16s Microbiome Analysis Workshop

Alana Schick Gut4Health

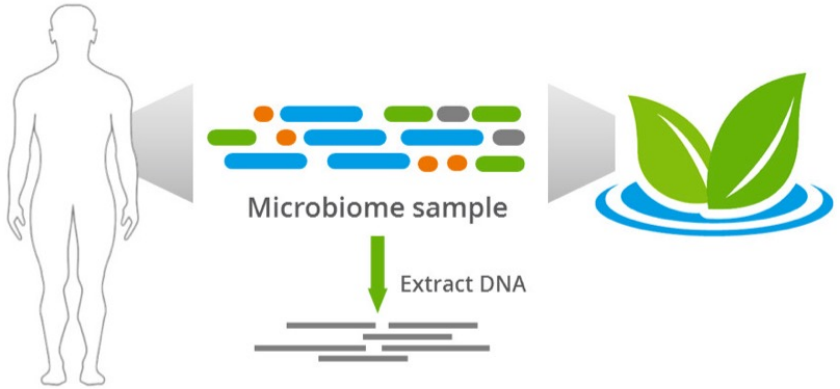
July 18-20, 2022

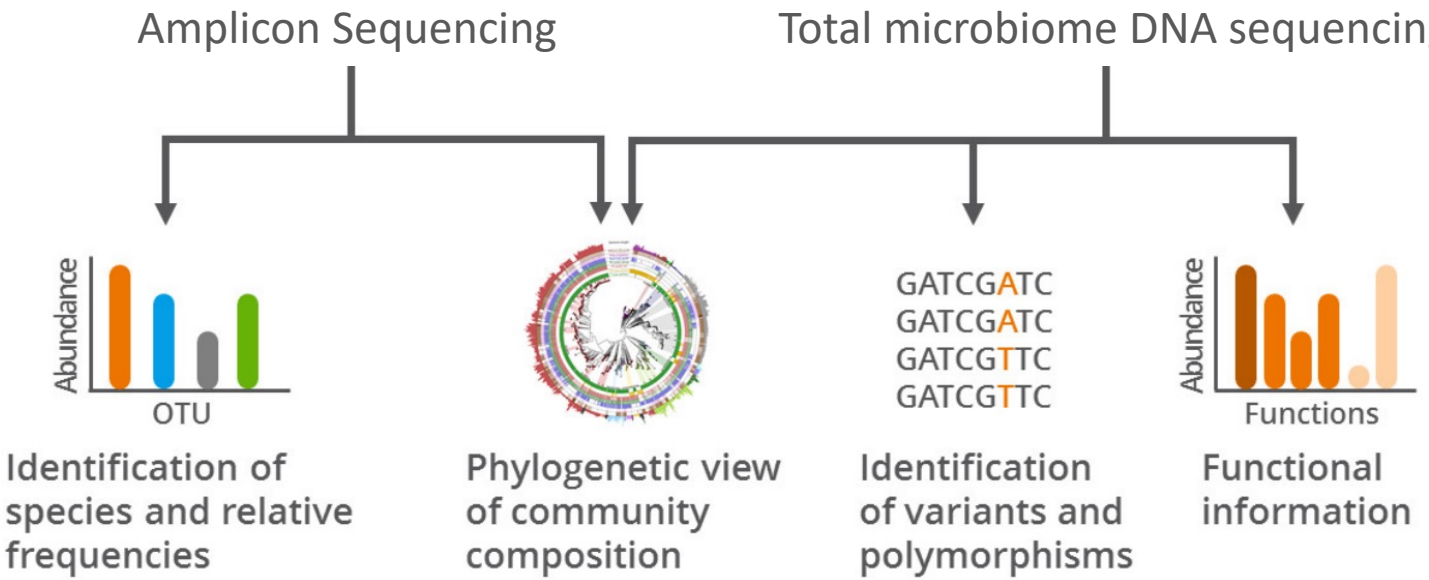
Disclaimer: These slides are taken from difference sources on the web and some are modified. I have added sources but if I have missed some, I apologize in advance to the original content creators.

Analyzing the microbiome

Who lives here?

Over time + Treatment



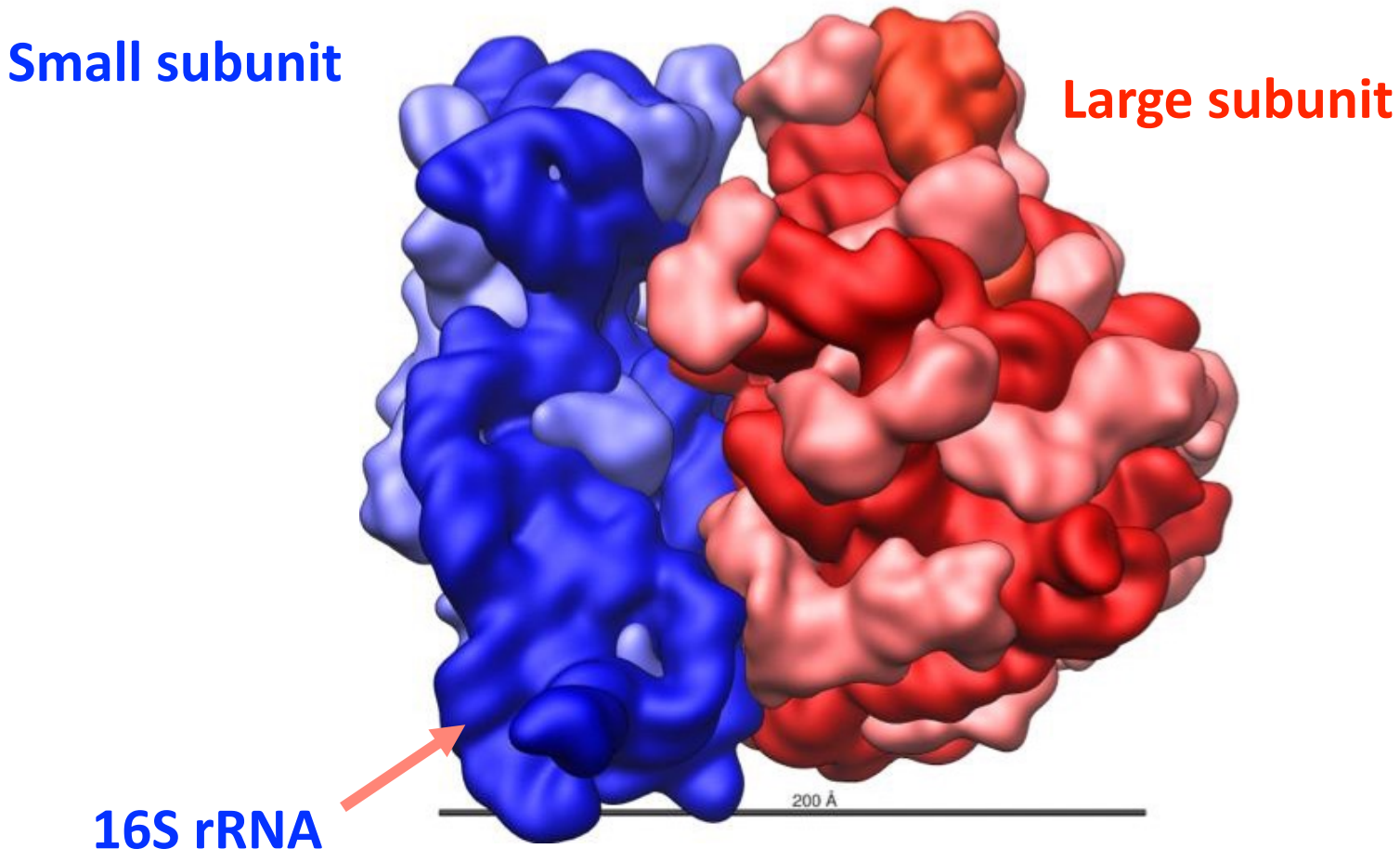


https://www.gatc-biotech.com/en/expertise/targeted-sequencing/16s-rrna-analysis.html

Amplicon Sequencing

What should we amplify?

An *E. coli* ribosome

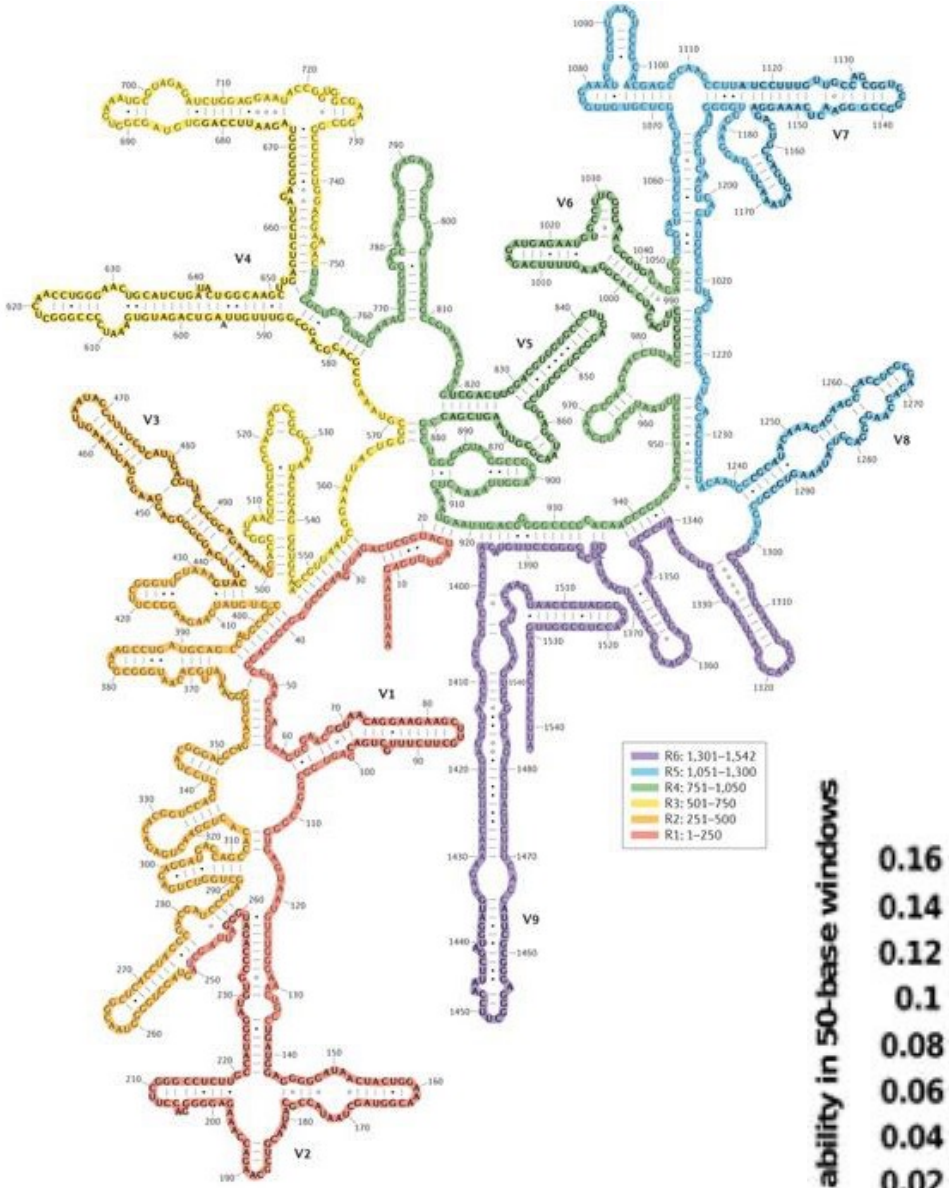


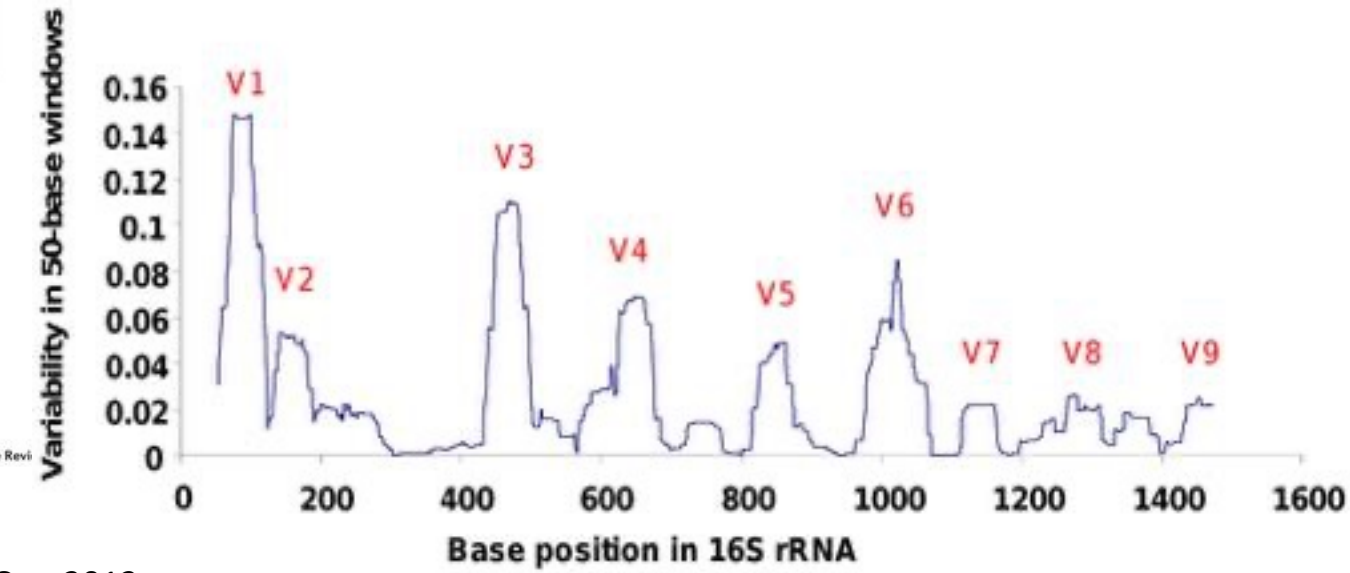
https://commons.wikimedia.org/wolo/File:Ribosome_shape.png

Why 16S rRNA

- rRNA is one of the few gene products present in all cells
- 16S rRNA has 9 **hypervariable regions** allowing species identification, as well as conserved regions allowing **primer construction**
- conserved function
- sequence has been characterized for many species

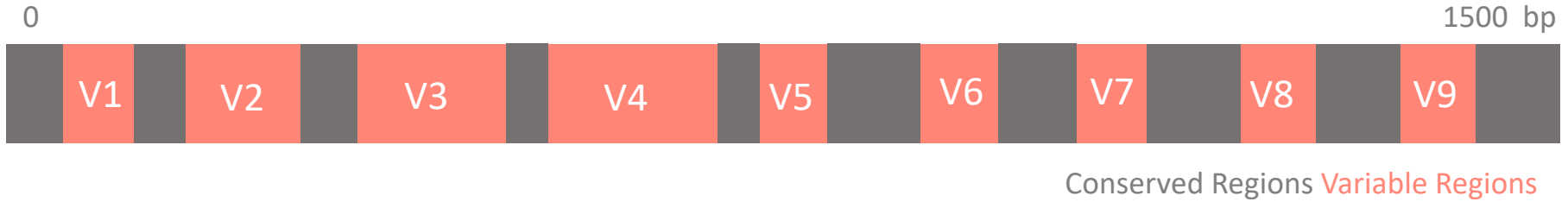
16S rRNA





Yarsa et al., Nature Rev Microbiology 2014; Bodilis et al., PLoS One 2012

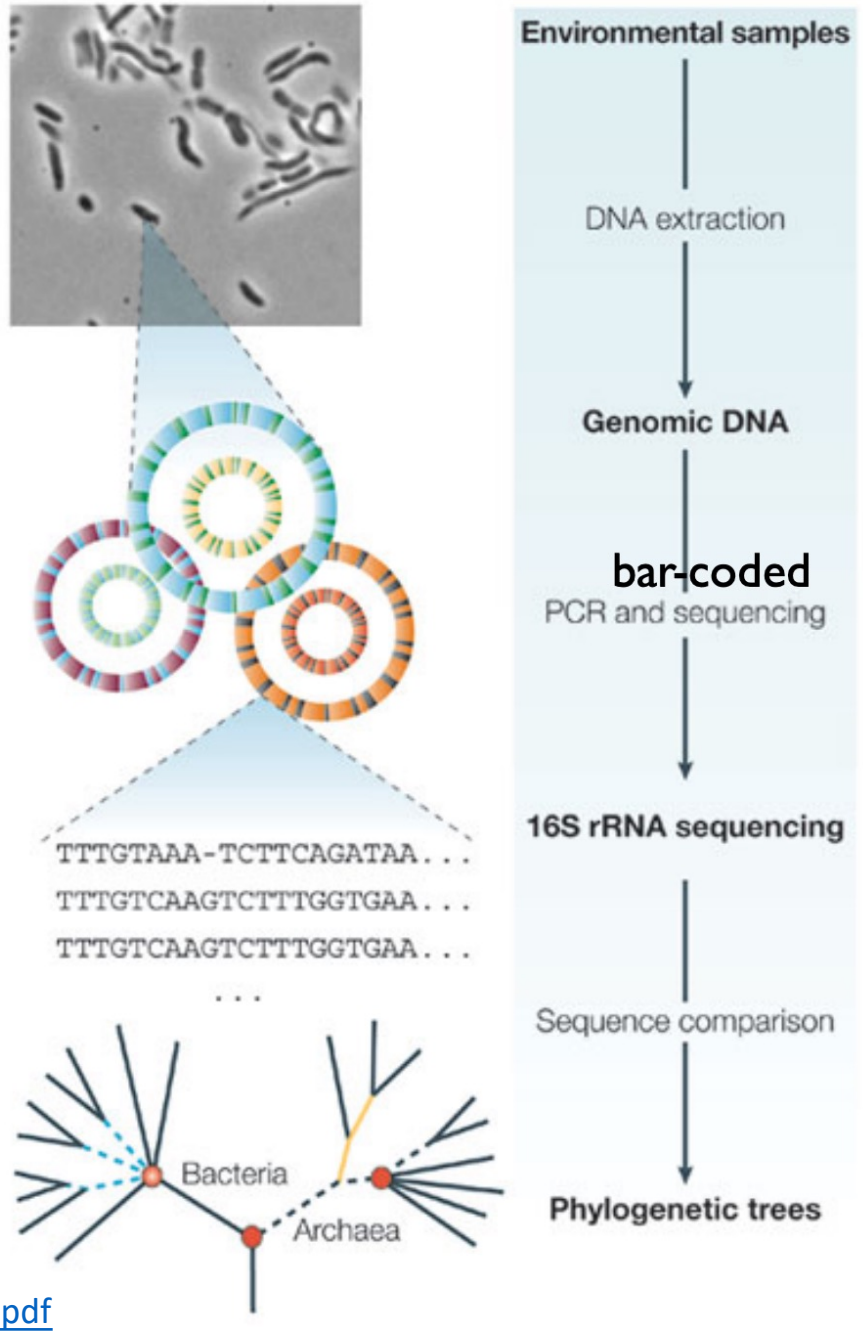
16S is not perfect



- 16S doesn’t capture all differences between the full DNA sequences
- Different species can have similar 16S sequences
- A single species can have paralogs that are not identical
- Results can depend on which variable region is considered

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



https://www.biostat.washington.edu/sites/default/files/modules/2018-SISMID-01b.pdf

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

Negative Controls

Was a negative control included?

Feasibility

Will this data answer the questions asked by the investigator?

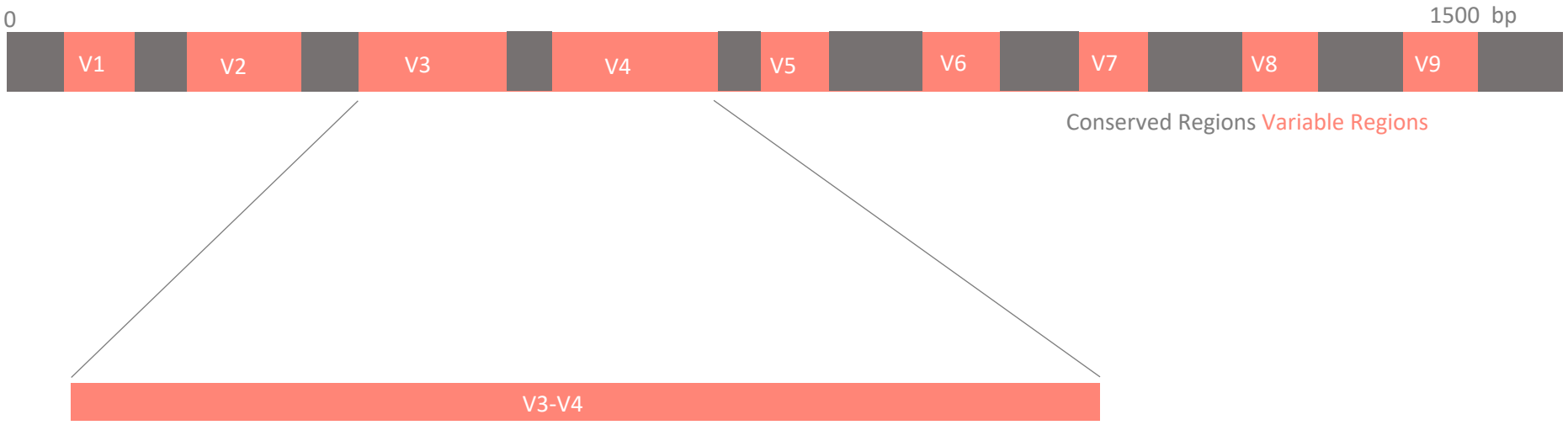
Read Assignment

Are the reads assigned correctly to each sample?

Sample size

Are there enough samples for down stream analysis

16 rRNA

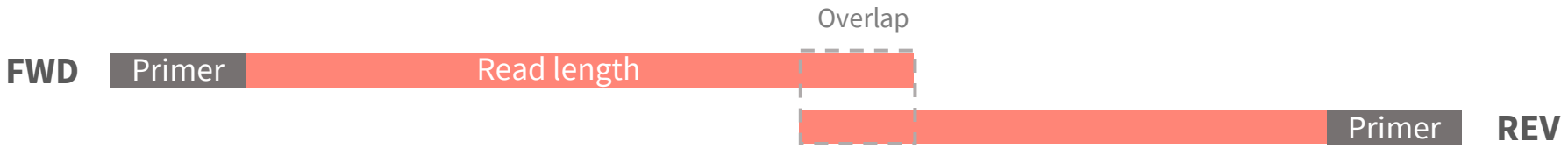


Pick a variable region

Pick a sequencing technique

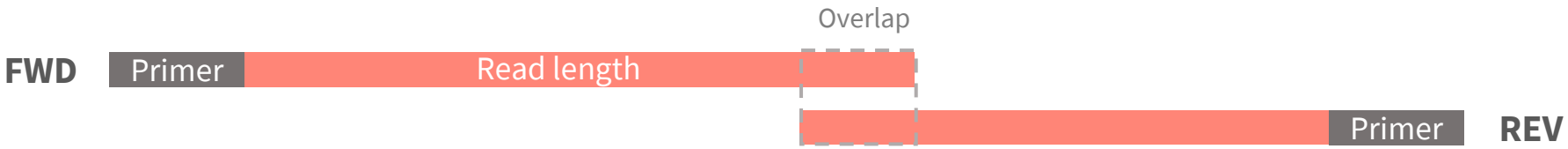
Illumina (MiSeq)

Select primers & Read lengths



Make sure there is an overlap!!!!

How to calculate V region overlap



Overlap = Read length X 2 – primer_F – primer_R – VR_length

Example: V4

VR_length = 254 bp

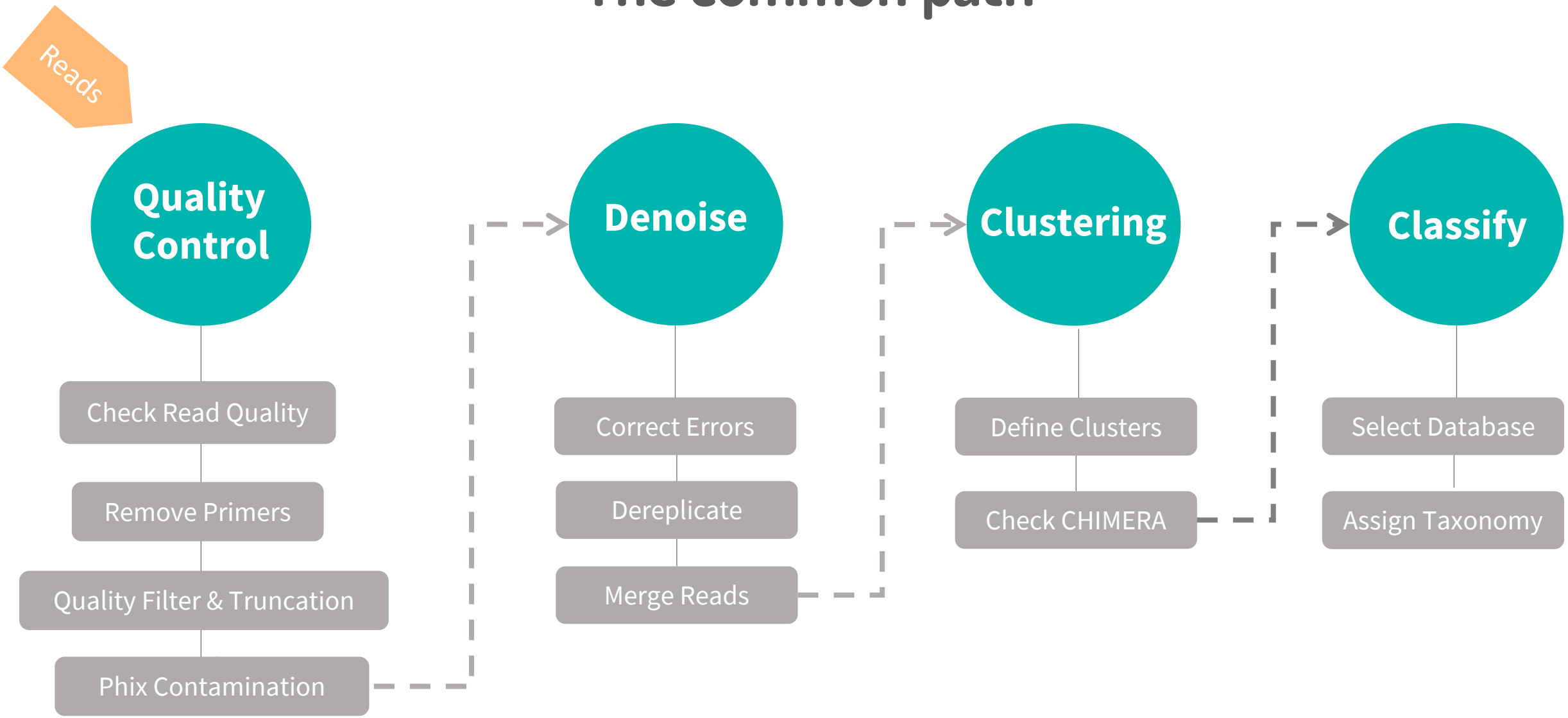
read length = 250 x 2 bp and

primers = 20 bp

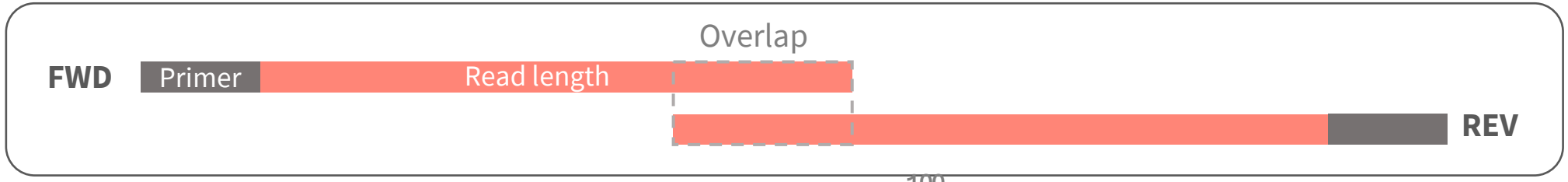
Overlap= 250 *2 – 20 – 20 – 254 = 206

Typical Workflow

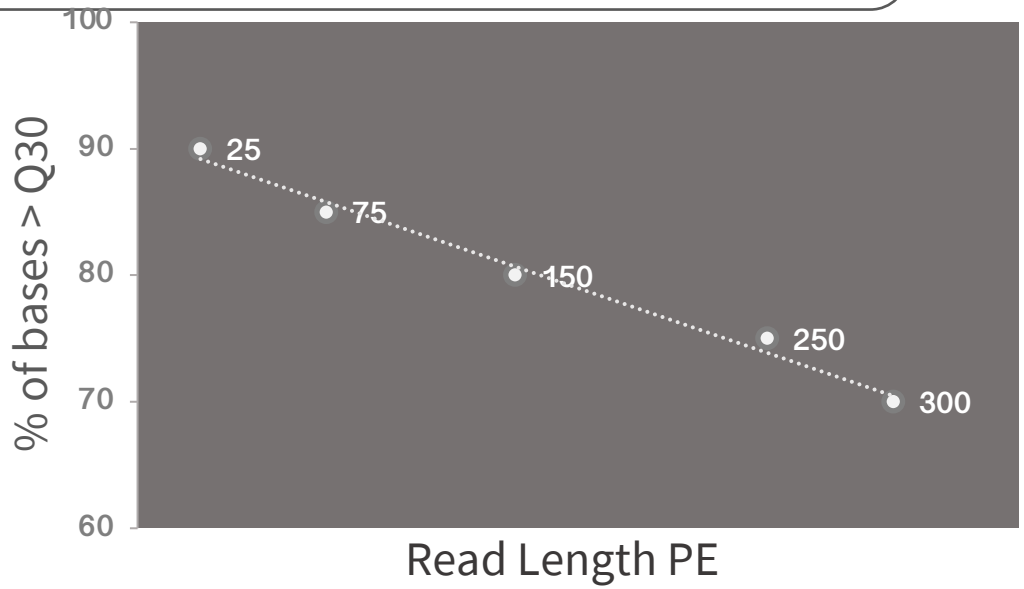
The common path



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



Source: Illumina

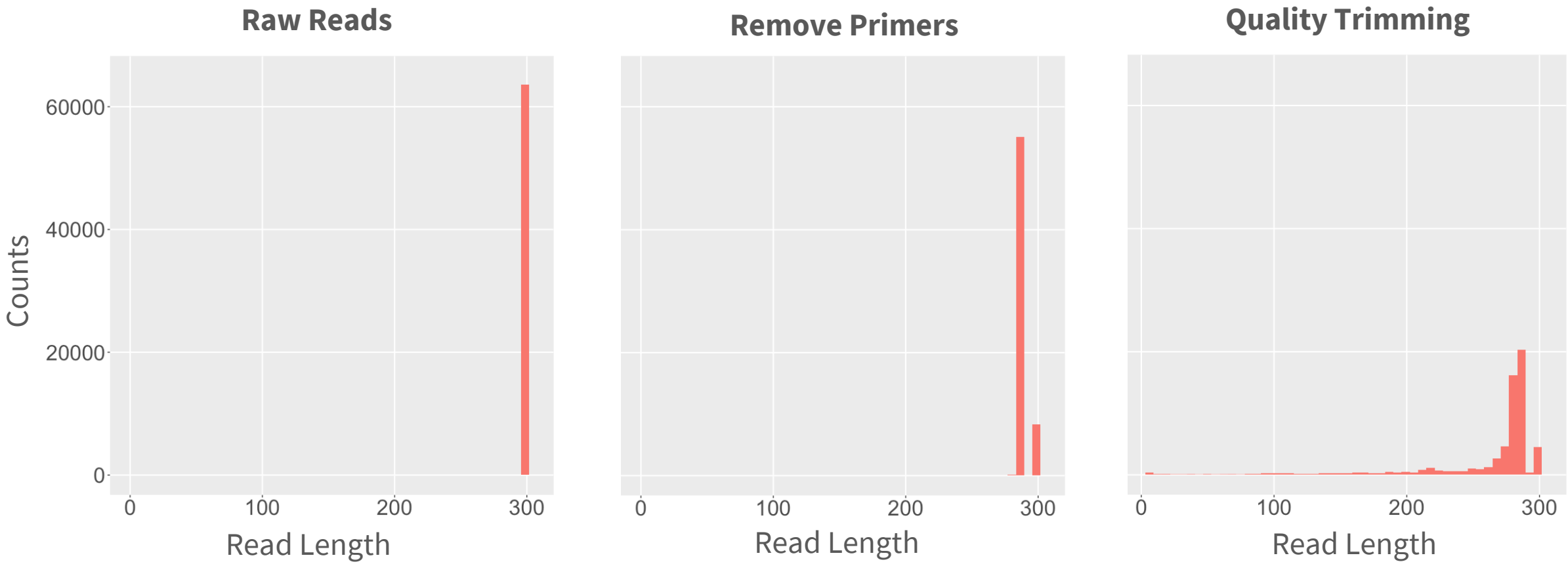
Base Quality differs between Fwd and Rev Reads

Tools

FastQC FastQp MultiQC & Specialized 16s rRNA packages

Source: Dada2 R package

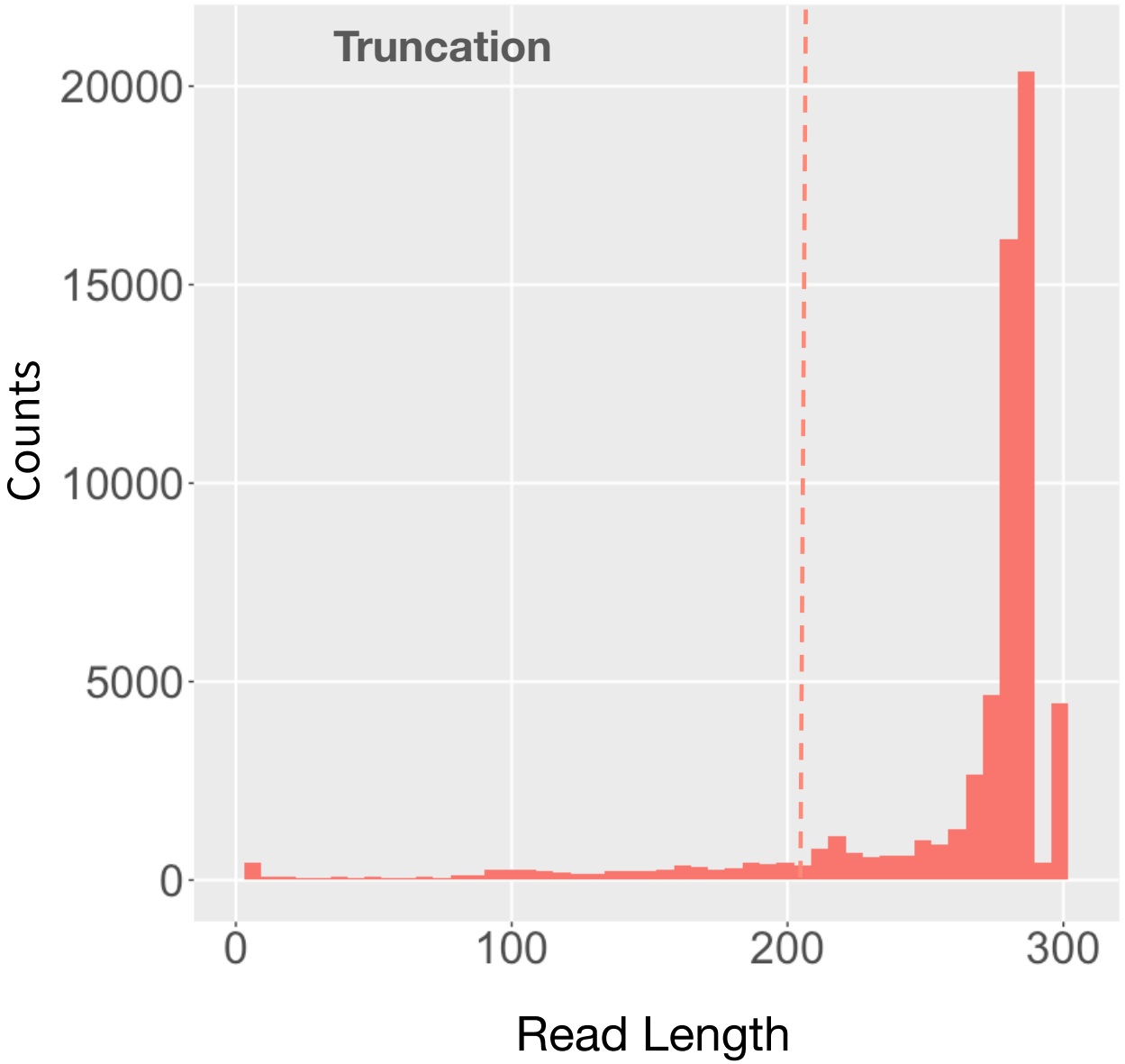
Trimming & Quality Filtering



Tools

CutAdapt

Trimming & Quality Filtering



Available Methods

Minimum Q (Q >= 20)

Truncate if 3 consecutive bases are Q < 3

Expected Errors

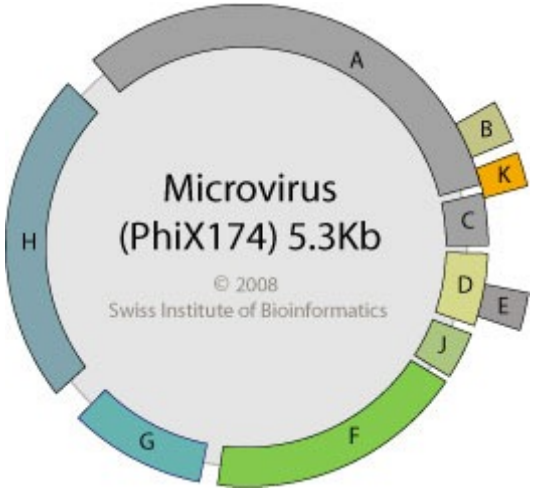
Read length	% of reads
10	98.96
100	96.97
200	89.3
250	80.5

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

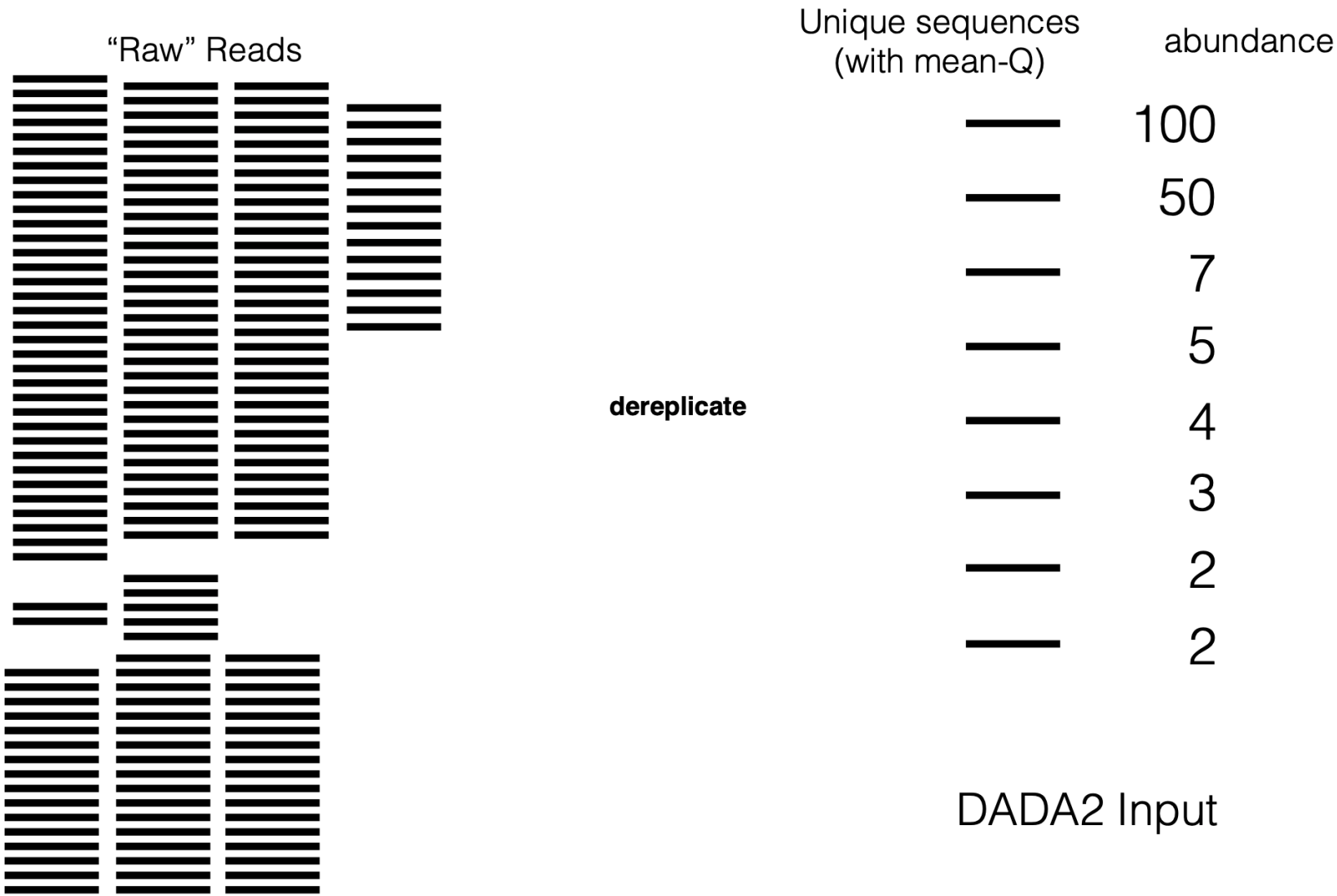
Blast sample reads against PhiX genome

Source: Mukherjee et al.,2015

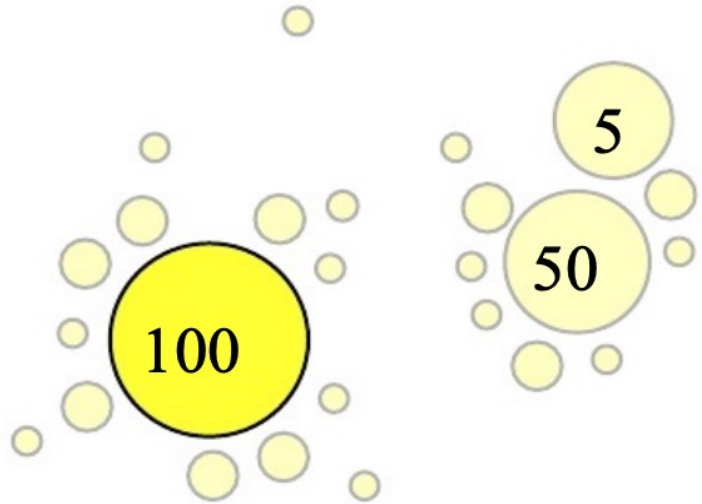
Denoising reads

DADA2 algorithm

Input: unique sequences, their quality values, and abundances



DADA2 algorithm

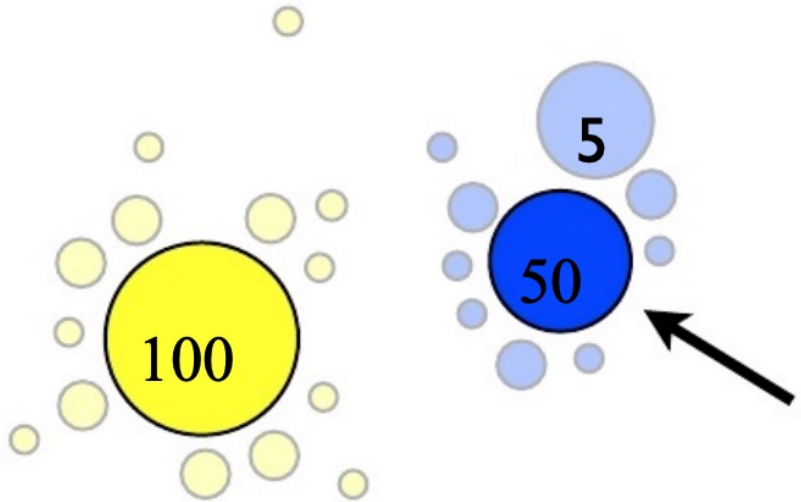


Infer initial *error model* under this assumption.

$\Pr(i \rightarrow j) =$

	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

DADA2 algorithm

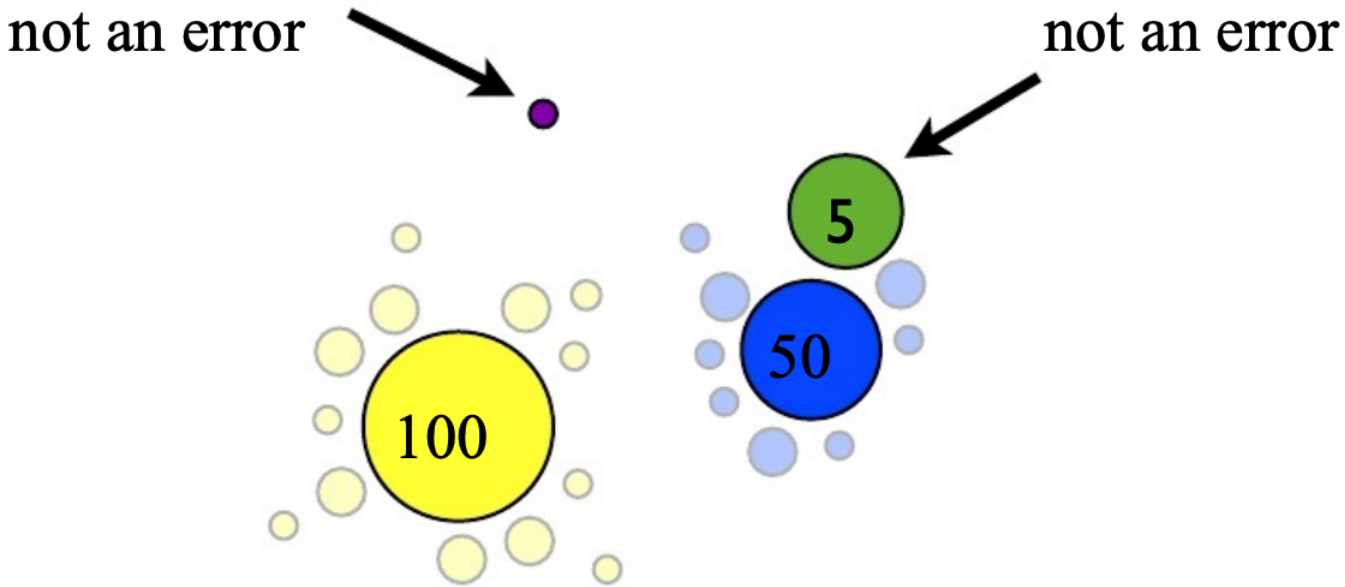


not an error

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

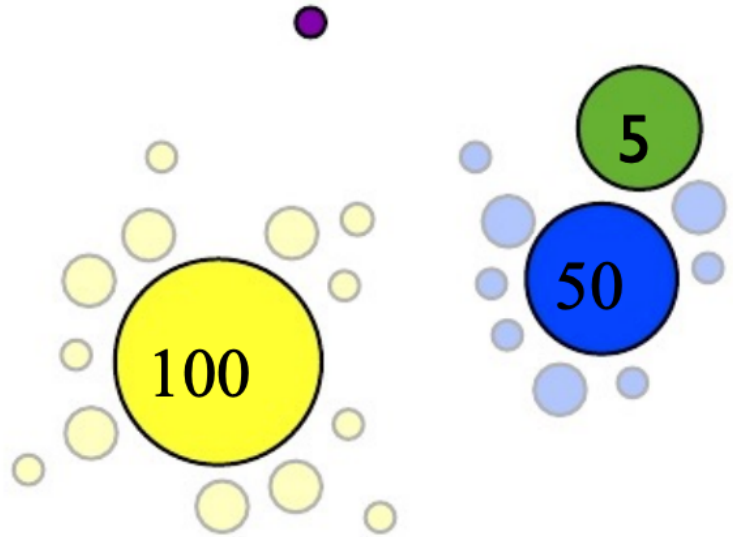
DADA2 algorithm



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

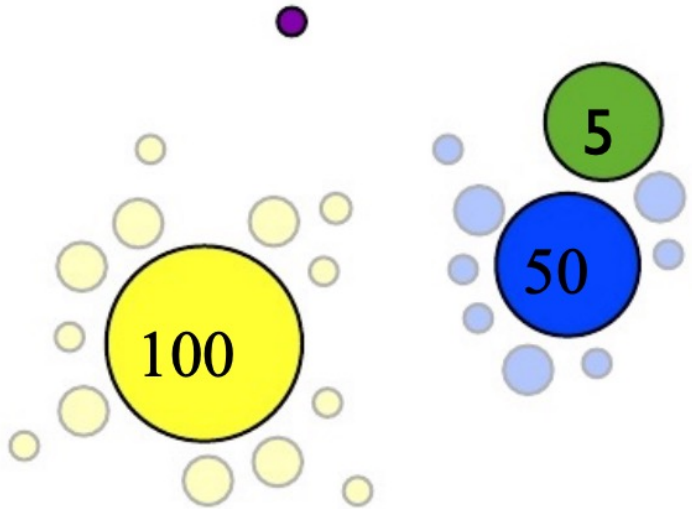
DADA2 algorithm



Update model again

	A	C	G	T
A	0.998	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$	
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$	
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$	
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998	

DADA2 algorithm

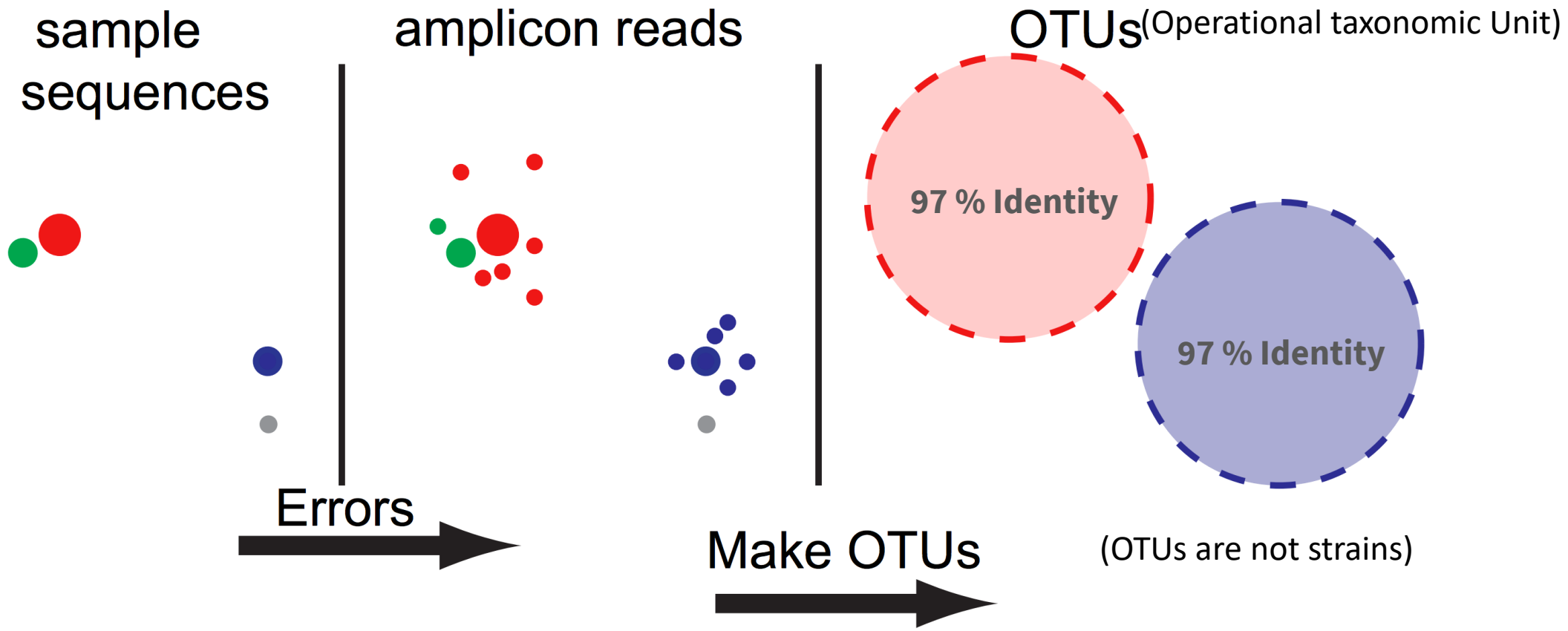


Convergence: all errors are plausible

	A	C	G	T
A	0.998	$1x10^{-4}$	$2x10^{-3}$	$2x10^{-4}$	
C	$6x10^{-5}$	0.999	$3x10^{-6}$	$1x10^{-3}$	
G	$1x10^{-3}$	$3x10^{-6}$	0.999	$6x10^{-5}$	
T	$2x10^{-4}$	$2x10^{-3}$	$1x10^{-4}$	0.998	

Clustering reads

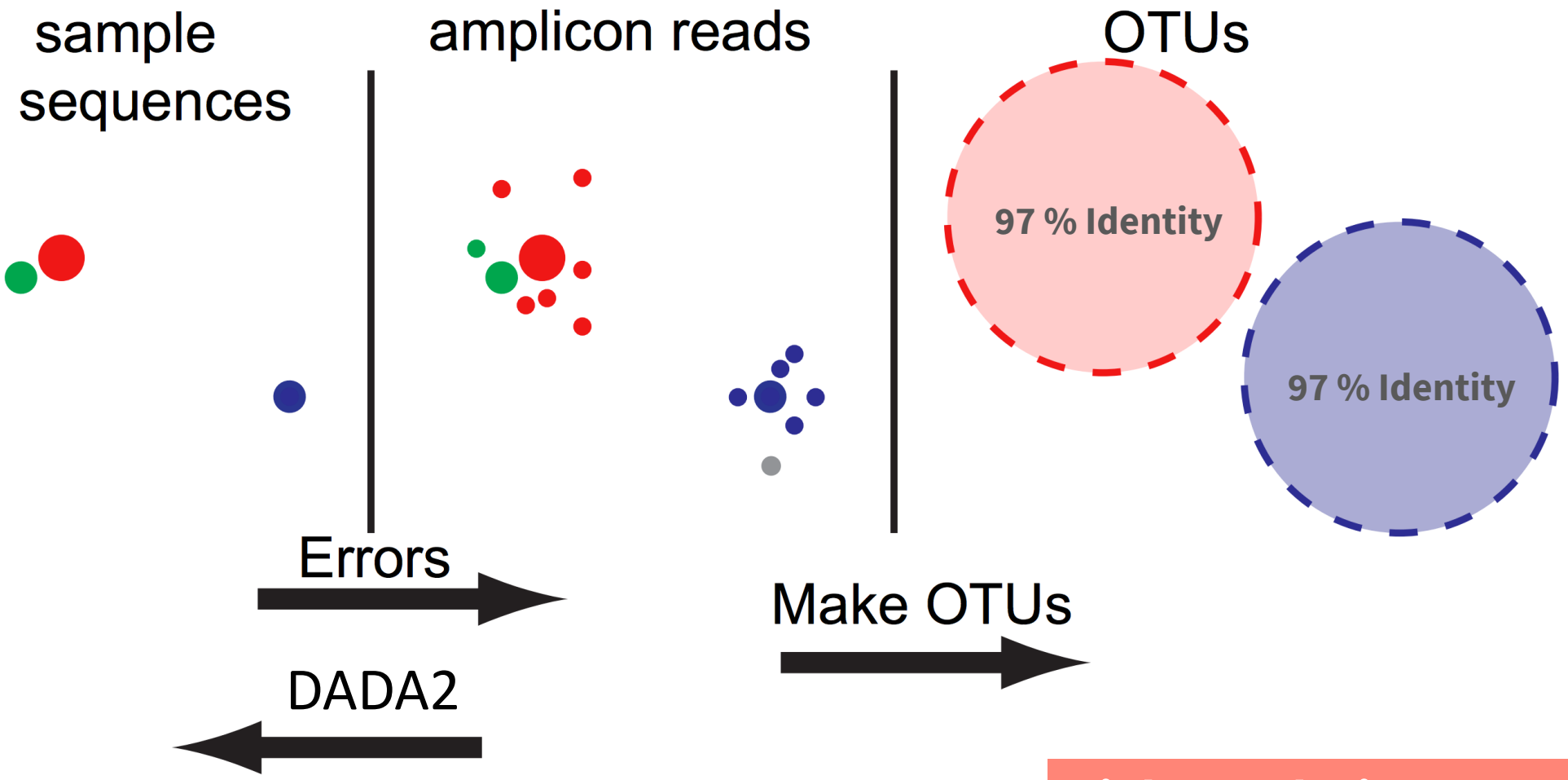
Clustering



OTUs: Lump similar sequences together
DADA2: Statistically infer the sample sequences (Amplicon sequence variants: ASVs)

Source: Callahan et al.,2016

Clustering

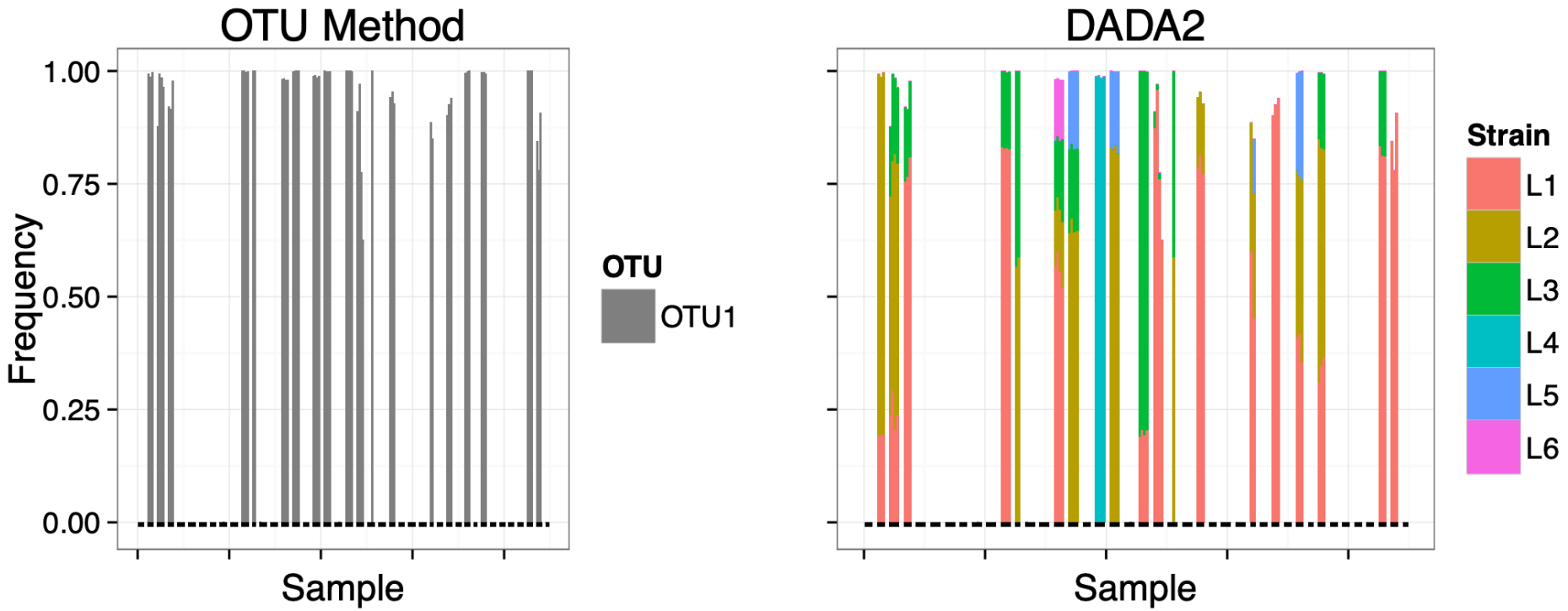


High Resolution & Accuracy

Source: Callahan et al.,2016

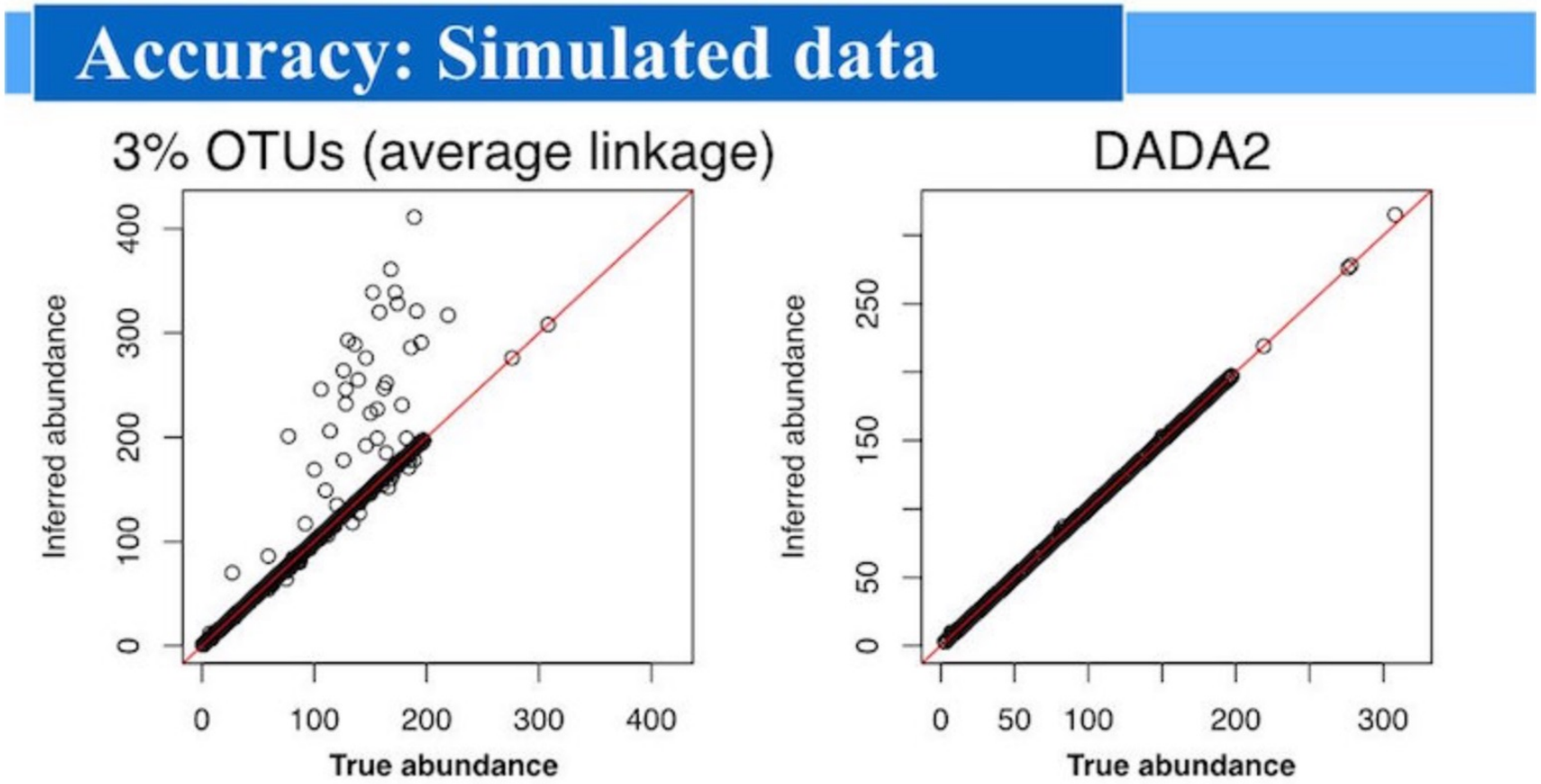
Real example, exact sequence resolution

Lactobacillus crispatus sampled from
vaginal microbiome 42 pregnant women



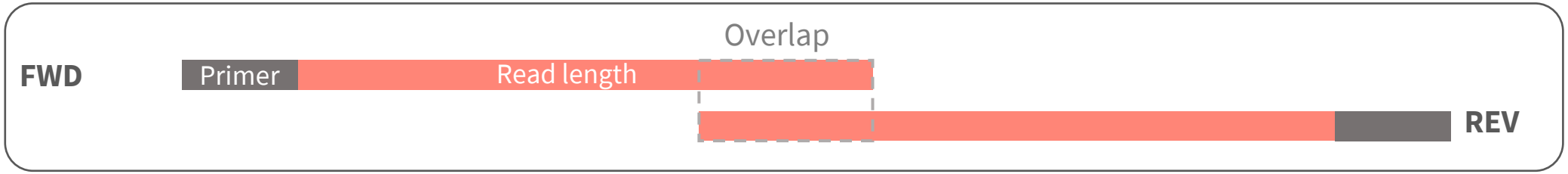
Data: MacIntyre et al. Scientific Reports, 2015.

Clustering

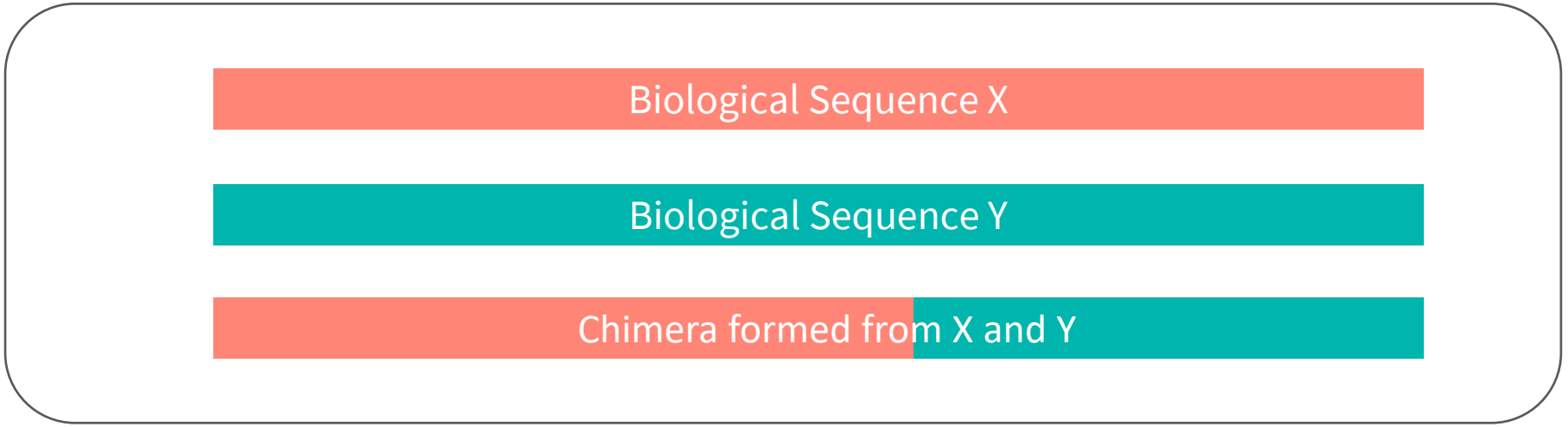


Don’t cluster unless you have a very good reason to do so.

Merge Paired End Reads



Chimeric Sequences



Created during PCR

Fragment primes different extension

About 1-5 % of reads are chimeric

Source: Lahr & Katz, 2009

Reference Databases
&
Taxonomy Assignment

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

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 Clustering
Unless you have a very good reason, avoid

Document
Make sure your workflow is well documented and reproducible

Know Why
Understand why you are using a tool or method

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

– Ronald Coase, *Economist*