

16s Microbiome Analysis workshop

International Microbiome Centre

June 15-16, 2020

Disclaimer: These slides are taken from difference sources on the web and some are modified. I have added sources but if I have missed some, I apologize in advance to the original content creators.

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



THE HUMAN MICROBIOME PROJECT

CONTAINS

10 TIMES

MORE

THAN

MICROBIAL CELLS

HUMAN CELLS.



LYMPH

BLOOD

MICROORGANISMS,
THEIR GENOMES
& ENVIRONMENTAL
INTERACTIONS

250



HEALTHY
PEOPLE

SPOTS SAMPLED:

VAGINA

G.T.

ABDOMEN

AIRWAYS

ORAL

EYE

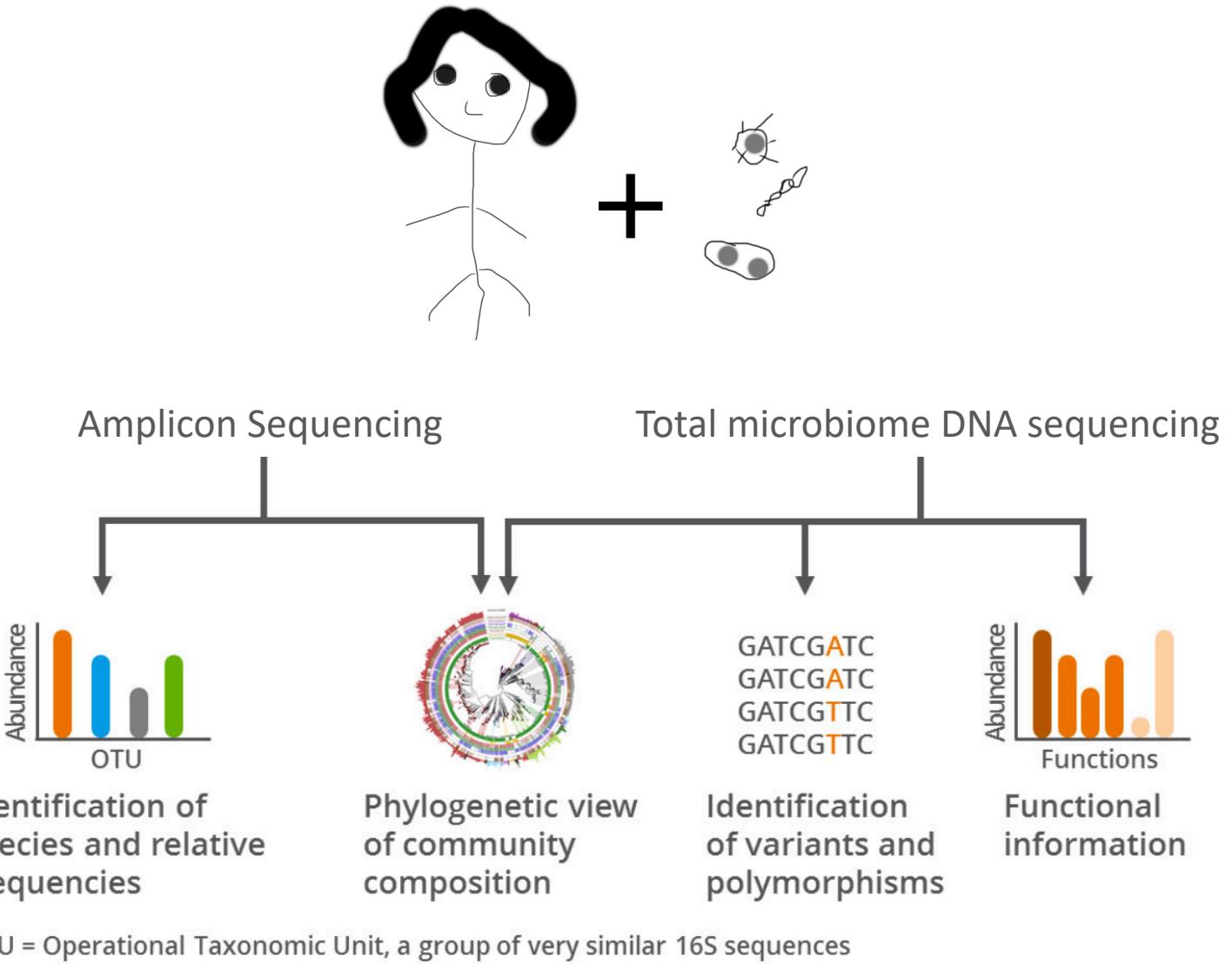
PRI

HAND

HEART

Who lives here?

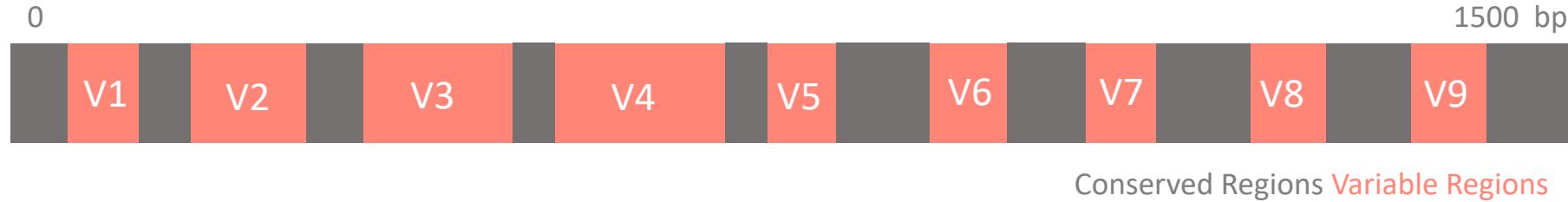
Over time + Treatment



Amplicon Sequencing

What should we amplify?

The Small Subunit '16s' ribosomal RNA gene

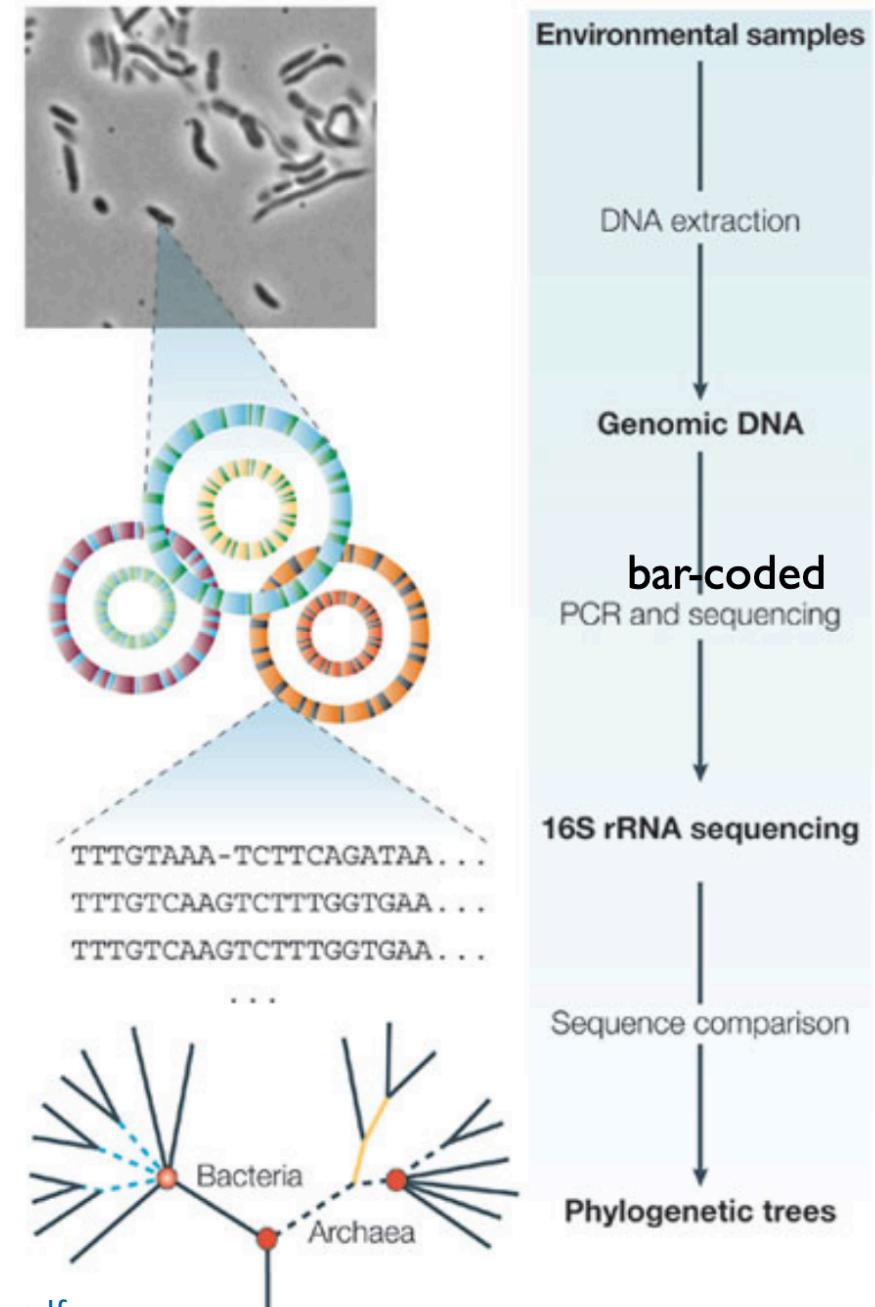


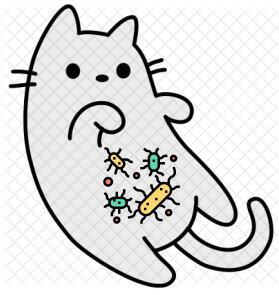
Because this gene is conserved and found in most bacteria, it is used for reconstructing phylogeny.

Depending on the hypervariable region picked, you can classify to different level of taxonomy

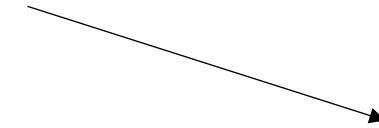
Many microbiomes in parallel

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





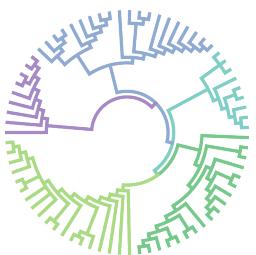
Amplify 16s region



ASV Table

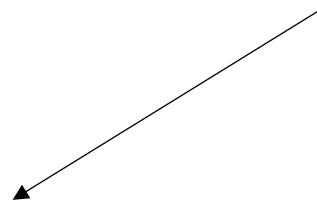
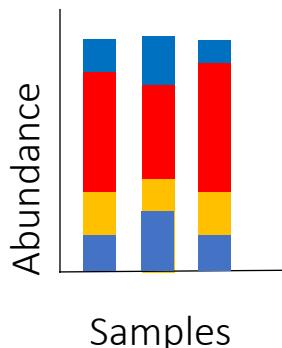
	Sample 1	Sample 2	Sample 3
ASV 1	0.5	0.4	0.6
ASV 2	0.2	0.3	0.1
ASV 3	0.2	0.1	0.2
ASV 4	0.1	0.2	0.1

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

Bioinformatics to the rescue !!!

1



Sequencing results

2



Woah! That is a lot of data

3



Bioinformatics Tools

4



Biological Knowledge

Source: LEGO+Johnathan Irish

**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?

Negative Controls

Was a negative control included?

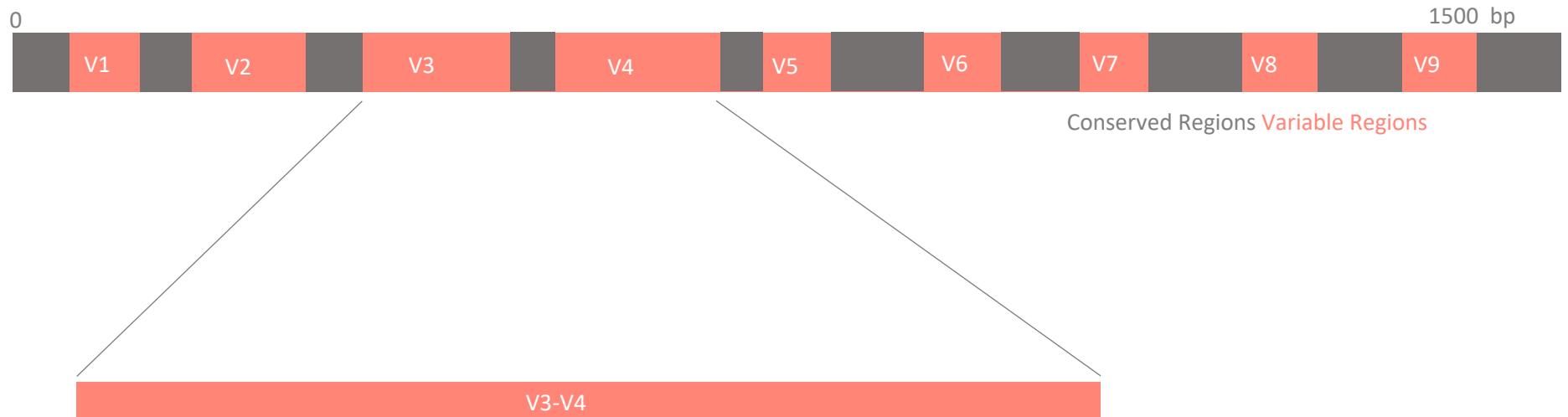
Sample size

Are there enough samples for downstream analysis

Feasibility

Will this data answer the questions asked by the investigator?

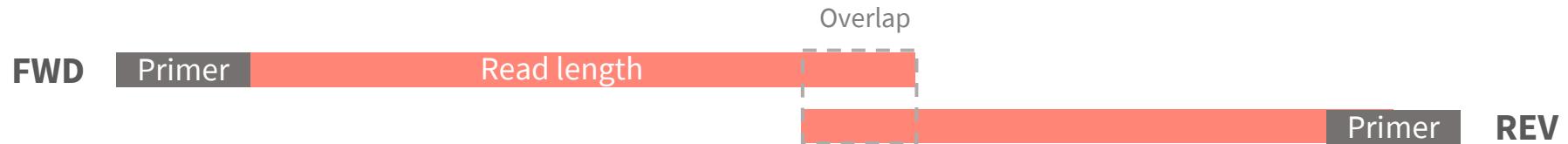
16 rRNA



Pick a sequencing technique

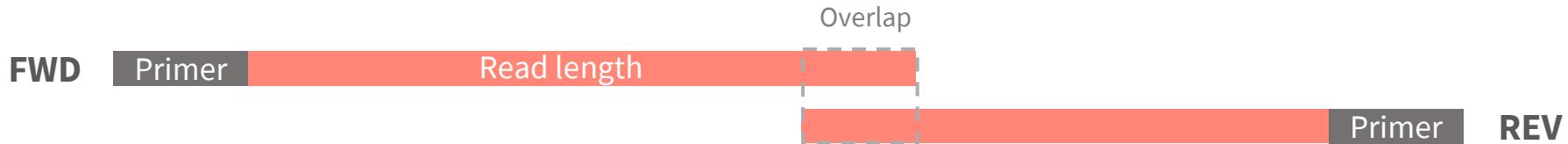
Illumina (MiSeq)

Select primers & Read lengths



Make sure there is an overlap!!!!

How to calculate V region overlap



$$\text{Overlap} = \text{Read length} * 2 - \text{primer_F} - \text{primer_R} - \text{VR_length}$$

Example: V4 VR_length=254 bp, read length =250 x2 bp and primers=20 bp

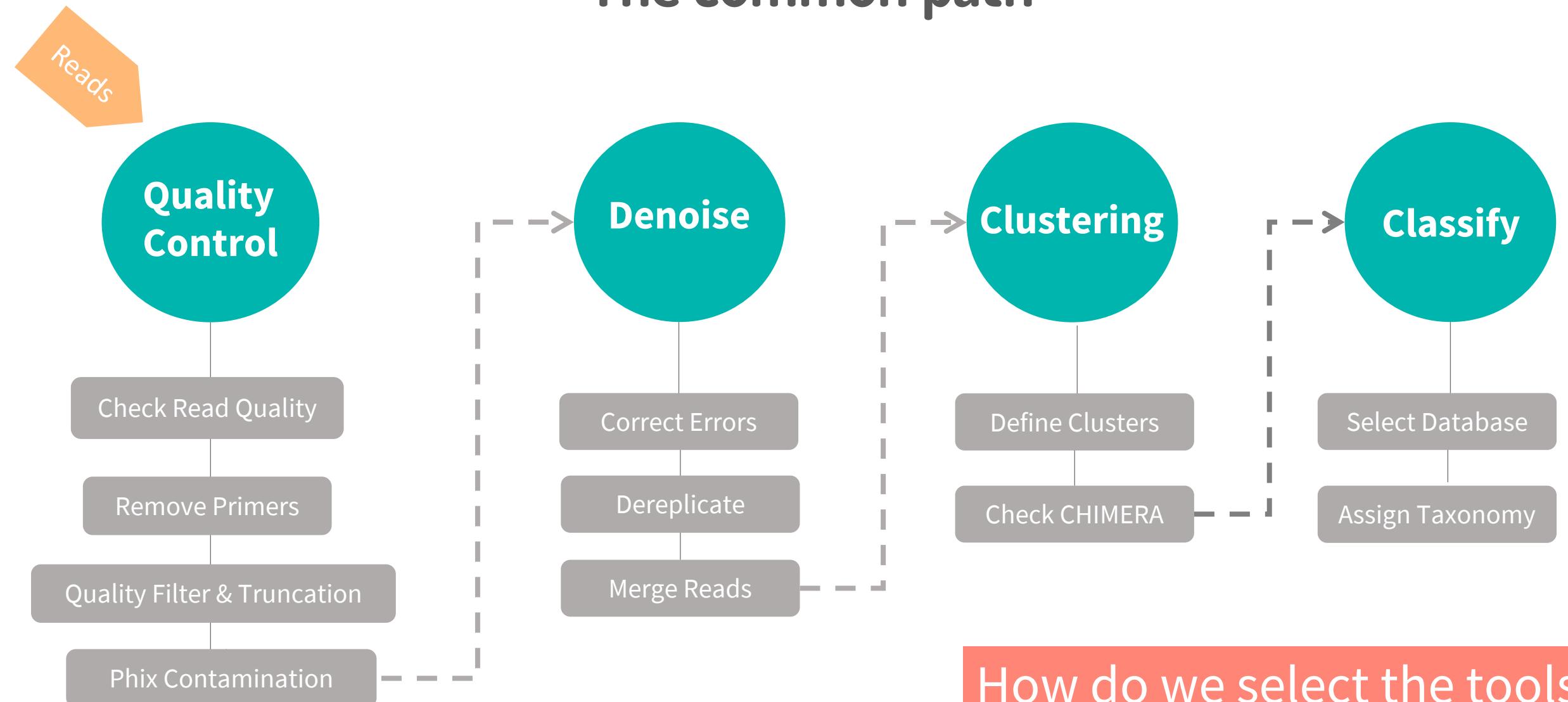
$$\begin{aligned}\text{Overlap} &= 250 * 2 - 20 - 20 - 254 \\ &= 206\end{aligned}$$

Always have a blank sample sequenced with your samples



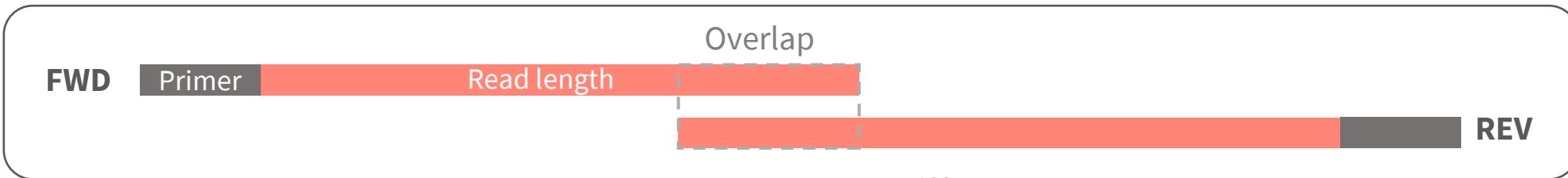
Typical Workflow

The common path

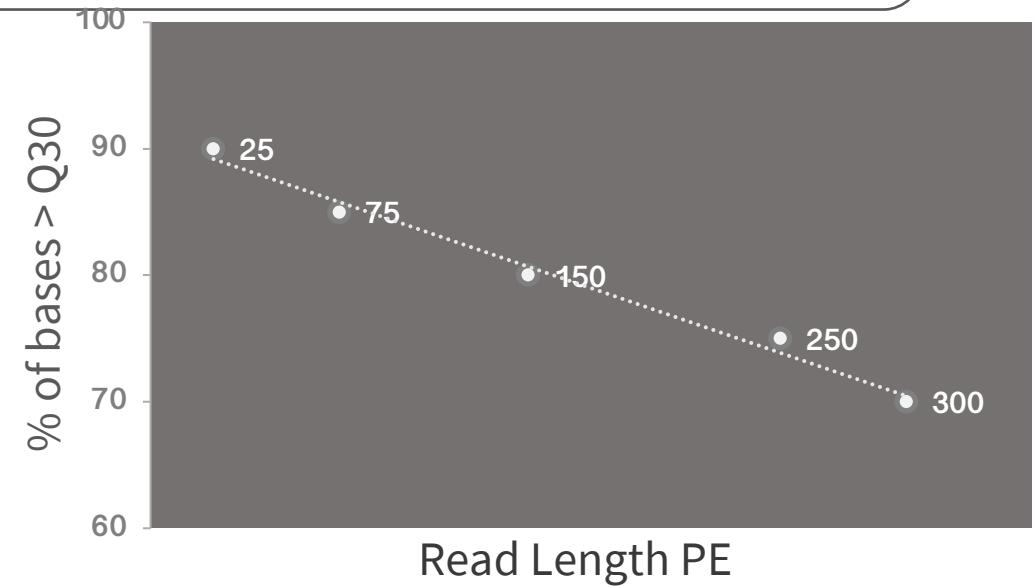


How do we select the tools
and use them well ?

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

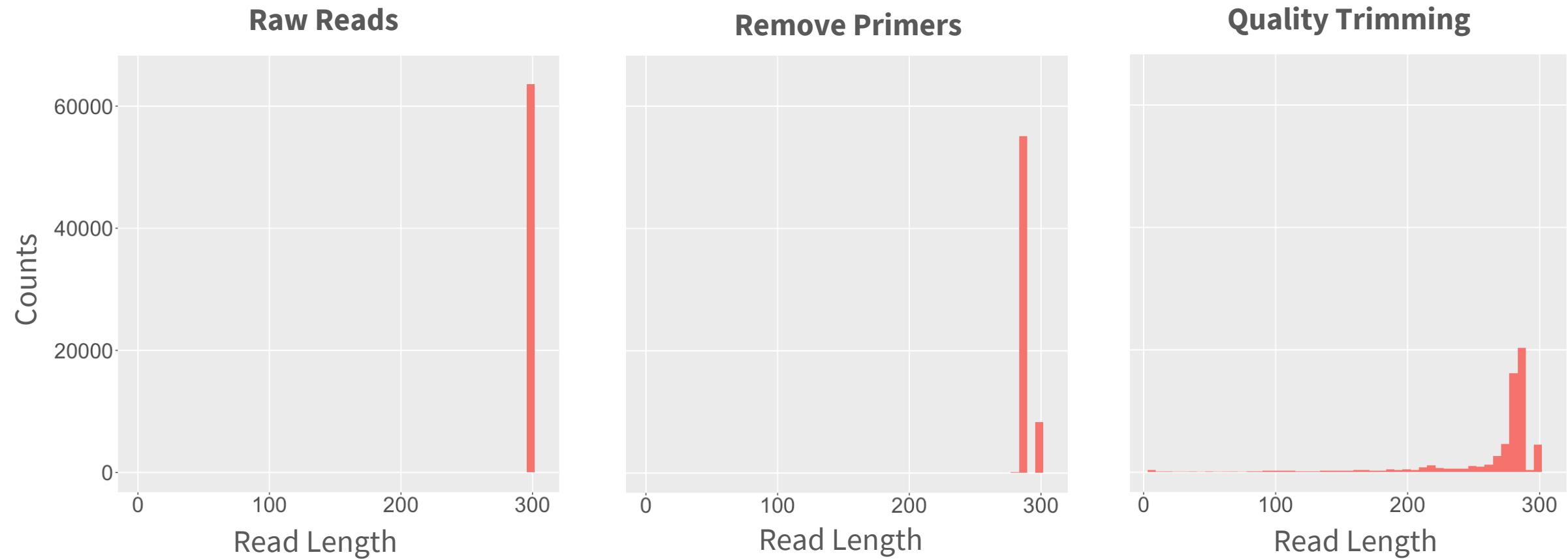


Tools

FastQC FastQp MultiQC & Specialized 16s rRNA packages

Source: Dada2 R package

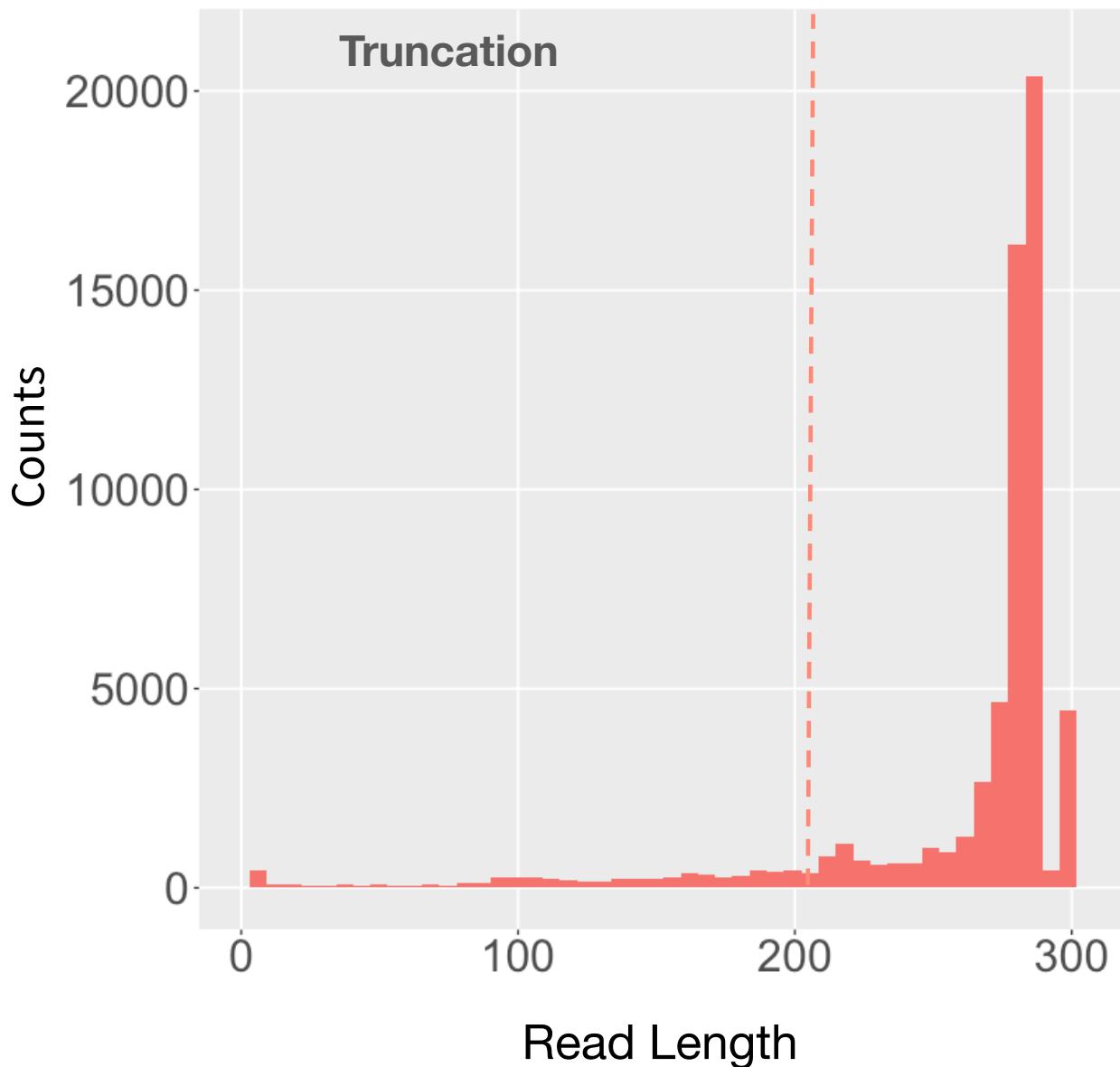
Trimming & Quality Filtering



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

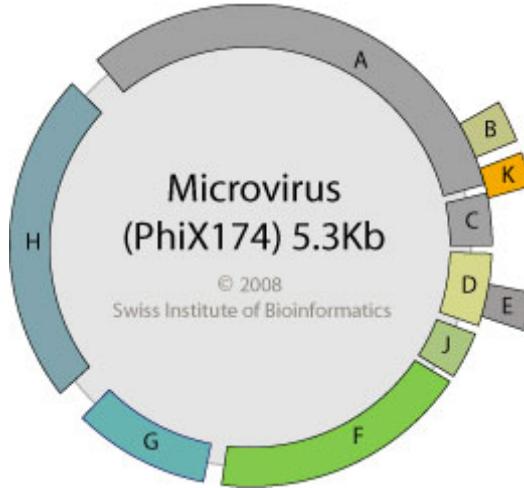
Why PhiX?

Low diversity

Significant number of reads have the same sequence

Quality Control

Cluster Generation
Phasing & Prephasing



Source: [Illumina](#)

Caution

PhiX Contamination

Illumina removes PhiX
Assigned to samples
(1%-12%)

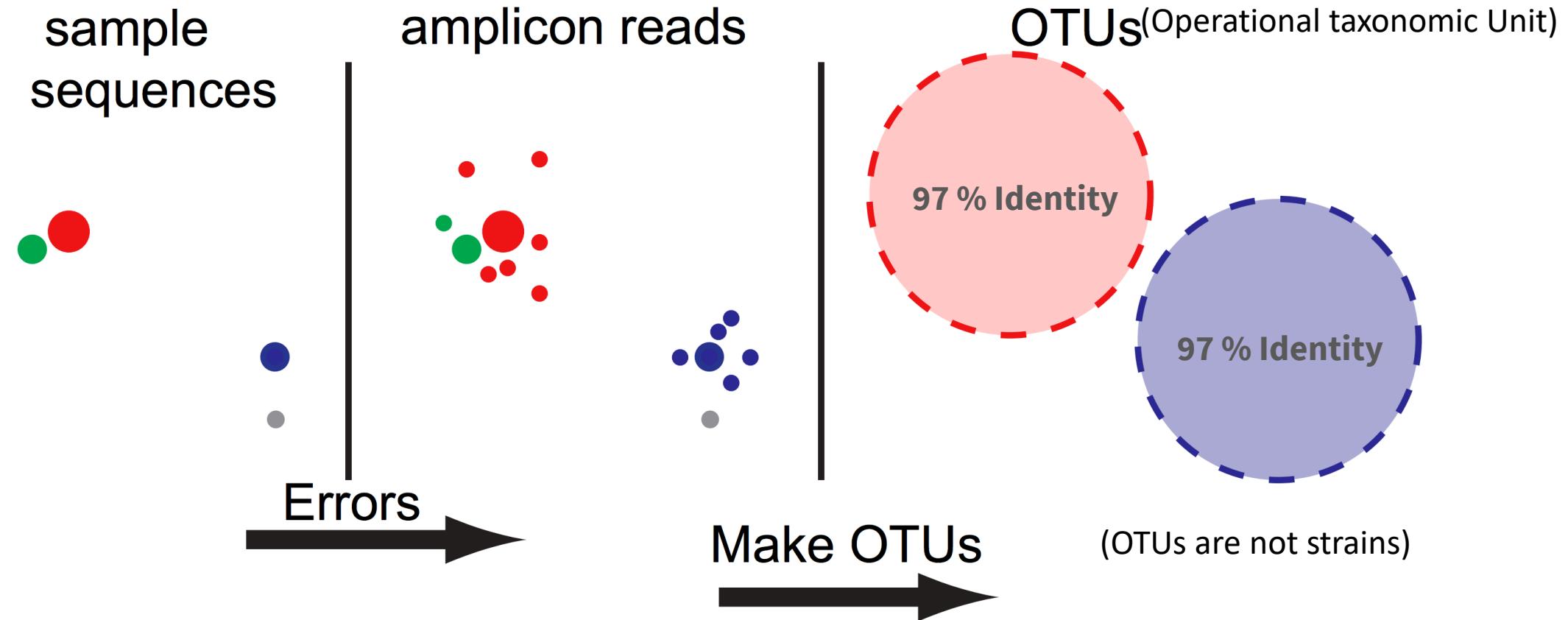
Removal

Blast sample reads against
PhiX genome

Source: [Mukherjee et al., 2015](#)

Denoising reads

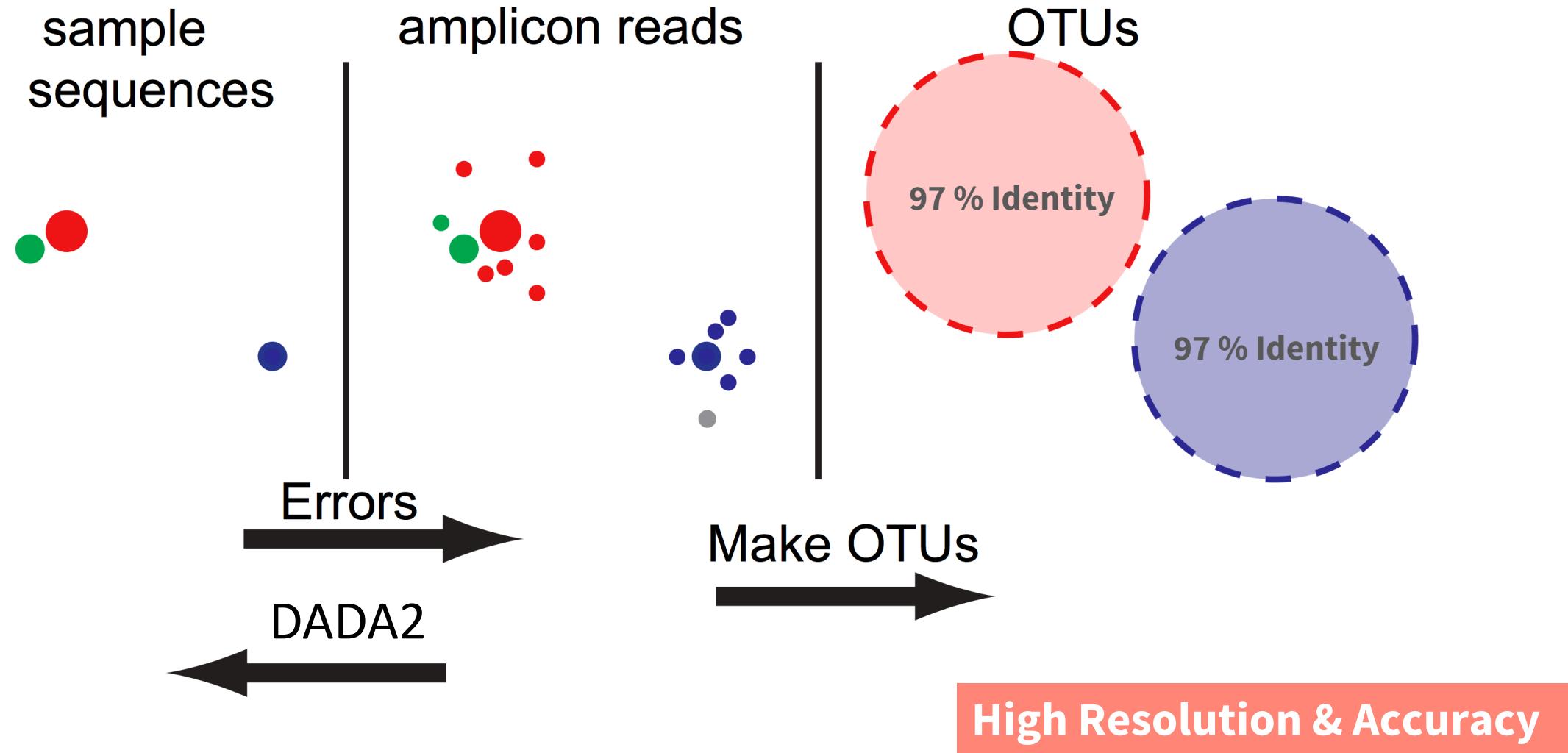
Clustering



OTUs: Lump similar sequences together

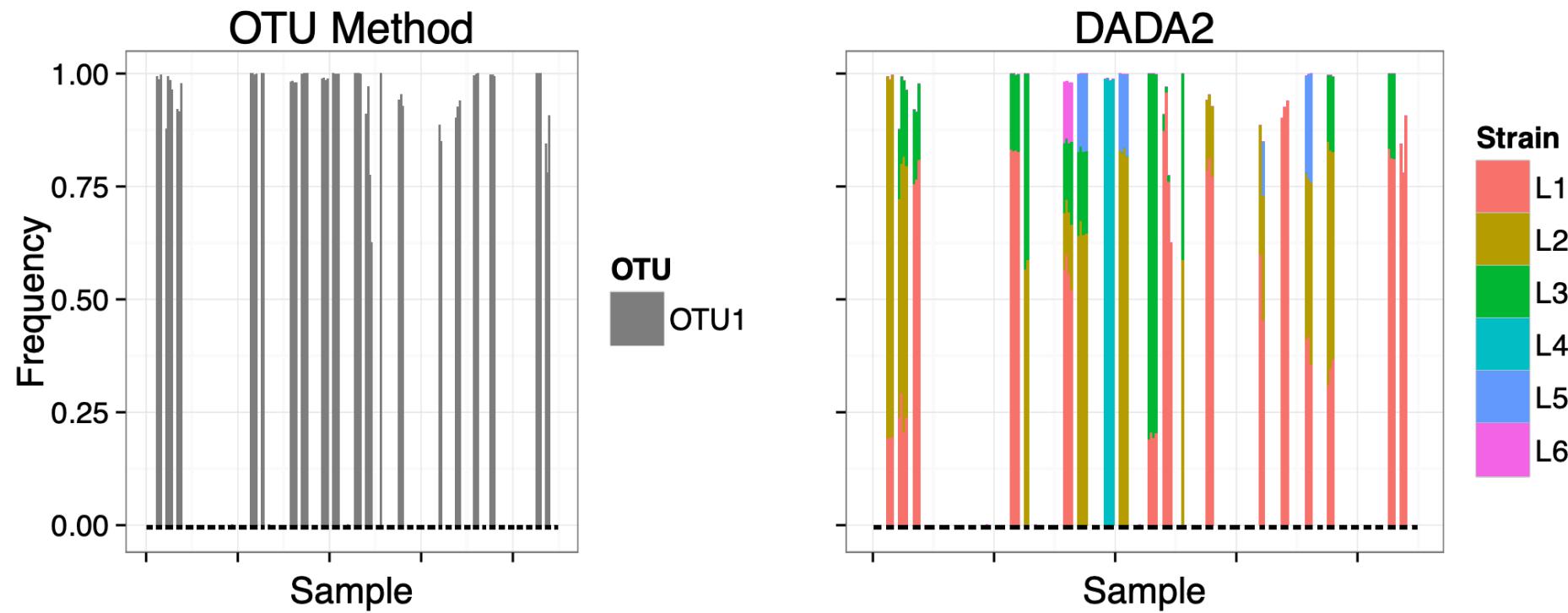
DADA2: Statistically infer the sample sequences (Amplicon sequence variants: ASVs)

Clustering



Real example, exact sequence resolution

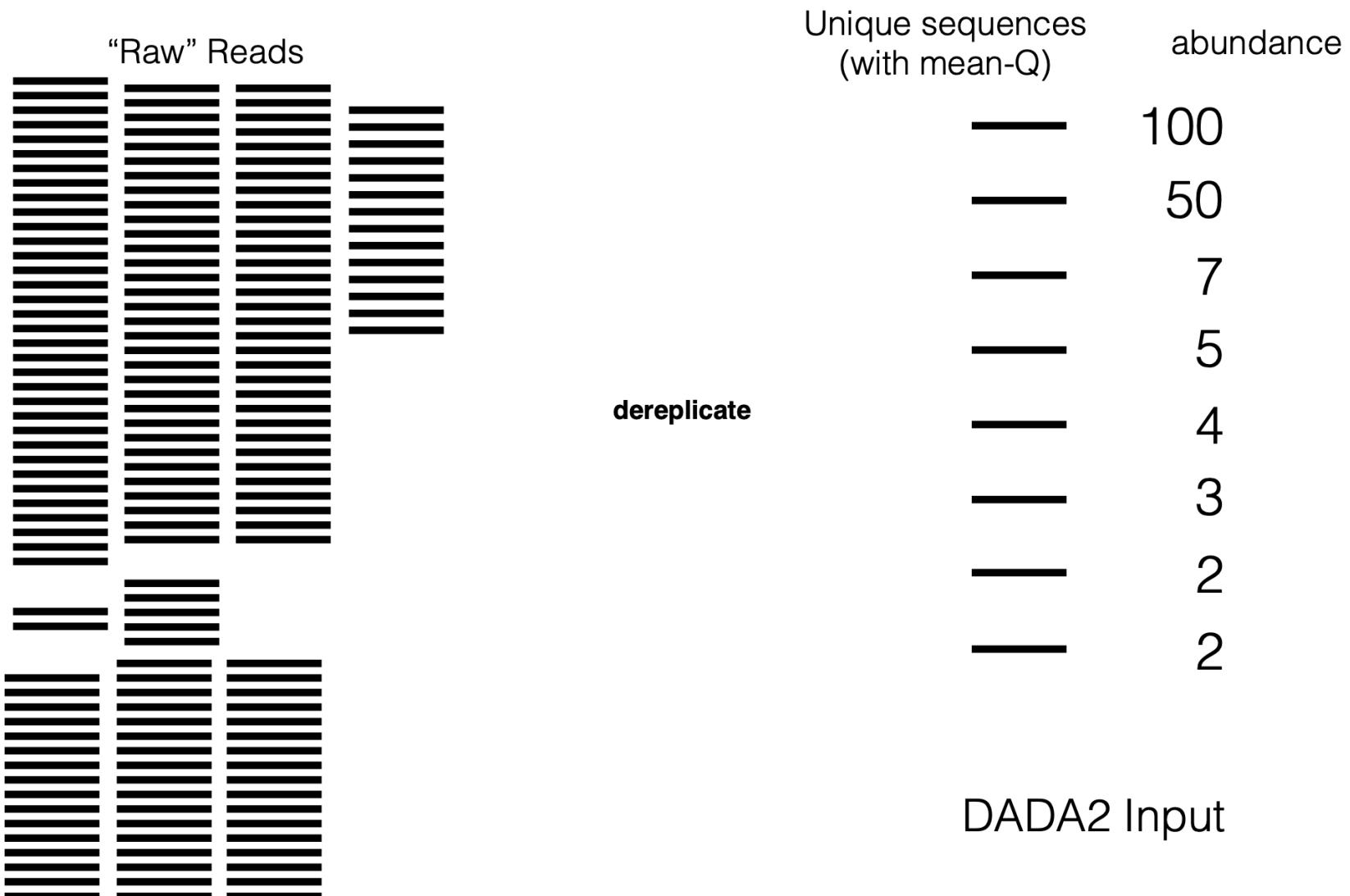
Lactobacillus crispatus sampled from vaginal microbiome 42 pregnant women



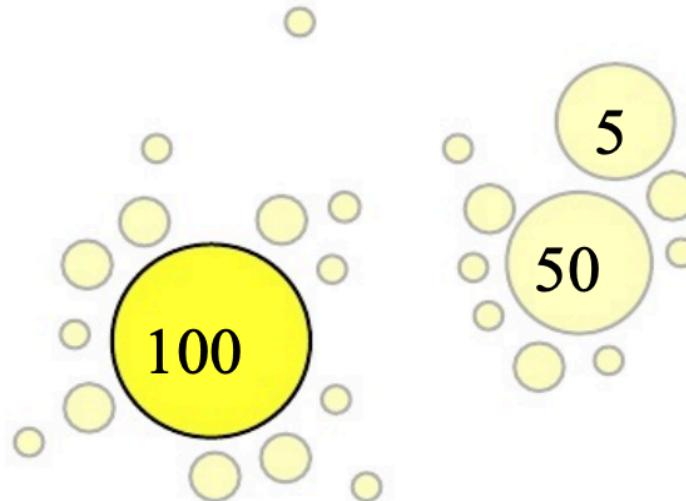
Data: MacIntyre et al. Scientific Reports, 2015.

DADA2 algorithm cartoon

Input: unique sequences, their quality values, and abundances



DADA2 algorithm cartoon



Infer initial *error model* under this assumption.

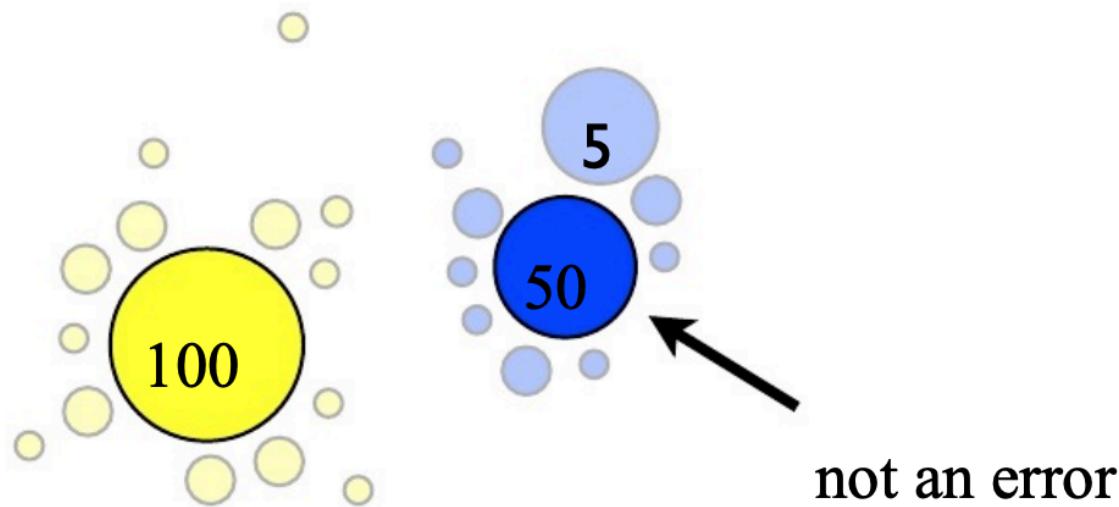
$$\Pr(i \rightarrow j) =$$

	A	C	G	T
A	0.97	10^{-2}	10^{-2}	10^{-2}
C	10^{-2}	0.97	10^{-2}	10^{-2}
G	10^{-2}	10^{-2}	0.97	10^{-2}
T	10^{-2}	10^{-2}	10^{-2}	0.97

Two models:

- Error Model
- Abundance model (p-value)

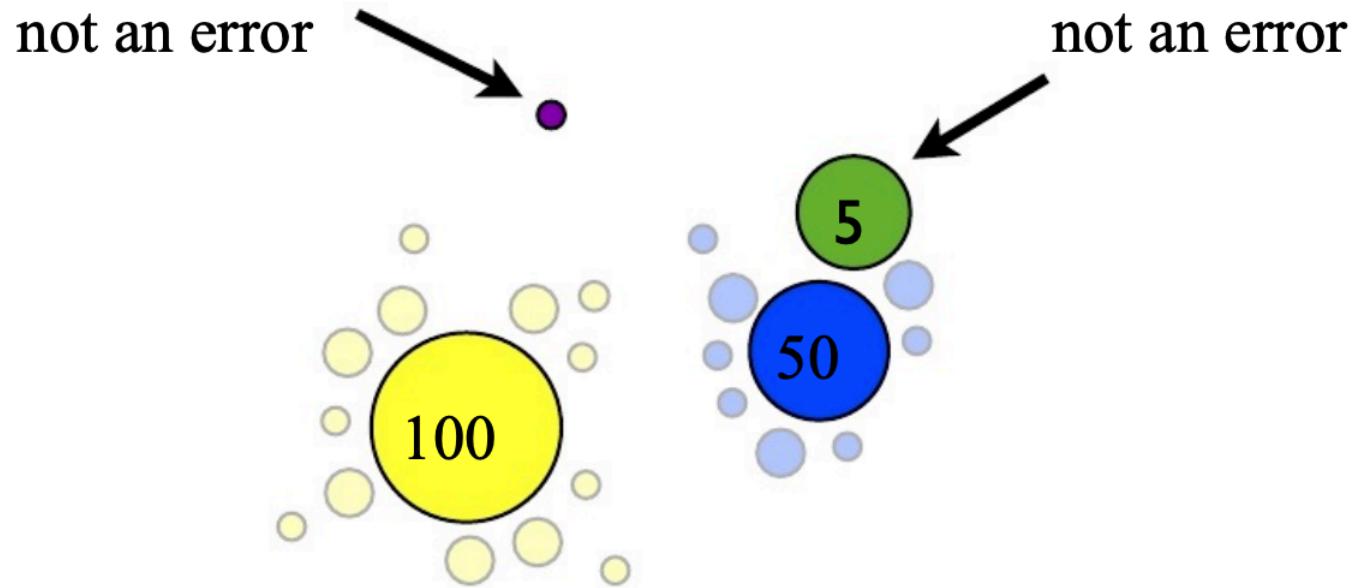
DADA2 algorithm cartoon



Reject unlikely error under model. **Recruit** errors.

	A	C	G	T
A	0.97	10^{-2}	10^{-2}	10^{-2}
C	10^{-2}	0.97	10^{-2}	10^{-2}
G	10^{-2}	10^{-2}	0.97	10^{-2}
T	10^{-2}	10^{-2}	10^{-2}	0.97

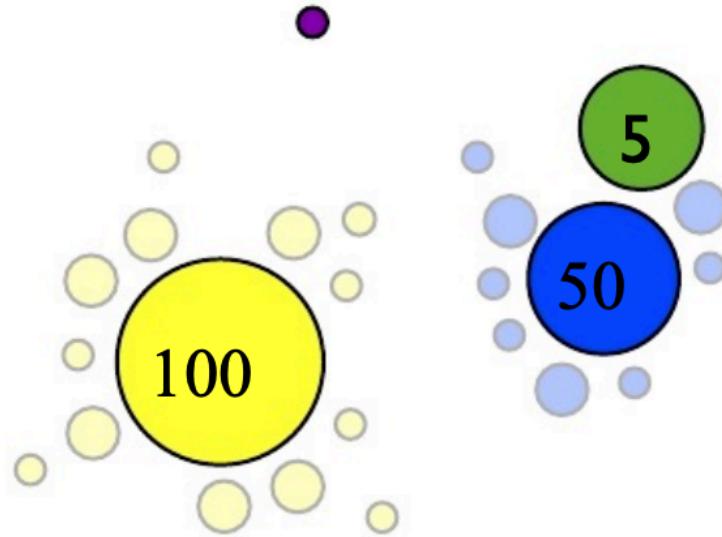
DADA2 algorithm cartoon



Reject more sequences under *new* model

	A	C	G	T
A	0.997	10^{-3}	10^{-3}	10^{-3}
C	10^{-3}	0.997	10^{-3}	10^{-3}
G	10^{-3}	10^{-3}	0.997	10^{-3}
T	10^{-3}	10^{-3}	10^{-3}	0.997

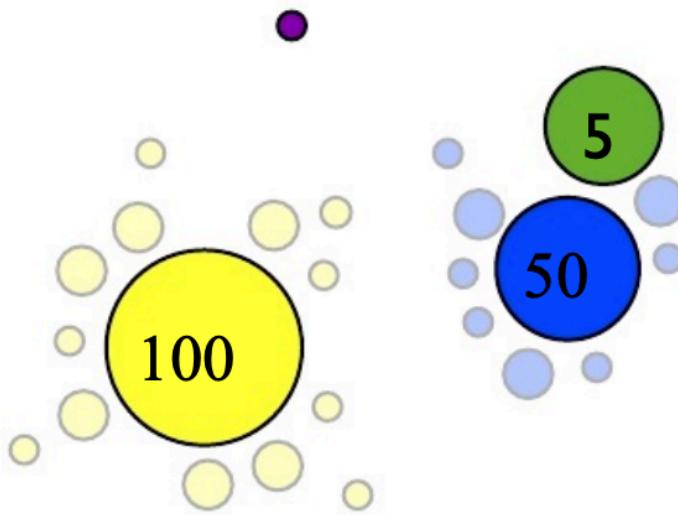
DADA2 algorithm cartoon



Update model again

	A	C	G	T
A	0.998	1×10^{-4}	2×10^{-3}	2×10^{-4}
C	6×10^{-5}	0.999	3×10^{-6}	1×10^{-3}
G	1×10^{-3}	3×10^{-6}	0.999	6×10^{-5}
T	2×10^{-4}	2×10^{-3}	1×10^{-4}	0.998

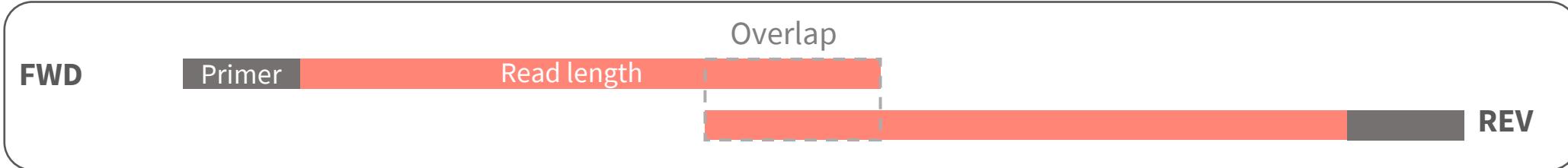
DADA2 algorithm cartoon



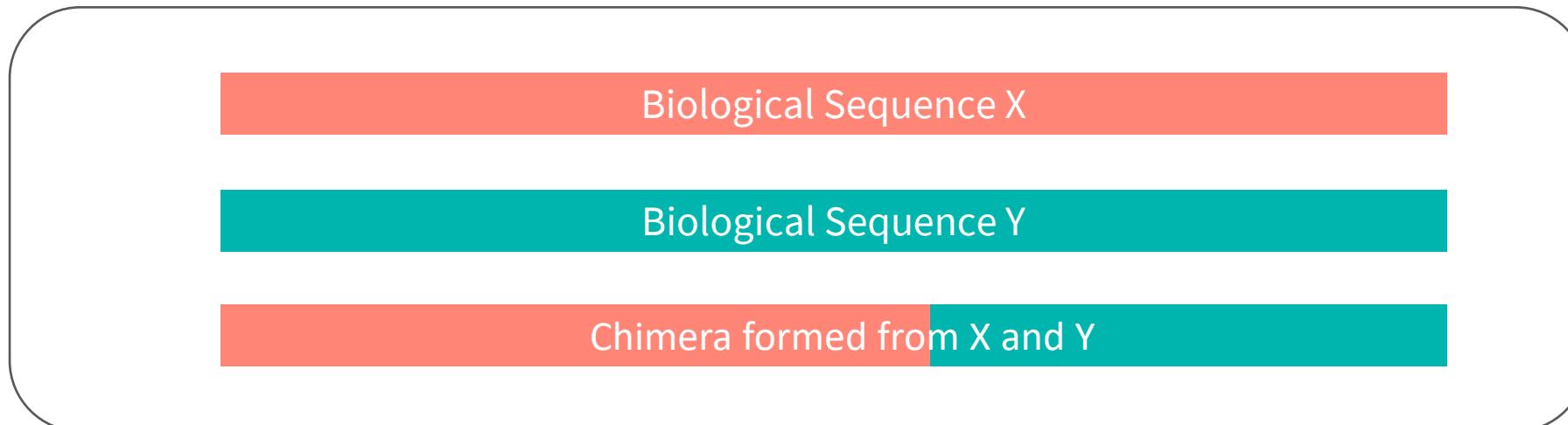
Convergence: all errors are plausible

	A	C	G	T
A	0.998	1×10^{-4}	2×10^{-3}	2×10^{-4}
C	6×10^{-5}	0.999	3×10^{-6}	1×10^{-3}
G	1×10^{-3}	3×10^{-6}	0.999	6×10^{-5}
T	2×10^{-4}	2×10^{-3}	1×10^{-4}	0.998

Merge Paired End Reads



Chimeric Sequences



Created during PCR
Fragment primes different extension
About 1-5 % of reads are chimeric

Important:

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

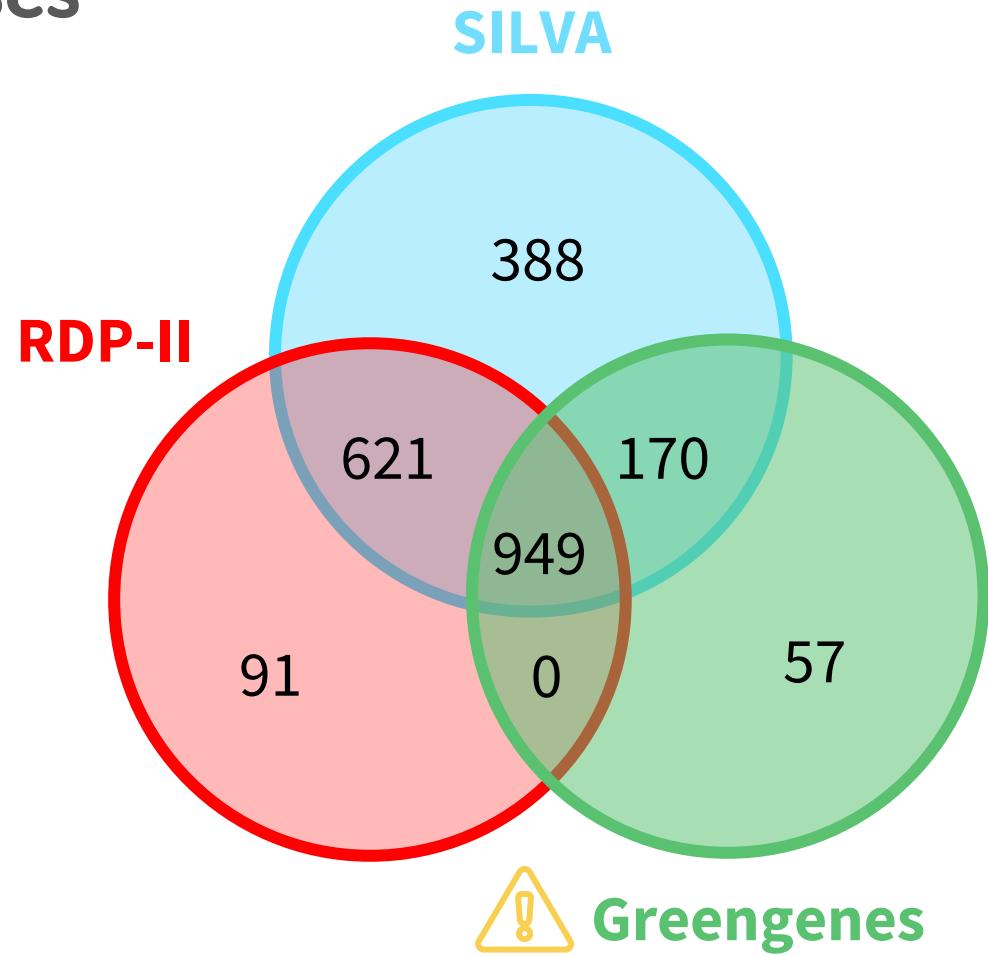
Reference Databases & Taxonomy Assignment

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 ^a

Source: [Balvočiūtė et al 2017](#)

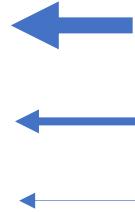
Use a small set of authoritatively classified sequences such as RDP training set or SILVA LTP subset



Source: [Yilmaz et al 2013](#)

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 ^a

Source: Balvočiūtė et al 2017

Use a small set of authoritatively classified sequences such as RDP training set or SILVA

SUGGESTIONS

Length & Depth

Know which region, read length & depth suits your experiment

Negative Controls

Helps identify contaminants

Sample size

Are there enough samples for down stream analysis

Denoise (DADA2)

Remove errors

Avoid C.R. Clustering

Unless you have a very good reason, avoid.

Be consistent

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

Know Why

Understand why you are using a tool or method

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

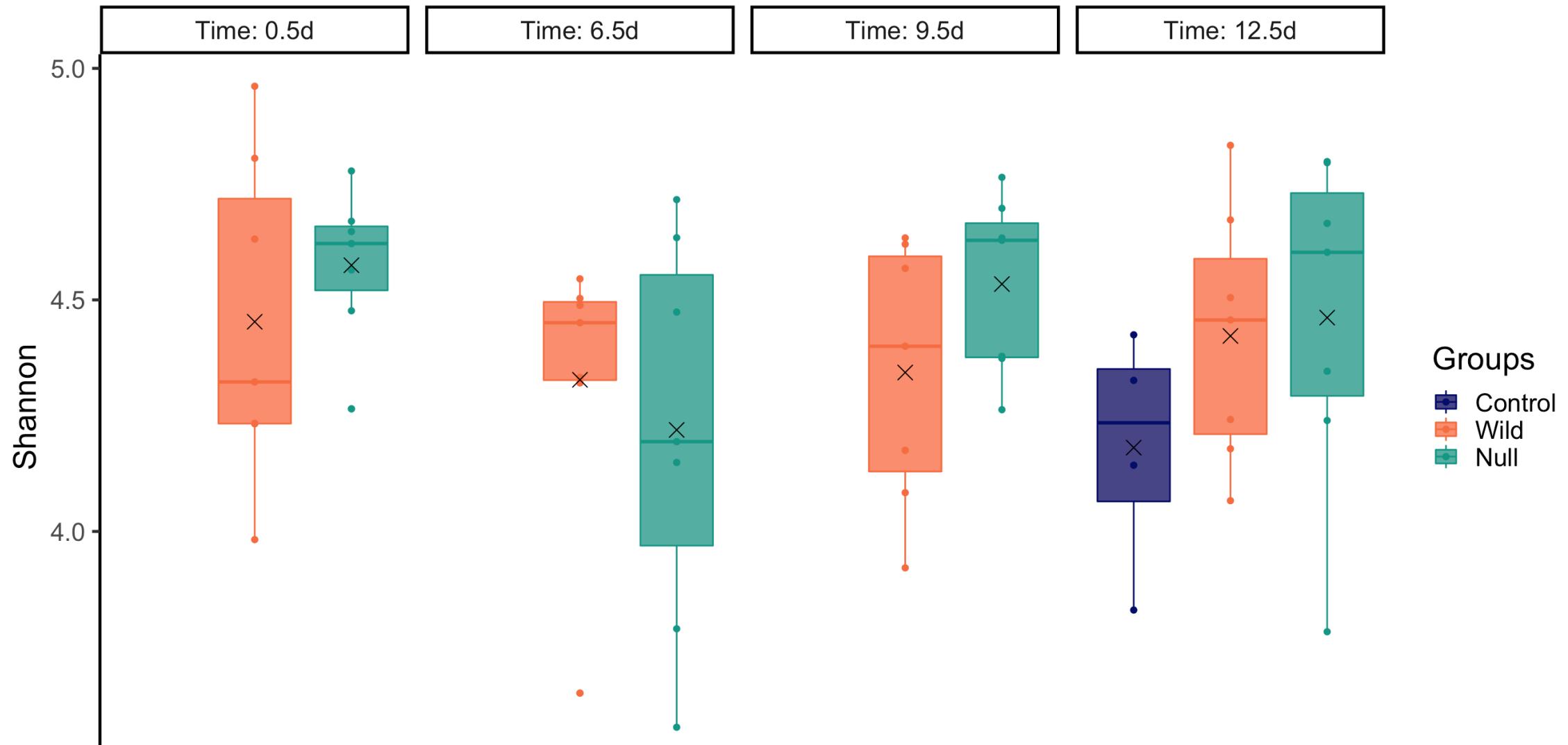
Absent/Present

Bray-Curtis

Weighted-
Unifrac

Abundance

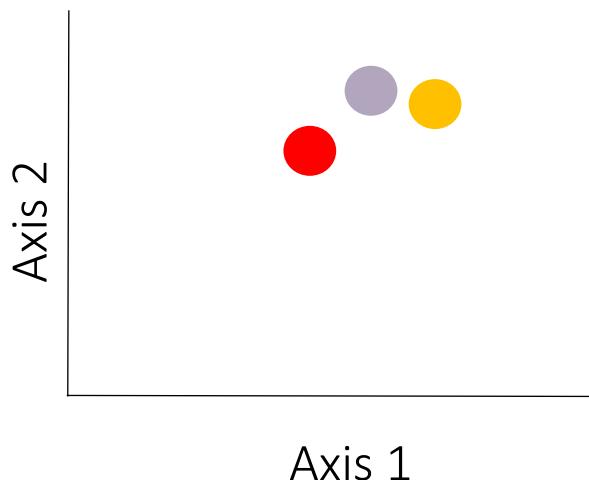
Shannon Index



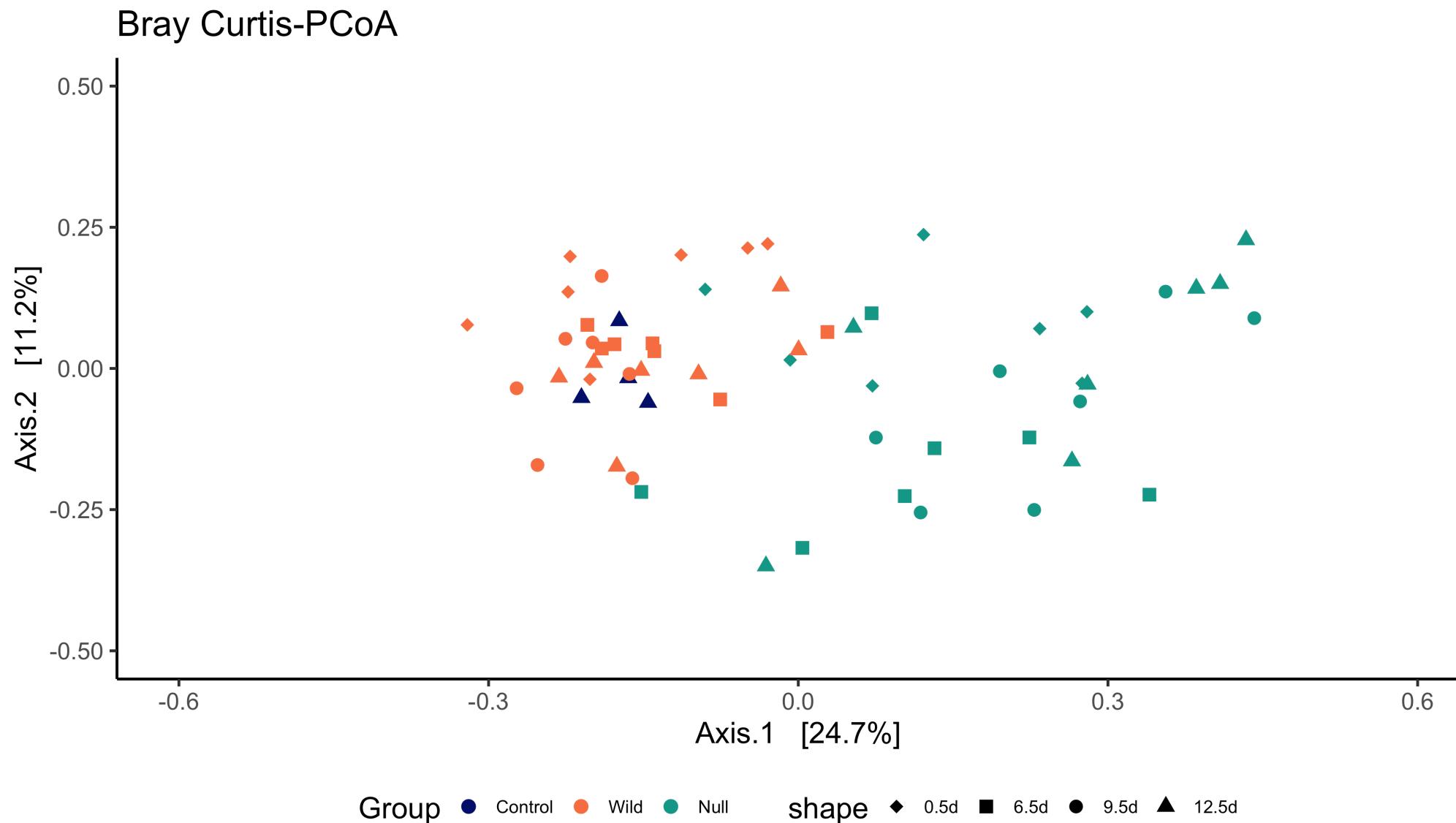
Beta Diversity

	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



What next?

Differential abundance analysis

Network analysis

Predictive functional analysis

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

- Ronald Coase, *Economist*

Taxonomy Assignment

	Alignment-Based	Composition-Based
Method of reference	Sequence Alignment	Shared Feature Vectors (K-mers)
Data used for classification	Reference Sequences	Shared Feature Vectors & probability of taxonomic inclusion
Taxonomic Inference	Identity thresholds	Quantity or proportion of feature vectors shared between reference and query sequence
Available Tools	MEGAN5, RTAX	RDP Classifier, UTAX

Source : [Richardson et al., 2016](#)