# Amplicon Sequencing Data Analysis with QIIME 2

Christian Diener, Gibbons Lab


from the **2021 ISB Virtual Microbiome Series**

CC-BY-SA gibbons.isbscience.org gibbons-lab @thaasophobia

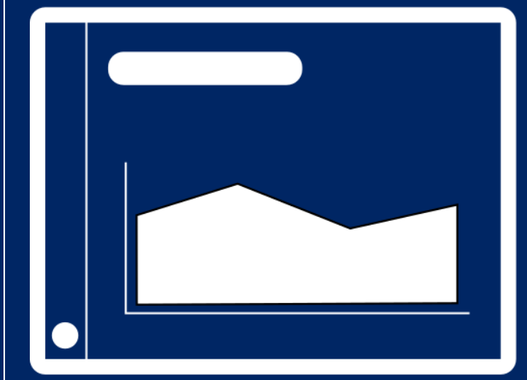
# Hold your horses 🐴

Let's get the slides first (use your computer, phone, TV, fridge, anything with a 16:9 screen)

**https://gibbons-lab.github.io/isb_course_2021/16S**

# Organization of the course

## Presentation

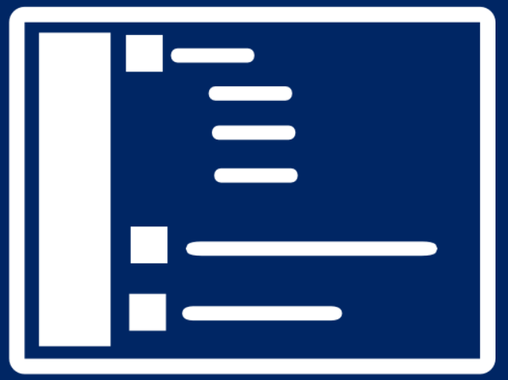


logic
explanations
links

## Notebook


technical aspects
materials
visualizations

## Chat



support
Q&A

# Setup

 Let’s switch to the notebook and get started

Click me to open the notebook!

## Wait... what?


All output we generate can be found in the `treasure_chest` folder at

https://github.com/gibbons-lab/isb_course_2021/treasure_chest

or `materials/treasure_chest` in the Colaboratory notebook.

# QIIME

Pronounced like wind **chime**.

Created ~2010 during the Human Microbiome Project (2007 - 2016) under the leadership of Greg Caporaso and Rob Knight.

# What is QIIME?

> *QIIME 2 is a powerful, extensible, and decentralized microbiome analysis package with a focus on data processing and analysis transparency.*

Quantitative Insights into Microbial Ecology

# So what is it really?

Essentially, QIIME is a set of commands to transform microbiome data into intermediate outputs and visualizations.

```
cdiener@moneta [ubc2018]
```

It's commonly used via the command line.

QIIME 2 was introduced in 2016 and improves upon QIIME 1, based on user experiences during the HMP.

Major changes:

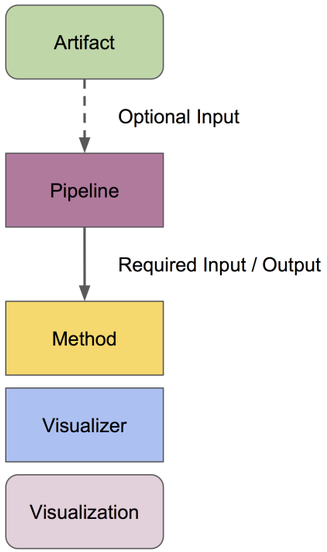
- integrated tracking of data provenance
- semantic type system
- extendable plugin system
- multiple user interfaces (in progress)

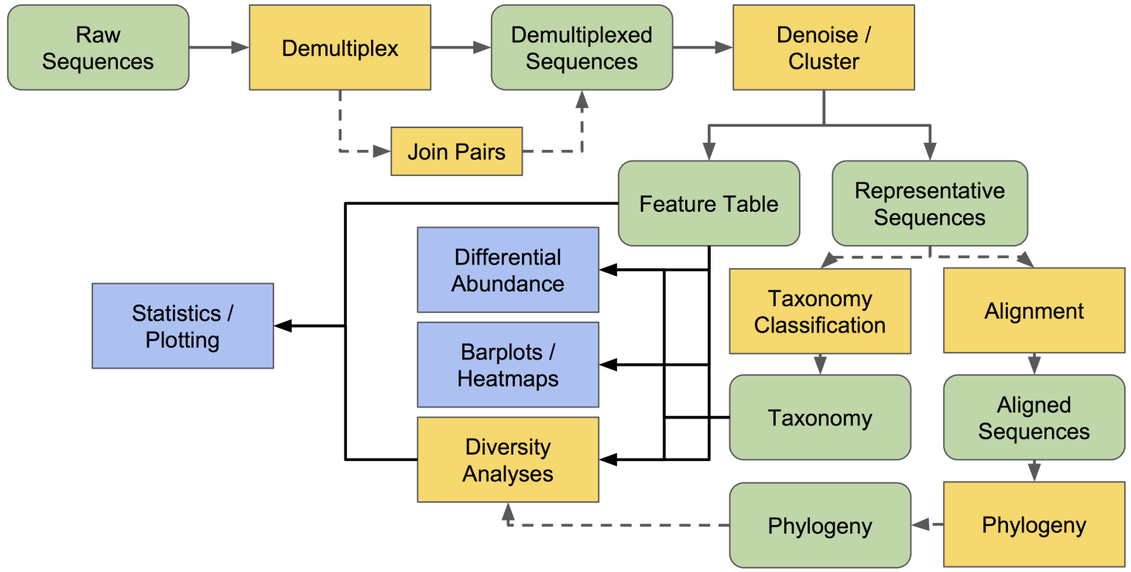
# Where to find help?

QIIME 2 comes with a lot of help, including a wide range of tutorials, general documentation and a user forum where you can ask questions.

# Artifacts, actions and visualizations

QIIME 2 manages artifacts, which are basically intermediate data that feed into actions to either produce other artifacts or visualizations.



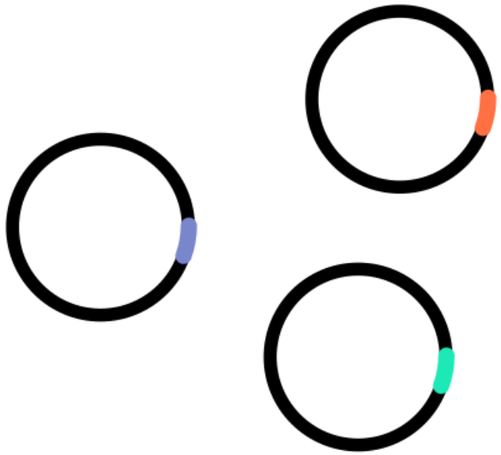

https://docs.qiime2.org/2021.8/tutorials/overview/

## Remember

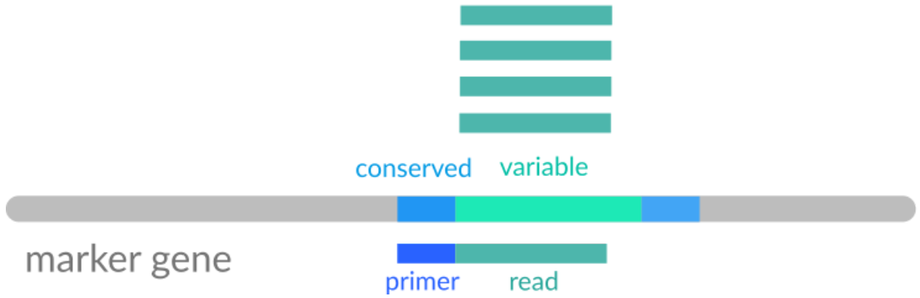
Artifacts often represent intermediate steps, but Visualizations are end points meant for human consumption ☝️ .

# What is amplicon sequencing?

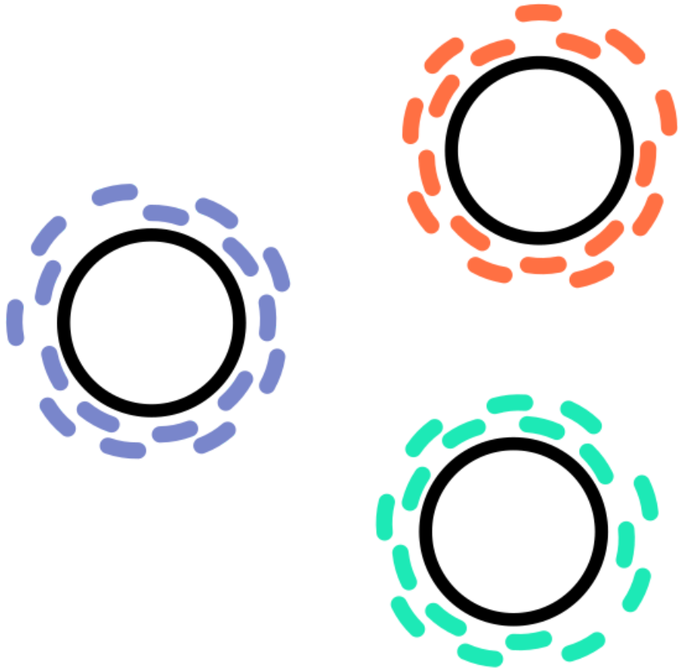
## amplicon sequencing



DNA extraction
PCR amplification
sequencing

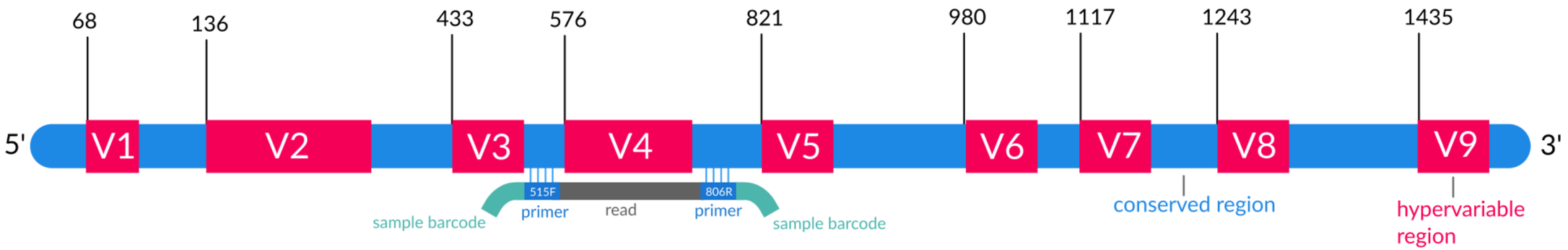



## shotgun metagenomics



DNA extraction
fragmentation
adapter ligation
sequencing

# Why the 16S gene?

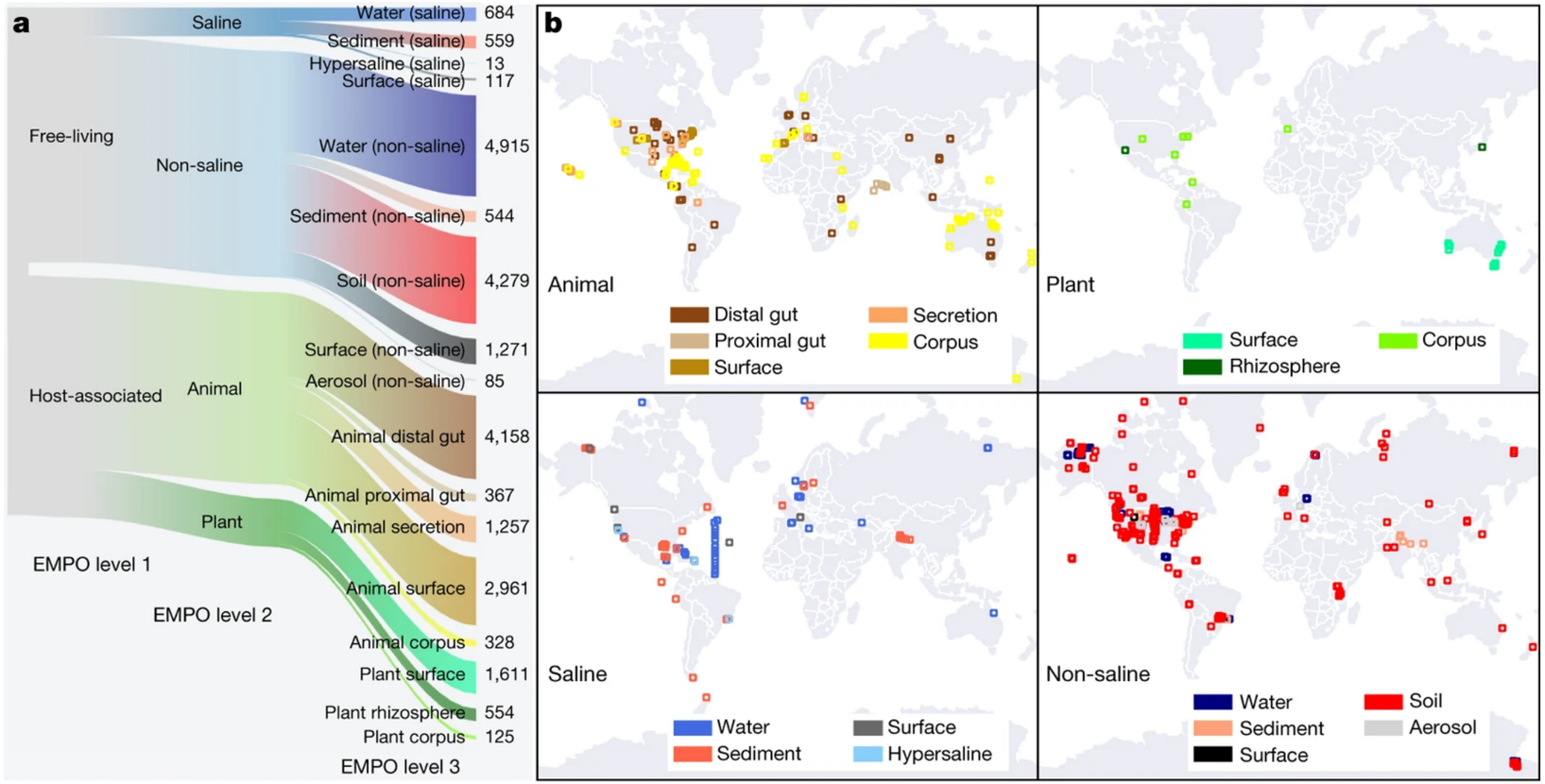



The 16S gene is universal and contains interspersed conserved regions perfect for PCR priming and hypervariable regions with phylogenetic heterogeneity.

# The Earth Microbiome Project

Photo by Nathan Jennings.

# A communal catalogue reveals Earth's multiscale microbial diversity




https://doi.org/10.1038/nature24621

# Our samples

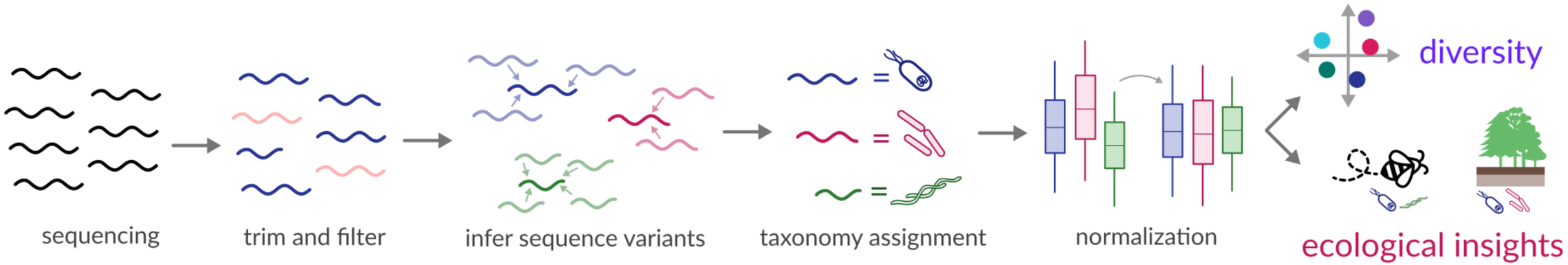




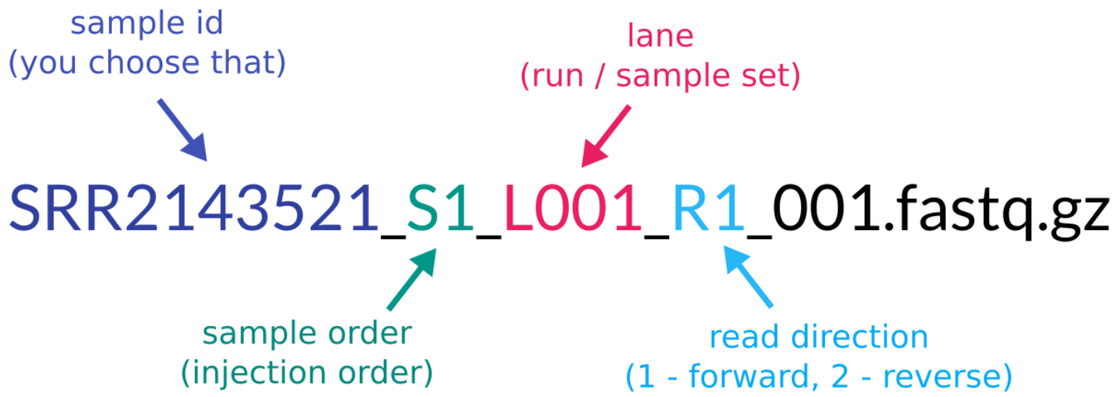
- 15 samples from 5 environments
- honey bee gut, cenotes in Yucatan (freshwater), ocean, human gut, soil

Photos by Dmitry Grigoriev, Jared Rice, Matt Hardy, Alex Block, and Roman Synkevich.

# What will we do today?



# Illumina FastQ files (Basespace)




```
@SRR2143527.13917 13917 length=251
TACGTAGGTGGCGAGCGTTATCCGGAATTATTGGGCGTAAA...
+
BBBBAF?A@D2BEEEGGGFGGGHGGGCFGFHHCFHCEFGGH...
```

We have our raw sequencing data, but QIIME 2 only operates on artifacts. How do we convert our data into an artifact??

 or  ?

# Our first QIIME 2 commands

 Let’s switch to the notebook and get started

# Time to bring out the big guns 💣 ⚡

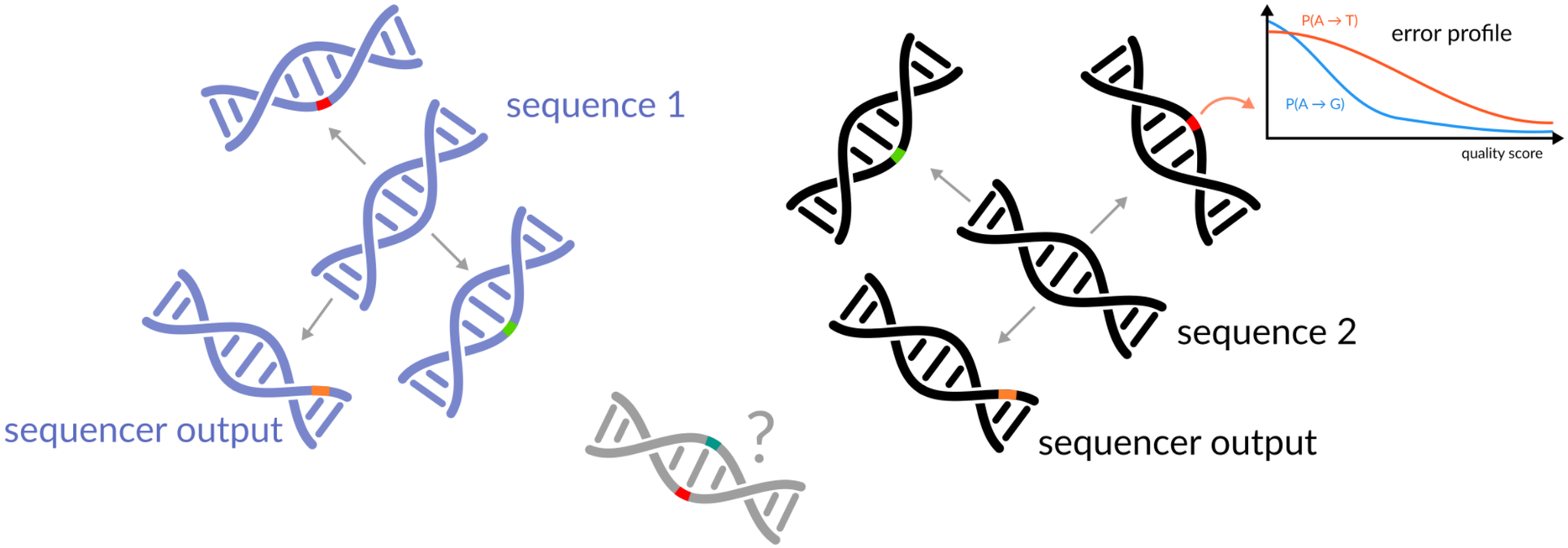
We will now run the DADA2 plugin, which will do 3 things:

1. filter and trim the reads
2. find the most likely original sequences in the sample (ASVs)
3. remove chimeras
4. count the abundances

# Preprocessing sequencing reads

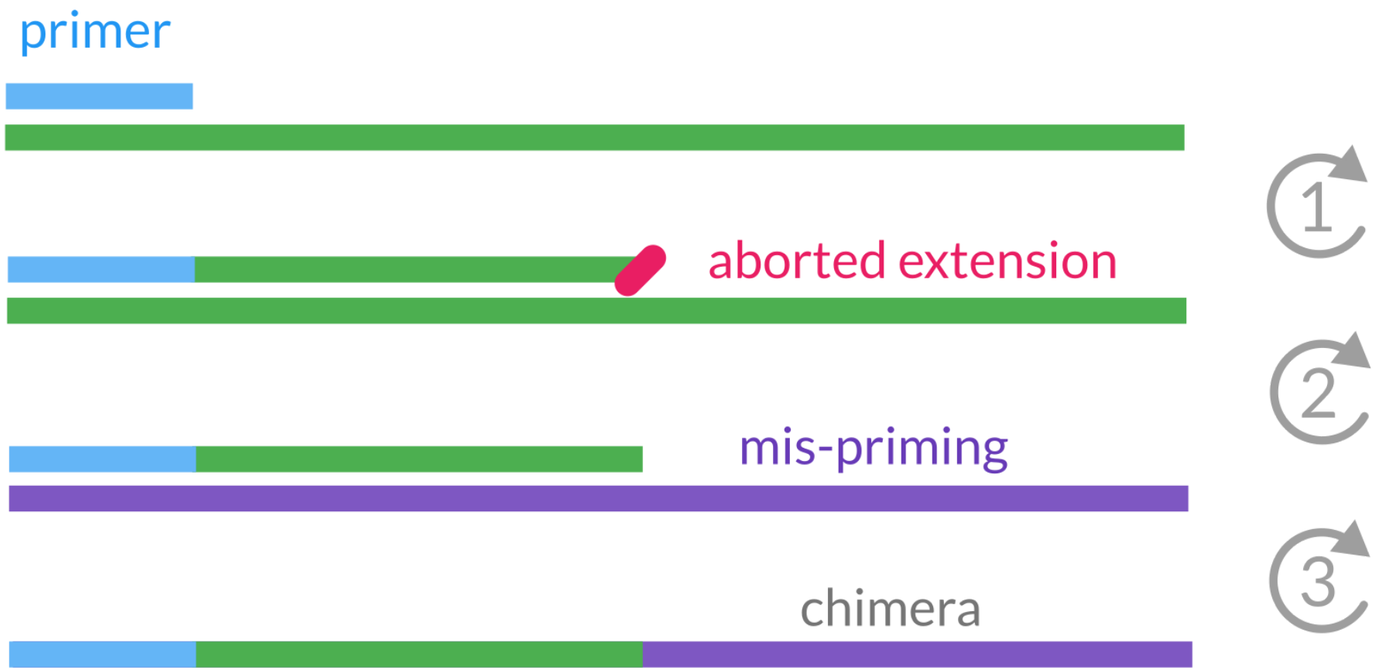
1. trim low quality regions
2. remove reads with low average quality
3. remove reads with ambiguous bases (Ns)
4. remove PhiX (bacteriophage genome commonly added as a control to sequencing runs)

# Identifying alternative sequence variants (ASVs)




Expectation-Maximization (EM) algorithm used to build a dataset-specific error model and find true amplicon sequence variants (ASVs), all at once.

# PCR chimeras




The primers used in this study were F515/R806. The numbers denote positions along the 16S gene. So, how long is the amplified fragment?

We now have a table containing the counts for each ASV in each sample. We also have a list of ASVs.

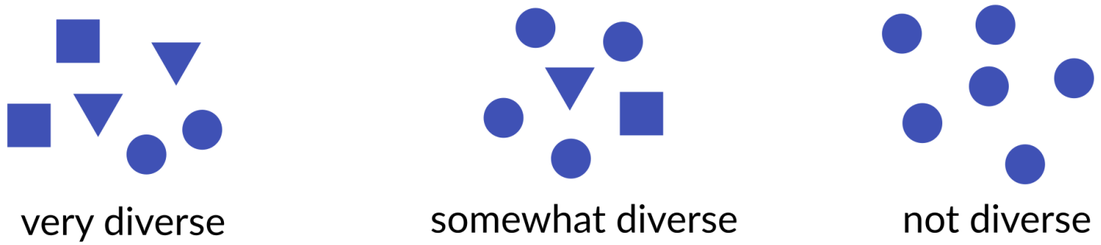
🤔 Do you have an idea for what we could do with these two data sets? What quantities might we be interested in?

# Diversity metrics

In microbial community analysis we are usually interested in two different families of diversity metrics, **alpha diversity** (ecological diversity within a sample) and **beta diversity** (ecological differences between samples).

# Alpha diversity

How diverse is a single sample?




- richness: how many taxa do we observe (richness)?
  → total number of observed taxa
- evenness: how evenly are abundances distributed across taxa?
  → Evenness index
- mixtures: metrics that combine both richness and evenness
  → Shannon index, Simpson's Index

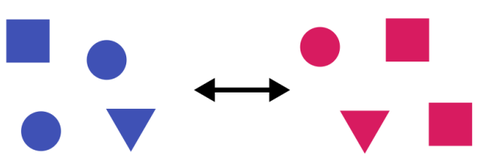
# Statistical tests for alpha diversity

Alpha diversity will provide a single value/covariate for each sample.

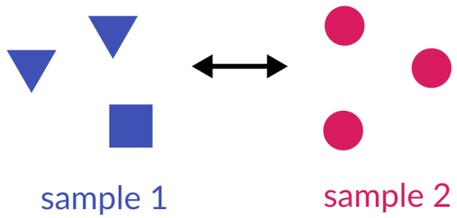
It can be treated as any other sample measurement and is suitable for classic univariate tests (t-test, Mann-Whitney U test).

# Beta diversity

How different are two or more samples/donors/sites from one another other?




very similar

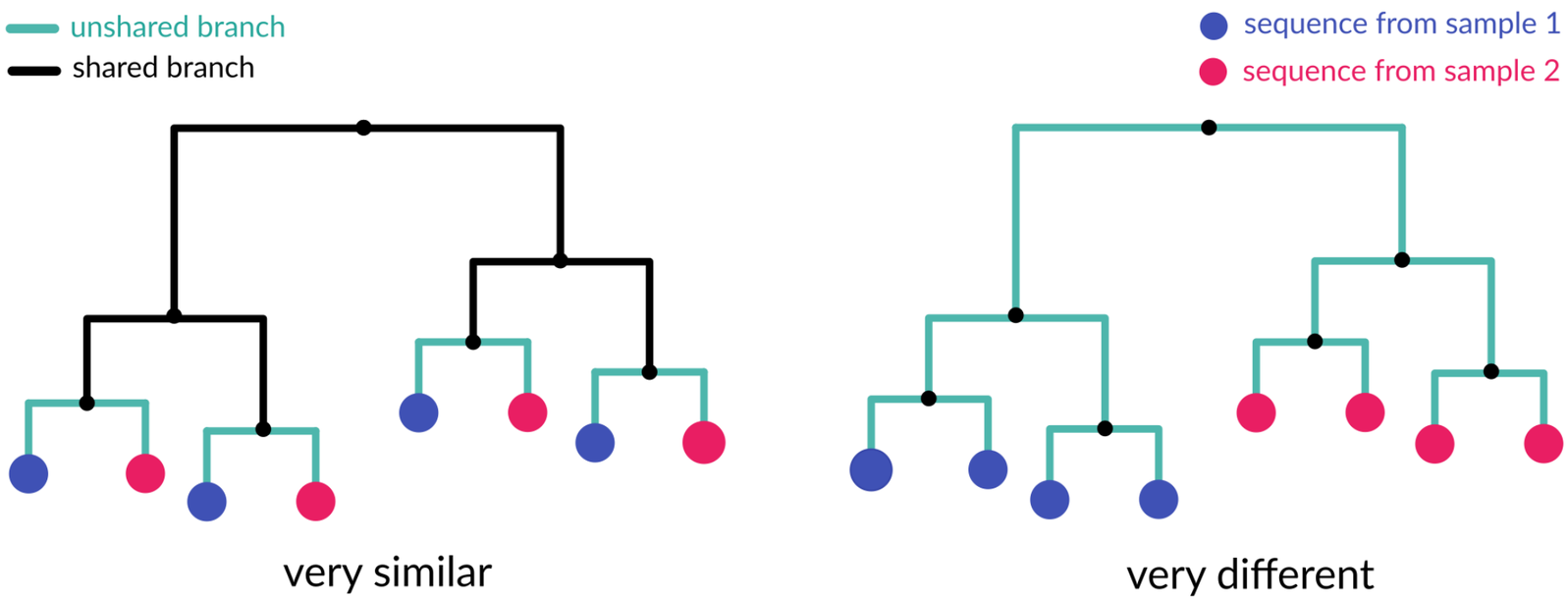



very different

- unweighted: how many taxa are shared between samples?
  → Jaccard index, unweighted UniFrac
- weighted: do shared taxa have similar abundances?
  → Bray-Curtis distance, weighted UniFrac

# UniFrac

Do samples share genetically similar taxa?



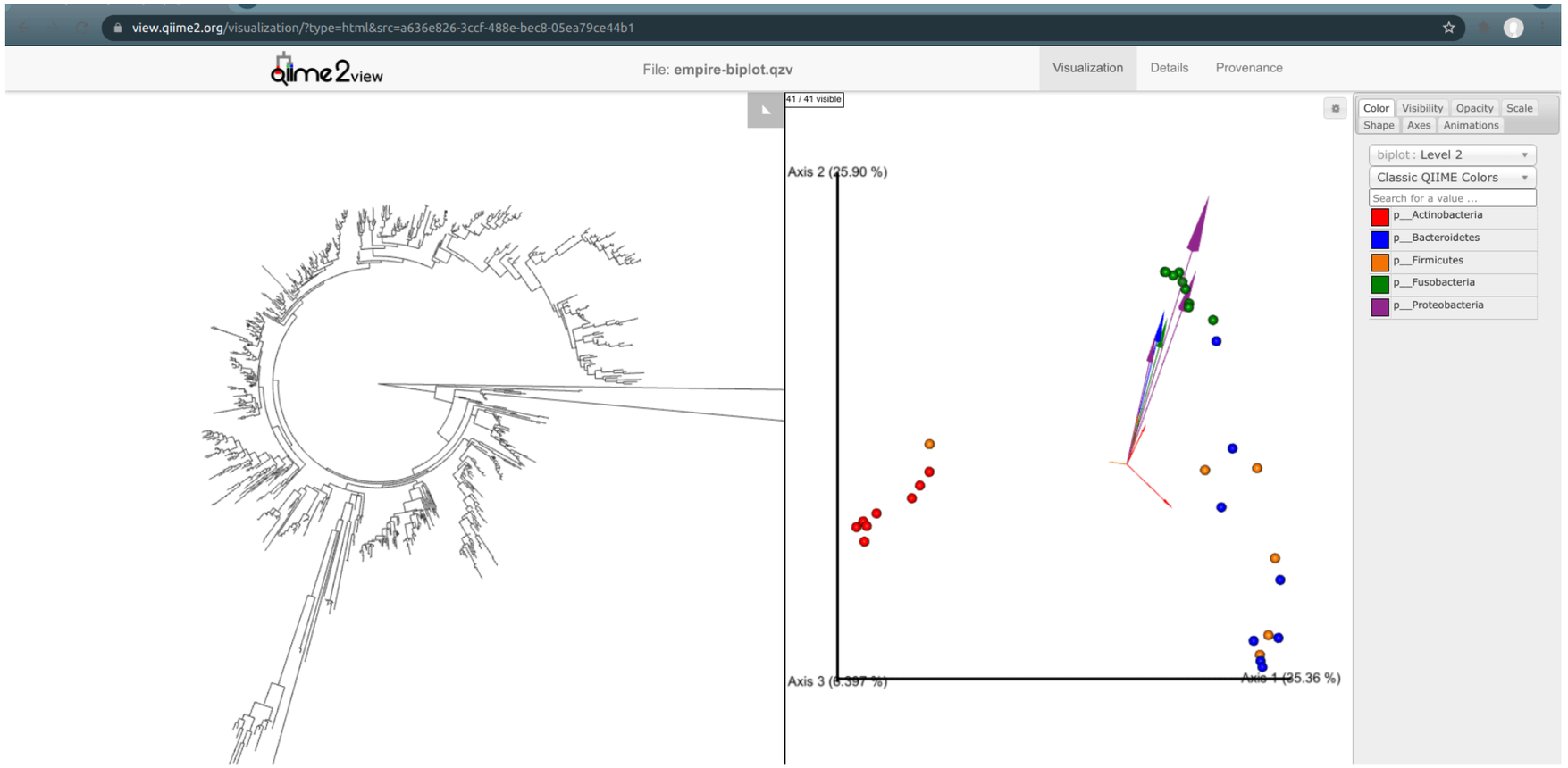

Weighted UniFrac further scales phylogenetic branch lengths by abundances.

# How to build a phylogenetic tree?

One of the basic things we might want to look at is how the ASVs across all samples are evolutionarily related to one another. That is, we are often interested in their phylogeny.

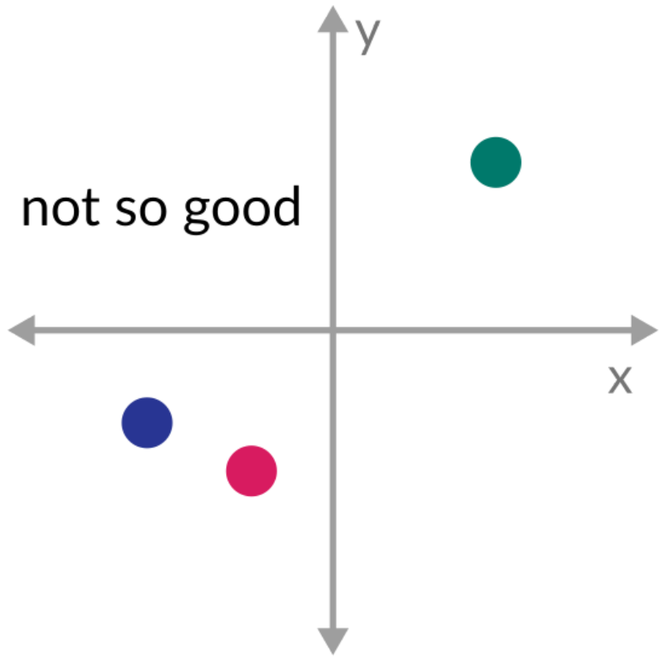
Phylogenetic trees are built from multiple sequence alignments and sequences are arranged by sequence similarity (branch length).

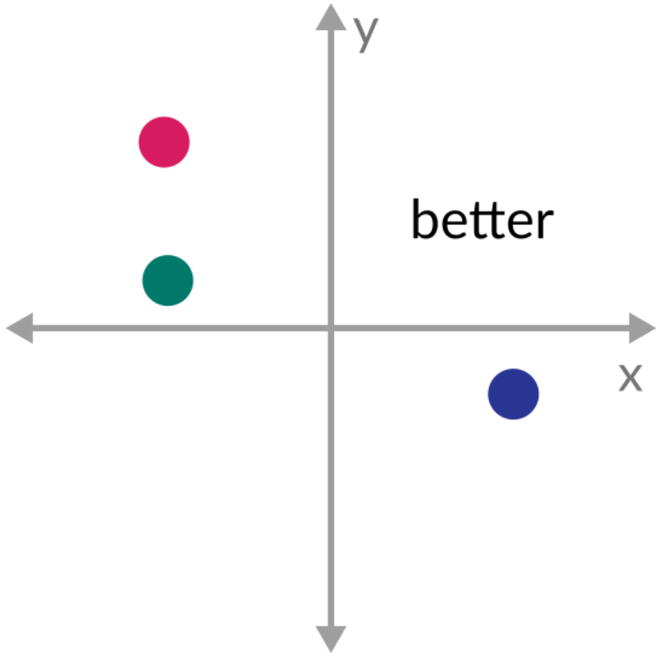
We can visualize this tree with EMPRESS.



# Principal Coordinate Analysis

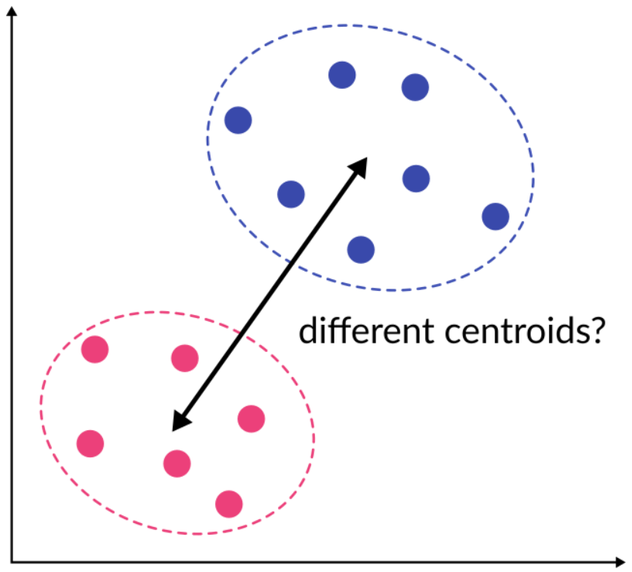
| β div. | sample 1 | sample 2 | sample 3 |
|---|---|---|---|
| sample 1 | 0 | | |
| sample 2 | 0.8 | 0 | |
| sample 3 | 0.2 | 0.1 | 0 |

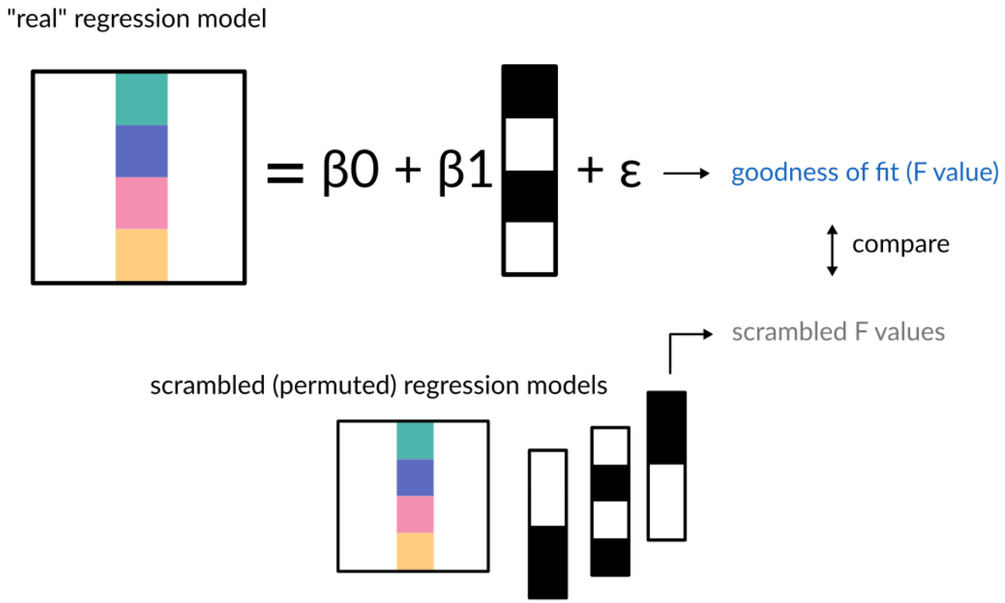


# Statistical tests for beta diversity

More complicated. Usually not normal and very heterogeneous. PERMANOVA can deal with that.



# Run the diversity analyses

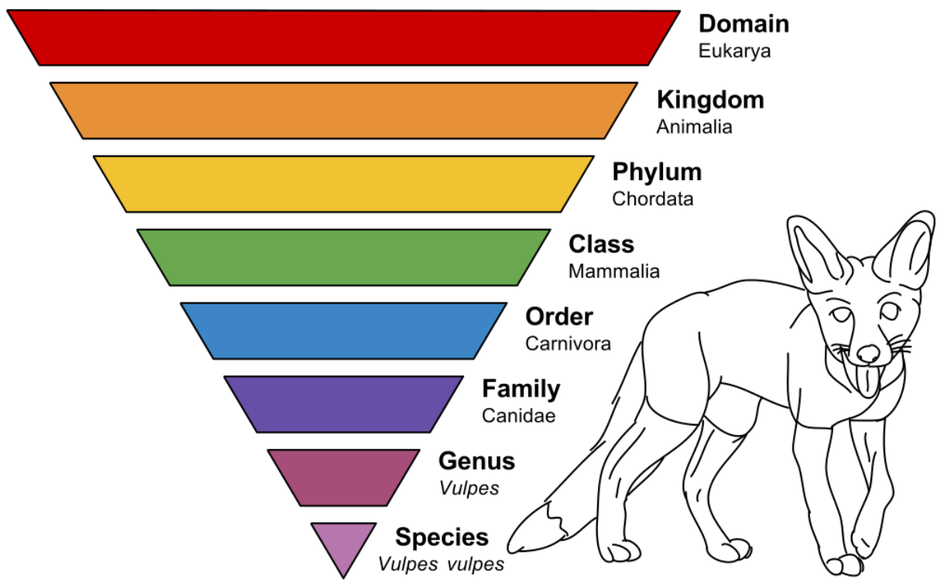
 Let’s switch to the notebook and calculate the diversity metrics

# What organisms are present in our samples?

We are still just working with sequences and we have no idea what **organisms** those sequences correspond to.

🤔 What would you do to go from a sequence to an organism's name?

# Taxonomic ranks




Red fox (*Vulpes vulpes*)

Even though directly aligning our sequences to a database of known genes seems most intuitive, this does not always work well in practice. Why?

# Multinomial Naive Bayes

query sequence

ACGCGC
ACG
CGC
GCG
CGC

reference model

| taxon 1 | taxon 2 | |
|---|---|---|
| *P*(taxon 1) = 0.2 | *P*(taxon 2) = 0.1 | prior |
| *P*(ACG) = 0.25 | *P*(ACG) = 0.4 | k-mer frequencies |
| *P*(CGC) = 0.25 | *P*(CGC) = 0.2 | |
| *P*(GCG) = 0.5 | *P*(GCG) = 0.4 | |

P(taxon 1 | query) ~ $0.2 \cdot 0.25 \cdot 0.25^2 \cdot 0.5 = 0.0016$

P(taxon 2 | query) ~ $0.1 \cdot 0.4 \cdot 0.2^2 \cdot 0.4 = 0.0006$

choose highest taxon

methods differ here

$$\mathbb{P}(t|q) = \frac{\mathbb{P}(t) \cdot \mathbb{P}(q|t)}{\mathbb{P}(q)}$$

Instead, use subsequences (k-mers) and their counts to predict the lineage/taxonomy with machine learning methods. For 16S amplicon fragments, this approach often provides better generalization and faster results.

# Let’s assign taxonomy to our samples

 Let’s switch to the notebook and assign taxonomy to our ASVs

# Your turn

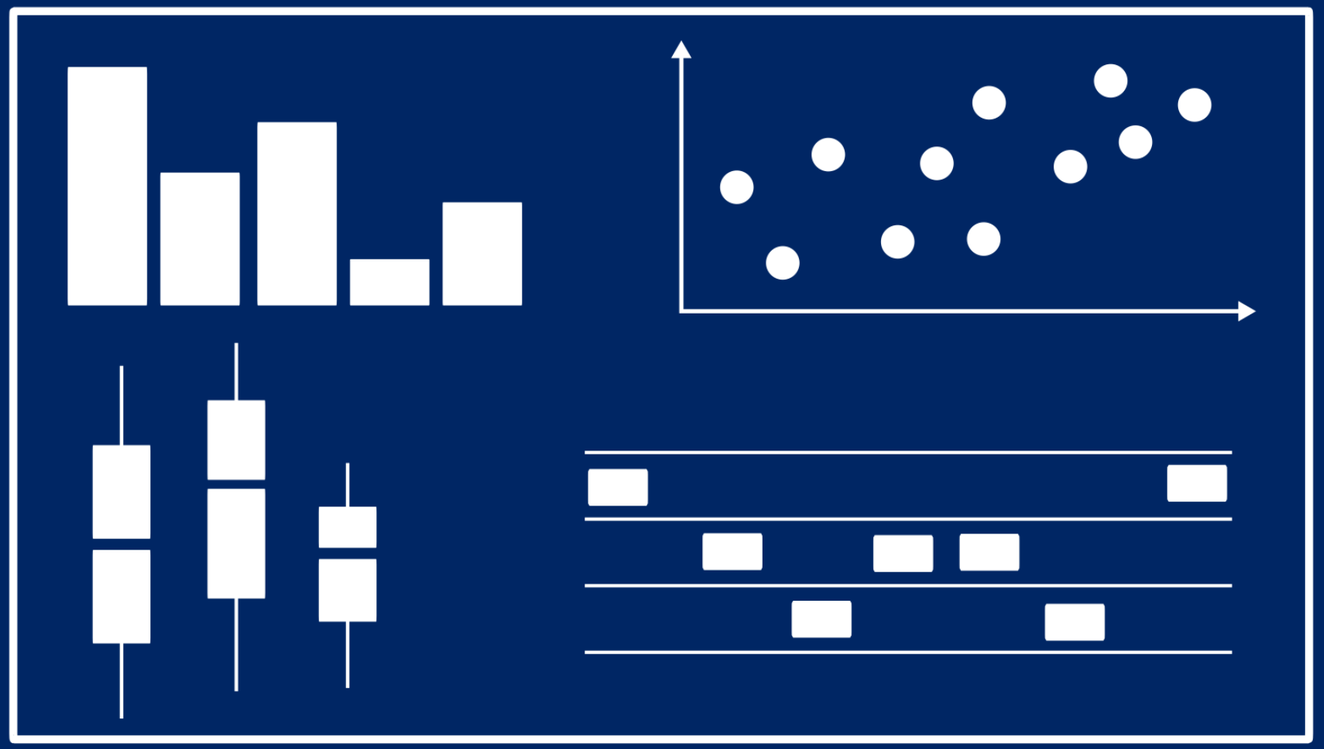
Are certain taxa only found in one environment? Are others more widely distributed?

# The project challenge 🦠

**Figure**

only uses Qiime 2 visualizations

4 panels max.



**Post (text)**

indicate region

provide a caption in post

**Region: South Africa**
**Caption:** Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod nulla non urna egestas, in mattis ipsum imperdiet. Fusce lorem sem, varius eget bibendum id, interdum et libero. Fusce volutpat sem in maximus ultrices. Pellentesque finibus lacinia nulla, eget lacinia justo maximus in.

Submit your figure here.

And we are done 👏

**ISB team**

Joey Petosa
Allison Kudla
Christian Diener
Sean Gibbons
Priyanka Baloni
Tomasz Wilmanski
Noa Rappaport
Alex Carr
Audri Hubbard
Renee Duprel
Joe Myxter
Thea Swanson

Thanks!