Amplicon Sequencing Data Analysis with QIIME 2

Alyssa Easton, Gibbons Lab

from the **ISB Microbiome Course 2024**

 gibbons.isbscience.org gibbons-lab

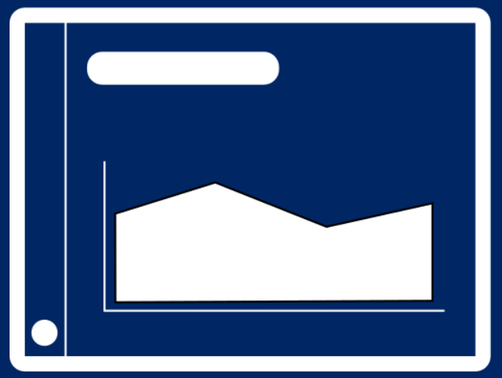
Let's Start Learning 📚

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

https://gibbons-lab.github.io/isb_course_2024/16S

Organization of the course

Presentation

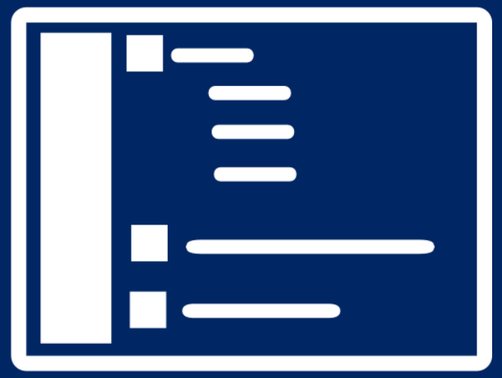


logic
explanations
links

Notebook

technical aspects
materials
visualizations

Chat



support
Q&A

Guts, Camera, Action 🎥

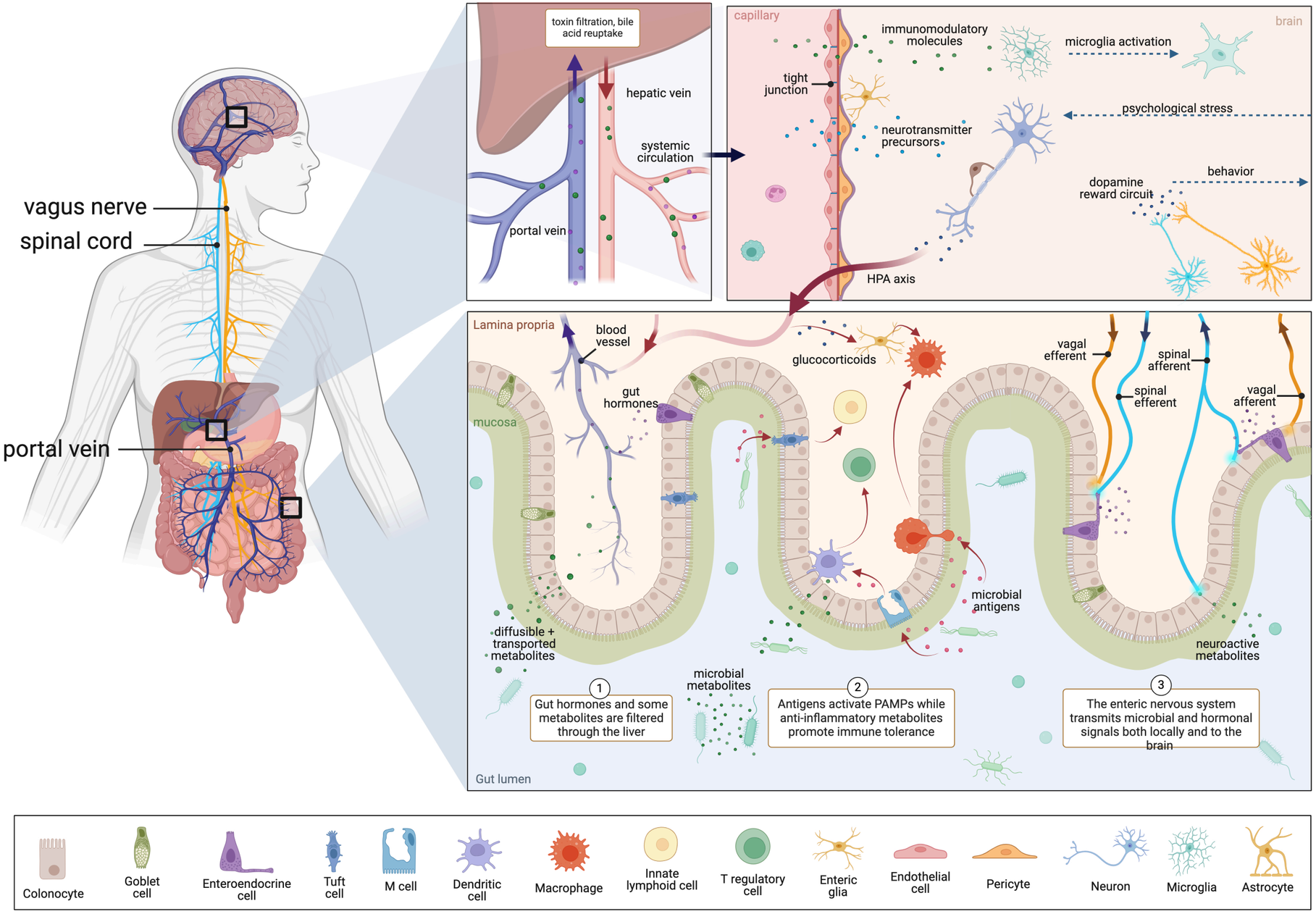
 Let's switch to the notebook and get started. Step 1: save a copy of this notebook in your Google Drive.

Click me to open the notebook!

In case you get lost, all output we generate can be found on Github, or in `materials/treasure_chest` in Colab.

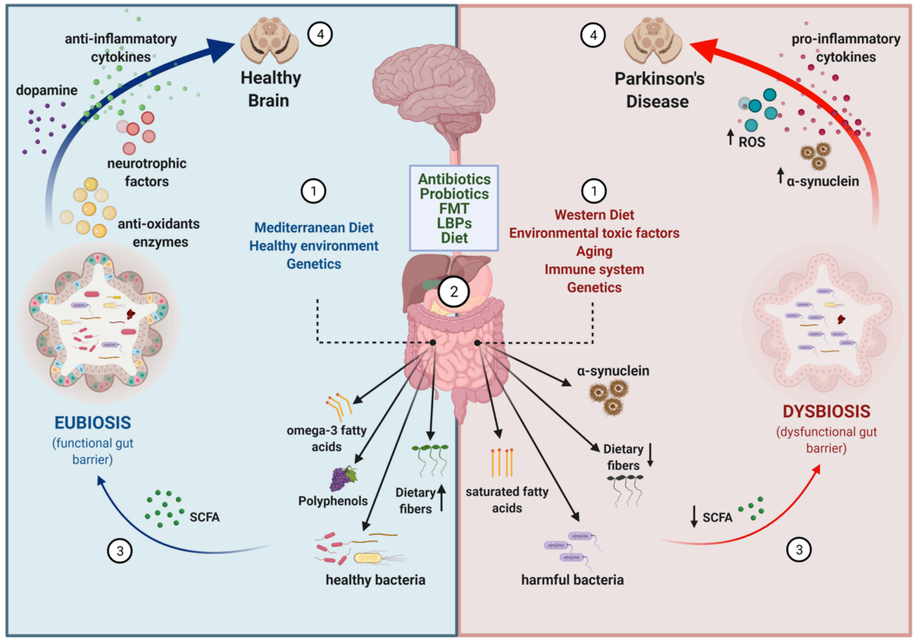
The Gut Microbiome 🦠

- 30-40 trillion bacterial cells
- Heterogenous between individuals
- Helps digest food and produces metabolites
- Affects our entire body, including the brain 🧠


vagus nerve
spinal cord
portal vein
toxin filtration, bile acid reuptake
hepatic vein
systemic circulation
portal vein
capillary
brain
immunomodulatory molecules
microglia activation
tight junction
psychological stress
neurotransmitter precursors
dopamine reward circuit
behavior
HPA axis
Lamina propria
blood vessel
glucocorticoids
vagal efferent
spinal afferent
spinal efferent
vagal afferent
gut hormones
mucosa
microbial antigens
diffusible + transported metabolites
neuroactive metabolites
microbial metabolites
1
Gut hormones and some metabolites are filtered through the liver
2
Antigens activate PAMPs while anti-inflammatory metabolites promote immune tolerance
3
The enteric nervous system transmits microbial and hormonal signals both locally and to the brain
Gut lumen
Colonocyte
Goblet cell
Enteroendocrine cell
Tuft cell
M cell
Dendritic cell
Macrophage
Innate lymphoid cell
T regulatory cell
Enteric glia
Endothelial cell
Pericyte
Neuron
Microglia
Astrocyte

Why Parkinson's Disease?

- Parkinson's Disease (PD) is characterized by aggregation of alpha-synuclein protein and degeneration of of dopaminergic neurons, leading to widespread neuroinflammation and progressive motor impairment.
- Motor symptoms are often preceded by gastrointestinal symptoms like constipation, increased gut permeability and inflammation.



Meta-analysis of the Parkinson's Disease Microbiome suggests alterations linked to intestinal inflammation (Romano et. al. 2021)

Figure from Lorente-Picón & Laguna, 2021

Today's dataset:

Case-control observational study of PD (Hill-Burns et. al., 2017), which was later included in a 2021 meta-analysis (Romano et. al. 2021).

- 197 PD cases, 130 healthy controls
- 16S rRNA Amplicon sequencing of stool
- included demographic information, health history, and medication use
- found small (but significant) independent effects of PD and PD drugs on microbiome composition

Today, we'll process a small subset of the original data: 5 PD patients and 5 healthy controls.

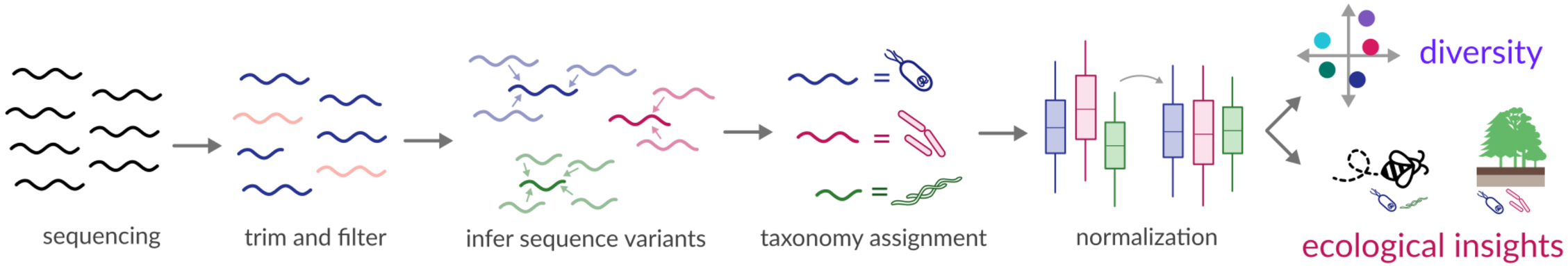
Free full-text manuscript of Hill-Burns, 2017 available at Europe PMC

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

What will we do today?



QIIME 2: Quantitative Insights into Microbial Ecology

Pronounced like chime 🔔

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

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

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.

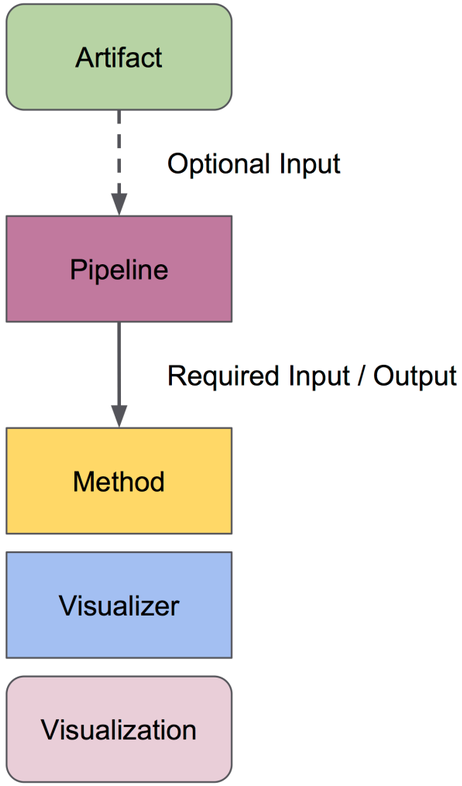
But what is QIIME2, really?

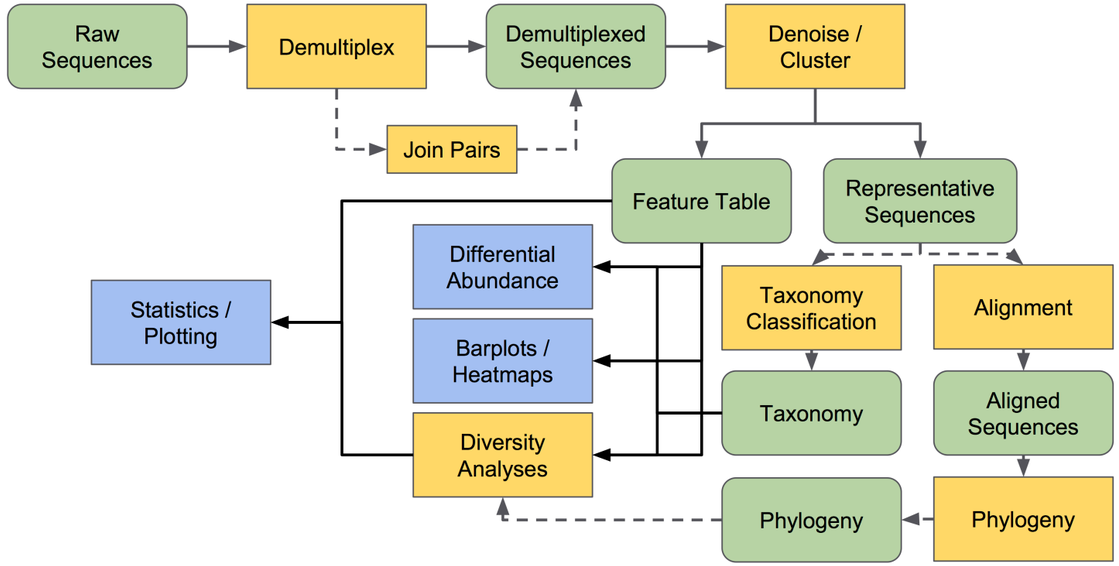
Essentially, QIIME 2 is a set of commands to transform microbiome data into intermediate outputs and visualizations.

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

QIIME2 Workflow

When we run a QIIME2 command, we specify the inputs and action to perform, and QIIME2 will output artifacts (.qza) and/or visualizations (.qzv).





https://docs.qiime2.org/2024.5/concepts/ https://docs.qiime2.org/2024.5/tutorials/overview/

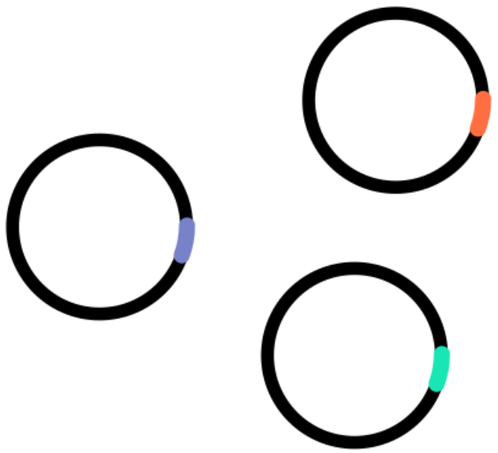
Let's make an artifact

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

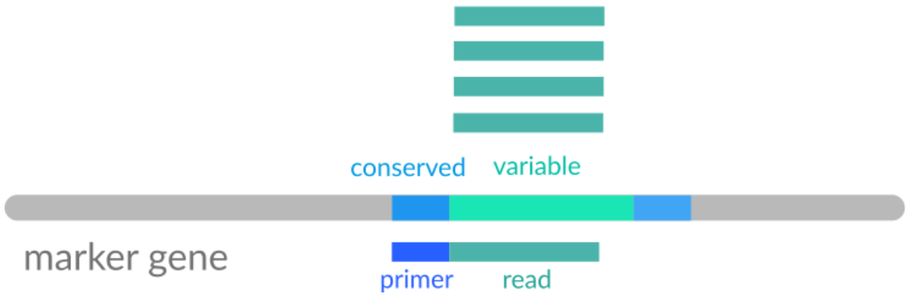
 Let's switch to the notebook and get started

What is amplicon sequencing?

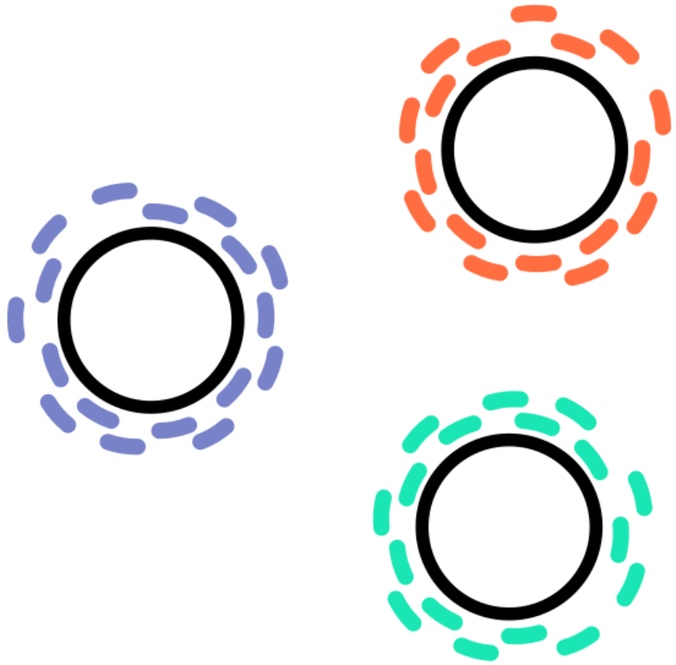
amplicon sequencing



DNA extraction
PCR amplification
sequencing

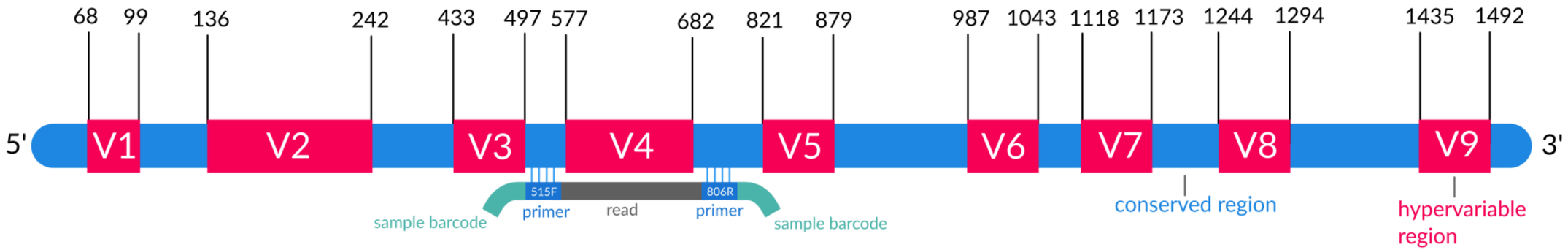


shotgun metagenomics



DNA extraction
fragmentation
adapter ligation
sequencing

Why the 16S gene?



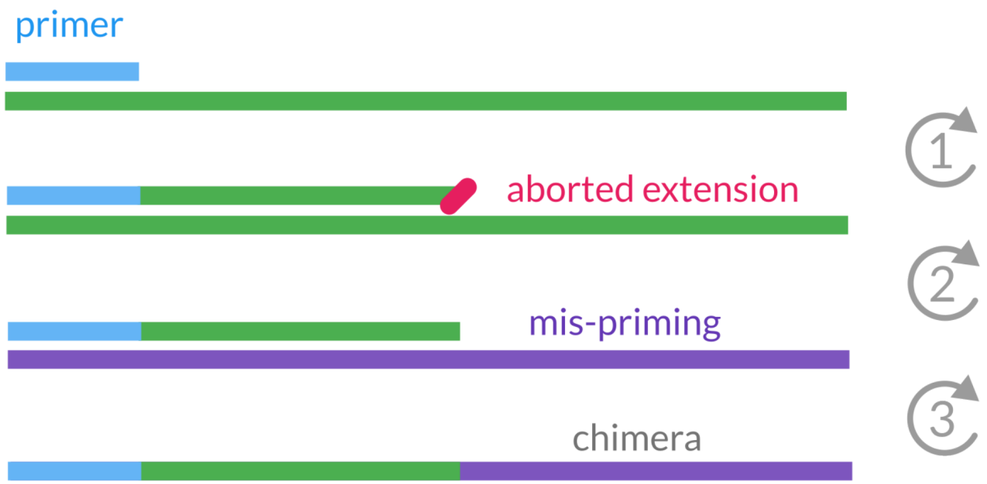
https://dx.doi.org/10.1016%2Fj.mimet.2007.02.005

The 16S gene is universal and contains interspersed conserved regions perfect for PCR priming and hypervariable regions with phylogenetic heterogeneity. Our data used the V4 region.

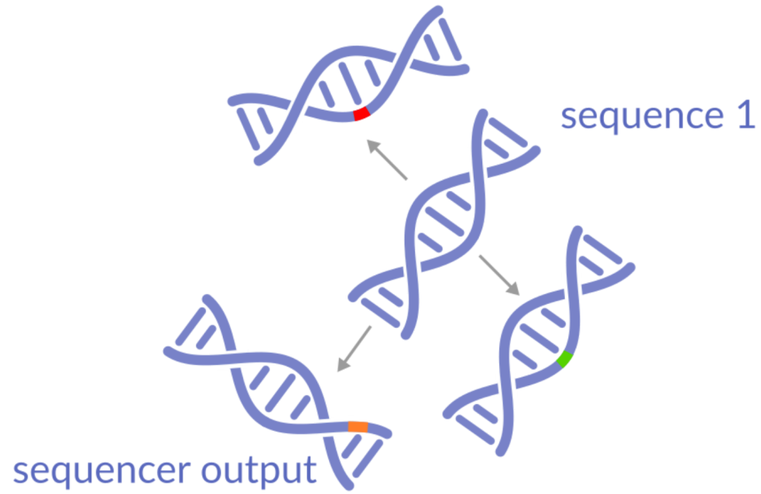
The V4-specific primers used in this study were F515/R806. How long is the amplified fragment, and how long are the reads?

Errors during PCR and sequencing generate noise

chimeras form during PCR amplification



next-generation sequencing methods have low, but non-negligible error rates

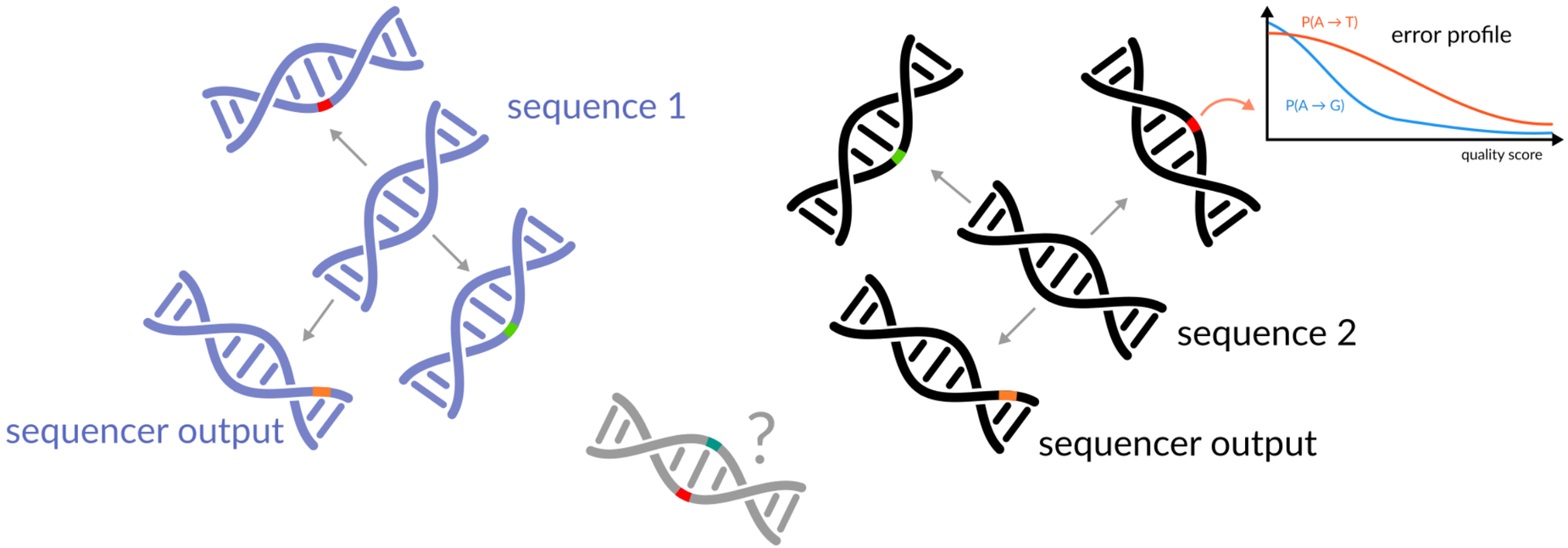


DADA2 to the rescue!

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

1. filter and trim the reads
   a. trim low quality regions
   b. remove reads with low average quality
   c. remove reads with ambiguous bases (Ns)
   d. remove PhiX (added to sequencing)
2. find the most likely original sequences in the sample (ASVs)
3. remove chimeras
4. count ASV abundances

Identifying Amplicon Sequence Variants (ASVs)



Expectation-Maximization (EM) algorithm simultaneously assigns ASVs and models error (Callahan, 2016).

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?

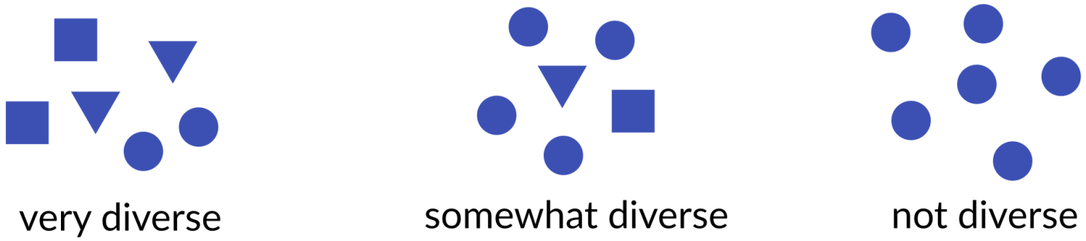
Diversity metrics

In microbial community analysis we are usually interested in two different families of diversity metrics:

- **alpha diversity** (ecological diversity within a sample)
- **beta diversity** (ecological differences between samples)

Alpha diversity

How diverse is a single sample?

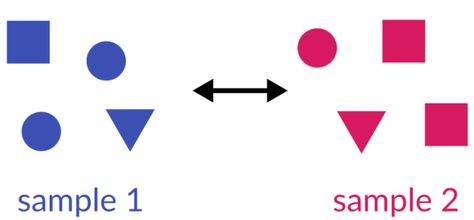


- richness: how many taxa do we observe (richness)?
  → #observed taxa
- evenness: how evenly are abundances distributed across taxa?
  → Evenness index
- mixtures: metrics that combine both richness and evenness
  → Shannon Index, Simpson's Index

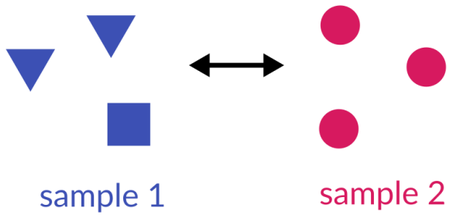
Each sample has **1** Shannon Index.

Beta diversity

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



very similar



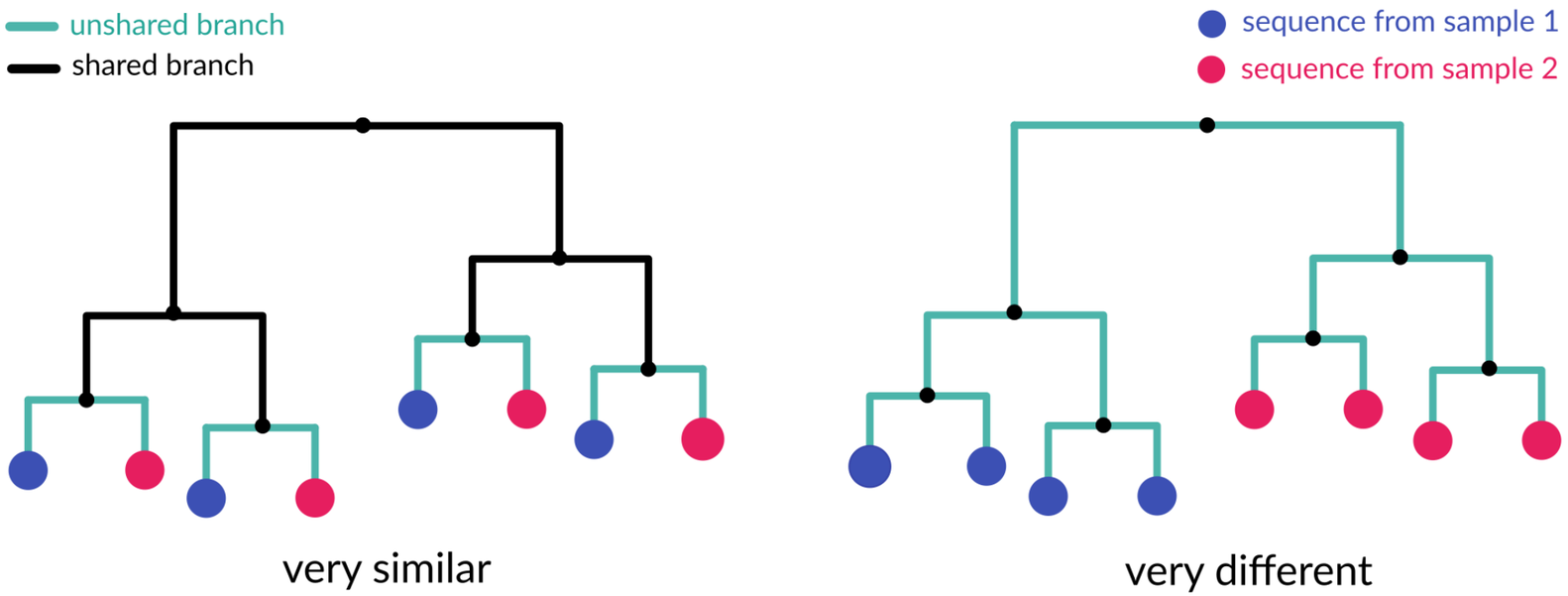
very different

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

Each sample has **n** Bray-Curtis distances, where n = number of samples.

UniFrac beta diversity

Do samples share genetically similar taxa? UniFrac distance = branch length



Weighted UniFrac **scales branches by abundance**, so the presence of one distant member does not skew diversity.

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).

Let's make one!

Statistical tests for alpha diversity

Alpha diversity can be treated as any other sample measurement and is suitable for classic univariate tests.

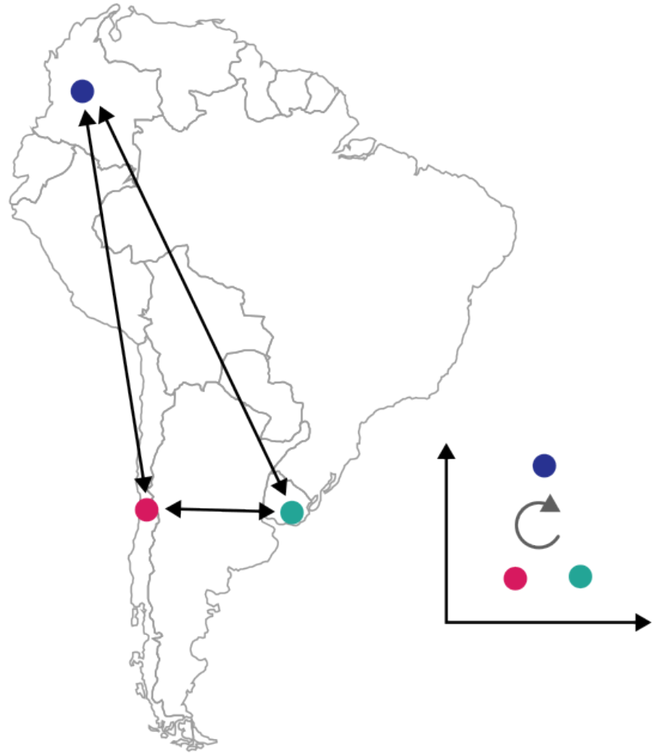
```
→ Nonparametric: Mann-Whitney U test<br>
→ Parametric: t-test
```

Visualizing beta diversity with PCoA

	Bogotá	Santiago	Montevideo
Bogotá	0		
Santiago	0.7	0	
Montevideo	0.5	0.2	0

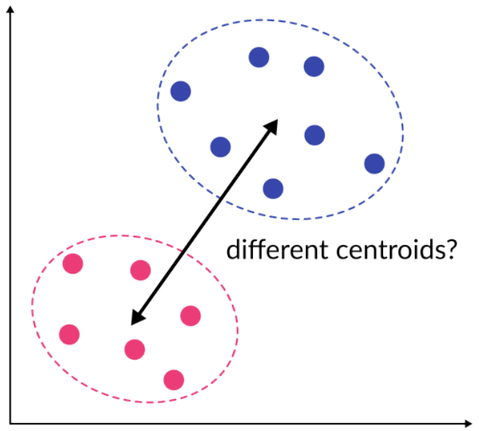
β div.

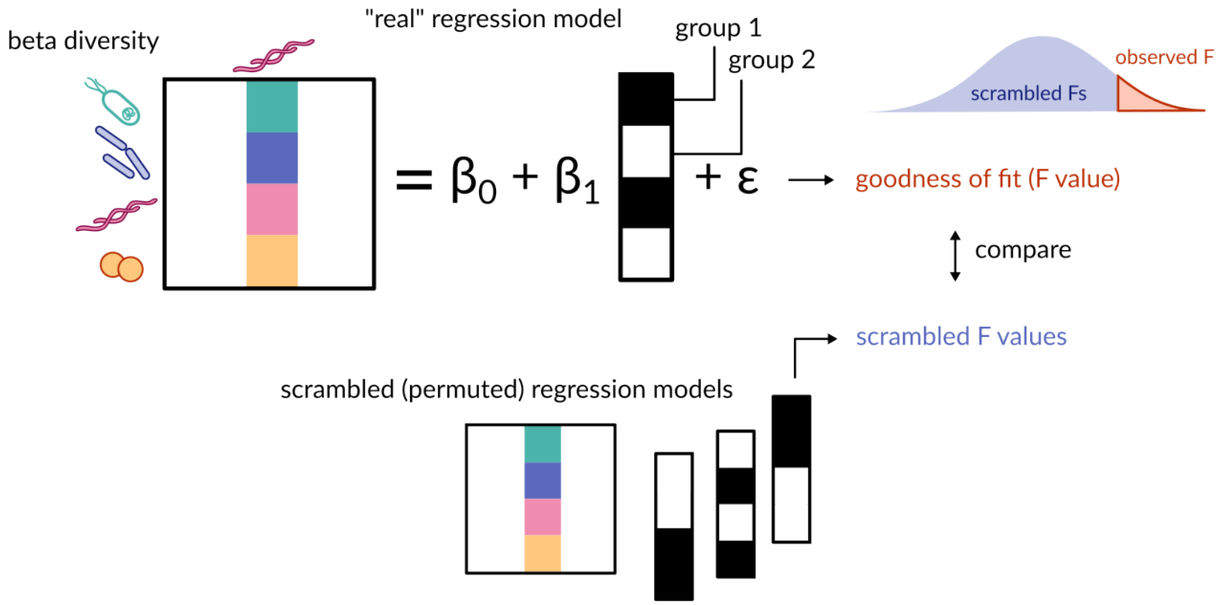
consistent representation(s)
many different projections
reduce dimensions



Statistical tests for beta diversity

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





Run the diversity analyses

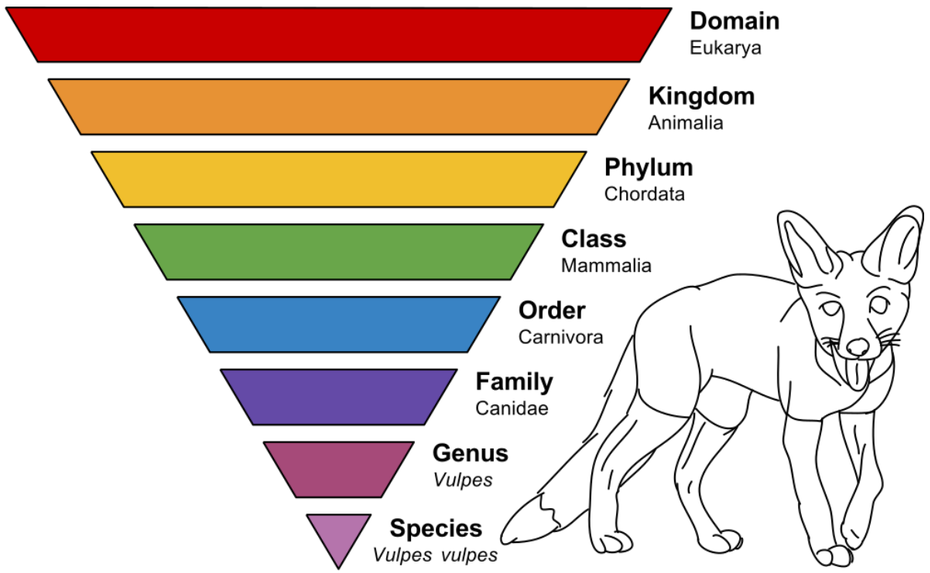
 Let’s switch to the notebook and calculate the diversity metrics

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

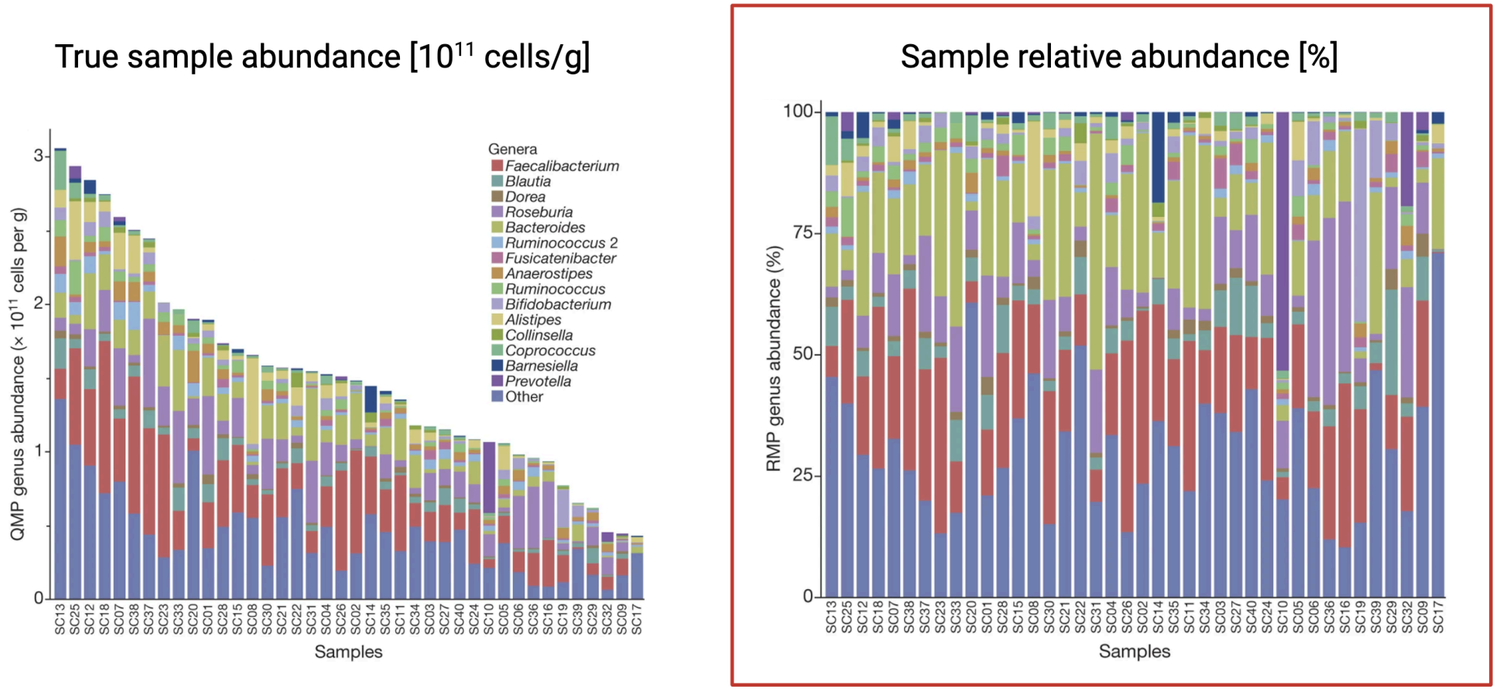
 Let’s switch to the notebook and assign taxonomy to our ASVs

Differential abundance

How can we compare abundance between groups of samples (like case-control)?

Sequencing is a random sample of an ecosystem

Sampling fraction varies between samples.



Even if we knew the concentration of bacteria in a sample, we don't know how much bacterial biomass is in each person. What we do know is the **proportions** of the bacteria **within a sample** (assuming these are not confounded by a taxon-specific sequencing bias).

Figures from Vandeputte, 2017

What taxonomic level should we use?

16S rRNA amplicon sequencing generally has good resolution at the genus level.

What statistical tests can we use?

Before we do any tests, we need to LOOK at the data 🔍

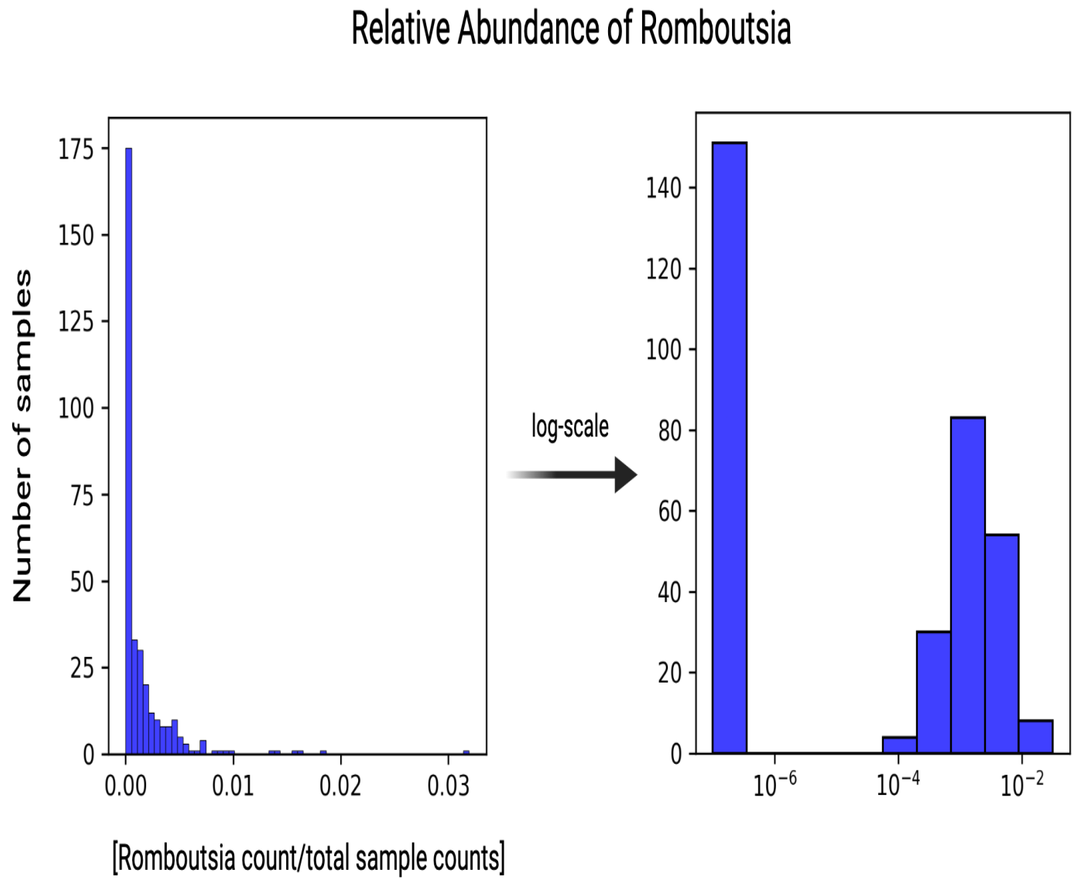
How is it distributed? What is the variance?

Relative abundance data structure

Microbiome relative abundance data is:

- compositional
- not normally distributed
- zero-inflated: contains both true and sampling zeros
- more variable than expected by a Poisson model (overdispersed)
- heteroscedastic

These features violate the assumptions of parametric statistical tests.



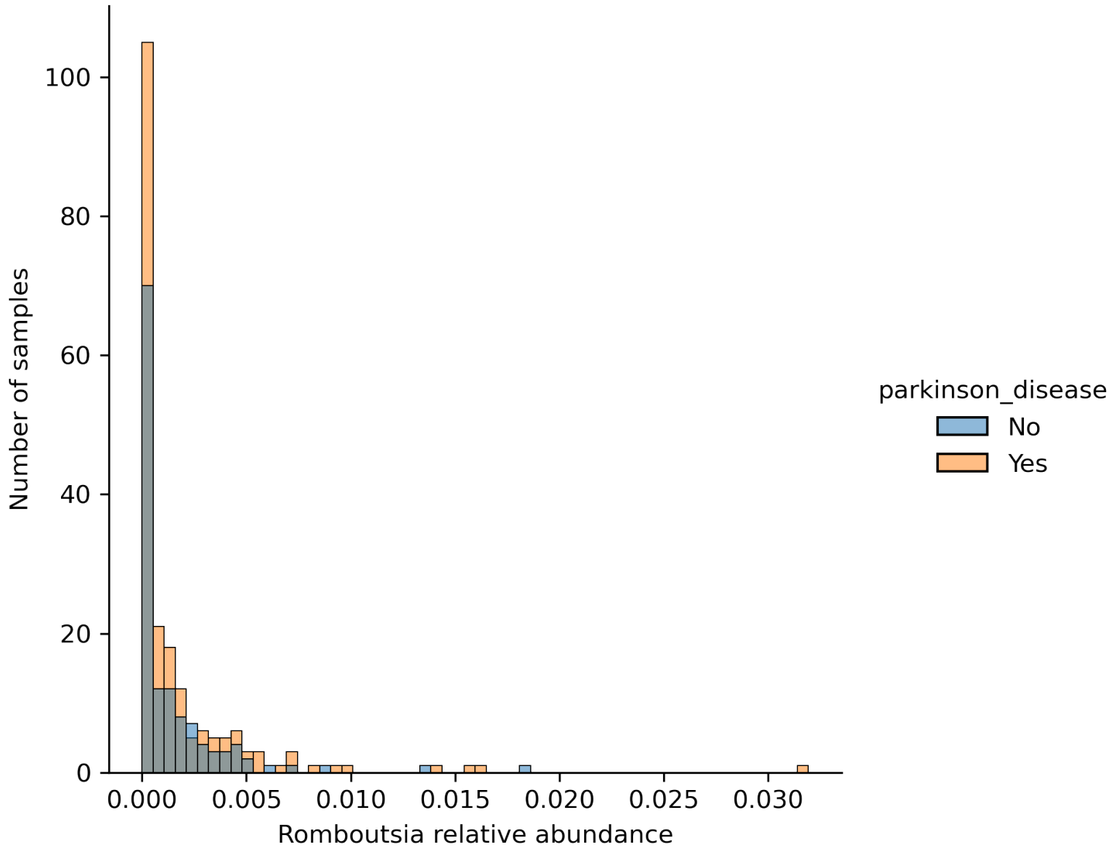
So what can we do?

We have a few options:

1. Use relative abundance data; run **nonparametric, rank-based tests** (underpowered)
2. Transform data (will require us to impute or discard zeros); run **parametric tests**
3. Use a more complex modeling package (each with their own caveats + assumptions)

Wilcoxon Rank-Sum Test (a.k.a. Mann-Whitney U test)

- Nonparametric test for difference in a continuous variable between two groups
- Uses ranks, rather than counts
- Underpowered, because we are not assuming an ideal distribution shape

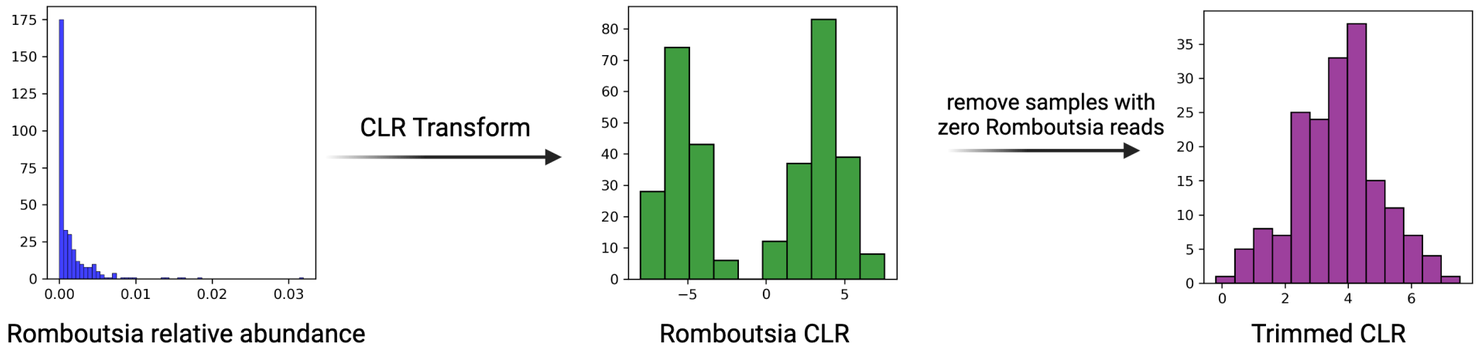


Normalization + Parametric tests

The center-log ratio (CLR) transform (Aitchison, 1982), is a transformation for compositional data that normalizes by the sample geometric mean, which is less sensitive to outliers and gives more weight to smaller values.

$$clr(x) = \left[log\frac{x_1}{g(x)}, \ldots, log\frac{x_n}{g(x)} \right]$$

$$\longrightarrow g(x) = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$



The caveat: zeros are still a problem. We can either impute or discard them.

After hypothesis testing, we need to correct our p-values

- We perform p-value correction to minimize the false discovery rate (FDR).
- We also look at the pre-correction p-value distribution as a sanity check.

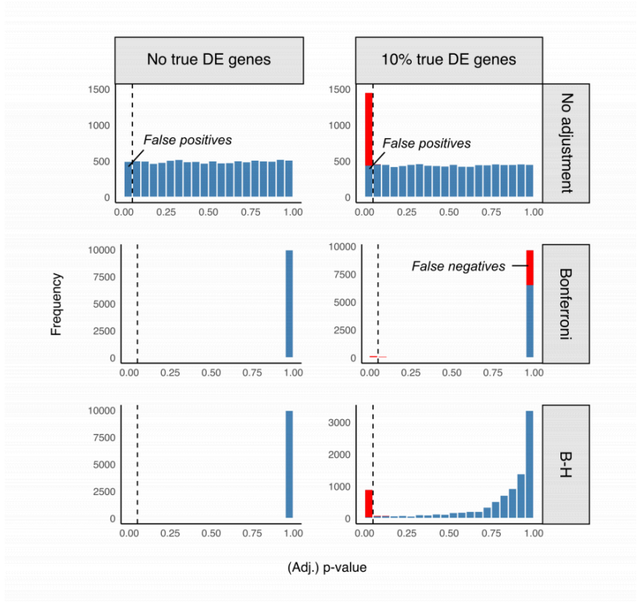


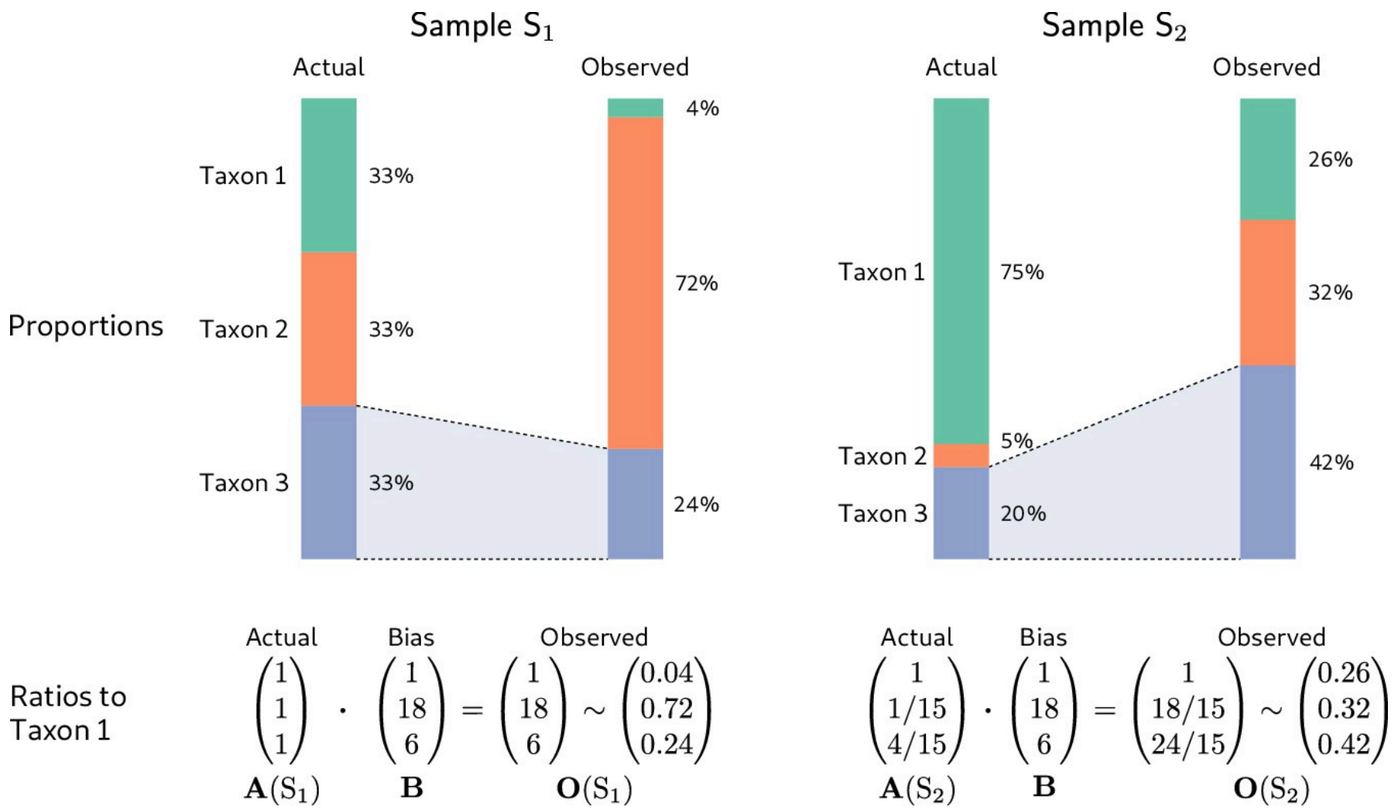
Figure from Genevia Technologies

Limitations

The assumption that taxonomic proportions of an ecosystem are conserved in sequencing data may not be true.

Analysis of defined bacterial communities suggests that bacterial taxa have different sequencing efficiencies, which can distort differences in abundance (McLaren et. al., 2019).

A new method that accounts for this bias is currently in review. radEmu, developed by David Clausen and Amy Willis at UW, is available as an R package on Github.



Let’s try it!

 We will now switch to the notebook to do some normalization and statistical testing.

In conclusion, always look at your processed data!

Your turn!

Put taxonomic assignments on your tree!

And we are done 👏

Alex Carr
Jacob Cavon
Christian Diener
Alyssa Easton
Karl Gaisser
Sean Gibbons
Crystal Perez
Nick Quinn-Bohmann
Noa Rappaport
Shanna Braga
Greg Caporaso
Audri Hubbard
Connor Kelly
Allison Kudla
Dominic Lewis
Joe Myxter
Thea Swanson
Victoria Uhl
ISB Facilities Team

Thanks! ❤️