Amplicon Sequencing Data Analysis with QIIME 2

Christian Diener, Gibbons Lab

from the **ISB Microbiome Course 2020**

 gibbons.isbscience.org gibbons-lab @thaasophobia

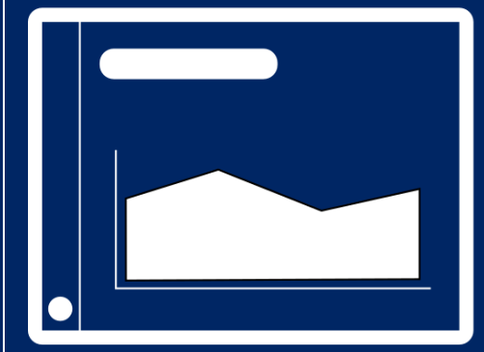
Hold your horses 

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_2020/16S

Organization of the course

Presentation

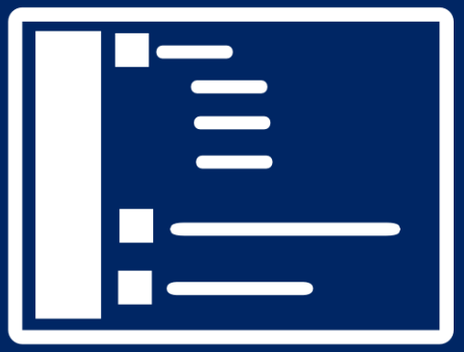


logic
explanations
links

Notebook

technical aspects
materials
visualizations

Chat



support
Q&A

Click me to open the notebook!

Setup

 Let’s switch to the notebook and get started

Wait... what?

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

https://github.com/gibbons-lab/isb_course_2020/treasure_chest

or `materials/treasure_chest` in the Colaboratory notebook.

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

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:

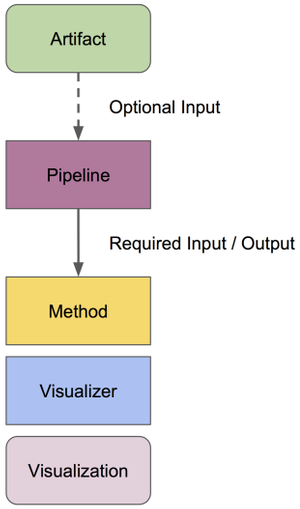
- integrated tracking of data provenance
- semantic type system
- extendable plugin system
- multiple user interfaces (in progress)

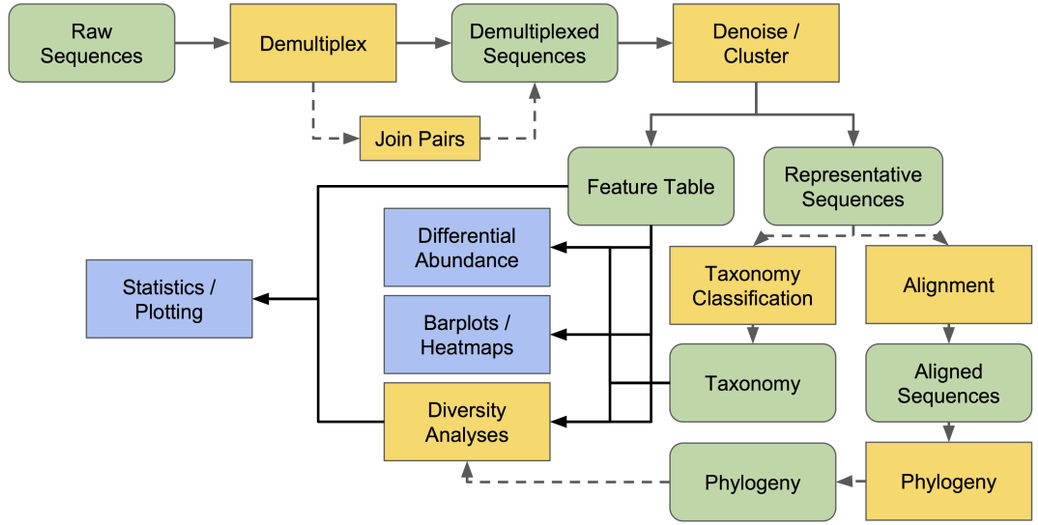
Where to find help?

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.





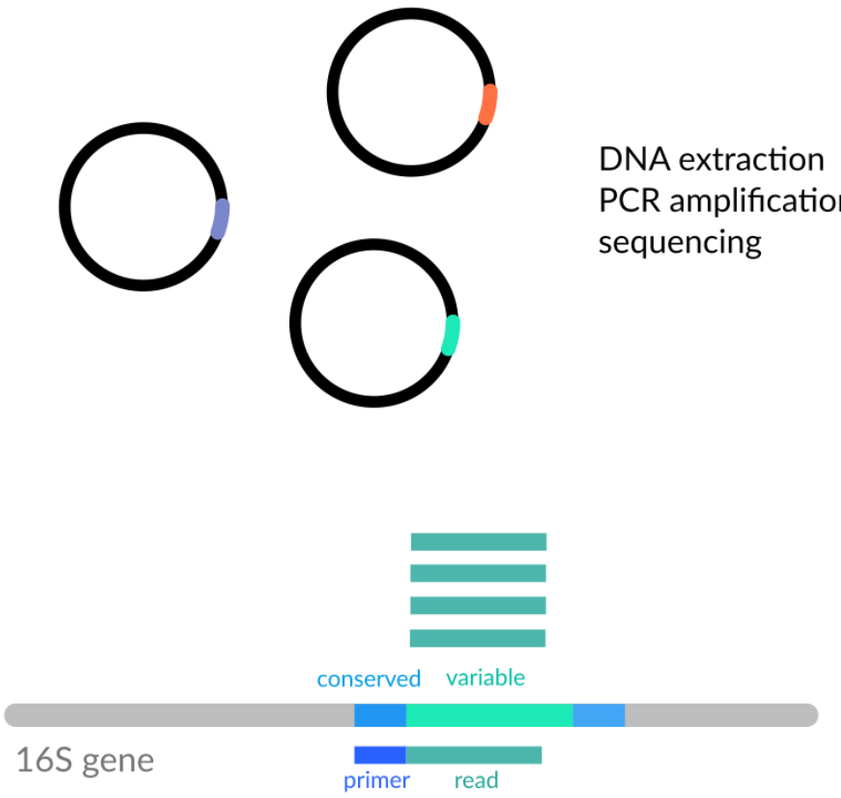
https://docs.qiime2.org/2020.8/tutorials/overview/

Remember

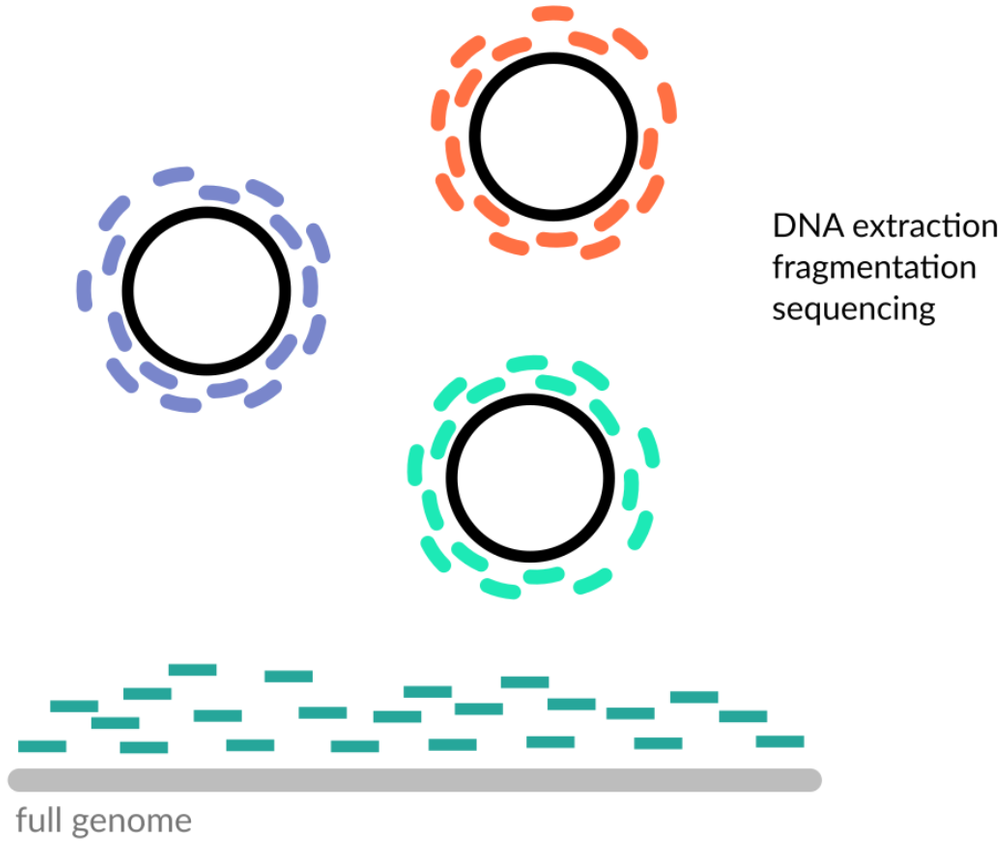
Artifacts often represent intermediate steps, but Visualizations are end points meant for human consumption ☝ .

What is amplicon sequencing?

16S amplicon sequencing



shotgun metagenomics



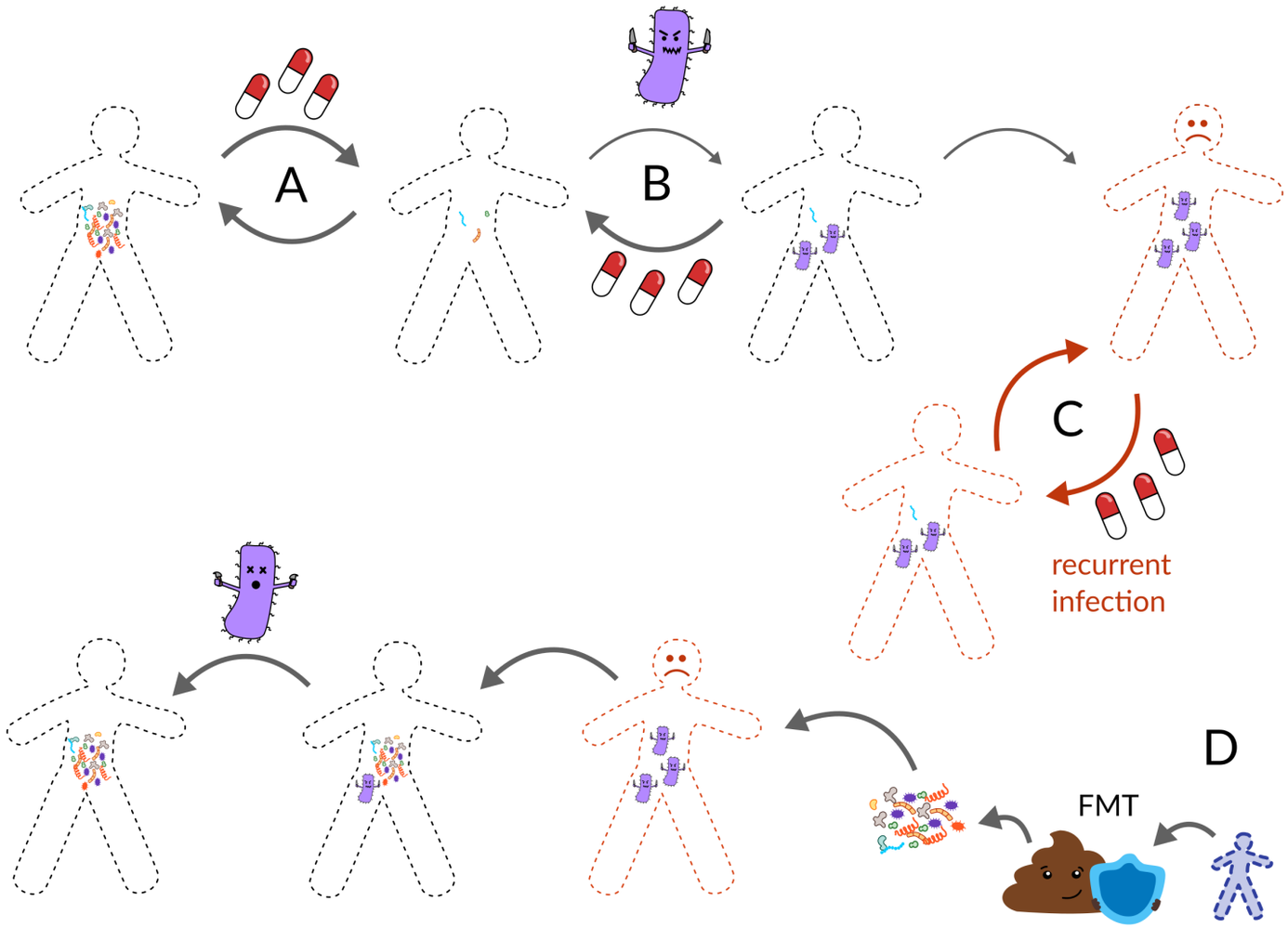
Analyzing gut microbial composition during recurrent *C. diff* infections

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

4 healthy donors and 4 individuals with recurrent infection.

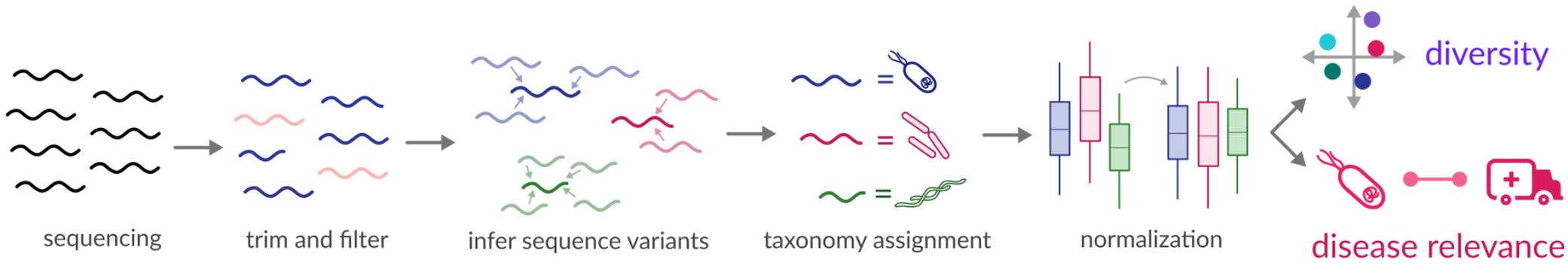
https://doi.org/10.1186/s40168-015-0070-0

The *C. diff* infection cycle

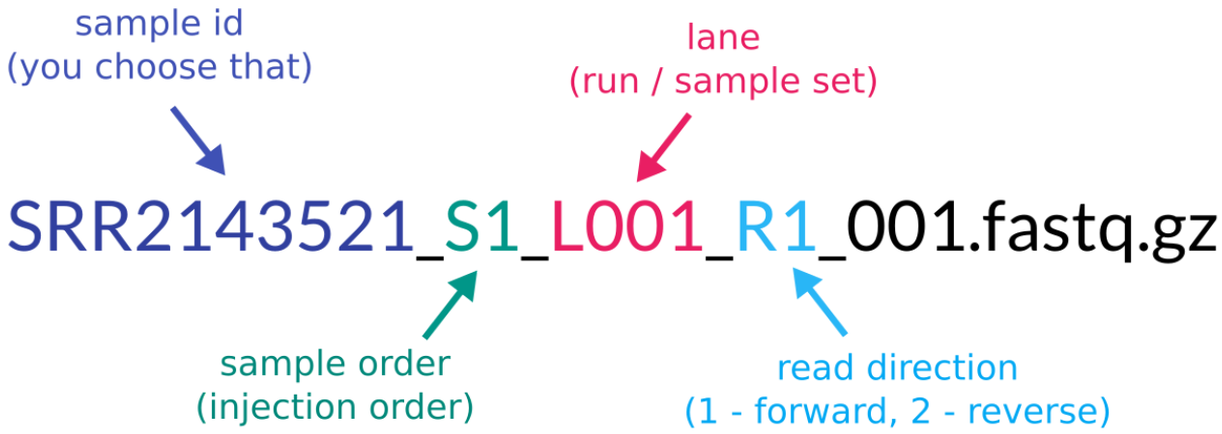


courtesy of Stephanie Swegle

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

Preprocessing sequencing reads

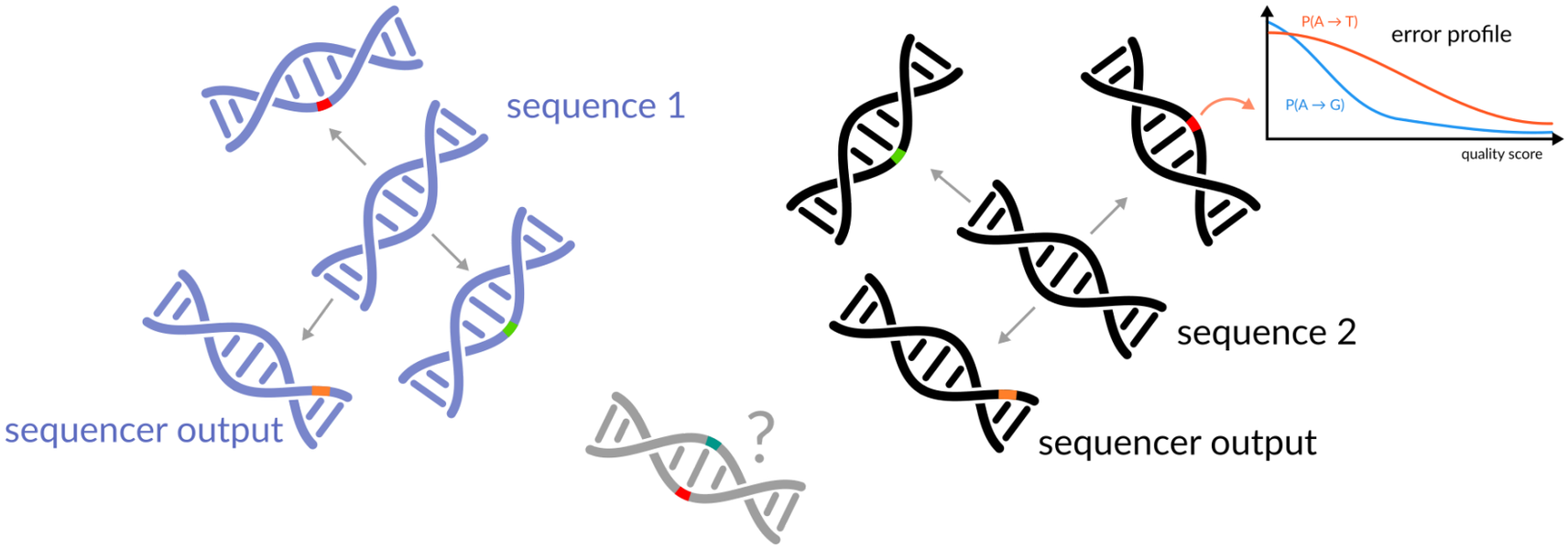
1. trim low quality regions
2. remove reads with low average quality
3. remove reads with ambiguous bases (Ns)
4. remove PhiX (added to sequencing)

Time to bring in the big guns 💣 ⚡

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

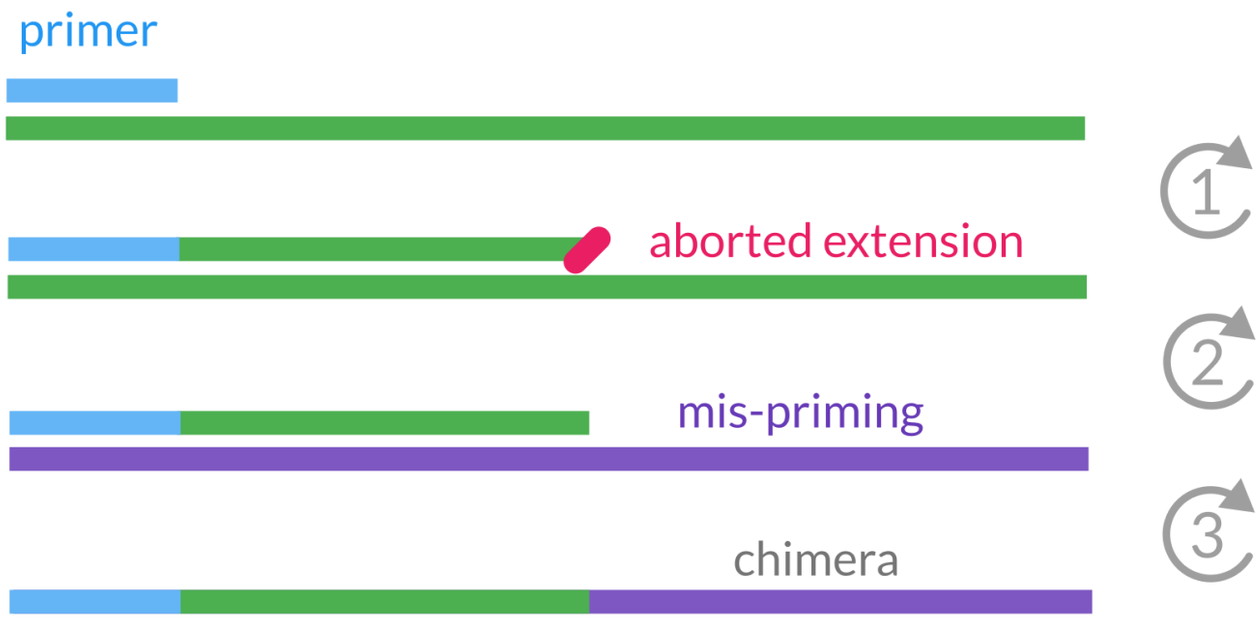
1. filter and trim the reads
2. find the most likely original sequences in the sample (ASVs)
3. remove chimeras
4. count the abundances

Identifying alternative sequence variants (ASVs)



Expectation-Maximization (EM) algorithm to find alternative 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.

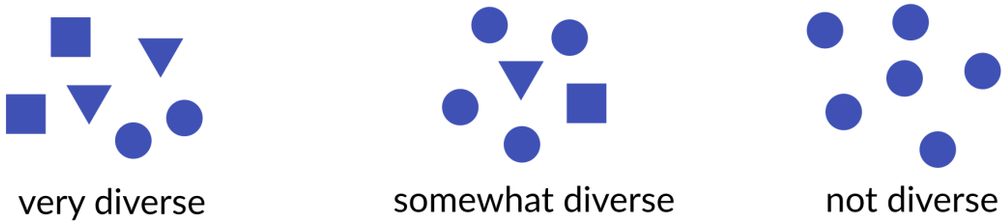
🤔 Do you have an idea for what we could do with those two data sets? What quantities might we be interested in?

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, Simpson index
- evenness: how evenly are abundances distributed across taxa?
  → Evenness index
- mixtures: metrics that combine both richness and evenness
  → Shannon 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).

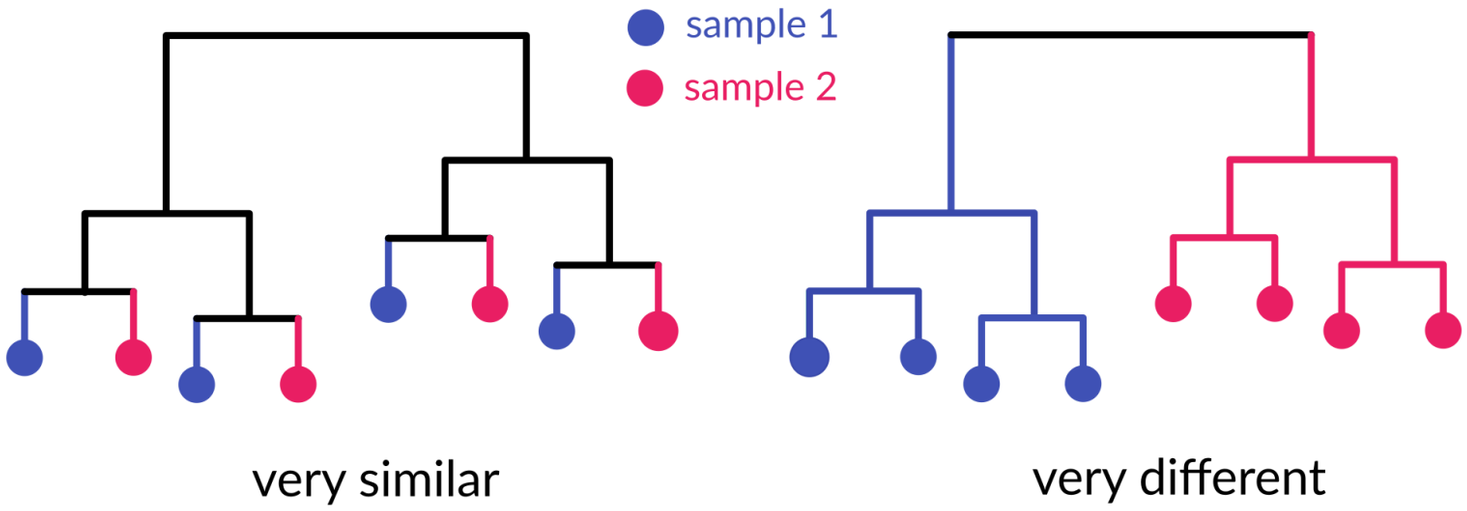
Beta diversity

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

- 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?



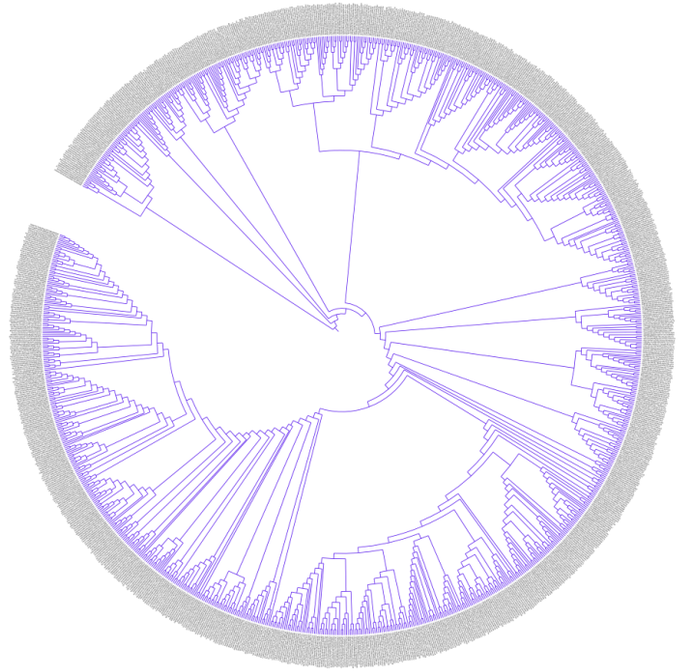
Weighted UniFrac scales branches by abundance.

How to build a phylogenetic tree?

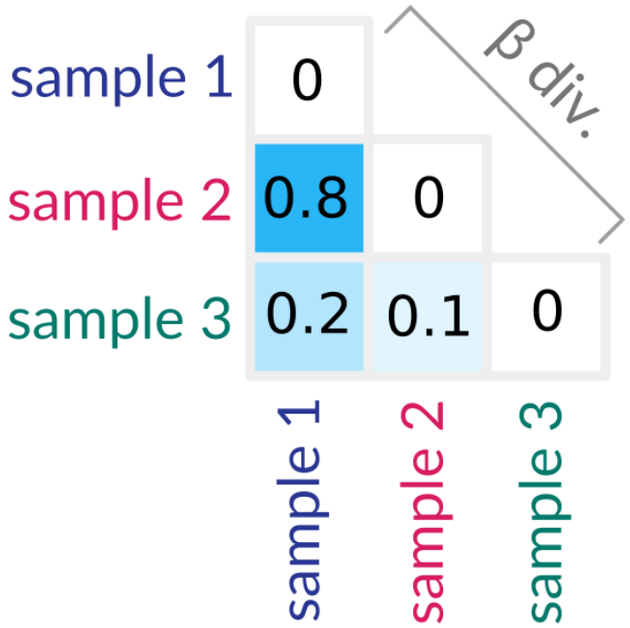
One of the basic things we might want to look at is how the sequences across all samples are related to one another. That is, we are often interested in their phylogeny.

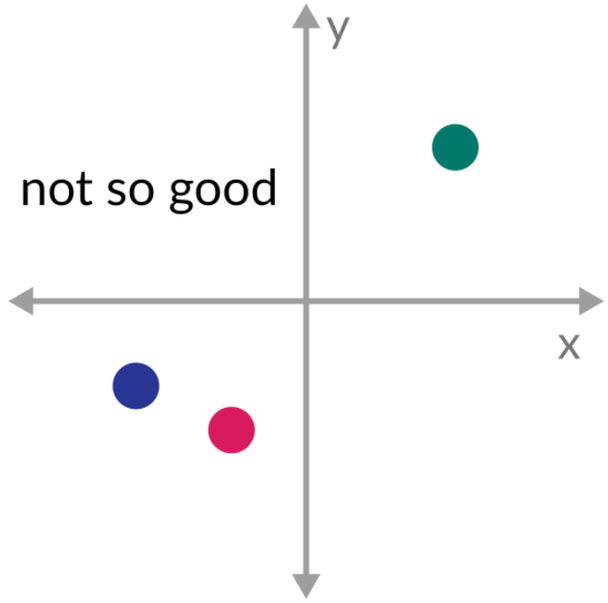
Phylogenetic trees are built from multiple sequence alignments and sequences are arranged by sequence similarity (branch length).

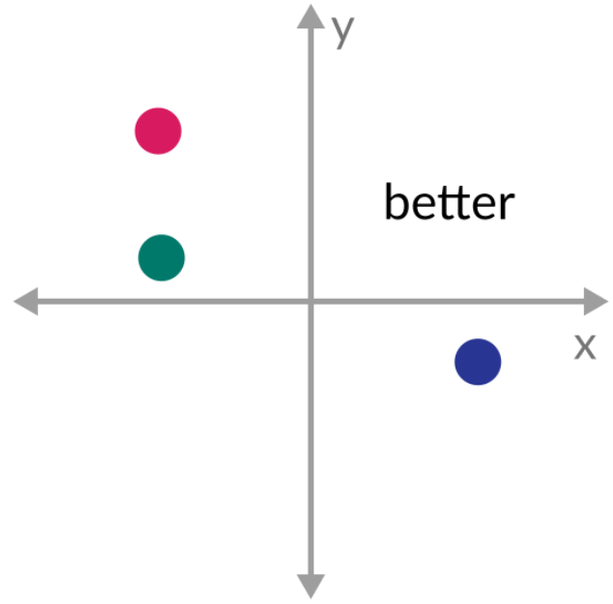
You can visualize your tree using iTOL (https://itol.embl.de/).



Principal Coordinate Analysis

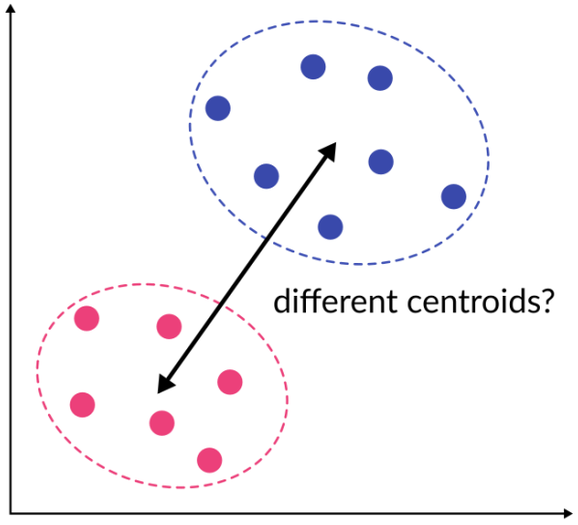


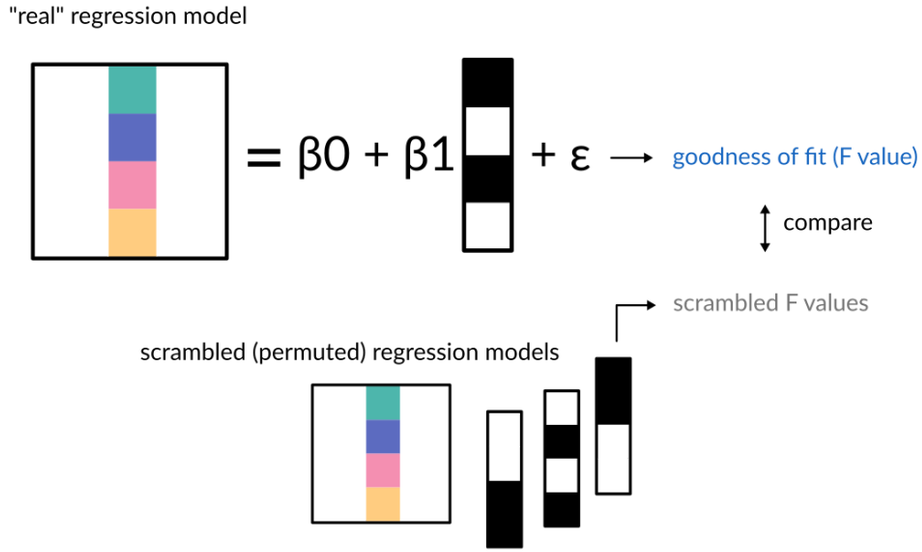




Statistical tests for beta diversity

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





Run the diversity analyses

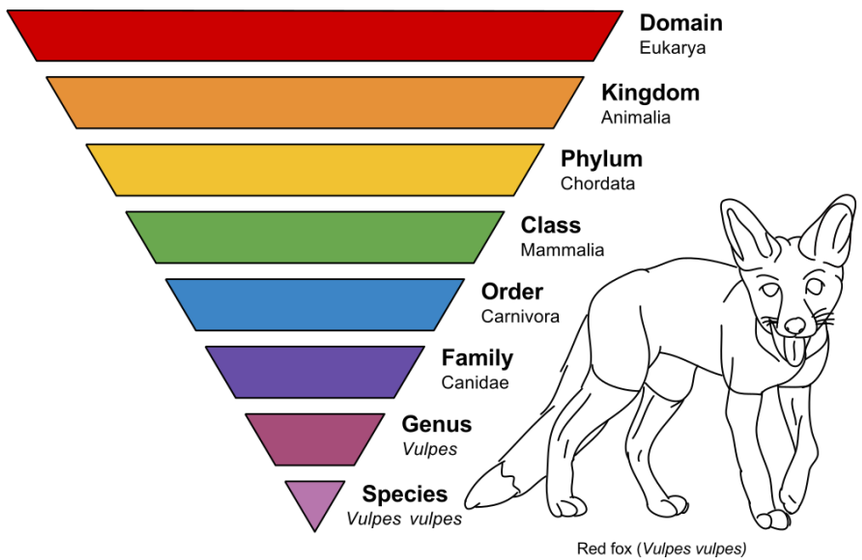
 Let’s switch to the notebook and calculate the diversity metrics

But what organisms are there in our sample?

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

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

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

Let’s assign taxonomy to the samples

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

Your turn

What is the relationship between particular taxa and the disease state?

And we are done 👏

Thanks!