MB590-012
Microbiome Analysis

Identifying ASVs using DADA2

Dr. Christine Hawkes

NC STATE UNIVERSITY

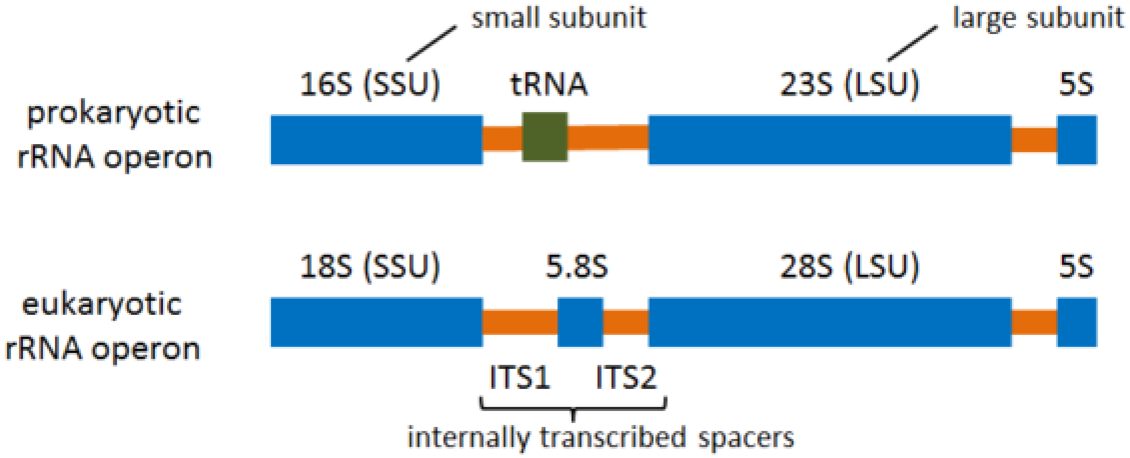
Ribosomal sequences for microbial identification

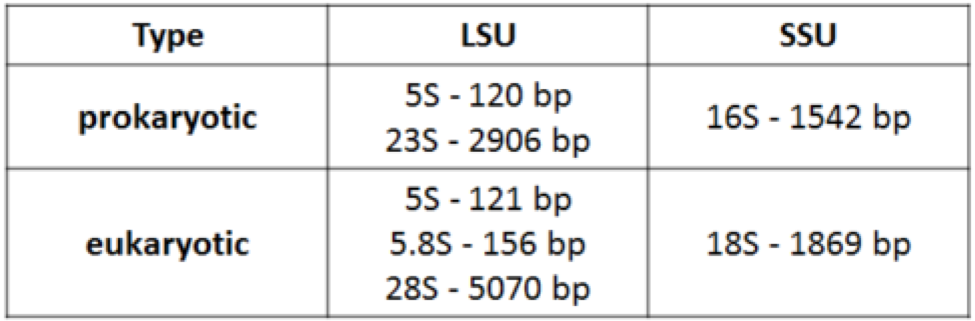
- Practical species concept - taxa are “operationally” defined based on short sequence similarity

- Concept developed exclusively for microorganisms
  - contrast with biological or morphological concepts for macroorganisms

- Other concepts useful for microorganisms:
  - Phylogenetic (shared ancestry)
  - Genomic (shared genome)
  - Both also based on sequence data

Ribosomal sequences for microbial identification

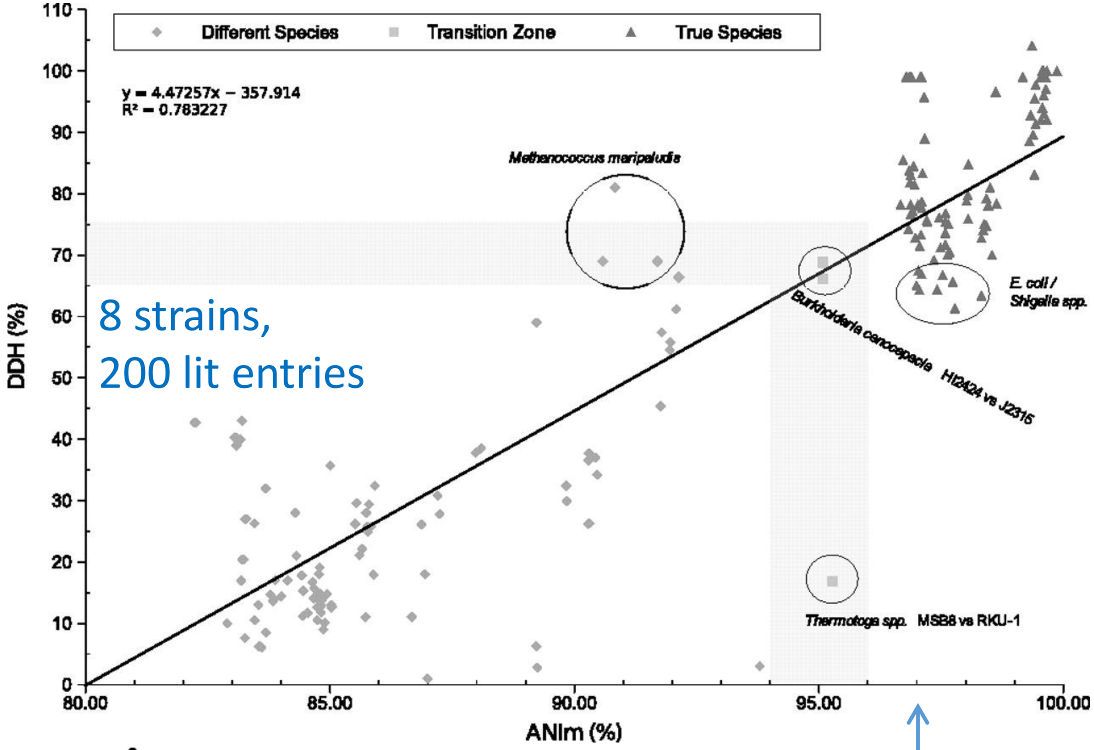




Type	LSU	SSU
prokaryotic	5S - 120 bp 23S - 2906 bp	16S - 1542 bp
eukaryotic	5S - 121 bp 5.8S - 156 bp 28S - 5070 bp	18S - 1869 bp

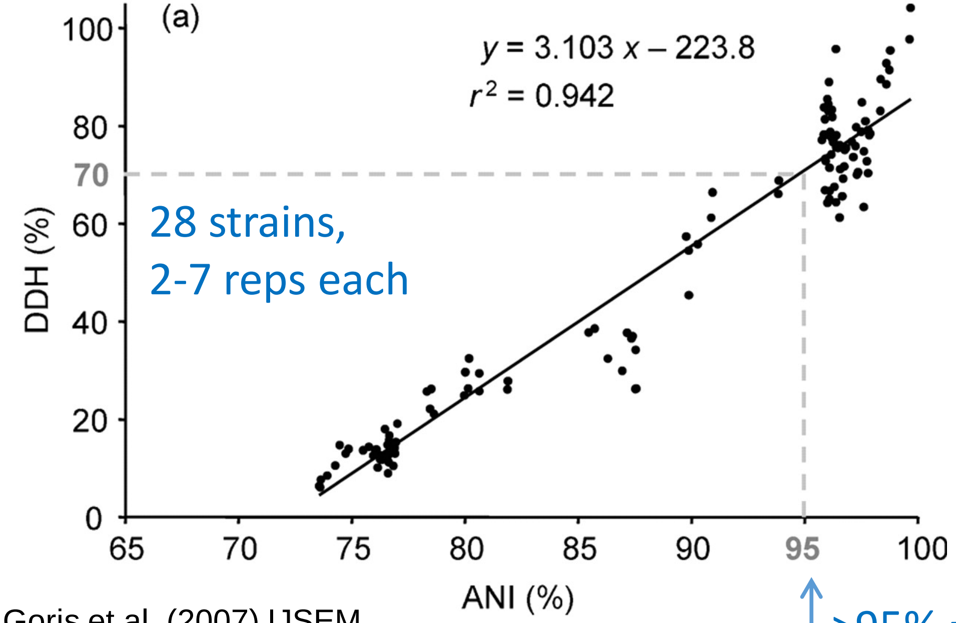
http://www.sixthresearcher.com/amplicon-sequencing-and-high-throughput-genotyping-metagenomics/

Ribosomal sequences for microbial identification



8 strains,
200 lit entries

97% = >70% DDH
(defined here as
“true” species)



28 strains,
2-7 reps each

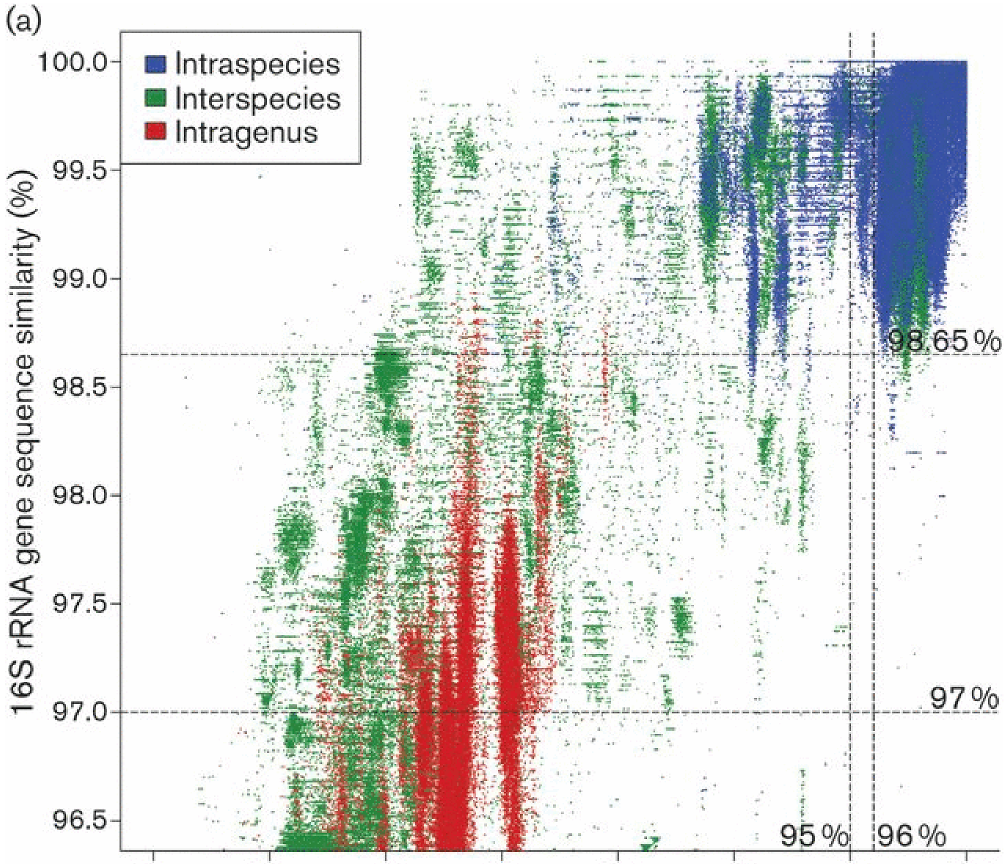
DDH = DNA-DNA hybridization
ANI = average nucleotide similarity

>95% => 70% DDH

Goris et al. (2007) IJSEM
https://doi.org/10.1099/ijs.0.64483-0

Richter & Rosselló-Móra (2009) PNAS
https://doi.org/10.1073/pnas.0906412106

Ribosomal sequences for microbial identification



Genome similarity
(Average nucleotide identity %)

More accurate cutoff for species delineation

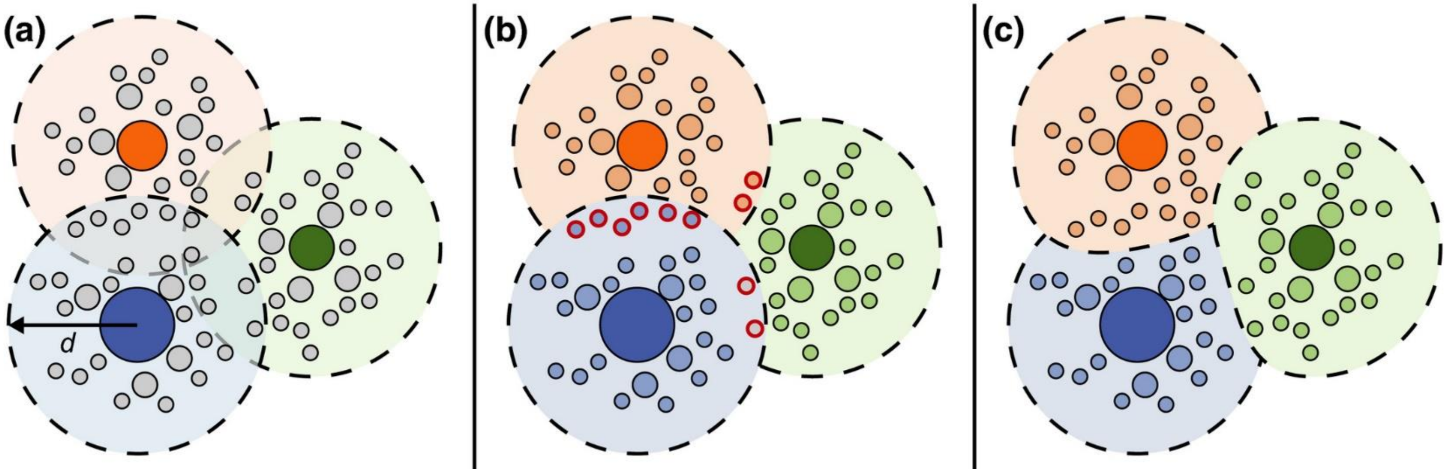
Traditional OTU cutoff

Updated with 6787 prokaryotic genomes

Kim et al. 2014 Int J Syst Evo Micro 64: 10.1099/ijs.0.059774-0

Operational Taxonomic Units

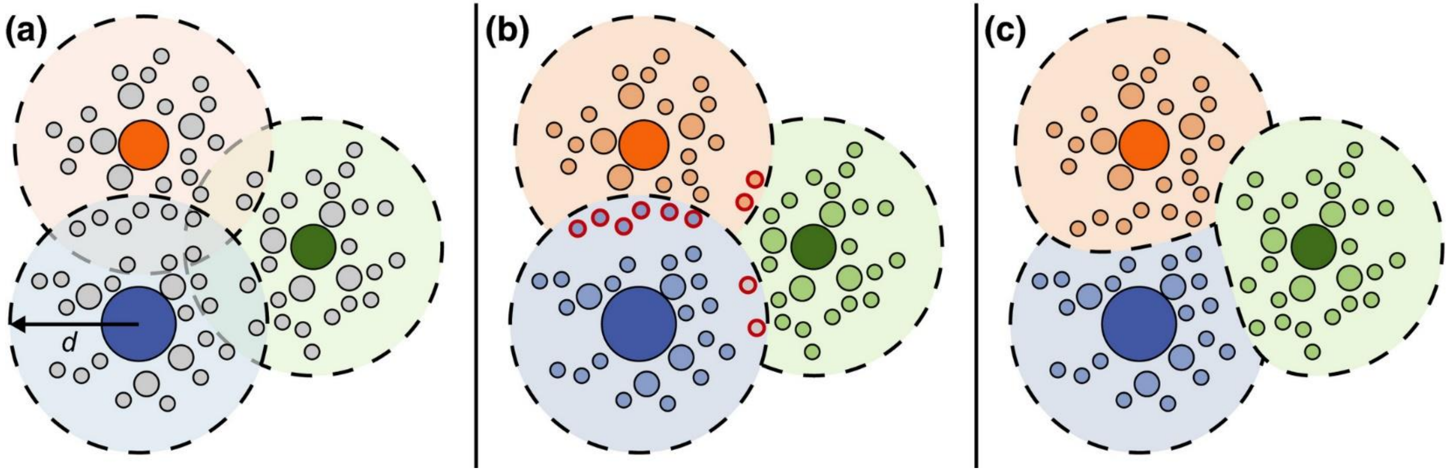
Various clustering algorithms used based on % similarity to approximate species as an “operational taxonomic unit”



Terrat et al. 2019 Methods Eco Evo doi.org/10.1111/2041-210X.13316

Operational Taxonomic Units

Various clustering algorithms used based on % similarity to approximate species as an “operational taxonomic unit”

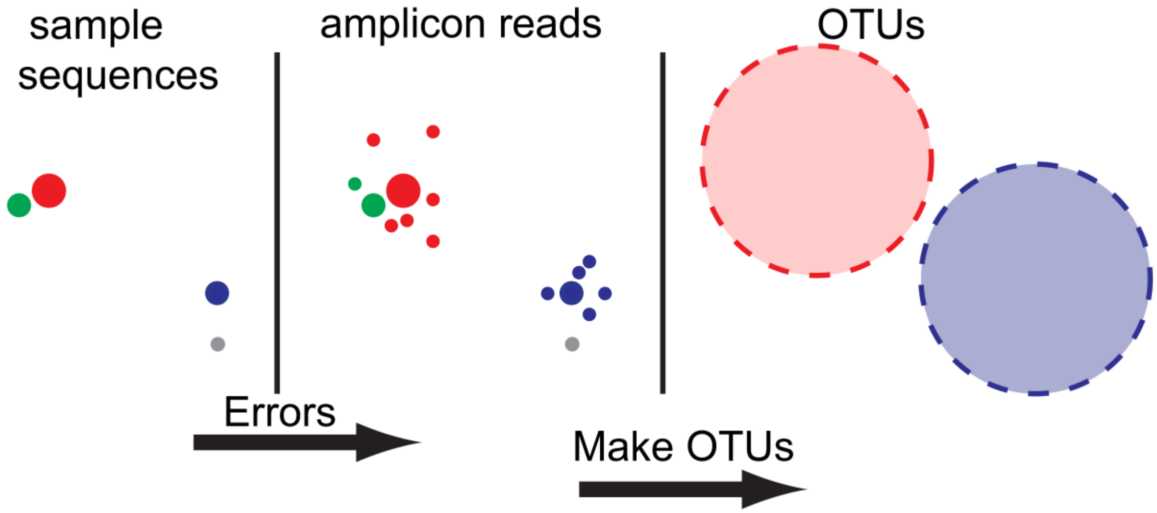


Problems:

- arbitrary classification – OTUs have no meaning in real world
- low repeatability of classification
- cannot compare across studies

Terrat et al. 2019 Methods Eco Evo doi.org/10.1111/2041-210X.13316

Why DADA2?

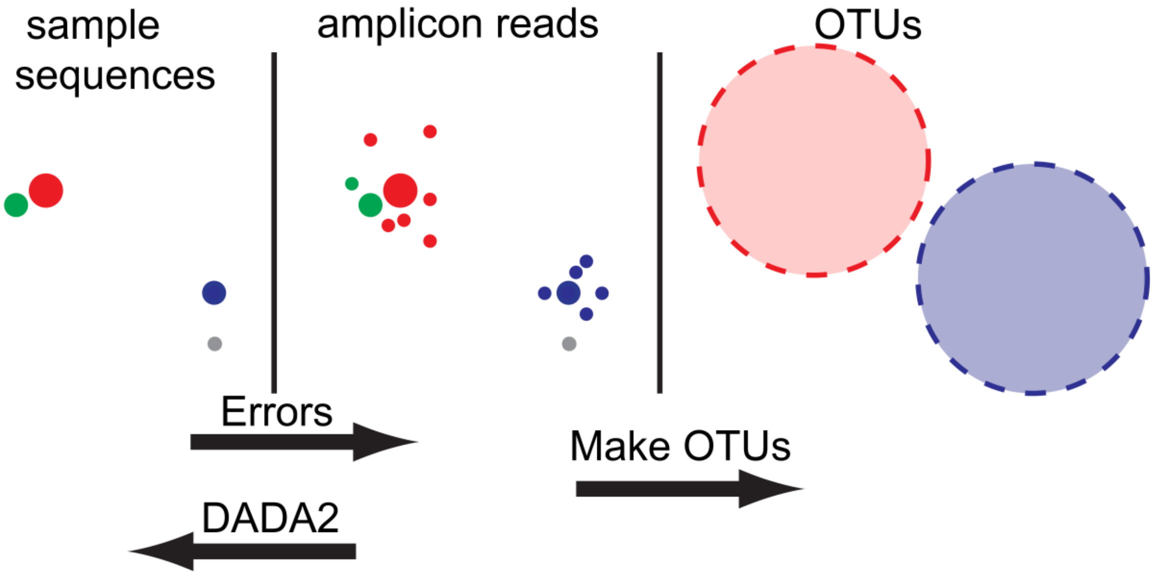


OTU methods limit false positives, but in doing so also lump distinct taxa (=false negatives)

Callahan et al. 2016. Nature Methods 13: 581-583 https://www.nature.com/articles/nmeth.3869

Why DADA2?

Divisive Amplicon Denoising Algorithm 2
Defines exact amplicon sequence variants (ASVs or ESVs) based on error rates.



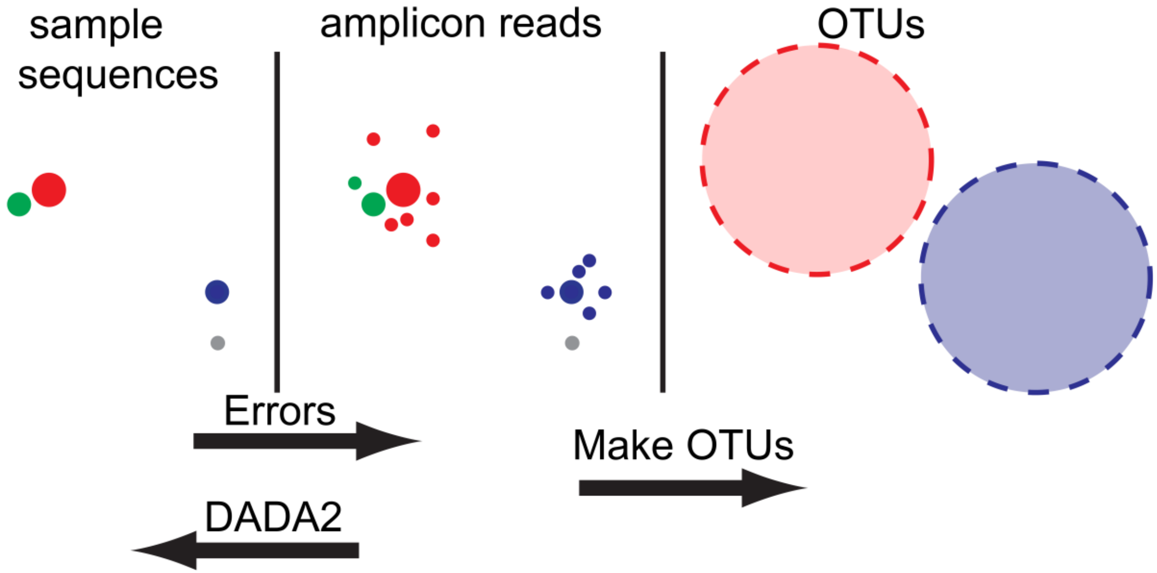
Dada2 attempts to identify errors to “denoise” the data and infer true sequences

OTU methods limit false positives, but in doing so also lump distinct taxa (=false negatives)

Callahan et al. 2016. Nature Methods 13: 581-583 https://www.nature.com/articles/nmeth.3869

Why DADA2?

Divisive Amplicon Denoising Algorithm 2
Defines exact amplicon sequence variants (ASVs or ESVs) based on error rates.



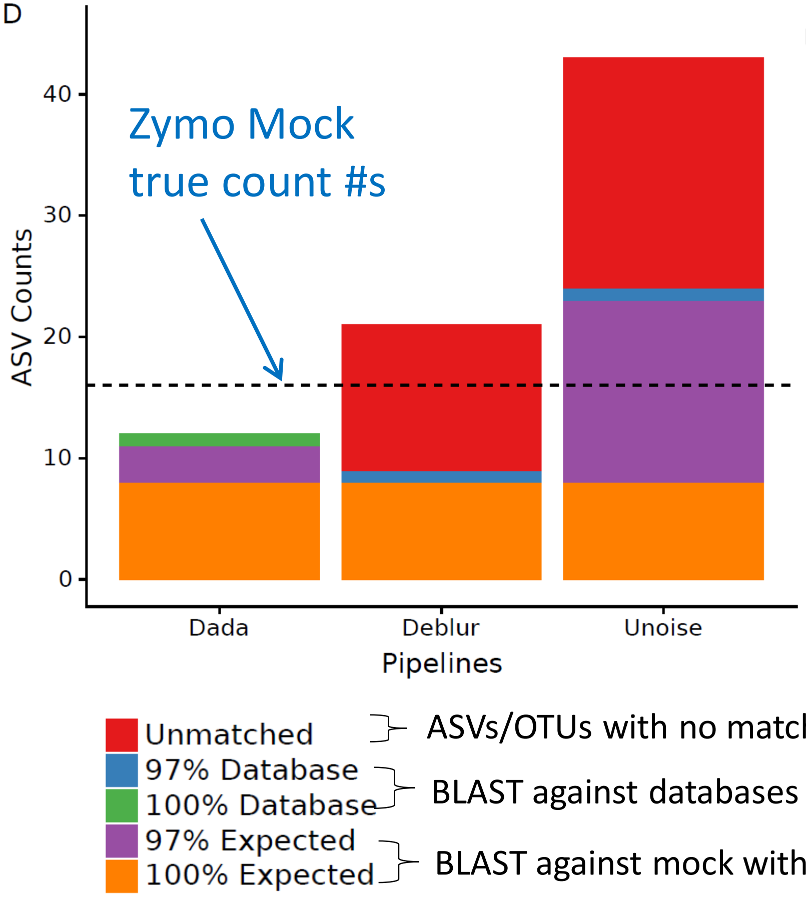
Solves most problems of OTUs: highly accurate, reproducible, comparable across studies

But very high resolution: intra-genomic/specific variation vs. interspecific resolution; still some disagreement about best methods

Also: listen to Ben Callahan talk about this https://bioinformatics.chat/amplicon-sequence-variants

Callahan et al. 2016. Nature Methods 13: 581-583 https://www.nature.com/articles/nmeth.3869

How accurate are DADA2 ASVs?



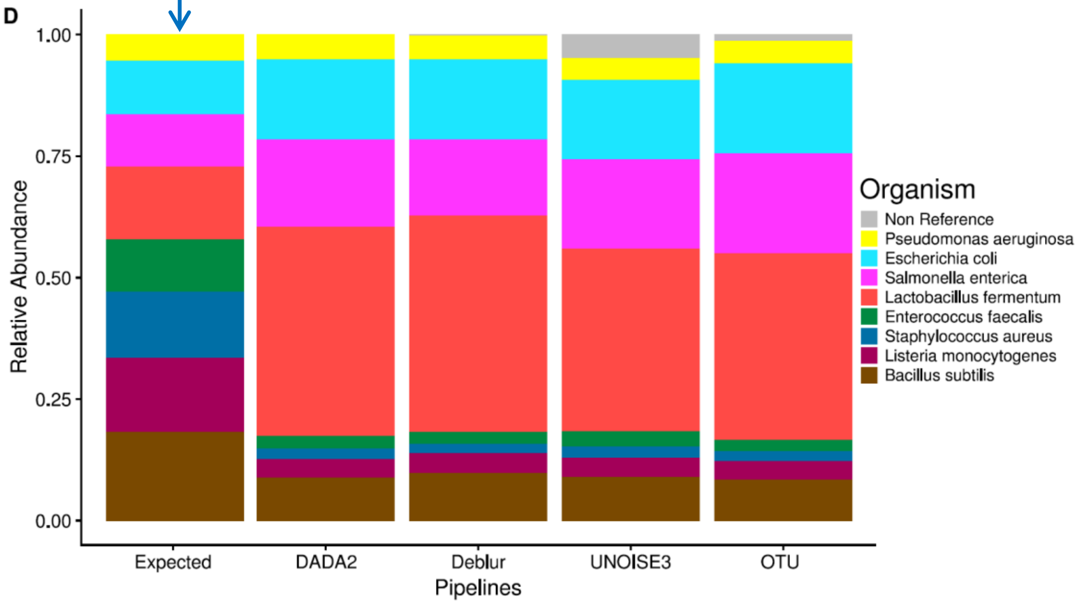
Zymo Mock true count #s

ASVs/OTUs with no matches

BLAST against databases with % cutoff

BLAST against mock with % cutoff

Zymo Mock true community

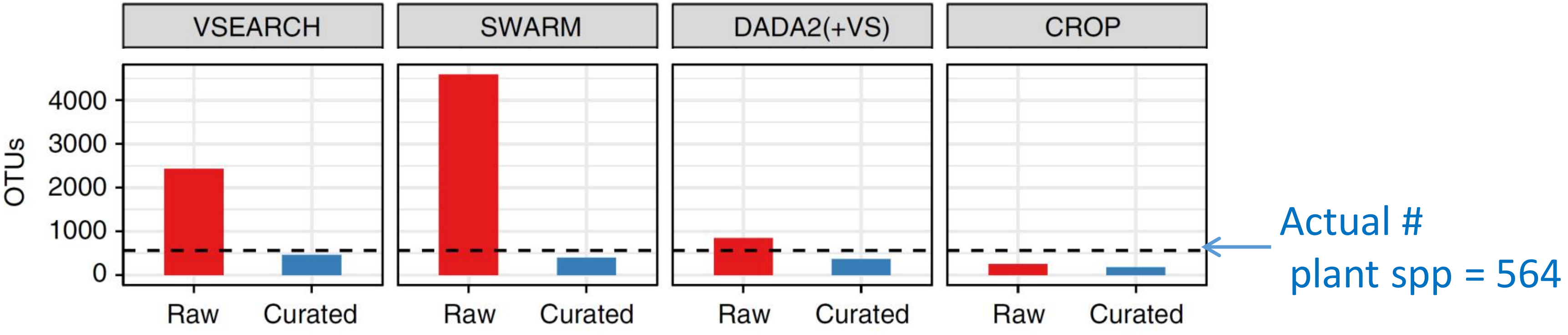


Nearing et al. 2018 Peer J 6:e5364

How accurate are DADA2 ASVs? Curation with LULU reduces artefacts

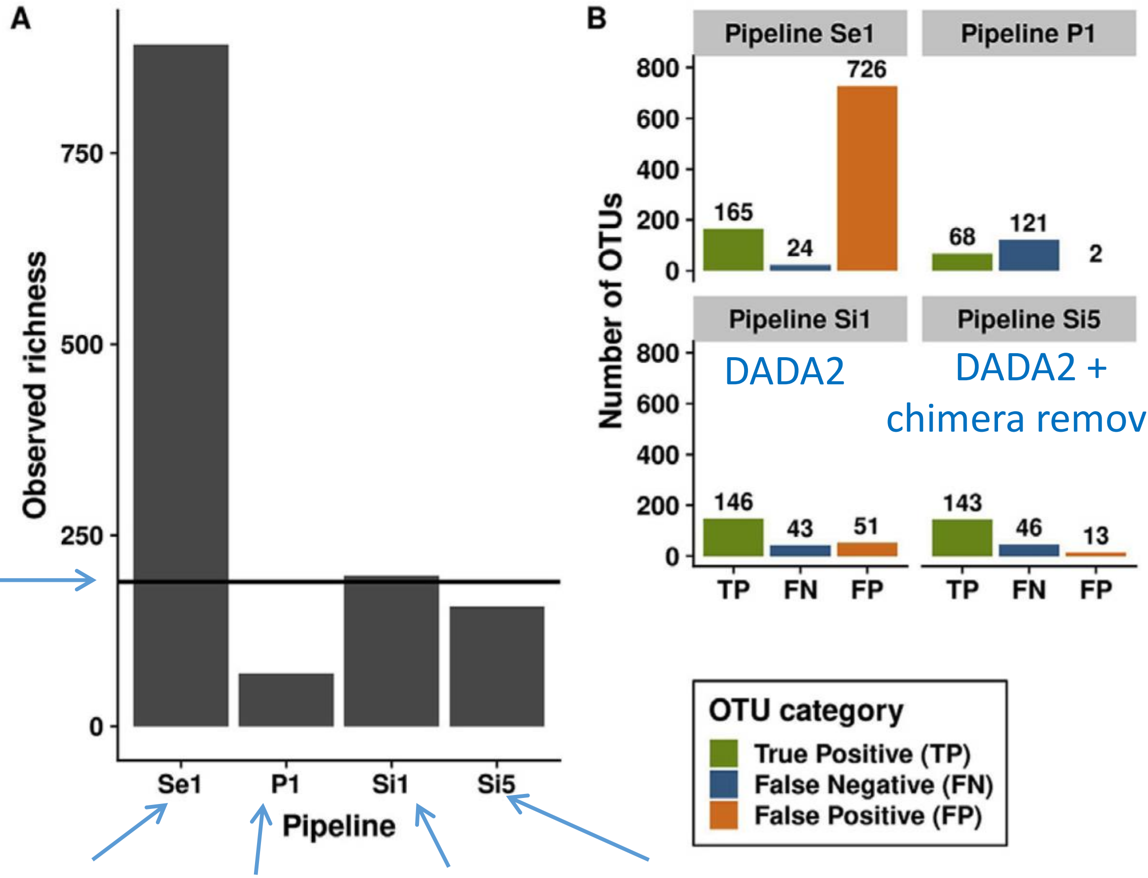
- LULU identifies and merges “daughter” with “parent” OTUs that are consistently co-occurring but more abundant

- Assumes “daughter” OTUs are artefacts



Froslev et al. 2017 Nature Comm 1118

How accurate are DADA2 ASVs? Curation with LULU reduces artefacts but is conservative



VSearch

DADA2 + LULU

DADA2

DADA2 + chimera removal

Actual # fungal taxa in mock = 189

VSearch

DADA2 + LULU

DADA2

DADA2 + chimera removal

Pauvert et al. 2019 Fung Ecol 41: 23-33

Use the DADA2 tutorial

- https://benjjneb.github.io/dada2/tutorial.html

- Note current version is 1.18, tutorial is still for 1.16
    - Mostly minor updates and bug fixes

Sequence processing steps

1. Preprocess
2. Filter and trim
3. Learn error rates
4. Sample inference
5. Merge paired-end reads
6. Create ASV table
7. Remove Chimeras
8. Assign taxonomy

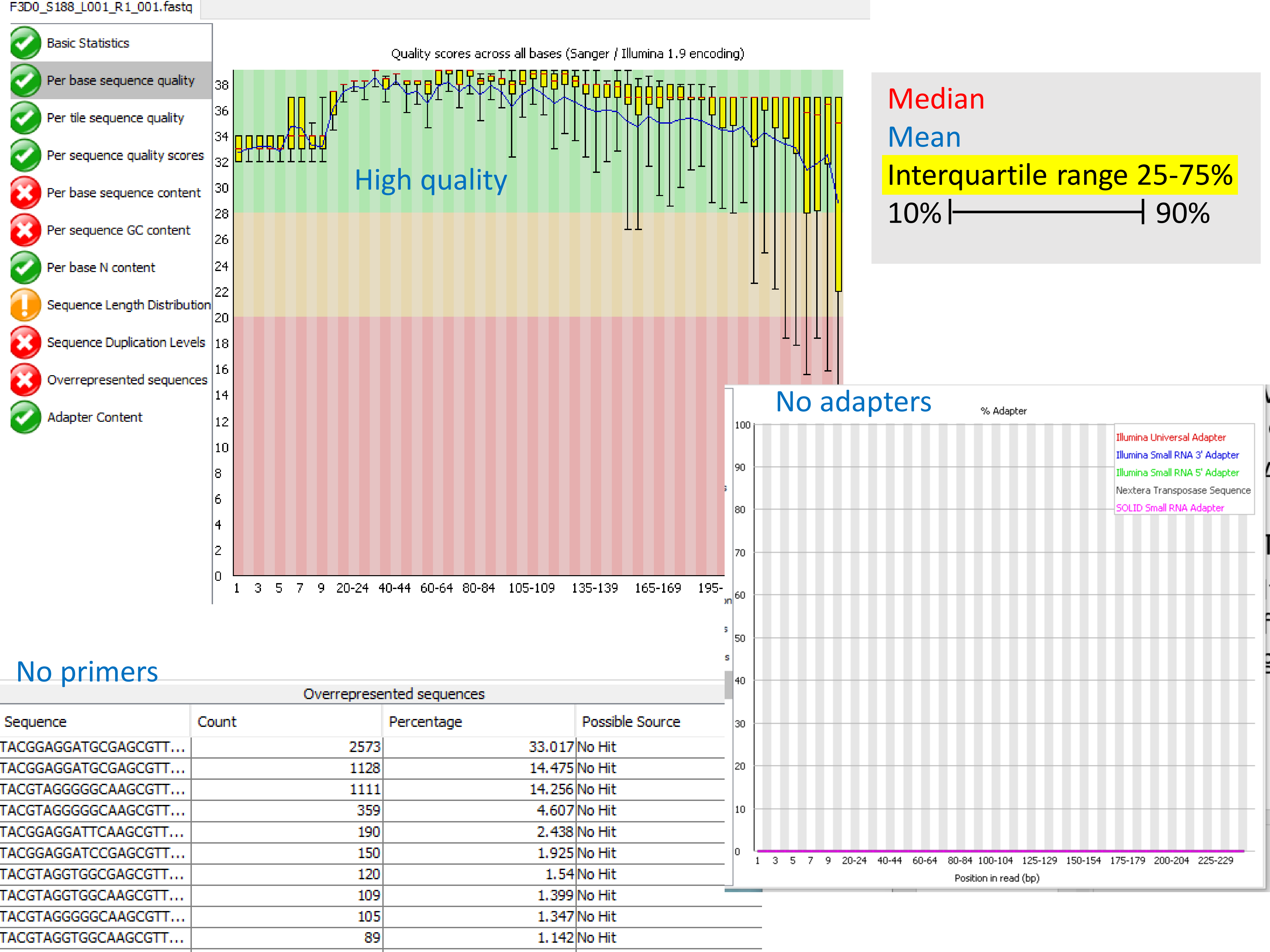
1. Preprocess (i.e., get ready!)

- Make sure current version of dada2 package is installed, then load with the library command
- Download the MiSeq_SOP sequence data from the tutorial and place in your synced project folder
- Download the taxonomy files from the tutorial and place in your synced project folder
- Check that samples are demultiplexed (split into individual per-sample fastq files)

2. Filter and Trim

- How long is your amplicon? Does length vary?
- What are your amplicon primer sequences and how long are they? Are they included in the data?
- Are Illumina adapters or indices still attached?

- Today's tutorial data have no primers/adapters
- If the primer/adapter information isn't available from your core facility, use FASTQC to check (also gives you a good idea of quality)
  - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/


F3D0_S188_L001_R1_001.fastq
✓ Basic Statistics
✓ Per base sequence quality
✓ Per tile sequence quality
✓ Per sequence quality scores
☒ Per base sequence content
☒ Per sequence GC content
✓ Per base N content
! Sequence Length Distribution
☒ Sequence Duplication Levels
☒ Overrepresented sequences
✓ Adapter Content
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
3
5
7
9
20-24
40-44
60-64
80-84
105-109
135-139
165-169
195-
% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
100
90
80
70
60
50
40
30
20
10
0
1
3
5
7
9
20-24
40-44
60-64
80-84
100-104
125-129
150-154
175-179
200-204
225-229
Position in read (bp)

High quality

Median

Mean

Interquartile range 25-75%

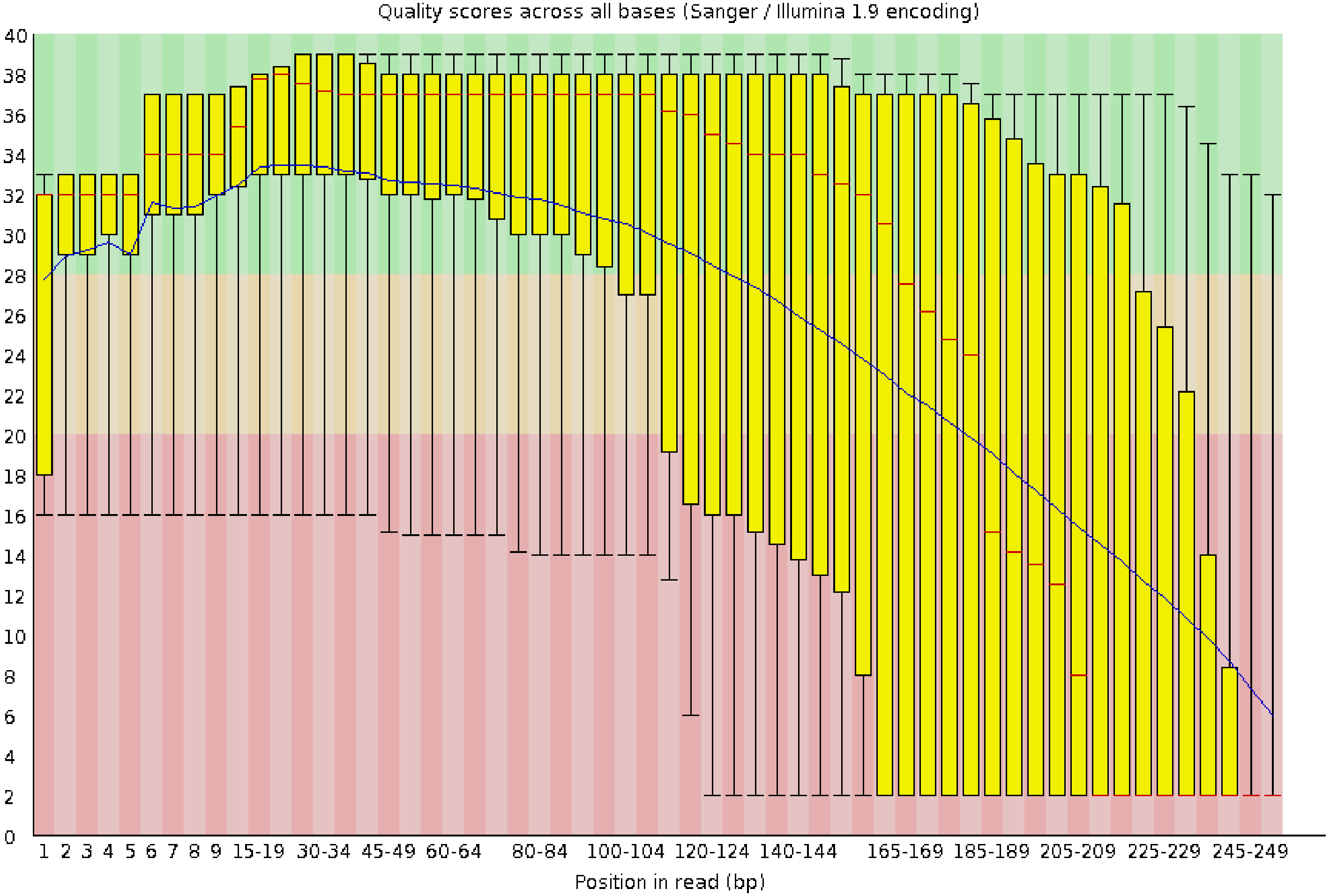
10% ├──────┤ 90%

No adapters

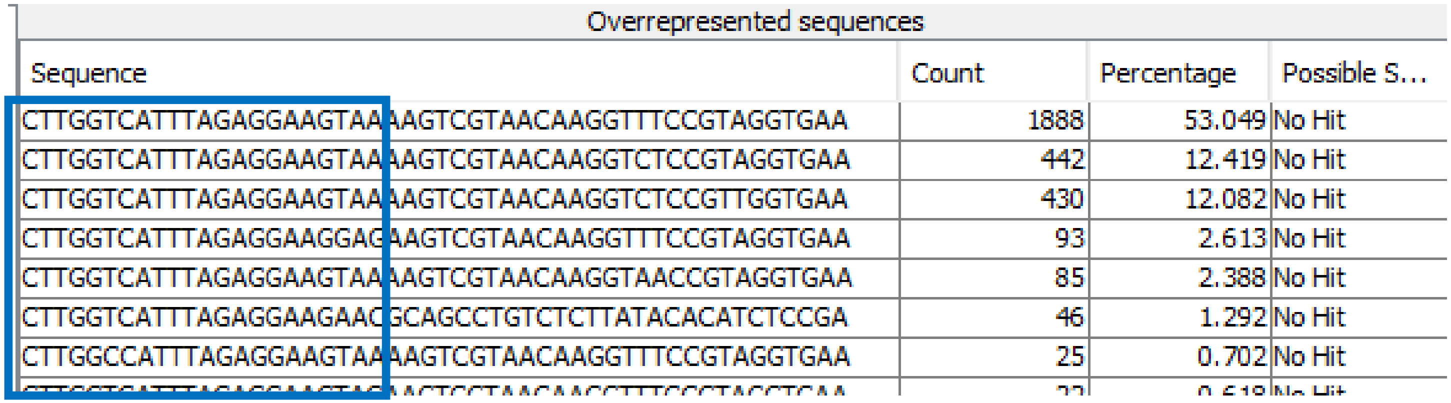
No primers

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TACGGAGGATGCGAGCGTT...	2573	33.017	No Hit
TACGGAGGATGCGAGCGTT...	1128	14.475	No Hit
TACGTAGGGGGCAAGCGTT...	1111	14.256	No Hit
TACGTAGGGGGCAAGCGTT...	359	4.607	No Hit
TACGGAGGATTCAAGCGTT...	190	2.438	No Hit
TACGGAGGATCCGAGCGTT...	150	1.925	No Hit
TACGTAGGTGGCGAGCGTT...	120	1.54	No Hit
TACGTAGGTGGCAAGCGTT...	109	1.399	No Hit
TACGTAGGGGGCAAGCGTT...	105	1.347	No Hit
TACGTAGGTGGCAAGCGTT...	89	1.142	No Hit



Low quality



Overrepresented sequences

Sequence	Count	Percentage	Possible S...
CTTGGTCATTTAGAGGAAGTAAAAGTCGTAACAAGGTTTCCGTAGGTGAA	1888	53.049	No Hit
CTTGGTCATTTAGAGGAAGTAAAAGTCGTAACAAGGTCTCCGTAGGTGAA	442	12.419	No Hit
CTTGGTCATTTAGAGGAAGTAAAAGTCGTAACAAGGTCTCCGTTGGTGAA	430	12.082	No Hit
CTTGGTCATTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAA	93	2.613	No Hit
CTTGGTCATTTAGAGGAAGTAAAAGTCGTAACAAGGTAACCGTAGGTGAA	85	2.388	No Hit
CTTGGTCATTTAGAGGAAGAACGCAGCCTGTCTCTTATACACATCTCCGA	46	1.292	No Hit
CTTGGCCATTTAGAGGAAGTAAAAGTCGTAACAAGGTTTCCGTAGGTGAA	25	0.702	No Hit
[illegible]	22	0.618	No Hit

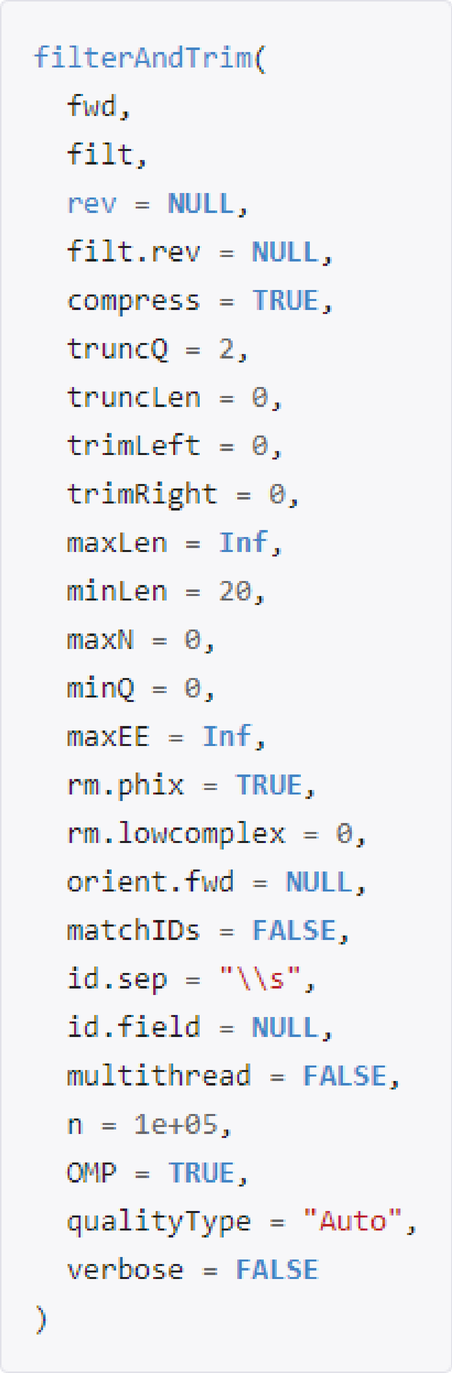
First 22bp are the the primer (need to trim)

2. Filter and Trim

- If primers are at the start of reads and constant length, remove in dada2 with the “trimLeft” function
- If you have variable length sequences, use fastqcleaner, BBDuk, trimmomatic, or cutadapt to remove primers/adapters

2. Filter and Trim

- Can set many parameters for filtering in dada2
- Today we will focus on
  - Trimming primers (trimLeft)
  - Trimming for quality (truncLen)
  - Filtering for quality (maxEE)

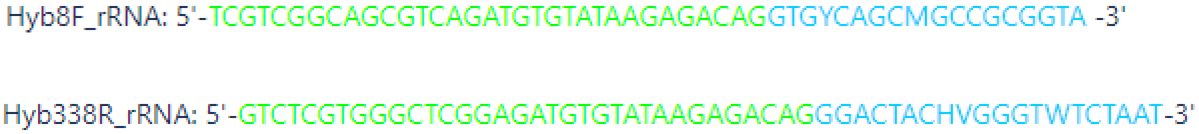


```
filterAndTrim(
  fwd,
  filt,
  rev = NULL,
  filt.rev = NULL,
  compress = TRUE,
  truncQ = 2,
  truncLen = 0,
  trimLeft = 0,
  trimRight = 0,
  maxLen = Inf,
  minLen = 20,
  maxN = 0,
  minQ = 0,
  maxEE = Inf,
  rm.phix = TRUE,
  rm.lowcomplex = 0,
  orient.fwd = NULL,
  matchIDs = FALSE,
  id.sep = "\\s",
  id.field = NULL,
  multithread = FALSE,
  n = 1e+05,
  OMP = TRUE,
  qualityType = "Auto",
  verbose = FALSE
)
```

https://rdrr.io/bioc/dada2/man/filterAndTrim.html

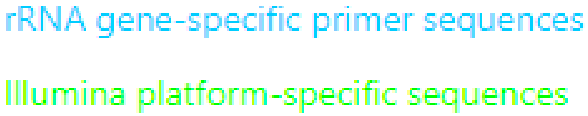
2. Filter and Trim:
Typical Illumina amplicon sequencing

First round: Illumina adapters + amplicon PCR primers



Hyb8F_rRNA: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGYCAGCMGCCGCGGTA -3'

Hyb338R_rRNA: 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT-3'



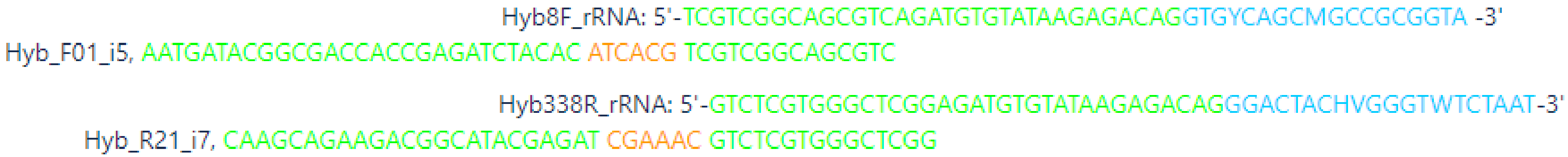
rRNA gene-specific primer sequences

Illumina platform-specific sequences

https://wikis.utexas.edu/display/GSAF/rRNA+bacterial+gene+and+fungal+ITS+metagenomics+samples

2. Filter and Trim:
Typical Illumina amplicon sequencing

First round: Illumina adapters + amplicon PCR primers



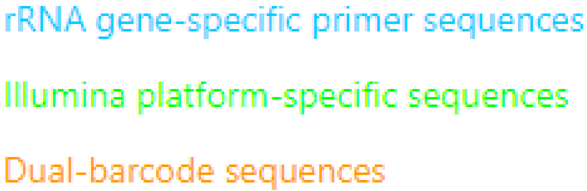
Hyb8F_rRNA: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGYCAGCMGCCGCGGTA -3'

Hyb_F01_i5, AATGATACGGCGACCACCGAGATCTACAC ATCACG TCGTCGGCAGCGTC

Hyb338R_rRNA: 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGACTACHVGGGTWTCTAAT-3'

Hyb_R21_i7, CAAGCAGAAGACGGCATACGAGAT CGAAAC GTCTCGTGGGCTCGG

Second round: Illumina PCR primers + barcode indices



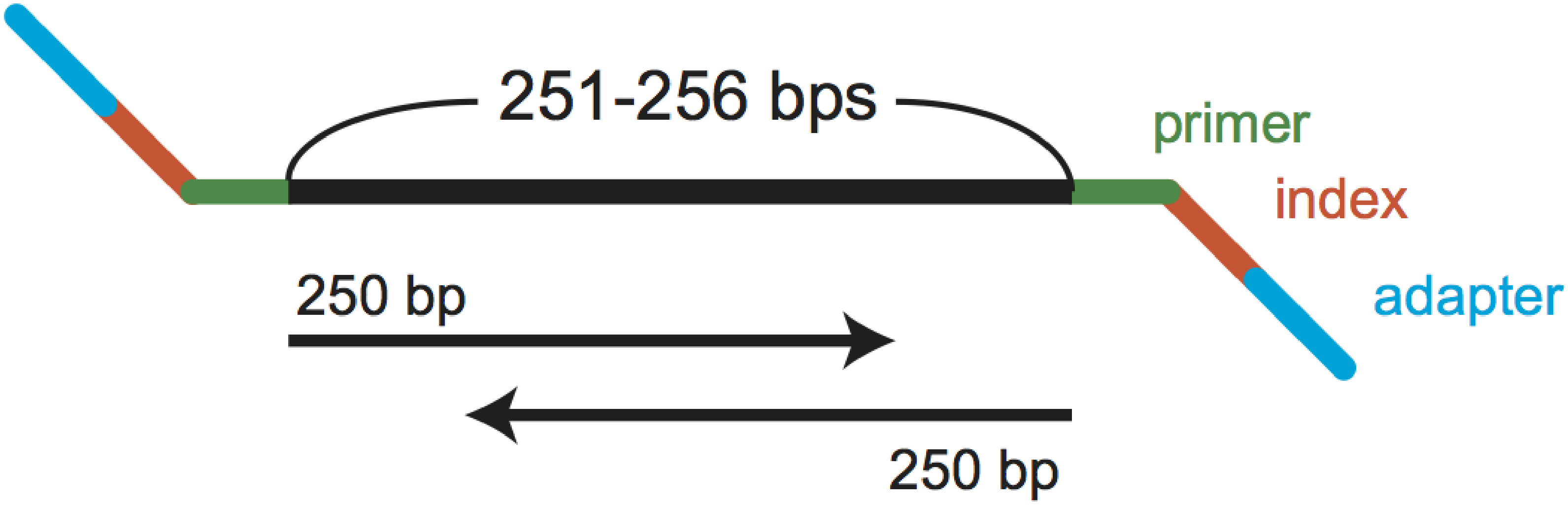
rRNA gene-specific primer sequences

Illumina platform-specific sequences

Dual-barcode sequences

https://wikis.utexas.edu/display/GSAF/rRNA+bacterial+gene+and+fungal+ITS+metagenomics+samples

2. Filter and Trim: Today's Data V4 region 2x250 PE



There are no primers in today's data set, but if you need to trim primers, use trimLeft to indicate the number of nucleotides (based on primer length) to remove from start of each read

```
#single reads
filterAndTrim(…, trimLeft=FWD_PRIMER_LENGTH, REV_PRIMER_LENGTH)

#paired-end reads
filterAndTrim(…, trimLeft=c(FWD_PRIMER_LENGTH, REV_PRIMER_LENGTH))
```

2. Filter and Trim: Trim on Illumina Quality Scores

Phred Quality Score Q = -10 $\log_{10}$ P	P = Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

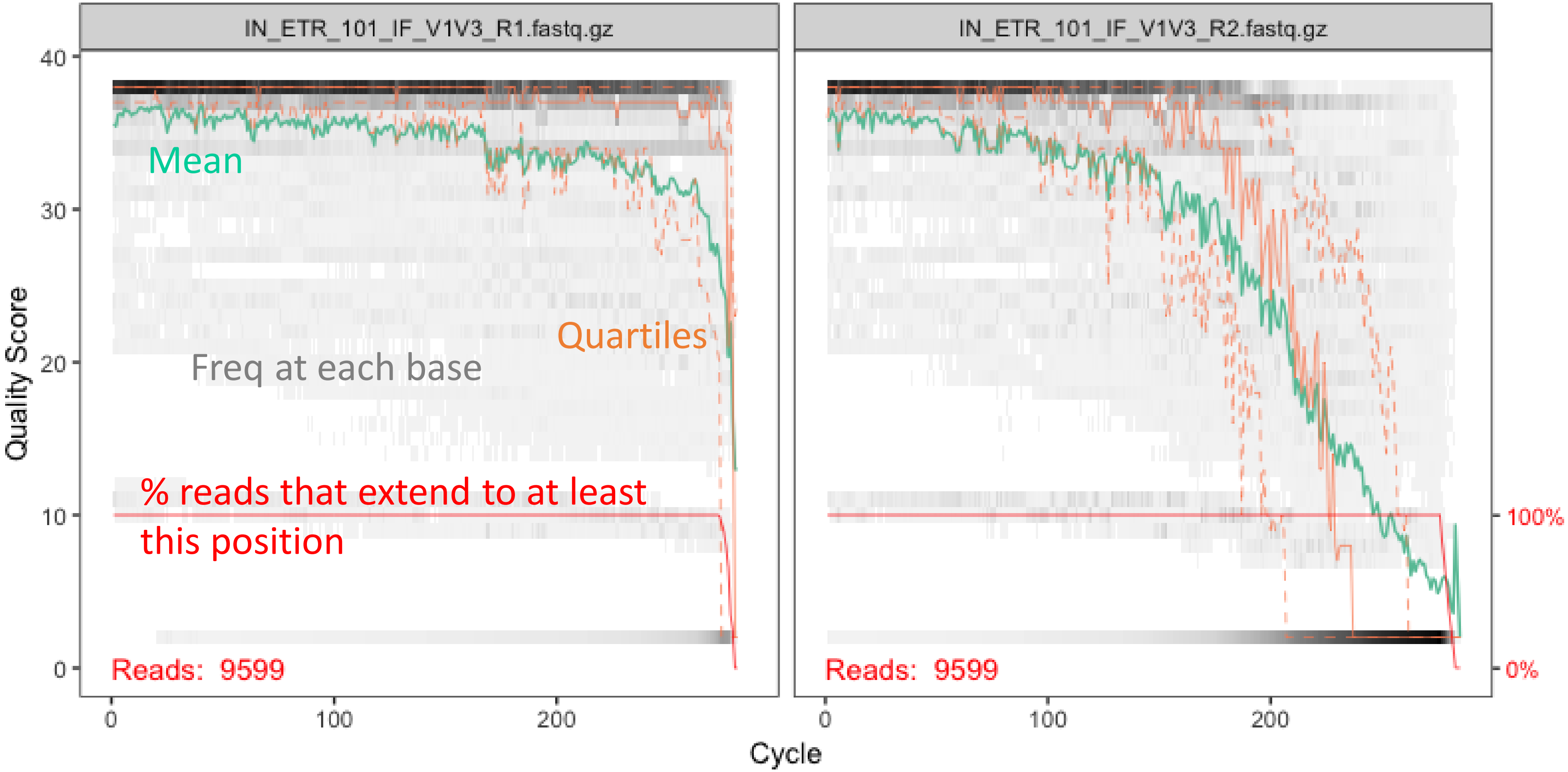
What cutoff would you use?

truncLen command to truncate to specified lengths
(default is 0, assign based on quality scores)
filterAndTrim(…, truncLen=0)
BUT DO NOT USE FOR ITS, where read lengths vary

Note: if both trimLeft and truncLen are used, filtered reads will have
length = truncLen - trimLeft

2. Filter and Trim: Trim on Illumina Quality Scores

2x300, Amplicon length 400-420 nucleotides, primers sequenced

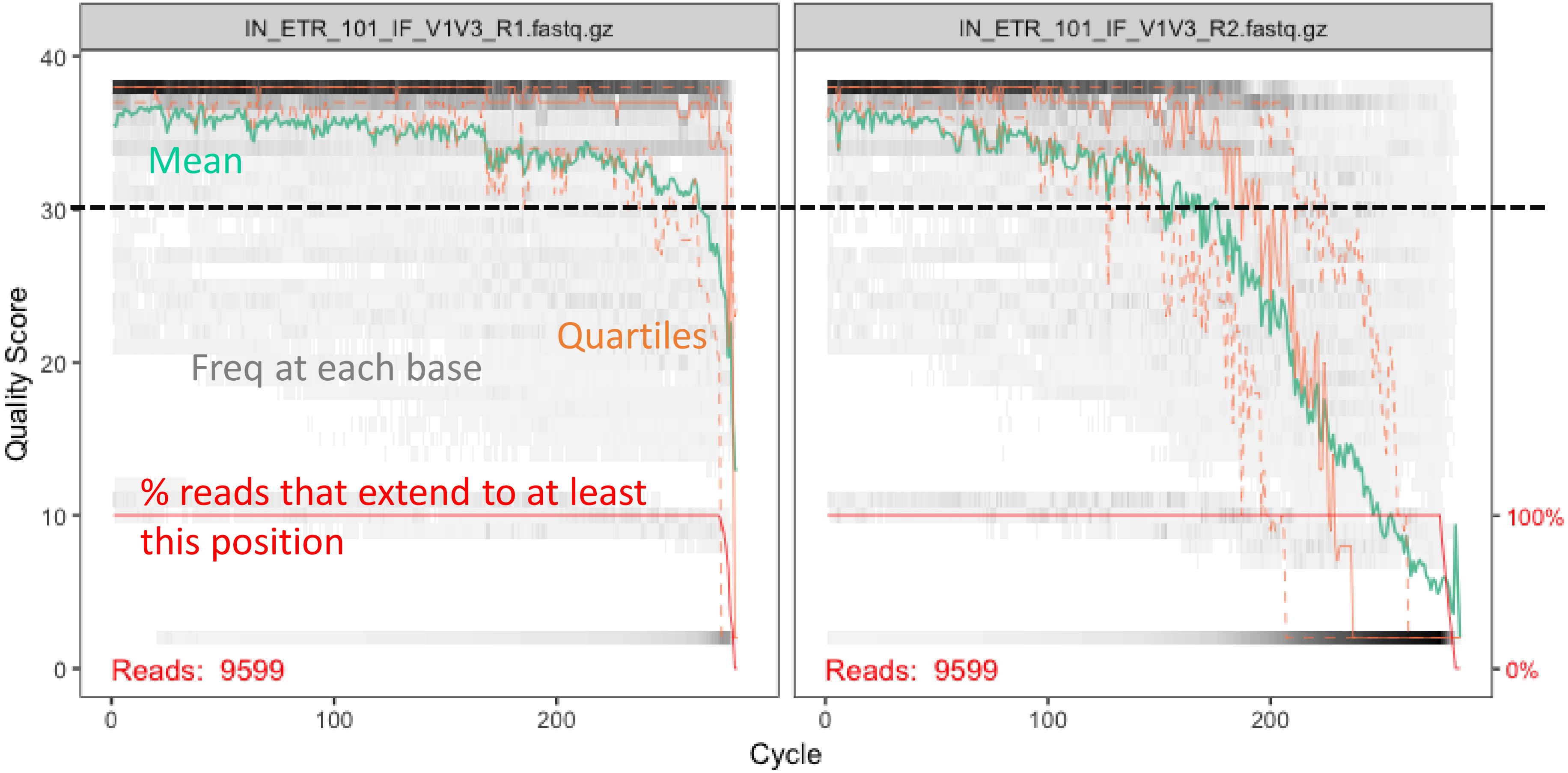


Strategy: maintain >20+ nucleotide overlap, truncate where quality crashes

Dada2 uses quality info as well, but trimming helps detect rare seq variants

2. Filter and Trim: Trim on Illumina Quality Scores

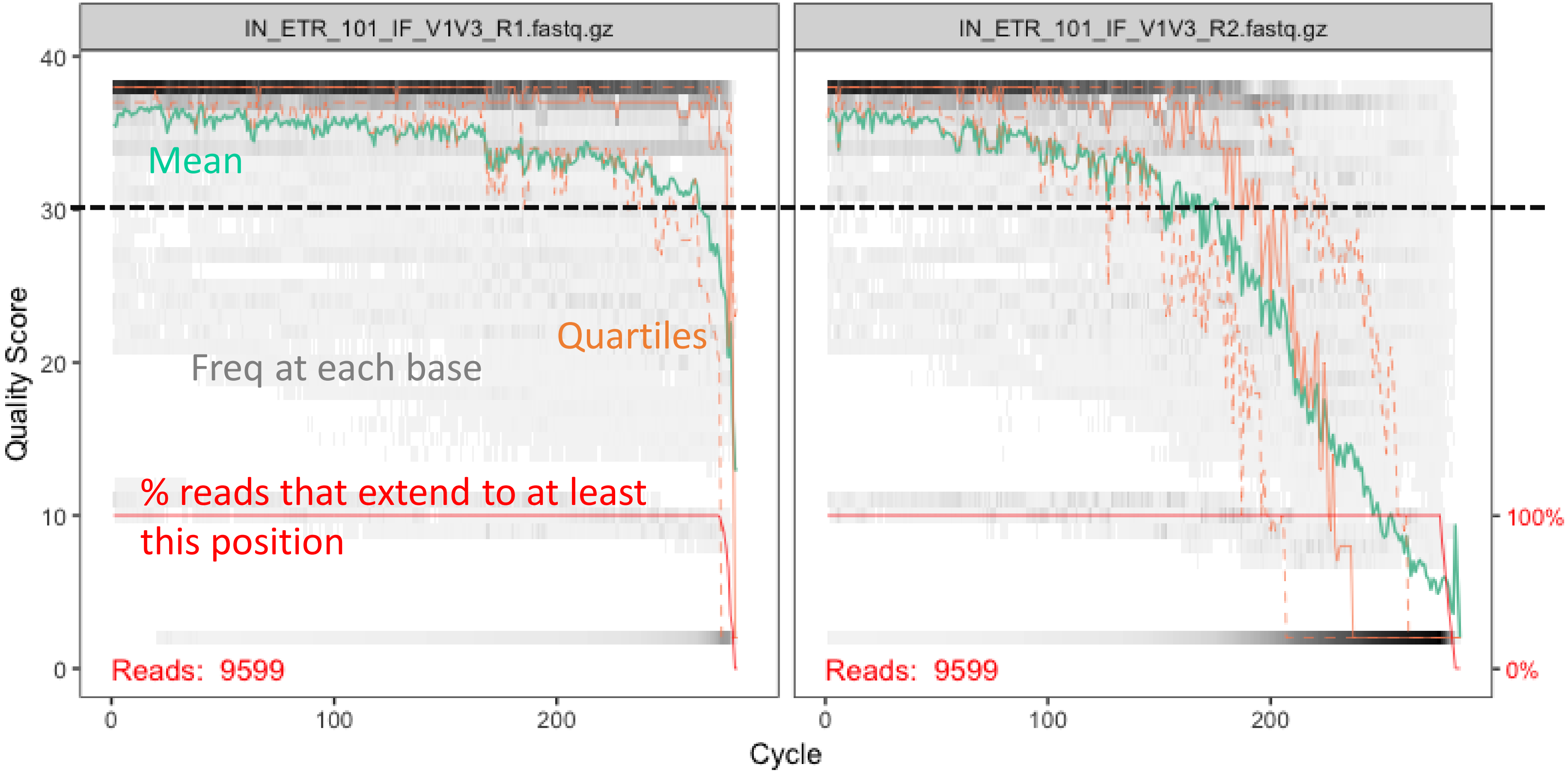
2x300, Amplicon length 400-420 nucleotides, primers sequenced (17, 21)



Strategy: maintain >20+ nucleotide overlap, truncate where quality crashes
Dada2 uses quality info as well, but trimming helps detect rare seq variants

2. Filter and Trim: Trim on Illumina Quality Scores

2x300, Amplicon length 400-420 nucleotides, primers sequenced (17, 21)

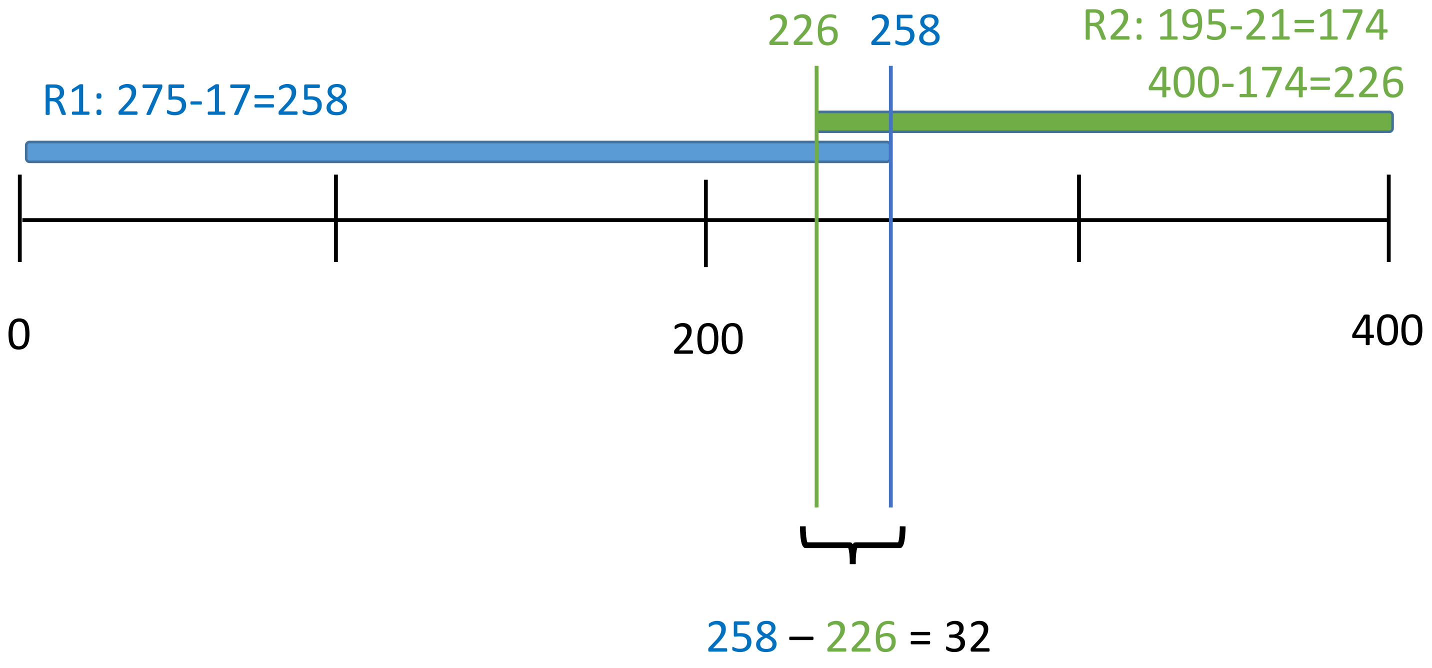


trimLeft = c(17, 21)
truncLen=c(275, 195) ←

Overlap is 12-32b

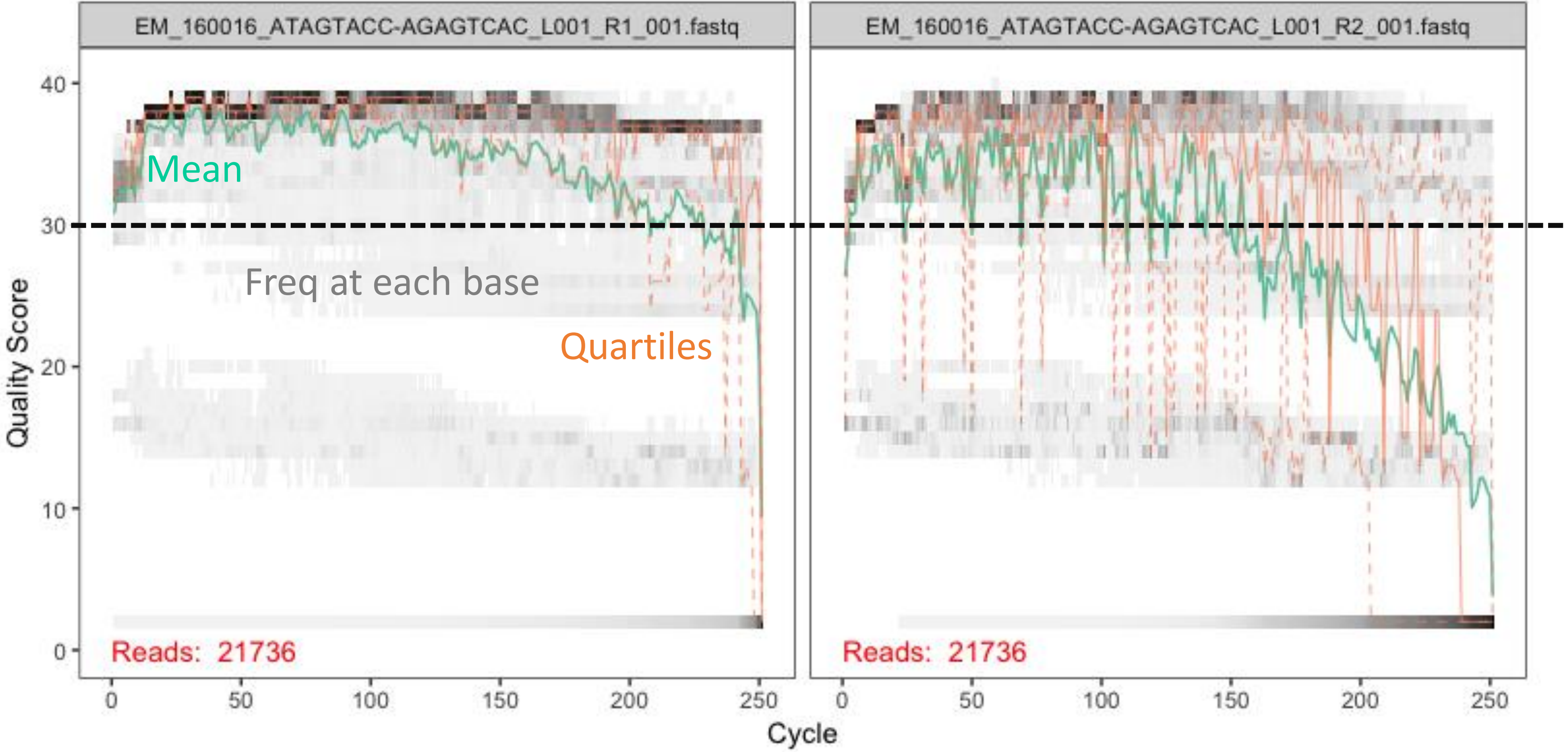
2. Filter and Trim: calculating overlap

- 400bp amplicon
- trimLeft = c(17, 21)
- truncLen=c(275, 195)



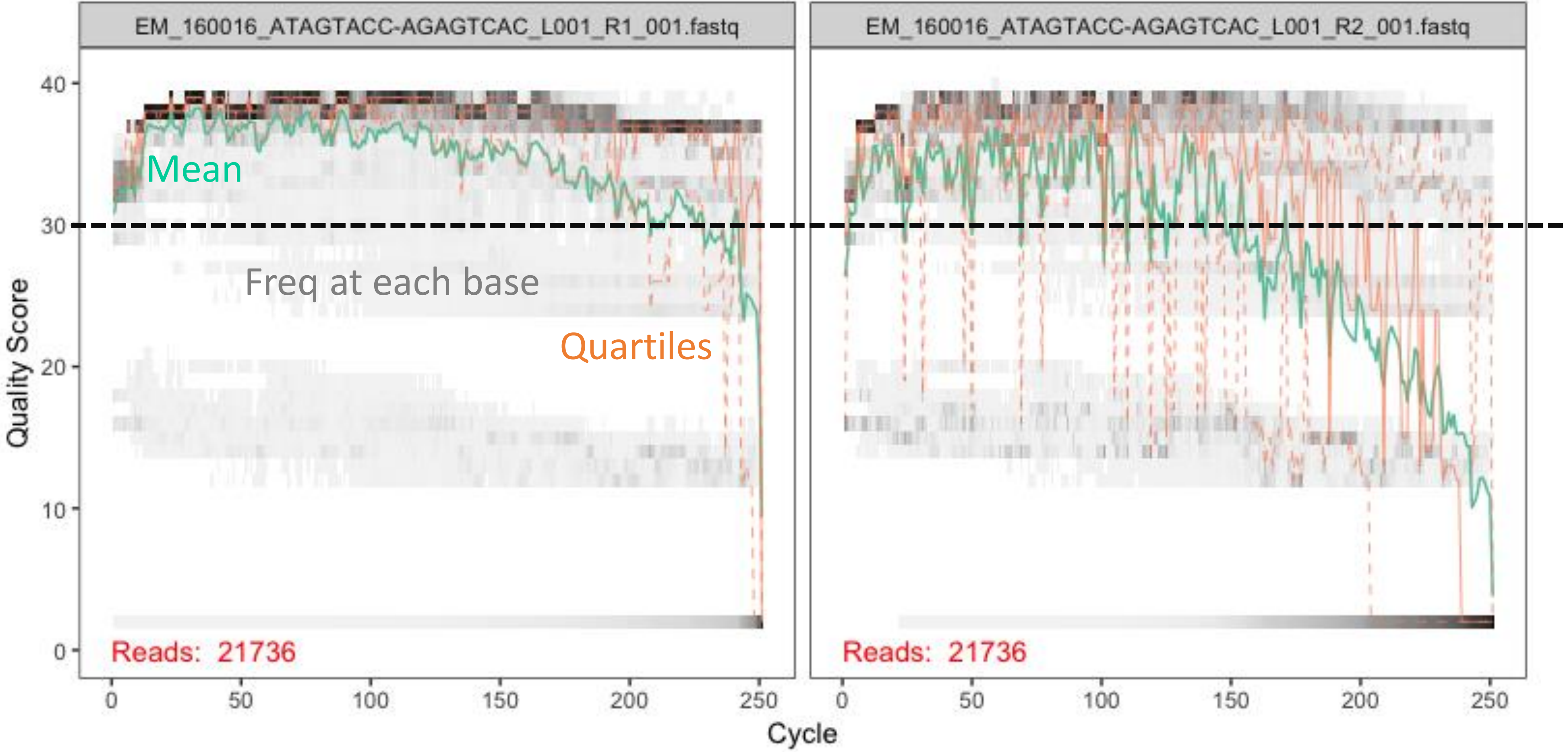
2. Filter and Trim:
Trim on Illumina Quality Scores

2x250, Amplicon length 250-260 nucleotides, primers sequenced (14, 17)



2. Filter and Trim: Trim on Illumina Quality Scores

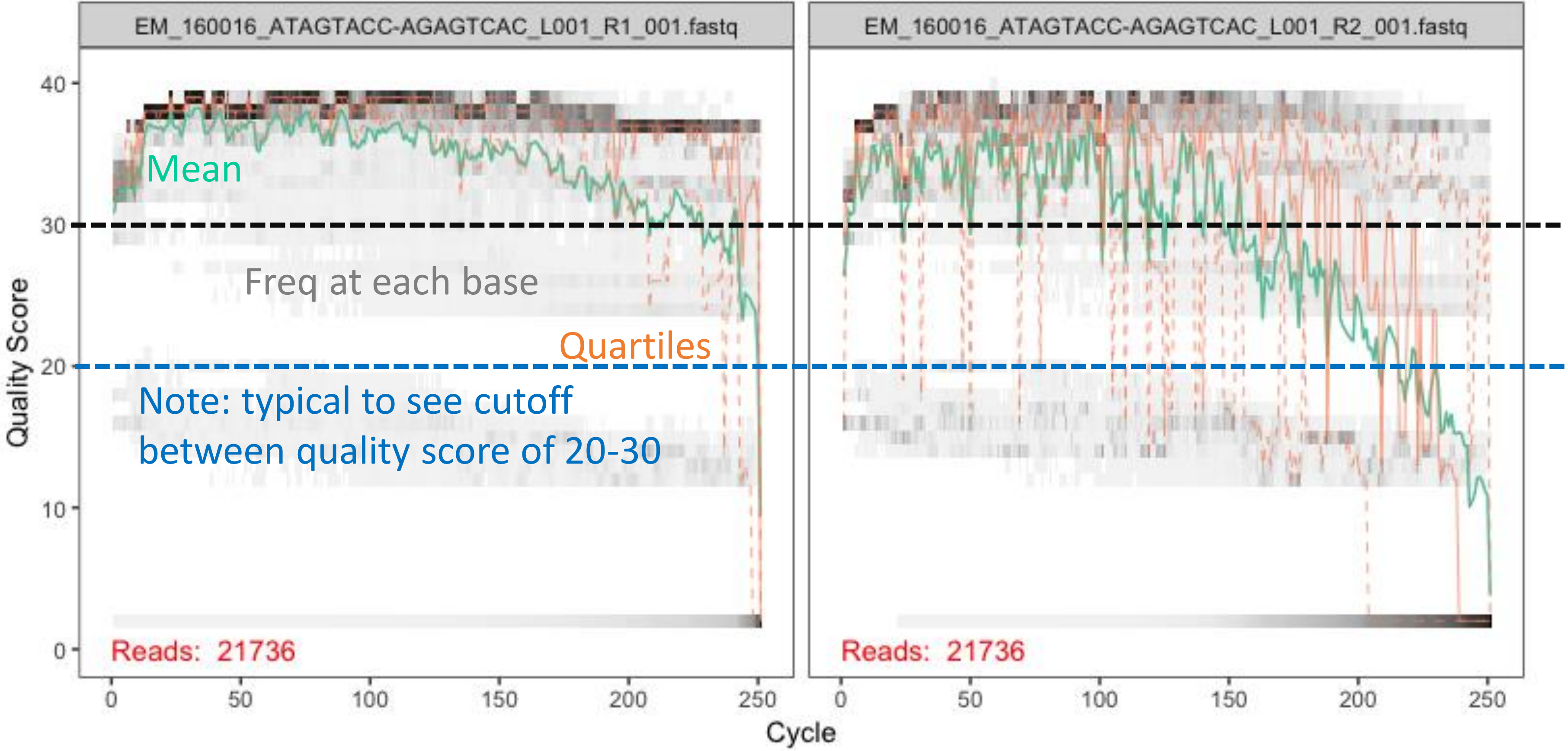
2x250, Amplicon length 250-260 nucleotides, primers sequenced (14, 17)



trimLeft = c(14, 17)
truncLen=c(220, 140) ← Overlap is 79-89b

2. Filter and Trim: Trim on Illumina Quality Scores

2x250, Amplicon length 250-260 nucleotides, primers sequenced (14, 17)



Note: typical to see cutoff between quality score of 20-30

```
trimLeft = c(14, 17)
truncLen=c(220, 140)
```

← Overlap is 79-89b

2. Filter and Trim: Quality Filtering Options

- maxEE: Max # of expected errors, usually this is the only quality filter needed
  - default is 2,2
  - Start with default, adjust if needed
    - Too few reads – consider relaxing reverse to 5
    - Too many reads – constrain further to 0 or 1
  - Can also try to optimize manually or with Zymo's FIGARO tool https://github.com/Zymo-Research/figaro#figaro

```
filterAndTrim(…, maxEE=c(2,2))
```

https://rdrr.io/bioc/dada2/man/filterAndTrim.html

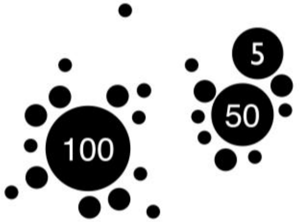
2. Filter and Trim: Quality Filtering

- Run the tutorial filter and trim steps
- Rerun the quality plots post-filtering if you want to see how they’ve changed

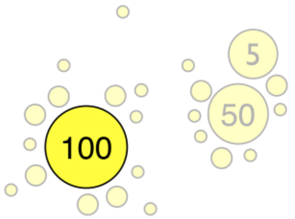
3. Learn Error Rates

- Parametric error model estimated from data
- Alternates estimation of error rates and inference of sample composition until they converge on a single consistent solutions
- Algorithm begins with an initial “guess” with max error (assumes only the most abundant sequence is correct and all others are errors)

3. Learn Error Rates



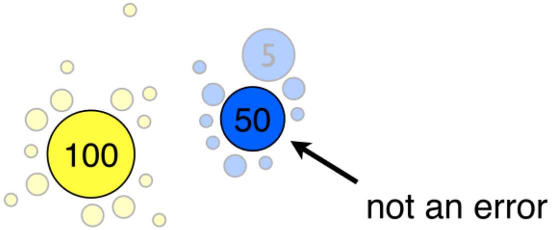
Initial guess: one real sequence + errors



Infer initial *error model* under this assumption.

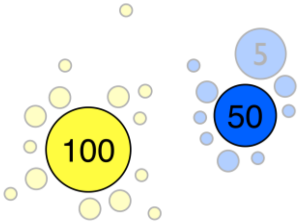
$\Pr(i \to j) =$

	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



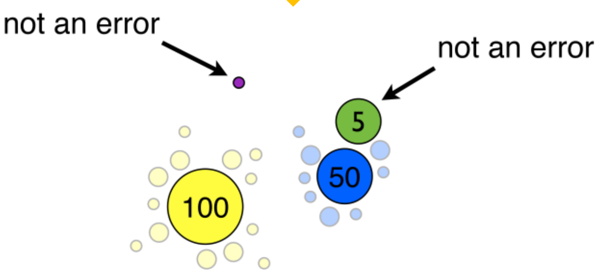
Reject unlikely error under model. **Recruit** errors.

	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



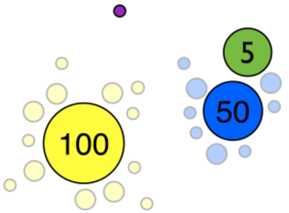
Update the model.

	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



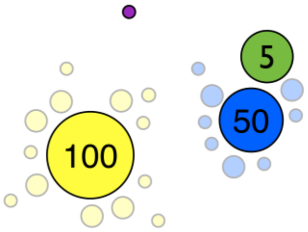
Reject more sequences under ***new*** model

	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



Update model again

	A	C	G	T
A	0.998	$1\text{x}10^{-4}$	$2\text{x}10^{-3}$	$2\text{x}10^{-4}$
C	$6\text{x}10^{-5}$	0.999	$3\text{x}10^{-6}$	$1\text{x}10^{-3}$
G	$1\text{x}10^{-3}$	$3\text{x}10^{-6}$	0.999	$6\text{x}10^{-5}$
T	$2\text{x}10^{-4}$	$2\text{x}10^{-3}$	$1\text{x}10^{-4}$	0.998



Convergence: all errors are plausible

	A	C	G	T
A	0.998	$1\text{x}10^{-4}$	$2\text{x}10^{-3}$	$2\text{x}10^{-4}$
C	$6\text{x}10^{-5}$	0.999	$3\text{x}10^{-6}$	$1\text{x}10^{-3}$
G	$1\text{x}10^{-3}$	$3\text{x}10^{-6}$	0.999	$6\text{x}10^{-5}$
T	$2\text{x}10^{-4}$	$2\text{x}10^{-3}$	$1\text{x}10^{-4}$	0.998

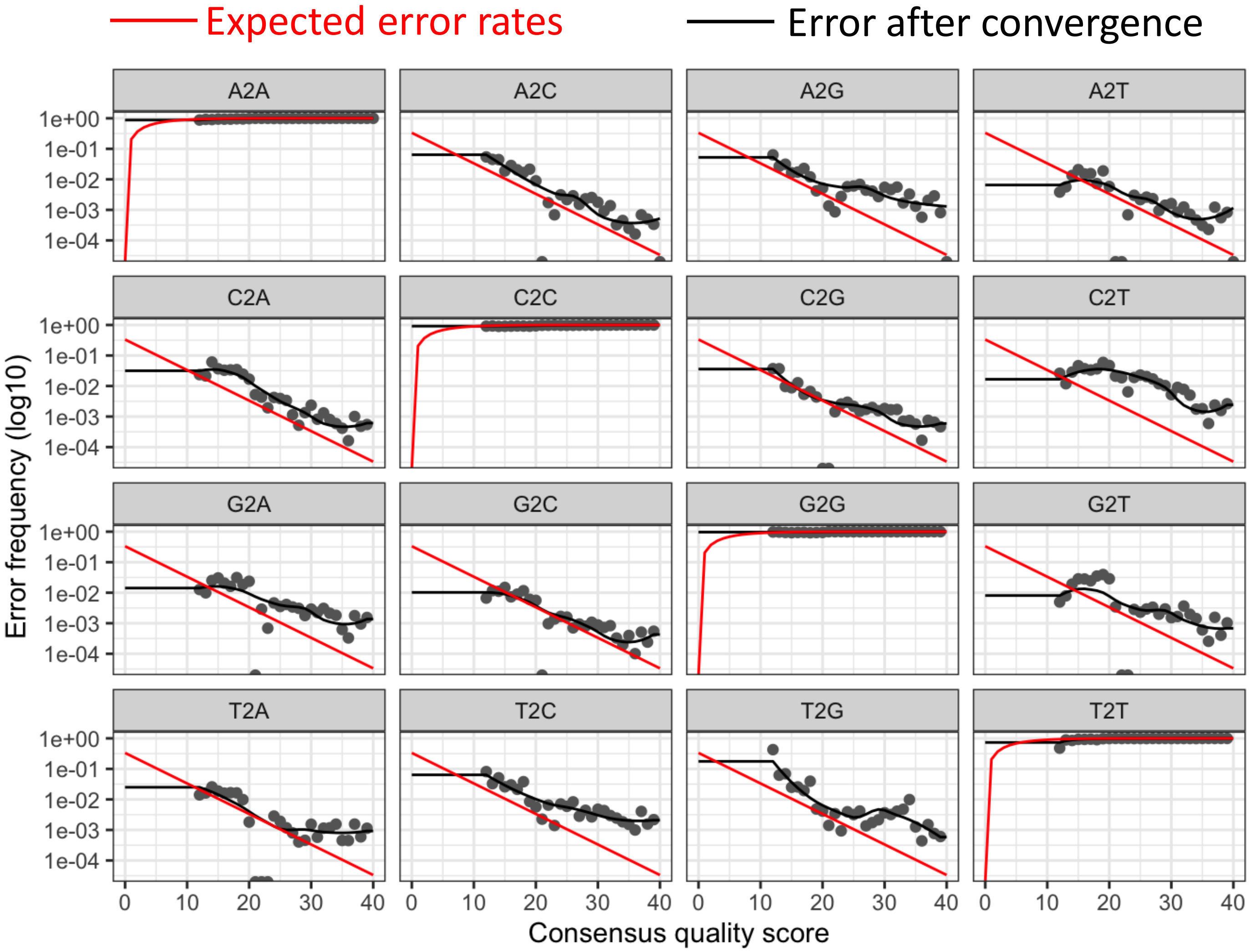
3. Learn Error Rates: Visualize

Error rates should:

-decrease with quality score

-have a good fit between model (black line) and actual observations (black points)

-goal is “good” not “perfect”



3. Learn Error Rates

- Run this portion of the tutorial
- Visualize and assess the error rates/models

4. Sample Inference

- Disentangle errors from true genetic variation to select ASVs
- Uses the parametric error model + clustering/partitioning algorithm
  - Partition reads into clusters consistent with error model
  - Overly abundant sequences are centers for initial clusters and all seqs are compared to the center
  - Iterative partitioning, clustering, and shuffling takes place until most likely partitioning is found
  - Final set of partitions is taken to represent the denoised composition of the sample

Callahan et al. 2016. Nature Methods 13: 581-583 https://www.nature.com/articles/nmeth.3869#methods

4. Sample Inference

- Run the sample inference portion of the tutorial

Callahan et al. 2016. Nature Methods 13: 581-583 https://www.nature.com/articles/nmeth.3869#methods

5. Merge Paired Reads

- Merge F and R reads together to get the full denoised sequences
- Pairs with insufficient overlap will not merge

- Run tutorial

6. Construct the ASV table

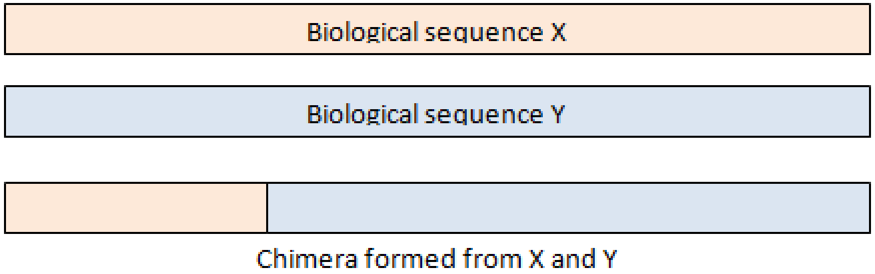
- Produce a table of samples by ASVs - this is the foundation for further analysis
- Run the tutorial

	ATACGGTT	CCTGTTAA	GAGTCCAT	…
Sample1	321	204	0	
Sample2	44	0	19	
Sample3	0	83	0	
…				

7. Remove chimeras

- DADA2 only addresses indel and substitution errors
- Chimeric sequences are those where two or more independent sequences join together
- Commonly occurs during amplicon sequencing when there are closely related sequences
- Removal ID's chimeras that consist of paired segments matching more abundant "parents"
- Run the tutorial

Note: if you lose a large proportion of sequences to chimeras, go back and check that your primers were correctly removed (esp. those with ambiguous nucleotides)



7.5 Track reads through the entire DADA2 pipeline

- This is important information that you will report in manuscript methods
- But also allows you to check how many sequences were lost at each stage.

- Run the tutorial

8. Assign taxonomy

- Multiple databases are available

Database	rRNA Regions	Web
SILVA	16S, 23S, 18S, 28S	https://www.arb-silva.de/
RDP	16S, ITS, 28S	https://rdp.cme.msu.edu/
UNITE	ITS	https://unite.ut.ee/
Greengenes (2013)	16S	http://greengenes.lbl.gov/

- Database annotations do have errors and conflicts (https://peerj.com/articles/5030/)
- What to do if you have many novel sequences that are unclassified?

- Run the tutorial

Homework: lulu curation of ASVs

- If you haven’t already, install local blast following NCBI instructions

  - Windows https://www.ncbi.nlm.nih.gov/books/NBK52637/#!po=8.33333

  - Mac https://www.ncbi.nlm.nih.gov/books/NBK279671/?report=reader

Homework: Run lulu curation

- Read: Froslev et al. 2017 Nature Communications 8:1188
  - https://www.nature.com/articles/s41467-017-01312-x
- Follow lulu tutorial instructions but use our wk2 data files
  - https://github.com/tobiasgf/lulu#tutorial
- **Submit on Github by Jan 26**
  - **Raw .Rmd file for DADA2 script**
  - **Knitted .Rmd file for LULU**