



UNIVERSITY OF  
CALGARY

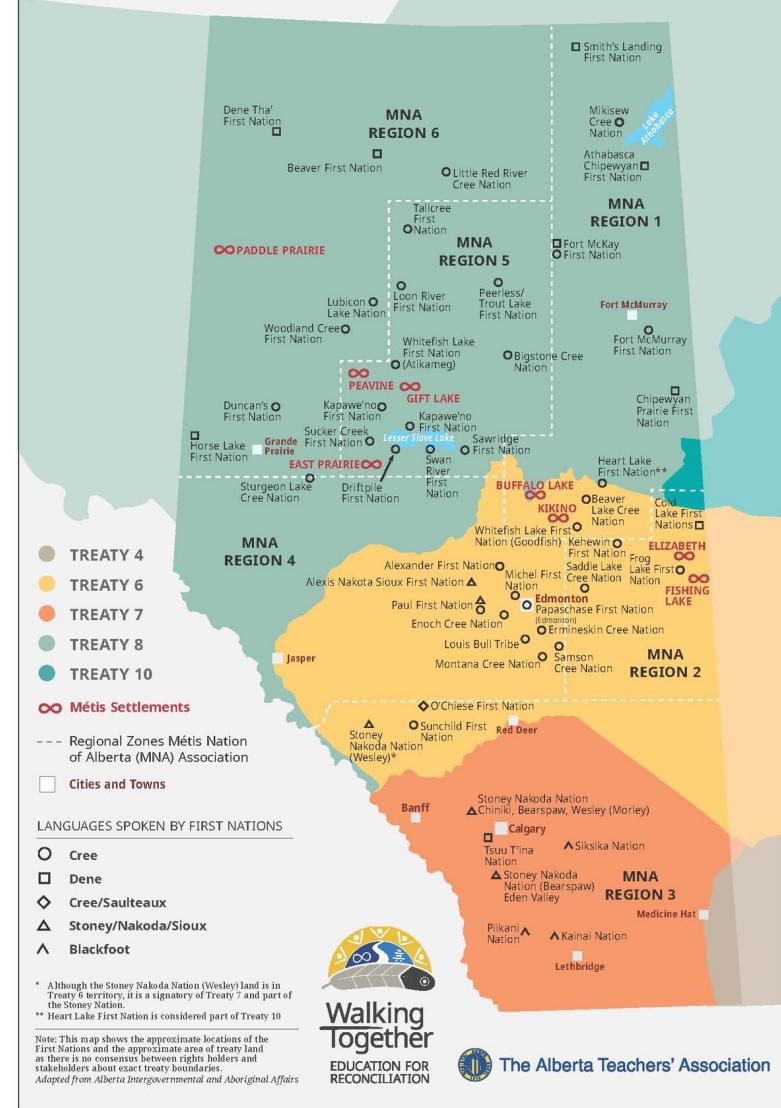
International  
**Microbiome**  
Centre

# Microbiome Workshop

**Hena R. Ramay**  
Bioinformatician  
International Microbiome Centre  
University of Calgary

I'd like to acknowledge that we are on Treaty 7 territory, the traditional territories of the Blackfoot Nations, including Siksika (Sick-sick-ah), Piikani (Pee-can-ee), and Kainai (Kigh-a-nigh), the Tsuut'ina (Soot- ina), Nation and Stoney Nakoda First Nations. We acknowledge all the many First Nations, Métis, and Inuit whose footsteps have marked these lands for centuries.

## ACKNOWLEDGING LAND AND PEOPLE



Why do we care about microbes?

How are they communicating with us?

# Omics

# Biomaterial

Meta-genome

DNA

Meta-Transcriptome

RNA

Protein

Proteins

Metabolome

Sugar Amino Acids Nucleotide SCFA

# Meta-genome

## Shotgun Sequencing Output

### Amplicon Sequencing Output

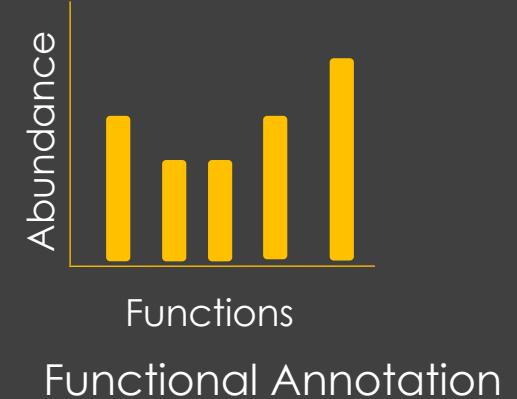


Phylogeny

+

GATC**G**ATC  
GATC**G**ATC  
GATC**C**CATC  
GATC**C**CATC

Polymorphism



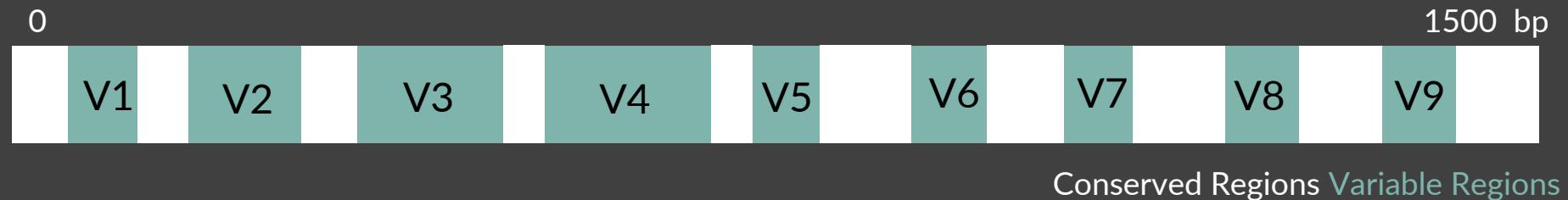
Functional Annotation

**ASV:** Amplicon Sequence Variant

# Amplicon Sequencing

What should we amplify?

# The Small Subunit 16s ribosomal RNA gene

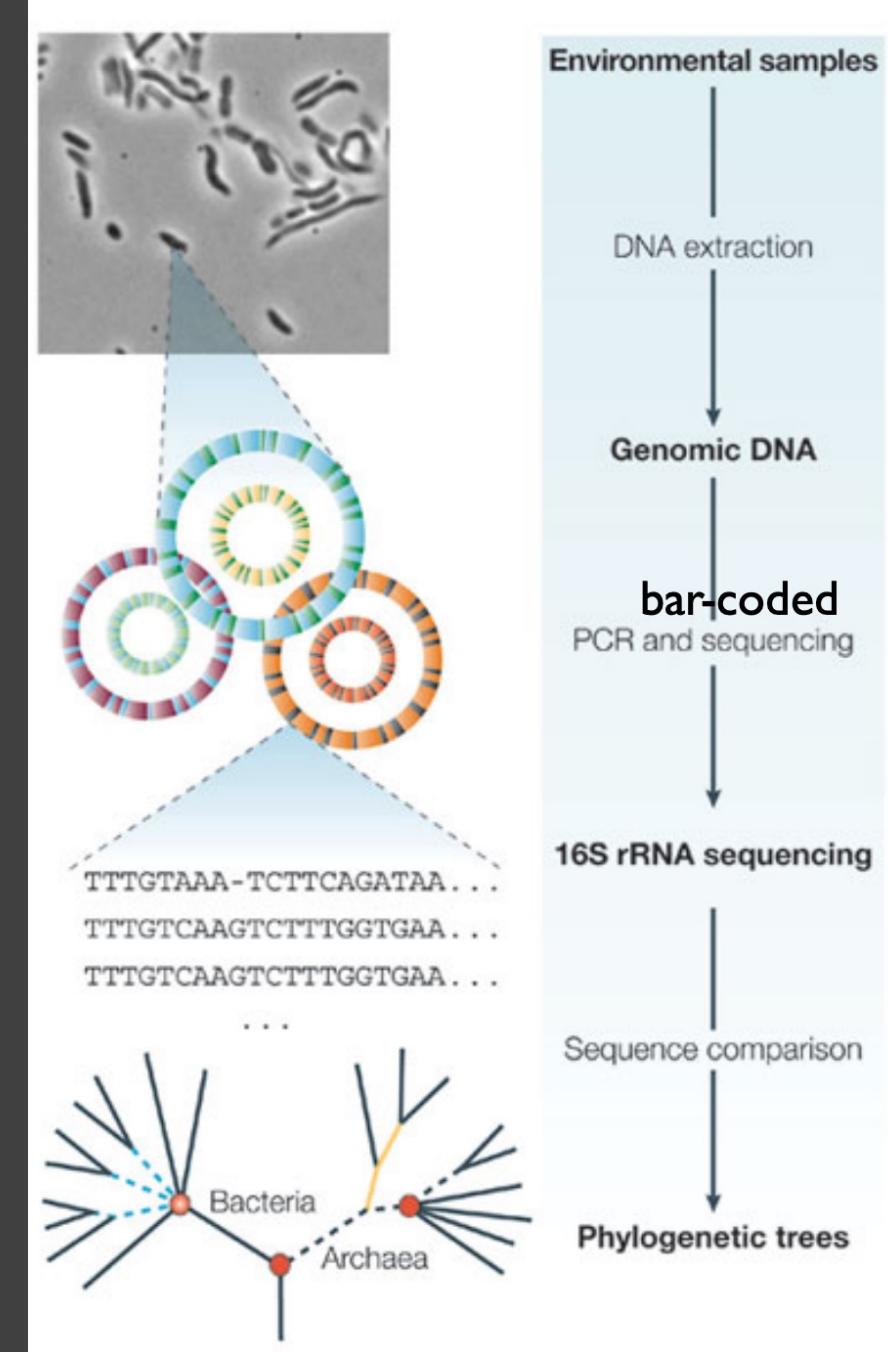


Highly conserved gene found in most bacteria

A variety of hypervariable region

# Many microbiomes in parallel

1. Break all cells, extract all DNA
2. PCR-amplify 16s rRNA genes using bar-coded primers
3. Sequence samples
4. Cluster sequences after De-multiplexing for each sample
5. Count each species





Amplify 16s region →



**ASV Table**

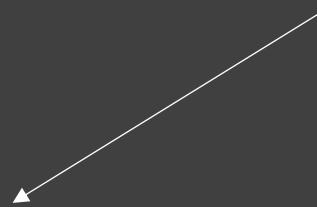
	Sample 1	Sample 2	Sample 3
ASV 1	0.5	0.4	0.6
ASV 2	0.2	0.3	0.1
ASV 3	0.2	0.1	0.2
ASV 4	0.1	0.2	0.1

## Amplicon Sequencing Cycle



Phylogeny

**Bacteria + Archaea**



**ASV 1 = Bug X**

**ASV:** Amplicon Sequence Variant  
is a sequence detected with a  
certain abundance in one or  
more samples

# Typical Workflow

**Thoughtful data analysis is  
critical for successful  
taxonomic assignment**

# What do we know about our data?

## 16s Region

Which region was sequenced, read length & depth

## Read Assignment

Are the reads assigned correctly to each sample?

## Controls

Did you use negative (extraction) and positive (mocks) ?

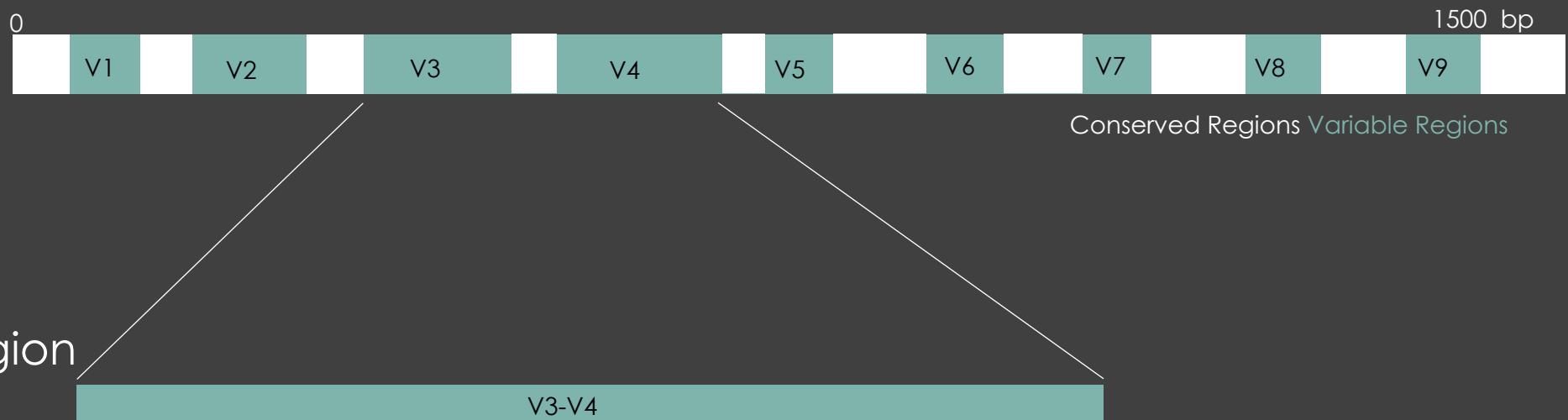
## Sample size

Are there enough samples for downstream analysis

## Feasibility

Will this data answer the questions asked by the investigator?

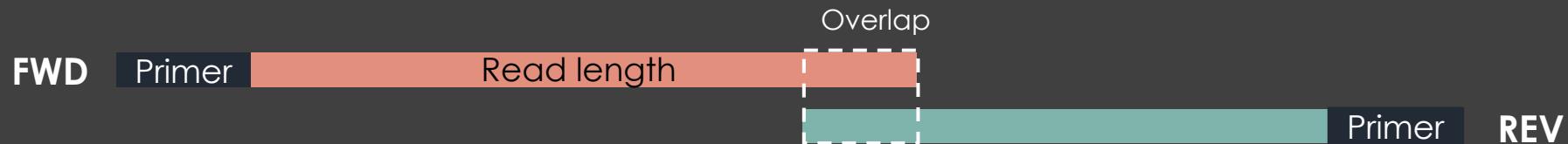
16 rRNA



Pick a sequencing technique

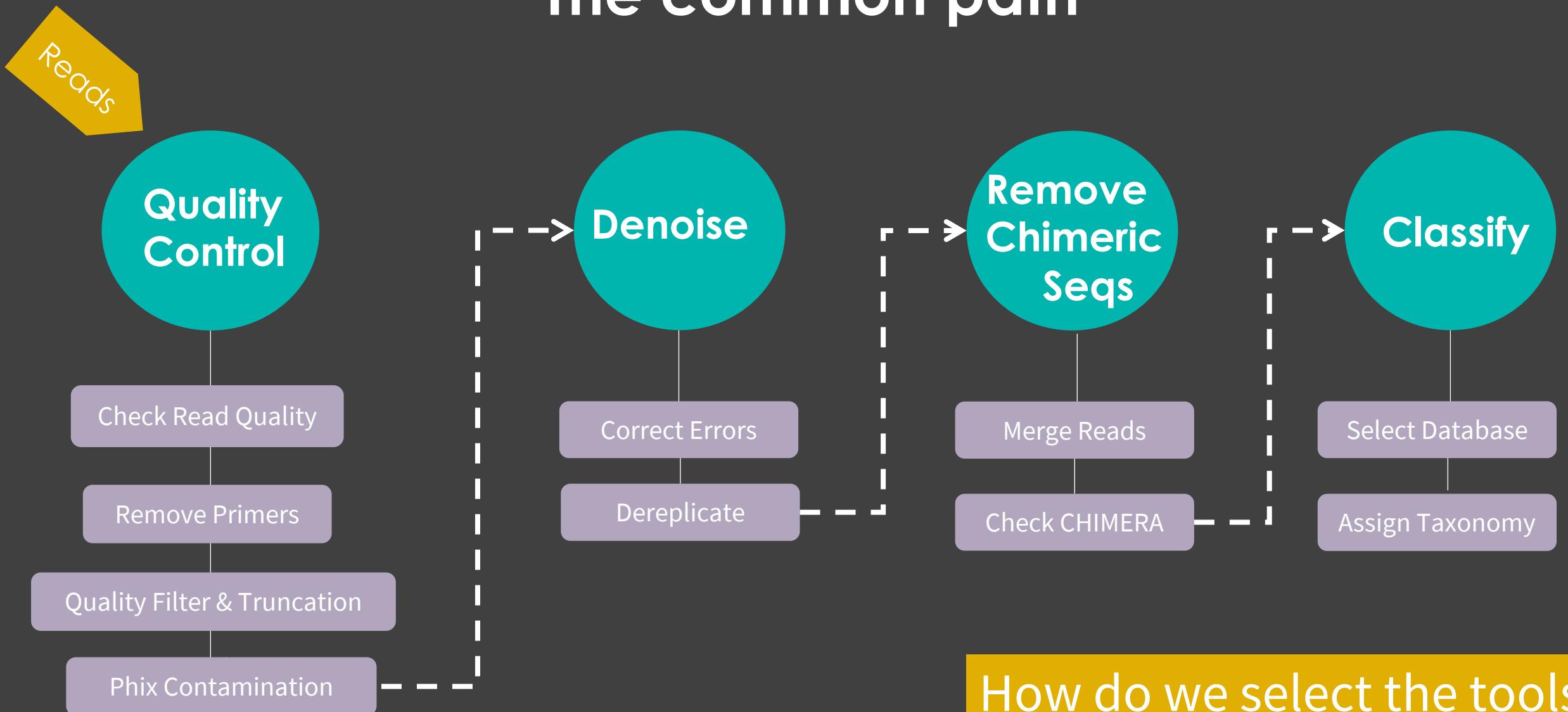
Illumina (MiSeq)

Select primers & Read lengths



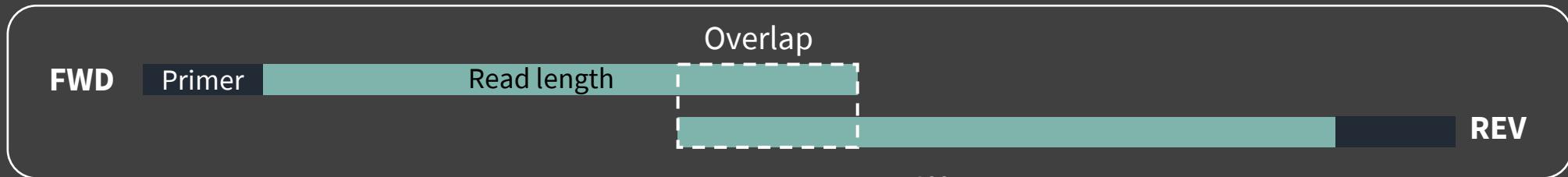
Make sure there is an overlap!!!!

# The common path

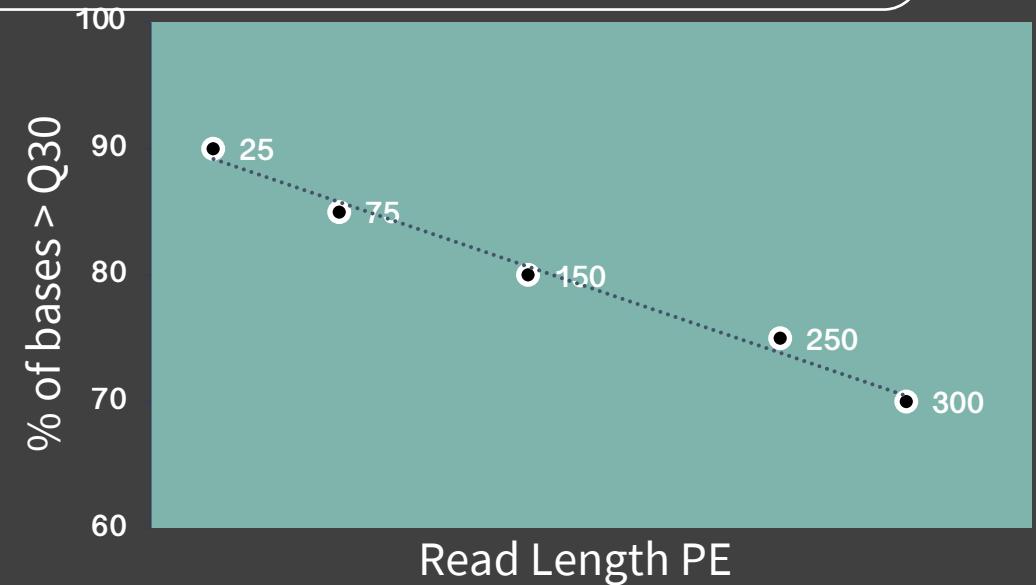


How do we select the tools  
and use them well ?

# Paired End Reads



Region	Read length	Amplicon Length	Overlap
V3	150	~170	130
V3-V4	300	~462	133
V4	150	~254	46
V4	250	~254	246



# Base Quality differs between Fwd and Rev Reads



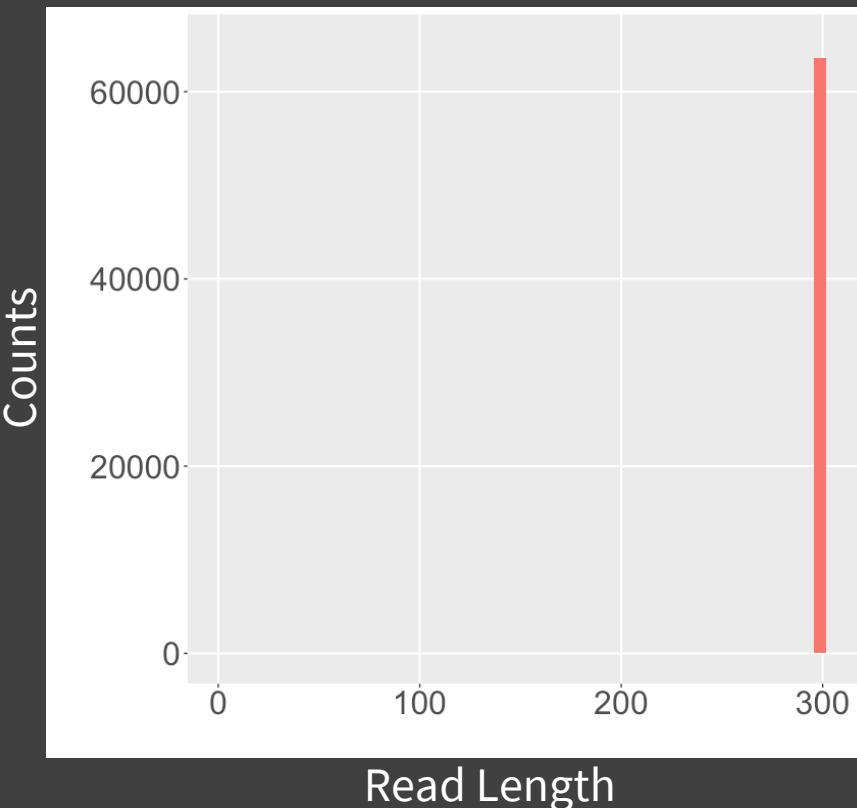
## Tools

FastQC FastQp MultiQC & Specialized 16s rRNA packages

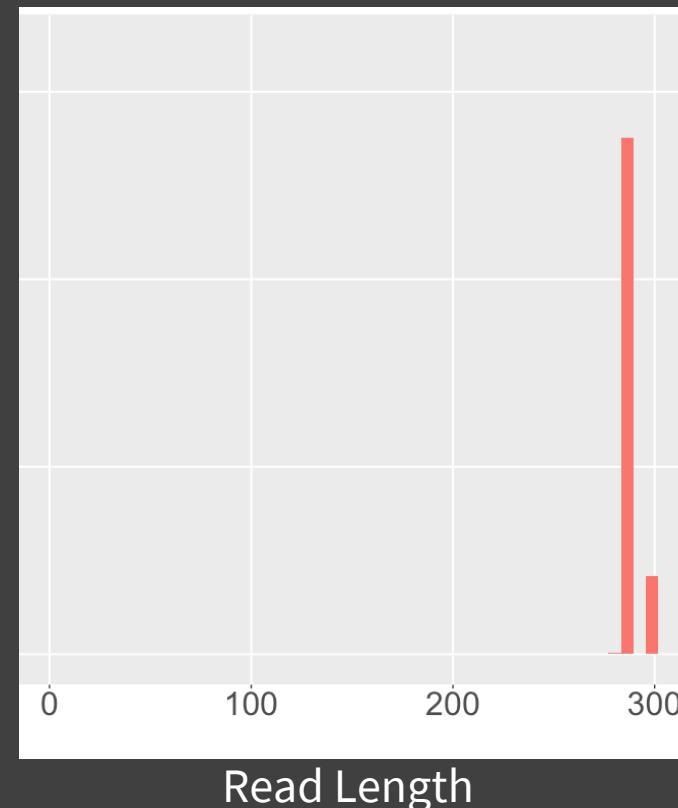
Source: Dada2 R package

# Trimming & Quality Filtering

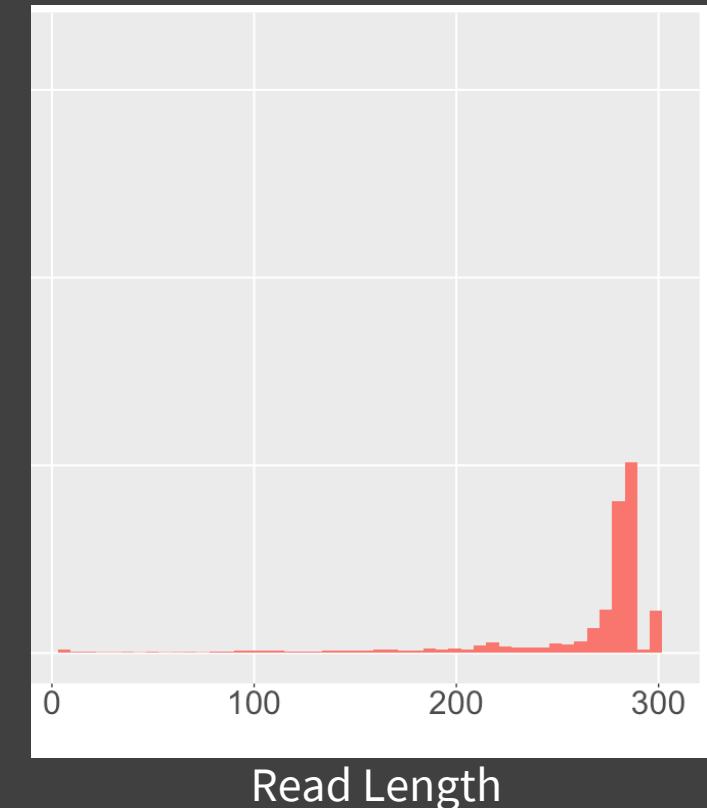
**Raw Reads**



**Remove Primers**



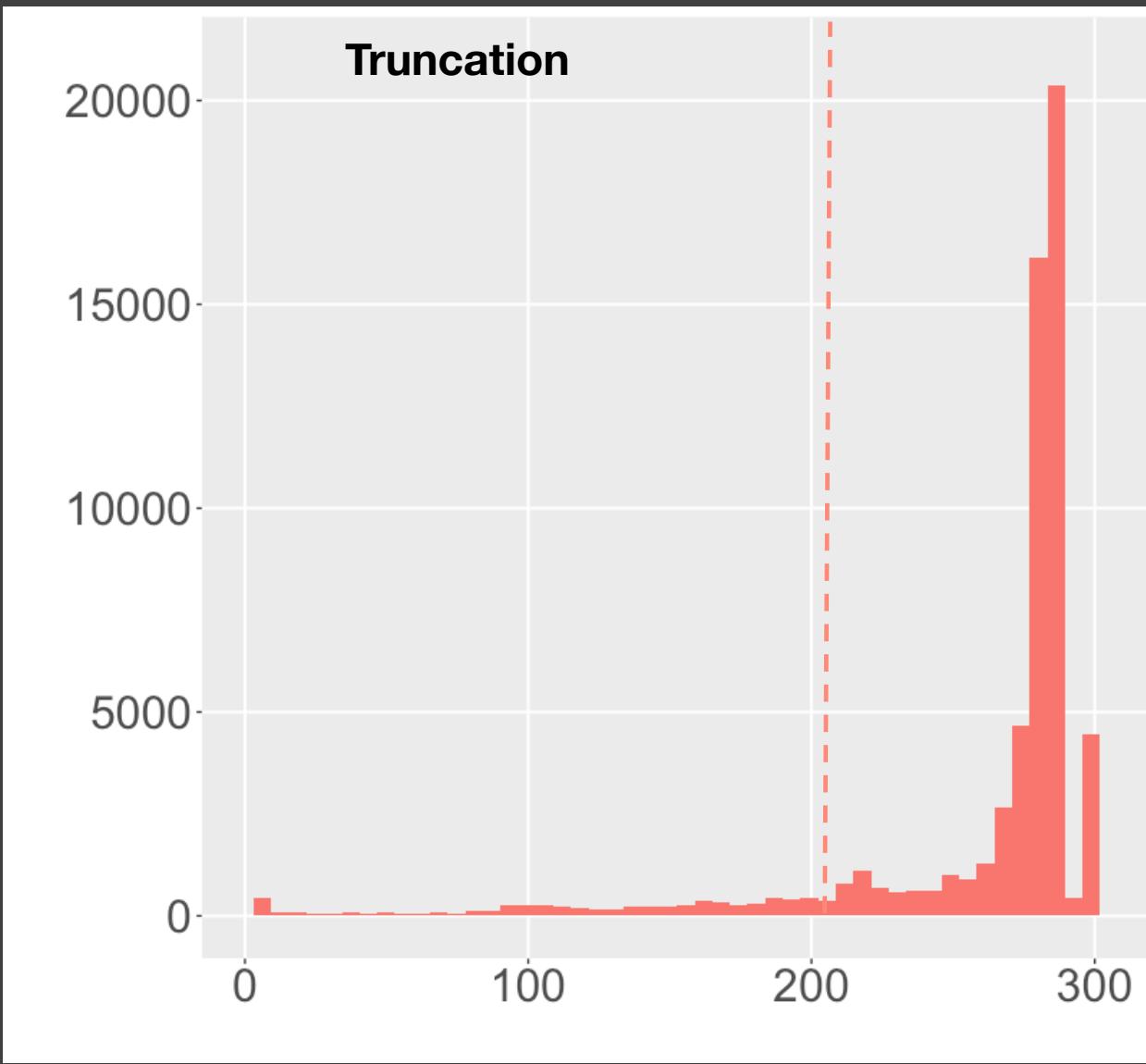
**Quality Trimming**



**Tools**

CutAdapt

# Trimming & Quality Filtering



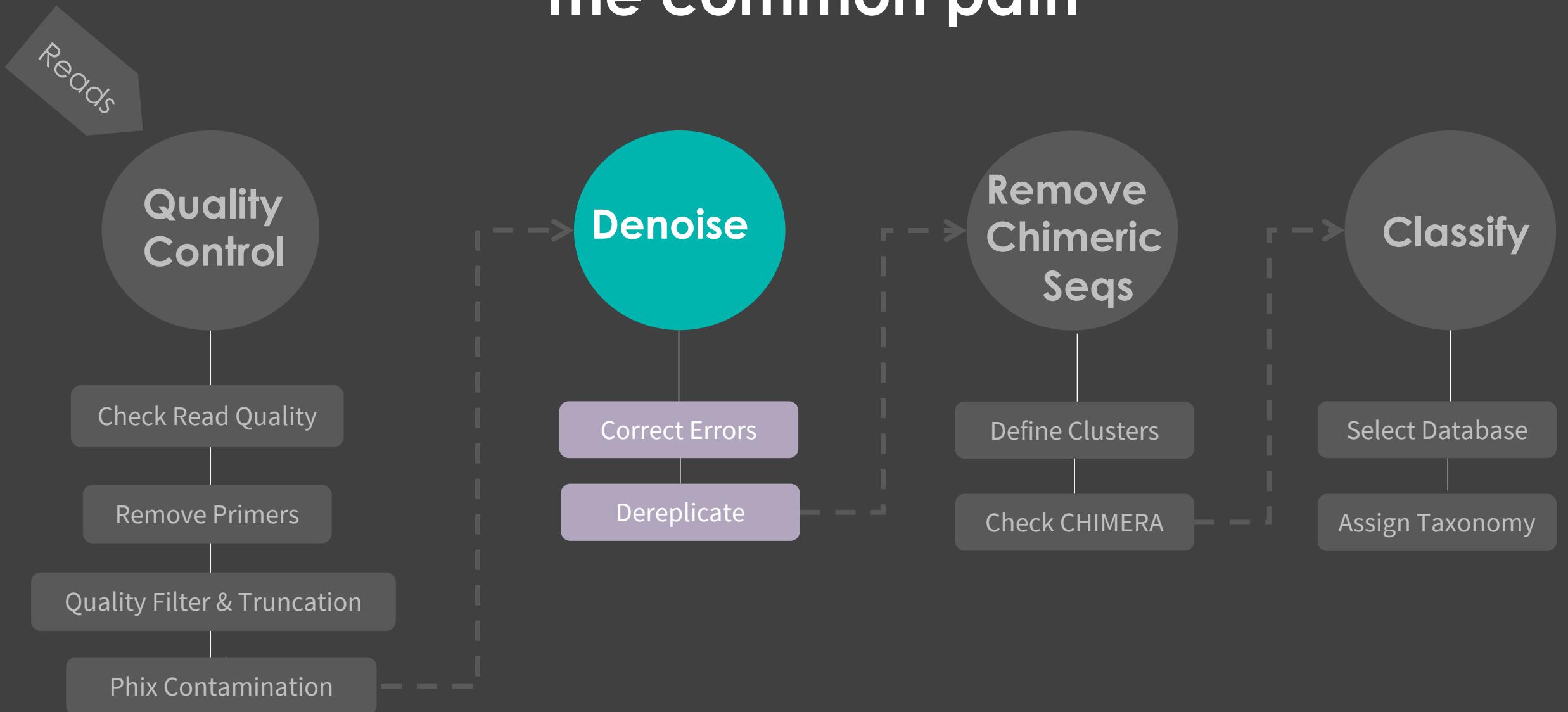
## Available Methods

- Minimum Q ( $Q \geq 20$ )
- Truncate if 3 consecutive bases are  $Q < 3$
- Expected Errors

Read length	% of reads
10	98.96
100	96.97
200	89.3
250	80.5

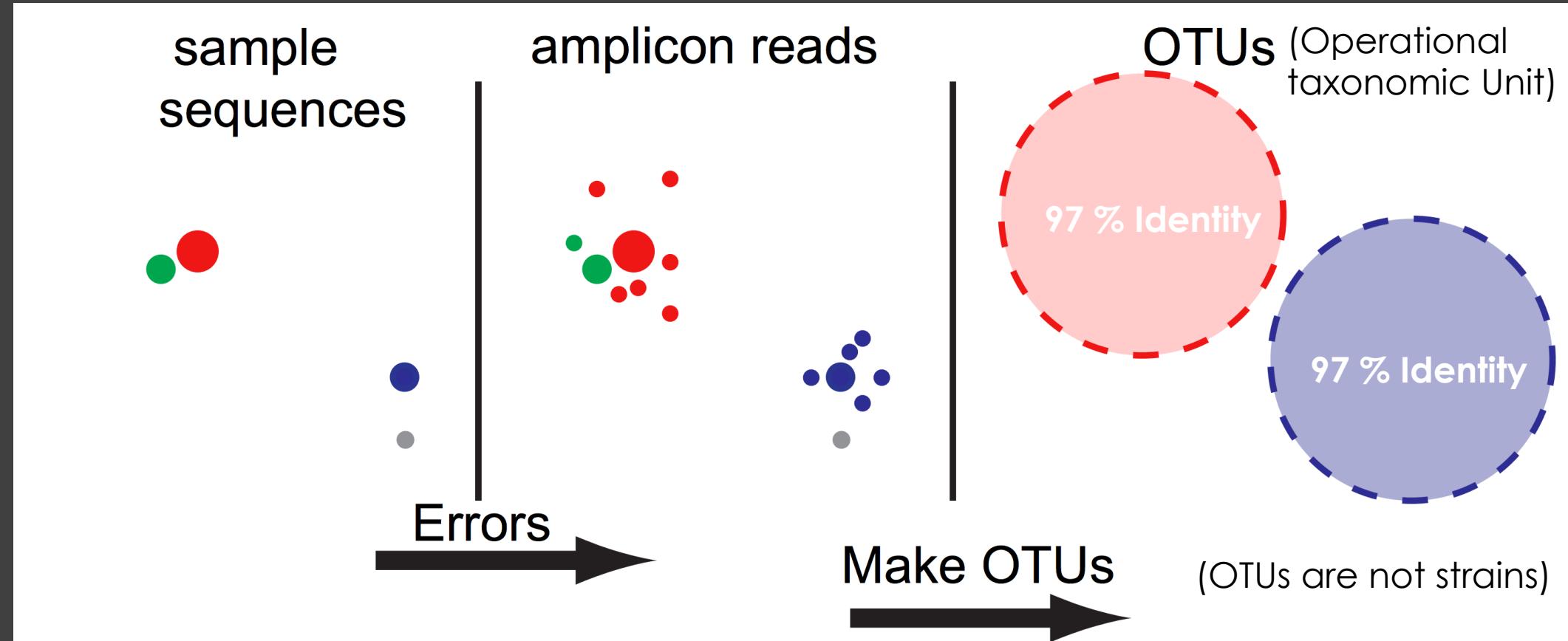
Source: Edgar et al., 2015

# The common path



# Denoising reads

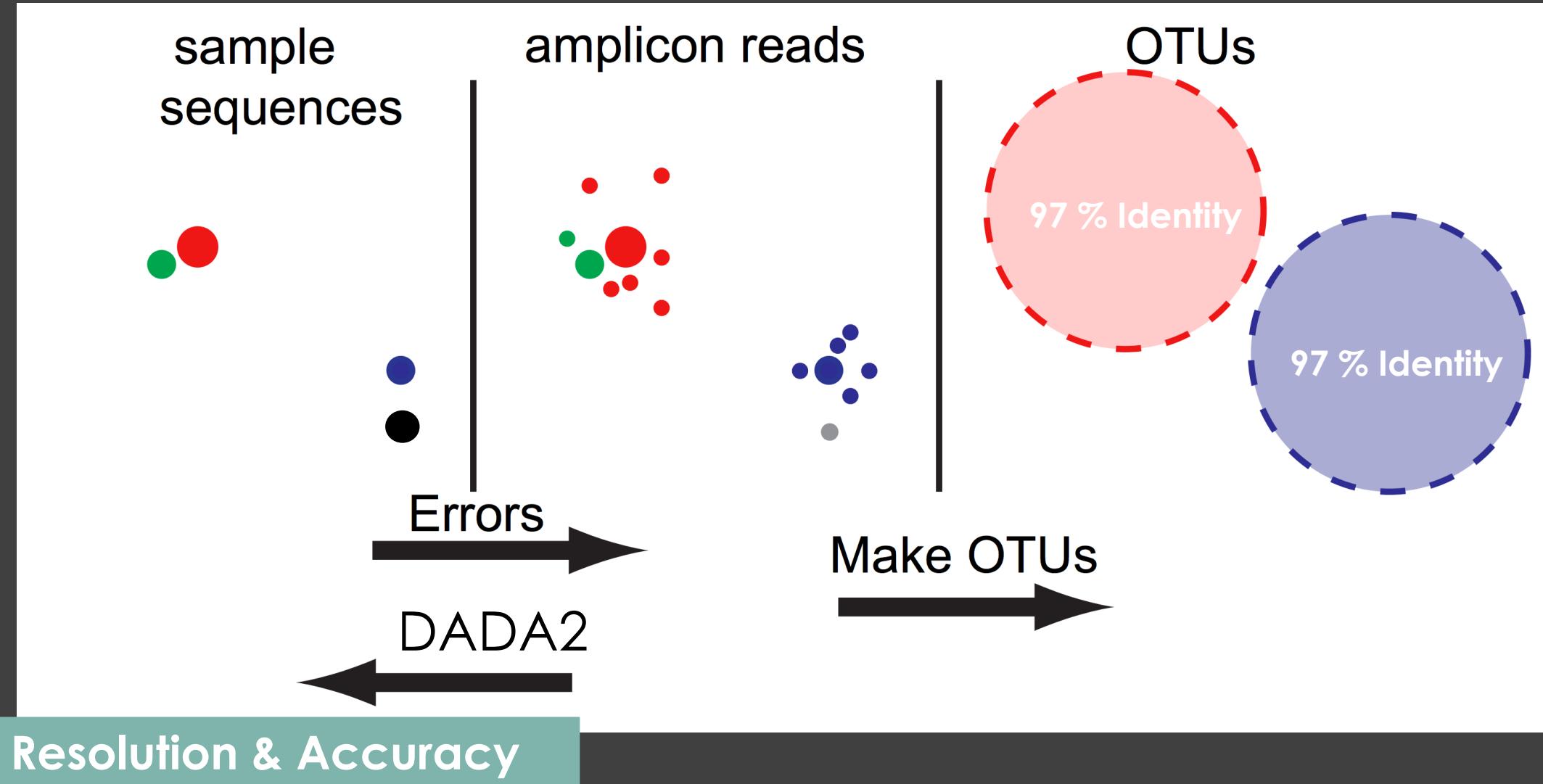
# Clustering



OTUs: Lump similar sequences together

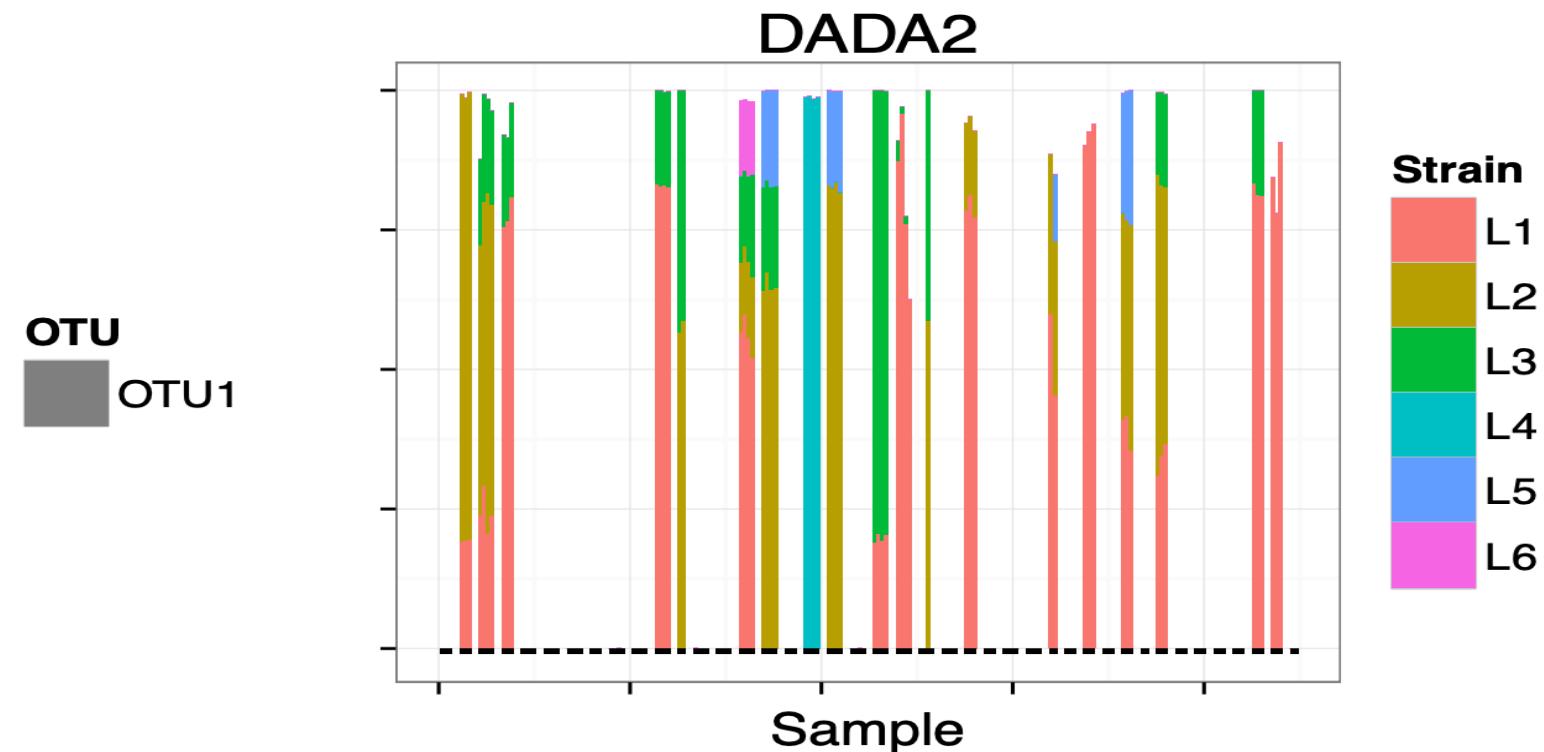
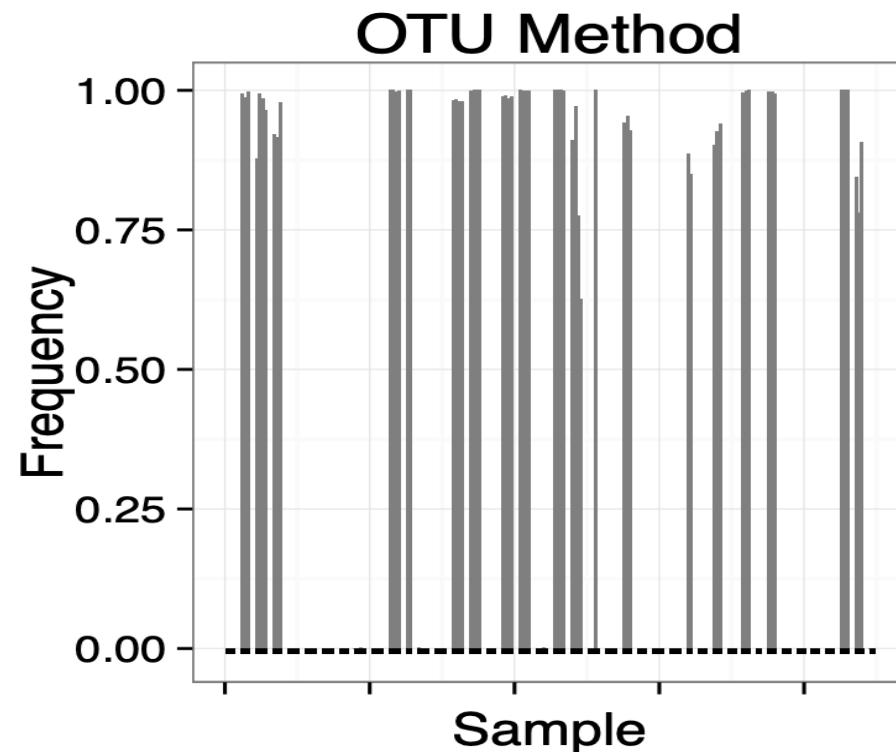
DADA2: Statistically infer the sample sequences (Amplicon sequence variants: ASVs)

# Clustering



# Real example, exact sequence resolution

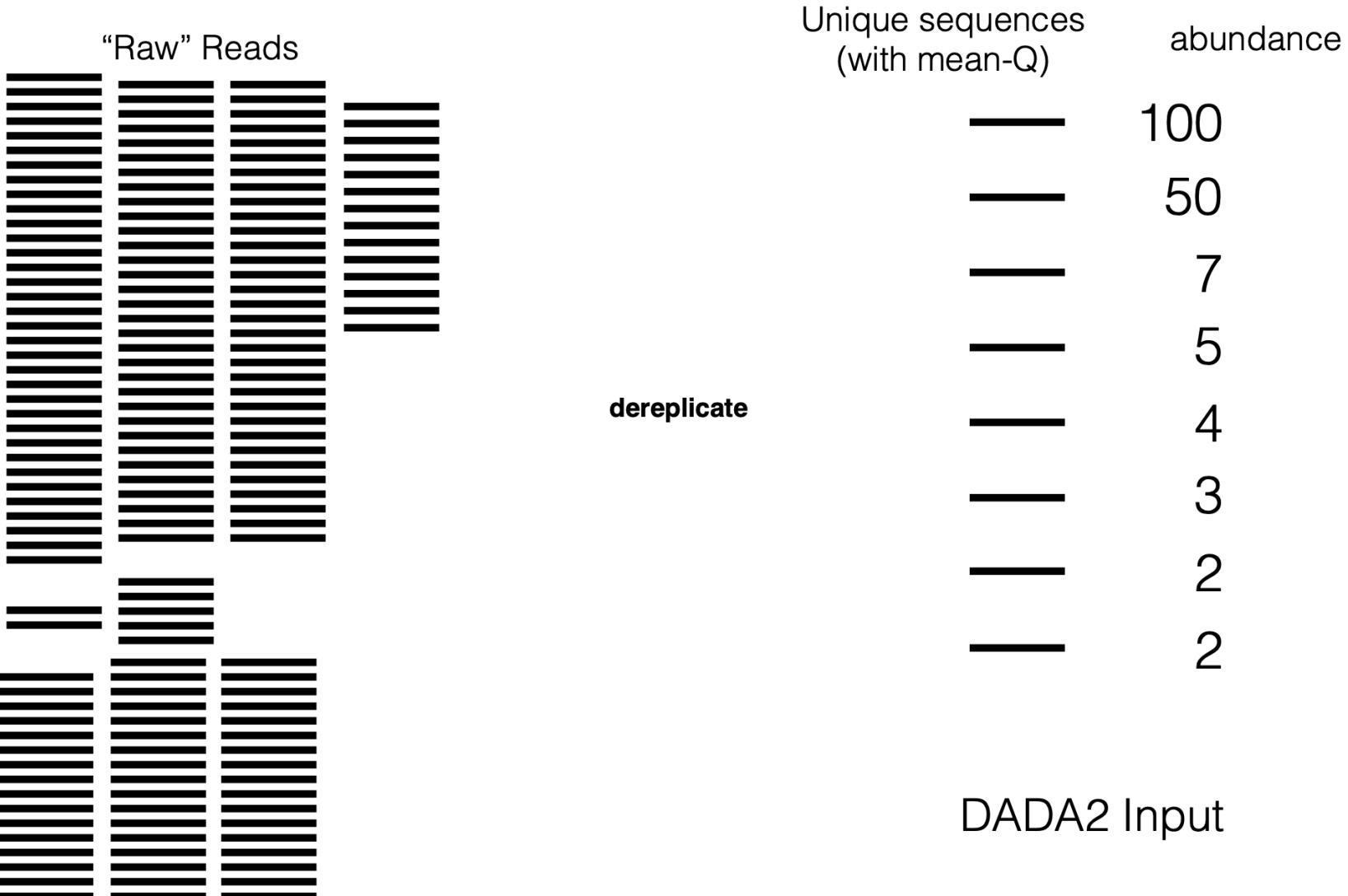
*Lactobacillus crispatus* sampled from vaginal microbiome 42 pregnant women



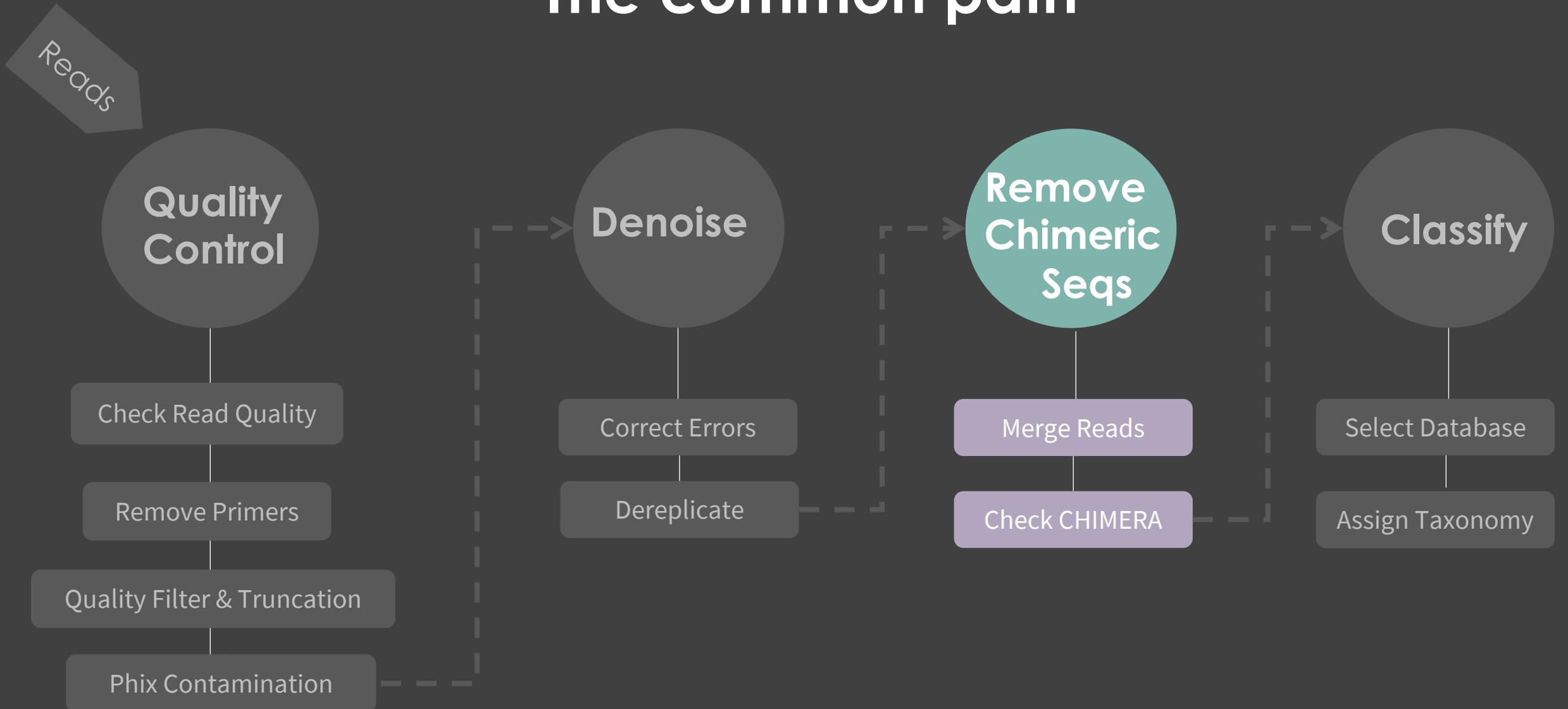
Data: MacIntyre et al. Scientific Reports, 2015.

# DADA2 algorithm cartoon

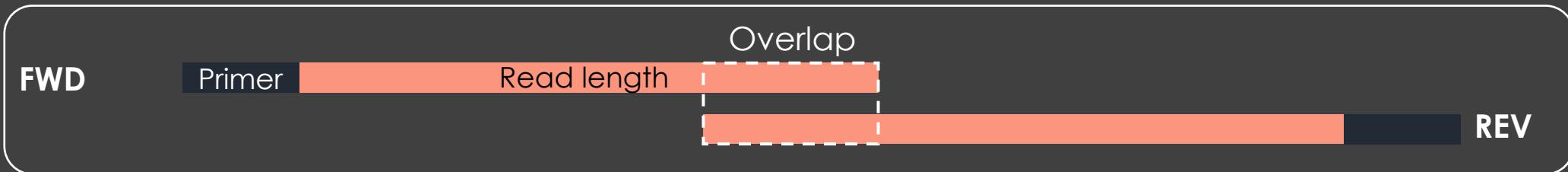
Input: unique sequences, their quality values, and abundances



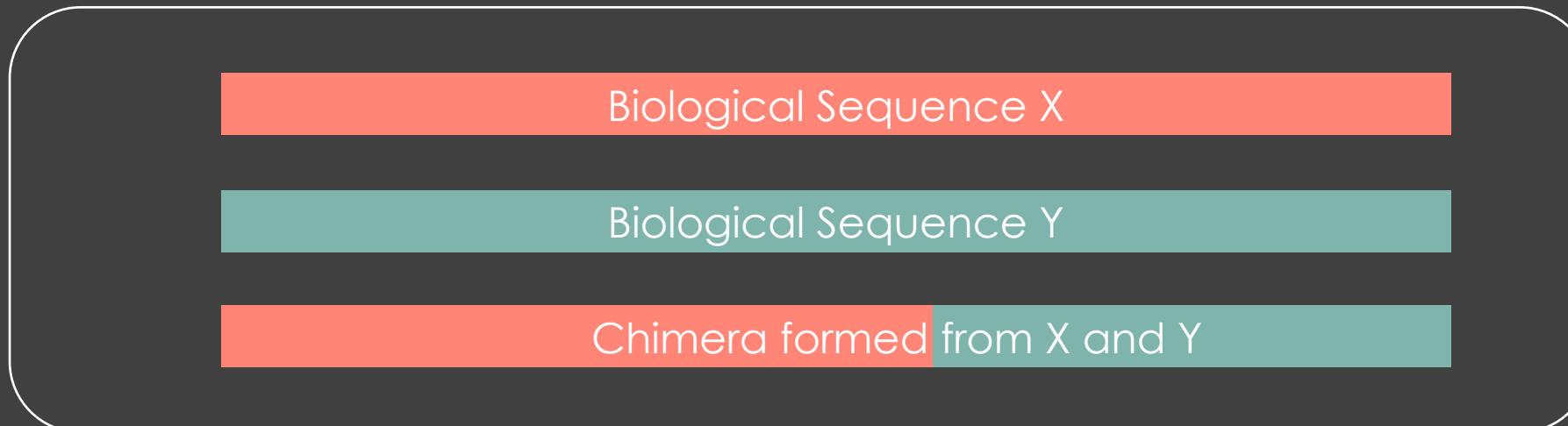
# The common path



# Merge Paired End Reads



## Chimeric Sequences



Created during PCR  
Fragment primes different extension

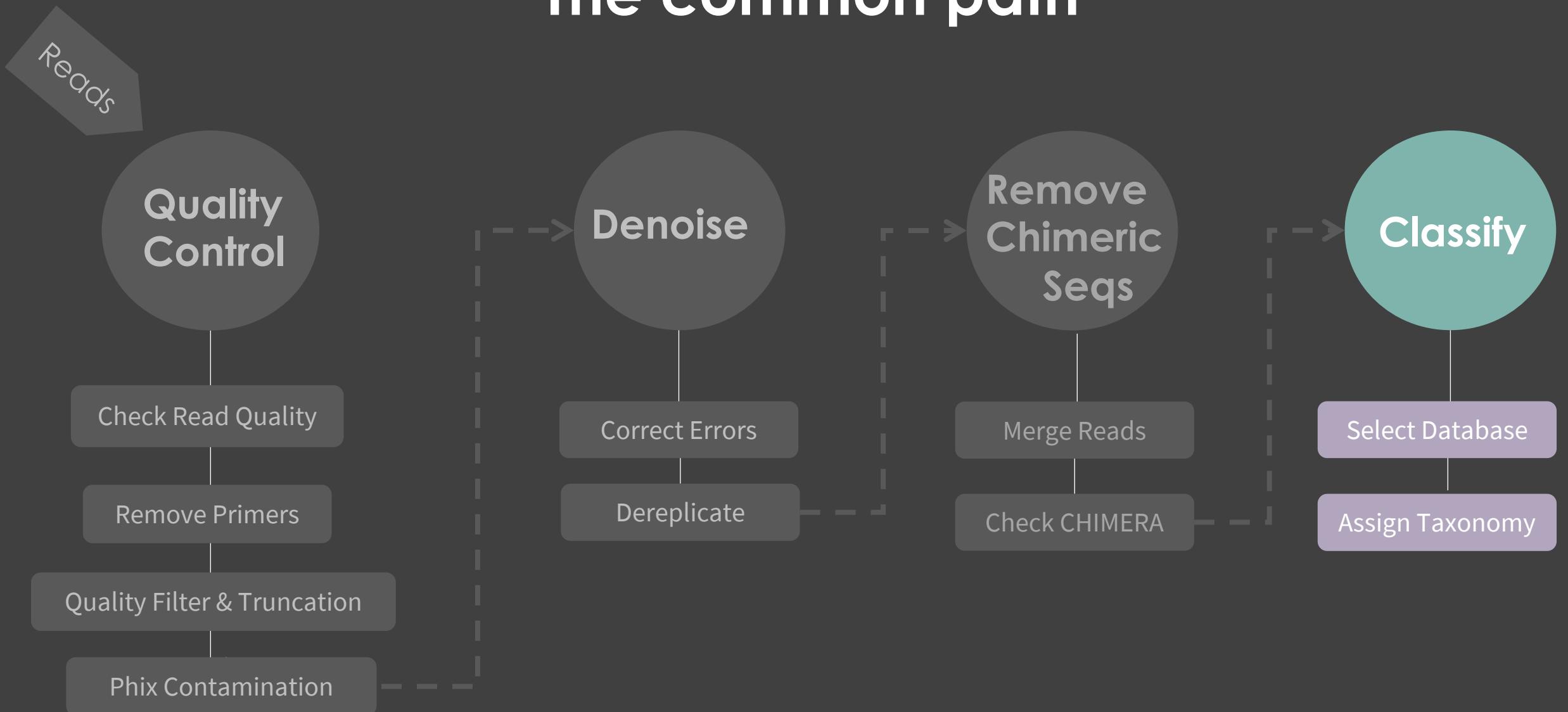
**About 1-5 % of reads are chimeric**

Source: Lahr & Katz, 2009

# Important

DADA2 gives us Amplicon sequence Variants  
(ASVs) not operational taxonomic units OTUs

# The common path



## Taxa levels

Kingdom  
Phylum  
Class  
Order  
Family  
Genus  
Species



## Reference Databases

Taxonomy	Type	No. of nodes	Lowest rank	Latest release
SILVA	Manual	12,117	Genus	Dec 2017
RDP-II	Semi	6,128	Genus	Sep 2016
Greengenes	Automatic	3,093	Species	May 2013
GTDB	Automatic	143,512	Species	Today <sup>a</sup>

Source: Balvočiūtė et al 2017

Use frequently updated and well curated databases!

# Diversity

## Alpha diversity

Within sample diversity  
Who is in the sample

Uses Abundance and/or observed number of each ASV

Shannon, Simpson, Chao1

Richness , Evenness

## Beta diversity

Between sample diversity  
How similar at the samples

## Distance metric & clustering

Jaccard	Unifrac
Bray-Curtis	Weighted-Unifrac

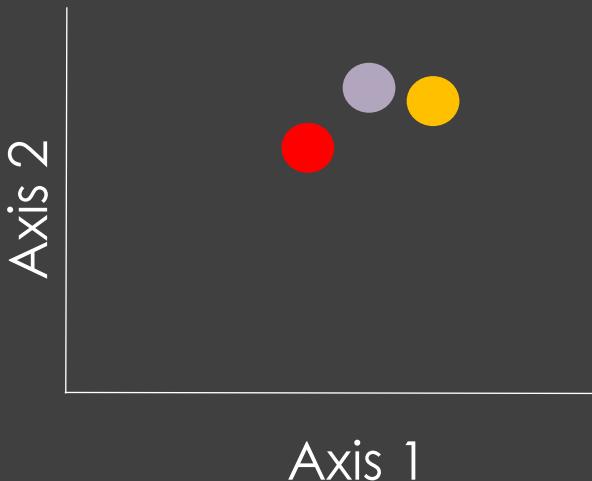
Absent/Present

Abundance

# Beta Diversity

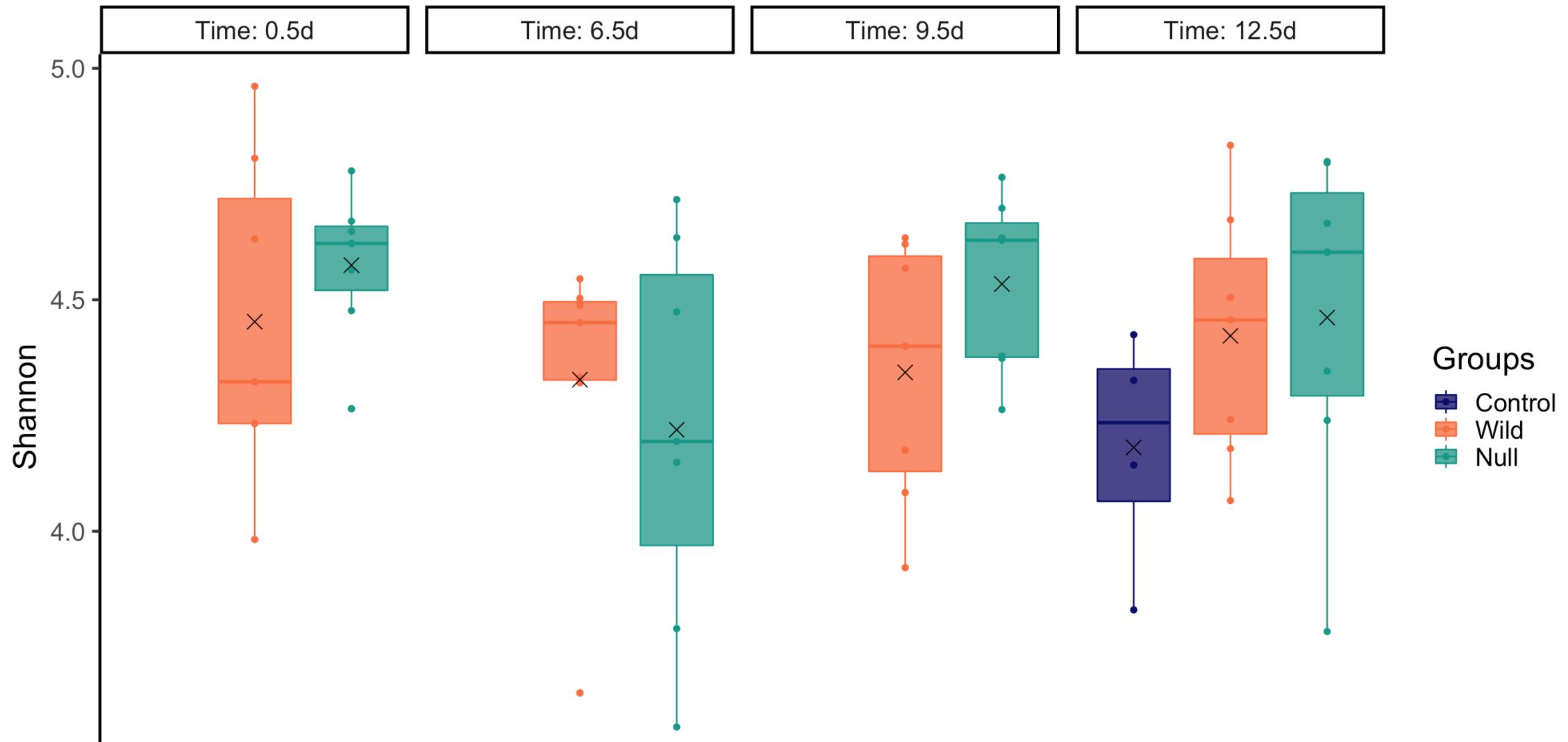
	Sample 1	Sample 2	Sample 3
Sample 1		0.2	0.6
Sample 2	0.2		0.5
Sample 3	0.6	0.5	

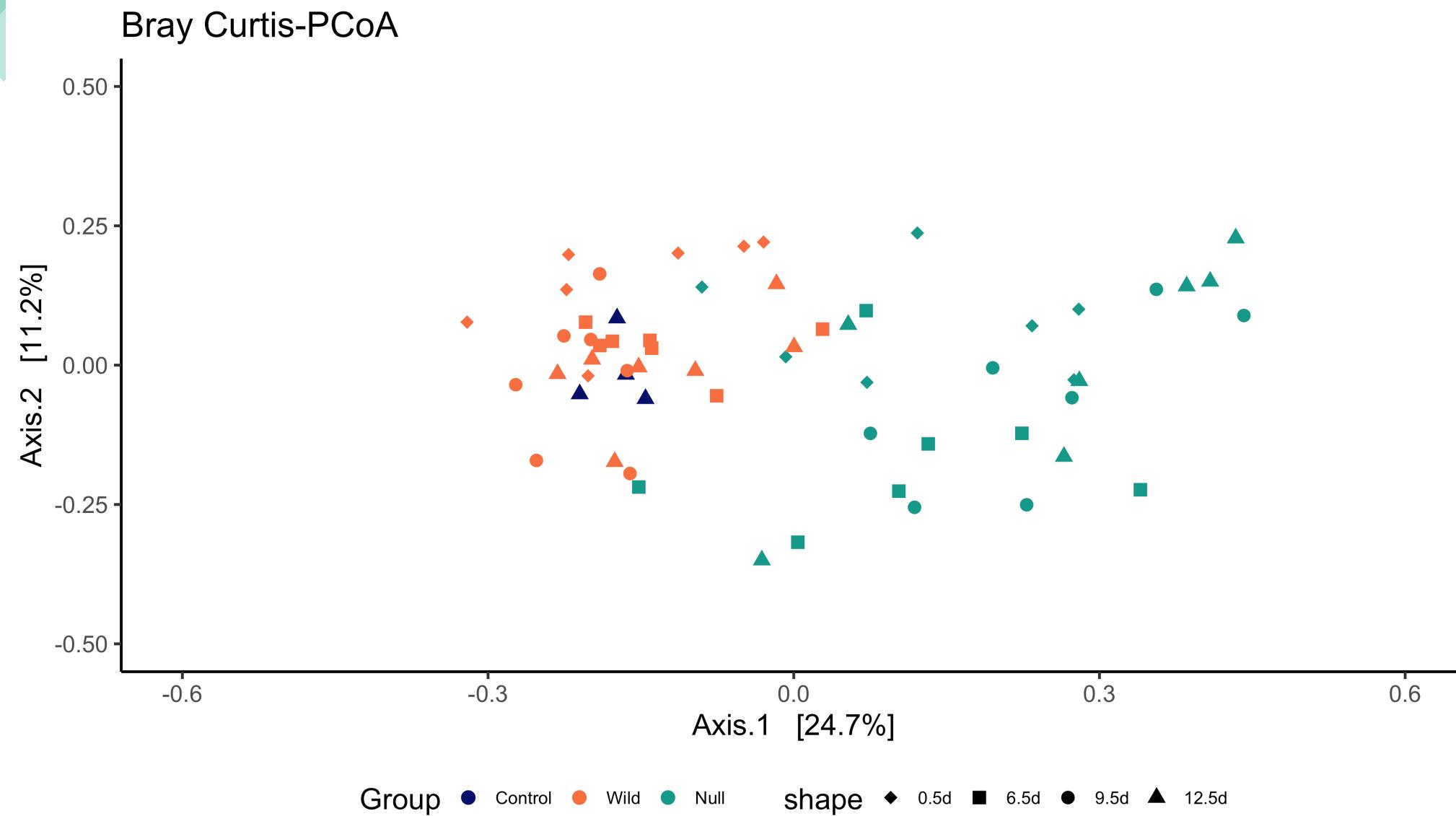
Each Axis represents a percentage of variability it explains of the data.



**Methods**  
PCoA  
NMDS  
CCA

# Shannon Index





# What next?

Differential abundance analysis

Network analysis

Predictive functional analysis

"If you **torture** the data long enough, it will confess."

- Ronald Coase, *Economist*

# Tips

## Sample size

Are there enough samples for down stream analysis

## Controls + & -

Helps identify contaminants

## Length & Depth

Know which region, read length & depth suits your experiment

## Be consistent

It is much better to use the same methods than to change methods frequently

**Think before you start and ask why!**

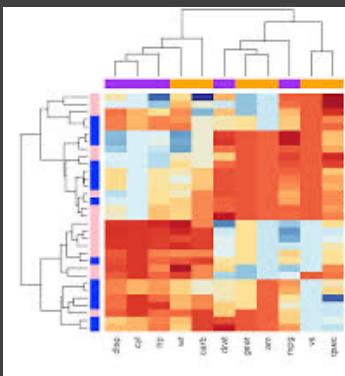
# Shotgun Metagenomics



Total DNA →



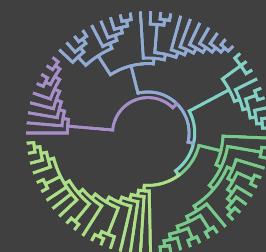
Functional Annotation  
+ Metabolite prediction



Assembly



Phylogeny



Taxonomic profile

Marker Gene  
Identification or  
Kmer exact match  
classification



Samples

## Taxa levels

Kingdom

Phylum

Class

Order

Family

Genus

Species

Strain



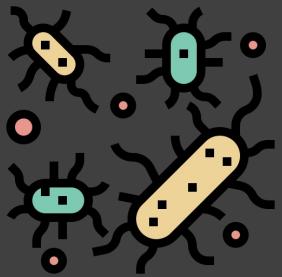
## Limitations

- Avoids PCR bias but has its own
- Computationally intense
- Did we get enough data to identify the whole community ?

# Metabolomics

Identification and quantification of metabolites

Who



Microbes

How



Metabolites



Host

Sugar

Amino Acids

Nucleotide

SCFA

# Metabolomics

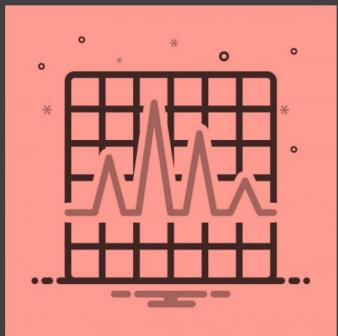
## Untargeted

Global profiling  
Qualitative

## Targeted

Measure a specific known set  
of metabolites  
Quantitative

### Separate



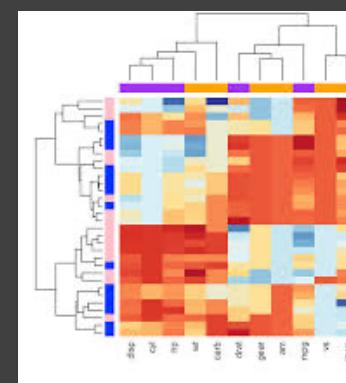
Chromatography  
By size and charge

### Detect



NMR or Mass spectrometry

### Analyze



# Detection options

## LC/MS

Separation is required

Can resolve metabolites in complex mixtures

High Sensitivity and range

Expensive and serious batch effects

## NMR

No separation required

Smaller number of biomarkers is detected

Lower sensitivity

Reproducible and free of batch effects

# Typical Analysis

**Data Acquisition:**  
Spectra,  
NMR

**Data Processing:** QC,  
normalization and  
statistical analysis

**Data Investigation:** Enrichment analysis  
or pathway analysis

**Data Integration:**  
Putting together  
with other omics  
data

# Multi-Omics Approach

Putting it all together

