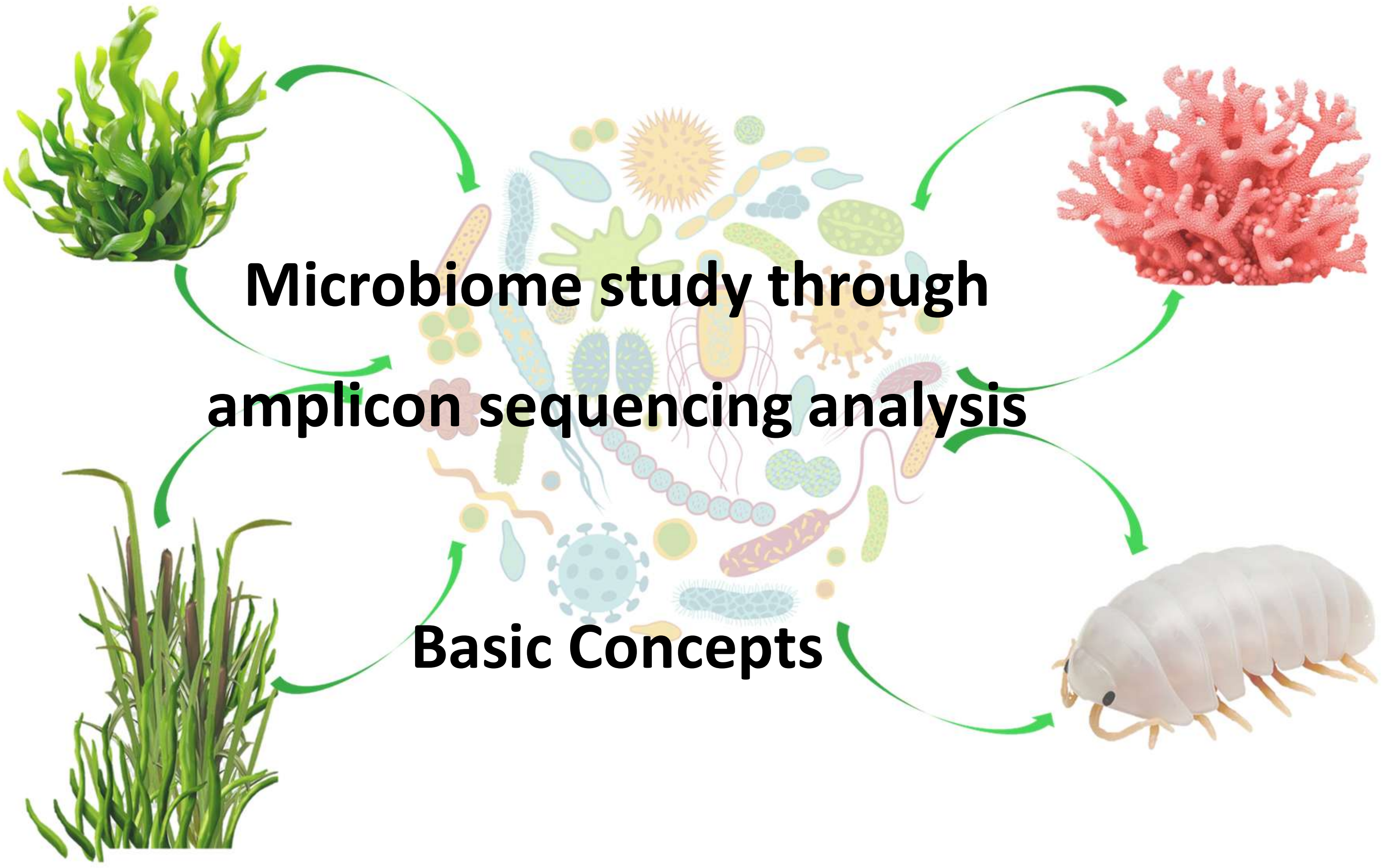


Microbiome study through amplicon sequencing analysis

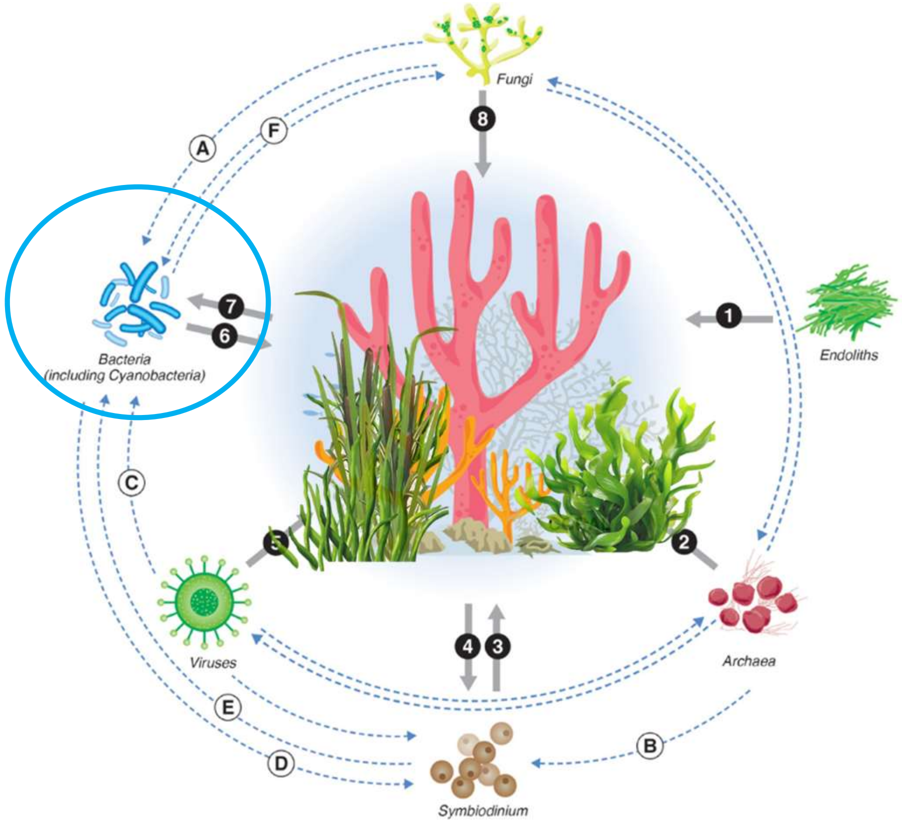
Basic Concepts

Tania Aires

CCMAR
Centro de Ciências do Mar
Algarve

MMEB
Marine Microbial Ecology & Biotechnology

Holobiont



- Settlement /reproduction
- Morphogenesis
- Nutrition
- Nutrient cycling
- Protection from abiotic factors
- Protection from grazers
- Defence against pathogens

Why it is important to study the microbiome of marine organisms?

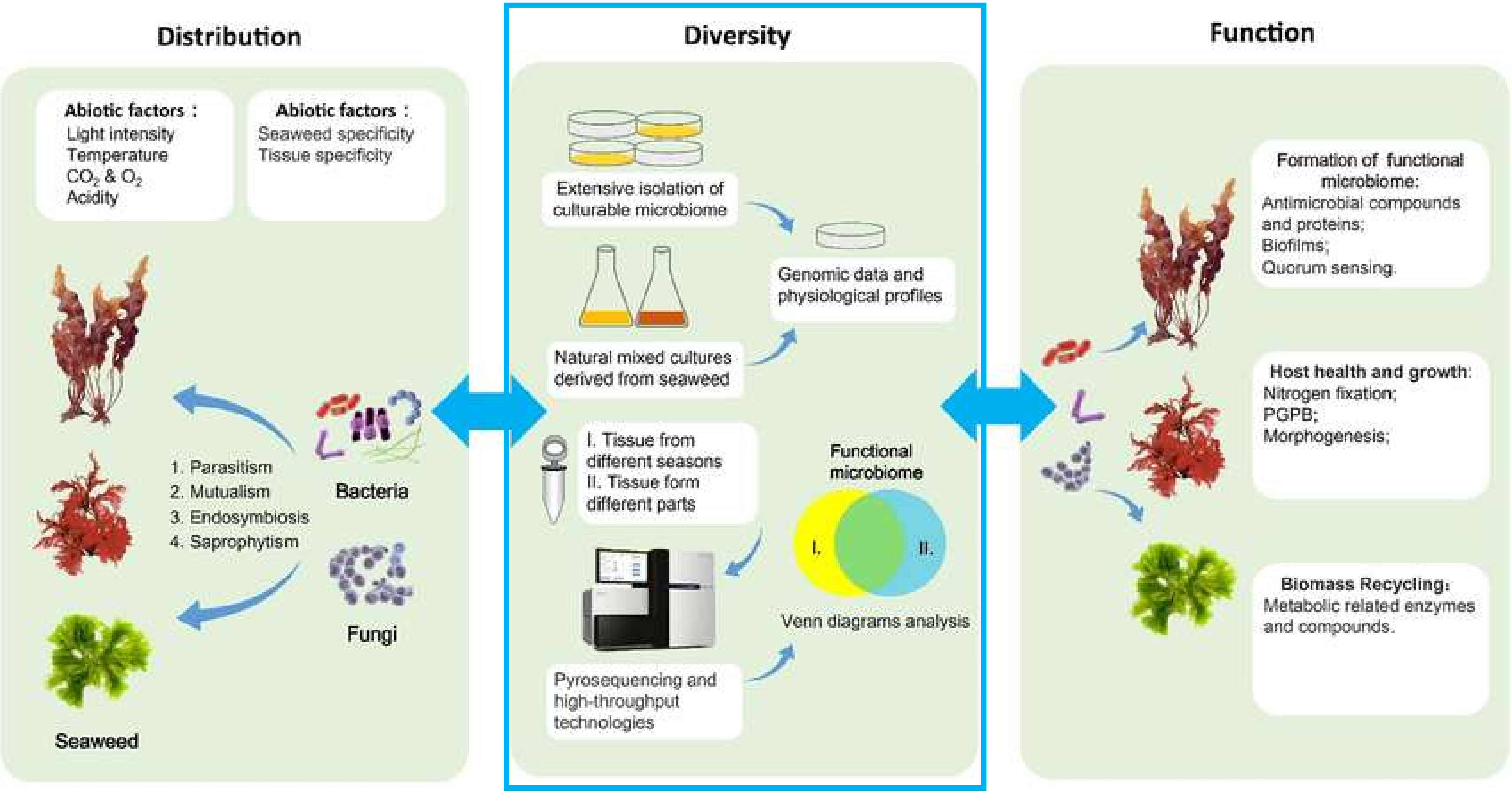
Seaweed aquaculture industry

- Increase stress tolerance
- Disease/pest control
- Improved growth and yield
- Better nutrition
- Product optimization

Seagrass meadows/coral reef restauration

- Increase stress tolerance
- Disease resistance
- Seedling survival
- Increased nutrient acquisition

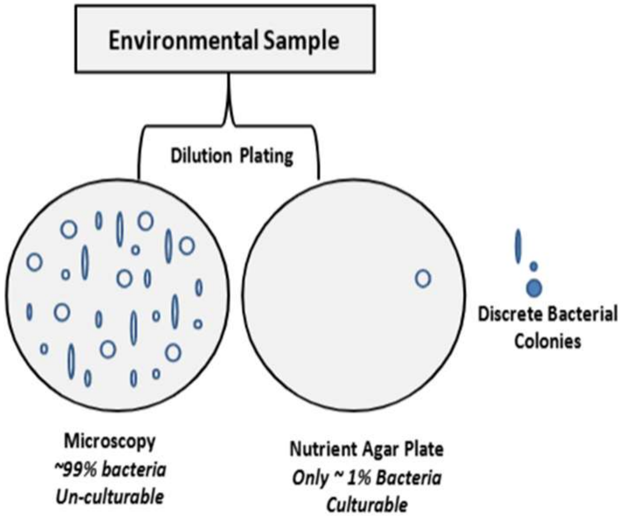
How do we know these microorganisms?



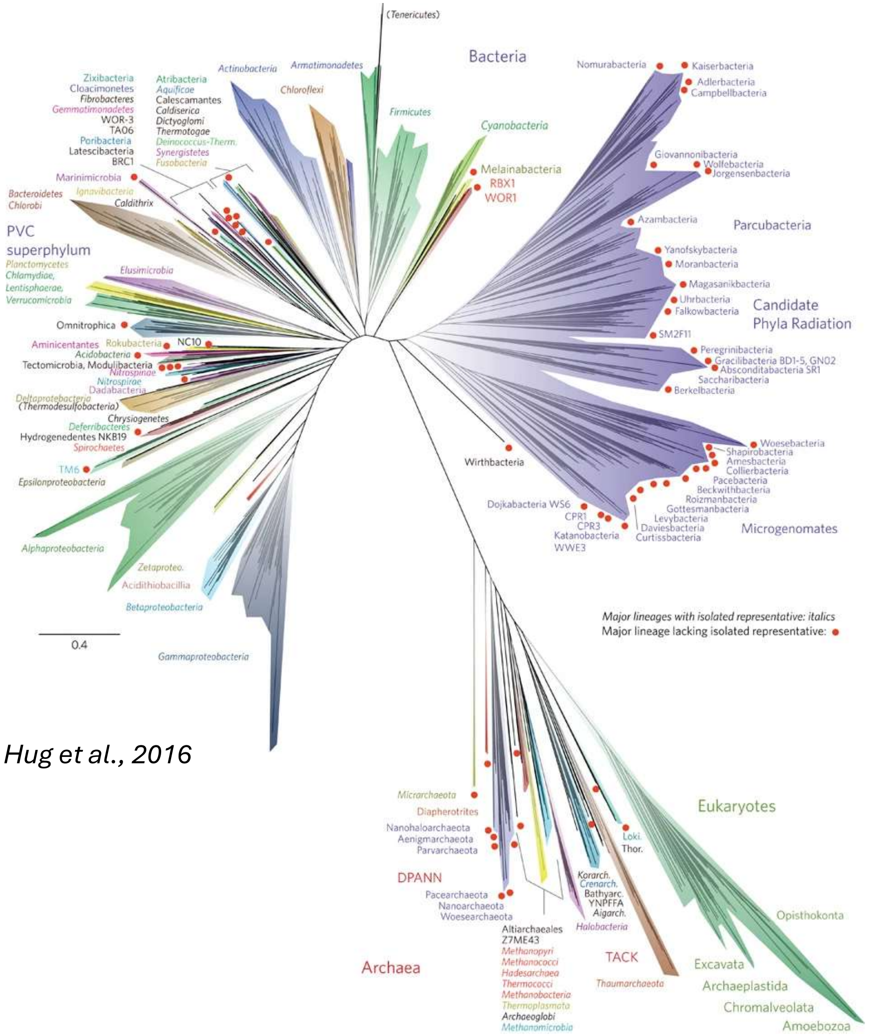
Source: https://enviromicro-journals.onlinelibrary.wiley.com/doi/10.1111/1751-7915.14014

How to study microbiome diversity?

“The Great Plate Count Anomaly”



Staley and Konopka, 1985



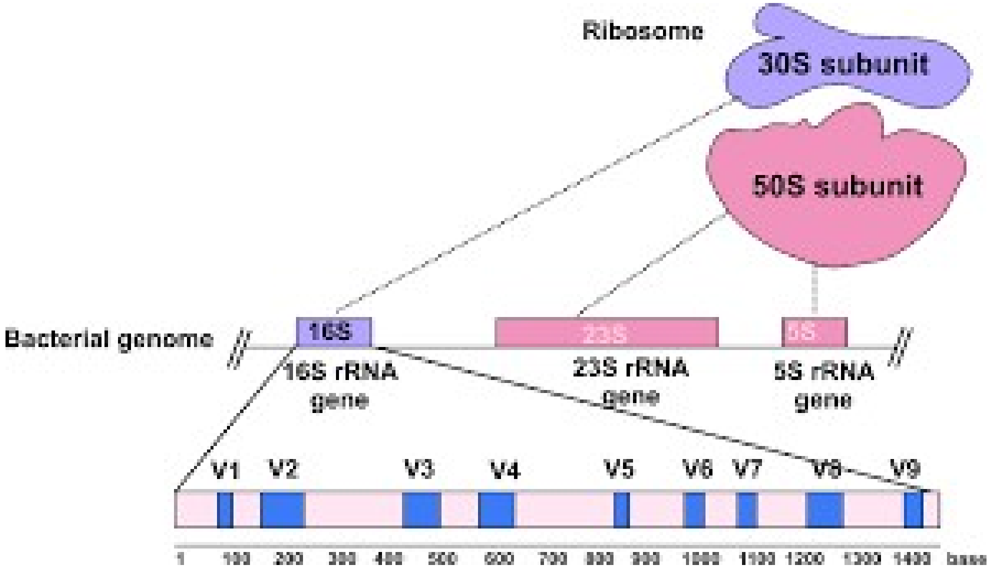
Hug et al., 2016

Problem

1-10% of the microorganisms are culturable

Solution

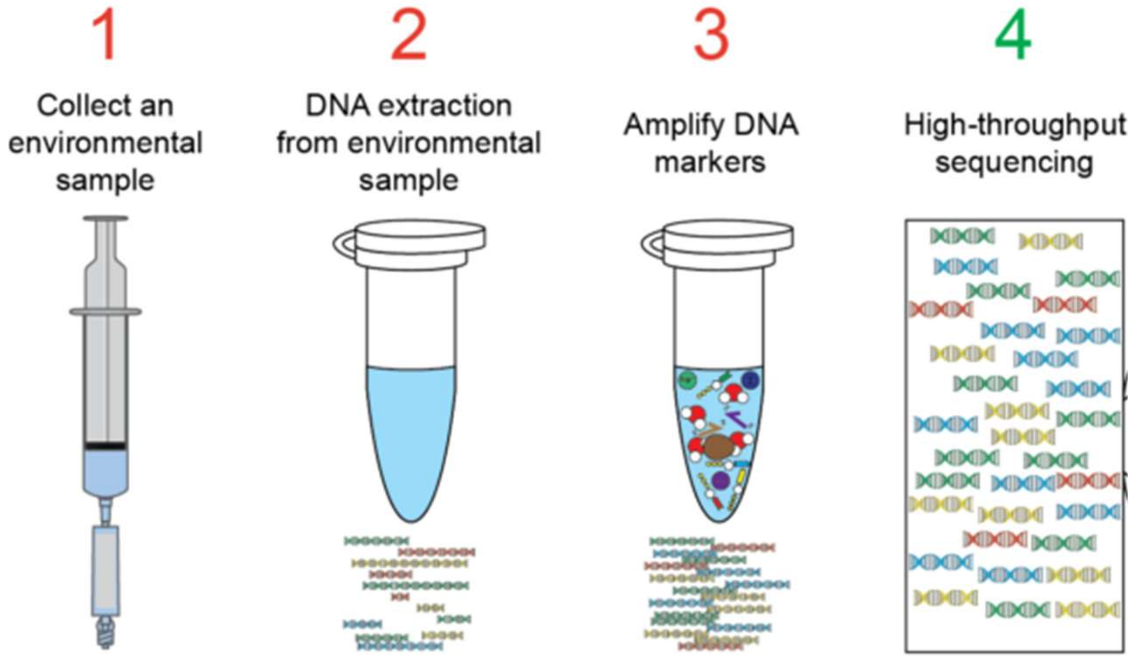
Metabarcoding: Large-scale taxonomic identification of complex environmental samples via analysis of DNA sequences for short regions of one or a few genes.

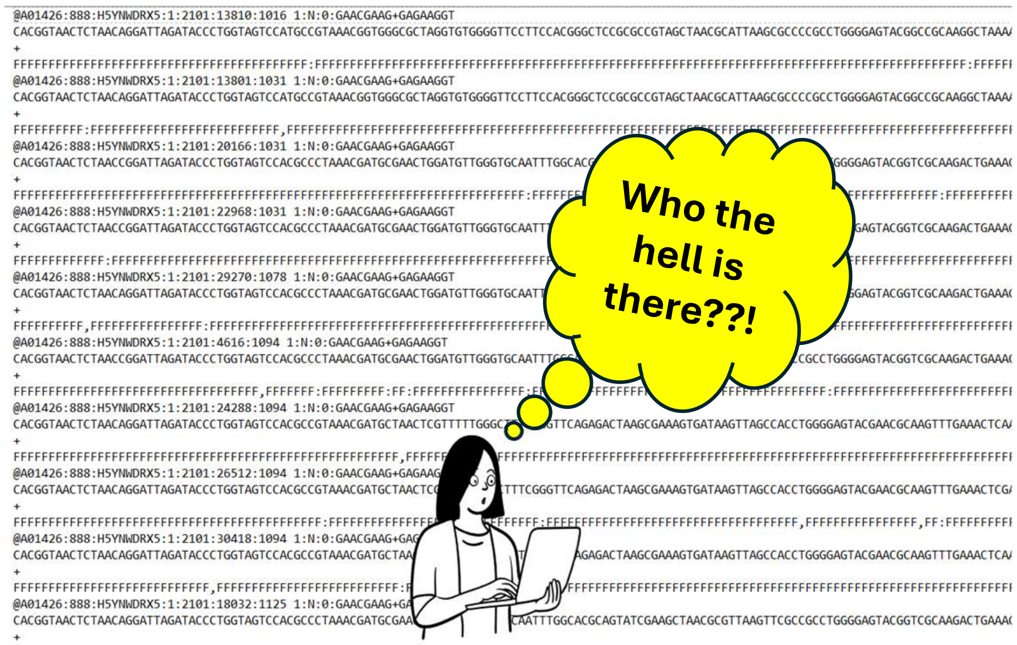


- **Illumina**
- **Nanopore**
- **PacBio**
- **Ion Torrent**

16S rRNA gene is the most common universal **DNA barcode** (marker) used to identify with great accuracy bacterial species from across the Tree of Life

Metabarcoding Workflow


1
Collect an environmental sample
2
DNA extraction from environmental sample
3
Amplify DNA markers
4
High-throughput sequencing

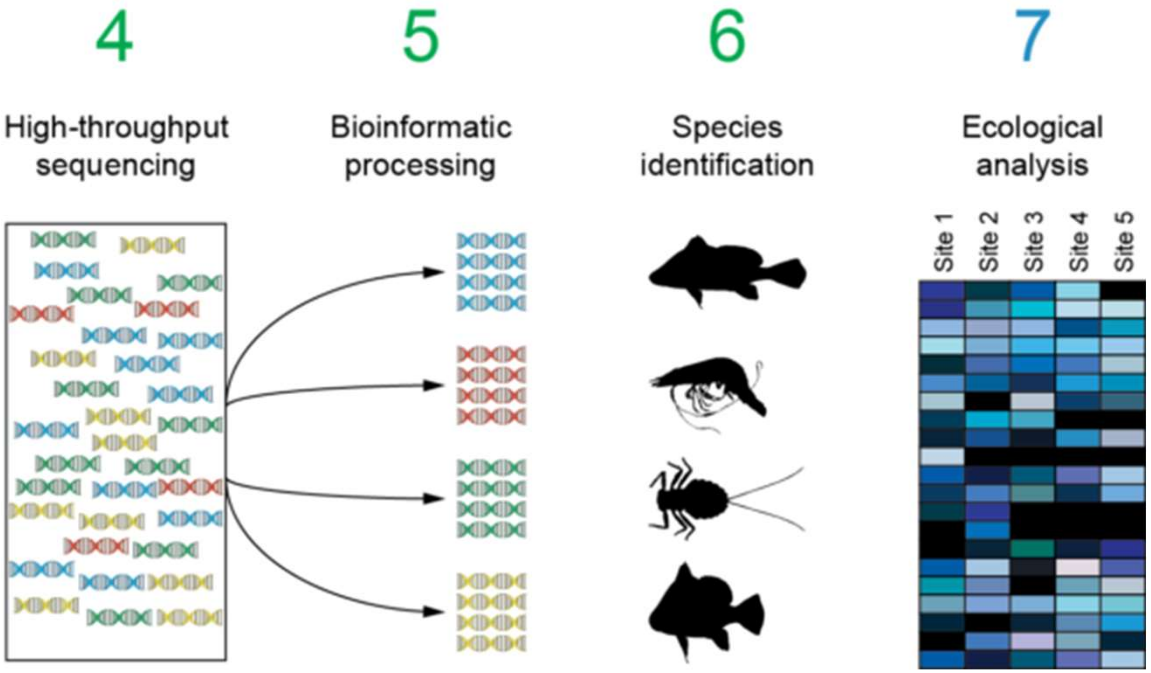


```
@A01426:888:H5YNWDRX5:1:2101:13810:1016 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCATGCCGTAAACGGTGGGCGCTAGGTGTGGGGTTCCTTCCACGGGCTCCGCGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF
@A01426:888:H5YNWDRX5:1:2101:13801:1031 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCATGCCGTAAACGGTGGGCGCTAGGTGTGGGGTTCCTTCCACGGGCTCCGCGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAA
+
FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:20166:1031 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACCGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGATGTTGGGTGCAATTTGGCACG[illegible]GGGGAGTACGGTCGCAAGACTGAAA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFF[illegible]FFFFFFFFFFFFF:FFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:22968:1031 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACCGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGATGTTGGGTGCAATT[illegible]GAGTACGGTCGCAAGACTGAAA
+
FFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF[illegible]FFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:29270:1078 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGATGTTGGGTGCAATT[illegible]GGAGTACGGTCGCAAGACTGAAA
+
FFFFFFFFFF,FFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF[illegible]FFFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:4616:1094 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACCGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAACTGGATGTTGGGTGCAATTTG[illegible]CCGCCTGGGGAGTACGGTCGCAAGACTGAAA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFF:FFFFFFFFF:FF:FFFFFFFFFFFFFFF:F[illegible]FFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:24288:1094 1:N:0:GAACGAAG+GAGAAGGT
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGCTAACTCGTTTTTGGGC[illegible]GTTCAGAGACTAAGCGAAAGTGATAAGTTAGCCACCTGGGGAGTACGAACGCAAGTTTGAAACTCA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFF[illegible]FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:26512:1094 1:N:0:GAACGAAG+GAGAAG[illegible]
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGCTAACTC[illegible]TTTCGGGTTCAGAGACTAAGCGAAAGTGATAAGTTAGCCACCTGGGGAGTACGAACGCAAGTTTGAAACTCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFF[illegible]FFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFF,FF:FFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:30418:1094 1:N:0:GAACGAAG+GAG[illegible]
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGCTAA[illegible]AGAGACTAAGCGAAAGTGATAAGTTAGCCACCTGGGGAGTACGAACGCAAGTTTGAAACTCA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFF:F[illegible]FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A01426:888:H5YNWDRX5:1:2101:18032:1125 1:N:0:GAACGAAG+GA[illegible]
CACGGTAACTCTAACAGGATTAGATACCCTGGTAGTCCACGCCCTAAACGATGCGAA[illegible]CAATTTGGCACGCAGTATCGAAGCTAACGCGTTAAGTTCGCCGCCTGGGGAGTACGGTCGCAAGACTGAAA
+
```

Source: http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Metabarcoding Workflow

(...)



Source: http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Quantitative Insights Into Microbial Ecology - Go to: https://qiime2.org/

Python program, open-source, continuous community development. It's a bioinformatics and data science platform particularly for microbiome multi-omics analysis, built upon a framework that enables reproducible biological data science. **It works through plugins** – developers create 3rd party plugins for Qiime2 as needed - https://amplicon-docs.qiime2.org/en/latest/references/available-plugins.html

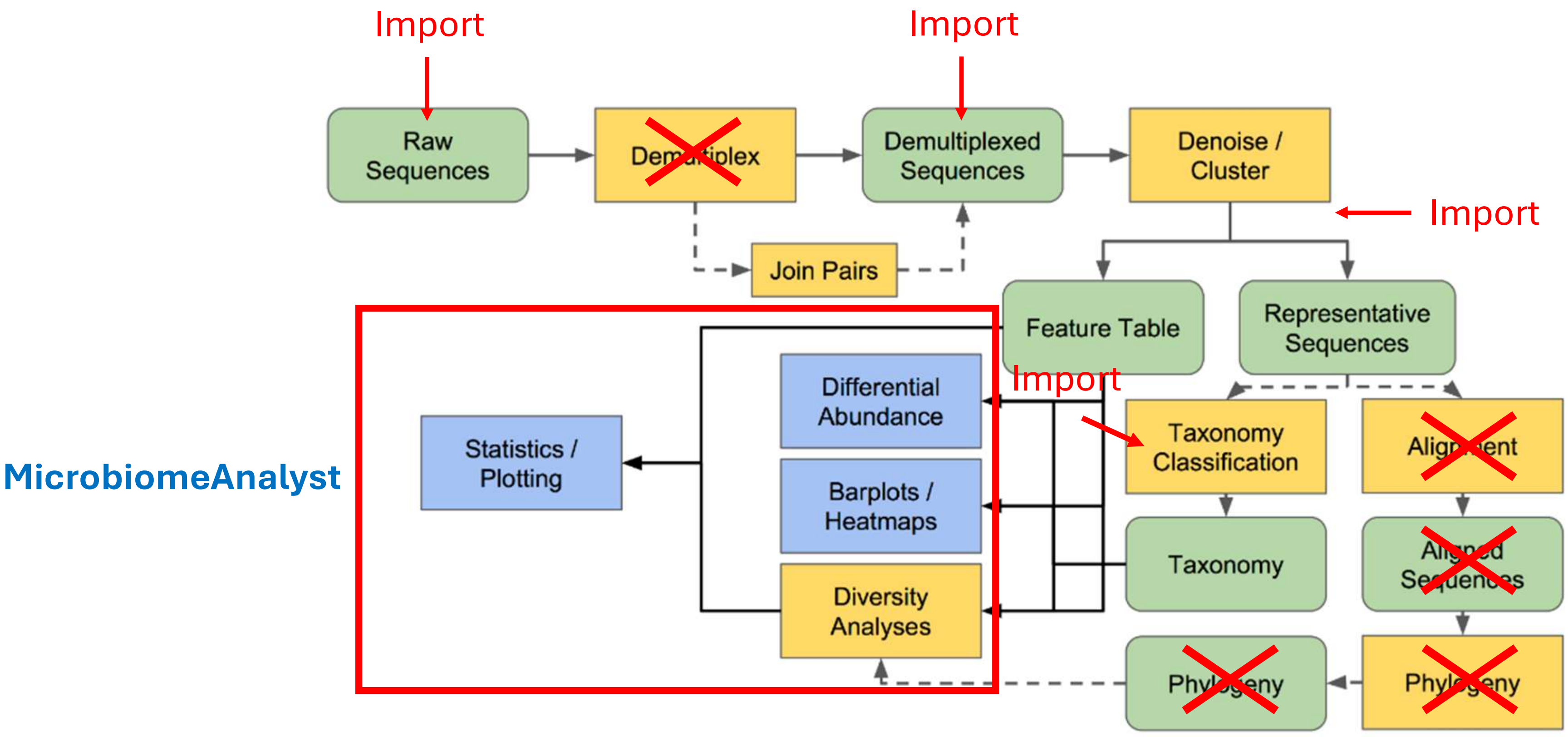
1. **Perform quality control and produce an ASV abundance table from raw amplicon sequences**
2. **Perform taxonomic classification using the SILVA 16S database**

Generated tables can be used for downstream analysis like community profiling or diversity analysis – **MicrobiomeAnalyst**

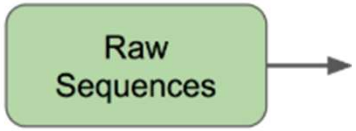
We will focus on the bacteria but this can be applied to other taxonomic groups such as animals, plants, fungi, using other markers and databases

Metabarcoding analysis using qiime2

Conceptual overview



Data preparation



IMPORT your data

- Import in a way that Qiime2 can read it

Output: demux.qza file

- Contains all our sequences
- Contains quality scores
- Qiime can read it

Import Data

https://docs.qiime2.org/2024.10/tutorials/importing/

The format you receive your data depends on the Sequencing Company you work with

With QIIME 2, there are different functions to import different types of FASTQ data:

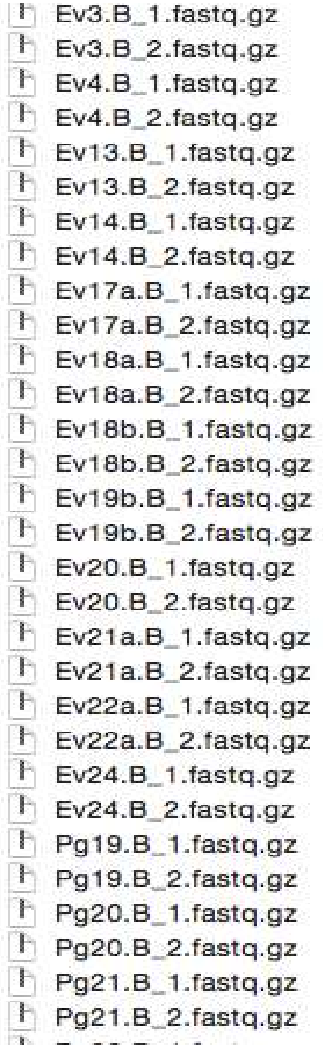
1.FASTQ data with the **EMP Protocol format** – **Multiplexed** Single-end or Paired-end reads – We receive two different files: Sequence file + Barcode file

2.FASTQ data with barcodes in sequences – **Multiplexed** Single-end or Paired-end reads - Sequence file+Metadata file – **Use a different program for demultiplexing**

3.FASTQ data in the **Casava 1.8 demultiplexed format** – **Demultiplexed** Single-end or Paired-end reads - The file name includes the sample identifier and should look like **L2S357_15_L001_R1_001.fastq.gz**

4.Any **demultiplexed** FASTQ data not represented in the list items above – None of the above formats – Use a **Manifest file**

Import Data

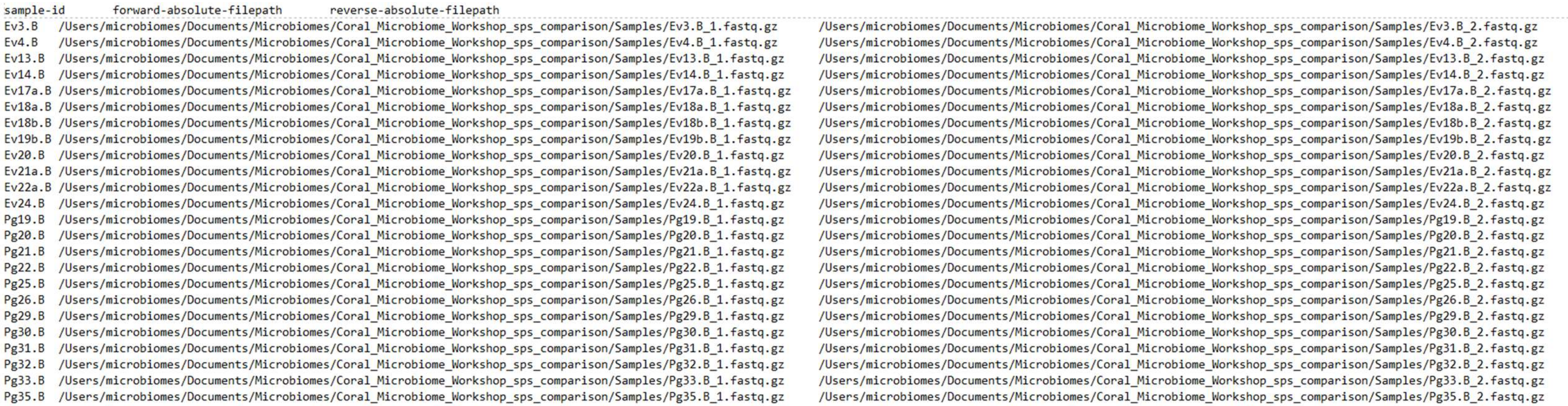


```
Ev3.B_1.fastq.gz
Ev3.B_2.fastq.gz
Ev4.B_1.fastq.gz
Ev4.B_2.fastq.gz
Ev13.B_1.fastq.gz
Ev13.B_2.fastq.gz
Ev14.B_1.fastq.gz
Ev14.B_2.fastq.gz
Ev17a.B_1.fastq.gz
Ev17a.B_2.fastq.gz
Ev18a.B_1.fastq.gz
Ev18a.B_2.fastq.gz
Ev18b.B_1.fastq.gz
Ev18b.B_2.fastq.gz
Ev19b.B_1.fastq.gz
Ev19b.B_2.fastq.gz
Ev20.B_1.fastq.gz
Ev20.B_2.fastq.gz
Ev21a.B_1.fastq.gz
Ev21a.B_2.fastq.gz
Ev22a.B_1.fastq.gz
Ev22a.B_2.fastq.gz
Ev24.B_1.fastq.gz
Ev24.B_2.fastq.gz
Pg19.B_1.fastq.gz
Pg19.B_2.fastq.gz
Pg20.B_1.fastq.gz
Pg20.B_2.fastq.gz
Pg21.B_1.fastq.gz
Pg21.B_2.fastq.gz
```

- Demultiplexed
- Primers and Barcodes already removed
- Paired-end sequences

Import Data

Manifest File



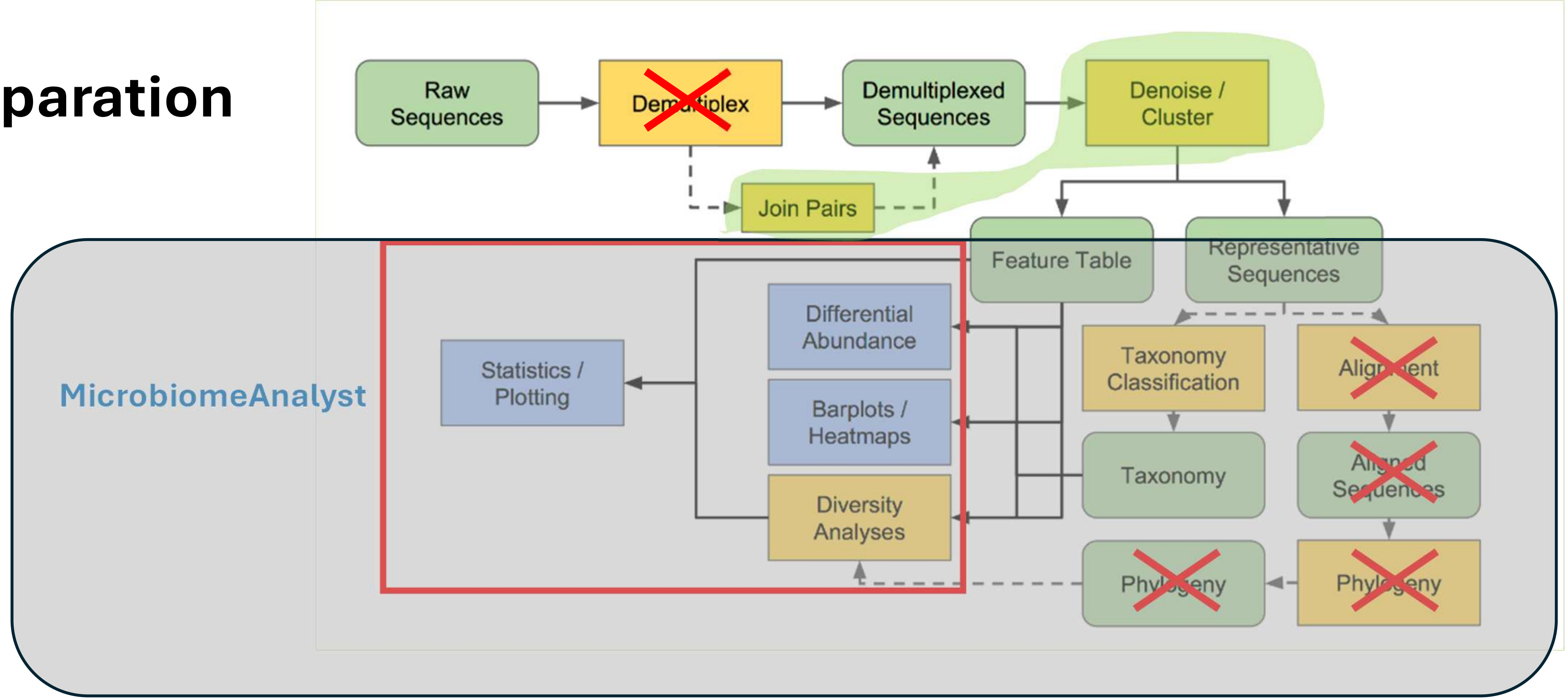
```
sample-id       forward-absolute-filepath       reverse-absolute-filepath
Ev3.B   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev3.B_1.fastq.gz     /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev3.B_2.fastq.gz
Ev4.B   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev4.B_1.fastq.gz     /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev4.B_2.fastq.gz
Ev13.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev13.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev13.B_2.fastq.gz
Ev14.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev14.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev14.B_2.fastq.gz
Ev17a.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev17a.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev17a.B_2.fastq.gz
Ev18a.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev18a.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev18a.B_2.fastq.gz
Ev18b.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev18b.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev18b.B_2.fastq.gz
Ev19b.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev19b.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev19b.B_2.fastq.gz
Ev20.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev20.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev20.B_2.fastq.gz
Ev21a.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev21a.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev21a.B_2.fastq.gz
Ev22a.B /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev22a.B_1.fastq.gz   /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev22a.B_2.fastq.gz
Ev24.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev24.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Ev24.B_2.fastq.gz
Pg19.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg19.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg19.B_2.fastq.gz
Pg20.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg20.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg20.B_2.fastq.gz
Pg21.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg21.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg21.B_2.fastq.gz
Pg22.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg22.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg22.B_2.fastq.gz
Pg25.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg25.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg25.B_2.fastq.gz
Pg26.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg26.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg26.B_2.fastq.gz
Pg29.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg29.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg29.B_2.fastq.gz
Pg30.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg30.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg30.B_2.fastq.gz
Pg31.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg31.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg31.B_2.fastq.gz
Pg32.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg32.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg32.B_2.fastq.gz
Pg33.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg33.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg33.B_2.fastq.gz
Pg35.B  /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg35.B_1.fastq.gz    /Users/microbiomes/Documents/Microbiomes/Coral_Microbiome_Workshop_sps_comparison/Samples/Pg35.B_2.fastq.gz
```

Contains

- Sample ID (the exact name of the sample)
- Absolute filepath for each one of your forward and reverse reads

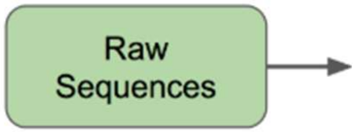
- Is a text file that function as a coordinates file for the program to know where to find your sequences in your computer

Data preparation



qiime2

Data preparation



IMPORT your data

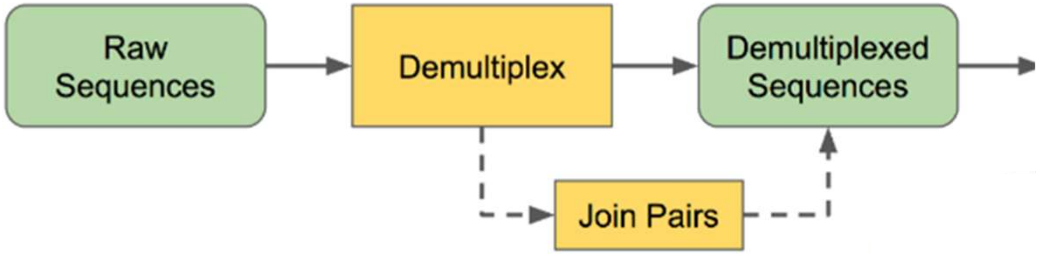
- Import in a way that Qiime2 can read it

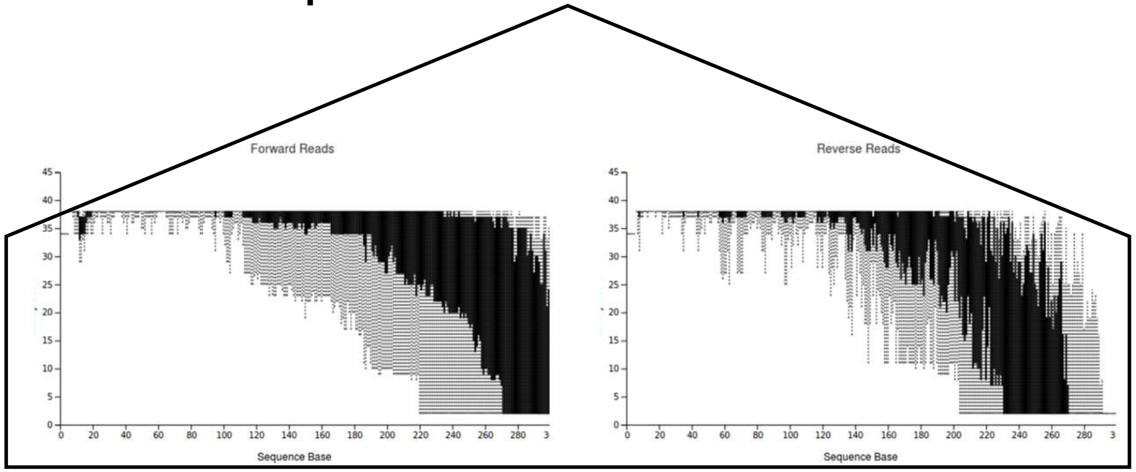
Output: demux.qza file

- Contains all our sequences
- Contains quality scores
- Qiime can read it

Data preparation

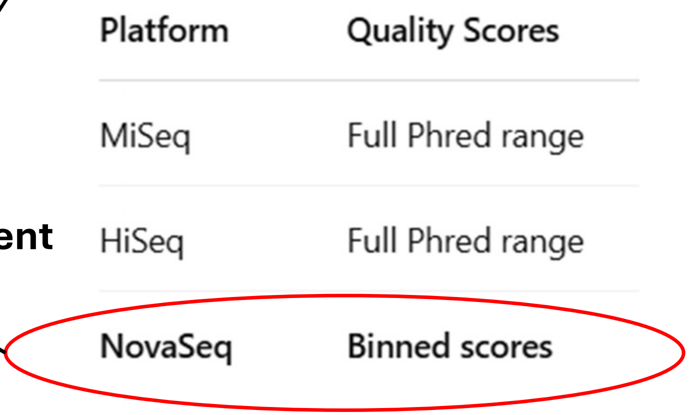
demux.qza file → demux.qzv file → https://view.qiime2.org/





- Import our sequences in a way that Qiime2 can read it ✓
- Demultiplex ✓
- Join read pairs – DADA2
- Quality filter/trim – DADA2

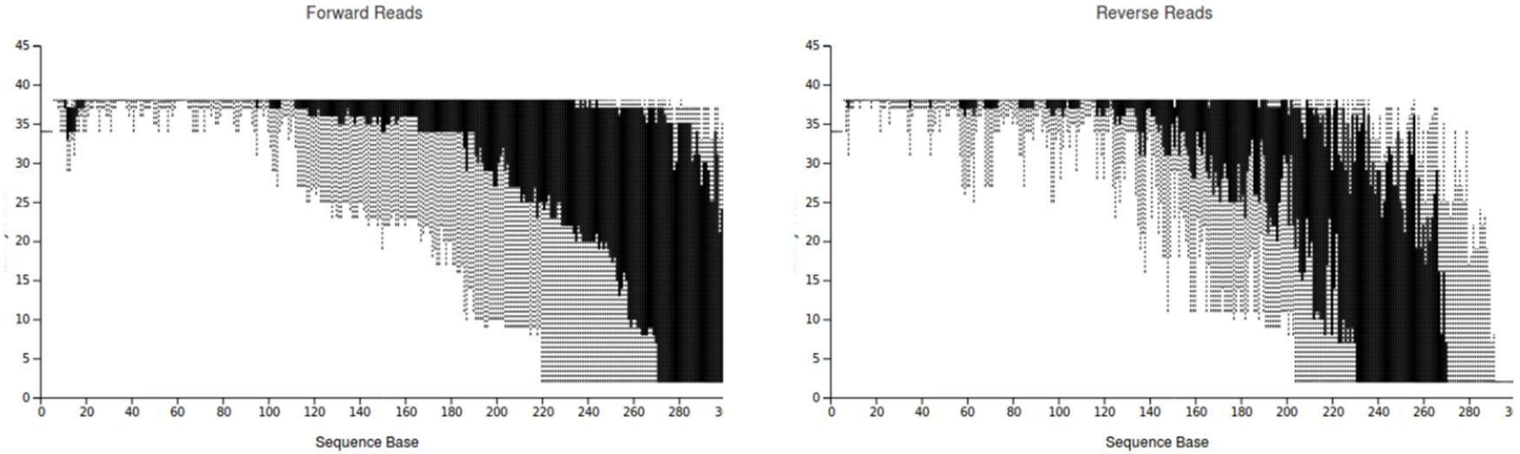
Our graph will be slightly different



Platform	Quality Scores
MiSeq	Full Phred range
HiSeq	Full Phred range
NovaSeq	Binned scores

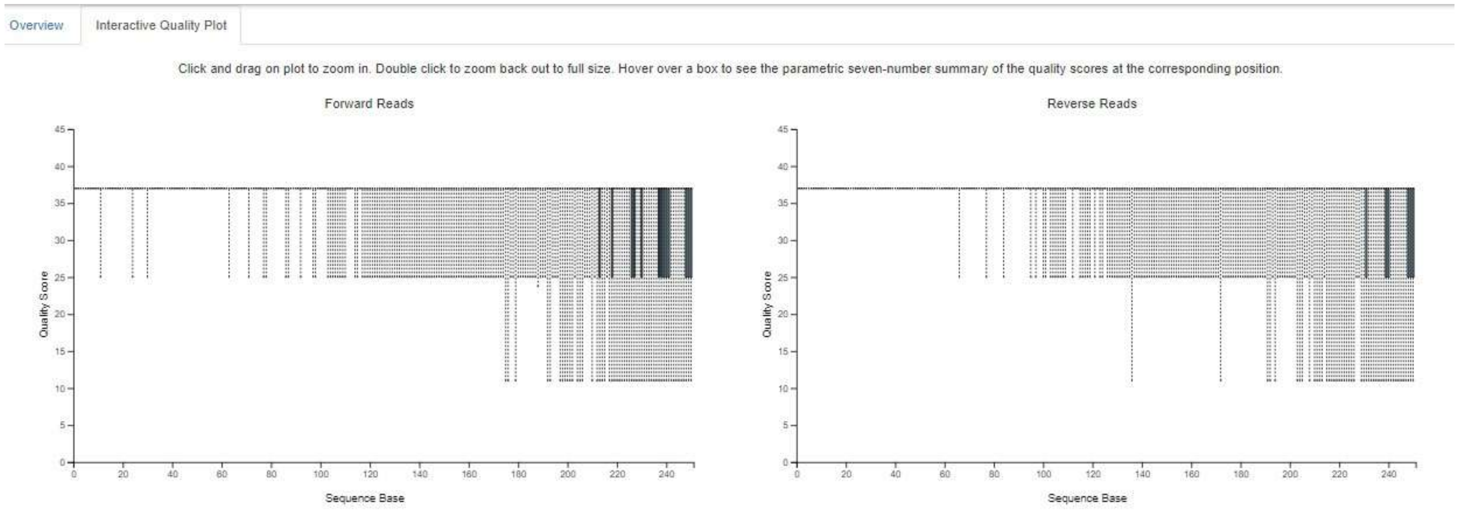
Illumina MiSeq and HiSeq – Full range Phred scores

Illumina MiSeq/HiSeq

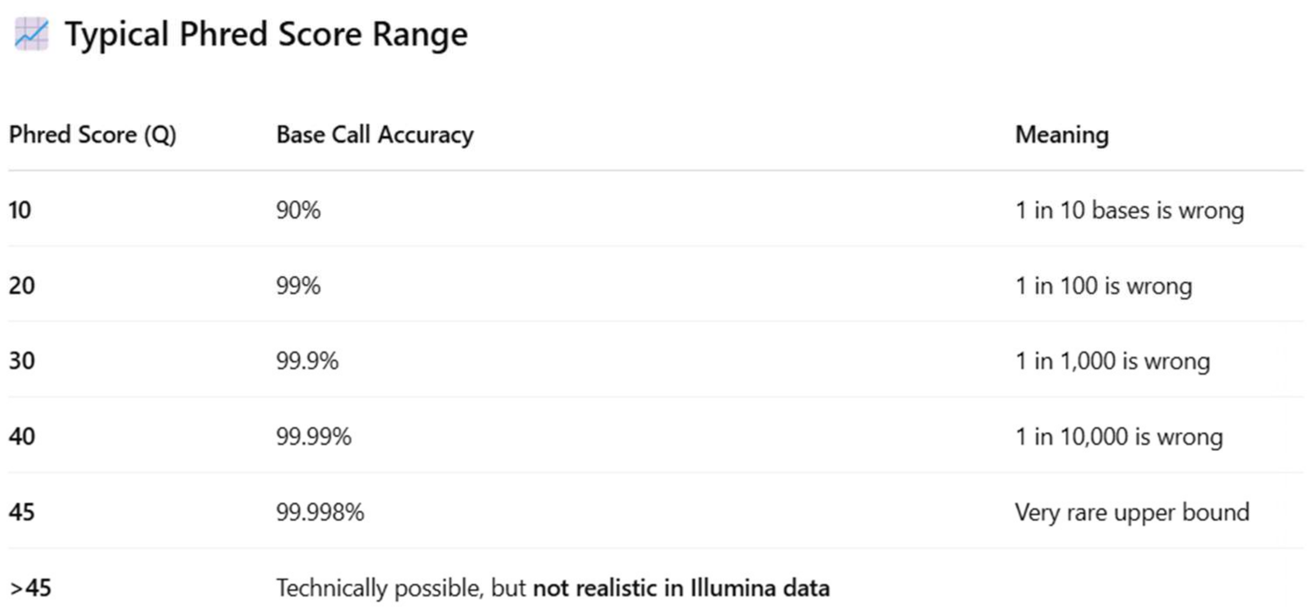


Illumina NovaSeq – Binned Phred scores

Illumina NovaSeq

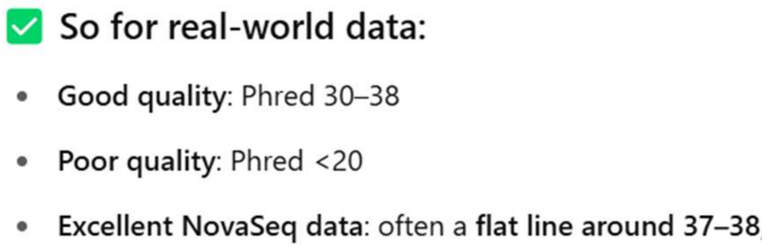


Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.



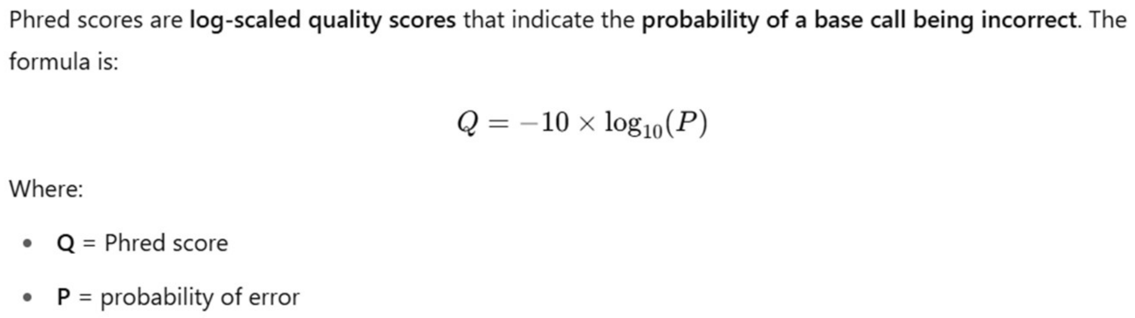
📈 Typical Phred Score Range

Phred Score (Q)	Base Call Accuracy	Meaning
10	90%	1 in 10 bases is wrong
20	99%	1 in 100 is wrong
30	99.9%	1 in 1,000 is wrong
40	99.99%	1 in 10,000 is wrong
45	99.998%	Very rare upper bound
>45	Technically possible, but **not realistic in Illumina data**	



✅ So for real-world data:

- **Good quality**: Phred 30–38
- **Poor quality**: Phred <20
- **Excellent NovaSeq data**: often a **flat line around 37–38**



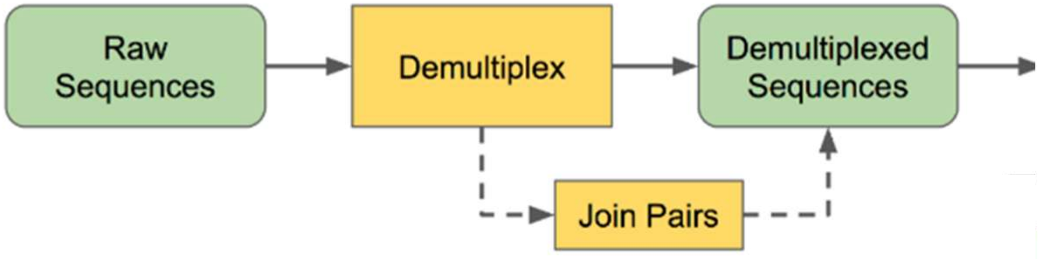
Phred scores are **log-scaled quality scores** that indicate the **probability of a base call being incorrect**. The formula is:

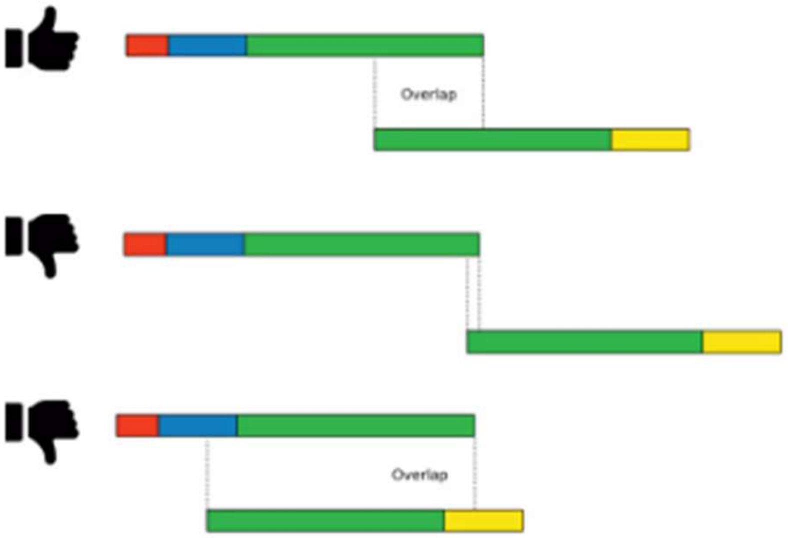
$$Q = -10 \times \log_{10}(P)$$

Where:

- **Q** = Phred score
- **P** = probability of error

Data preparation





- Import in a way that Qiime2 can read it ✓
- Demultiplex ✓
- Join read pairs – DADA2

  - Use overlaps between Forw and Rev reads

  - Too little overlap is bad but too much also

  - How much is enough? Depends on quality! A good overlap can be ~ 20–100 bp

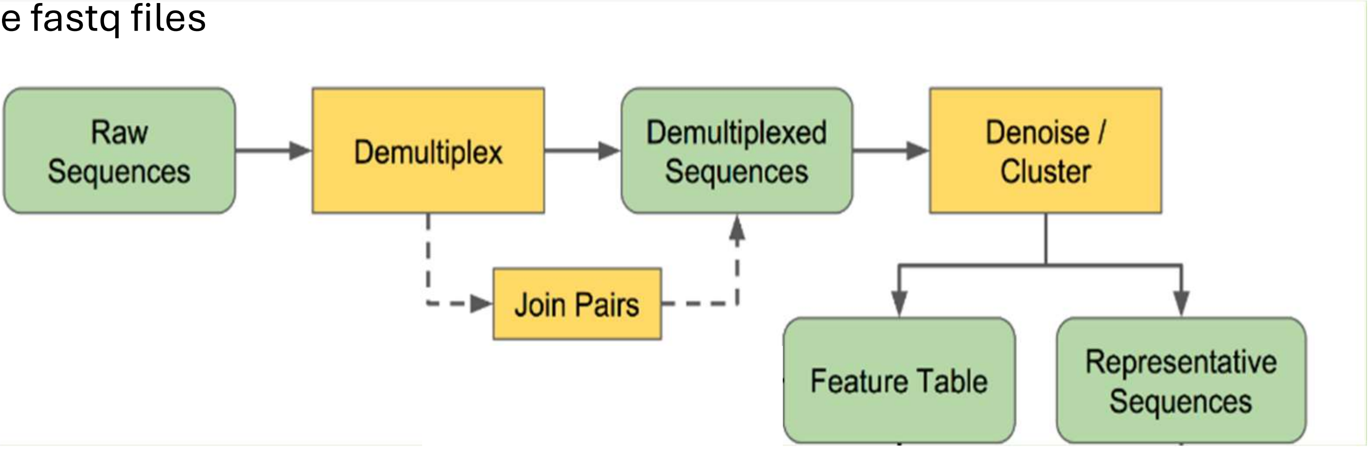
Quality filter/trim – DADA2

Abundance table construction

- Import in a way that Qiime2 can read it ✓
- Demultiplex ✓
- Join read pairs – DADA2 ✓
- Quality filter/trim – DADA2

  - Use Phred scores to purge or trim low quality reads

  - Phred scores are encoded in the fastq files



Clustering vs Denoising

Are very different strategies for dealing with sequencing noise and biological variation.

Spoiler alert! We are going to use the <u>Denoising strategy</u>

qiime2

Abundance table construction

Clustering vs Denoising

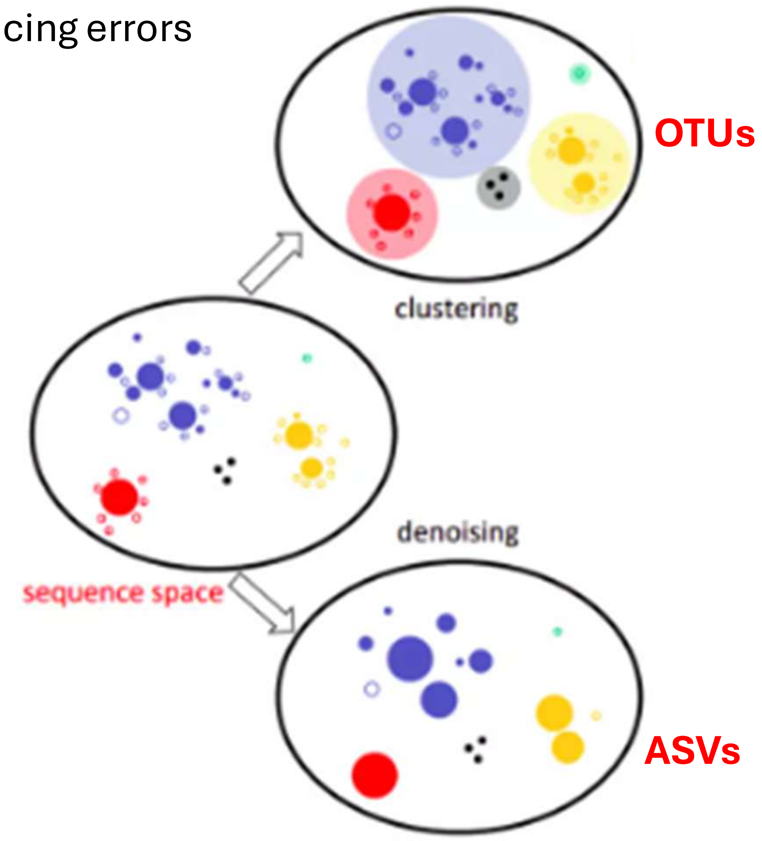
Purpose: distinguish biologically real nucleotide differences from sequencing errors

Clustering

- Traditional approach – less accurate
- Cluster sequences that fall above fixed similarity thresholds (e.g. 97%)
- Operational Taxonomic Units ~ species (OTUs)

Denoising

- Distinguish sequencing errors from true sequencing variants
- Up to single nucleotide resolution
- Amplicon Sequencing Variants (ASVs)



Abundance table construction

Denoising Methodologies - Deblur denoise vs DADA2 denoise

Filtering erroneous ASVs

- In **DADA2** sequences are changed to match the sequence they are more likely to belong to
- In **Deblur** sequences are removed

Filtering rare ASVs

- **DADA2** retains all sequences, no matter how rare
- **Deblur** discards everything under a frequency of 10 (default, you can change it!)

Use of reference database

- **DADA2** does not use a reference database to identify valid amplicon sequences
- **Deblur** uses Greengenes database as a reference

qiime2

Abundance table construction

Deblur denoise vs DADA2 denoise

Read quality requirements

- **DADA2** more sensitive to low quality reads – fail to join
- **Deblur** higher joining success even at low quality reads

Meta analysis

- **DADA2** cannot be used for meta analysis where individual data sets are pre-processed separately
- **Deblur** can be used for meta analysis

Taxonomy Assignment (https://docs.qiime2.org/2024.10/tutorials/overview/#taxonomy-flowchart)

Naive Bayes Classifier (sklearn method)

The Naive Bayes classifier outperforms other methods tested based on several criteria for classification of 16S rRNA gene, 18S and fungal ITS sequences

💪 Use a pre-trained classifier

- **SILVA** – for Bacteria (16S), Archaea (16S) and Eukaryotes (18S) most comprehensive, regularly updated
- **Greengenes** – for Bacteria (16S), older, but still common in some workflows
- **UNITE** - for fungi (ITS)

https://docs.qiime2.org/2024.10/data-resources/

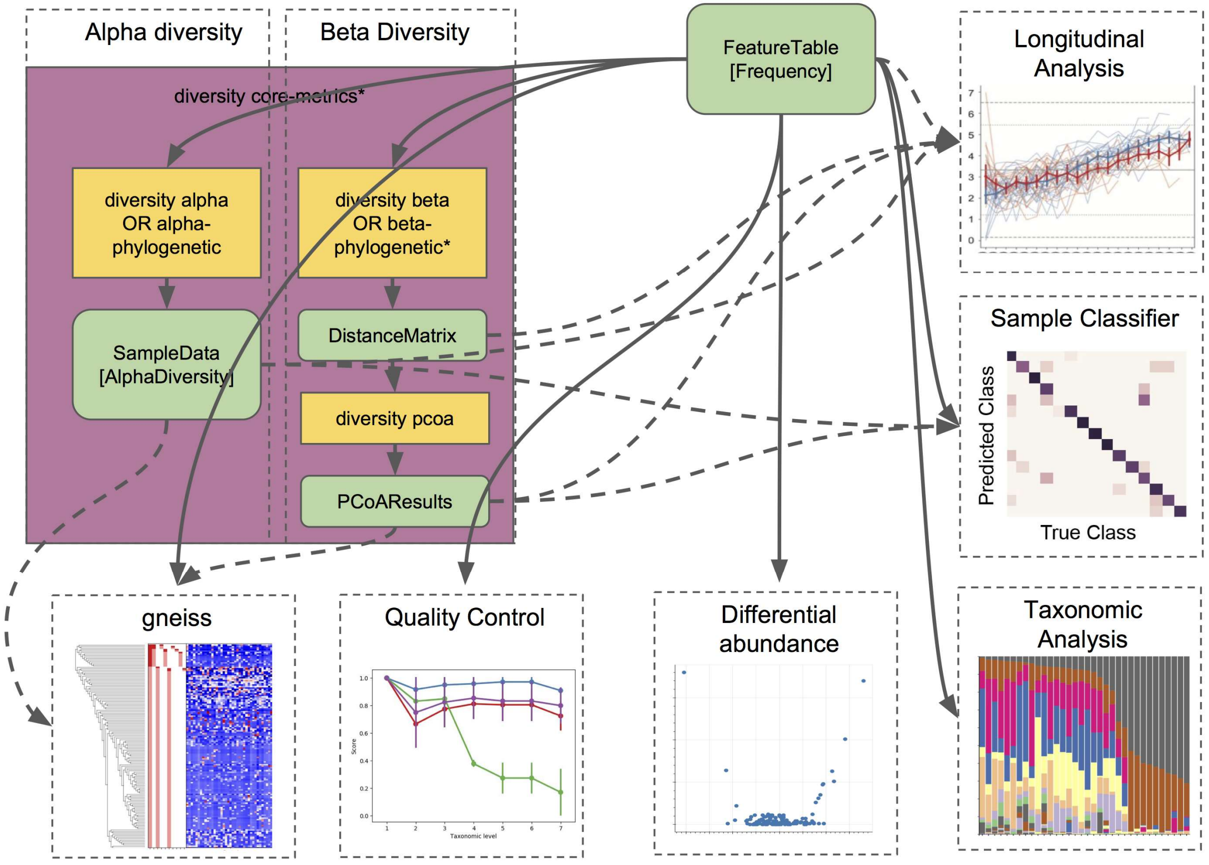
🏋 Train your classifier

https://docs.qiime2.org/2024.10/tutorials/overview/#diversity-analysis

Diversity analysis / Statistical Analysis / Taxonomic Analysis

In microbiome experiments, investigators frequently wonder about things like:

- How many different species//ASVs are present in my samples?
- How much phylogenetic diversity is present in each sample?
- How similar/different are individual samples and groups of samples?
- What factors (e.g. geography, host species, temperature, etc) associate with differences in microbial composition and biodiversity?



MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

MicrobiomeAnalyst is a **free online platform** designed to help you to analyse and visualize **microbiome data** — even if you don't have advanced programming skills!

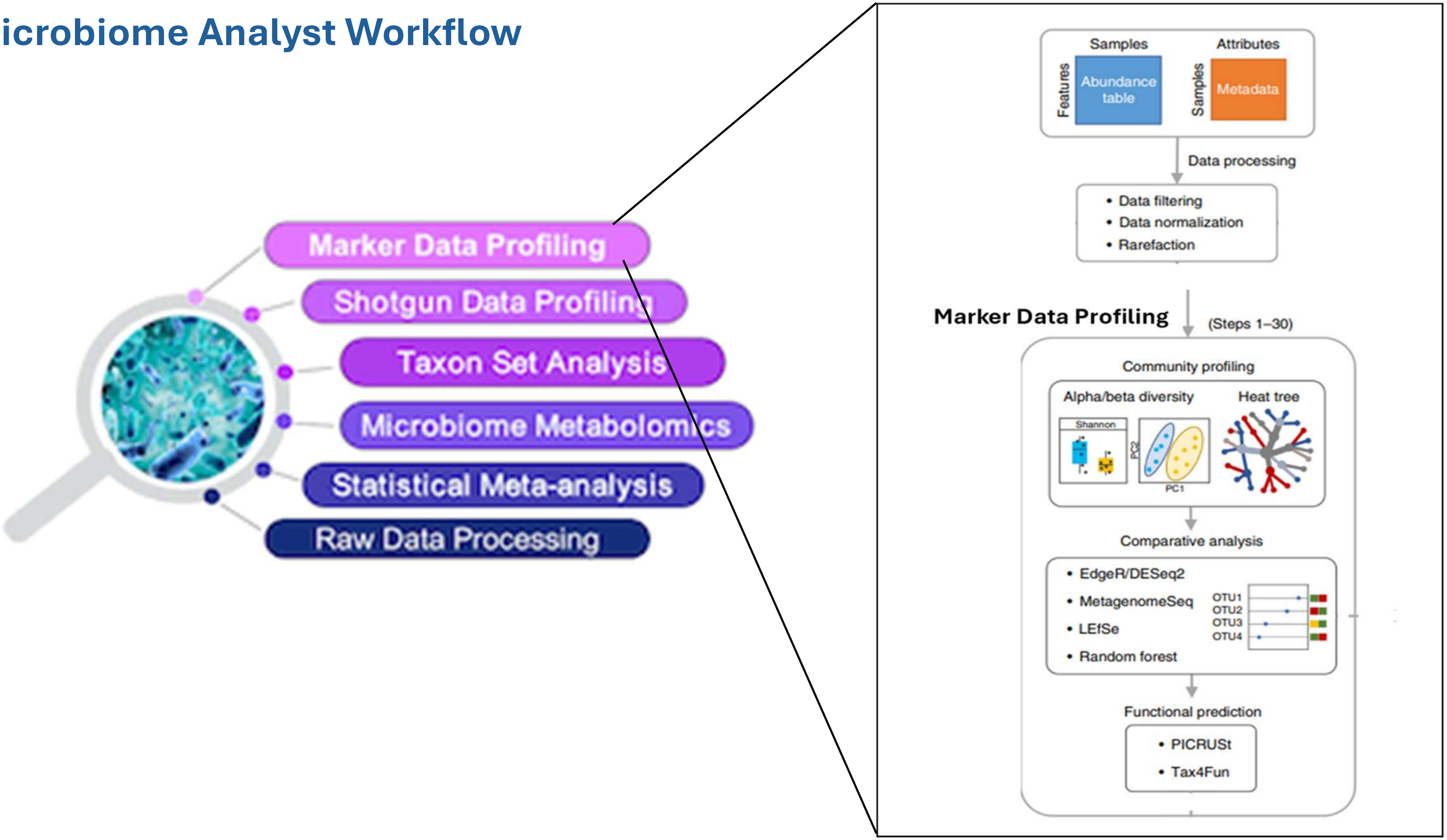
Although it runs in your browser, MicrobiomeAnalyst is **powered by R**. Using well-established **R packages for statistical analysis** and **data visualization** behind the scenes. This means it's doing serious analysis under the hood, even if you don't have to write any code.

With MicrobiomeAnalyst, you can:

- **Clean and filter** your data
- **Explore diversity** (alpha/beta diversity)
- **Identify key microbes** driving differences between groups
- **Predict functions** (like metabolic pathways)
- **Create interactive plots** for presentations or papers

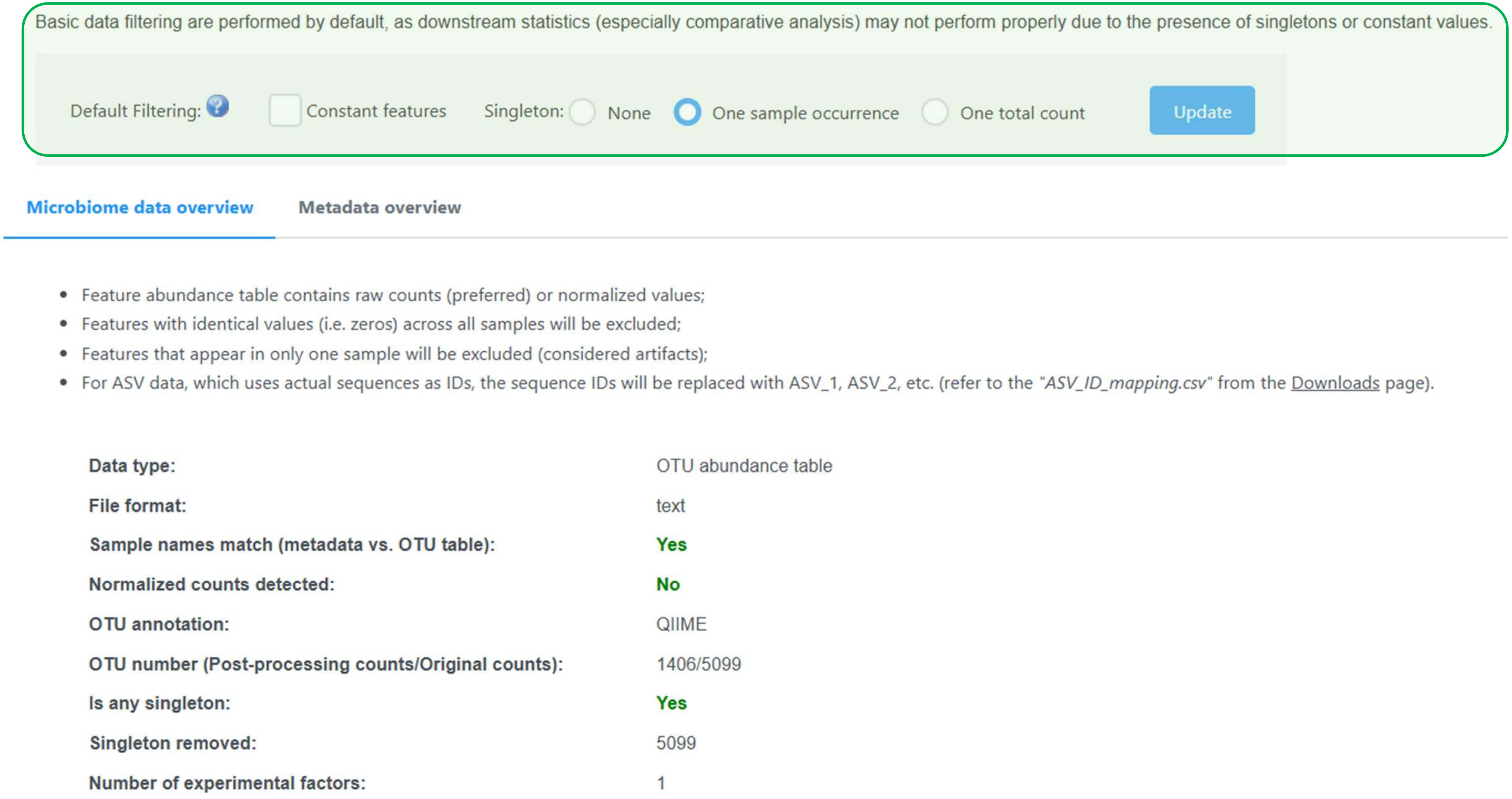
It supports common input formats (like QIIME2 outputs or OTU/ASV tables) and has **step-by-step workflows** — so you don't need to be a bioinformatics expert to use it.

Microbiome Analyst Workflow



https://www.nature.com/articles/s41596-019-0264-1

Data Integrity Check



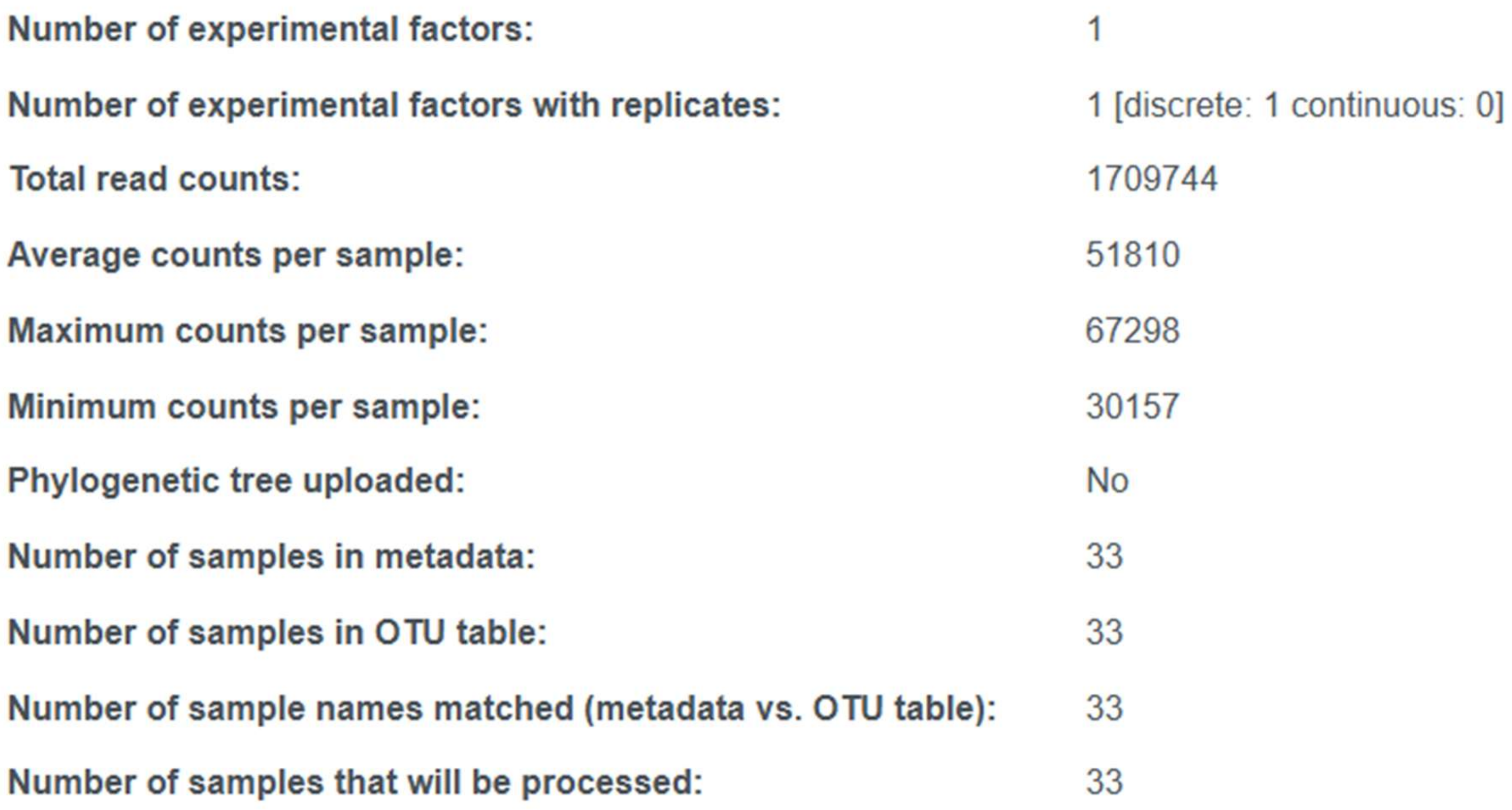
Basic data filtering are performed by default, as downstream statistics (especially comparative analysis) may not perform properly due to the presence of singletons or constant values.

Default Filtering: □ Constant features Singleton: ○ None ● One sample occurrence ○ One total count

- Feature abundance table contains raw counts (preferred) or normalized values;
- Features with identical values (i.e. zeros) across all samples will be excluded;
- Features that appear in only one sample will be excluded (considered artifacts);
- For ASV data, which uses actual sequences as IDs, the sequence IDs will be replaced with ASV_1, ASV_2, etc. (refer to the "*ASV_ID_mapping.csv*" from the Downloads page).

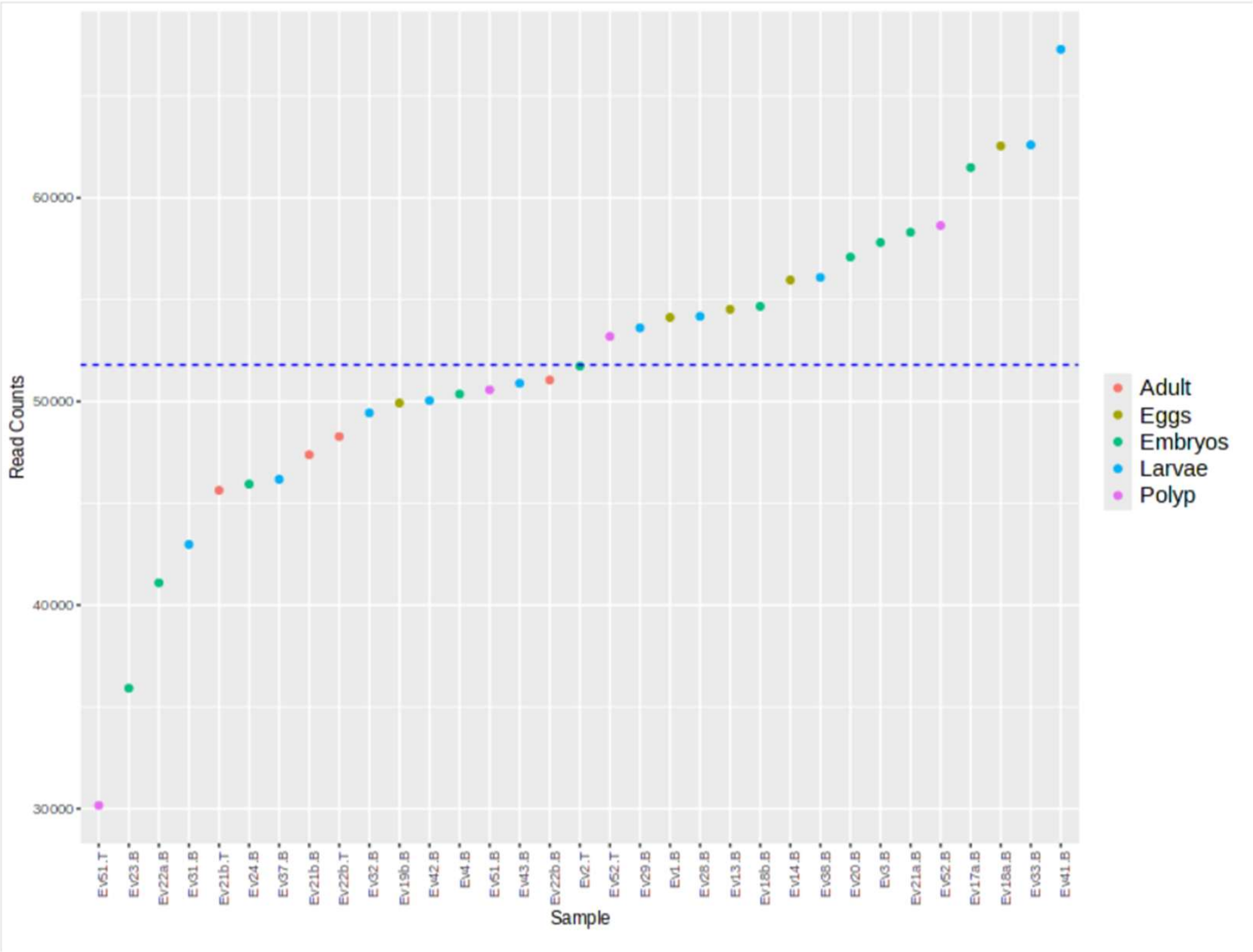
Data type:	OTU abundance table
File format:	text
Sample names match (metadata vs. OTU table):	**Yes**
Normalized counts detected:	**No**
OTU annotation:	QIIME
OTU number (Post-processing counts/Original counts):	1406/5099
Is any singleton:	**Yes**
Singleton removed:	5099
Number of experimental factors:	1

Data Integrity Check (cont.)

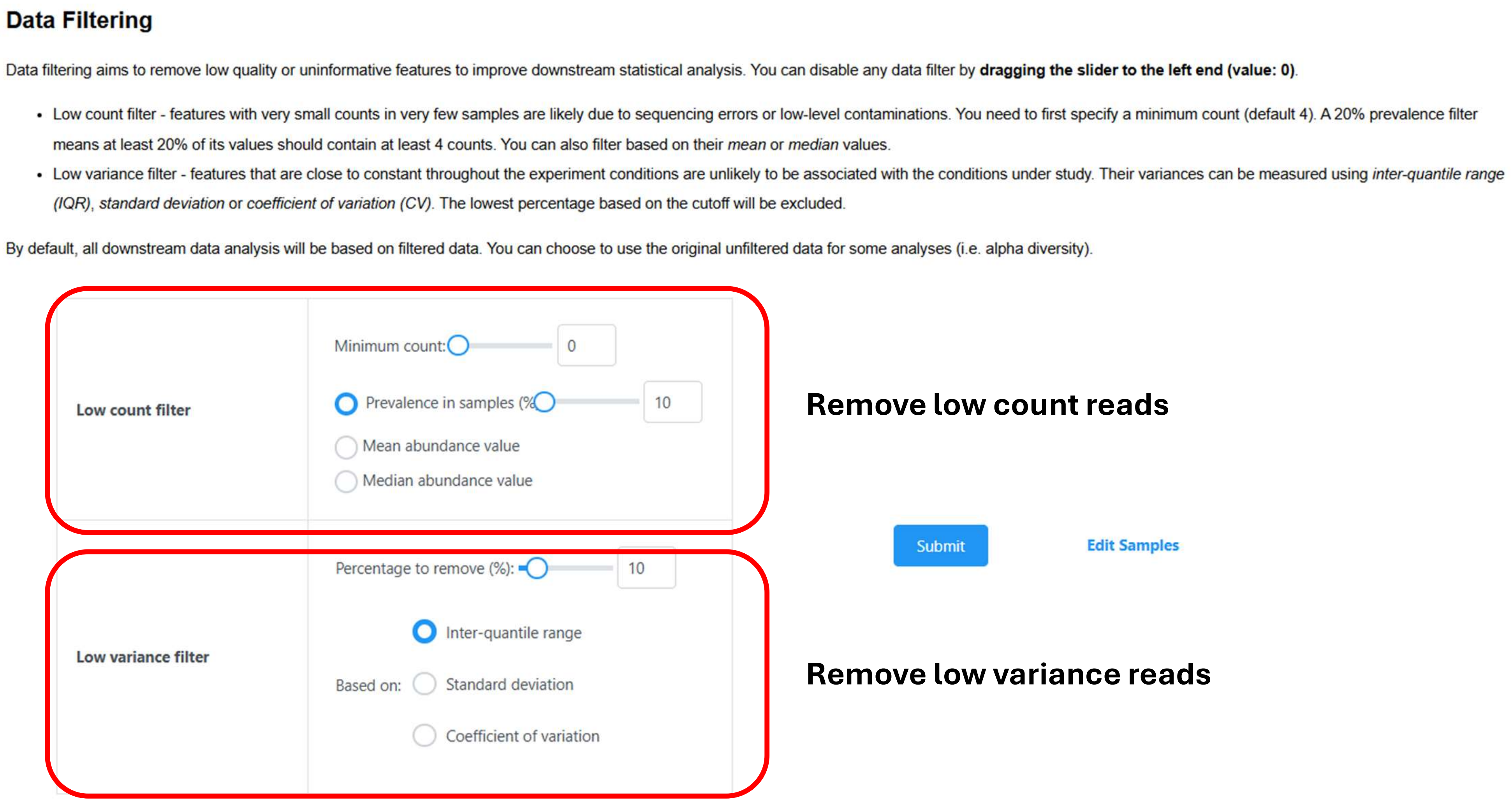


Number of experimental factors:	1
Number of experimental factors with replicates:	1 [discrete: 1 continuous: 0]
Total read counts:	1709744
Average counts per sample:	51810
Maximum counts per sample:	67298
Minimum counts per sample:	30157
Phylogenetic tree uploaded:	No
Number of samples in metadata:	33
Number of samples in OTU table:	33
Number of sample names matched (metadata vs. OTU table):	33
Number of samples that will be processed:	33

Data Integrity Check (cont.)



Data Filtering



Data Filtering

Data filtering aims to remove low quality or uninformative features to improve downstream statistical analysis. You can disable any data filter by **dragging the slider to the left end (value: 0)**.

- Low count filter - features with very small counts in very few samples are likely due to sequencing errors or low-level contaminations. You need to first specify a minimum count (default 4). A 20% prevalence filter means at least 20% of its values should contain at least 4 counts. You can also filter based on their *mean* or *median* values.
- Low variance filter - features that are close to constant throughout the experiment conditions are unlikely to be associated with the conditions under study. Their variances can be measured using *inter-quantile range (IQR)*, *standard deviation* or *coefficient of variation (CV)*. The lowest percentage based on the cutoff will be excluded.

By default, all downstream data analysis will be based on filtered data. You can choose to use the original unfiltered data for some analyses (i.e. alpha diversity).

Low count filter

Minimum count: 0

● Prevalence in samples (% 10

○ Mean abundance value

○ Median abundance value

Low variance filter

Percentage to remove (%): 10

Based on:

● Inter-quantile range

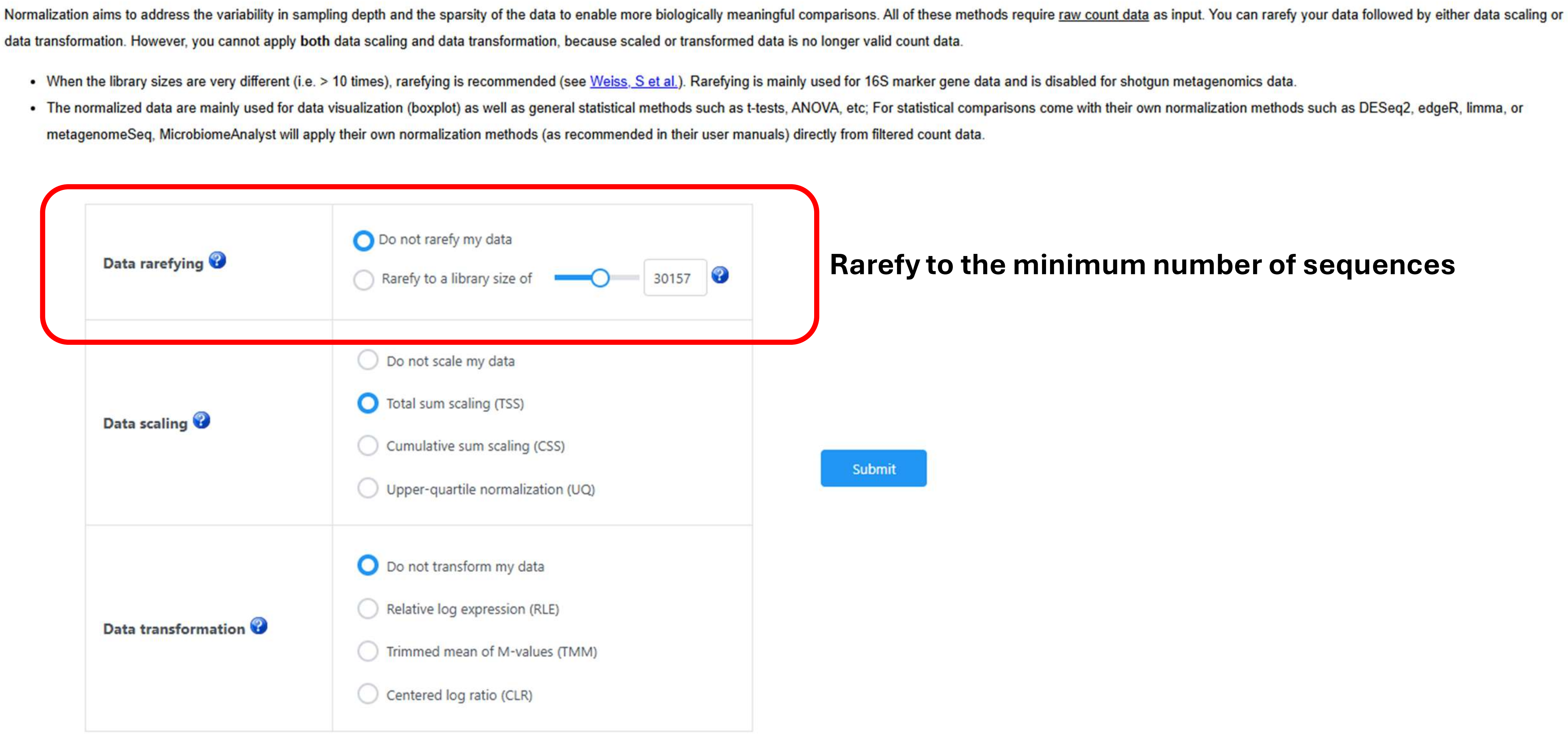
○ Standard deviation

○ Coefficient of variation

Remove low count reads

Remove low variance reads

Data Normalization

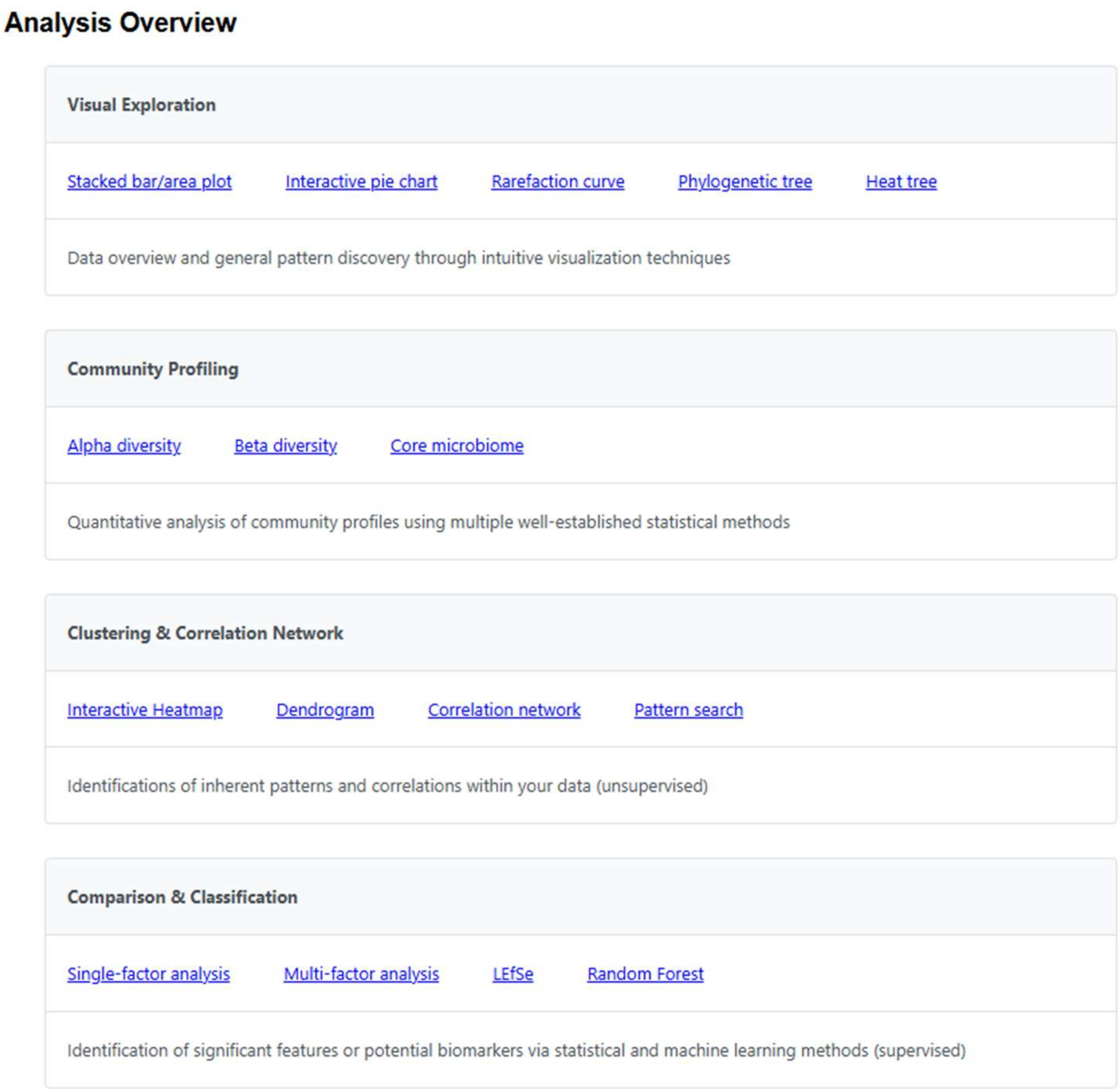


Normalization aims to address the variability in sampling depth and the sparsity of the data to enable more biologically meaningful comparisons. All of these methods require raw count data as input. You can rarefy your data followed by either data scaling or data transformation. However, you cannot apply **both** data scaling and data transformation, because scaled or transformed data is no longer valid count data.

- When the library sizes are very different (i.e. > 10 times), rarefying is recommended (see Weiss, S et al.). Rarefying is mainly used for 16S marker gene data and is disabled for shotgun metagenomics data.
- The normalized data are mainly used for data visualization (boxplot) as well as general statistical methods such as t-tests, ANOVA, etc; For statistical comparisons come with their own normalization methods such as DESeq2, edgeR, limma, or metagenomeSeq, MicrobiomeAnalyst will apply their own normalization methods (as recommended in their user manuals) directly from filtered count data.

Data rarefying	● Do not rarefy my data ○ Rarefy to a library size of 30157
Data scaling	○ Do not scale my data ● Total sum scaling (TSS) ○ Cumulative sum scaling (CSS) ○ Upper-quartile normalization (UQ)
Data transformation	● Do not transform my data ○ Relative log expression (RLE) ○ Trimmed mean of M-values (TMM) ○ Centered log ratio (CLR)

Rarefy to the minimum number of sequences



Analysis Overview

Visual Exploration

Stacked bar/area plot   Interactive pie chart   Rarefaction curve   Phylogenetic tree   Heat tree

Data overview and general pattern discovery through intuitive visualization techniques

Community Profiling

Alpha diversity   Beta diversity   Core microbiome

Quantitative analysis of community profiles using multiple well-established statistical methods

Clustering & Correlation Network

Interactive Heatmap   Dendrogram   Correlation network   Pattern search

Identifications of inherent patterns and correlations within your data (unsupervised)

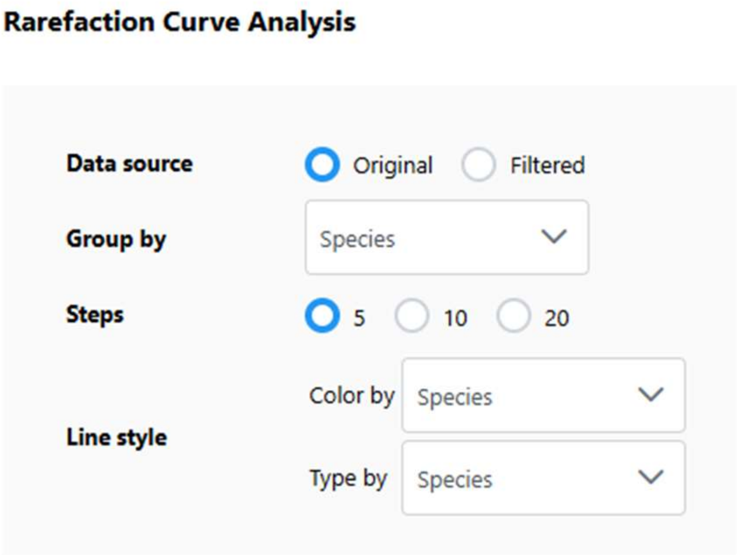
Comparison & Classification

Single-factor analysis   Multi-factor analysis   LEfSe   Random Forest

Identification of significant features or potential biomarkers via statistical and machine learning methods (supervised)

There's also Functional Prediction but we will not do it!

A. Visual Exploration



Rarefaction Curve Analysis

Data source ● Original ○ Filtered

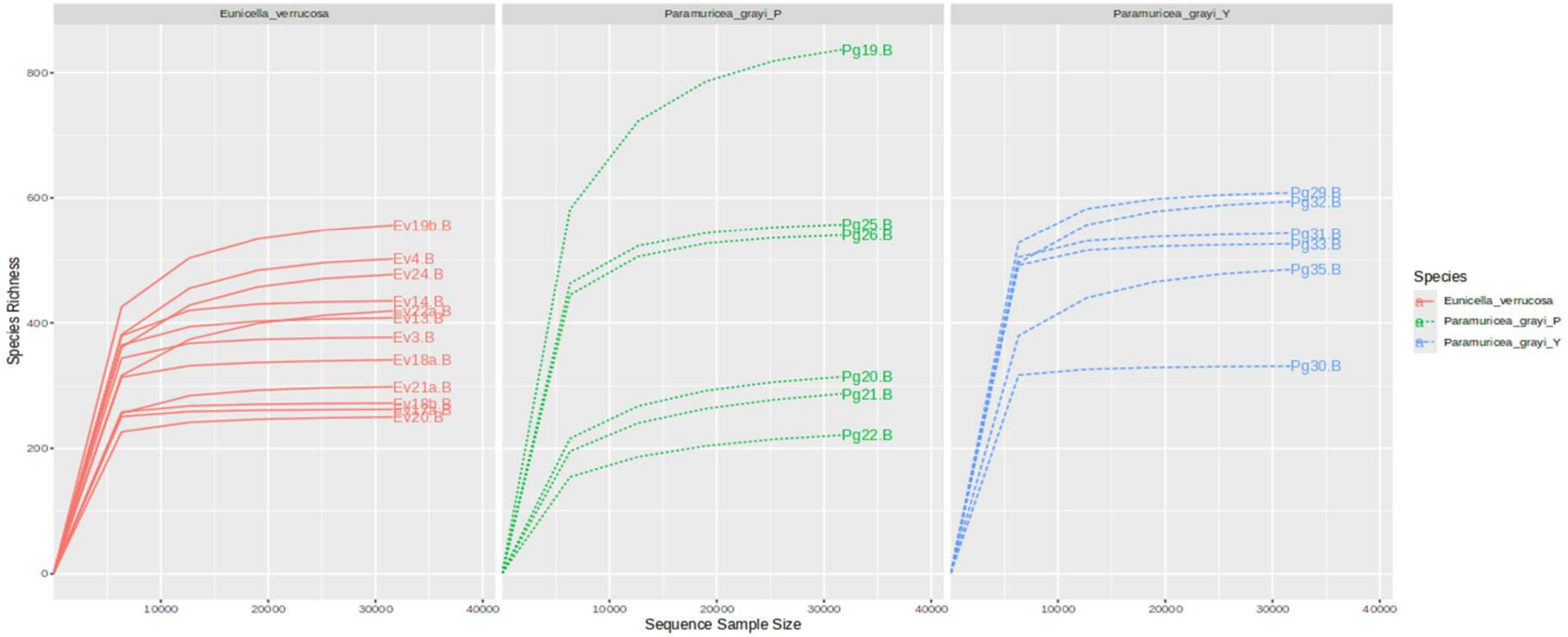
Group by: Species

Steps ● 5 ○ 10 ○ 20

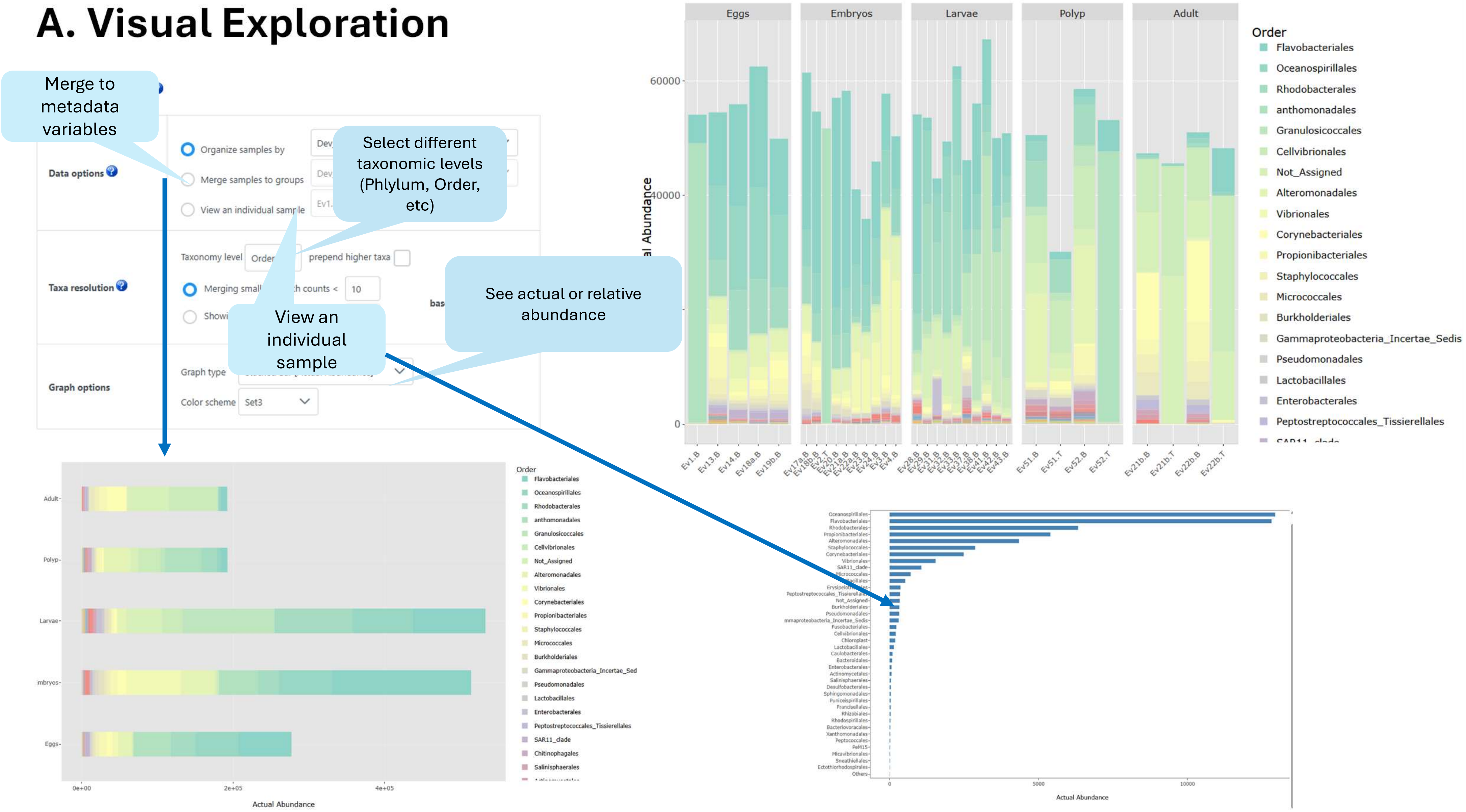
Line style

Color by: Species

Type by: Species



A. Visual Exploration



Merge to metadata variables

Select different taxonomic levels (Phylum, Order, etc)

See actual or relative abundance

View an individual sample

Data options

● Organize samples by Dev

○ Merge samples to groups Dev

○ View an individual sample Ev1.

Taxa resolution

Taxonomy level Order

prepend higher taxa □

● Merging small … h counts < 10

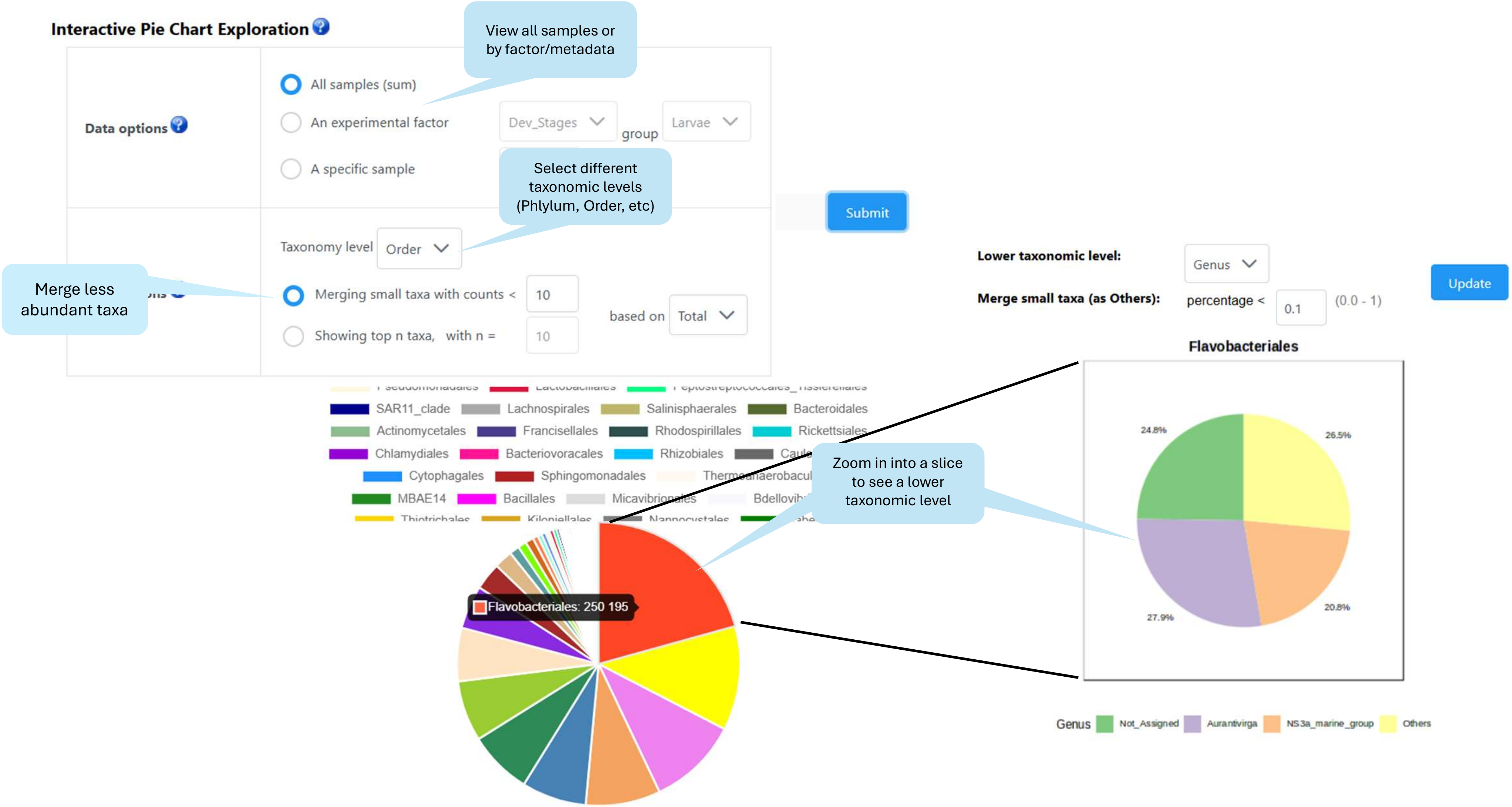
○ Showi…

Graph options

Graph type

Color scheme Set3

A. Visual Exploration



Interactive Pie Chart Exploration

Data options

- ● All samples (sum)
- ○ An experimental factor: Dev_Stages, group: Larvae
- ○ A specific sample

Taxonomy level: Order

- ● Merging small taxa with counts < 10
- ○ Showing top n taxa, with n = 10

based on: Total

Submit

View all samples or by factor/metadata

Select different taxonomic levels (Phlylum, Order, etc)

Merge less abundant taxa

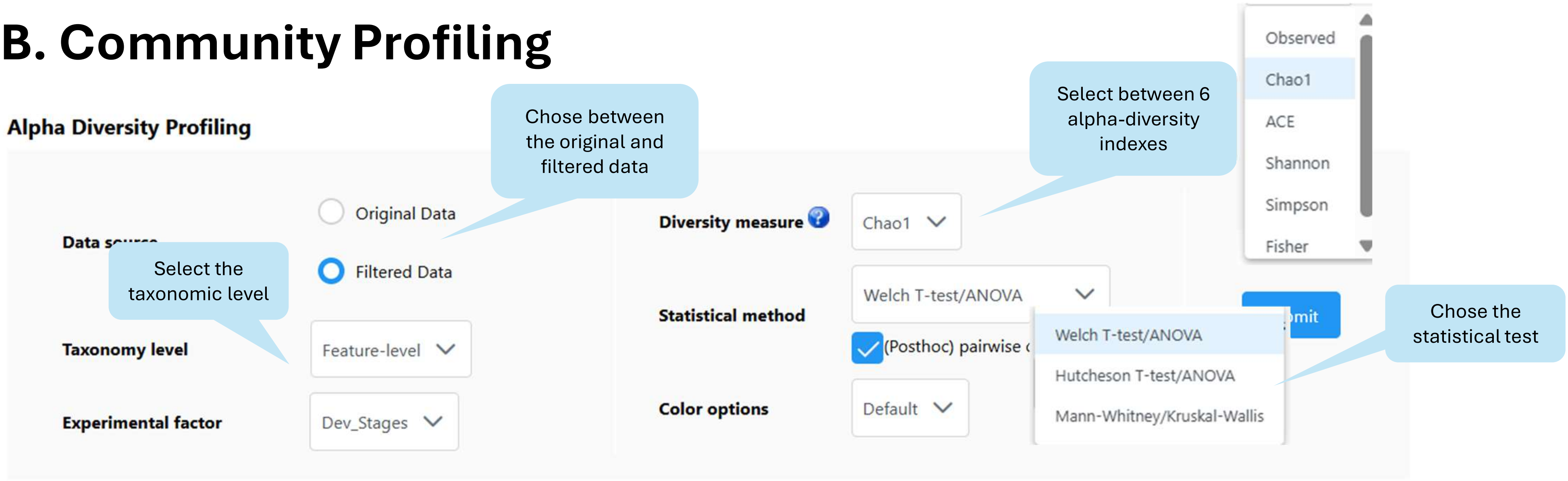
Zoom in into a slice to see a lower taxonomic level

Lower taxonomic level: Genus

Merge small taxa (as Others): percentage < 0.1 (0.0 - 1)

Update

B. Community Profiling



Alpha Diversity Profiling

Data source: ○ Original Data ● Filtered Data

Taxonomy level: Feature-level

Experimental factor: Dev_Stages

Diversity measure: Chao1

Statistical method: Welch T-test/ANOVA

☑ (Posthoc) pairwise c

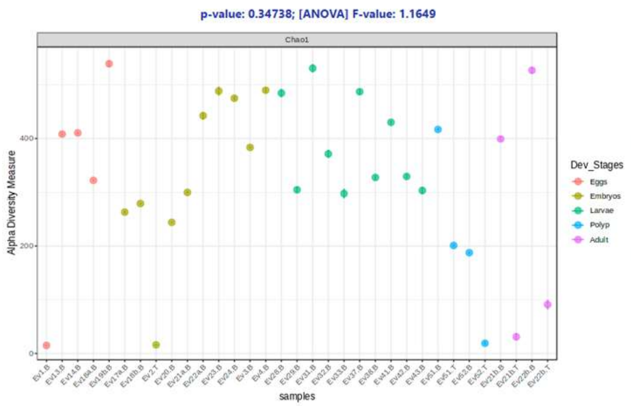
Color options: Default

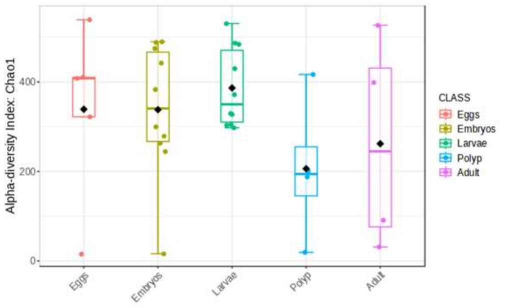
Chose between the original and filtered data

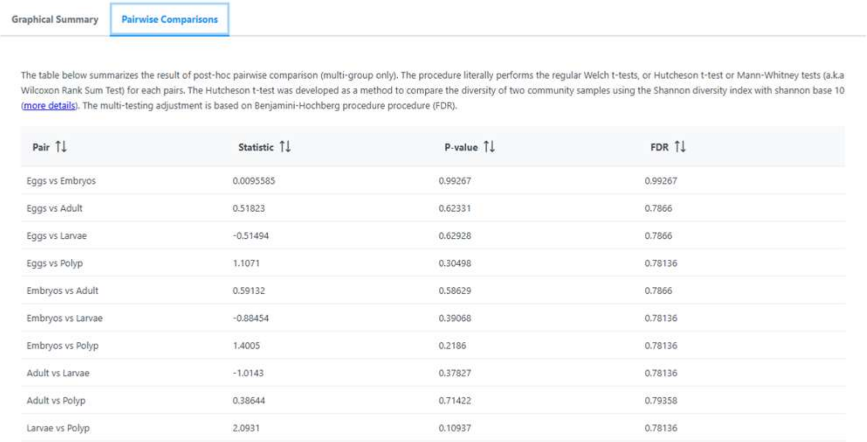
Select the taxonomic level

Select between 6 alpha-diversity indexes

Chose the statistical test



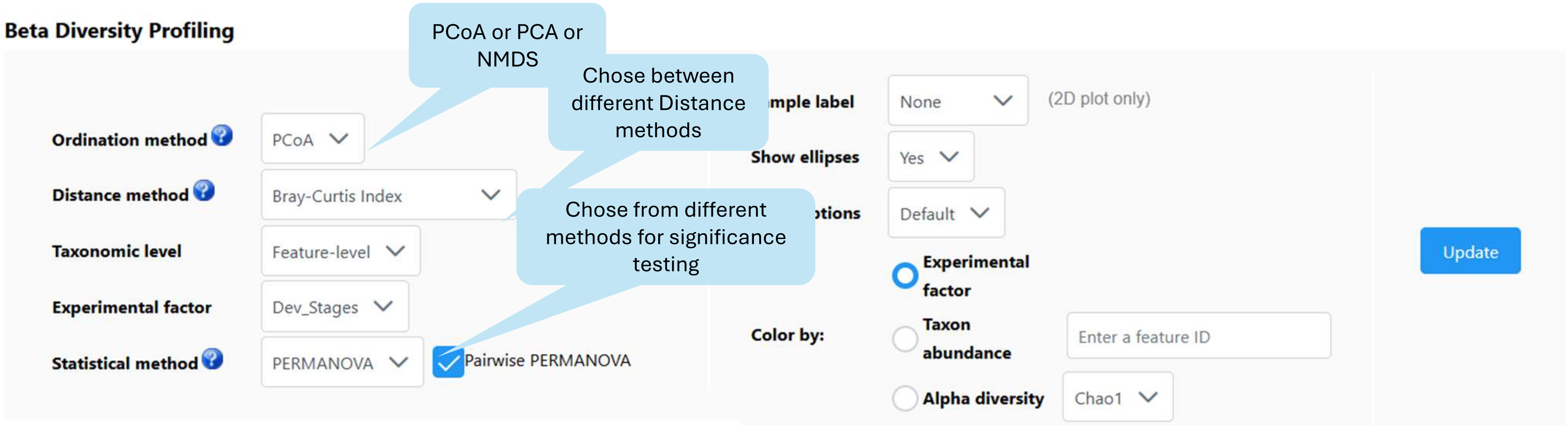




The table below summarizes the result of post-hoc pairwise comparison (multi-group only). The procedure literally performs the regular Welch t-tests, or Hutcheson t-test or Mann-Whitney tests (a.k.a Wilcoxon Rank Sum Test) for each pairs. The Hutcheson t-test was developed as a method to compare the diversity of two community samples using the Shannon diversity index with shannon base 10 (more details). The multi-testing adjustment is based on Benjamini-Hochberg procedure procedure (FDR).

Pair ⇅	Statistic ⇅	P-value ⇅	FDR ⇅
Eggs vs Embryos	0.0095585	0.99267	0.99267
Eggs vs Adult	0.51823	0.62331	0.7866
Eggs vs Larvae	-0.51494	0.62928	0.7866
Eggs vs Polyp	1.1071	0.30498	0.78136
Embryos vs Adult	0.59132	0.58629	0.7866
Embryos vs Larvae	-0.88454	0.39068	0.78136
Embryos vs Polyp	1.4005	0.2186	0.78136
Adult vs Larvae	-1.0143	0.37827	0.78136
Adult vs Polyp	0.38644	0.71422	0.79358
Larvae vs Polyp	2.0931	0.10937	0.78136

B. Community Profiling



Beta Diversity Profiling

PCoA or PCA or NMDS

Chose between different Distance methods

Chose from different methods for significance testing

Ordination method PCoA

Distance method Bray-Curtis Index

Taxonomic level Feature-level

Experimental factor Dev_Stages

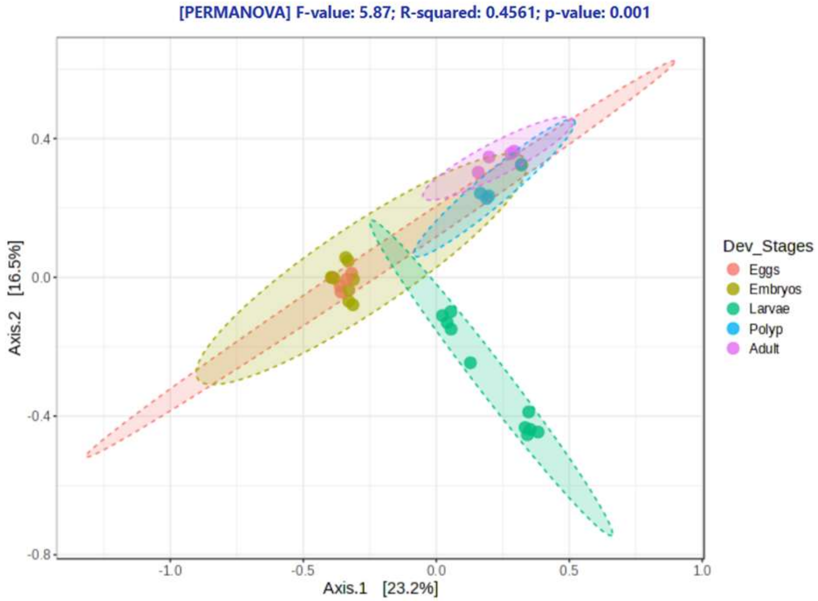
Statistical method PERMANOVA ☑ Pairwise PERMANOVA

mple label None (2D plot only)

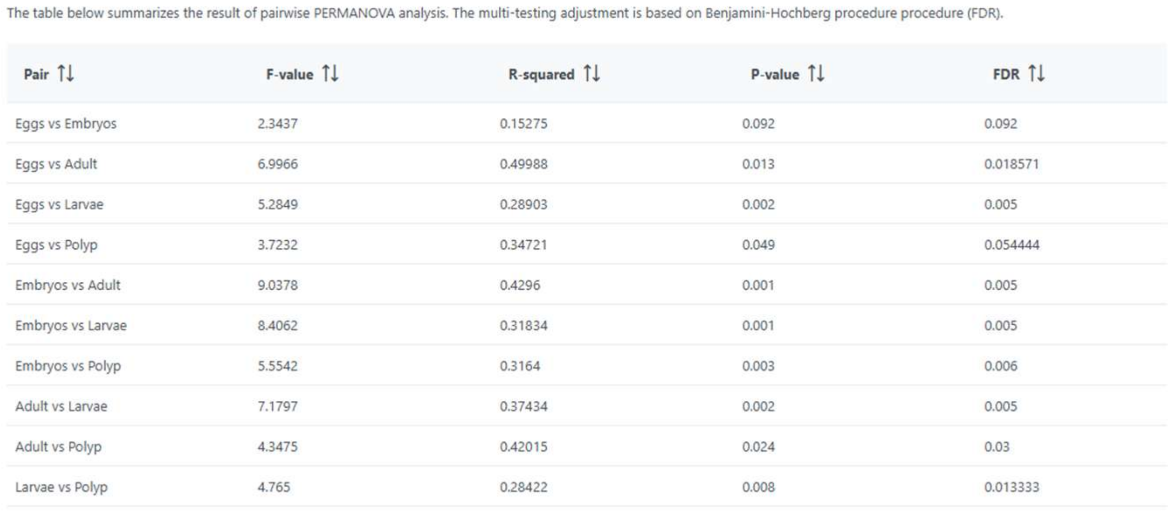
Show ellipses Yes

ptions Default

Color by: ● **Experimental factor** ○ **Taxon abundance** Enter a feature ID ○ **Alpha diversity** Chao1



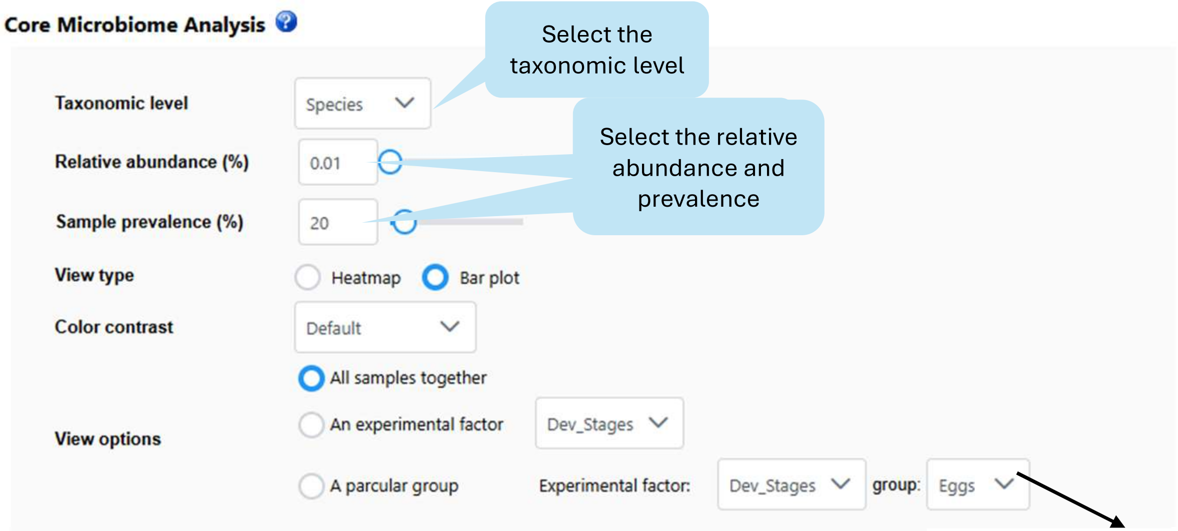
[PERMANOVA] F-value: 5.87; R-squared: 0.4561; p-value: 0.001



The table below summarizes the result of pairwise PERMANOVA analysis. The multi-testing adjustment is based on Benjamini-Hochberg procedure procedure (FDR).

Pair	F-value	R-squared	P-value	FDR
Eggs vs Embryos	2.3437	0.15275	0.092	0.092
Eggs vs Adult	6.9966	0.49988	0.013	0.018571
Eggs vs Larvae	5.2849	0.28903	0.002	0.005
Eggs vs Polyp	3.7232	0.34721	0.049	0.054444
Embryos vs Adult	9.0378	0.4296	0.001	0.005
Embryos vs Larvae	8.4062	0.31834	0.001	0.005
Embryos vs Polyp	5.5542	0.3164	0.003	0.006
Adult vs Larvae	7.1797	0.37434	0.002	0.005
Adult vs Polyp	4.3475	0.42015	0.024	0.03
Larvae vs Polyp	4.765	0.28422	0.008	0.013333

B. Community Profiling



Core Microbiome Analysis

Select the taxonomic level

Select the relative abundance and prevalence

Taxonomic level: Species

Relative abundance (%): 0.01

Sample prevalence (%): 20

View type: ○ Heatmap ● Bar plot

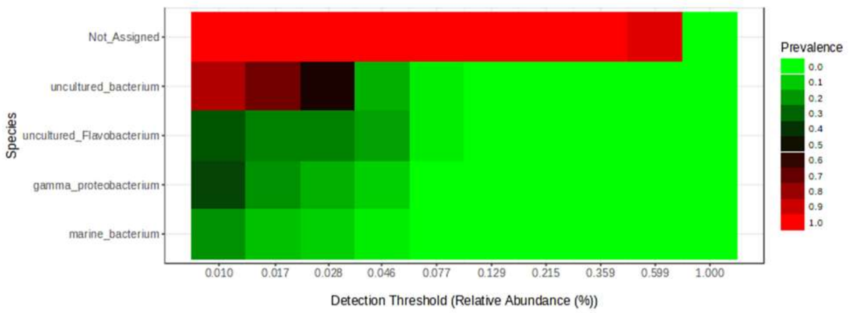
Color contrast: Default

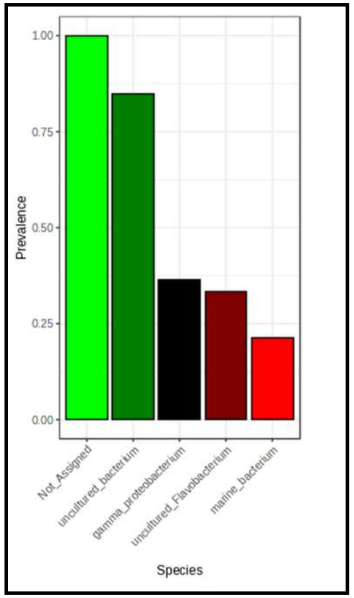
View options:

● All samples together

○ An experimental factor Dev_Stages

○ A parcular group Experimental factor: Dev_Stages group: Eggs

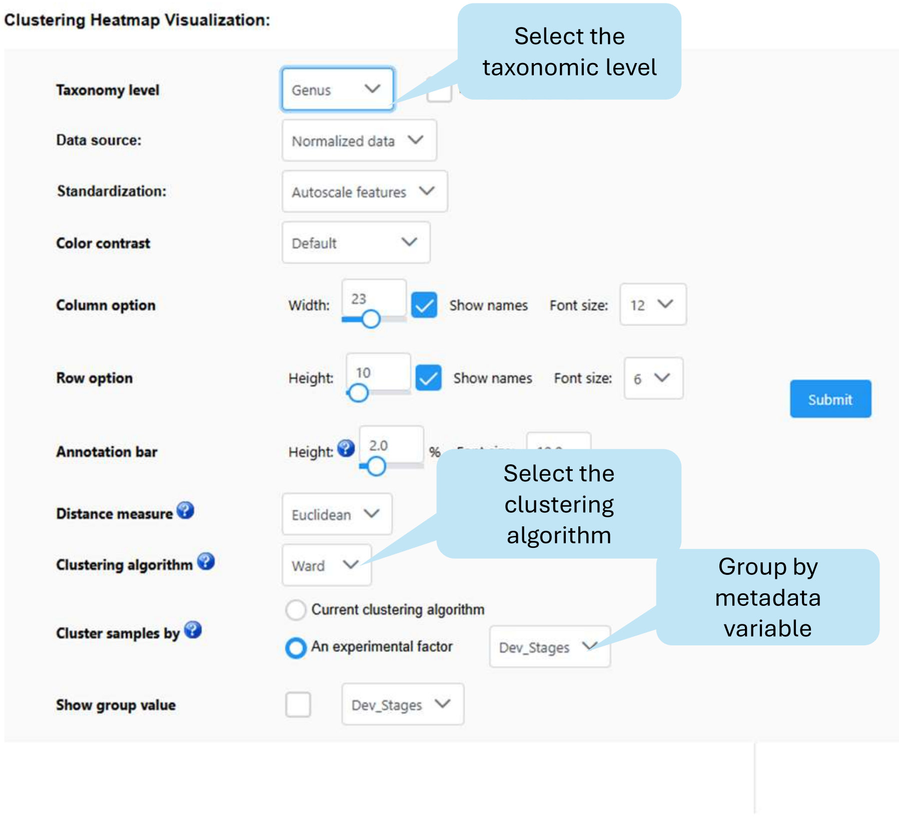




- Identifies the bacteria that are common to all the samples, to a certain group of samples

- Can be performed at different taxonomic levels

C. Clustering Analysis



Clustering Heatmap Visualization:

Taxonomy level: Genus

Data source: Normalized data

Standardization: Autoscale features

Color contrast: Default

Column option — Width: 23 ☑ Show names Font size: 12

Row option — Height: 10 ☑ Show names Font size: 6

Annotation bar — Height: 2.0 %

Distance measure: Euclidean

Clustering algorithm: Ward

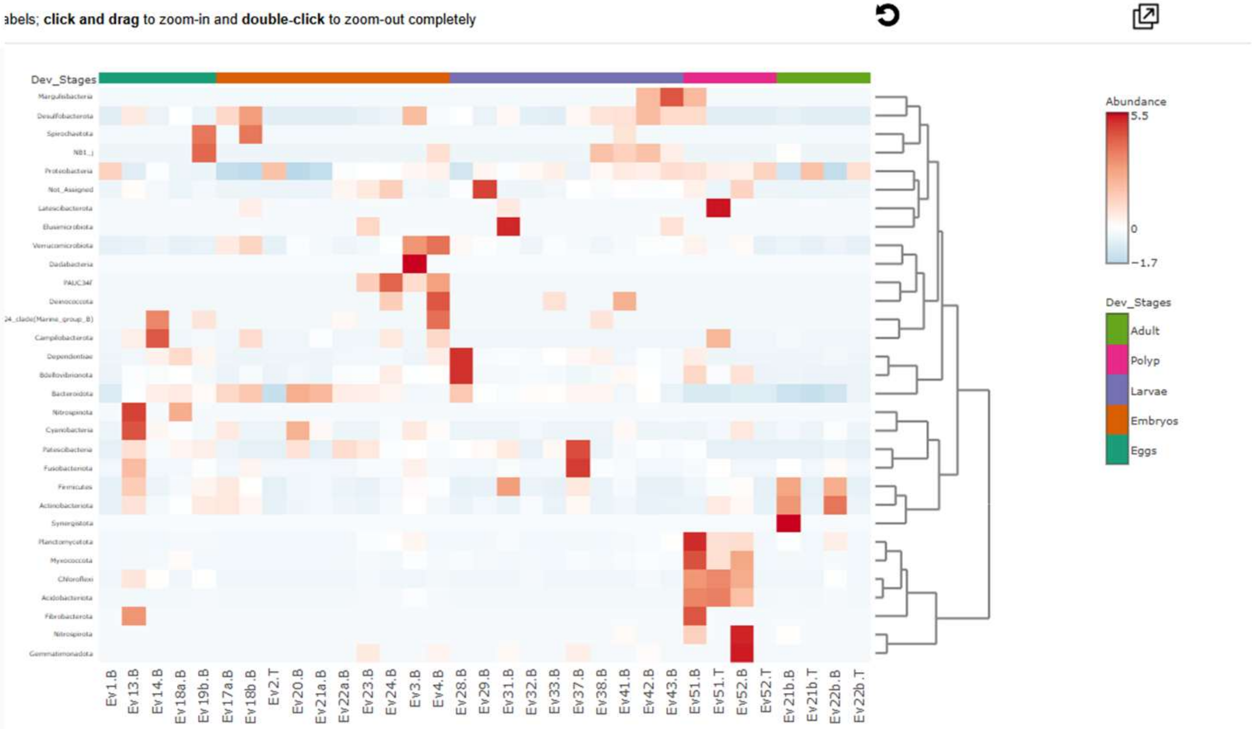
Cluster samples by: ○ Current clustering algorithm ● An experimental factor Dev_Stages

Show group value: □ Dev_Stages

Select the taxonomic level

Select the clustering algorithm

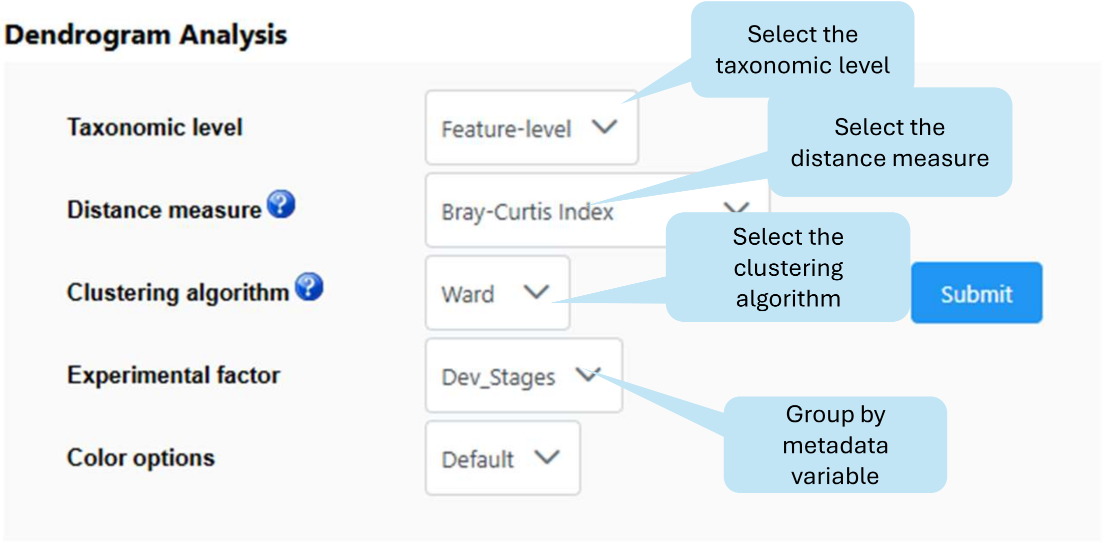
Group by metadata variable



abels; **click and drag** to zoom-in and **double-click** to zoom-out completely

Allows you to identify abundance patterns/clusters

C. Clustering Analysis



Dendrogram Analysis

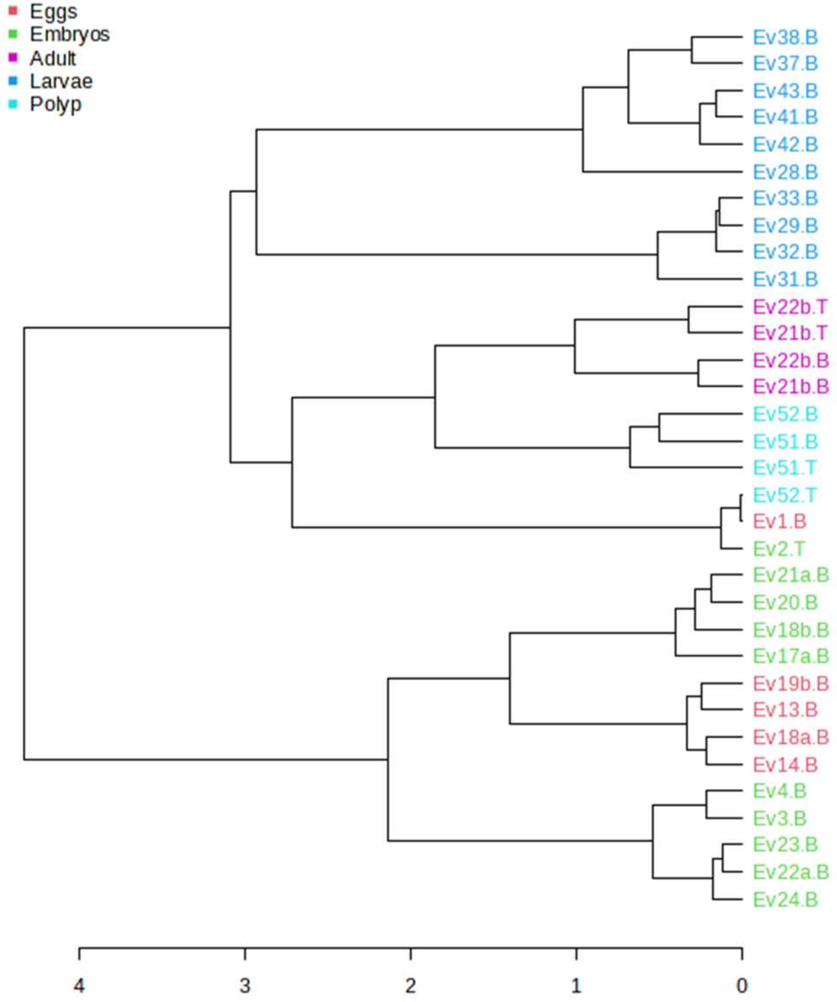
Taxonomic level	Feature-level
Distance measure	Bray-Curtis Index
Clustering algorithm	Ward
Experimental factor	Dev_Stages
Color options	Default

Select the taxonomic level

Select the distance measure

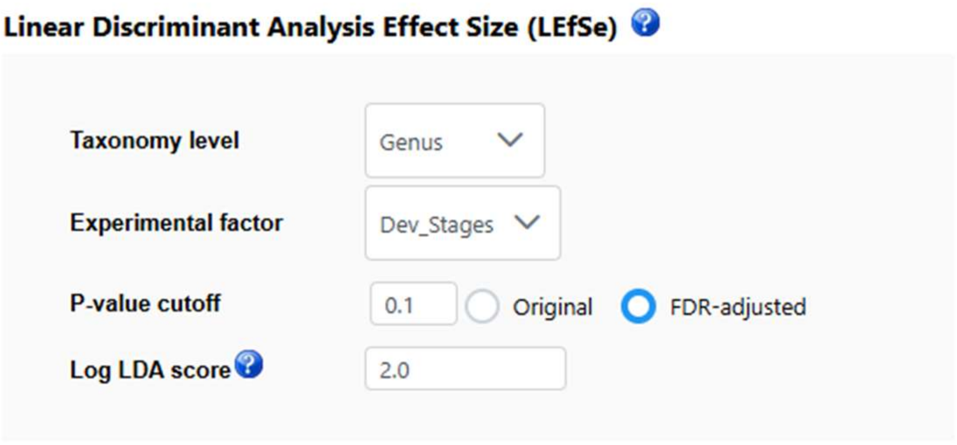
Select the clustering algorithm

Group by metadata variable



- Performs phylogenetic analysis on samples using either various phylogenetic or nonphylogenetic distance measures

D. Biomarker Analysis



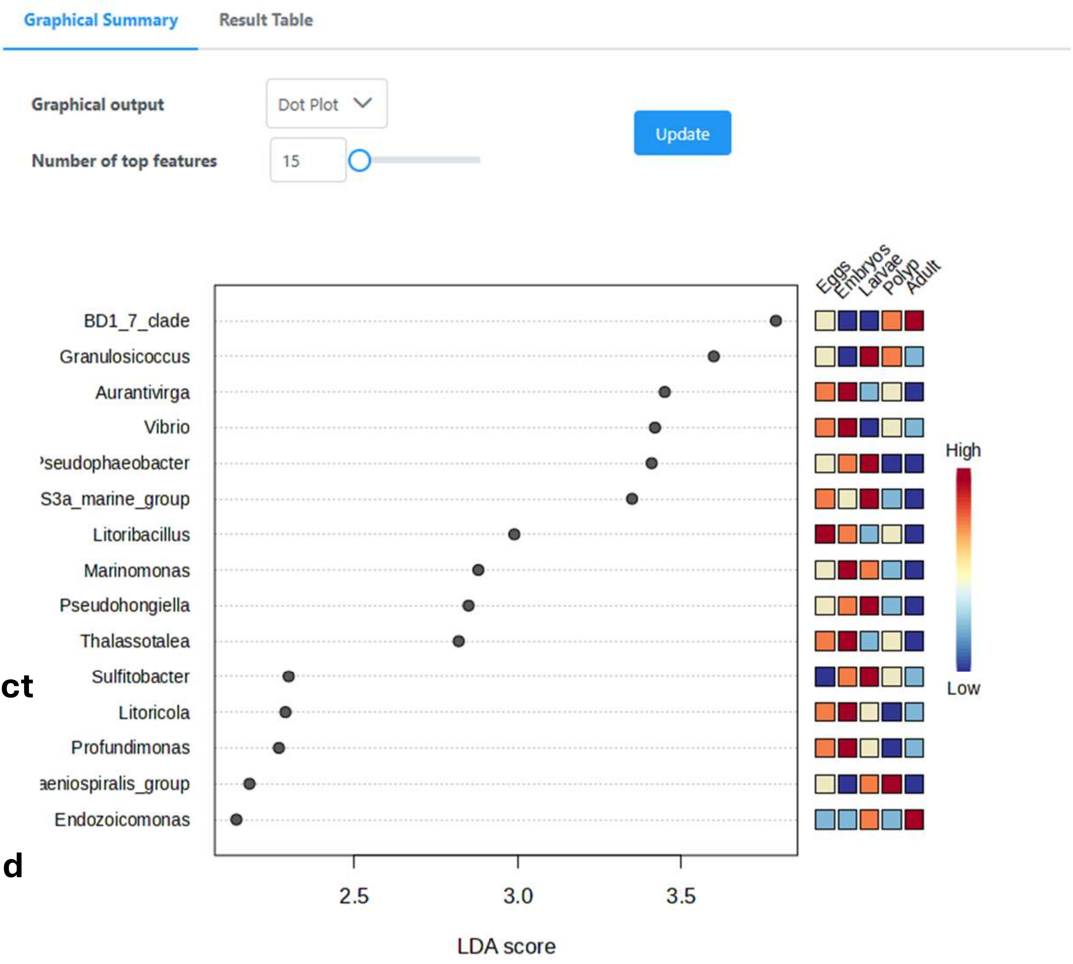
Linear Discriminant Analysis Effect Size (LEfSe)

Taxonomy level: Genus

Experimental factor: Dev_Stages

P-value cutoff: 0.1 ○ Original ● FDR-adjusted

Log LDA score: 2.0



Graphical Summary | Result Table

Graphical output: Dot Plot

Number of top features: 15

Update

- LEfSe focuses on identifying **microbes that can *discriminate* between groups**, with an emphasis on **effect size** and **consistency**.

- **Finds "who matters most"** — taxa that are not just statistically different but also **biologically meaningful and predictive**.

- **First tests for statistical differences for detecting differentially abundant features**, then **uses LDA (Linear Discriminant Analysis) to estimate effect size, which helps highlight potential biomarkers**.

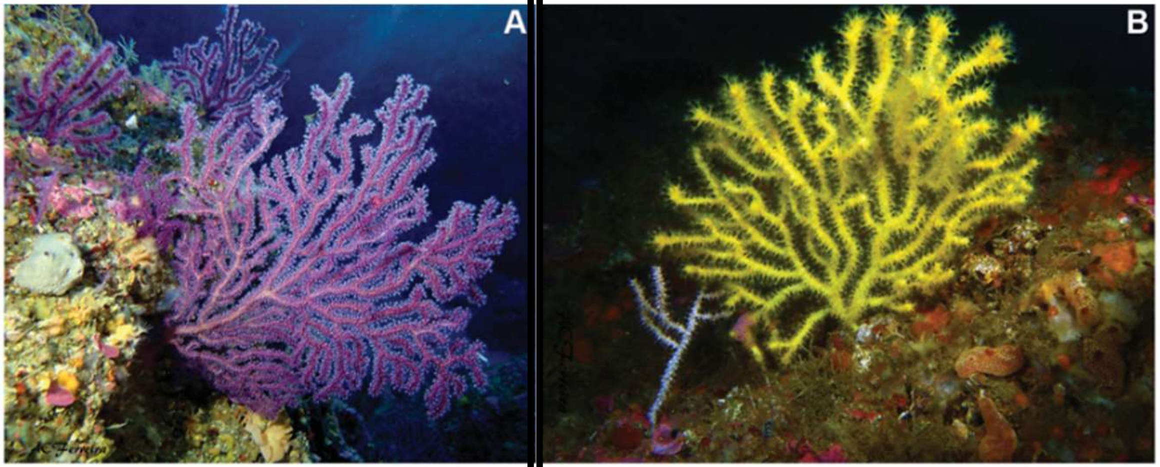
Our data set (Illumina NovaSeq 6000 sequencing data – V5-V7 region of the 16S)

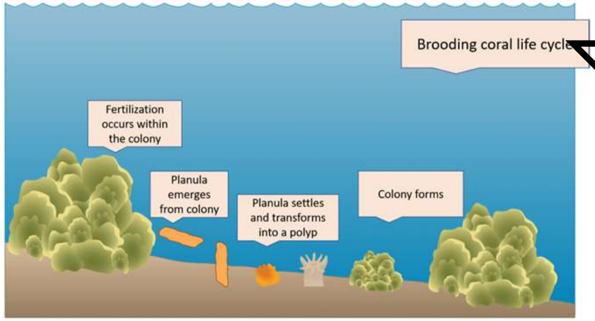
Paramuricea cf. grayi

Eunicella verrucosa

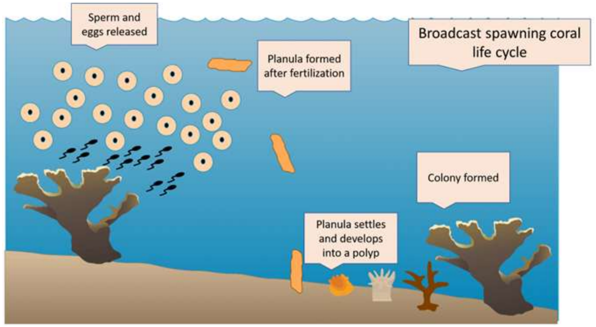
purple morphotype

yellow morphotype




Brooding coral life cycle
Fertilization occurs within the colony
Planula emerges from colony
Planula settles and transforms into a polyp
Colony forms

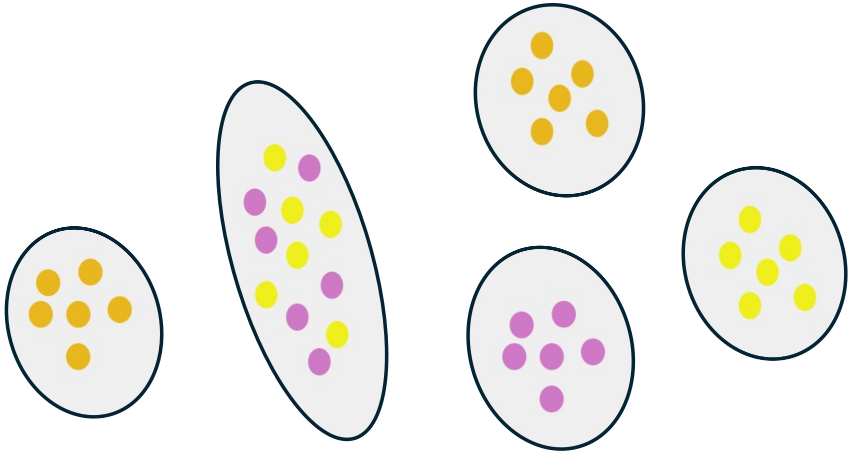
Surface brooder


Sperm and eggs released
Planula formed after fertilization
Broadcast spawning coral life cycle
Colony formed
Planula settles and develops into a polyp

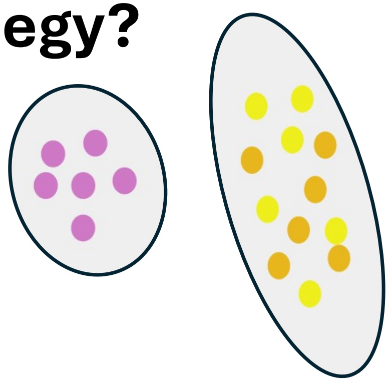
Broadcast spawners

Biological question

- Is the microbiome species-specific?



- Is the microbiome related to the reproductive strategy?



- Or is the microbiome shaped by both factors?