ASU Biodesign Institute
Arizona State University

ASU College of Health Solutions
Arizona State University

2025 HTM Workshop

Dr. Sterling L. Wright
Arizona State University

December 9th, 2025

Biodesign Institute

Learning Objectives

01

Understand the Nextflow's PacBio long-read workflow architecture

02

Be able to execute the full 16S PacBio Pipeline

03

Learn how to customize the pipeline

Overview of Microbiome Workflow

SAMPLE COLLECTION & STORAGE

DNA EXTRACTION

LIBRARY PREPARATION

SEQUENCING PLATFORM

BIOINFORMATIC PROCESSING

DATA ANALYSIS

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

Overview of Microbiome Study Design

SAMPLE COLLECTION & STORAGE

DNA EXTRACTION

LIBRARY PREPARATION

SEQUENCING PLATFORM

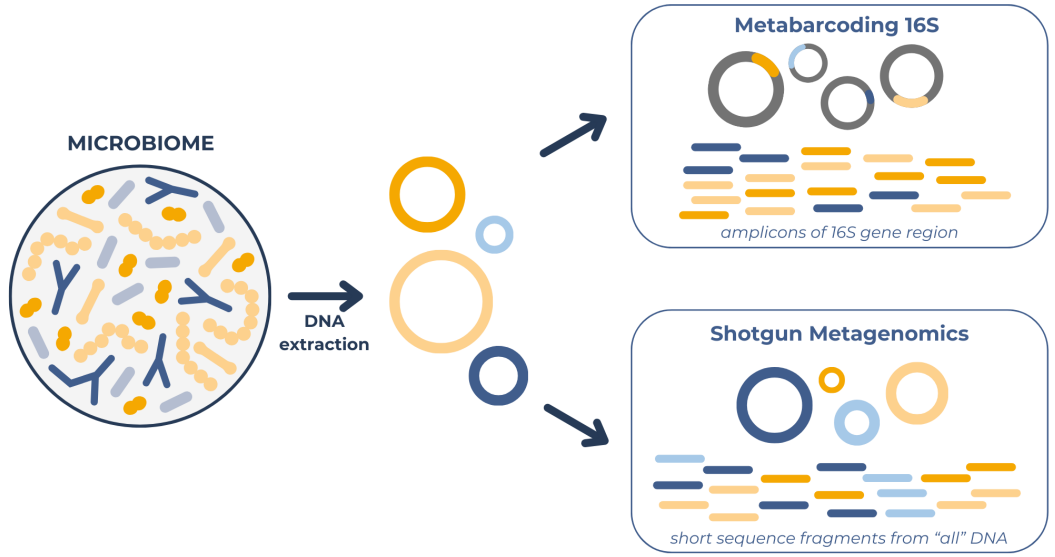
BIOINFORMATIC PROCESSING

DATA ANALYSIS

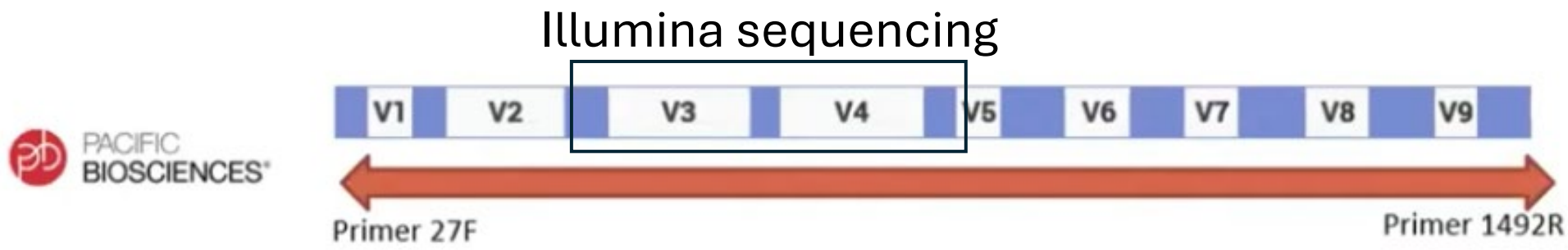
Forry et al. (2024); McGuinness et al. (2024); Duran-Pinedo et al. (2021)

16S vs Shotgun metagenomic basics

	16S rRNA sequencing	Shotgun Metagenomic Sequencing
Cost	~$50	~$150
Sample preparation	Utilizes primers	No primers
Functional profiling	Sort of	Better
Taxonomic resolution	Limited	Species and strains
Taxonomic coverage	Bacteria and archaea	All taxa, including viruses and host DNA
Databases	Established, well-curated	Less established and continually evolving
Bias	Requires *a priori*	Uses an untargeted approach



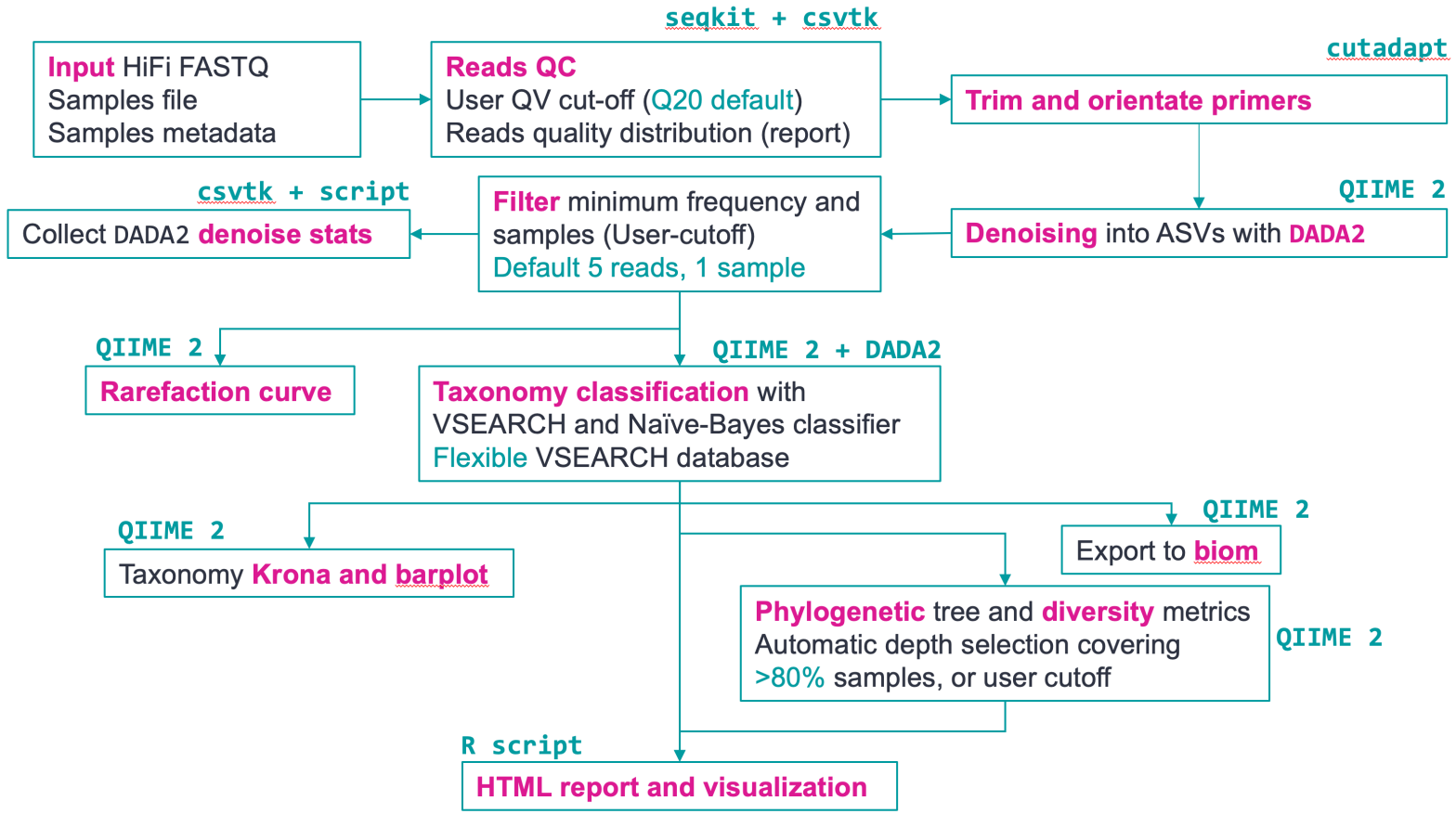
PacBio full-length 16S rRNA



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

PacBio 16S Nextflow Pipeline



FASTQ Files: Raw Sequencing Data Files

- **Line 1**: Read ID:

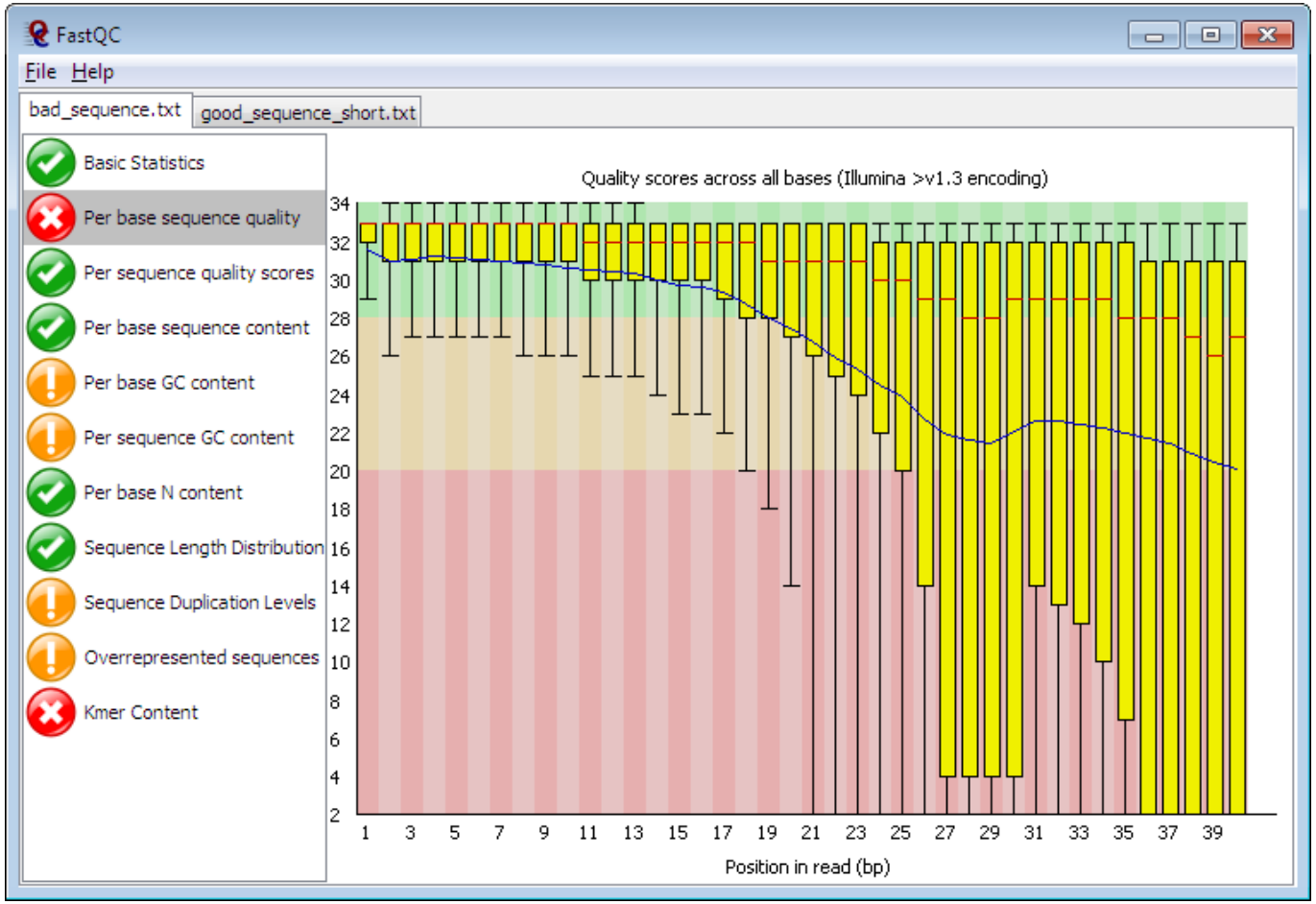
```
@PSQ01:25:FB0012915-ABB:1:01001:35:104 1:N:0:GTACTTCTACGTT:JB?EDBKGEIHB>
1 UMI:
@<instrumentID>:<runID>:<flowcell>:<lane>:<swathtile>:<x>:<y>
<read>:<filtered>:<0>:<UMI>:<UMI_qscores>
2 UMI:
@<instrumentID>:<runID>:<flowcell>:<lane>:<swathtile>:<x>:<y>
<read>:<filtered>:<0>:<UMI1>:<UMI2>:<UMI1_qscores>: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

FASTQC

What is FASTQC?

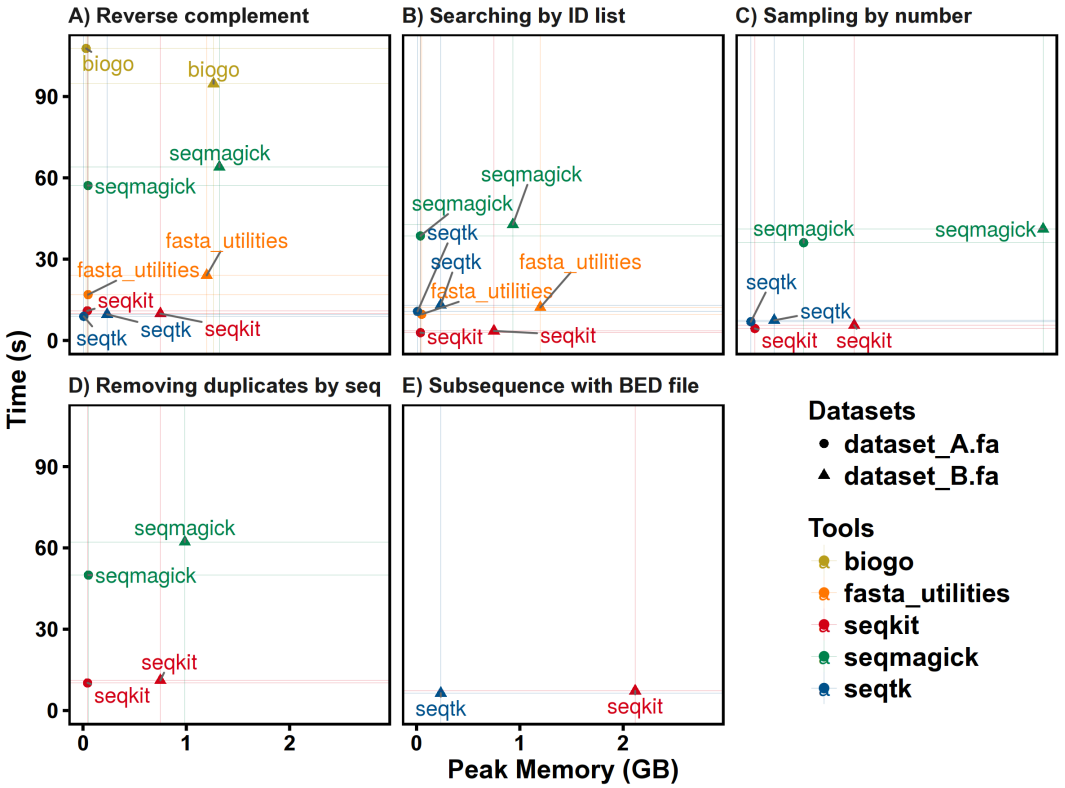
- Provides a detailed report of key metrics to evaluate the quality of raw sequencing reads
- Can identify adapter contamination, low-quality samples, or potential biases or artifacts



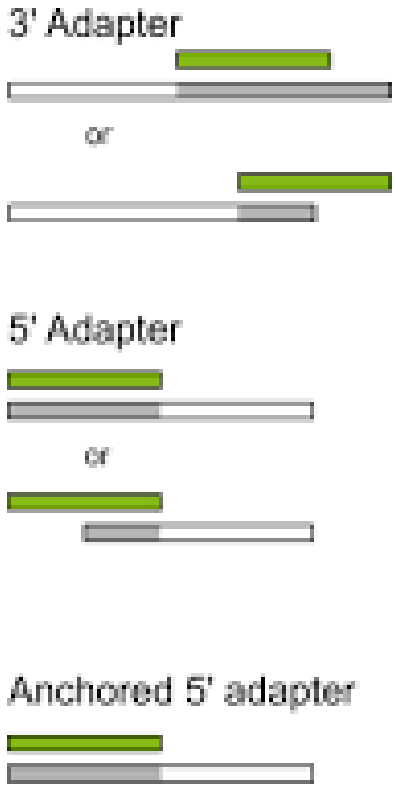
SEQKIT

What is seqkit?

- Highly efficient tool for handling and manipulating sequence data
- Broad range of functions
- Compared to other tools, user-friendly, versatile, scalable, and easily integrated in pipelines



Cutadapt



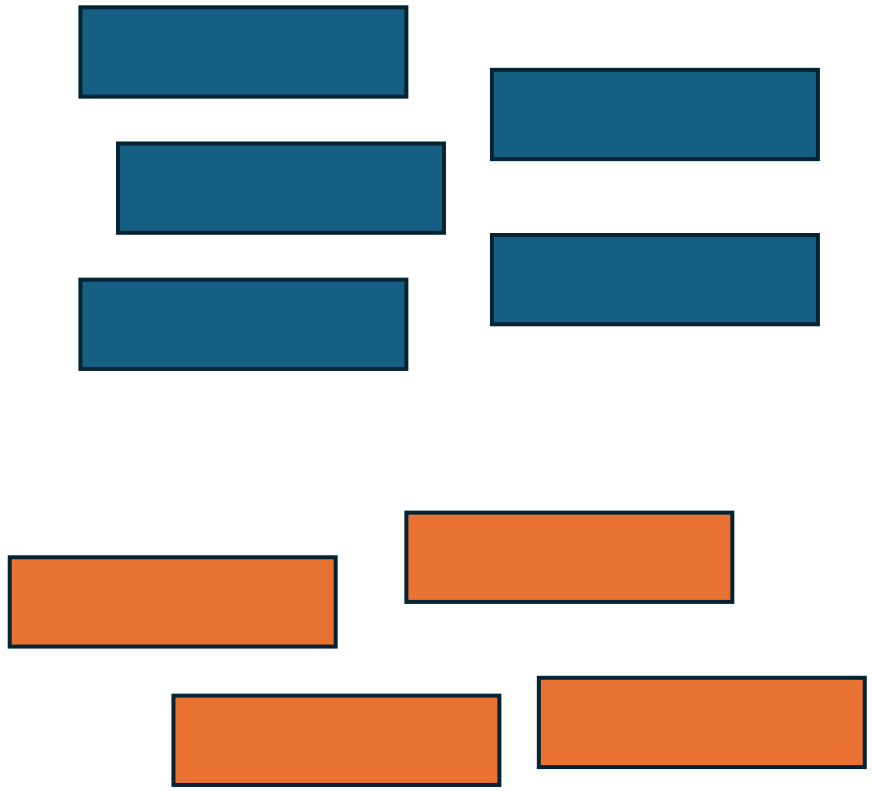
Remove unwanted sequences

- Primers
- Residual adapter sequences

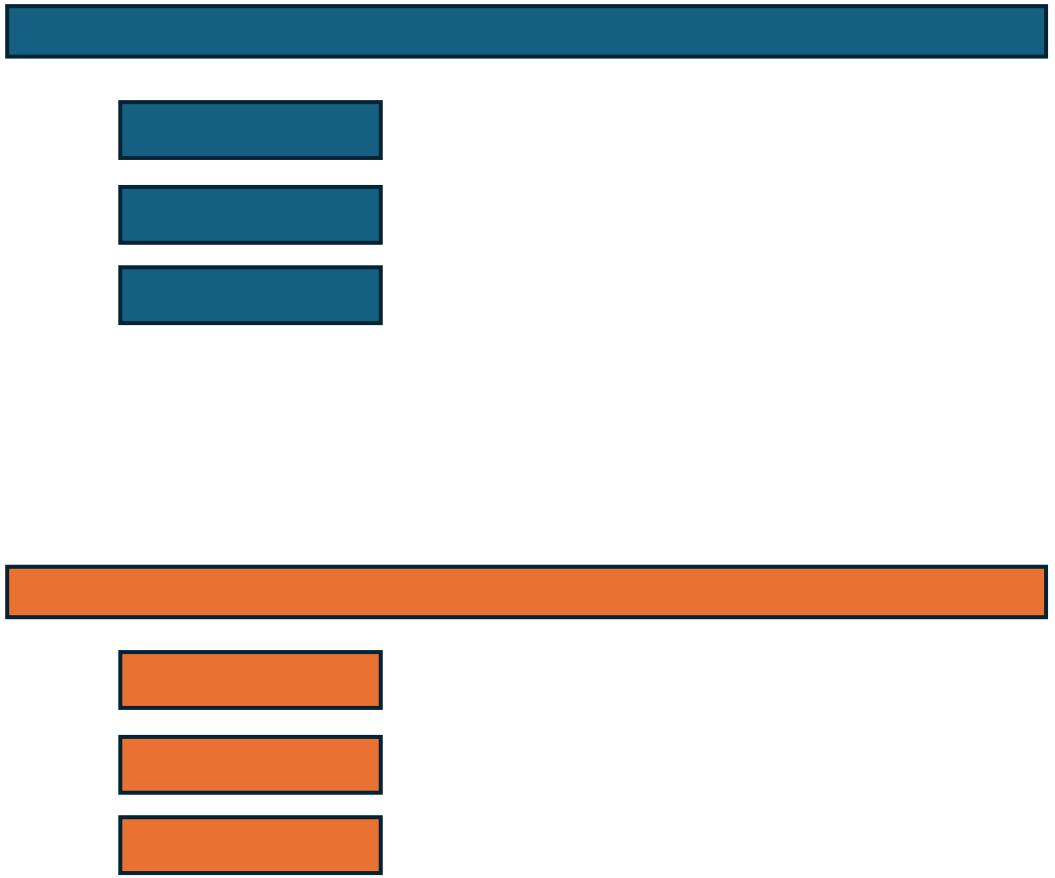
- Tool to preprocess 16S rRNA data
- Denoising: distinguish true biological sequences from errors
- Chimera removal: identify and remove chimeric sequences generated during PCR
- Generates Amplicon Sequence Variants (ASVs)

OTU Clustering vs ASV alignment

Operational Taxonomic Units
Cluster of microbes grouped according to their DNA sequence similarity



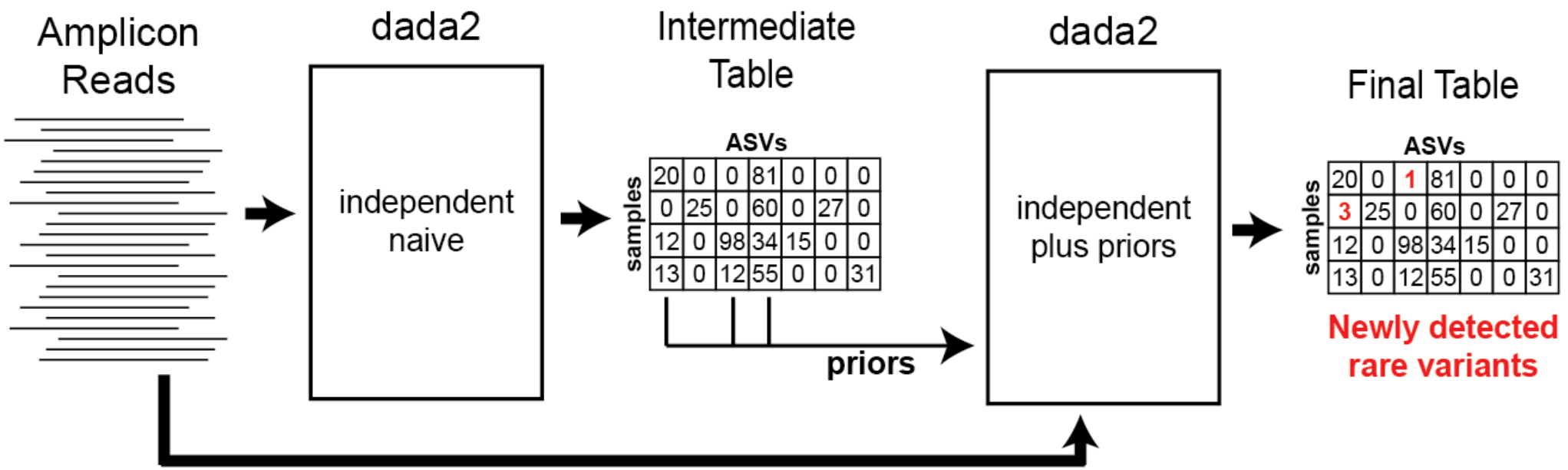
Amplicon Sequence Variants
Exact biological sequences



OTU vs ASV

Feature	OTU	ASV
Resolution	Cluster sequences at a percentage (97% or 99%)	Single-nucleotide precision
Error handling	Errors can be absorbed in clustering	Uses algorithms to denoise and correct errors
Reproducibility	May vary between studies based on clustering	Exact sequence variants, reproducible results
Computational cost	Less computationally demanding	Higher computational needs because of denoising
Interpretation	Groupings can mix closely related species	More precise, can distinguish variations

Pseudo-Pooling



Mode	How it works	Sensitivity	False Positives
Pseudo-pooling	Use cross-sample ASV candidates to refine sample-level inference	High	Low
Independent	Each sample alone	Lowest	Lowest
Full pooling	All samples treated as one big dataset	Highest	Higher

Taxonomic Classification

What is Taxonomic Classification?

- Identify and quantifying the microbial taxa present in a given sample
- Three main approaches
  - **Genome based approach**
    - Reads aligned to reference genomes
  - **Gene based approach**
    - Reads are aligned to reference genes (e.g., marker genes/16S)
  - **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)

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

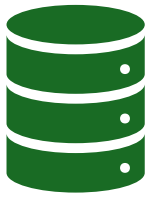
Supports hierarchical observations (e.g., nested taxonomies)

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

Key characteristics of the BIOM TABLE

Highly sparse:

Most datasets will have many taxa that are absent in several samples, leading to a high proportion of zero entries (i.e., zero-inflated)



Multidimensional:

Hundreds or even thousands of taxa present in a dataset makes it highly dimensional

Compositional nature:

Abundances are relative rather than absolute; violates assumptions of many classical statistical models (e.g., Euclidean distance, correlation), leading to spurious results

What Test Should I Perform?

- Different statistical methods rely on different assumptions
- Combining results lead to incompatible inferential frameworks
- Many tests without a hypothesis leads to interpretive incoherence
- Statistical models are not truth-generating machines; they quantify uncertainty
- Don’t confuse epistemic uncertainty with algorithmic diversity

Thank you!

Login to SOL

Prepare for coding part of the workshop