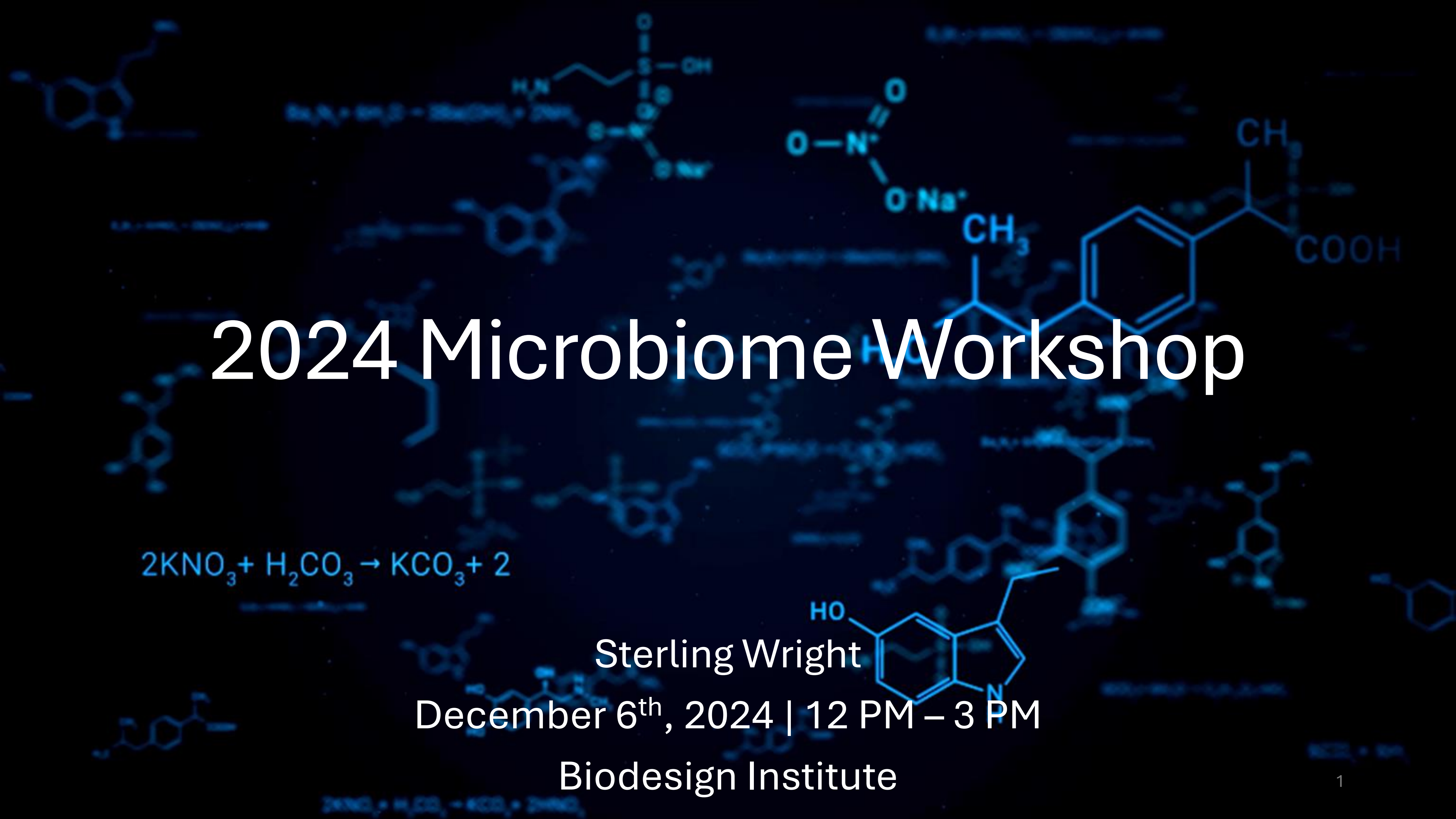
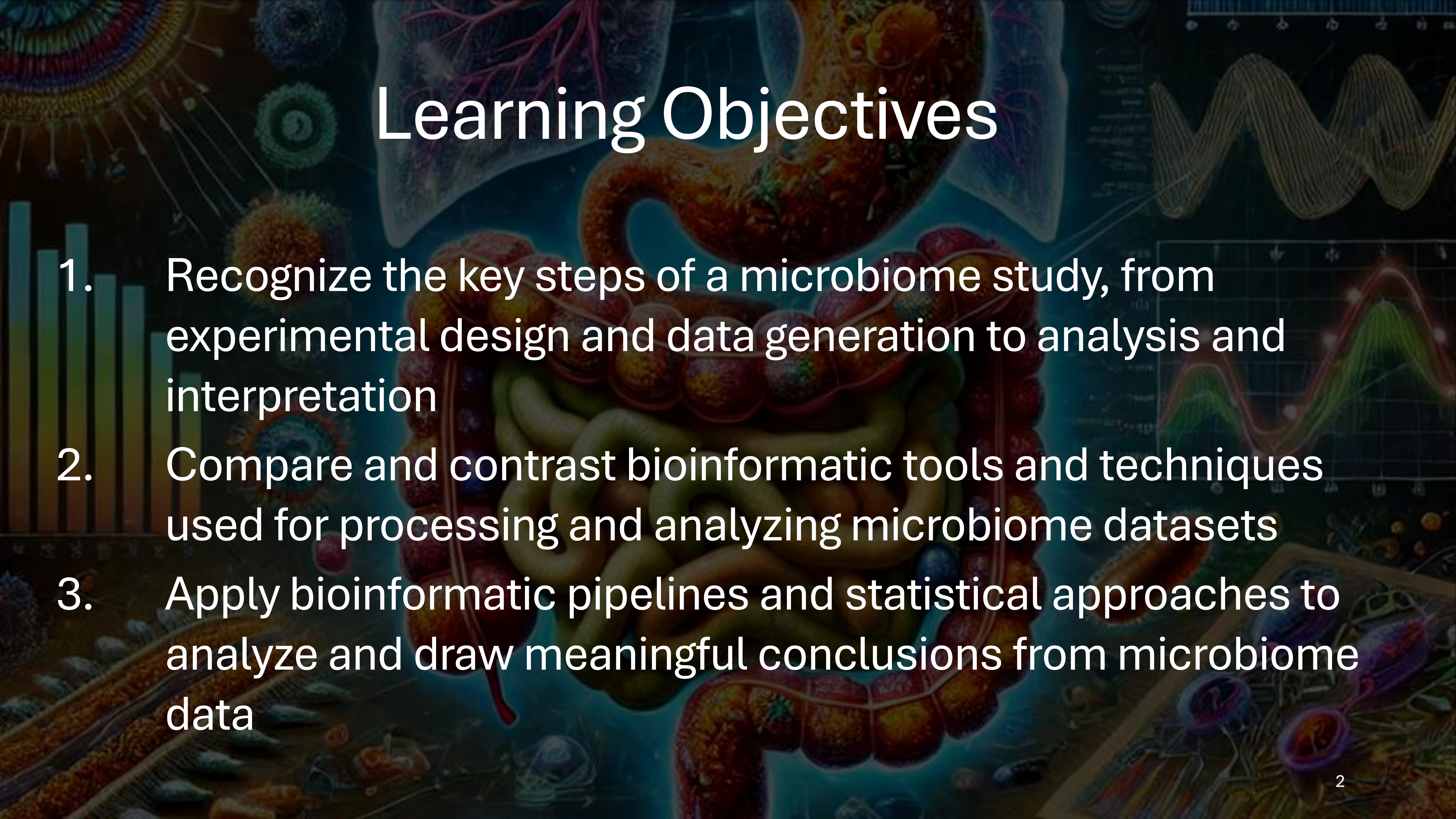


2024 Microbiome Workshop

Sterling Wright

December 6th, 2024 | 12 PM – 3 PM

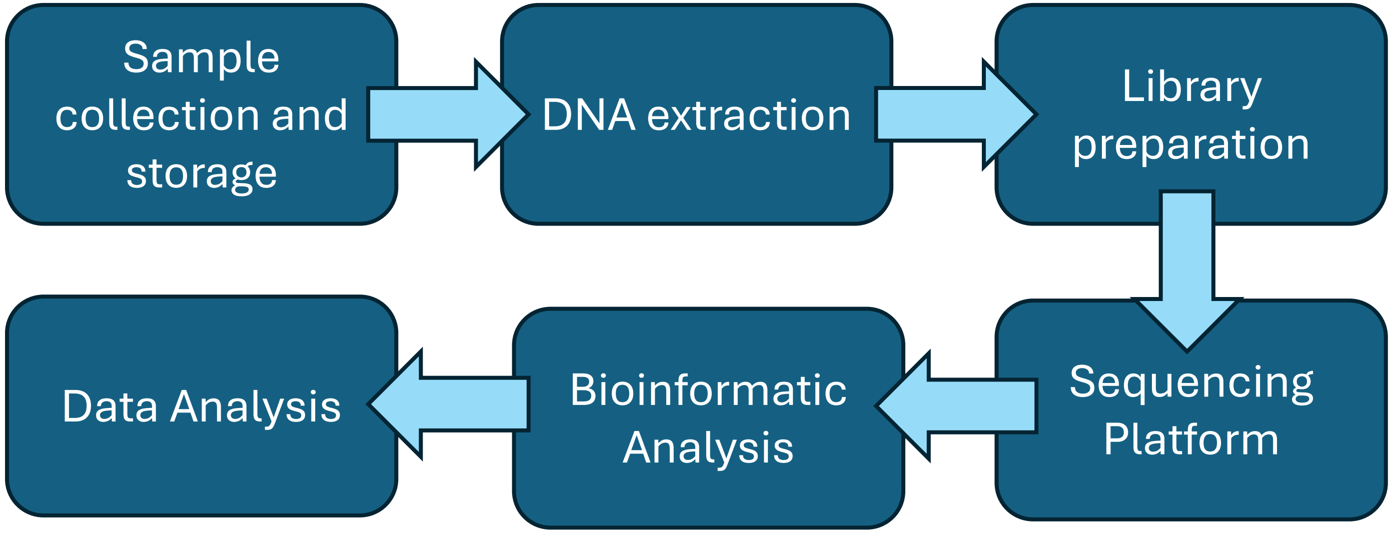
Biodesign Institute

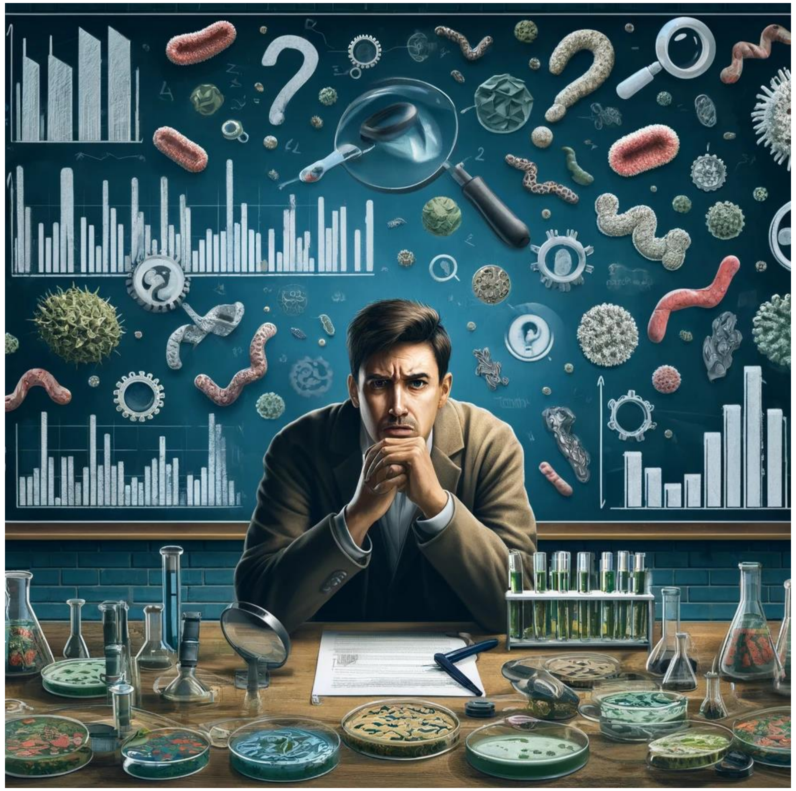


Learning Objectives

1. Recognize the key steps of a microbiome study, from experimental design and data generation to analysis and interpretation
2. Compare and contrast bioinformatic tools and techniques used for processing and analyzing microbiome datasets
3. Apply bioinformatic pipelines and statistical approaches to analyze and draw meaningful conclusions from microbiome data

Overview of Microbiome Study Design



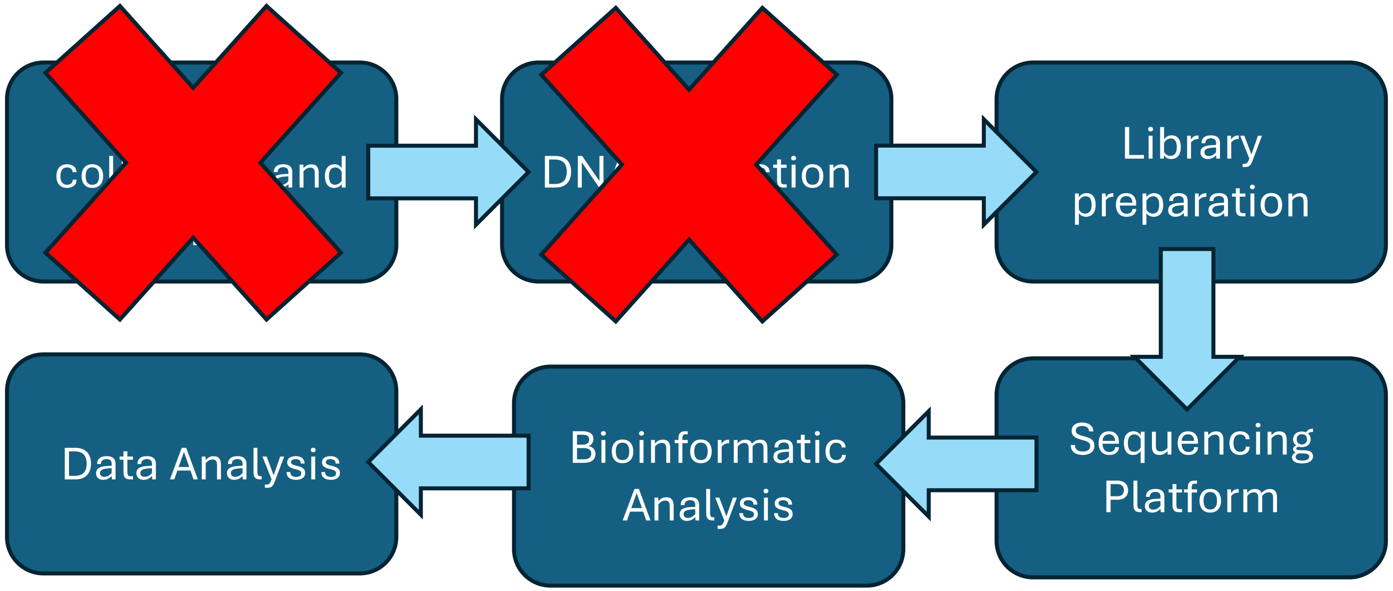


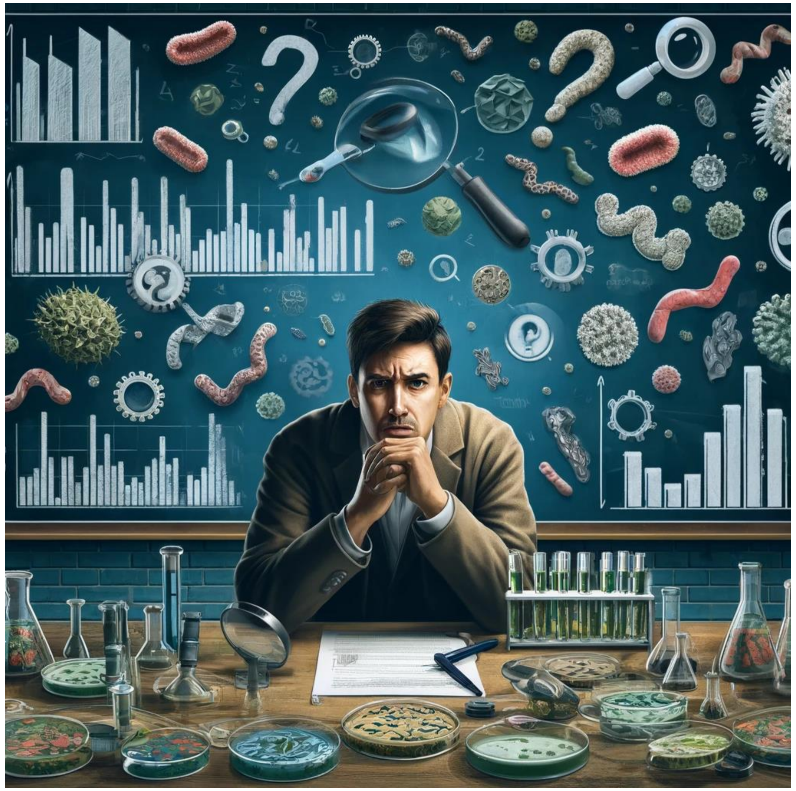
Measurement Bias: Deviation from ground truth

Measurement Noise: Experimental variability

Forry et al. (2024); McGuinness et al. (2024); Duran-Pinedo et al. (2021)

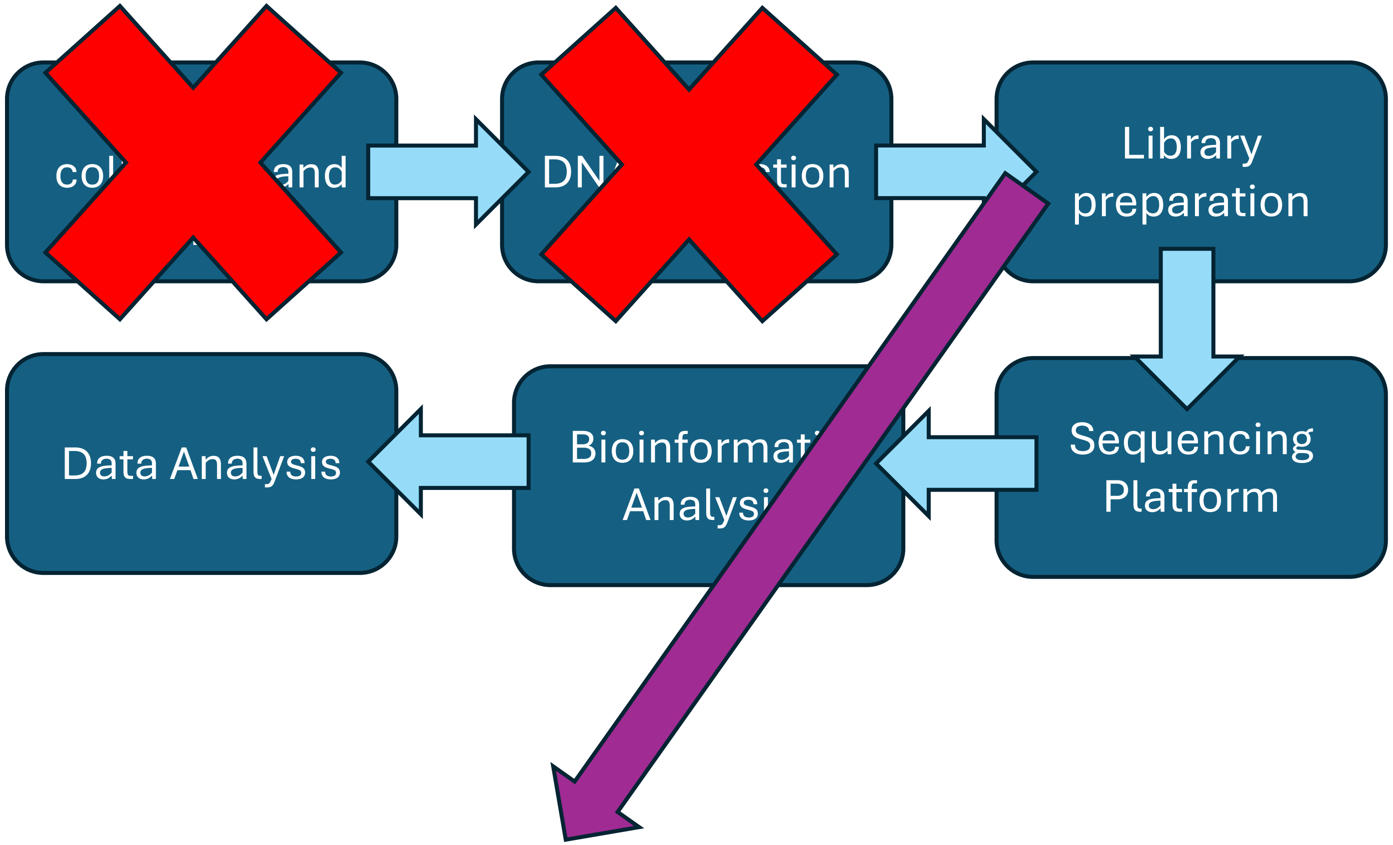
Overview of Microbiome Study Design





Focus on key steps from library preparation to downstream analyses

Overview of Microbiome Study Design



col and

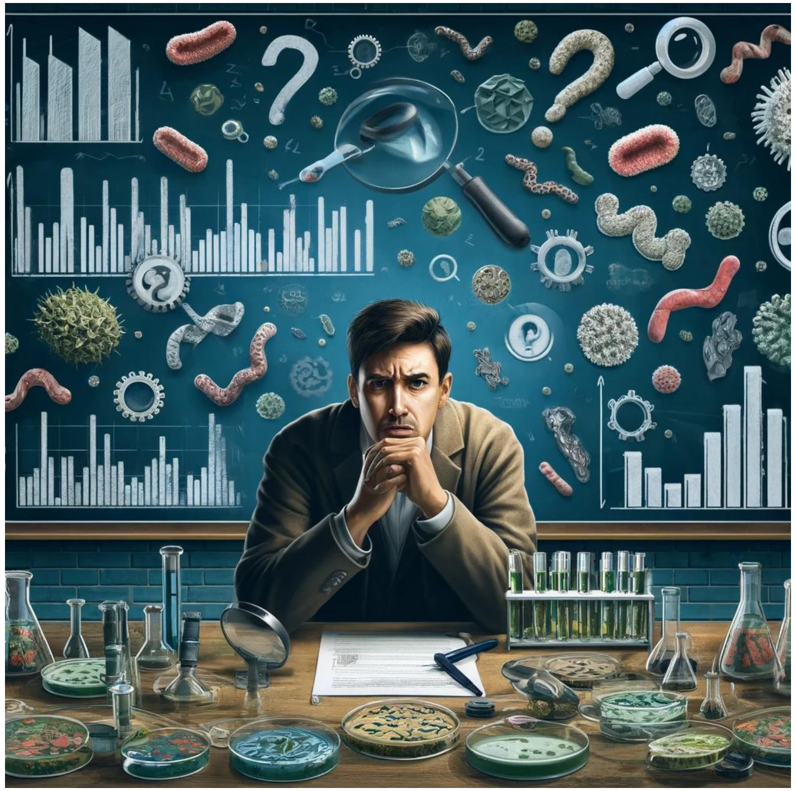
DN tion

Library preparation

Sequencing Platform

Bioinformat Analysi

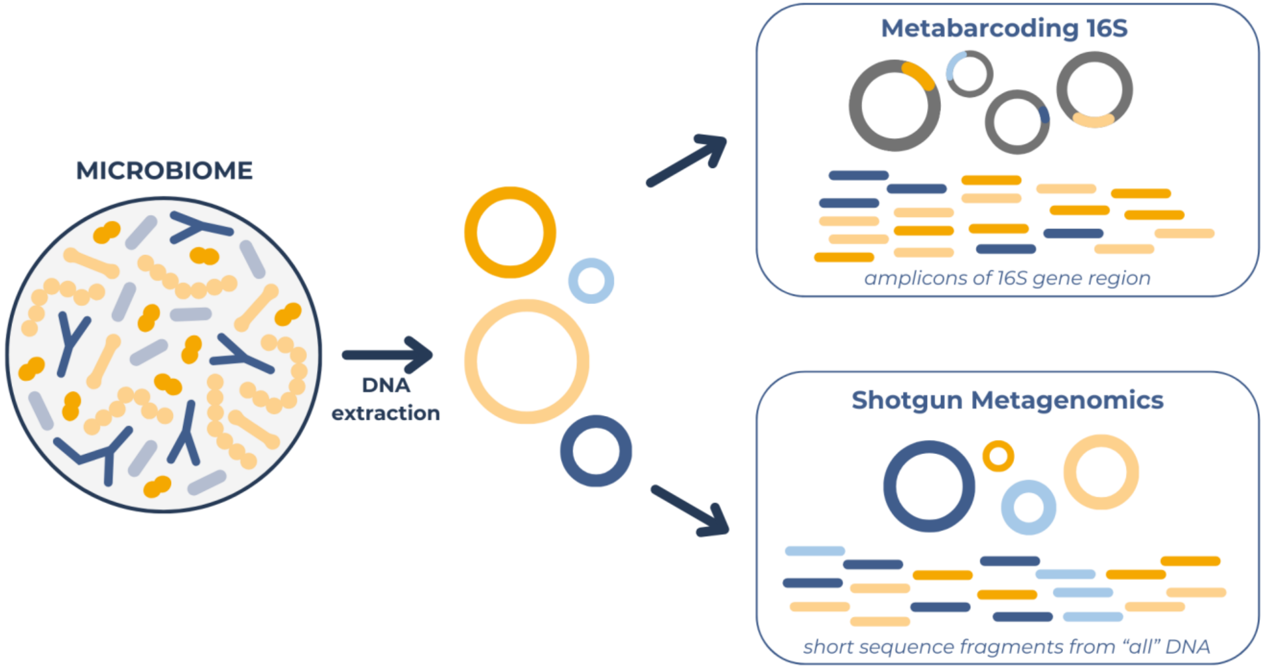
Data Analysis



What is the difference between 16S rRNA and shotgun metagenomic sequencing?

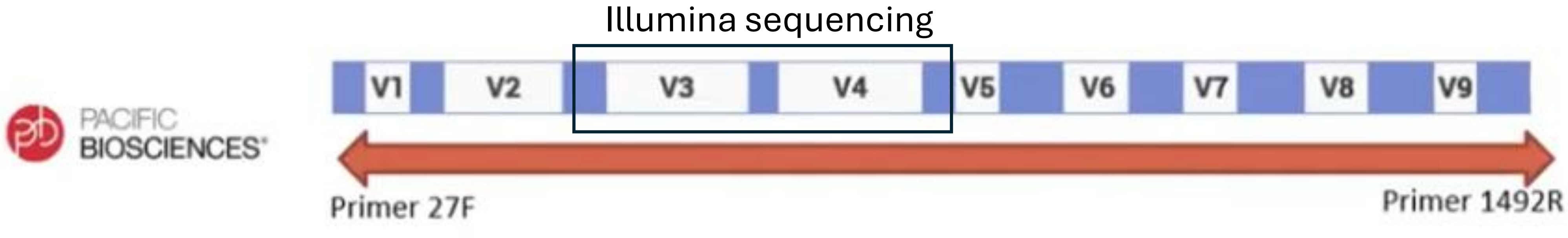
16S vs Shotgun metagenomic basics

	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Cost	~$50	~$150
Sample preparation	Utilizes primers	No primers
Functional profiling	No	Yes
Taxonomic resolution	Historically, only to the genus level	Species and strains
Taxonomic coverage	Bacteria and archaea	All taxa, including viruses
Databases	Established, well-curated	Less established and continually evolving
Bias	Requires *a priori*	Uses an untargeted approach


MICROBIOME
DNA extraction
Metabarcoding 16S
amplicons of 16S gene region
Shotgun Metagenomics
short sequence fragments from "all" DNA

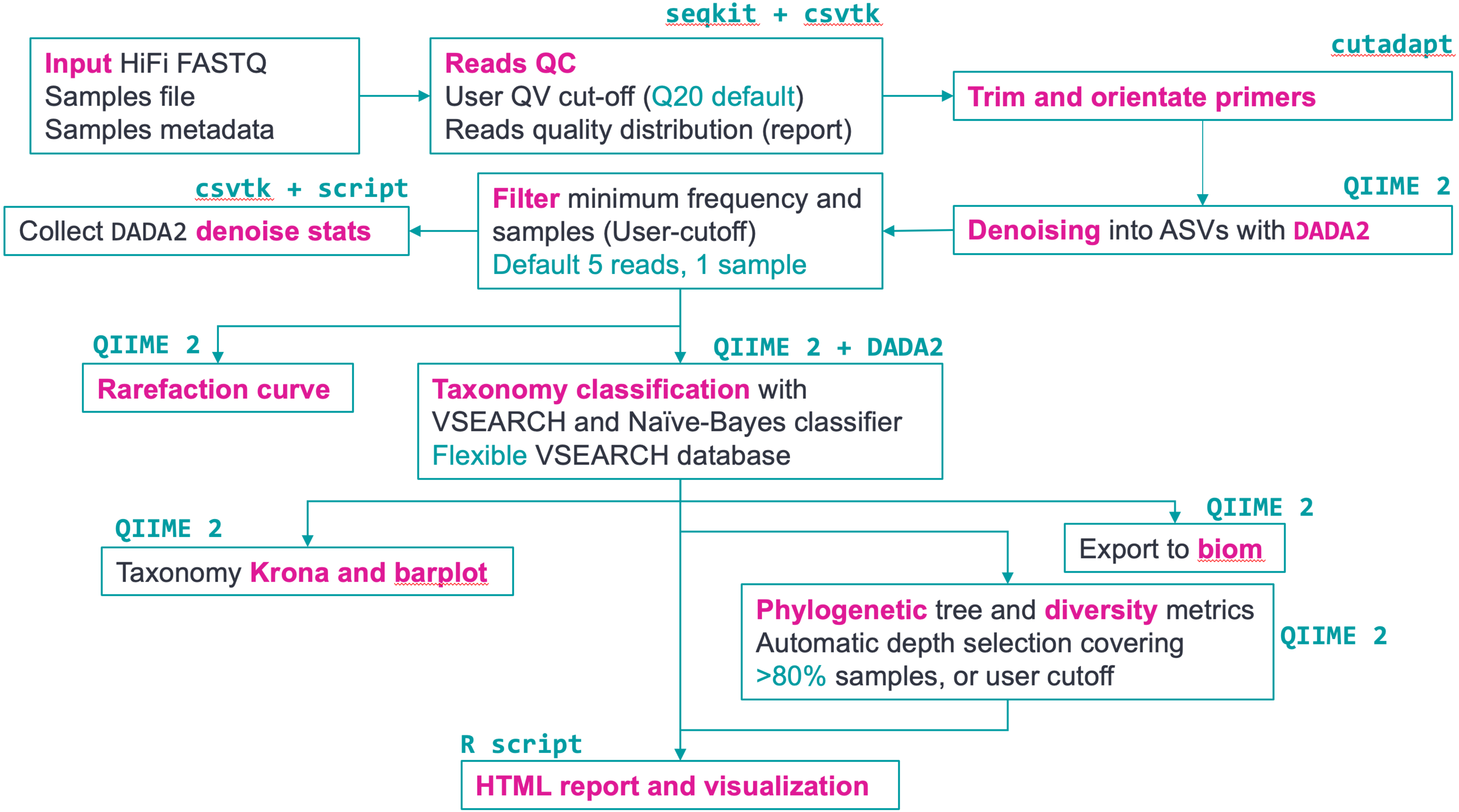
Eisenhofer, Wright, et al. (2024); biofortis.fr

PacBio full-length 16S rRNA



Advantages of full-length 16S rRNA

- Greater taxonomic and phylogenetic resolution
- More reliable for species and strain identification
- May have lower throughput than short-read methods



Input HiFi FASTQ
Samples file
Samples metadata

seqkit + csvtk

Reads QC
User QV cut-off (Q20 default)
Reads quality distribution (report)

cutadapt

Trim and orientate primers

QIIME 2

Denoising into ASVs with **DADA2**

Filter minimum frequency and samples (User-cutoff)
Default 5 reads, 1 sample

csvtk + script

Collect DADA2 **denoise stats**

QIIME 2

Rarefaction curve

QIIME 2 + DADA2

Taxonomy classification with VSEARCH and Naïve-Bayes classifier
Flexible VSEARCH database

QIIME 2

Taxonomy **Krona and barplot**

QIIME 2

Export to **biom**

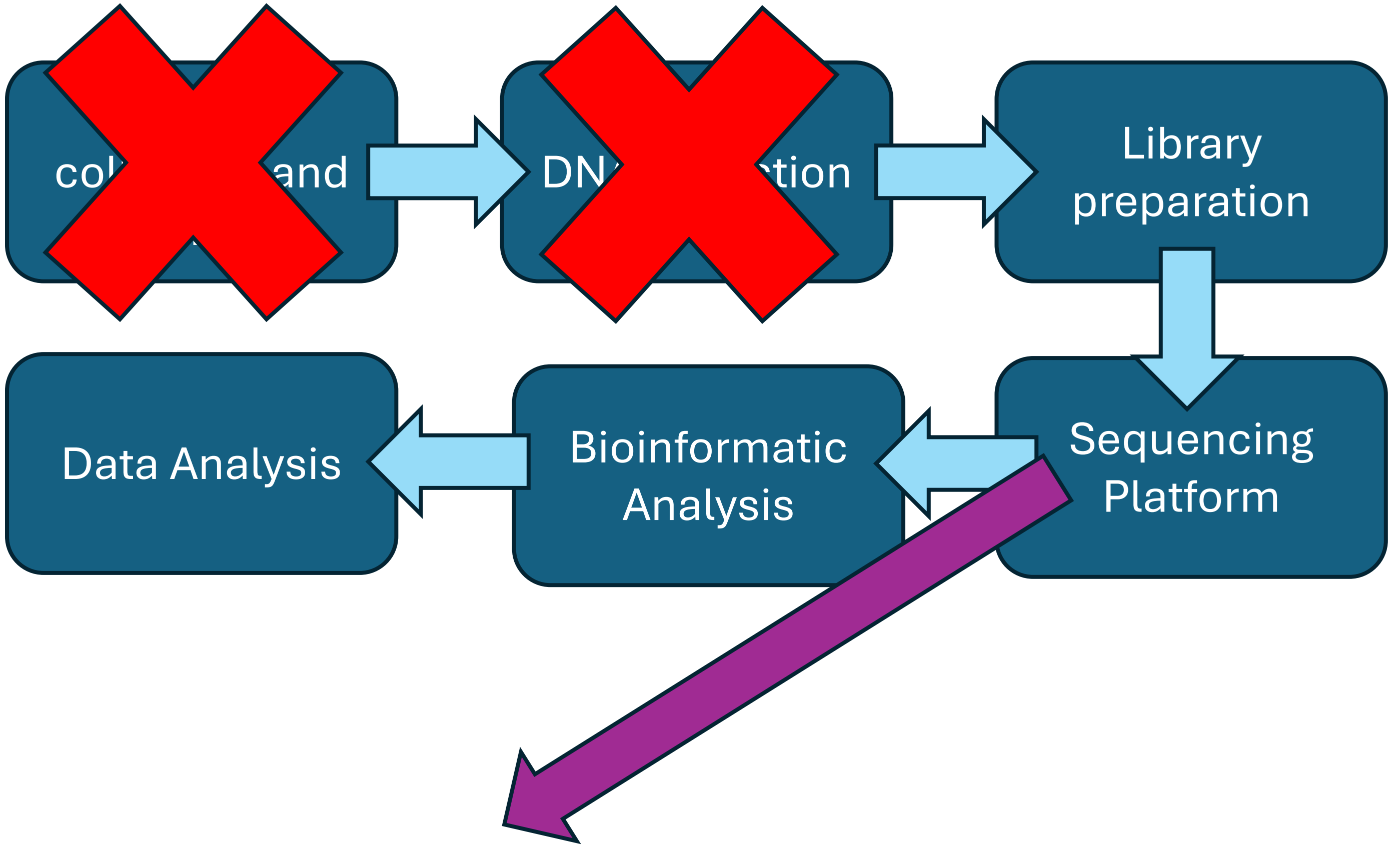
QIIME 2

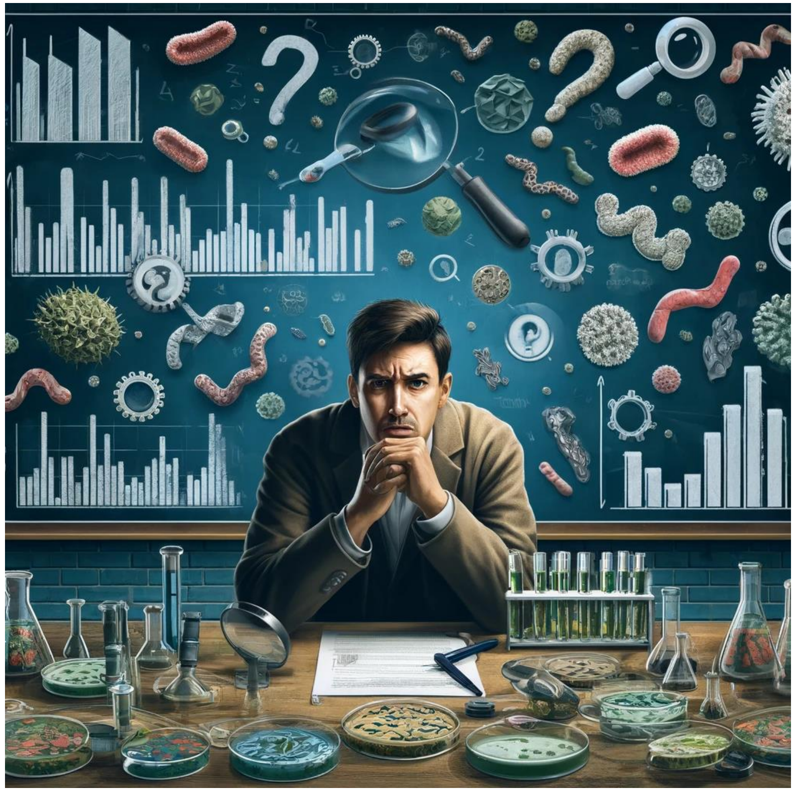
Phylogenetic tree and **diversity** metrics
Automatic depth selection covering >80% samples, or user cutoff

R script

HTML report and visualization

Overview of Microbiome Study Design

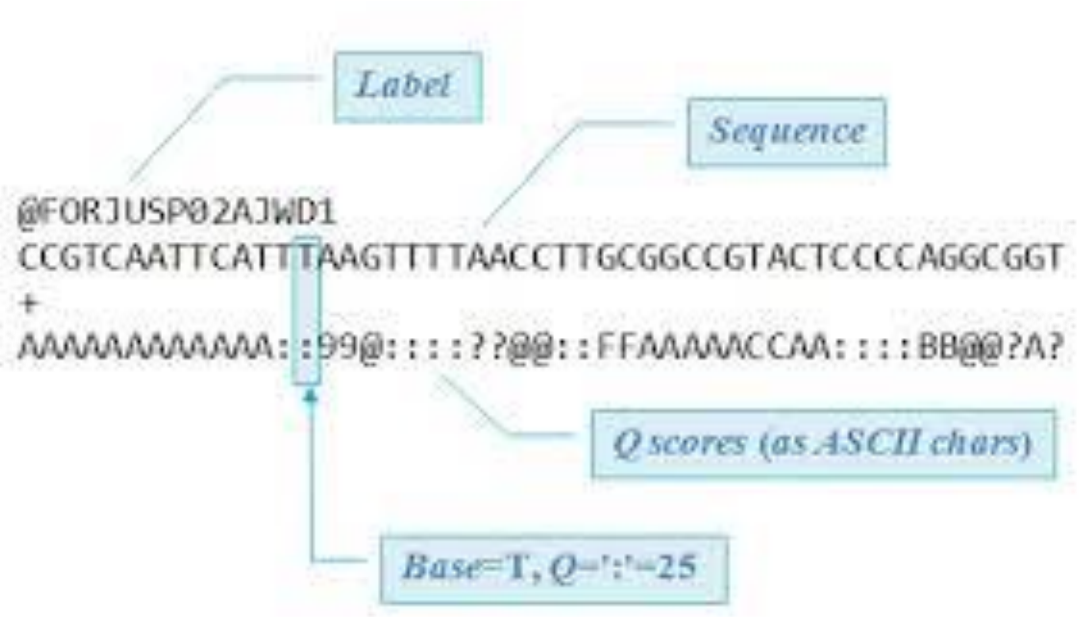




What does the sequencing data look like once it has been generated?

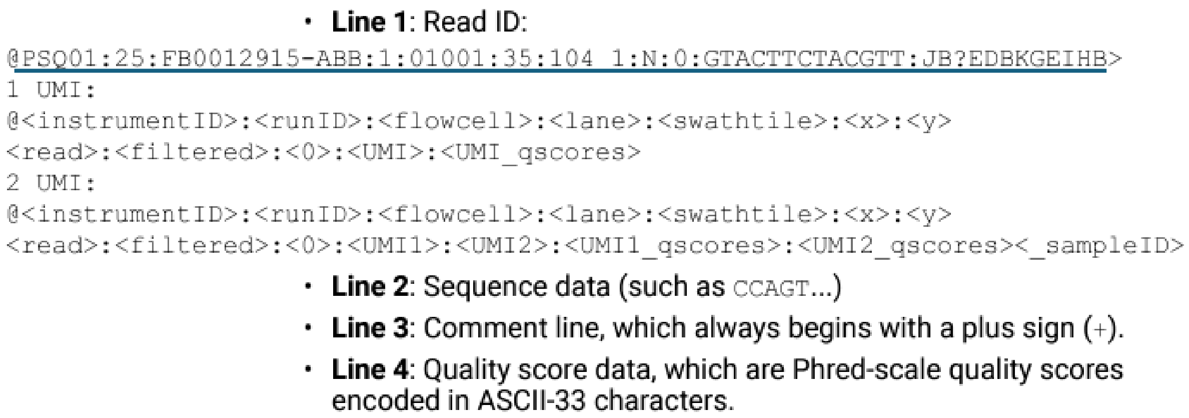
FASTQ Files: Starting point for bioinformatic analysis

Illumina FASTQ file



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

PacBio FASTQ file

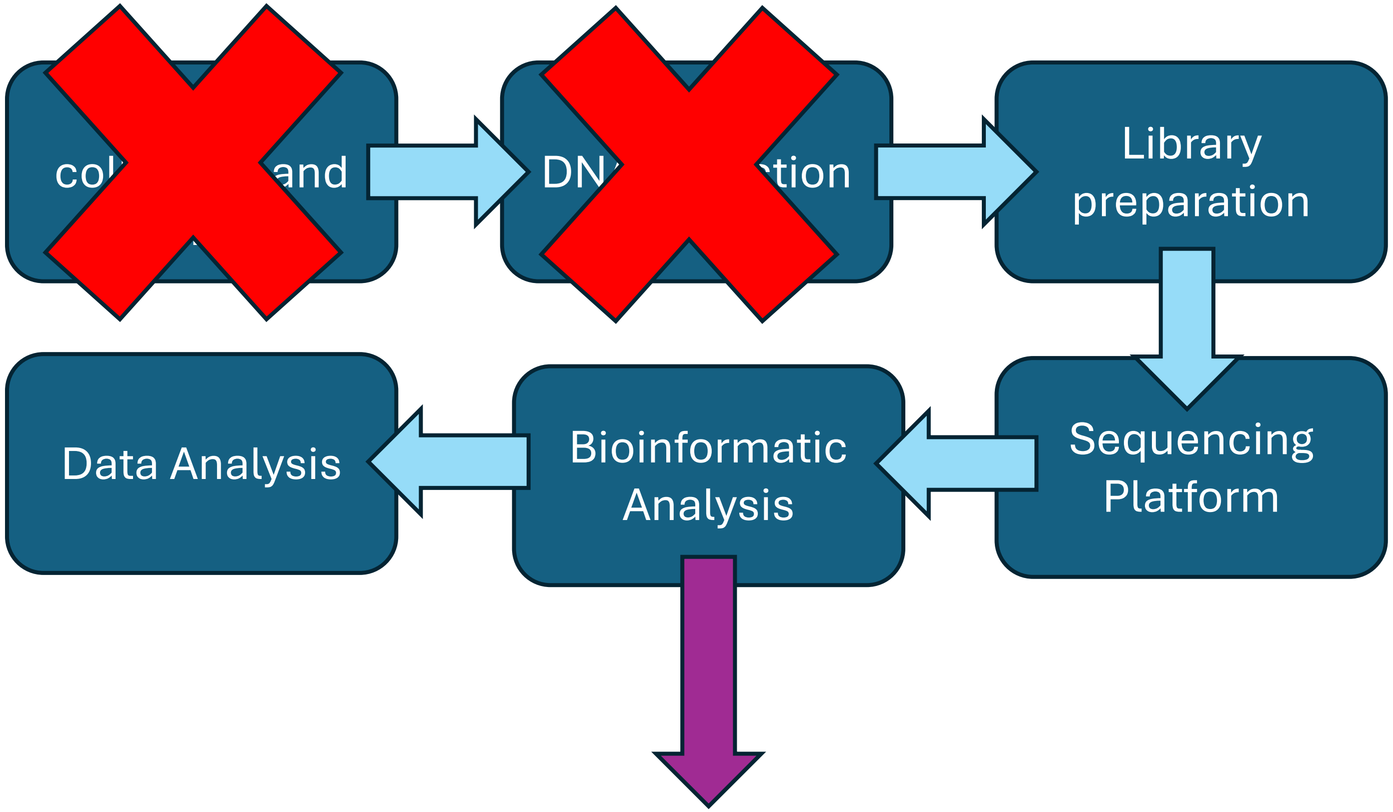


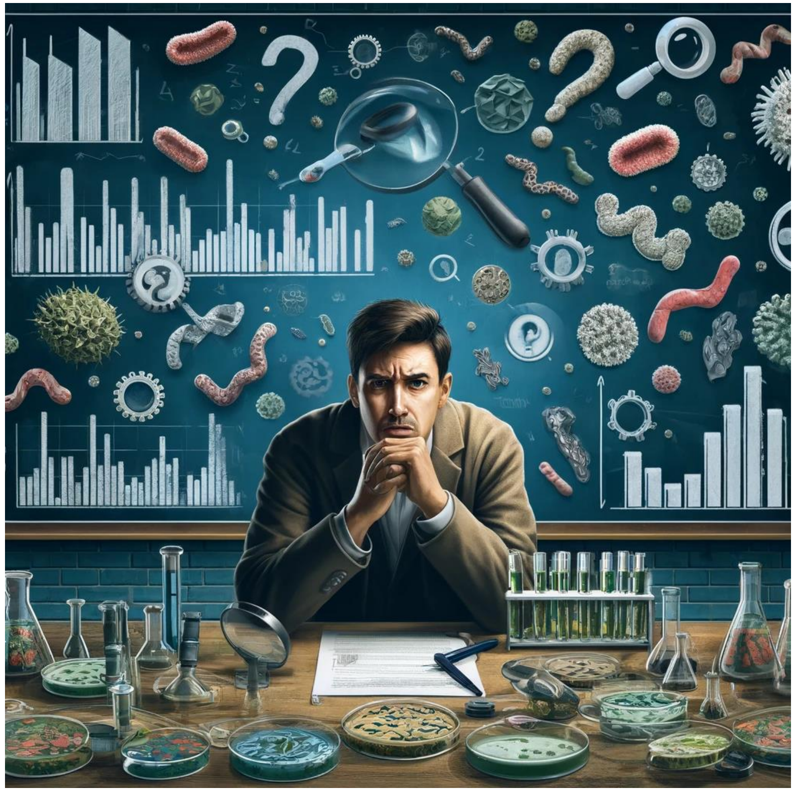
- **Line 1**: Read ID:

```
@PSQ01:25:FB0012915-ABB:1:01001:35:104 1:N:0:GTACTTCTACGTT:JB?EDBKGEIHB>
1 UMI:
@<instrumentID>:<runID>:<flowcell>:<lane>:<swathtile>:<x>:<y>
<read>:<filtered>:<0>:<UMI>:<UMI_qscores>
2 UMI:
@<instrumentID>:<runID>:<flowcell>:<lane>:<swathtile>:<x>:<y>
<read>:<filtered>:<0>:<UMI1>:<UMI2>:<UMI1_qscores>:<UMI2_qscores><_sampleID>
```

- **Line 2**: Sequence data (such as CCAGT...)
- **Line 3**: Comment line, which always begins with a plus sign (+).
- **Line 4**: Quality score data, which are Phred-scale quality scores encoded in ASCII-33 characters.

Overview of Microbiome Study Design





What to do with the sequencing data?

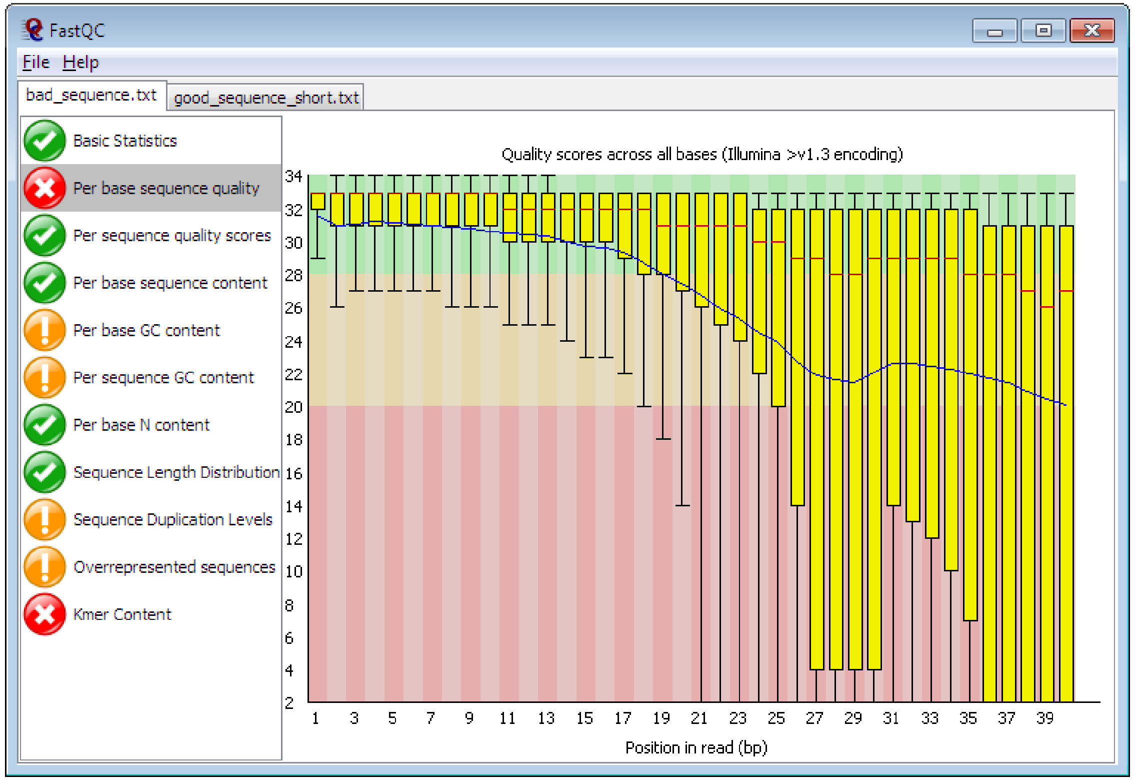
FASTQC

What is FASTQC?

- Provides a detailed report of key metrics to evaluate the quality of raw sequencing reads, including per base quality scores, GC content, sequence length distribution

Why use it?

- Ensures sequencing data is of sufficient quality
- Can identify adapter contamination, low-quality samples, or potential biases or artifacts



FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

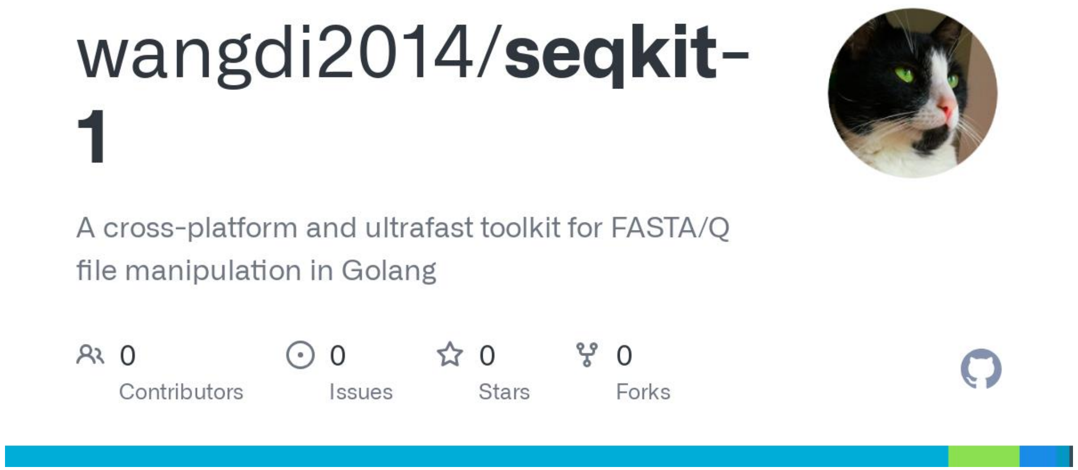
SEQKIT

What is seqkit?

- Highly efficient tool for handling and manipulating sequence data
- Broad range of functions
  - Subsample FASTQ files
  - Remove duplicates
  - Calculate basic stats

Why use it?

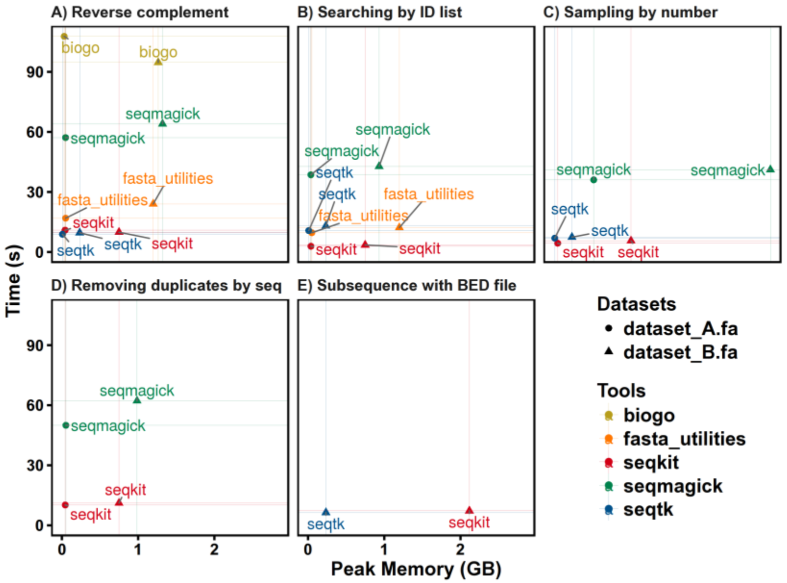
- Compared to other tools, user-friendly, versatile, scalable, and easily integrated in pipelines
- Further assess quality of sequencing data



wangdi2014/**seqkit-1**

A cross-platform and ultrafast toolkit for FASTA/Q file manipulation in Golang

0 Contributors · 0 Issues · 0 Stars · 0 Forks



What is QIIME2?

- A widely-used microbiome pipeline for processing and analyzing microbiome sequencing data
- Can carry out many processes including
  - Demultiplexing, denoising, taxonomic profiling, diversity analyses, statistical analyses, data visualizations, and extensibility

Why use it?

- Comprehensive workflow: provides full pipeline; from processing raw data to generating downstream results
- Reproducibility: easily able to track analyses
- Many plugins available for specific tasks
- Community support

Other available programs

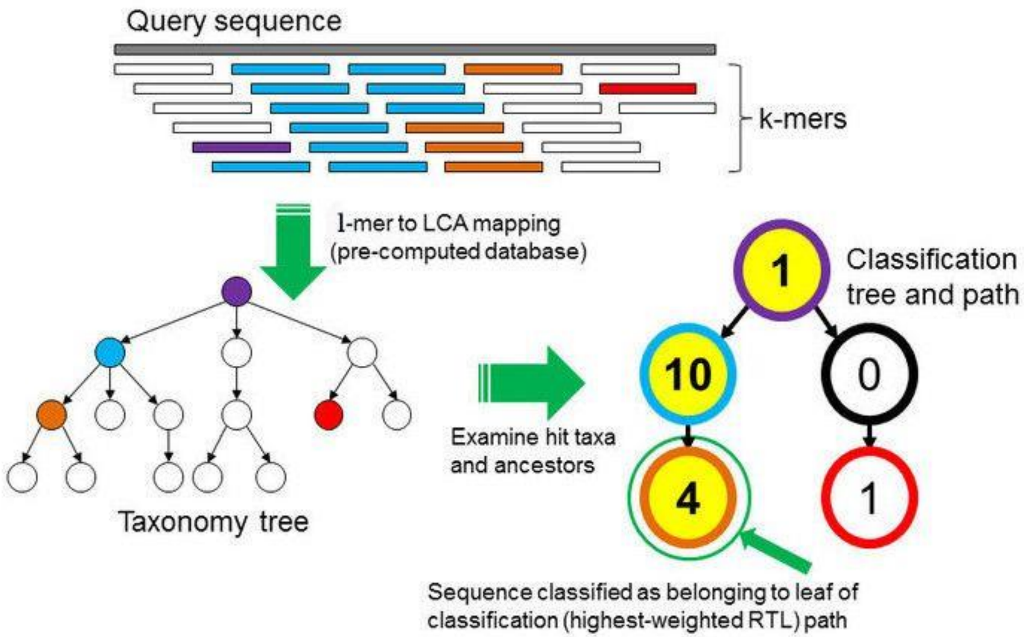
	QIIME2	Phyloseq	Vegan
Description	End-to-end bioinformatics pipeline for microbiome sequencing data using Linux	R package for microbiome data integration, visualization, and analysis	R package for general ecological analysis, including microbiomes
Accessibility	User-friendly command-line interface with interactive visualizations	Requires familiarity with R programming; offers excellent flexibility for experienced users	Steeper learning curve for non-R users; requires combining with other tools for microbiome-specific tasks
Community & Support	Large and active	Growing	Moderate
Limitations	Limited flexibility for visualizations; requires exporting data to R or python	Must use other pipelines for certain steps/analyses	Requires manual setup and integration with other tools for full microbiome analysis

DADA^2

Amplicon Sequencing. **Exactly.**

- Tool to preprocess 16S rRNA data
- Generates ASVs
  - Exact biological sequences inferred after denoising sequencing data
- Denoising: distinguish true biological sequences from errors
- Chimera removal: identify and remove chimeric sequences generated during PCR
- Produce unique sequences, ready for taxonomic profiling

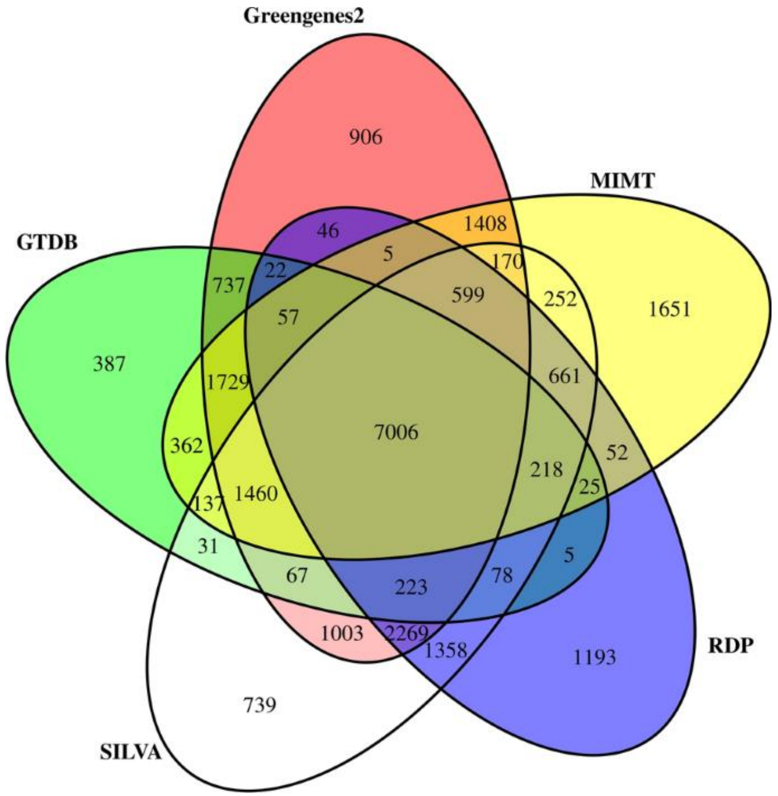
Taxonomic Profiling



- What is Taxonomic Profiling?
  - Identify and quantifying the microbial taxa present in a given sample
  - Three main approaches
    1. Genome based approach
       - Reads aligned to reference genomes
    2. Gene based approach
       - Reads are aligned to reference genes (e.g., marker genes)
    3. K-mer approach
       - Databases and DNA of samples are broken into length *k* for comparison

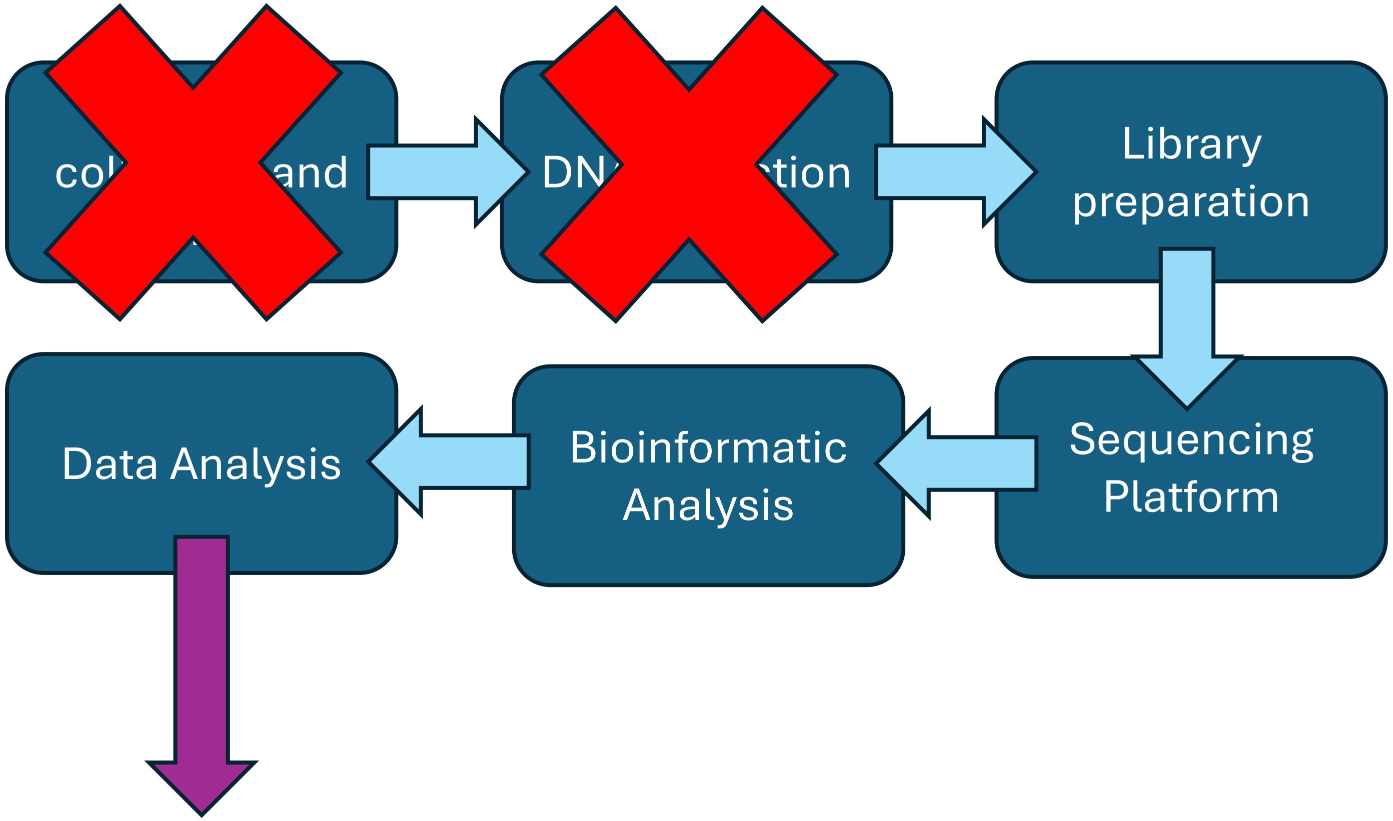
Databases for 16S rRNA sequencing

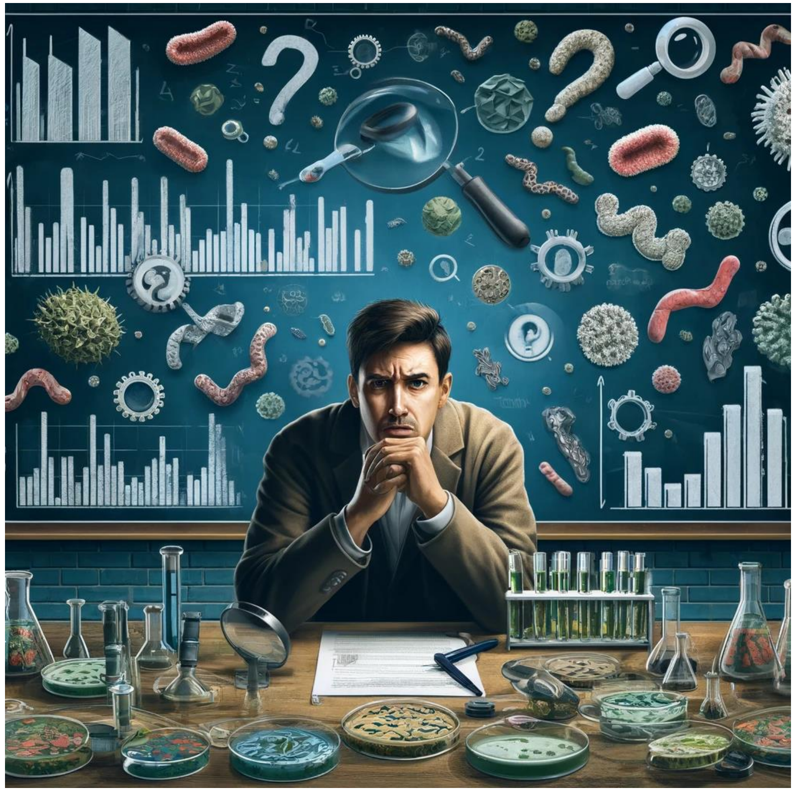
Database	Description
Greengenes2	Updated Greengenes database and includes full-length 16S genes
SILVA	A comprehensive database that has been widely used; last updated in 2020
Ribosomal Database Project (RDP)	Contains bacterial and archaeal SSU rRNA gene sequences and fungal large subunit gene sequences
Genome Taxonomy Database (GTDB)	Was developed to provide a standardized bacterial and archaeal taxonomy based on genome phylogeny; contains high redundancy; inflates species counts and errors in estimation methods



Cabezas et al (2024)

Overview of Microbiome Study Design





How do I analyze the data?

BIOM TABLE

	Sample 1	Sample 2	Sample 3
Species 1	7231	391	71818
Species 2	0	512	91009
Species 3	291	290	0
Species 4	192900	15900	0

Metadata	**Group #**	**Health**	**Factor 1**
Sample 1	Group 1	Healthy	0.01
Sample 2	Group 2	Diseased	0.42
Sample 3	Group 1	Healthy	0.55
Sample 4	Group 2	Diseased	0.75

BIOM (Biological Observation Matrix)

Cornerstone for microbiome analyses

Matrix Structure:

Each cell contains the abundance or count of a given feature in a sample

Each row includes taxonomic annotation

Each column contains sample ID

Widely supported by microbiome analysis tools

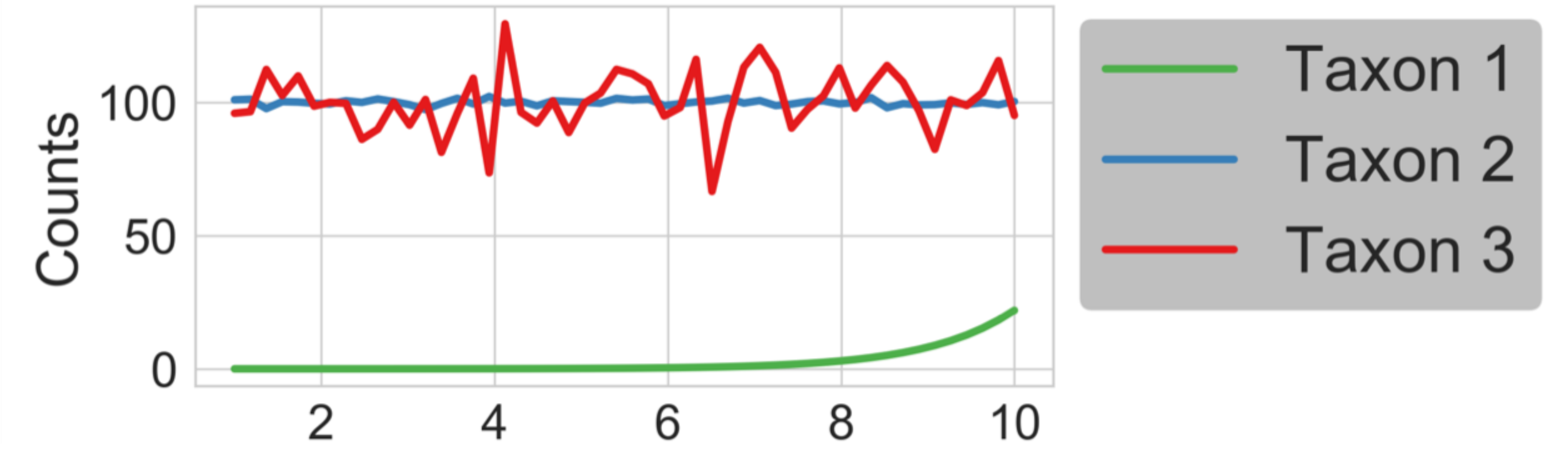
Supports hierarchical observations (e.g., nested taxonomies)

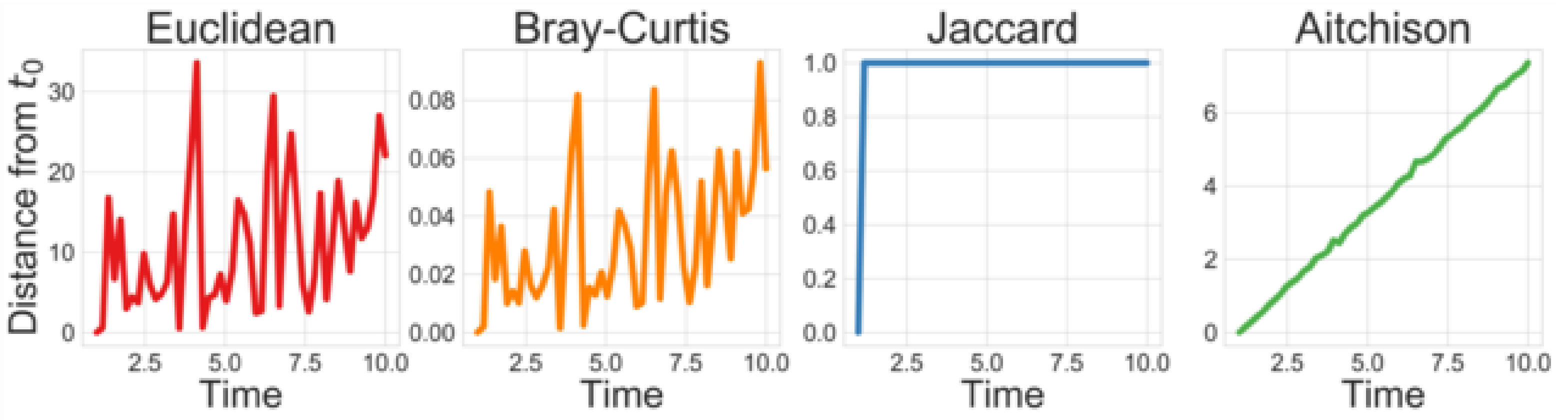
Anderson et al. (2001); Anderson (2014); Liu et al. (2021); Mallick et al (2021)

Key characteristics of the BIOM TABLE

1. **Highly sparse:**
   Most microbial communities will have many taxa that are absent in individual samples, leading to a high proportion of zero entries (zero-inflated)
2. **Multidimensional:**
   Hundreds or even thousands of taxa present in a dataset makes it highly dimensional
3. **Compositional nature:**
   Abundances are relative rather than absolute; violates assumptions of many classical statistical models (e.g., Euclidean distance, correlation), leading to spurious results

Accounting for Aitchison distance





Counts are proportional

Constant-sum constrain: increase the abundance of one taxon leads to the decrease abundance of another, regardless of whether their actual absolute abundances have changed

Leads to spurious correlations

Aitchison distance: metric designed for compositional data

Relies on log-ratio transformations

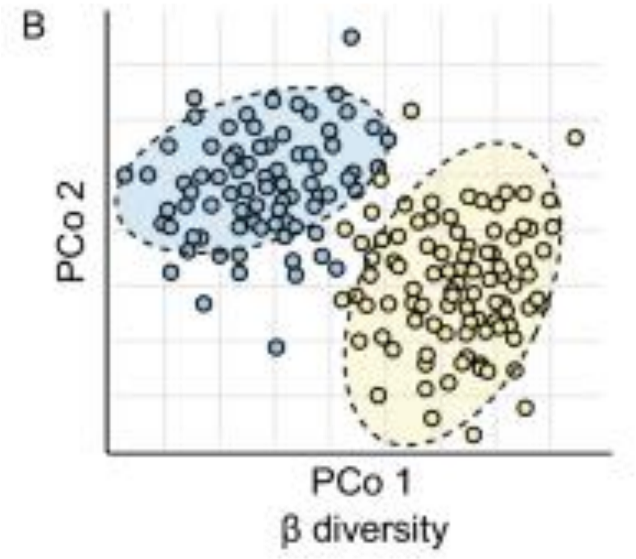
Useful for beta diversity, ordination, and statistical testing

Gloor et al. (2016)

Beta diversity

Aitchison distance matrix

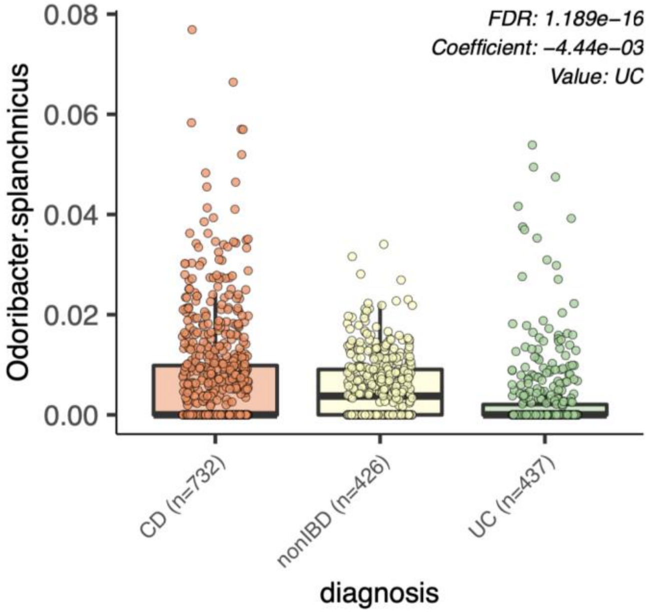
	Sample 1	Sample 2	Sample 3
Sample 1	0		
Sample 2	0.5	0	
Sample 3	0.7	0.45	0



Principal coordinates analysis (PCoA): visualization of sample similarity

Multifactorial PERMANOVA (adonis): statistical test to assess similarity between factors; provides a p-value and R2 value

	Sample 1	Sample 2	Sample 3
Species 1	7231	391	71818
Species 2	0	512	91009
Species 3	291	290	0
Species 4	192900	15900	0



Differential abundance analysis: Identify which species are enriched or depleted in groups

Many tools can carry out differential abundance analysis

Summary

- Microbiome datasets are highly complex and involve several steps for analyses
- Recognize the limitations of the tools that you use and how they could be impacting your downstream results
- Ask yourself how could your results be impacted and what needs to be done to make the results more robust
- Reproducibility is essential for microbiome research to move forward

Questions?

Next Up

15 Minute Break

Download the data and script from GitHub: https://github.com/asu-htm/Microbiome-Workshop-2024

Login to Sol