From .fastq to OTU: an overview

GC

```
TTCGAGGACTACTGGGGTATCTAATCCCGTTCGCTCCCCTAGCTTTCGC
CTTGATATCTACGCATTTCACCGCTACACCAAGAATTCCAGTTGCCCCT
GATTTCACAAGTGACTTGTCGCGCCGCCTACGCGCCCTTTACGCCCAG

FF:F,FF,FFF::FFFFF:F:FFFFF:FFFF:F,FF,FFFFFFFFFFF:
FFFFFFFFFFF:F:FFFFF,F:FFFF:FF:::F,,FF:FFFFFFF:FFF
F:FFF,:F,FFFF:FFF:FFF,:,FFF:FFFF:FFFFFFFF:,F,F:F
1010:95:HLWN2DRXX:1:2101:1090:1000 2:N:0:1
ATACCAGGACTACAGGGGTTTCTAATCCCGTTCGCTCCCCTGGCTTTCG
TCTCGATATCTACGCATTTCACCGCTACACCGAGAATTCCAGTAGCCCC
AGATTTCACAAGTGACTTGTTGCGCCGCCTACGCGCCCTTTACGCCCA

FF,F:FFF,FF:FFFFFF:FFFFFFFFFFFFFFFFF:F:FFFFFFFFF
F,FFFFFFF:FFFFF:FFFFFFFFFFFF:FFFFF,FFFFF:FFFFFFFF
FFFFF:FFFFFFFFFFF,FFF,FFFFF:FFFFFFFF,FFFF::FFF,
1010:95:HLWN2DRXX:1:2101:1127:1000 2:N:0:1
TTGGAGATGCGTTCTTCATCTATTCCAGAAGAAAGAGATCC
TTTGAAGTGTCGACCCGTTGCACACGACCGAAGGGAGGGTC
GCGTTGTTGTATTGTTCACTCCGCAGGTTCTCCTACGGTTA
```

	B1	B2	B3	B4	B5
AGTTGA	6	0	13	10	15
GTGGAC	4	3	7	2	15
CTGTAA	3	3	0	10	12

Many pipelines exist

- QIIME
- QIIME2 (uses DADA2)
- Usearch
- Vsearch
- DADA2
- PEAR
- And many more

Overview of steps with in DADA2 pipeline

Library prep and sequencing

Demultiplexing

Check for remaining primers

Filter and Trim: fastqFilter() or fastqPairedFilter()

Infer sample composition: dada()

Merge paired reads: mergePairs()

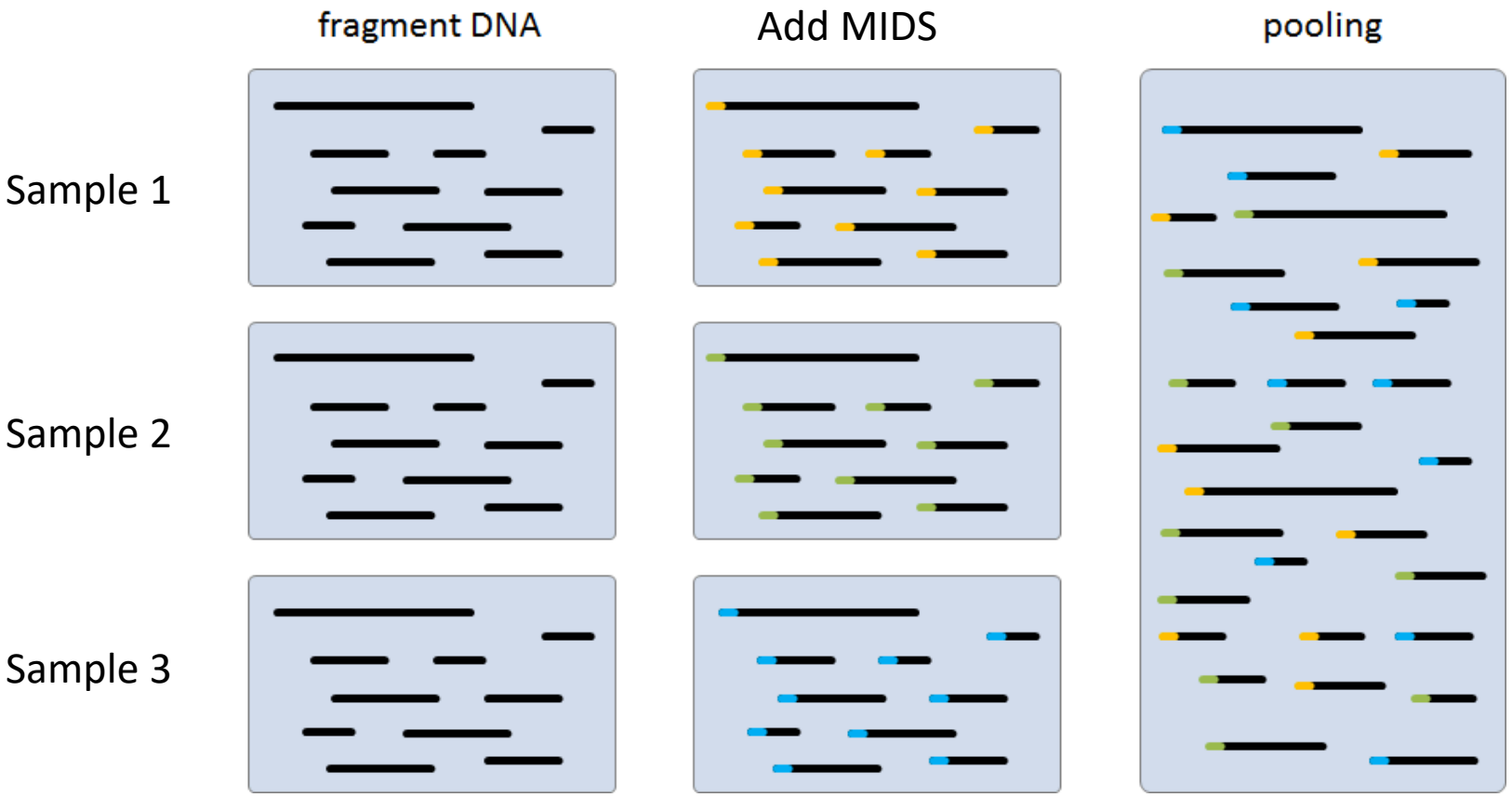
Dereplicate: derepFastq()

Make sequence table: makeSequenceTable()

Remove chimeras: isBimeraDenovo() or removeBimeraDenovo()

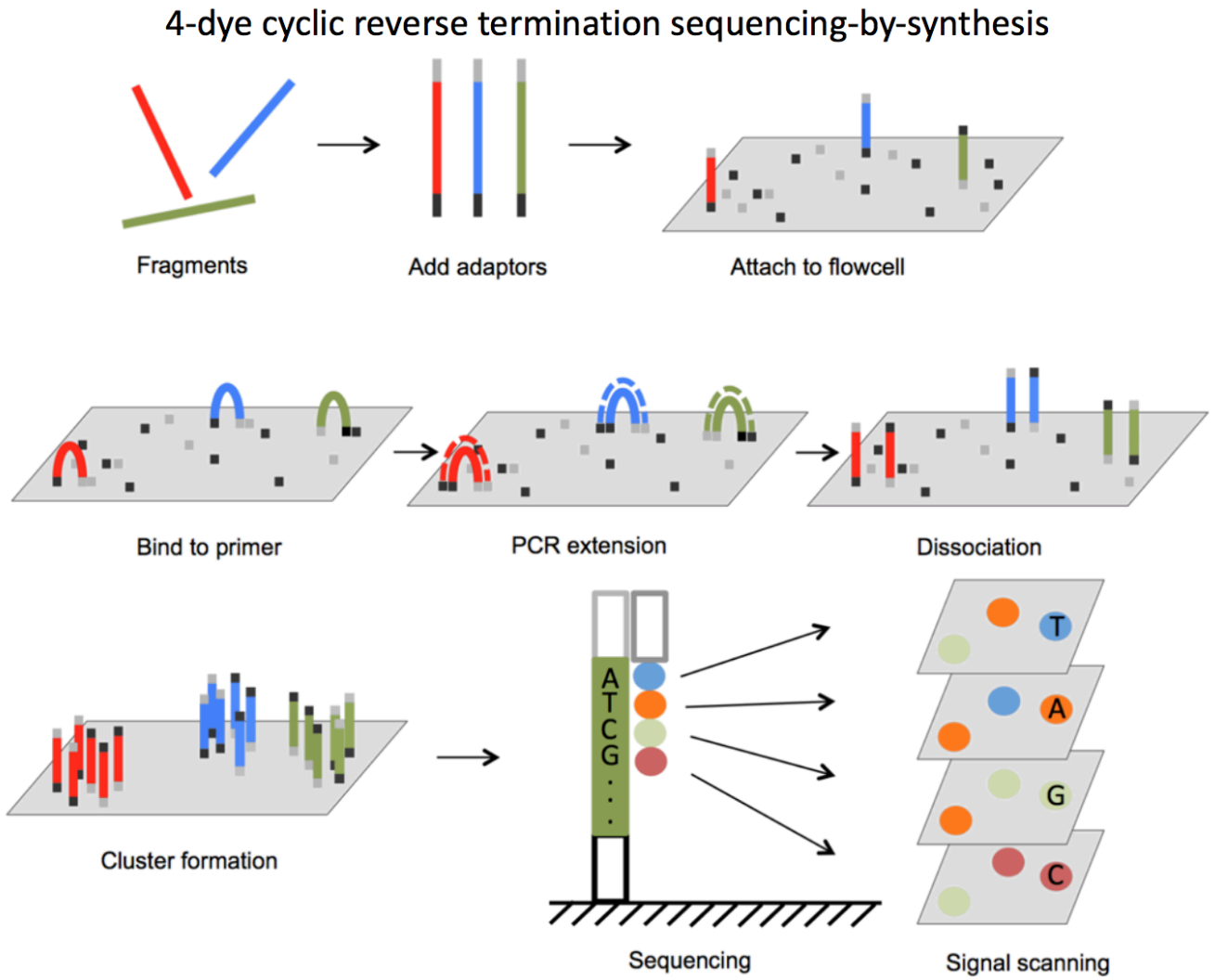
Assign taxonomy

Library Prep

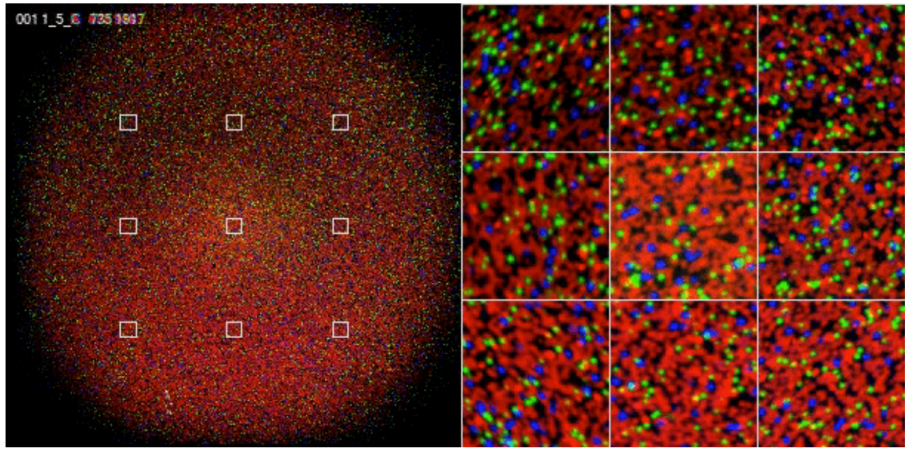


Sequencing

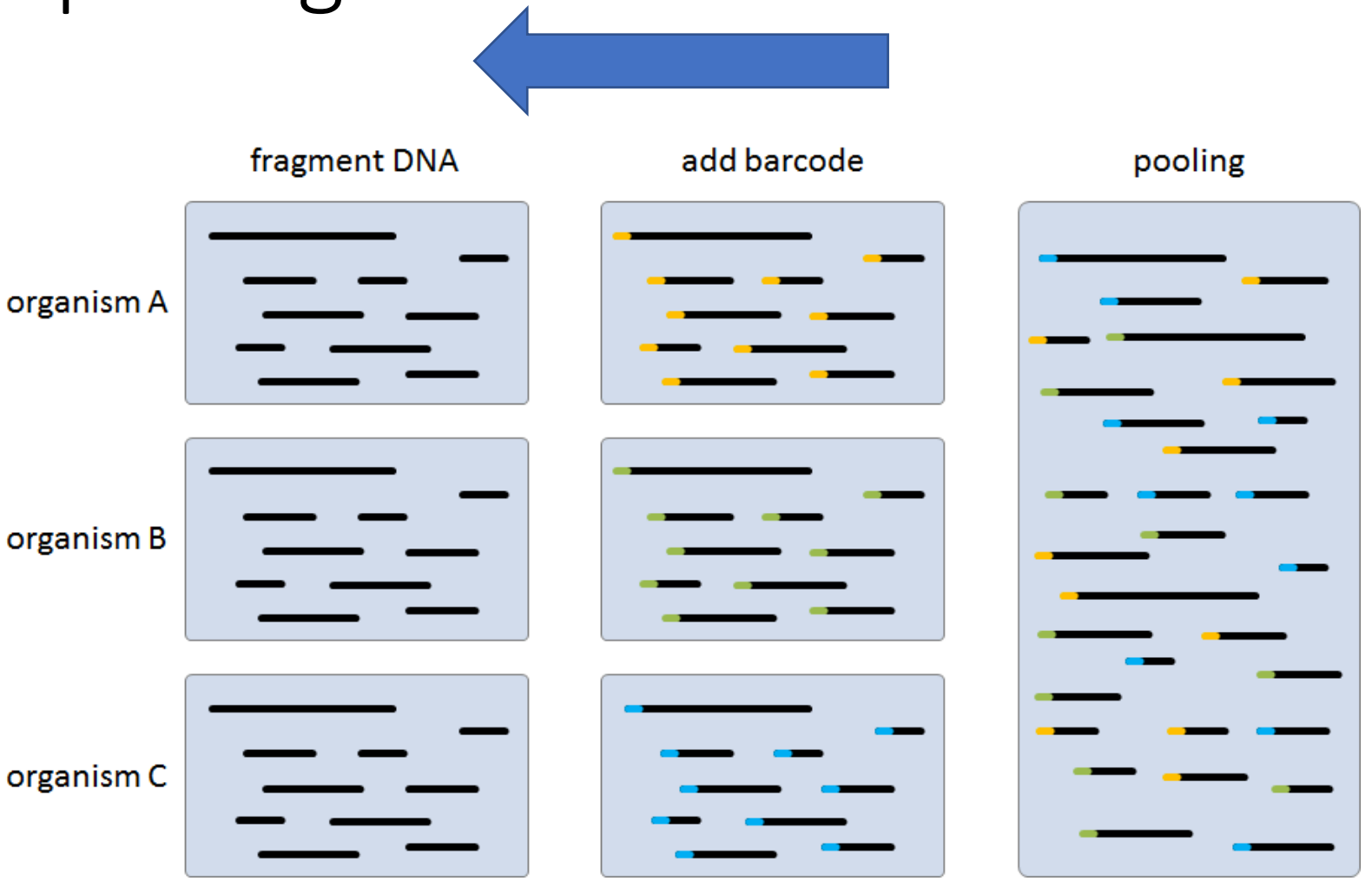
Illumina sequencing: how it works



Amplicon sequencing run on Illumina MiSeq



Demultiplexing



Removing non-biological nucleotides

- If not removed, non-biological nucleotides can cause problems with OTU picking
- Non-specific binding of primers can mask actual sequence of data

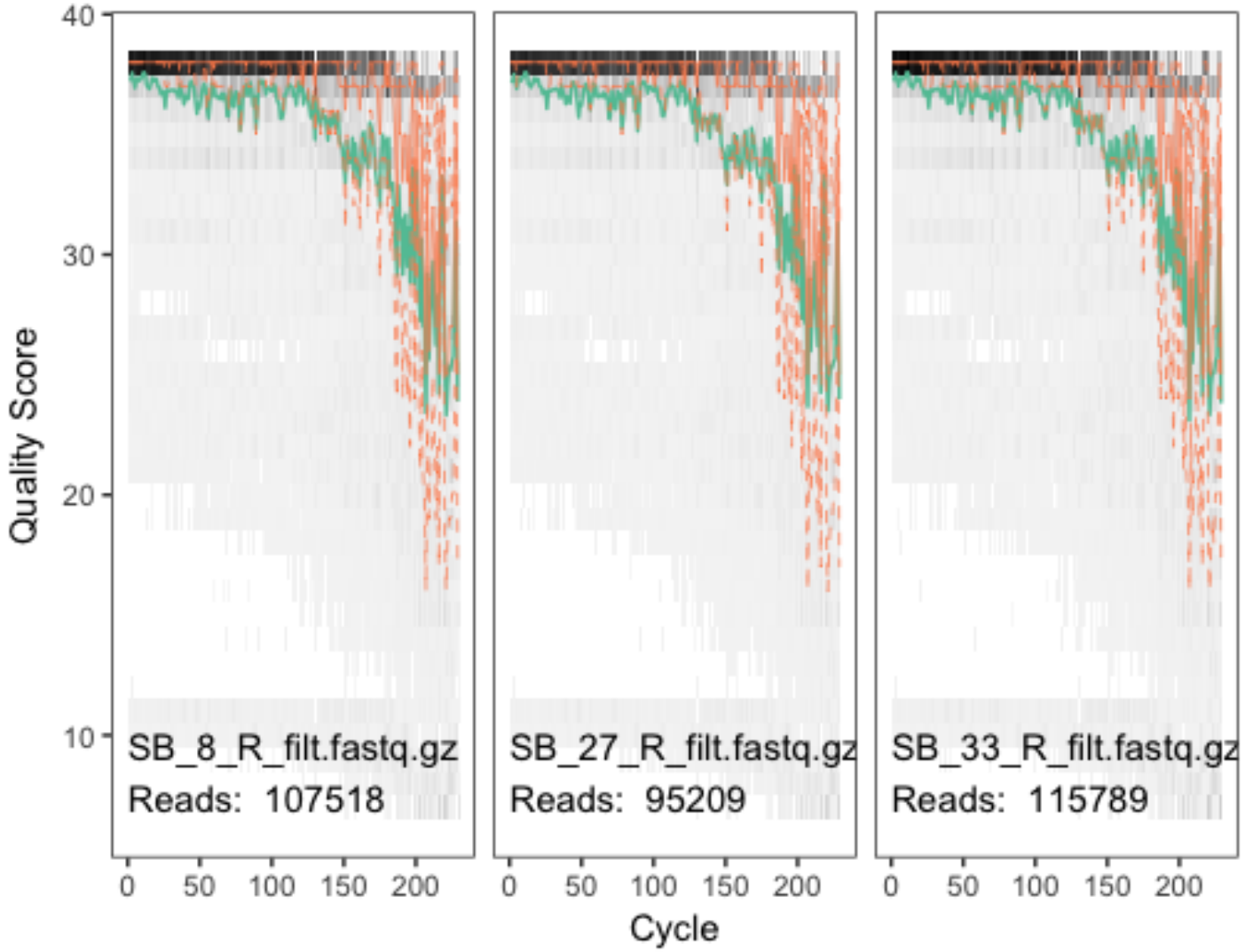
```
65 ```{r}
66 FWD <- "GTGYCAGCMGCCGCGGTAA"
67 REV <- "GGACTACNVGGGTWTCTAAT"
68 ```
```

ATTCGA
TAAGCT ✓

ATTCGA
TATGCT x

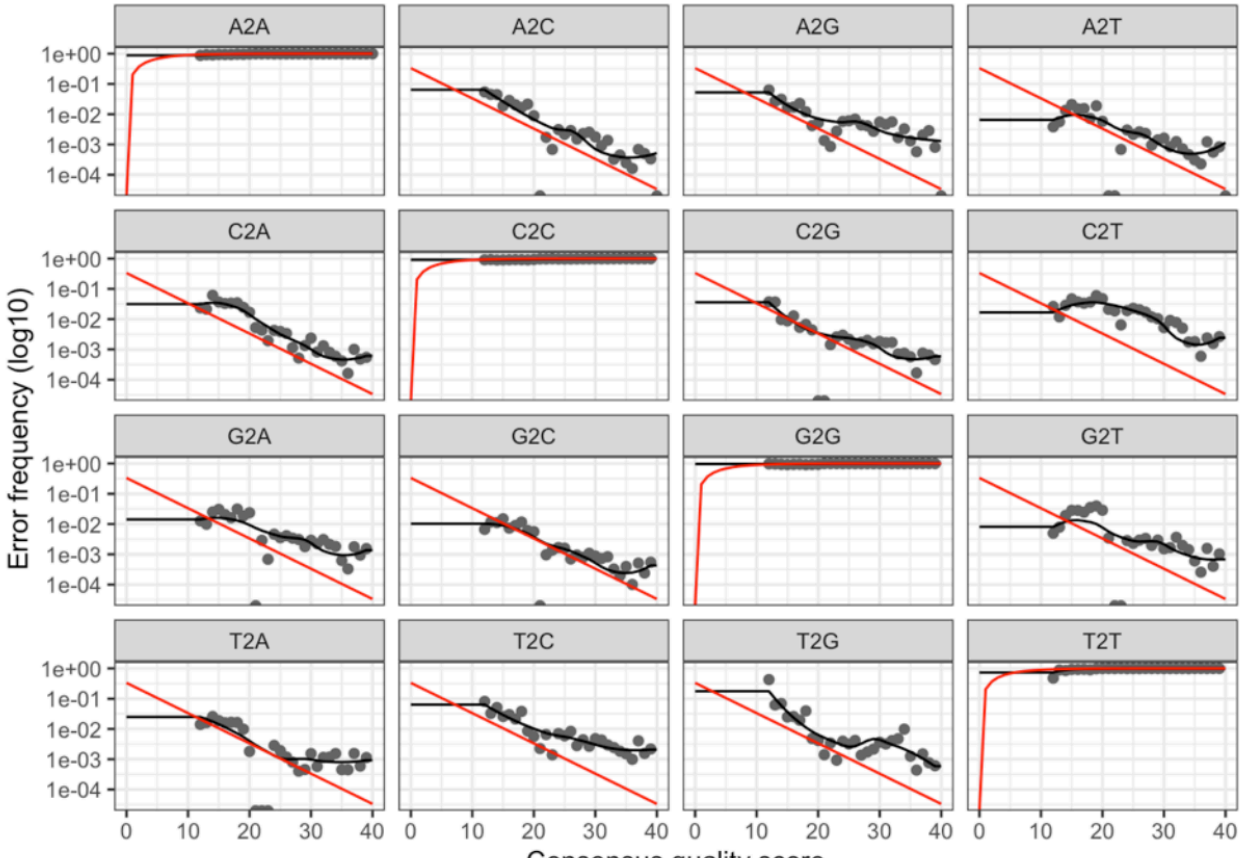
Quality filtering

- Different ways to accomplish this…
  - Minimum quality score required for inclusion
  - Minimum or maximum length

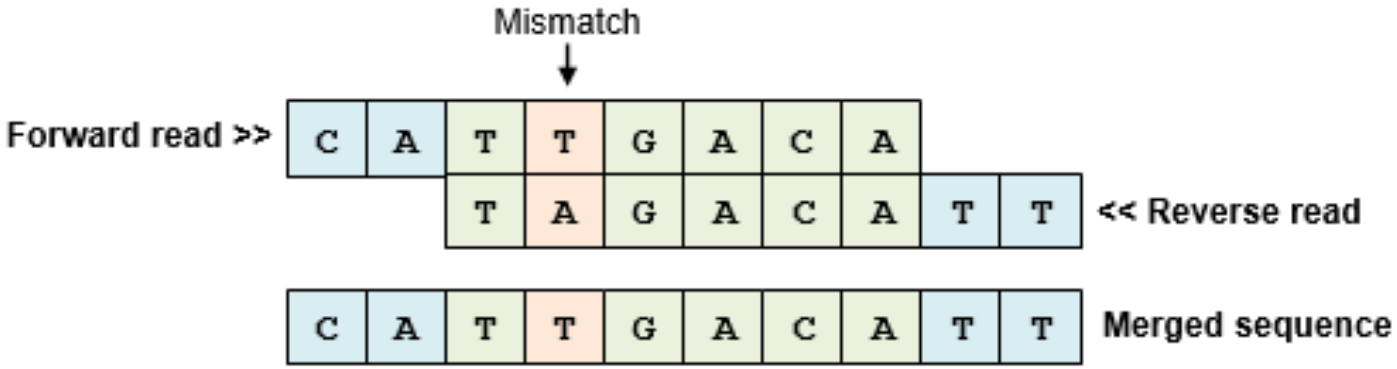


Infer sample composition

- Plotting probability of errors
- ”Fixes” sequencing errors
  - # of occurrences of each sequence



Merge reads



Why merge?

Serial No. of dollar bill

Considerations:

- Minimum number of overlapping base pairs
- Maximum number of mismatches

- What if you can’t get your reads to merge?

Dereplication

- During dereplication, we condense the data by collapsing together all reads that encode the same sequence.

- This step occurs in all pipelines, but DADA2 retains a summary of quality info here.

- Before there may be 10000s of repeats. After dereplication, we condense our dataset to a single representative sequence with counts.
  - Significantly reduces computation time

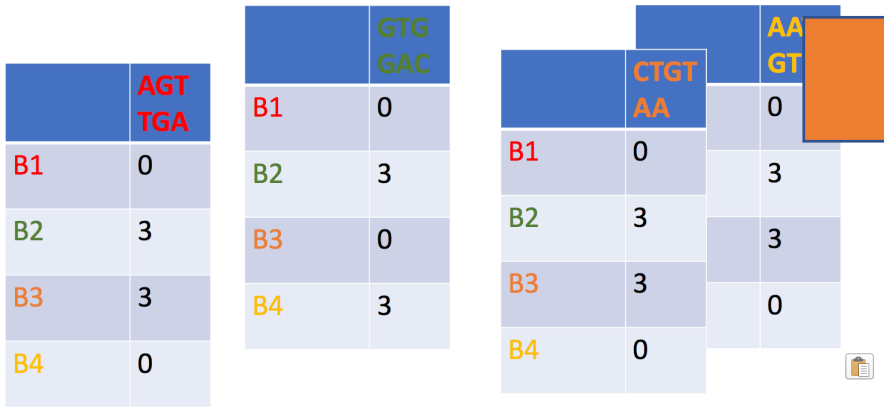
	AGTTGA
B1	0
B2	3
B3	3
B4	0

	GTGGAC
B1	0
B2	3
B3	0
B4	3

	CTGTAA
B1	0
B2	3
B3	3
B4	0

	AATAGT
	0
	3
	3
	0

Making sequence table



	B1	B2	B3	B4	B5
AGTTGA	6	0	13	10	15
GTGGAC	4	3	7	2	15
CTGTAA	3	3	0	10	12
TAAAGT	1	0	1	0	1

But we're not quite done…

Chimeras!

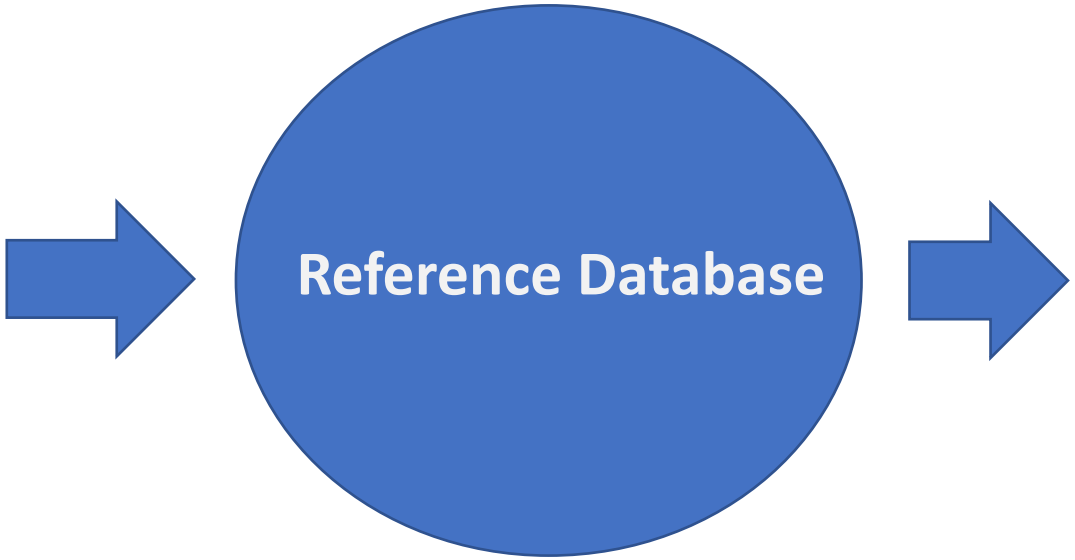
- What is a chimera?
- Why do we care about this in sequencing?

Chimera removal

	B1	B2	B3	B4	B5
AGTTGA	6	0	13	10	15
GTGGAC	4	3	7	2	15
CTGTAA	3	3	0	10	12
TAAAGT	1	0	1	0	1

Assigning taxonomy

Sequence
AGTTGA
GTGGAC
CTGTAA



Sequence	Taxonomy
AGTTGA	*Pseudomonas syringae*
GTGGAC	*Streptococcus sp.*
CTGTAA	*Mycobacterium tuberculosis*

	B1	B2	B3	B4	B5	Taxonomy
AGTTGA	6	0	13	10	15	*Pseudomonas syringae*
GTGGAC	4	3	7	2	15	*Streptococcus sp.*
CTGTAA	3	3	0	10	12	*Mycobacterium tuberculosis*

DADA2 vignette

https://www.bioconductor.org/packages/devel/bioc/vignettes/dada2/inst/doc/dada2-intro.html#dereplicate