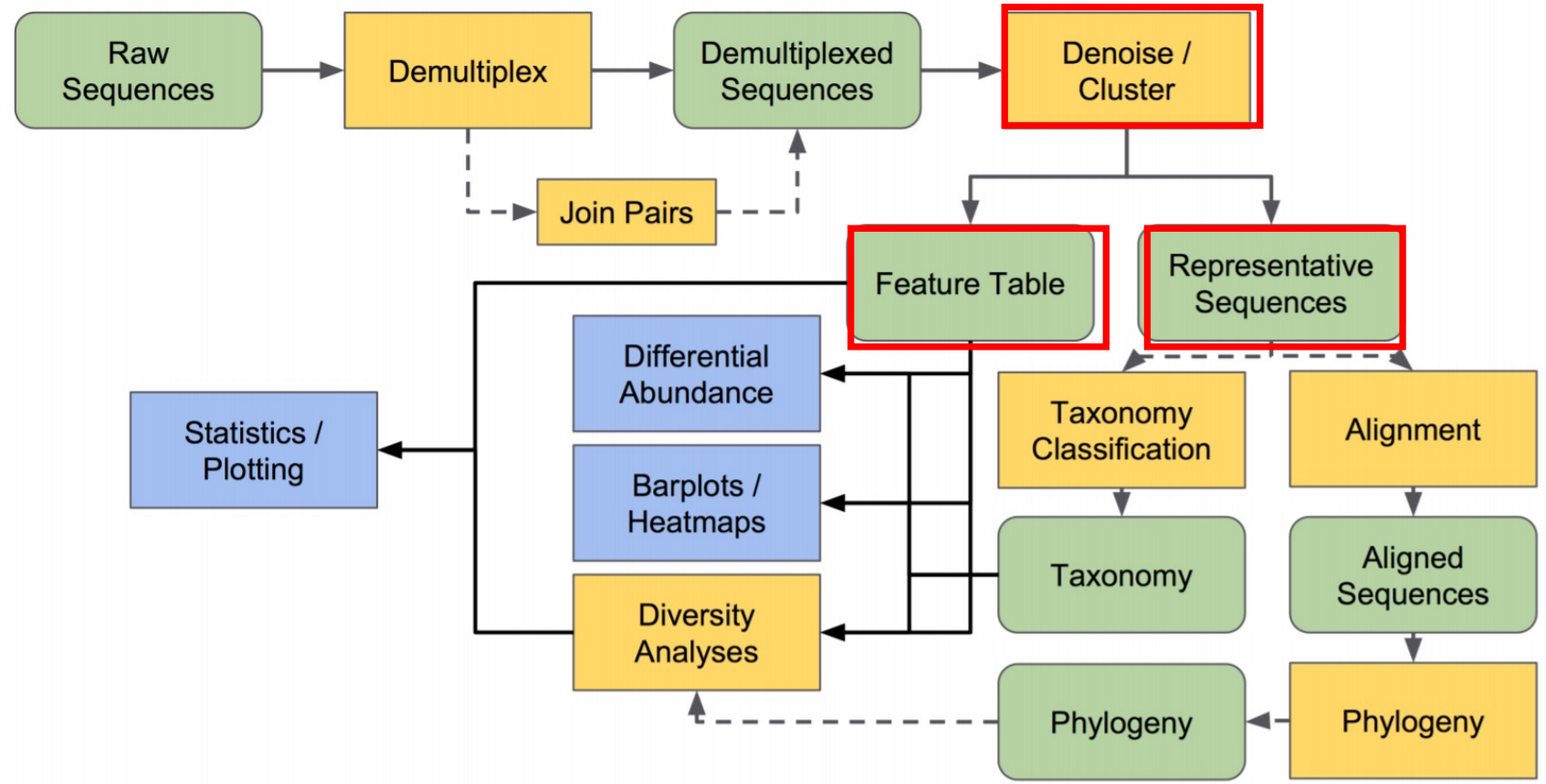
Module 5

Determine your ASVs

Module Outcomes

1. Denoise your data by trimming your reads and removing low quality reads using DADA2 or Deblur
2. Cluster unique reads in your data called Amplicon Sequence Variants (ASVs)
3. Distinguish between ASVs and OTUs

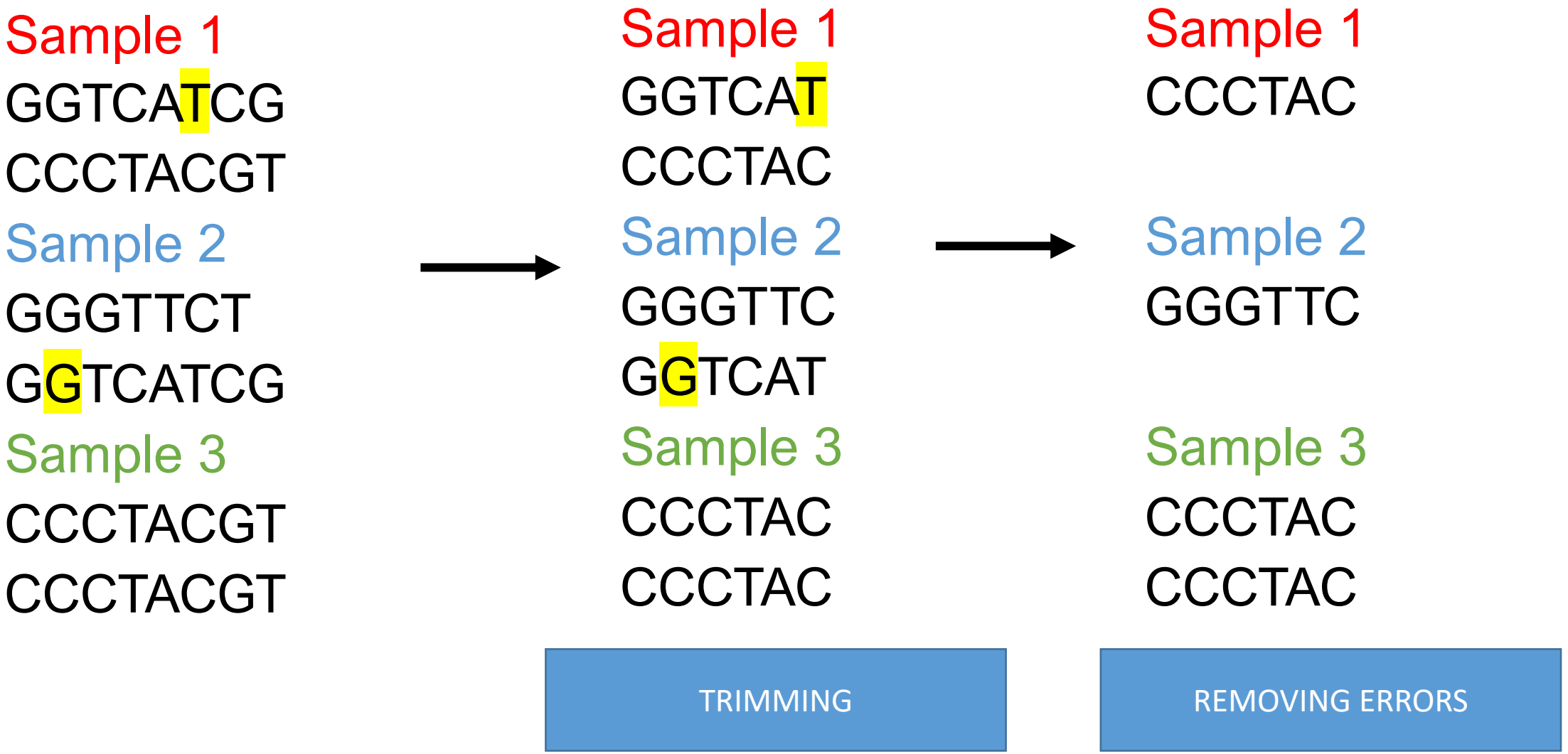
QIIME2 workflow



Yellow: processing steps
Green: inputs/outputs
Blue: R analysis

https://docs.qiime2.org/2018.8/tutorials/overview/#demultiplexing

Denoising (visual)



Denoising Tools

- DADA2 or Deblur (no consensus on which is better)
  - sequencing errors are detected and corrected (DADA2) or
  - detected and removed/filtered (Deblur)
- DADA2 and Deblur are unique algorithms that essentially achieve the same result in different ways
- We use DADA2 in the Moving Pics Tutorial

qiime2/**q2-deblur**

Denoising code

```
qiime dada2 denoise-single \
--i-demultiplexed-seqs demux.qza \
--p-trim-left 0 \
--p-trunc-len 120 \
--o-representative-sequences rep-seqs.qza \
--o-table table.qza \
--o-denoising-stats stats.qza
```

Calling on the DADA2 tool to denoise single end sequences

input file is your demultiplexed file

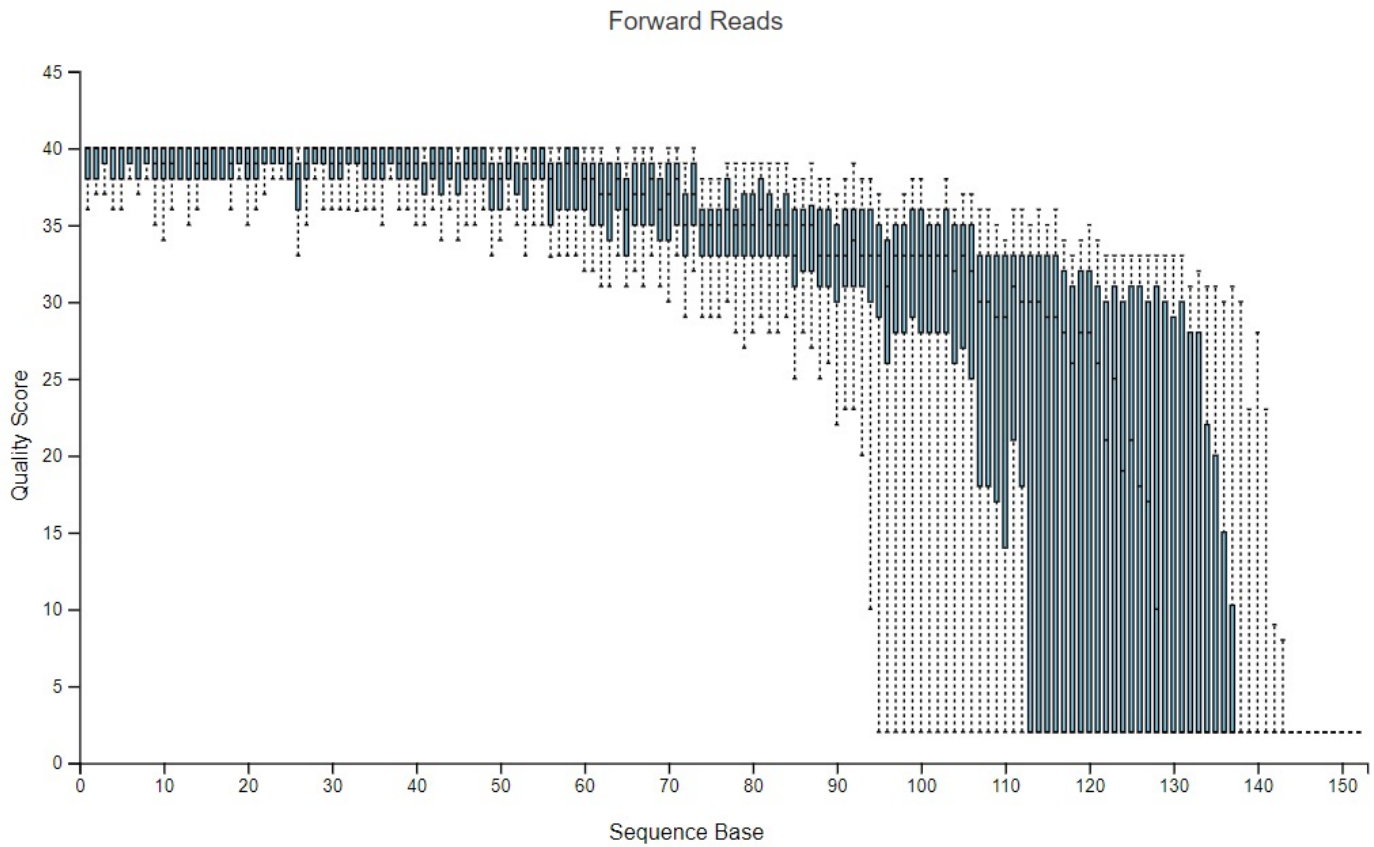
Your trimming parameters!

3 output files are generated

qiime2view File: **demux.qzv**

Overview | Interactive Quality Plot

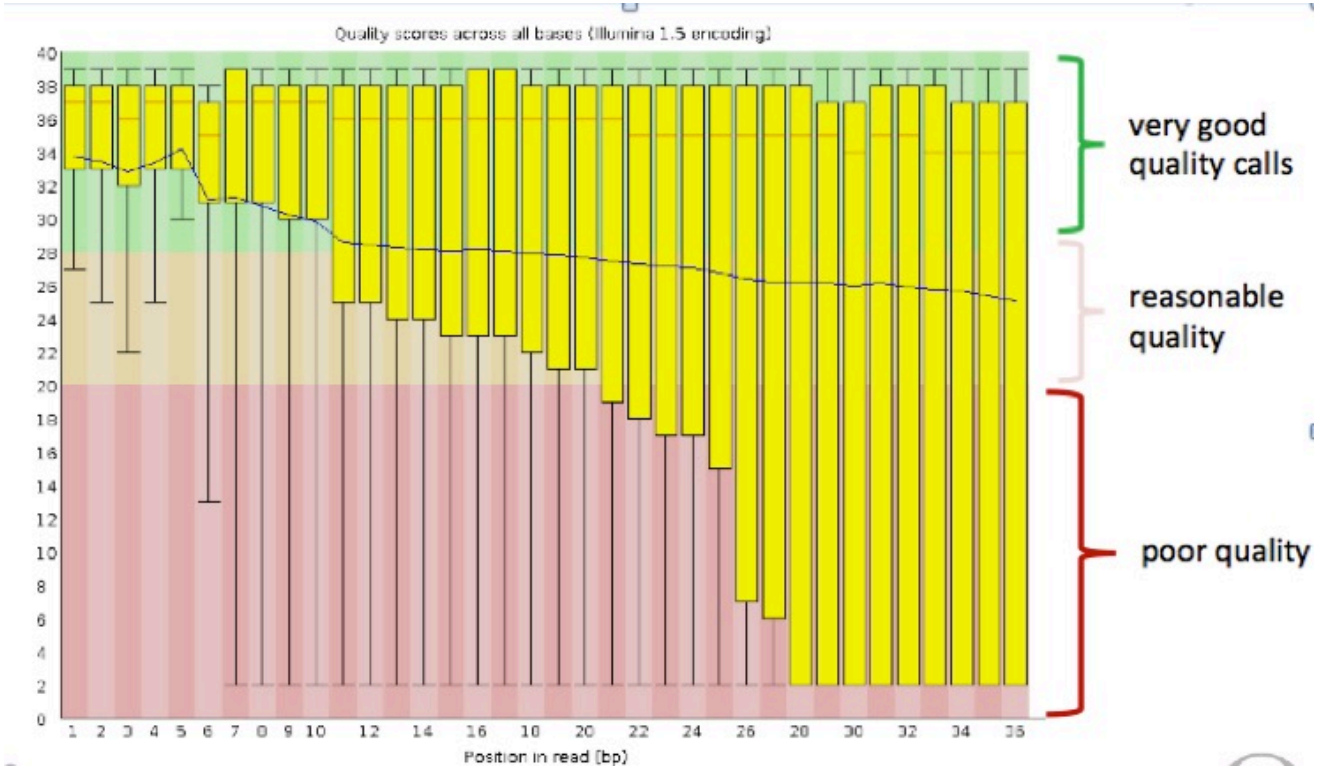
Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a bo



The plot at position 141 was generated using a random sampling of 10000 out of [object Object] sequences without replacement. The minimum sequence length identified during subsampling was 152 bases. Outlier quality scores are not shown in box plots for clarity.

Why does the quality drop at the end?

- Remember that you are reading bases in a cluster
- There is higher synchronization in the beginning
- More de-synchronization leading towards the end that contributes to poorer quality
- Of interest: Fuller et al., 2009, Nature Biotechnology



https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2017/Day1/Session4-seqIntro.html

Denoising code

```
qiime dada2 denoise-single \
--i-demultiplexed-seqs demux.qza \
--p-trim-left 0 \
--p-trunc-len 120 \
--o-representative-sequences rep-seqs.qza \
--o-table table.qza \
--o-denoising-stats stats.qza
```

Cluster into ASVs (visual)

Sample 1
CCCTAC

Sample 2
GGGTTC

Sample 3
CCCTAC
CCCTAC

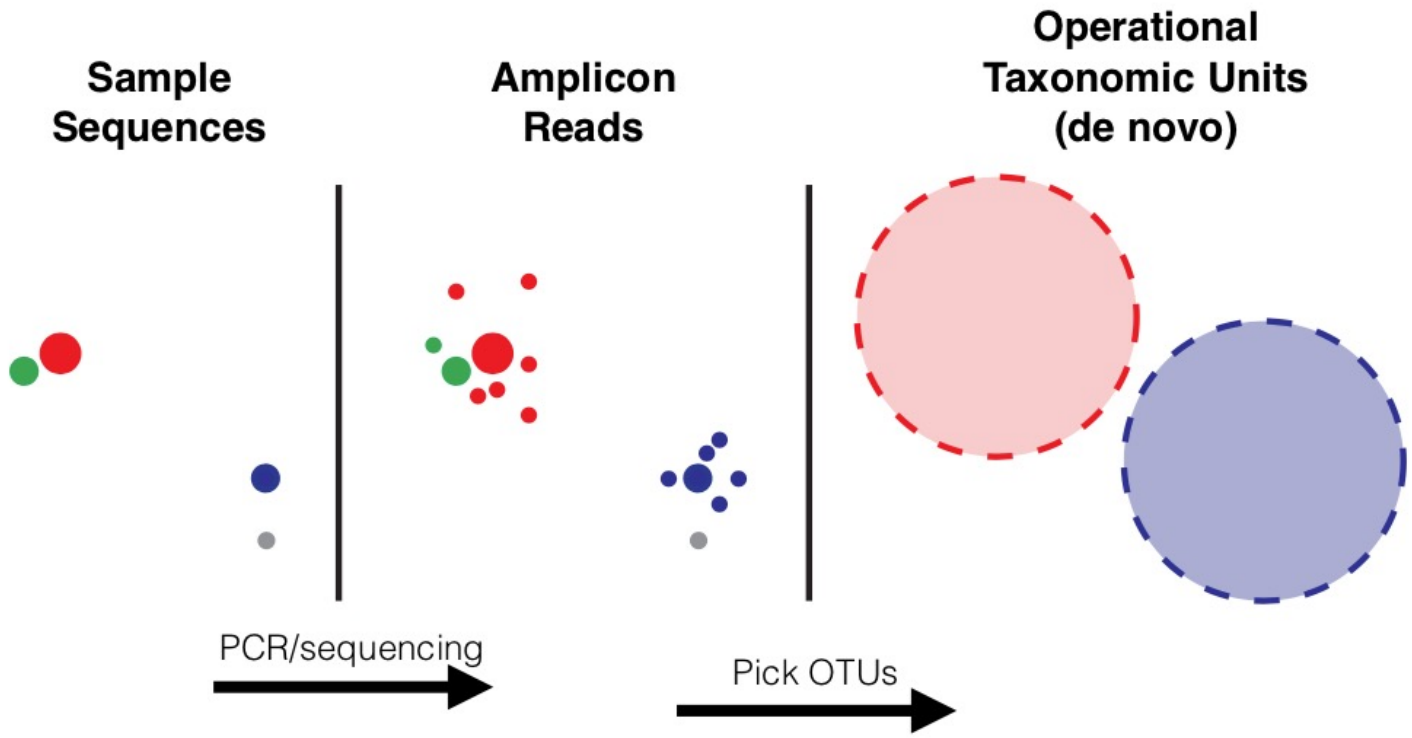
Only have 2 ASVs:

CCCTAC
abundant by 1 in sample 1 and 2 in sample 3

GGGTTC
abundant by 1 in sample 2

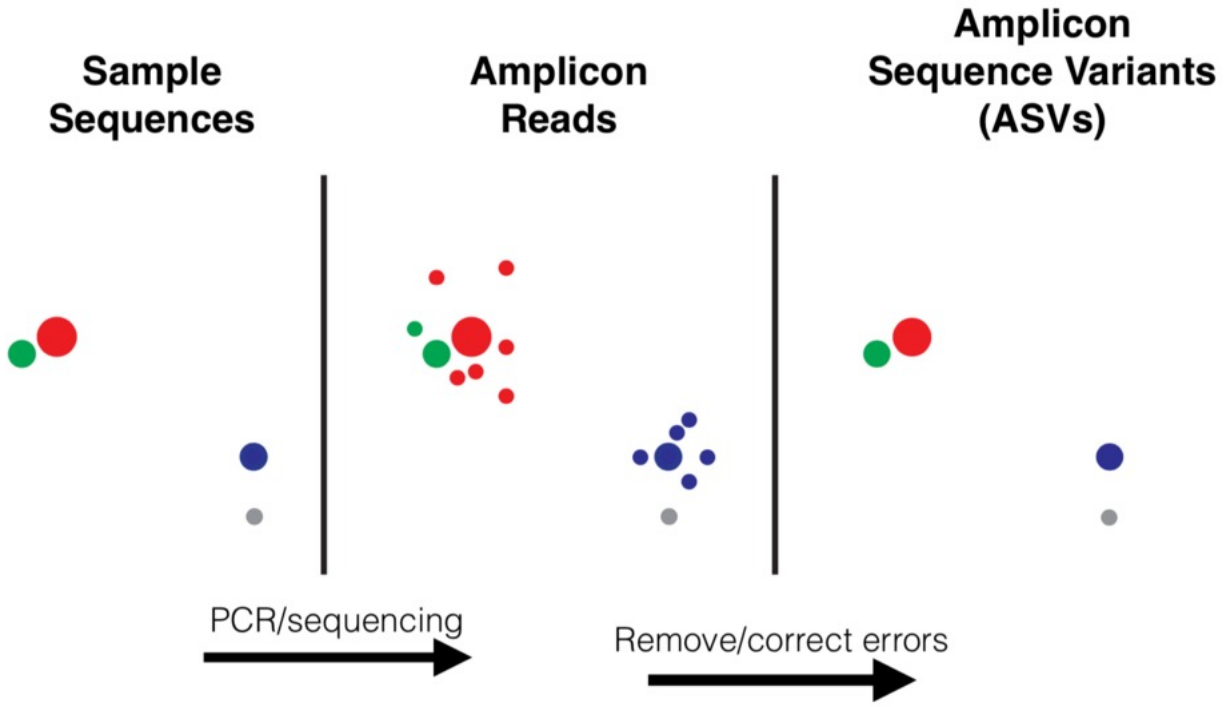
Operational Taxonomic Unit (OTU): the out approach

- Uses a reference sequence to find representative consensus sequence (subject to reference bias)
- Combines sequences that are very similar to compensate for potential errors (eg. 97% similar)
- Table then summarizes sequence “clusters”
- Lose some diversity information through this approach



Amplicon Sequence Variant (ASV)

- Describes “exact” sequences with high confidence (higher quality sequences only)
- Removes sequences that may have resulted from errors (eg. artifacts)
- No reference is used (until we start to assign taxonomy) so there is no reference bias
- Disadvantage: may throw out real sequences that were present in very low abundance
- Can also be called exact sequence variant (EZV) or zero-radius OTU (zOTU)



ASVs versus OTUs

ASVs	OTUs
No reference bias (reference not assigned until taxonomic analysis)	Subject to reference bias
Defined by exact sequences	Defined by consensus sequence of multiple variants (similar by 97%)
Keeps unique sequences separate	One sequence can represent multiple species
Can be used to compare between studies	Cannot be compared between studies

Clustering methods can impact inferred community structure

yellow stars = biological point mutations

red stars = sequencing errors

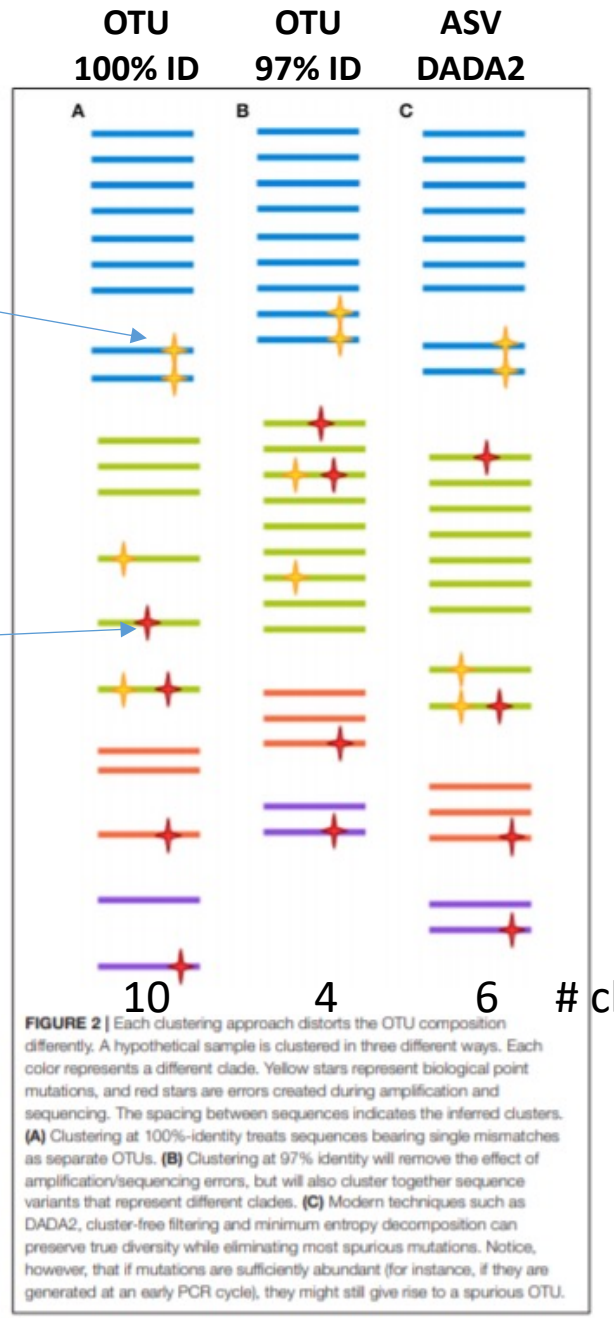


FIGURE 2 | Each clustering approach distorts the OTU composition differently. A hypothetical sample is clustered in three different ways. Each color represents a different clade. Yellow stars represent biological point mutations, and red stars are errors created during amplification and sequencing. The spacing between sequences indicates the inferred clusters. **(A)** Clustering at 100%-identity treats sequences bearing single mismatches as separate OTUs. **(B)** Clustering at 97% identity will remove the effect of amplification/sequencing errors, but will also cluster together sequence variants that represent different clades. **(C)** Modern techniques such as DADA2, cluster-free filtering and minimum entropy decomposition can preserve true diversity while eliminating most spurious mutations. Notice, however, that if mutations are sufficiently abundant (for instance, if they are generated at an early PCR cycle), they might still give rise to a spurious OTU.

https://www.frontiersin.org/articles/10.3389/fmicb.2017.01561/full

Open Access | Published: 21 July 2017

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan ✉, Paul J McMurdie & Susan P Holmes

The ISME Journal **11**, 2639–2643(2017) | Cite this article

Exact sequence variant = ASV = A unique post-quality-filtered sequence. It just takes one single base pair difference to define an entirely new ASV.

Denoising code

```
qiime dada2 denoise-single \
--i-demultiplexed-seqs demux.qza \
--p-trim-left 0 \
--p-trunc-len 120 \
--o-representative-sequences rep-seqs.qza \
--o-table table.qza \
--o-denoising-stats stats.qza
```

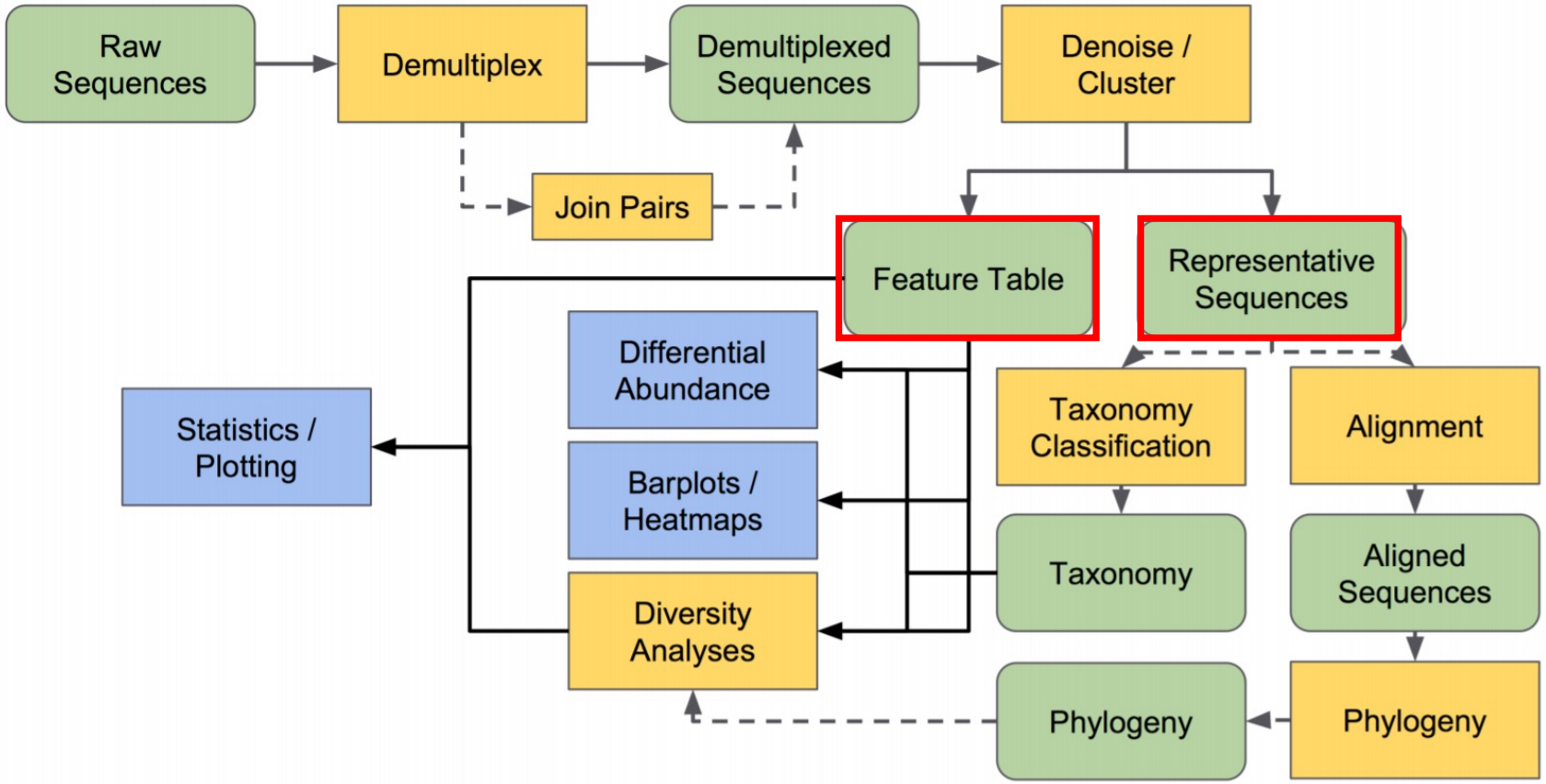
Converting your files to qzv

```
qiime feature-table summarize \
--i-table table.qza \
--o-visualization table.qzv \
--m-sample-metadata-file sample-metadata.tsv
```

you need your metadata file here

```
qiime feature-table tabulate-seqs \
--i-data rep-seqs.qza \
--o-visualization rep-seqs.qzv
```

QIIME2 workflow



https://docs.qiime2.org/2018.8/tutorials/overview/#demultiplexing