently though you will need to install any software needed.

## Required Software

For information about required software, click here.

You will need to use a Google Chrome or Mozilla Firefox web browser to view files in QIIME2 View.

## Required Data

- No additional data needs to be downloaded for this workshop - it is all located on the Nectar Instance. FASTQs are located in the directory `raw_data` and a metadata ( `metadata.tsv` ) file has also been provided.

- If you wish to analyse the data independently at a later stage, it can be downloaded from here. This zipped folder contains both the FASTQs and associated metadata file.

- If you are running this tutorial independently, you can also access the classifier that has been trained specifically for this data from here.

## Slides and workshop instructions

Click here for slides presented during this workshop.

Click here for a printer friendly PDF version of this workshop.

---

# Author Information

Written by: Ashley Dungan and Gayle Philip
School of Biosciences, University of Melbourne; Melbourne Bioinformatics

Created/Reviewed: August 2021

---

# Background

What is the influence of genotype (intrinsic) and environment (extrinsic) on anemone-associated bacterial communities?

## The Players

- *Exaiptasia diaphana* - a shallow-water, marine anemone that is often used in research as a model organism for corals. In this experiment, two genotypes (AIMS1 and AIMS4) of *E. diaphana* were grown in each of two different environments:

    1. sterile seawater *OR*
    2. unfiltered control seawater

- The anemone-associated bacterial communities or *microbiome* - these bacteria live on, or within *E. diaphana*, and likely consist of a combination of commensals, transients, and long-term stable members, and combined with their host, form a mutually beneficial, stable symbiosis.

## The Study

The anemone microbiome contributes to the overall health of this complex system and can evolve in tandem with the anemone host. In this data set we are looking at the impact of intrinsic and extrinsic factors on anemone microbiome composition. After three weeks in either sterile or control seawater (environment), anemones were homogenized and DNA was extracted. There are *23* samples in this data set - *5* from each anemone treatment combination (2 genotypes x 2 environments) and *3* DNA extraction blanks as controls. *This data is a subset from a larger experiment*.

Dungan AM, van Oppen MJH, and Blackall LL (2021) Short-Term Exposure to Sterile Seawater Reduces Bacterial Community Diversity in the Sea Anemone, *Exaiptasia diaphana*. *Front. Mar. Sci.* 7:599314. doi:10.3389/fmars.2020.599314 [Full Text].

## QIIME 2 Analysis platform

!!! attention The version used in this workshop is qiime2-2021.8. Other versions of QIIME2 may result in minor differences in results.

Quantitative Insights Into Microbial Ecology 2 (QIIME 2™) is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed. It allows researchers to:

- Automatically track analyses with decentralised data provenance
- Interactively explore data with beautiful visualisations
- Easily share results without QIIME 2 installed
- Plugin-based system — researchers can add in tools as they wish

### Viewing QIIME2 visualisations

!!! attention In order to use QIIME2 View to visualise your files, you will need to use a Google Chrome or Mozilla Firefox web browser (not in private browsing). For more information, click here.

As this workshop is being run on a remote Nectar Instance, you will need to download the visual files (*.qzv) to your local

[computer](#) and view them in [QIIME2 View](#) (q2view).

!!! attention We will be doing this step multiple times throughout this workshop to view visualisation files as they are generated.

Alternatively, *if you have QIIME2 installed and are running it on your own computer*, you can use `qiime tools view` to view the results from the command line (e.g. `qiime tools view filename.qzv`). `qiime tools view` opens a browser window with your visualization loaded in it. When you are done, you can close the browser window and press `ctrl-c` on the keyboard to terminate the command.

---

# Initial Set up on Nectar

## Byobu-screen

To ensure that commands continue to run should you get disconnected from your Nectar Instance, we'll [run a byobu-screen session](#).

### Starting a byobu-screen session

On Nectar, to start a `byobu-screen` session called `workshop`, type

```
byobu-screen -S workshop
```

### Reconnecting to a byobu-screen session

If you get disconnected from your Nectar Instance, follow the instructions [here](#) to resume your session.

---

## Symbolic links to workshop data

Data for this workshop is stored in a central location (`/mnt/shared_data/`) on the Nectar file system that we will be using. We will use symbolic links (`ln -s`) to point to it. Symbolic links (or symlinks) are just "virtual" files or folders (they only take up a very little space) that point to a physical file or folder located elsewhere in the file system. Sequencing data can be large, and rather than unnecessarily having multiple copies of the data which can quickly take up a lot of space, we will simply point to the files needed in the `shared_data` folder.

```
 cd
 ln -s /mnt/shared_data/raw_data raw_data
 ln -s /mnt/shared_data/metadata.tsv metadata.tsv
 ln -s /mnt/shared_data/silva_138_16s_v5v6_classifier_2021-4.qza silva_138_16s_v5v6_classifier_2021-4.q:
```

---

# Section 1: Importing, cleaning and quality control of the data

## Import data

These [samples](#) were sequenced on a single Illumina MiSeq run using v3 (2 × 300 bp) reagents at the Walter and Eliza Hall Institute (WEHI), Melbourne, Australia. Data from WEHI came as paired-end, demultiplexed, unzipped *.fastq files with adapters still attached. Following the [QIIME2 importing tutorial](#), this is the Casava One Eight format. The files have been renamed to satisfy the Casava format as SampleID_FWDXX-REVXX_L001_R[1 or 2]_001.fastq e.g. CTRLA_Fwd04-

Rev25_L001_R1_001.fastq.gz. The files were then zipped (.gzip).

Here, the data files (two per sample i.e. forward and reverse reads `R1` and `R2` respectively) will be imported and exported as a single QIIME 2 artefact file. These samples are already demultiplexed (i.e. sequences from each sample have been written to separate files), so a metadata file is not initially required.

!!! note To check the input syntax for any QIIME2 command, enter the command, followed by `--help` e.g. `qiime tools import --help`

!!! attention If you haven't already done so, make sure you are running the workshop in byobu-screen and have created the symbolic links to the workshop data.

Start by making a new directory `analysis` to store all the output files from this tutorial. In addition, we will create a subdirectory called `seqs` to store the exported sequences.

```
 cd
 mkdir -p analysis/seqs
```

Run the command to import the raw data located in the directory `raw_data` and export it to a single QIIME 2 artefact file, `combined.qza`.

```
 qiime tools import \
 --type 'SampleData[PairedEndSequencesWithQuality]' \
 --input-path raw_data \
 --input-format CasavaOneEightSingleLanePerSampleDirFmt \
 --output-path analysis/seqs/combined.qza
```

## Remove primers

!!! important Remember to ask you sequencing facility if the raw data you get has the primers attached - they may have already been removed.

These sequences still have the primers attached - they need to be removed (using `cutadapt`) before denoising.

```
 qiime cutadapt trim-paired \
 --i-demultiplexed-sequences analysis/seqs/combined.qza \
 --p-front-f AGGATTAGATACCCTGGTA \
 --p-front-r CRRCACGAGCTGACGAC \
 --p-error-rate 0.20 \
 --output-dir analysis/seqs_trimmed \
 --verbose
```

!!! attention The primers specified (784f and 1492r for the bacterial 16S rRNA gene) correspond to *this* specific experiment - they will likely not work for your own data analyses.

!!! attention The error rate parameter, `#!python --p-error-rate`, will likely need to be adjusted for your own sample data to get 100% (or close to it) of reads trimmed.

## Create and interpret sequence quality data

Create a viewable summary file so the data quality can be checked. Viewing the quality plots generated here helps determine trim settings.

!!! info "**Things to look for:**" 1. Where does the median quality drop below 30?
2. Do any of the samples have only a few sequences e.g. <1000? If so, you may want to omit them from the analysis later on in R.

Create a subdirectory in `analysis` called `visualisations` to store all files that we will visualise in one place.

```
mkdir analysis/visualisations
```

```
qiime demux summarize \
--i-data analysis/seqs_trimmed/trimmed_sequences.qza \
--o-visualization analysis/visualisations/trimmed_sequences.qzv
```

Copy `analysis/visualisations/trimmed_sequences.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisations: Read quality and demux output" Click to view the `trimmed_sequences.qzv` file in QIIME 2 View.

```
<br>
Make sure to switch between the "Overview" and "Interactive Quality Plot" tabs in the top left hand co


![OverviewQualPlotTabs](./media/q2view_OverviewQualPlotTabs.png)
```

## Denoising the data

Trimmed sequences are now quality assessed using the `dada2` plugin within QIIME2. `dada2` denoises data by modelling and correcting Illumina-sequenced amplicon errors, and infers exact amplicon sequence variants (*ASVs*), resolving differences of as little as 1 nucleotide. Its workflow consists of filtering, de-replication, reference-free chimera detection, and paired-end reads merging, resulting in a feature or *ASV* table.

!!! note This step may long time to run (i.e. hours), depending on files sizes and computational power.

```
Remember to adjust `p-trunc-len-f` and `p-trunc-len-r` values according to your own data.
```

!!! question "Question: Based on your assessment of the quality plots from the trimmed_sequences.qzv file generated in the previous step, what values would you select for `p-trunc-len-f` and `p-trunc-len-r` in the command below? *Hint: At what base pair does the median quality drop below 30?*"

```
??? answer
    For version qiime2-2021.8: `p-trunc-len-f 211` and `p-trunc-len-r 167`. Other QIIME2 versions may
```

*The specified output directory must not pre-exist.*

```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs analysis/seqs_trimmed/trimmed_sequences.qza \
--p-trunc-len-f xx \
--p-trunc-len-r xx \
--p-n-threads 0 \
--output-dir analysis/dada2out \
--verbose
```

## Generate summary files

A metadata file is required which provides the key to gaining biological insight from your data. The file `metadata.tsv` is provided in the home directory of your Nectar instance. This spreadsheet has already been verified using the plugin for Google Sheets, keemei.

!!! info "**Things to look for:**" 1. How many features (*ASVs*) were generated? Are the communities high or low diversity? 2. Do BLAST searches of the representative sequences make sense? Are the features what you would expect e.g. marine or terrestrial? 3. Have a large number (e.g. >50%) of sequences been lost during denoising/filtering? If so, the settings might be too stringent.

```
qiime metadata tabulate \
--m-input-file analysis/dada2out/denoising_stats.qza \
--o-visualization analysis/visualisations/16s_denoising_stats.qzv \
--verbose
```

Copy `analysis/visualisations/16s_denoising_stats.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Denoising Stats" Click to view the **16s_denoising_stats.qzv** file in QIIME 2 View.

```
qiime feature-table summarize \
--i-table analysis/dada2out/table.qza \
--m-sample-metadata-file metadata.tsv \
--o-visualization analysis/visualisations/16s_table.qzv \
--verbose
```

Copy `analysis/visualisations/16s_table.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisations: Feature/ASV summary" Click to view the **16s_table.qzv** file in QIIME 2 View.

```
Make sure to switch between the "Overview" and "Feature Detail" tabs in the top left hand corner.
![ASV_detailPNG](./media/q2view_ASV_detail.png)
```

```
qiime feature-table tabulate-seqs \
--i-data analysis/dada2out/representative_sequences.qza \
--o-visualization analysis/visualisations/16s_representative_seqs.qzv \
--verbose
```

Copy `analysis/visualisations/16s_representative_seqs.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Representative Sequences" Click to view the **16s_representative_seqs.qzv** file in QIIME 2 View.

---

# Section 2: Taxonomic Analysis

## Assign taxonomy

Here we will classify each identical read or *Amplicon Sequence Variant (ASV)* to the highest resolution based on a database. Common databases for bacteria datasets are Greengenes, SILVA, Ribosomal Database Project, or Genome Taxonomy Database. See Porter and Hajibabaei, 2020 for a review of different classifiers for metabarcoding research. The classifier chosen is dependent upon:

1. Previously published data in a field
2. The target region of interest
3. The number of reference sequences for your organism in the database and how recently that database was updated.

A classifier has already been trained for you for the V5V6 region of the bacterial 16S rRNA gene using the SILVA database. The next step will take a while to run. *The output directory cannot previously exist*.

n_jobs = 1 This runs the script using all available cores

!!! note The classifier used here is only appropriate for the specific 16S rRNA region that *this* data represents. You will need to train your own classifier for your own data. For more information about training your own classifier, see Section 6: Extra Information.

!!! fail "STOP - Workshop participants only" Due to time limitations in a workshop setting, please do NOT run the `qiime feature-classifier classify-sklearn` command below. You will need to access a pre-computed `classification.qza` file that this command generates by running the following: `cd; mkdir analysis/taxonomy; cp /mnt/shared_data/pre_computed/classification.qza analysis/taxonomy` . If you have accidentally run the command below, `ctrl-c` will terminate it.

```
qiime feature-classifier classify-sklearn \
--i-classifier silva_138_16s_v5v6_classifier_2021-4.qza \
--i-reads analysis/dada2out/representative_sequences.qza \
--p-n-jobs 1 \
--output-dir analysis/taxonomy \
--verbose
```

!!! warning "Warning" This step often runs out of memory on full datasets. Some options are to change the number of cores you are using (adjust `--p-n-jobs` ) or add `--p-reads-per-batch 10000` and try again. The QIIME 2 forum has many threads regarding this issue so always check there was well.

## Generate a viewable summary file of the taxonomic assignments.

```
qiime metadata tabulate \
--m-input-file analysis/taxonomy/classification.qza \
--o-visualization analysis/visualisations/taxonomy.qzv \
--verbose
```

Copy `analysis/visualisations/taxonomy.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Taxonomy" Click to view the **`taxonomy.qzv`** file in QIIME 2 View.

### Filtering

Filter out reads classified as mitochondria and chloroplast. Unassigned ASVs are retained. Generate a viewable summary file of the new table to see the effect of filtering.

According to QIIME developer Nicholas Bokulich, low abundance filtering (i.e. removing ASVs containing very few sequences) is not necessary under the ASV model.

```
qiime taxa filter-table \
--i-table analysis/dada2out/table.qza \
--i-taxonomy analysis/taxonomy/classification.qza  \
--p-exclude Mitochondria,Chloroplast \
--o-filtered-table analysis/taxonomy/16s_table_filtered.qza \
--verbose
```

```
qiime feature-table summarize \
--i-table analysis/taxonomy/16s_table_filtered.qza \
--m-sample-metadata-file metadata.tsv \
--o-visualization analysis/visualisations/16s_table_filtered.qzv \
--verbose
```

Copy `analysis/visualisations/16s_table_filtered.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: 16s_table_filtered" Click to view the **`16s_table_filtered.qzv`** file in QIIME 2 View.

---

# Section 3: Build a phylogenetic tree

The next step does the following:

1. Perform an alignment on the representative sequences.
2. Mask sites in the alignment that are not phylogenetically informative.
3. Generate a phylogenetic tree.
4. Apply mid-point rooting to the tree.

A phylogenetic tree is necessary for any analyses that incorporates information on the relative relatedness of community members, by incorporating phylogenetic distances between observed organisms in the computation. This would include any beta-diversity analyses and visualisations from a weighted or unweighted Unifrac distance matrix.

```
mkdir analysis/tree
```

Use one thread only (which is the default action) so that identical results can be produced if rerun.

```
qiime phylogeny align-to-tree-mafft-fasttree \
--i-sequences analysis/dada2out/representative_sequences.qza \
--o-alignment analysis/tree/aligned_16s_representative_seqs.qza \
--o-masked-alignment analysis/tree/masked_aligned_16s_representative_seqs.qza \
--o-tree analysis/tree/16s_unrooted_tree.qza \
--o-rooted-tree analysis/tree/16s_rooted_tree.qza \
--p-n-threads 1 \
--verbose
```

---

# Section 4: Basic visualisations and statistics

## ASV relative abundance bar charts

Create bar charts to compare the relative abundance of ASVs across samples.

```
qiime taxa barplot \
--i-table analysis/taxonomy/16s_table_filtered.qza \
--i-taxonomy analysis/taxonomy/classification.qza \
--m-metadata-file metadata.tsv \
--o-visualization analysis/visualisations/barchart.qzv \
--verbose
```

Copy `analysis/visualisations/barchart.qzv` to your local computer and view in QIIME 2 View (q2view). Try selecting different taxonomic levels and metadata-based sample sorting.

??? example "Visualisations: Taxonomy Barplots" Click to view the **barchart.qzv** file in QIIME 2 View.

```
Increase the "Bar Width", select Index in "Sort Samples By" drop-down menu and explore the resulting b

![barplot1](./media/q2view_barplot_levels.png)
```

## Rarefaction curves

Generate rarefaction curves to determine whether the samples have been sequenced deeply enough to capture all the community members. The max depth setting will depend on the number of sequences in your samples.

!!! info "**Things to look for:**" 1. Do the curves for each sample plateau? If they don't, the samples haven't been sequenced deeply enough to capture the full diversity of the bacterial communities, which is shown on the y-axis. 2. At what sequencing depth (x-axis) do your curves plateau? This value will be important for downstream analyses, particularly for alpha diversity analyses.

!!! note The value that you provide for --p-max-depth should be determined by reviewing the "Frequency per sample" information presented in the 16s_table_filtered.qzv file that was created above. In general, choosing a value that is somewhere around the median frequency seems to work well, but you may want to increase that value if the lines in the resulting rarefaction plot don't appear to be leveling out, or decrease that value if you seem to be losing many of your samples due to low total frequencies closer to the minimum sampling depth than the maximum sampling depth.

```
qiime diversity alpha-rarefaction \
--i-table analysis/taxonomy/16s_table_filtered.qza \
--i-phylogeny analysis/tree/16s_rooted_tree.qza \
--p-max-depth 9062 \
--m-metadata-file metadata.tsv \
--o-visualization analysis/visualisations/16s_alpha_rarefaction.qzv \
--verbose
```

Copy `analysis/visualisations/16s_alpha_rarefaction.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Rarefaction" Click to view the **16s_alpha_rarefaction.qzv** file in QIIME 2 View.

```
Select "Genotype" in the "Sample Metadata Column":

![rarefaction](./media/q2view_rarefaction.png)
```

## Alpha and beta diversity analysis

The following is taken directly from the Moving Pictures tutorial and adapted for this data set. QIIME 2's diversity analyses are available through the `q2-diversity` plugin, which supports computing alpha- and beta- diversity metrics, applying related statistical tests, and generating interactive visualisations. We'll first apply the core-metrics-phylogenetic method, which rarefies a FeatureTable[Frequency] to a user-specified depth, computes several alpha- and beta- diversity metrics, and generates principle

coordinates analysis (PCoA) plots using Emperor for each of the beta diversity metrics.

The metrics computed by default are:

- Alpha diversity (operate on a single sample (i.e. within sample diversity)).
    - Shannon's diversity index (a quantitative measure of community richness)
    - Observed OTUs (a qualitative measure of community richness)
    - Faith's Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features)
    - Evenness (or Pielou's Evenness; a measure of community evenness)
- Beta diversity (operate on a pair of samples (i.e. between sample diversity)).
    - Jaccard distance (a qualitative measure of community dissimilarity)
    - Bray-Curtis distance (a quantitative measure of community dissimilarity)
    - unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
    - weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)

An important parameter that needs to be provided to this script is `--p-sampling-depth`, which is the even sampling (i.e. rarefaction) depth that was determined above. As most diversity metrics are sensitive to different sampling depths across different samples, this script will randomly subsample the counts from each sample to the value provided for this parameter. For example, if `--p-sampling-depth 500` is provided, this step will subsample the counts in each sample without replacement, so that each sample in the resulting table has a total count of 500. If the total count for any sample(s) are smaller than this value, those samples will be excluded from the diversity analysis. Choosing this value is tricky. We recommend making your choice by reviewing the information presented in the 16s_table_filtered.qzv file that was created above. Choose a value that is as high as possible (so more sequences per sample are retained), while excluding as few samples as possible.

```
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny analysis/tree/16s_rooted_tree.qza \
  --i-table analysis/taxonomy/16s_table_filtered.qza \
  --p-sampling-depth 5583 \
  --m-metadata-file metadata.tsv \
  --output-dir analysis/diversity_metrics
```

Copy the `.qzv` files created from the above command into the `visualisations` subdirectory.

```
cp analysis/diversity_metrics/*.qzv analysis/visualisations
```

To view the differences between sample composition using unweighted UniFrac in ordination space, copy `analysis/visualisations/unweighted_unifrac_emperor.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisations: Unweighted UniFrac Emperor Ordination" Click to view the **unweighted_unifrac_emperor.qzv** file in QIIME 2 View.

```
On q2view, select the "Color" tab, choose "Environment" under the "Select a Color Category" dropdown m

![unweighted_unifrac_emperor2](./media/q2view_unweighted_unifrac_emperor2.png)
```

Next, we'll test for associations between categorical metadata columns and alpha diversity data. We'll do that here for the Faith Phylogenetic Diversity (a measure of community richness) and evenness metrics.

```
qiime diversity alpha-group-significance \
  --i-alpha-diversity analysis/diversity_metrics/faith_pd_vector.qza \
  --m-metadata-file metadata.tsv \
  --o-visualization analysis/visualisations/faith-pd-group-significance.qzv
```

Copy `analysis/visualisations/faith-pd-group-significance.qzv` to your local computer and view in QIIME 2 View

(q2view).

??? example "Visualisation: Faith Phylogenetic Diversity output" Click to view the `faith-pd-group-significance.qzv` file in QIIME 2 View.

```
Select "Environment" under the "Column" dropdown menu.

![faith](./media/q2view_faith.png)
```

```
qiime diversity alpha-group-significance \
  --i-alpha-diversity analysis/diversity_metrics/evenness_vector.qza \
  --m-metadata-file metadata.tsv \
  --o-visualization analysis/visualisations/evenness-group-significance.qzv
```

Copy `analysis/visualisations/evenness-group-significance.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Evenness output" Click to view the `evenness-group-significance.qzv` file in QIIME 2 View.

```
Select "Genotype" under the "Column" dropdown menu.

![evenness](./media/q2view_evenness.png)
```

Finally, we'll analyse sample composition in the context of categorical metadata using a permutational multivariate analysis of variance (PERMANOVA, first described in Anderson (2001)) test using the beta-group-significance command. The following commands will test whether distances between samples within a group, such as samples from the same genotype, are more similar to each other then they are to samples from the other groups. If you call this command with the `--p-pairwise` parameter, as we'll do here, it will also perform pairwise tests that will allow you to determine which specific pairs of groups (e.g., AIMS1 and AIMS4) differ from one another, if any. This command can be slow to run, especially when passing `--p-pairwise`, since it is based on permutation tests. So, unlike the previous commands, we'll run beta-group-significance on specific columns of metadata that we're interested in exploring, rather than all metadata columns to which it is applicable. Here we'll apply this to our unweighted UniFrac distances, using two sample metadata columns, as follows.

```
qiime diversity beta-group-significance \
  --i-distance-matrix analysis/diversity_metrics/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata.tsv \
  --m-metadata-column Genotype \
  --o-visualization analysis/visualisations/unweighted-unifrac-genotype-significance.qzv \
  --p-pairwise
```

Copy `analysis/visualisations/unweighted-unifrac-genotype-significance.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Genotype significance output" Click to view the `unweighted-unifrac-genotype-significance.qzv` file in QIIME 2 View.

```
qiime diversity beta-group-significance \
  --i-distance-matrix analysis/diversity_metrics/unweighted_unifrac_distance_matrix.qza \
  --m-metadata-file metadata.tsv \
  --m-metadata-column Environment \
  --o-visualization analysis/visualisations/unweighted-unifrac-environment-significance.qzv \
  --p-pairwise
```

Copy `analysis/visualisations/unweighted-unifrac-environment-significance.qzv` to your local computer and view in QIIME 2 View (q2view).

??? example "Visualisation: Environmental significance output" Click to view the **unweighted-unifrac-environment-significance.qzv** file in QIIME 2 View.

??? example "Provenance" Click to view the **unweighted-unifrac-environment-significance.qzv** provenance file in QIIME 2 View.

---

# Section 5: Exporting data for further analysis in R

You need to export your ASV table, taxonomy table, and tree file for analyses in R. Many file formats can be accepted.

Export unrooted tree as `.nwk` format as required for the R package `phyloseq`.

```
qiime tools export \
  --input-path analysis/tree/16s_unrooted_tree.qza \
  --output-path analysis/export
```

Create a BIOM table with taxonomy annotations. A FeatureTable[Frequency] artefact will be exported as a BIOM v2.1.0 formatted file.

```
qiime tools export \
  --input-path analysis/taxonomy/16s_table_filtered.qza \
  --output-path analysis/export
```

Then export BIOM to TSV

```
biom convert \
-i analysis/export/feature-table.biom \
-o analysis/export/feature-table.tsv \
--to-tsv
```

Export Taxonomy as TSV

```
qiime tools export \
--input-path analysis/taxonomy/classification.qza \
--output-path analysis/export
```

Delete the header lines of the .tsv files

```
sed '1d' analysis/export/taxonomy.tsv > analysis/export/taxonomy_noHeader.tsv
sed '1d' analysis/export/feature-table.tsv > analysis/export/feature-table_noHeader.tsv
```

Some packages require your data to be in a consistent order, i.e. the order of your ASVs in the taxonomy table rows to be the same order of ASVs in the columns of your ASV table. It's recommended to clean up your taxonomy file. You can have blank spots where the level of classification was not completely resolved.

# Section 6: Extra Information

!!! fail "STOP" This section contains information on how to train the classifier for analysing your **own** data. This will **NOT** be covered in the workshop.

### Train SILVA v138 classifier for 16S/18S rRNA gene marker sequences.

The newest version of the SILVA database (v138) can be trained to classify marker gene sequences originating from the 16S/18S rRNA gene. Reference files `silva-138-99-seqs.qza` and `silva-138-99-tax.qza` were downloaded from SILVA and imported to get the artefact files. You can download both these files from here.

Reads for the region of interest are first extracted. **You will need to input your forward and reverse primer sequences**. See QIIME2 documentation for more information.

```
qiime feature-classifier extract-reads \
--i-sequences silva-138-99-seqs.qza \
--p-f-primer FORWARD_PRIMER_SEQUENCE \
--p-r-primer REVERSE_PRIMER_SEQUENCE \
--o-reads silva_138_marker_gene.qza \
--verbose
```

The classifier is then trained using a naive Bayes algorithm. See QIIME2 documentation for more information.

```
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads silva_138_marker_gene.qza \
--i-reference-taxonomy silva-138-99-tax.qza \
--o-classifier silva_138_marker_gene_classifier.qza \
--verbose
```