# Amplicon Sequencing Data Analysis with QIIME 2

Nick Quinn-Bohmann, Gibbons Lab


from the **ISB Microbiome Course 2023**

 gibbons.isbscience.org gibbons-lab @BioBohmann

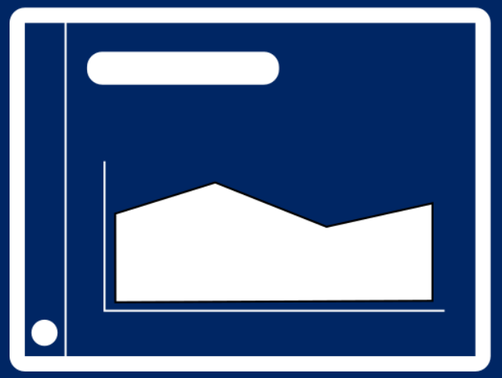
# Let’s Start Learning 

First, we’ll need the slides, full of digestible information

https://gibbons-lab.github.io/isb_course_2023/16S

# Organization of the course

## Presentation

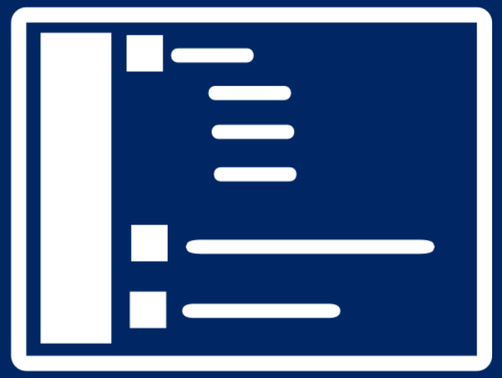


logic
explanations
links

## Notebook


technical aspects
materials
visualizations

## Chat



support
Q&A

# Guts, Camera, Action 🎥

📖 Let’s switch to the notebook and get started

Click me to open the notebook!

# In Case you Get Lost 🧭

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

https://github.com/Gibbons-Lab/isb_course_2023/tree/main/treasure_chest

or `materials/treasure_chest` in the Colaboratory notebook.

# The Gut Microbiome

- 30-40 trillion bacterial cells
- Heterogenous between individuals
- Helps digest food and produces metabolites
- Affects our health

# Measuring Your Microbes

How do we see what is in the microbiome?

- Hundreds to Thousands of taxa in each person
- Difficult to culture outside the resident environment
- We can sequence their DNA instead

# QIIME

Pronounced like **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

# What can we do with QIIME?

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

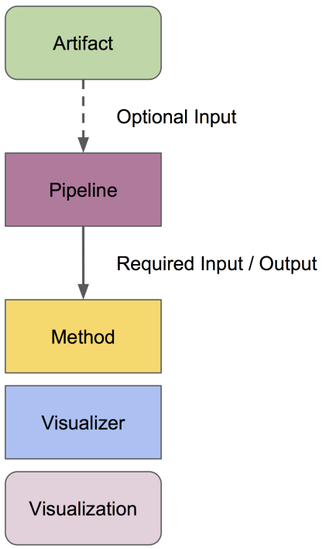

It's commonly used via the command line. We'll use it within the Colab Notebook

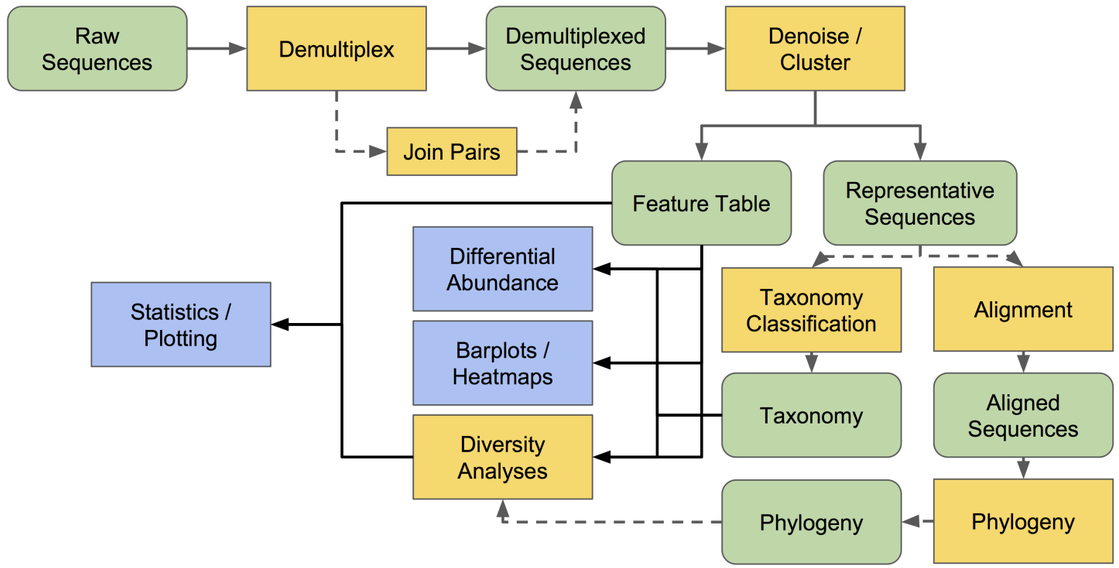
# Where to find help?

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/2023.7/tutorials/overview/

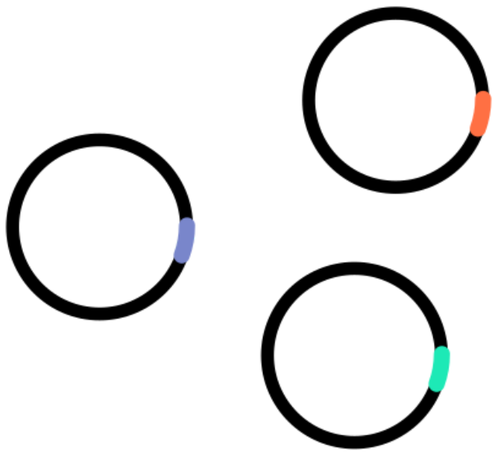
# Remember

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

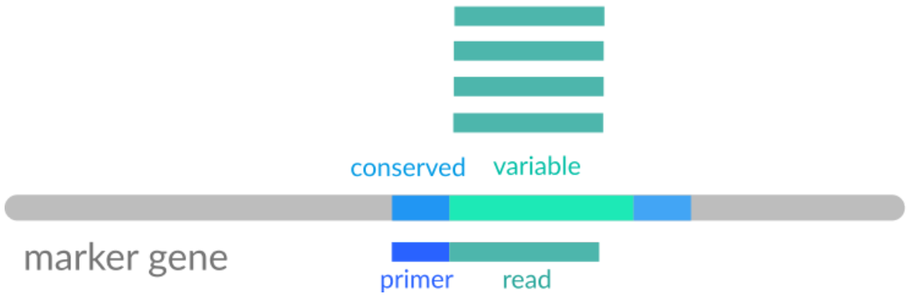
Visualizations cannot be used as inputs for additional commands

# What is amplicon sequencing?

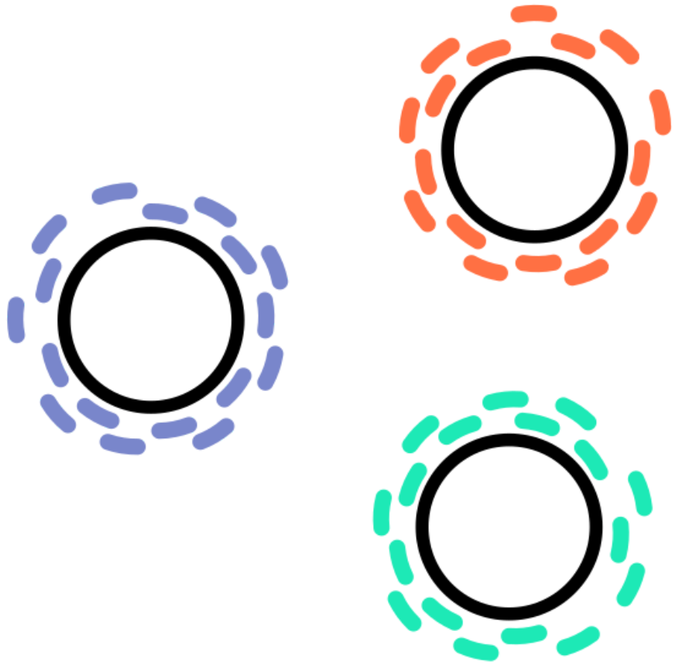
## amplicon sequencing



DNA extraction
PCR amplification
sequencing

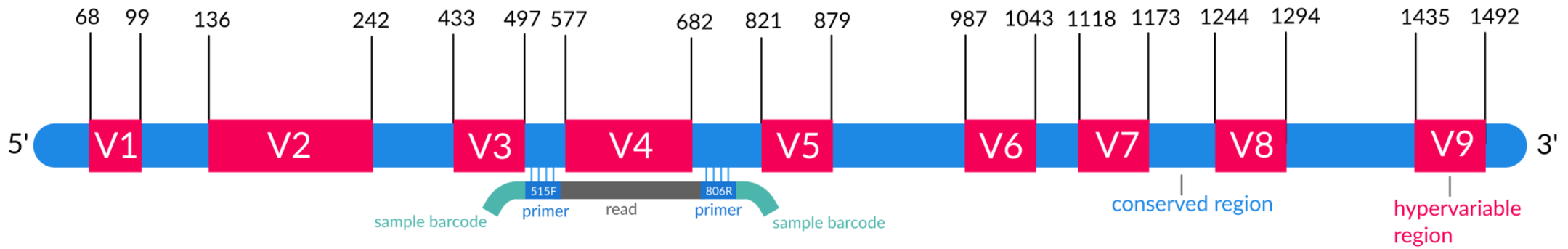



## shotgun metagenomics



DNA extraction
fragmentation
adapter ligation
sequencing

# Why the 16S gene?




https://dx.doi.org/10.1016%2Fj.mimet.2007.02.005

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

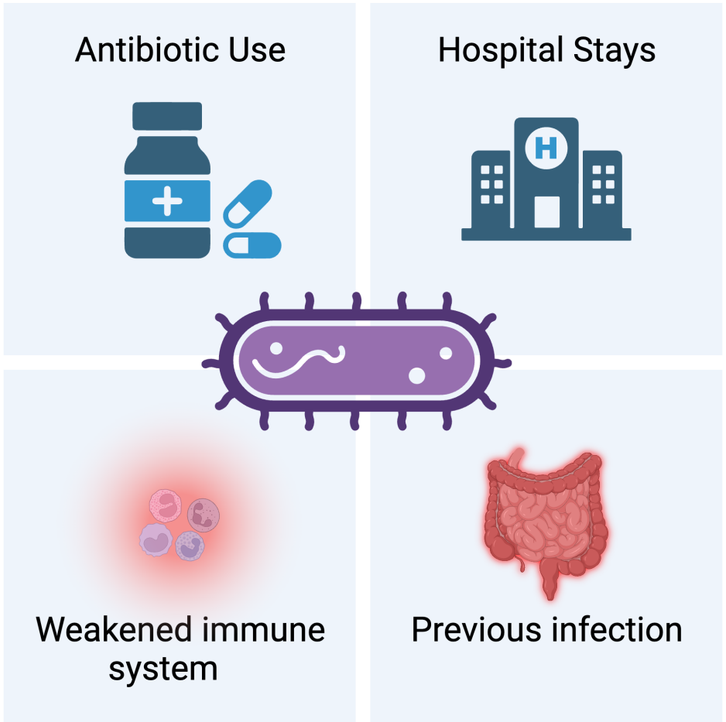
# Analyzing gut microbial composition during recurrent *C. diff* infections

16S amplicon sequencing data of the V4 region from human fecal samples

4 healthy donors and 4 individuals with recurrent infection.

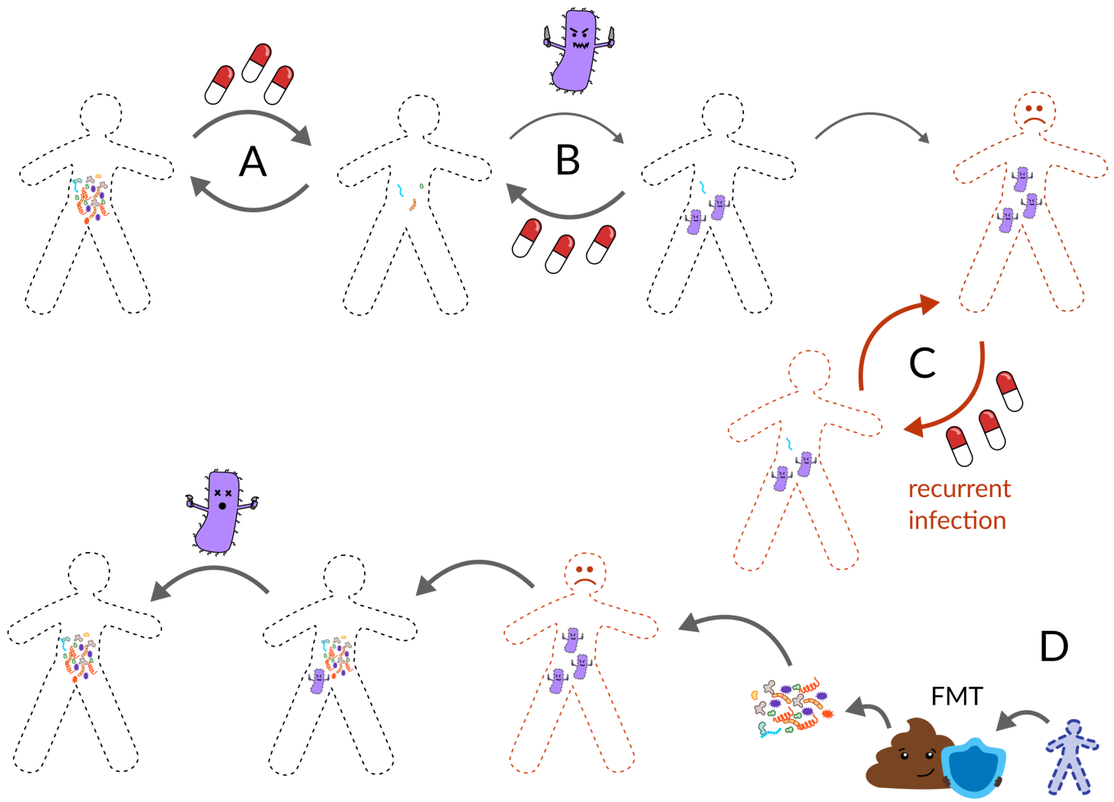
https://doi.org/10.1186/s40168-015-0070-0

# *C. difficile* Risk Factors



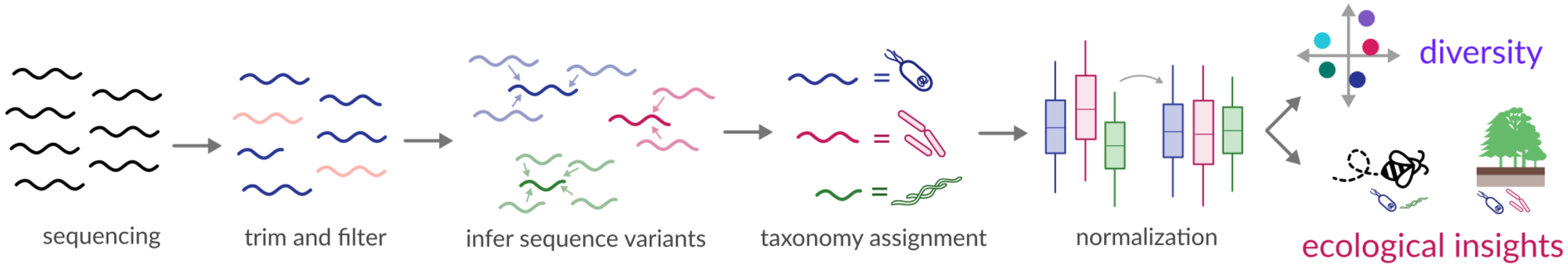

created with BioRender.com

# The *C. diff* infection cycle




courtesy of Stephanie Swegle

# What will we do today?



# Artifact Hunting

To start, we'll import our raw data into QIIME as an artifact.

 Let's switch to the notebook and get started

# Denoising with DADA 2

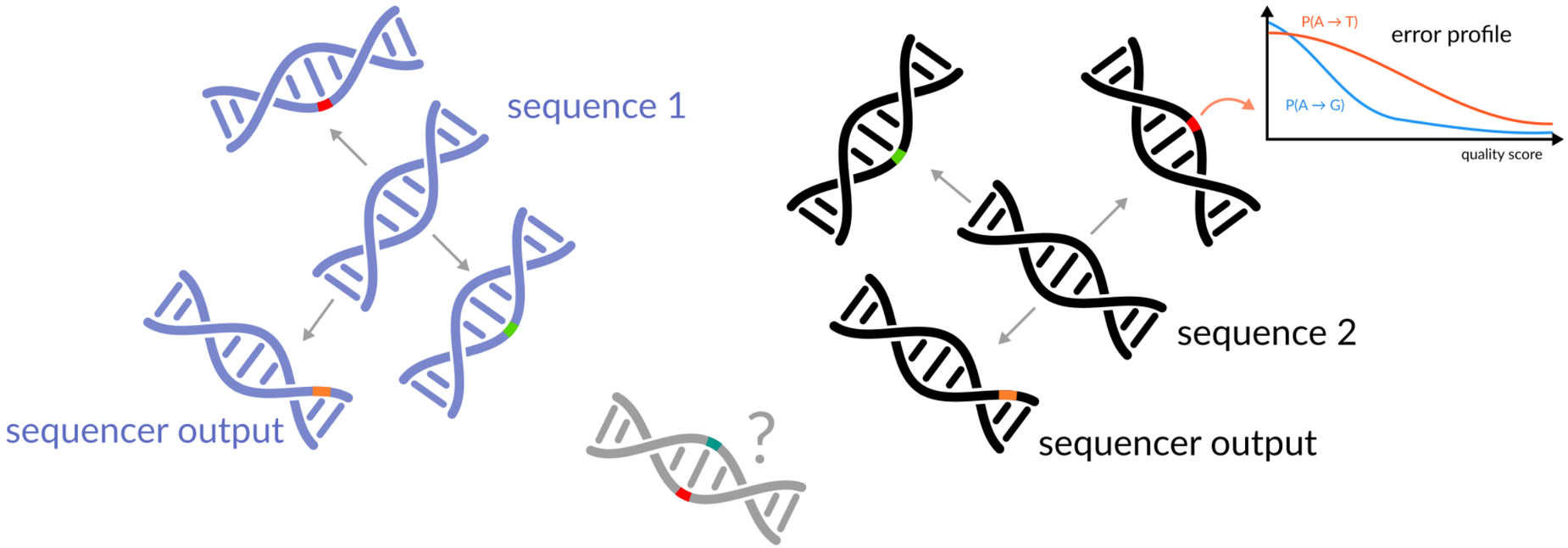
We just ran the DADA2 plugin for QIIME, which is doing 4 things:

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

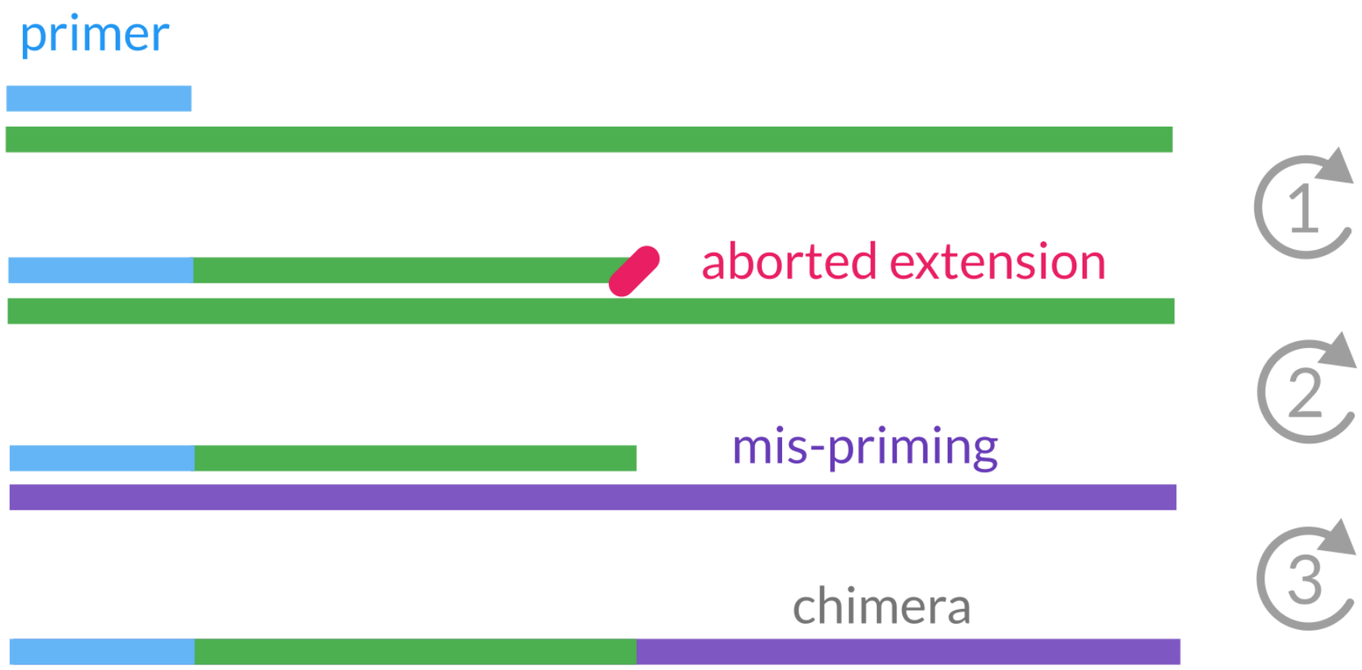
1. trim low quality regions
2. remove reads with low average quality
3. remove reads with ambiguous bases (Ns)
4. remove PhiX (added to sequencing)

# Identifying amplicon sequence variants (ASVs)




Expectation-Maximization (EM) algorithm to find amplicon sequence variants (ASVs) and the real error model at the same time.

# PCR chimeras




The primers used in this study were F515/R806. 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 those 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)?
  → #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

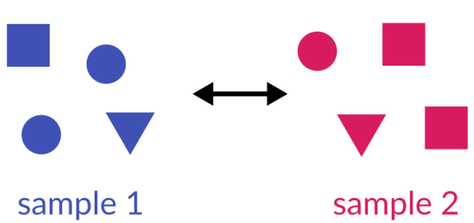
# Statistical tests for alpha diversity

Alpha diversity will provide a single value for each sample.

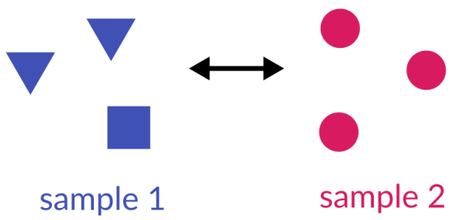
It can be treated as any other sample measurement and is suitable for classic univariate tests (t-test, Mann-Whitney U test).

# Beta diversity

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




very similar

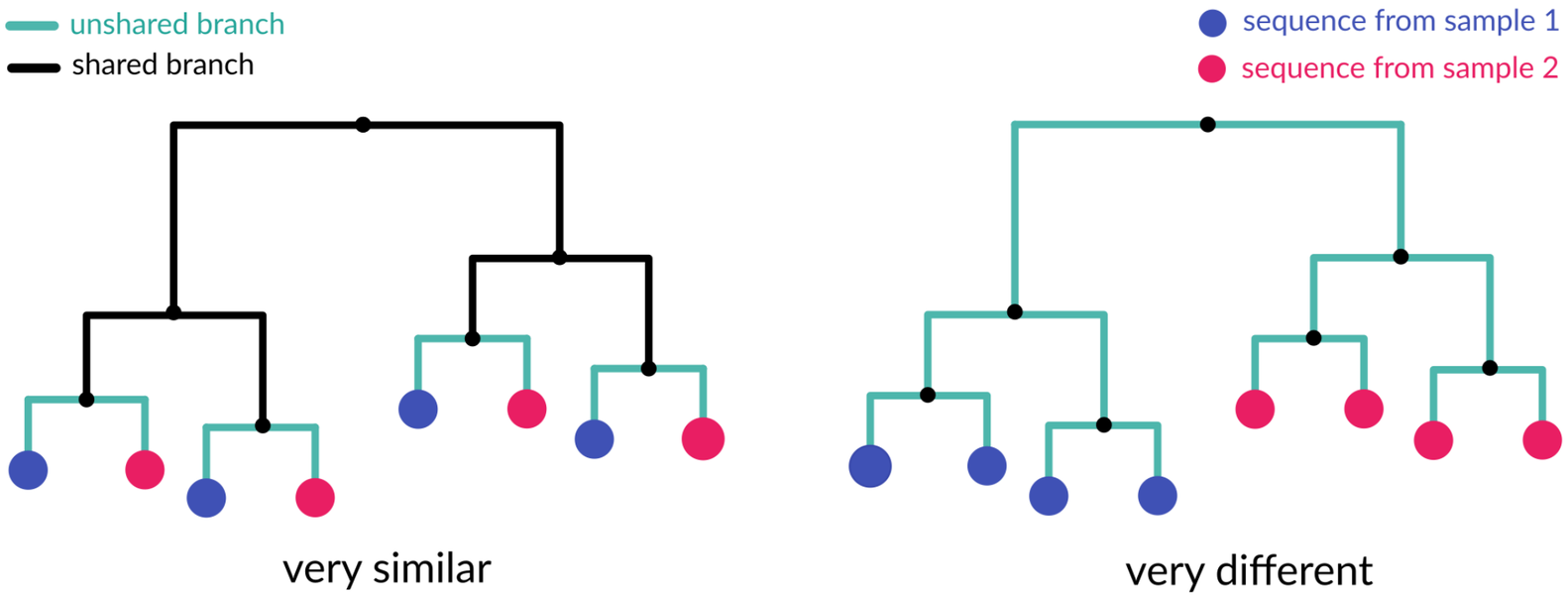



very different

- 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 scales branches by abundance.

# How to build a phylogenetic tree?

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

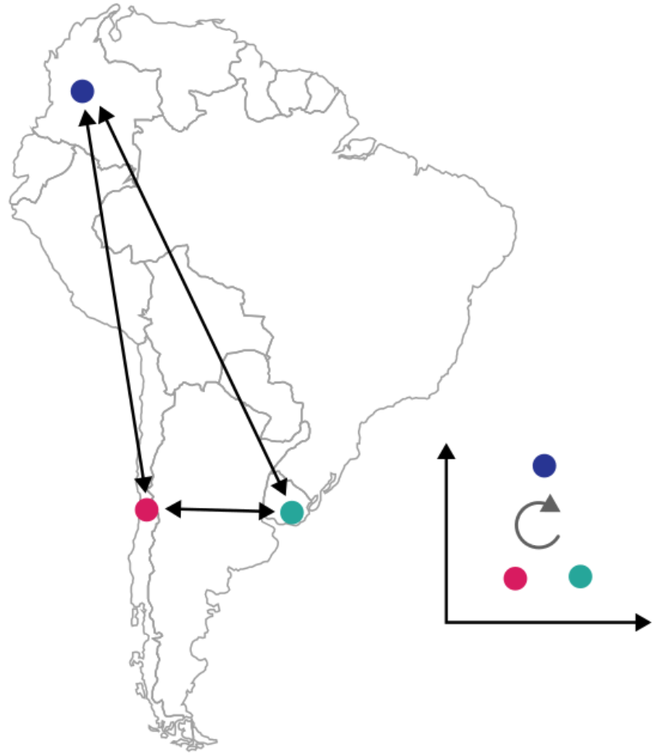
Phylogenetic trees are built from multiple sequence alignments and sequences are arranged by sequence similarity (branch length).

# Principal Coordinate Analysis

| | Bogotá | Santiago | Montevideo |
|---|---|---|---|
| Bogotá | 0 | | |
| Santiago | 0.7 | 0 | |
| Montevideo | 0.5 | 0.2 | 0 |

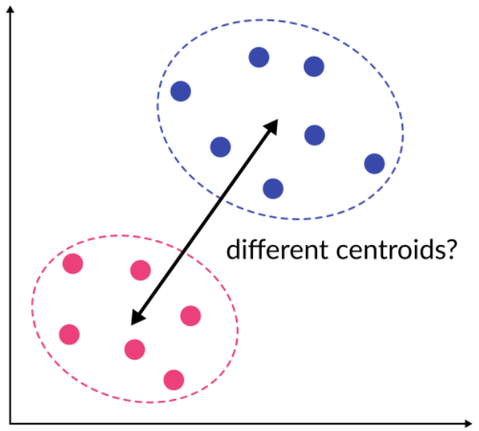
β div.

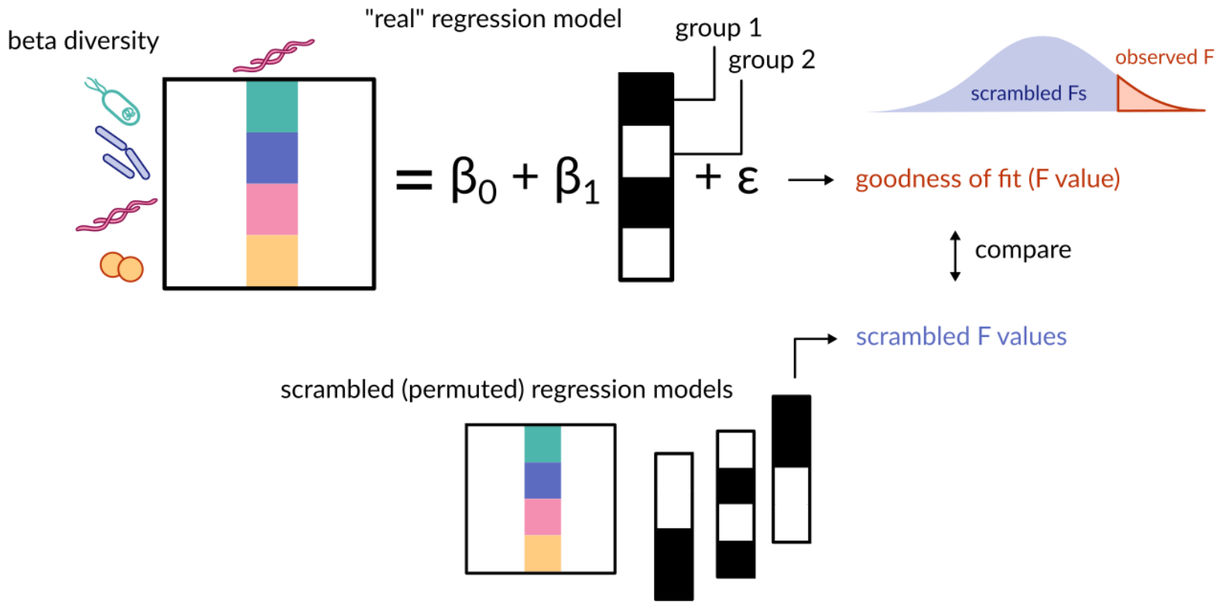
consistent representation(s)
many different projections
reduce dimensions



# Statistical tests for beta diversity

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



# Run the diversity analyses

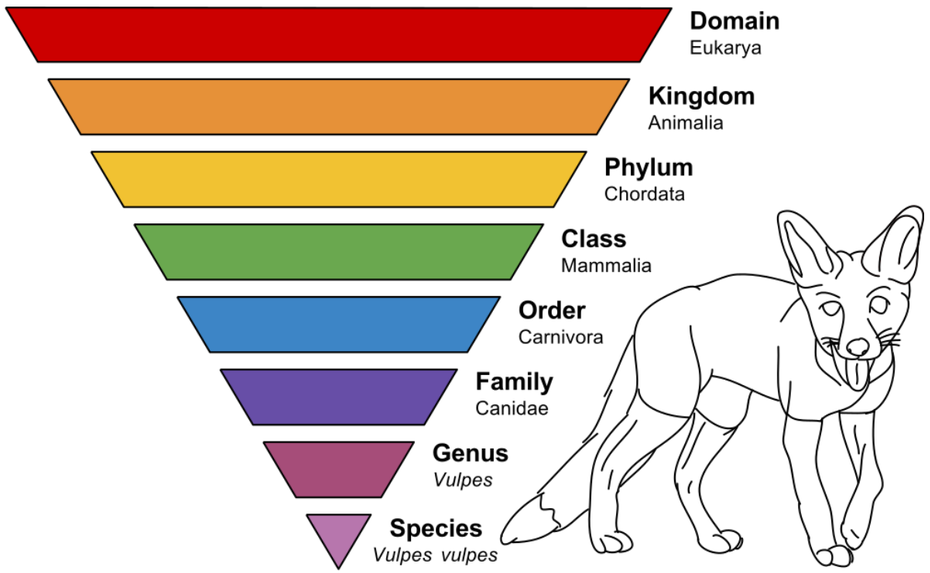
 Let’s switch to the notebook and calculate the diversity metrics

# But what organisms are there in our sample?

We are still just working with sequences and have no idea what **organisms** they 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

what we know before sequencing

| 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 | |

probability of taxon 1 given the query

probability of the query given it's taxon 1

1 ACG *and* 2 CGC *and* 1 GCG

$$P(\text{taxon 1} \mid \text{query}) \sim 0.2 \cdot 0.25 \cdot 0.25^2 \cdot 0.5 = 0.0016$$

probability of taxon 1

$$P(\text{taxon 2} \mid \text{query}) \sim 0.1 \cdot 0.4 \cdot 0.2^2 \cdot 0.4 = 0.0006$$

choose highest taxon

methods differ here

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

we usually ignore this

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

# Let’s assign taxonomy to the sequences

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

# Your turn

Which taxa are associated with the disease state?

And we are done 👏

Alex Carr
Sean Gibbons
Alyssa Easton
Katherine Ramos Sarmiento
Noa Rappaport
Karl Gaisser
Chloe Herman
Greg Caporaso
Christian Diener
Dominic Lewis
Allison Kudla
Audri Hubbard
Joe Myxter
Thea Swanson
Victoria Uhl
Connor Kelly
Shanna Braga
ISB Facilities Team

# Thanks! ❤️