ICG INSTITUTE FOR COMPARATIVE GENOMICS

DALHOUSIE UNIVERSITY

From Sequences to ASV Tables

Rob Beiko and Diana Haider

(with thanks to Jacob Nearing and others)

Dalhousie University

Learning Objectives

- By the end of this lecture, you will be able to:
  - **Understand** the main strengths and weaknesses of the 16S rRNA gene
  - **Read** and **interpret** the contents of sequence files
  - **Describe** the process of sequence clustering
  - **Understand** the key differences between OTUs, ASVs, and taxonomic summaries

The Plan

(1) Why 16S? (10 minutes)
(2) Sequence data and quality control (10 minutes)
(3) Clustering your sequences (25-ish minutes)

Why 16S?

Taxonomy

- The classification of organisms according to a predefined hierarchy

- “How do we assign individuals” has often been based on what was measurable at the time (morphology, biochemistry, DNA hybridization, marker-gene similarity, whole-genome similarity)

Taxonomy as a genomic game

- Great!
  - Lots of “objective” characters to compare (e.g., nucleotides, amino acids)
  - More phylogenetically informative than morphology
  - Can be assessed in an automated way
  - Less prone to convergence, maybe

- But:
  - We cannot profile the complete genomes of all organisms in any but the simplest of microbial communities
  - So we need to choose a smaller sequence that everyone has…
  - …with enough characters to distinguish different groups
  - Variation and taxonomy may not have much to do with function (so won’t tell us about important evolutionary events)
  - Molecular evolution is *weird*

Genes that everyone has, more or less

- AMPHORA for bacterial core-genome phylogeny uses *dnaG*, *frr*, *infC*, *nusA*, *pgk*, *pyrG*, *rplA*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *rplK*, *rplL*, *rplM*, *rplN*, *rplP*, *rplS*, *rplT*, *rpmA*, *rpoB*, *rpsB*, *rpsC*, *rpsE*, *rpsI*, *rpsJ*, *rpsK*, *rpsM*, *rpsS*, *smpB*, *tsf*

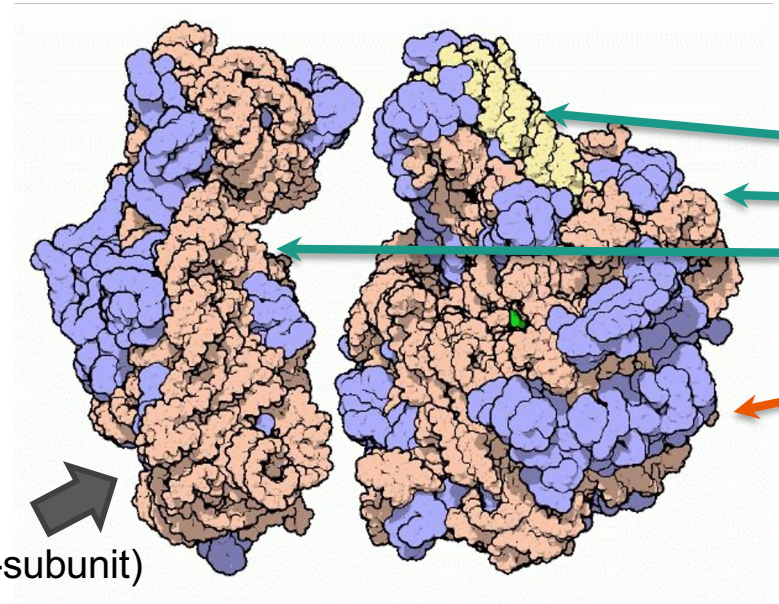
- 23 of these encode ribosomal proteins!

- Which are great, but there are…*issues* with using protein-coding genes.

- So…

Wu and Eisen (2009) *Genome Biology*

The prokaryotic ribosome (more or less)

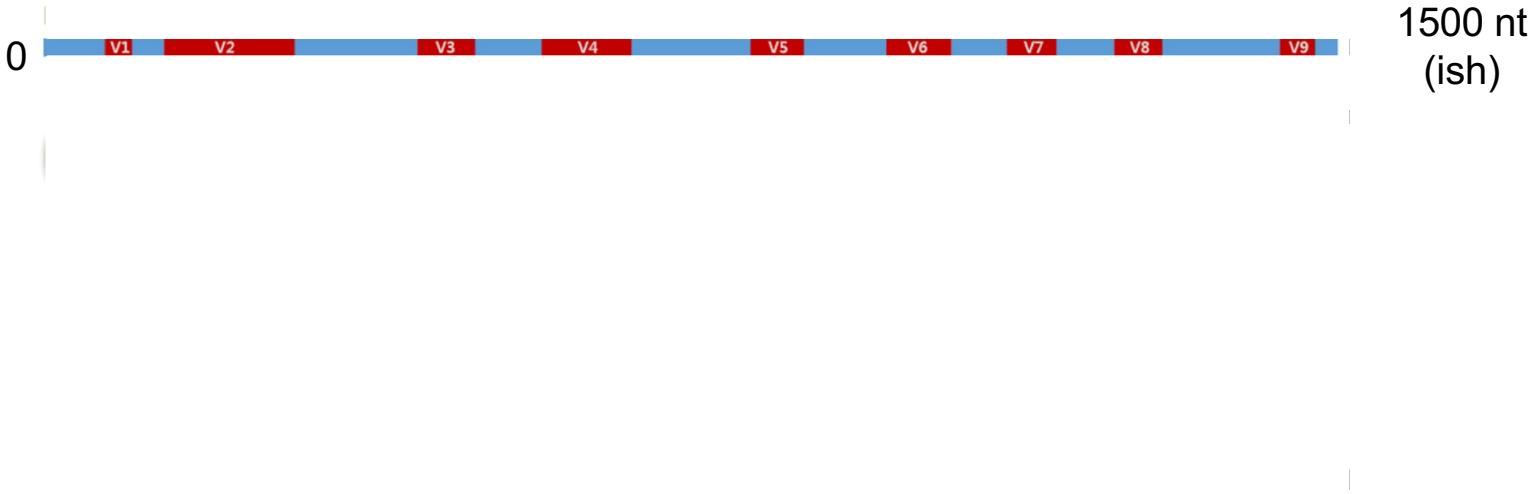


16S (small-subunit)

3 rRNAs

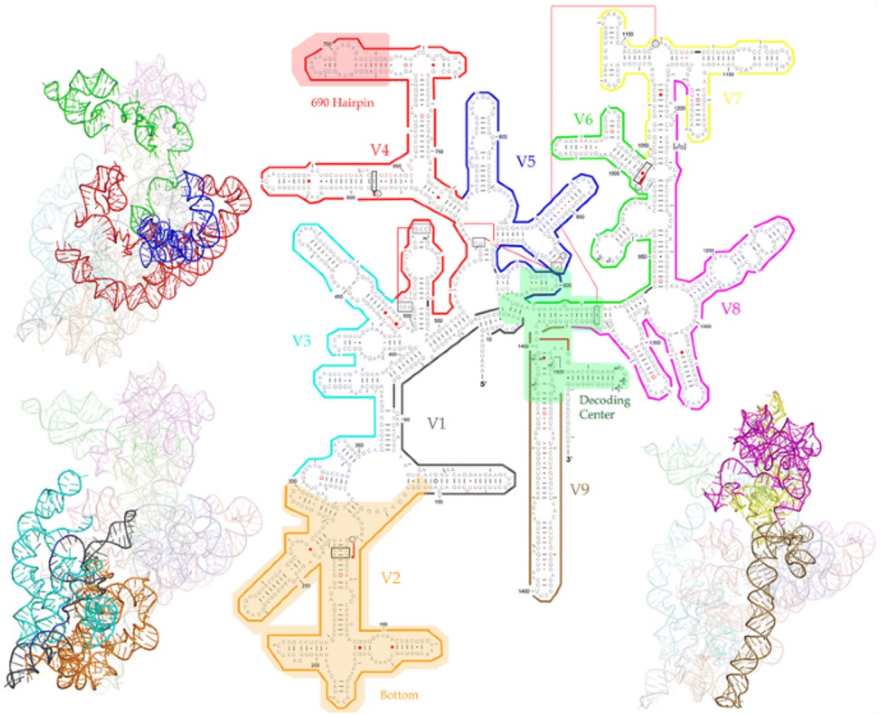
~45 proteins

The 16S rRNA gene



Modified from https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/

Not just a straight line!



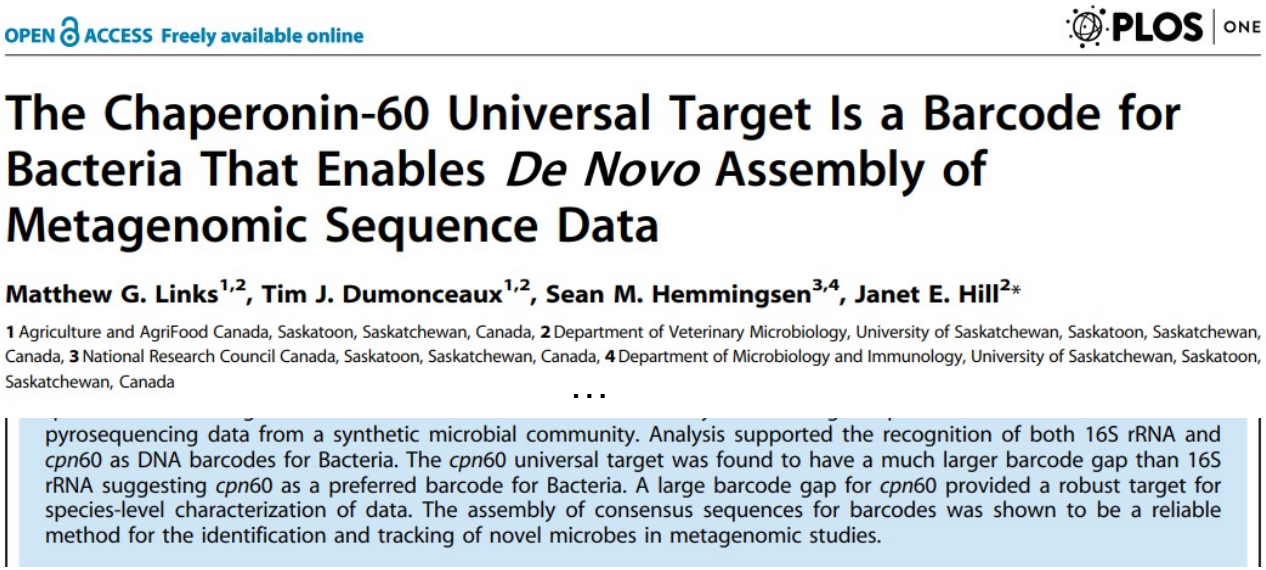
Yang et al (2016) *BMC Bioinformatics*

The Story of 16S

- Good!
  - Everyone has it
  - Conserved and variable regions, none of this genetic code nonsense
  - An absurd number of sequences can be found in public databases (9,469,124 in SILVA 138.1)

- Less good!
  - Found in multiple copies, recombination and gene transfer are things
  - Insufficient resolution at the finest (strain) levels
  - Differential amplification of groups depending on primer affinity (not 16S's fault)

16S is not the only option!



OPEN ACCESS Freely available online

PLOS ONE

The Chaperonin-60 Universal Target Is a Barcode for Bacteria That Enables *De Novo* Assembly of Metagenomic Sequence Data

Matthew G. Links[1,2], Tim J. Dumonceaux[1,2], Sean M. Hemmingsen[3,4], Janet E. Hill[2]*

1 Agriculture and AgriFood Canada, Saskatoon, Saskatchewan, Canada, **2** Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, **3** National Research Council Canada, Saskatoon, Saskatchewan, Canada, **4** Department of Microbiology and Immunology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

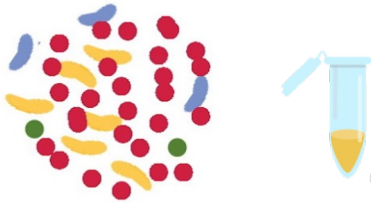
...

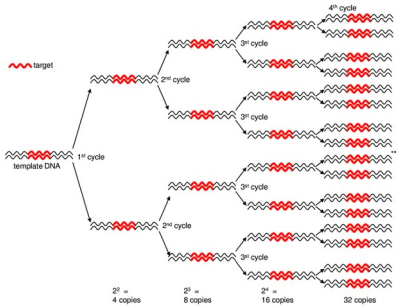
pyrosequencing data from a synthetic microbial community. Analysis supported the recognition of both 16S rRNA and *cpn*60 as DNA barcodes for Bacteria. The *cpn*60 universal target was found to have a much larger barcode gap than 16S rRNA suggesting *cpn*60 as a preferred barcode for Bacteria. A large barcode gap for *cpn*60 provided a robust target for species-level characterization of data. The assembly of consensus sequences for barcodes was shown to be a reliable method for the identification and tracking of novel microbes in metagenomic studies.

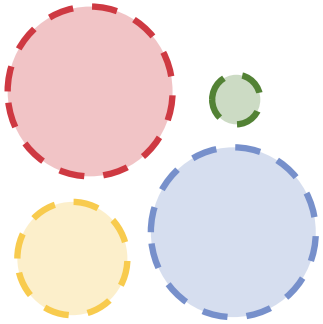
cpndb: https://www.cpndb.ca/publications.php

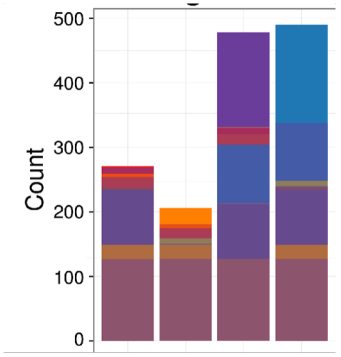
(2012)

The basic microbial diversity pipeline





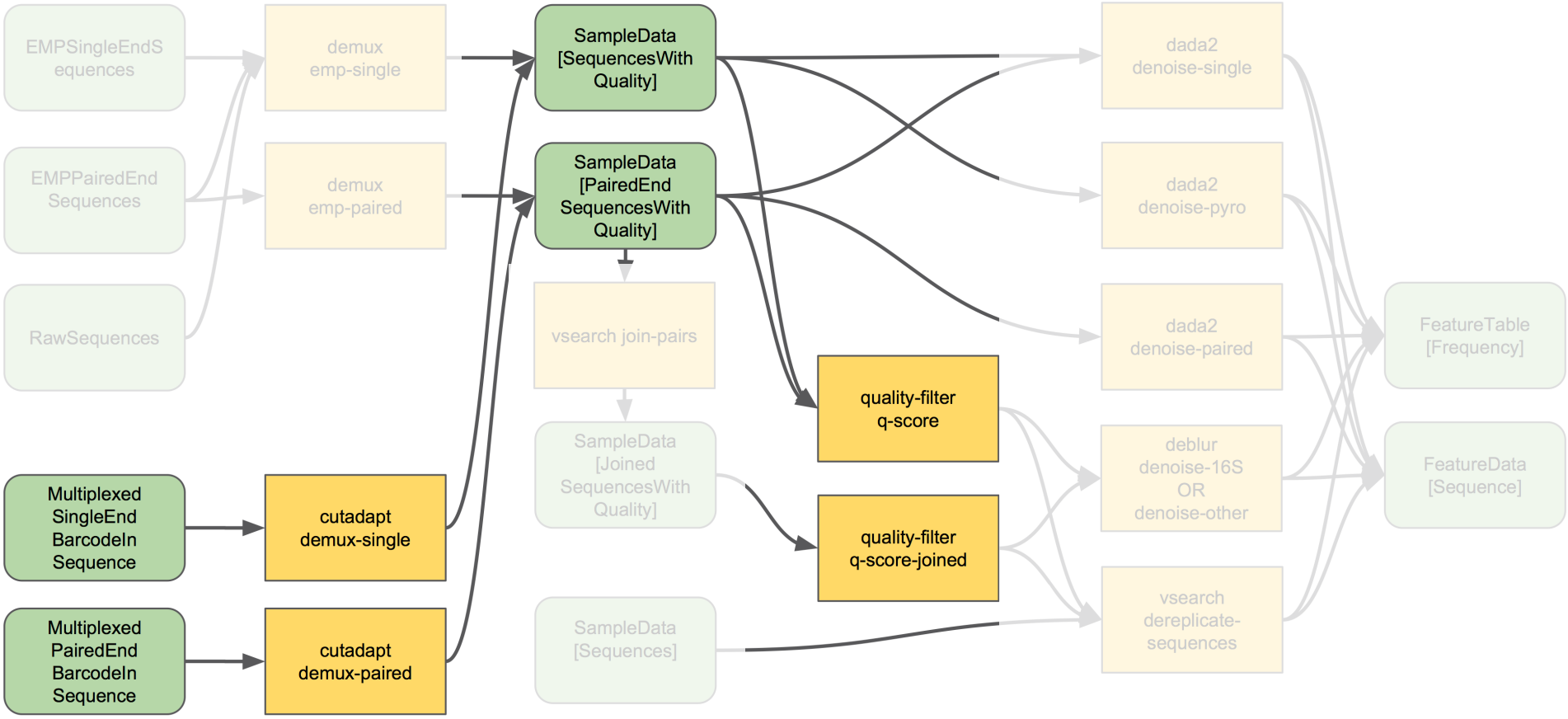




Sample → DNA → Amplify and sequence → Cluster → Diversity analysis

Plus statistics and possibly machine learning

Sequence retrieval and quality control


EMPSingleEndSequences
EMPPairedEndSequences
RawSequences
demux emp-single
demux emp-paired
SampleData [SequencesWithQuality]
SampleData [PairedEndSequencesWithQuality]
vsearch join-pairs
SampleData [JoinedSequencesWithQuality]
SampleData [Sequences]
Multiplexed SingleEnd BarcodeInSequence
Multiplexed PairedEnd BarcodeInSequence
cutadapt demux-single
cutadapt demux-paired
quality-filter q-score
quality-filter q-score-joined
dada2 denoise-single
dada2 denoise-pyro
dada2 denoise-paired
deblur denoise-16S OR denoise-other
vsearch dereplicate-sequences
FeatureTable [Frequency]
FeatureData [Sequence]

Anatomy of a FASTQ file

Header line
Sequence
+
PHRED Quality scores

```
@M03730:144:000000000-CG9WN:1:1101:16483:2296 1:N:0:377
ACGCGAAAAACCTTACCAGGTCTTGACATCTAG...
+
CCCCCDGGGGGGGGGGGGGGGGGGGGGGGGG...
```

'C': ASCII code 67, PHRED score 34 (good!)

Importing into QIIME2

- Sequencing data will generally be in FASTQ format
- Paired-end reads will have two associated files, usually designated R1 and R2:

```
-rwxrwxrwx 1 rbeiko rbeiko 14678002 Mar 27 09:15 1051C-M-D00-V_S353_L001_R1_001.fastq.gz
-rwxrwxrwx 1 rbeiko rbeiko 19940005 Mar 27 09:15 1051C-M-D00-V_S353_L001_R2_001.fastq.gz
```

- Files are zipped to save space; this is fine as you can still view contents (zmore) and QIIME2 works directly on the zipped files

FASTQ format

First four lines of ‘R1”:

```
@M03730:144:000000000-CG9WN:1:1101:16483:2296 1:N:0:377
ACGCGAAAAACCTTACCAGGTCTTGACATCTAGTGCCATTTGTAGAGATACAAAGTTCCCTTCGGGGACGCTAAGACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAG
ATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTTATTAGTTGCCAGCATTAAGTTGGGCACTCTAATGAGACTGCCGGTGACAAACCAGAGGAAGGTGGGGATGACGTC
AAGTCATCATGCCCCTTATGACCTGGGCTACACACGTGCTACAATGGGCAGGACAACGAGAAGCGAGCCTGCGAA
+
CCCCCDGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGFFGGGGGGGGGGGFGGGGGFGGGGGGGGEGGG
GGGGGGGGGGGGGGGGFGGGGGGDFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGFGGFGGFFFFGFBGGGFFFGFFGFFDGFFFFFFGAF>6>?FFFFD2<
@DFECC?E<).>;@D?FF<>ADFAFFF?F6:??FFB??B:B9BFFFFFF<?((3:3:>?B9?B(7<BB>1::961
```

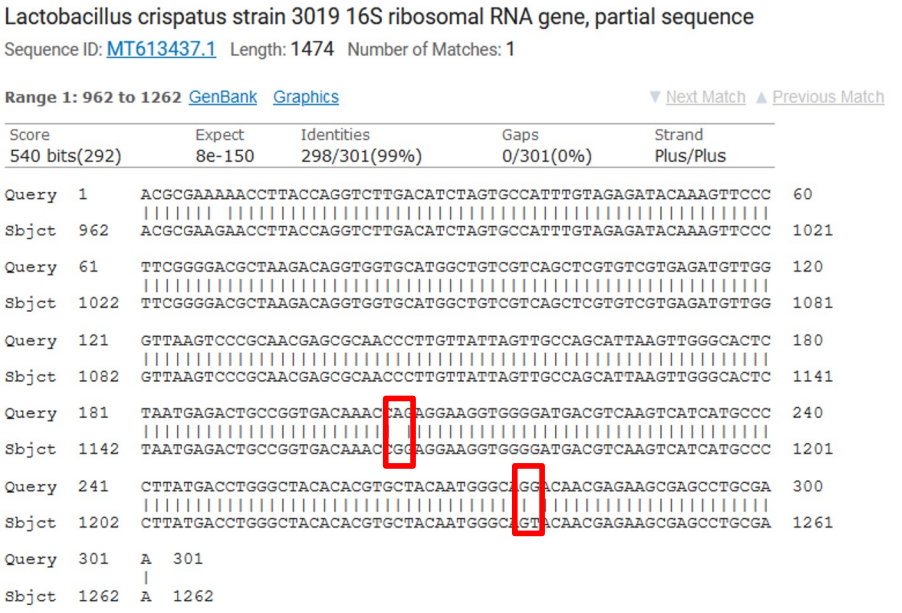
First four lines of ‘R2”:

```
@M03730:144:000000000-CG9WN:1:1101:16483:2296 2:N:0:377
ACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCGGCGTGCTGATCCGCGATTACTAGCGATTCCAGCTTCGTGCAGTCGAGTTGCAGACTGCAGTCCGAACTGAGAA
CAGCTTTCAGAGATTCGCTTGCCTTCGCAGGCTCGCTTCTCGTTGTACTGCCCATTGTAGCACGTGTGTAGCCCAGGTCATAAGGGGCATGATGACTTGACGTCATCCCCACC
TTCCTCTGGGTTGTCACCGGCAGTCTCATTAGAGTGCCCAACTTAATGCTGGCCACTAATAACAAGGGTTGCGC
+
CCCCCGGGGGFGGGGGGGGGGGGGGGGEGGFGGGGGGGGGGGGGGGGFFGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGFGGFGFGGGEGEGGGGFFGGGFFGGGGGGGGGGGGG
FGFGGFGGGGGGGGGGGEGGCGGGGGGGGGGGGGCFGGG>DF?DD>AAFG7,54?ACFFFFFFEFGFBAFDF=6?A:F8;@AFFF>8@F05(;;<FEFEC<=?D1<::<AFF79?
?09?9>)50(8>>?AFFFF(,(4:>)4)0<6:27<AA93>B<:93<<)0)),25@((.4<4:<4).2/26-(9>
```

Is it a real sequence?

I searched against the RefSeq database using the previous sequence as query (you generally won't do this). "Subject" is a 16S gene from *Lactobacillus plantarum*

Looks like it's legit…mismatches could be real, or sequencing errors



Lactobacillus crispatus strain 3019 16S ribosomal RNA gene, partial sequence

Sequence ID: MT613437.1 Length: 1474 Number of Matches: 1

Range 1: 962 to 1262 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
540 bits(292)	8e-150	298/301(99%)	0/301(0%)	Plus/Plus

```
Query  1     ACGCGAAAAACCTTACCAGGTCTTGACATCTAGTGCCATTTGTAGAGATACAAAGTTCCC  60
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  962   ACGCGAAGAACCTTACCAGGTCTTGACATCTAGTGCCATTTGTAGAGATACAAAGTTCCC  1021

Query  61    TTCGGGGACGCTAAGACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1022  TTCGGGGACGCTAAGACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGG  1081

Query  121   GTTAAGTCCCGCAACGAGCGCAACCCTTGTTATTAGTTGCCAGCATTAAGTTGGGCACTC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1082  GTTAAGTCCCGCAACGAGCGCAACCCTTGTTATTAGTTGCCAGCATTAAGTTGGGCACTC  1141

Query  181   TAATGAGACTGCCGGTGACAAACCAGGAGGAAGGTGGGGATGACGTCAAGTCATCATGCCC  240
             ||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||
Sbjct  1142  TAATGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGCCC  1201

Query  241   CTTATGACCTGGGCTACACACGTGCTACAATGGGCAGGACAACGAGAAGCGAGCCTGCGA  300
             |||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  1202  CTTATGACCTGGGCTACACACGTGCTACAATGGGCAGTACAACGAGAAGCGAGCCTGCGA  1261

Query  301   A  301
             |
Sbjct  1262  A  1262
```

Importing into QIIME2

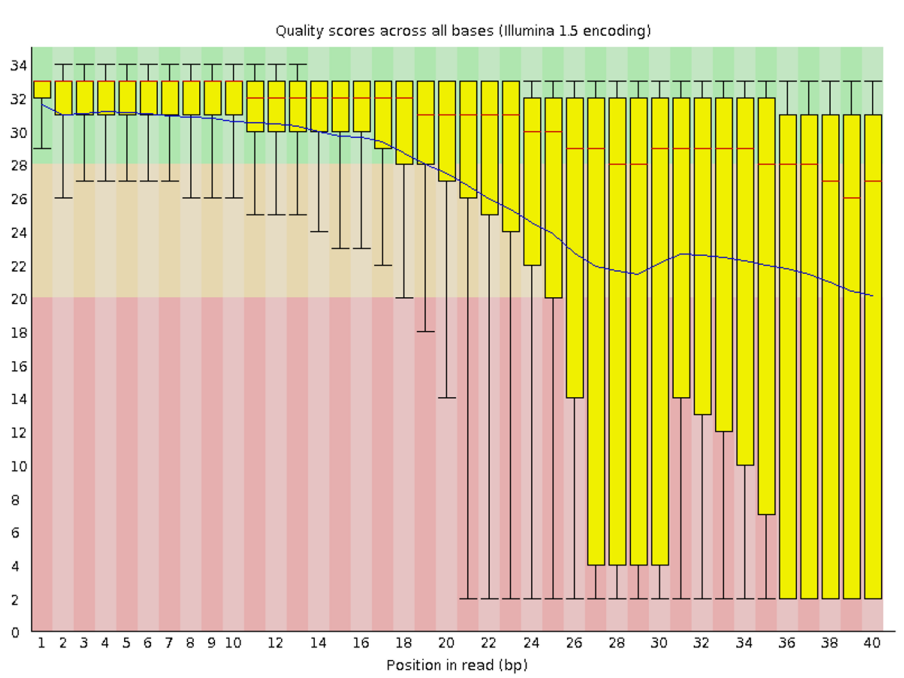
```
qiime tools import \
--type 'SampleData[PairedEndSequencesWithQuality]' \
--input-path import_to_qiime \
--output-path CBW_reads
```

QIIME wants all the fastq.gz files to be in the same directory ('import_to_qiime' in this case)

Visualizing sequence quality

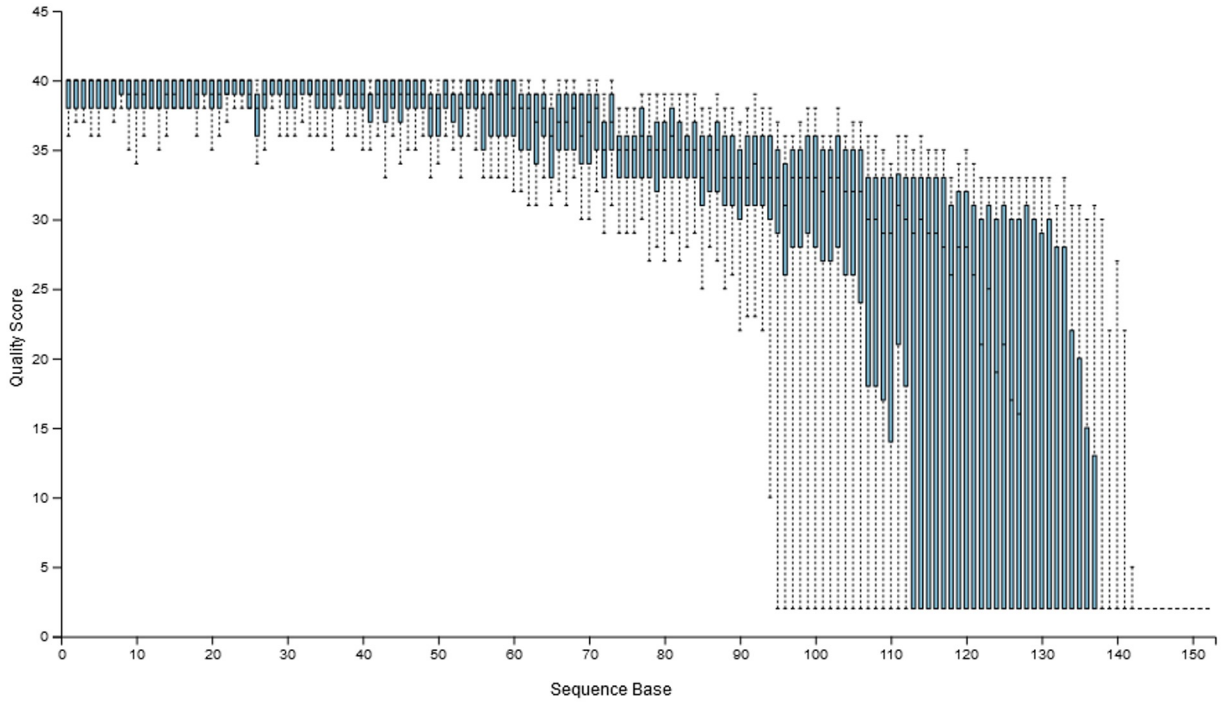
Average over all reads in a dataset

Can show red flags of poor sequencing run



fastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

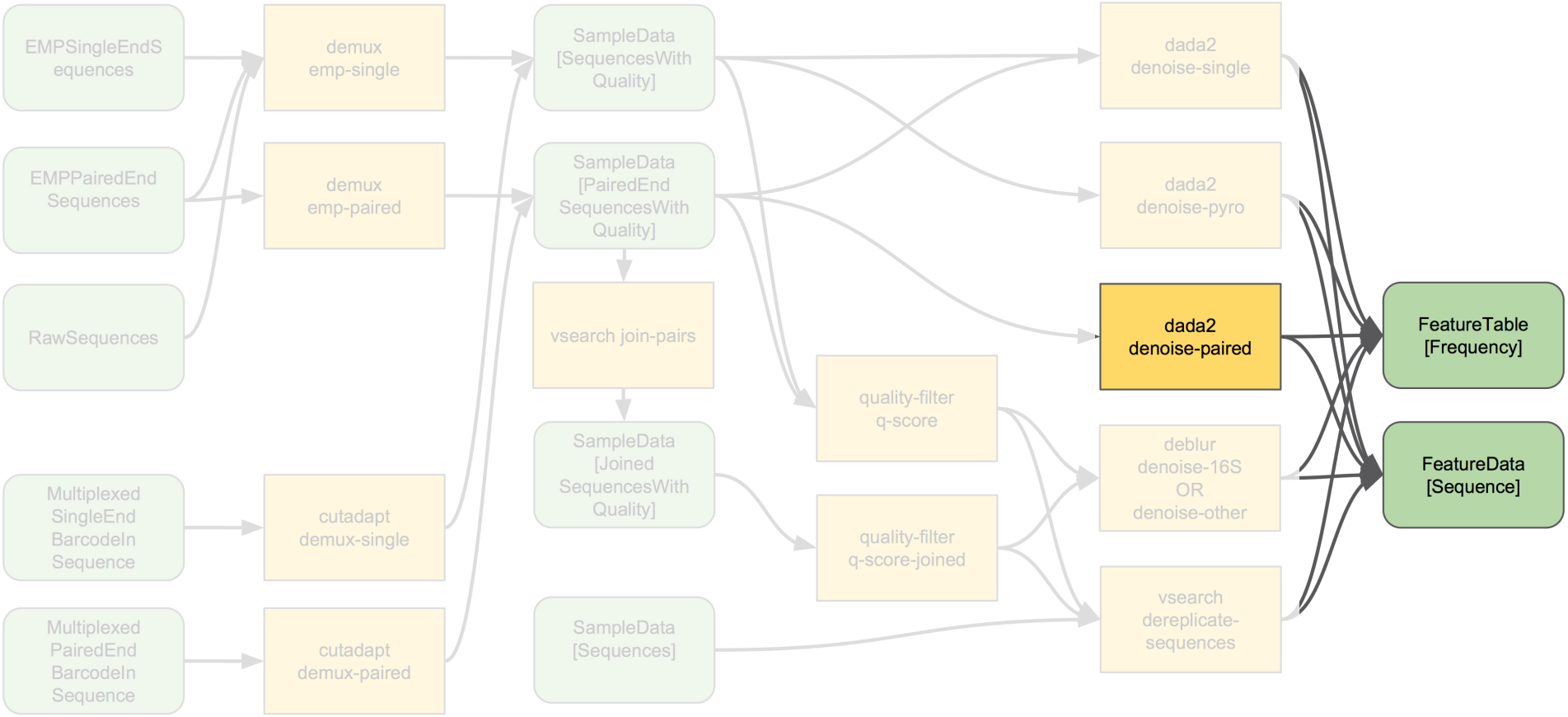
QIIME2 interactive quality plot (from “demux”)



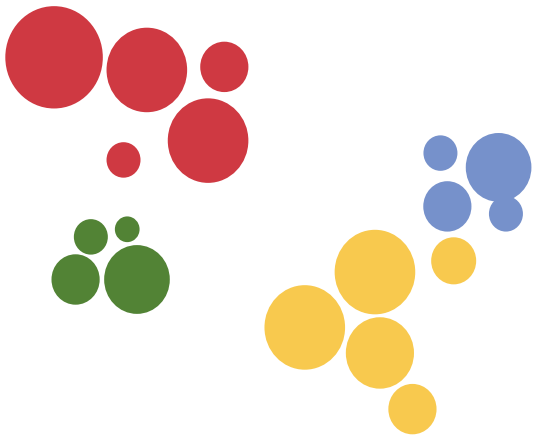
Files are imported. What now?

- If sequence files were multiplexed (i.e., multiple samples in a single file), ‘cutadapt’ can be used to assign sequences to samples according to provided barcodes.
- You can also remove reads below a certain length in ‘cutadapt’
- Additional processing and use of quality scores will be carried out in the DADA2 step (next module)

Sequence clustering

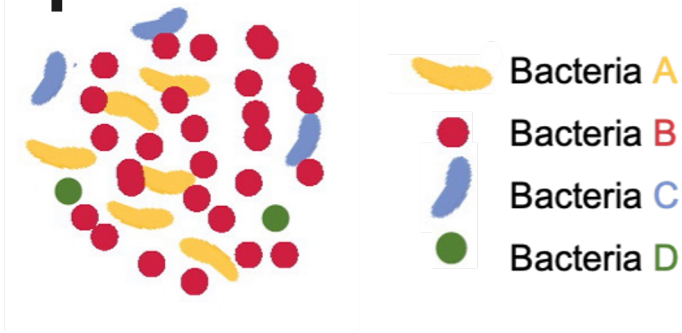

EMPSingleEndSequences
EMPPairedEndSequences
RawSequences
MultiplexedSingleEndBarcodeInSequence
MultiplexedPairedEndBarcodeInSequence
demux emp-single
demux emp-paired
cutadapt demux-single
cutadapt demux-paired
SampleData [SequencesWithQuality]
SampleData [PairedEndSequencesWithQuality]
vsearch join-pairs
SampleData [JoinedSequencesWithQuality]
SampleData [Sequences]
quality-filter q-score
quality-filter q-score-joined
dada2 denoise-single
dada2 denoise-pyro
dada2 denoise-paired
deblur denoise-16S OR denoise-other
vsearch dereplicate-sequences
FeatureTable [Frequency]
FeatureData [Sequence]

Why do we cluster sequences?



- Reduce dimensionality
- Remove sequencing artifacts
- Compensate for within-species variability

Operational Taxonomic Units (OTUs)



ATCGATCGATCGATGCTAGATATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGA

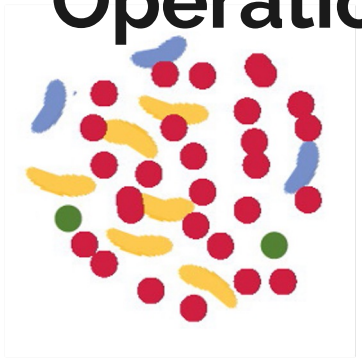
ATCGATCGATCGAGGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGA

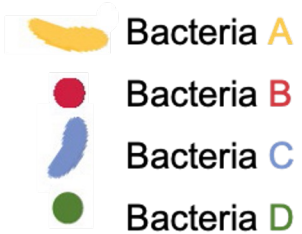
$$\% \; identity = \frac{\#\; matches}{length\; of\; sequence} \cdot 100$$

$$\% \; identity = \frac{98}{100} \cdot 100$$

$$\% \; identity = 98\%$$

Operational Taxonomic Units (OTUs)





ATCGATCGATCGATGCTAGATATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGAGGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA

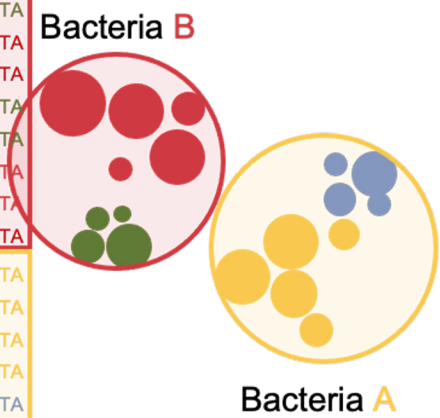
Operational Taxonomic Units (OTUs)

Bacteria A
Bacteria B
Bacteria C
Bacteria D

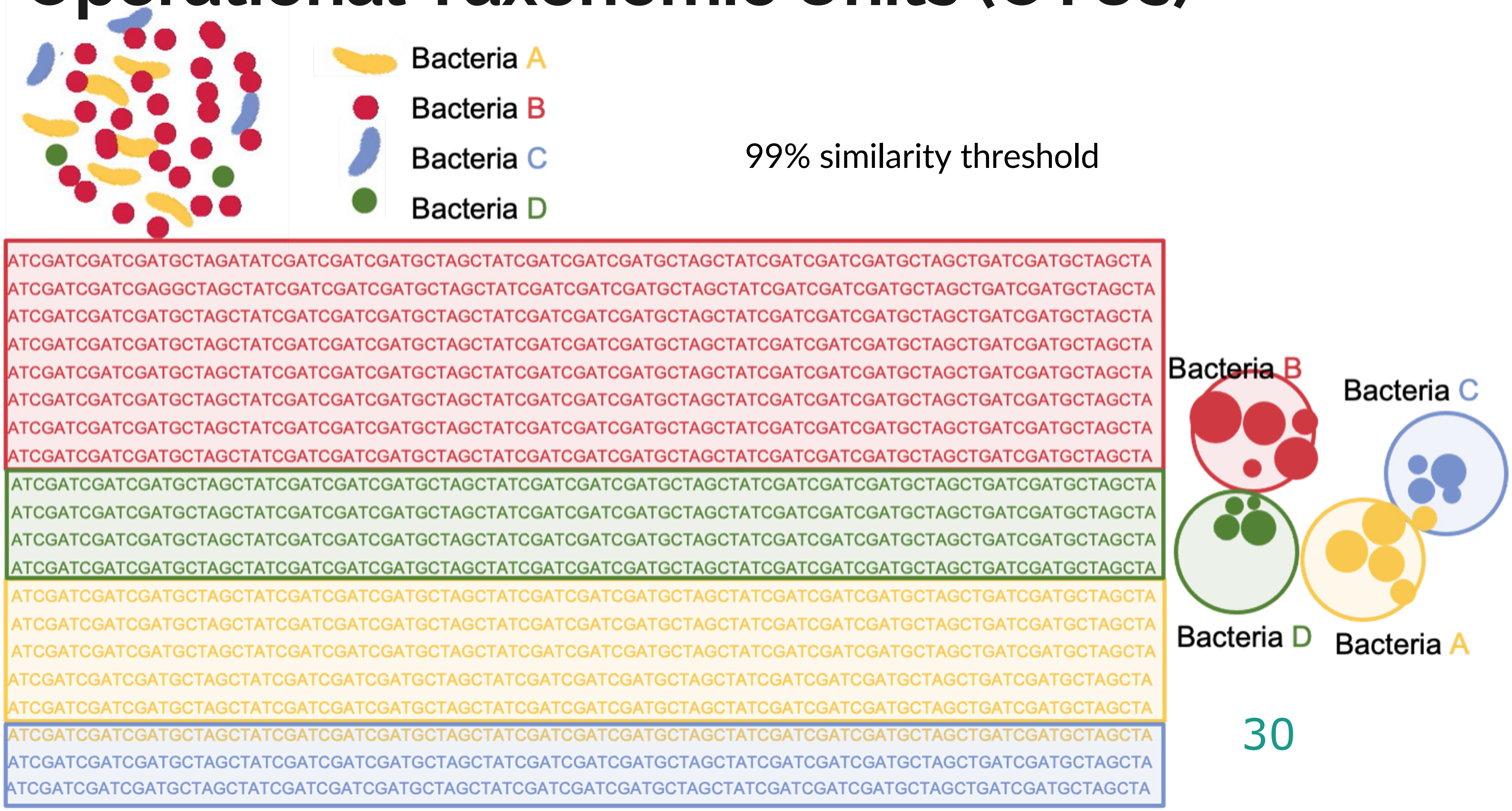
97% Similarity threshold

ATCGATCGATCGATGCTAGATATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGAGGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA

ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA
ATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTATCGATCGATCGATGCTAGCTGATCGATGCTAGCTA



Operational Taxonomic Units (OTUs)



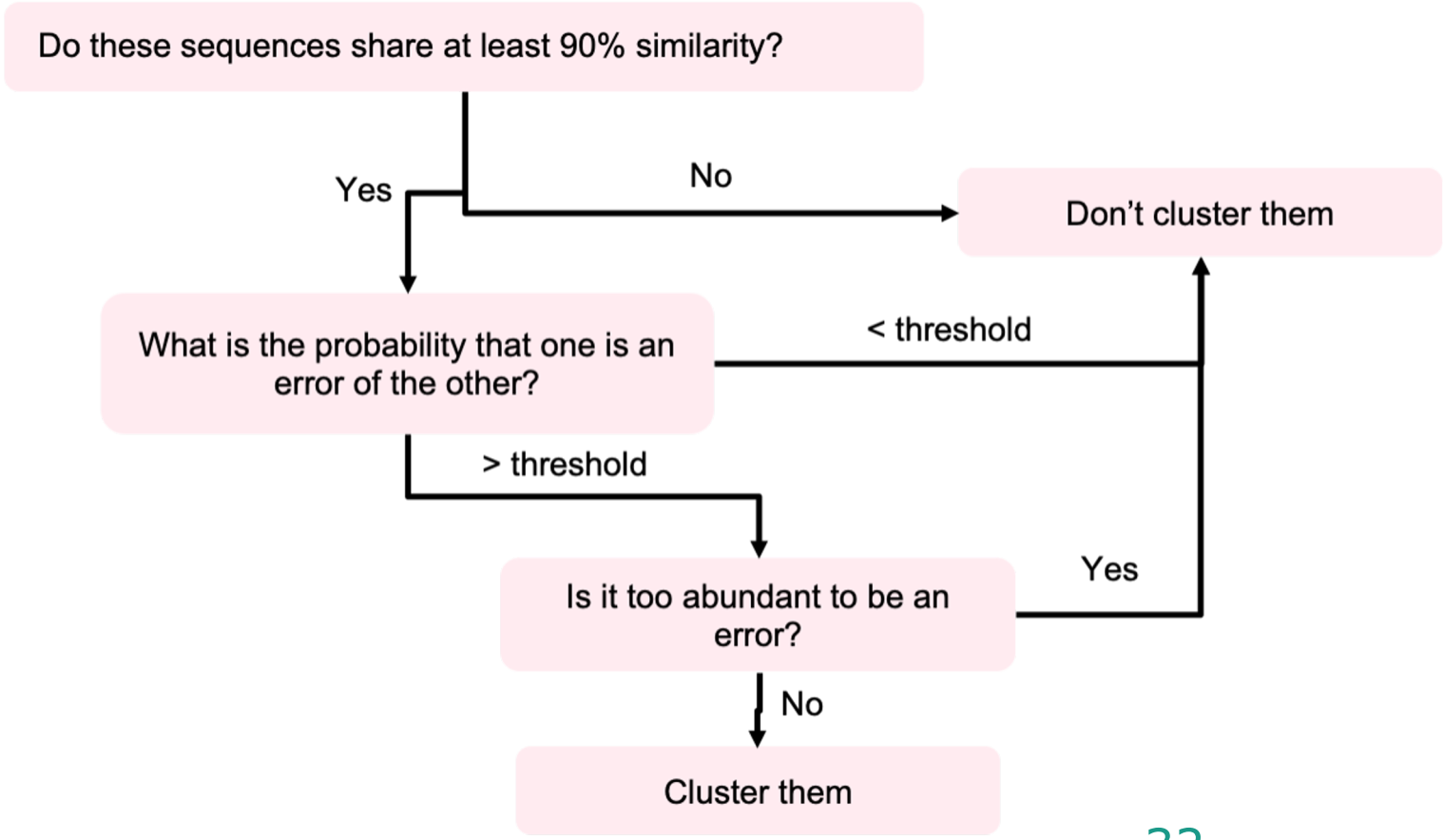
OTUs are a blunt instrument

- Grouping at arbitrary thresholds soaks up a lot of errors, but also loses a lot of information

- An alternative is calculating Amplicon Sequence Variants* that comprise “correct” sequences plus similar sequences with a high probability of sequencing errors

* or zero-radius OTUs, sub-OTUs

Inferring ASVs with DADA2

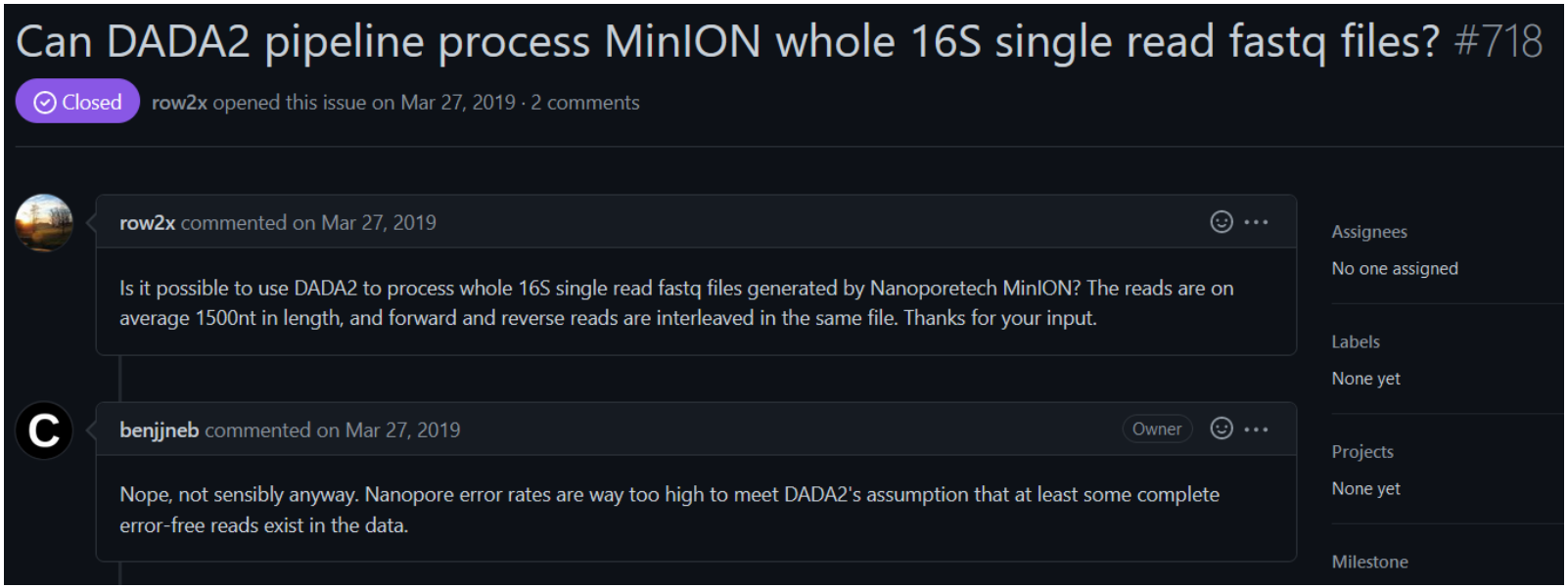


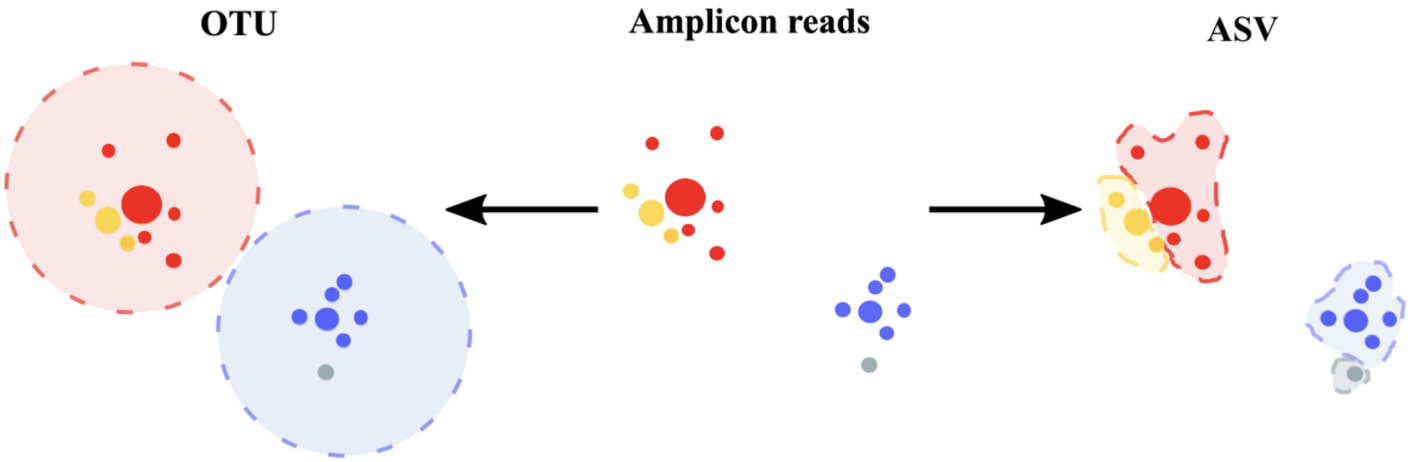
Callahan et al. (2016) *Nat Meth*

DADA2 1.10 RELEASE NOTES

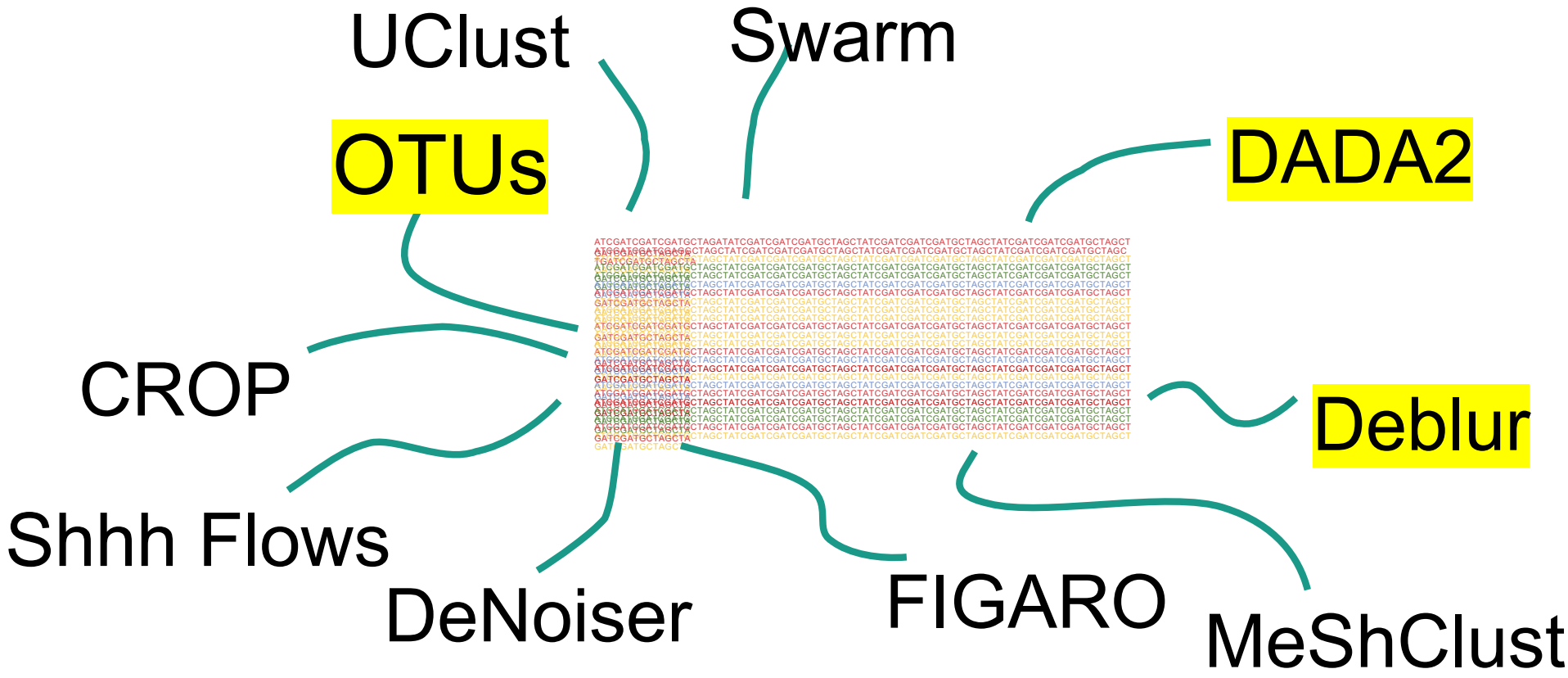
NEW FEATURES

- PacBio CCS reads up to 3 kilobases are now supported. See also `PacBioErrfun`, the new and recommended error-estimation function for PacBio CCS data. The preprint introducing DADA2's long-read functionality has information on accuracy and sub-species resolution, and the associated reproducible analyses show PacBio-specific workflows.



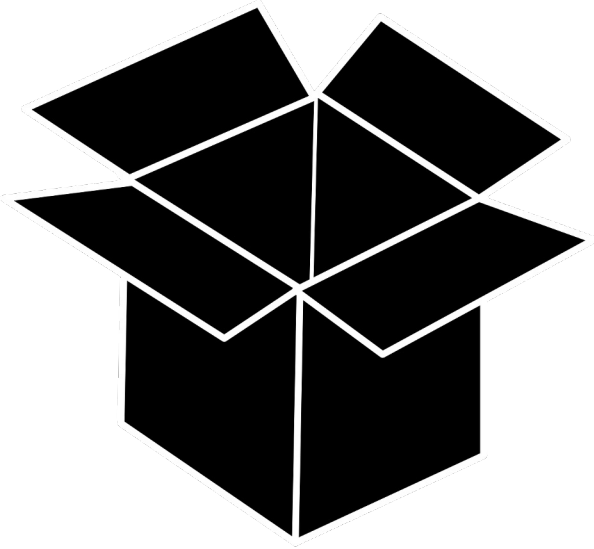


- Alternatives
  - Deblur, UNOISE
  - All draw on notions of errors and relative abundance
  - Comparison: Nearing et al. (2018) *PeerJ*


UClust
Swarm
OTUs
DADA2
CROP
Deblur
Shhh Flows
DeNoiser
FIGARO
MeShClust

How can I safely cluster my data?

- Try different methods,
- Understand them,
- Compare your results,
- Be able to defend your method



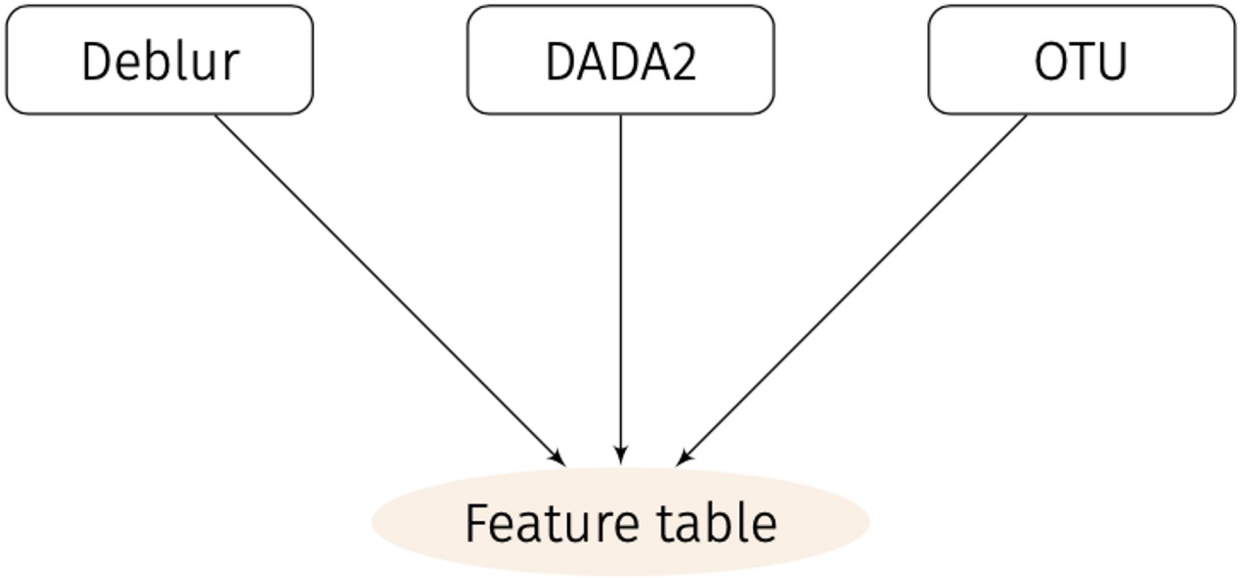


Table 1: Feature table

	Sample 1	Sample2	Sample 3	...
Microbe 1				
Microbe 2		Count		
...				

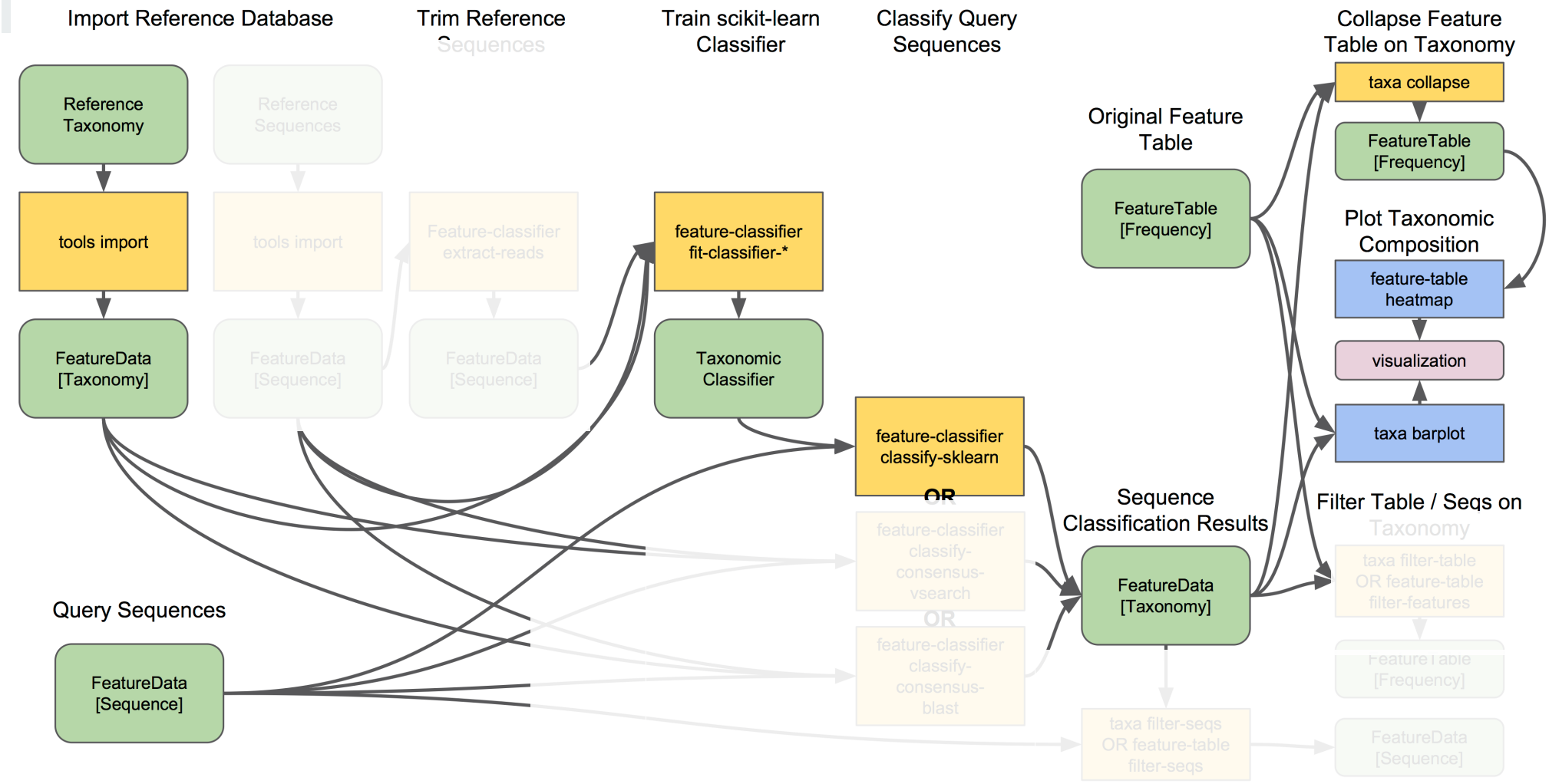
Taxonomic classification

38

From the feature table…

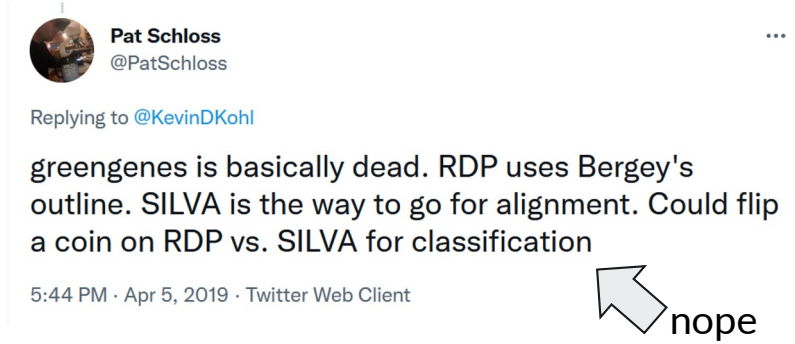
- Assign taxonomy to ASVs

feature_id	Taxon
00071b5a03322beb265e12561e62b7d2	d__Bacteria; p__Actinobacteriota; c__Acidimicrobiia; o__Actinomarinales; f__Actinomarinaceae; g__Candidatus_Actinoma
00071b5a03322beb265e12561e62b7d2	d__Bacteria; p__Actinobacteriota; c__Acidimicrobiia; o__Actinomarinales; f__Actinomarinaceae; g__Candidatus_Actinoma
0007c2bd79c09a2c5074f744573464e9	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales; f__Pseudohongiellaceae; g__Pseudoho
0007c2bd79c09a2c5074f744573464e9	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales; f__Pseudohongiellaceae; g__Pseudoho
0031efc0a916e54571c6ae4ddddbbd23	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Oceanospirillales; f__Pseudohongiellaceae; g__Pseudoho
0032e6d9726df6cf4176f545ca56623f	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Puniceispirillales; f__SAR116_clade; g__Candidatus_Punicei
0032e6d9726df6cf4176f545ca56623f	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Puniceispirillales; f__SAR116_clade; g__Candidatus_Punicei
003a1d54a176434c88b3e6488631bbe2	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria
003a1d54a176434c88b3e6488631bbe2	d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria


Import Reference Database
Trim Reference Sequences
Train scikit-learn Classifier
Classify Query Sequences
Collapse Feature Table on Taxonomy
Reference Taxonomy
Reference Sequences
tools import
tools import
Feature-classifier extract-reads
feature-classifier fit-classifier-*
FeatureData [Taxonomy]
FeatureData [Sequence]
FeatureData [Sequence]
Taxonomic Classifier
Original Feature Table
FeatureTable [Frequency]
taxa collapse
FeatureTable [Frequency]
Plot Taxonomic Composition
feature-table heatmap
visualization
taxa barplot
feature-classifier classify-sklearn
OR
feature-classifier classify-consensus-vsearch
OR
feature-classifier classify-consensus-blast
Sequence Classification Results
FeatureData [Taxonomy]
Filter Table / Seqs on Taxonomy
taxa filter-table OR feature-table filter-features
FeatureTable [Frequency]
Query Sequences
FeatureData [Sequence]
taxa filter-seqs OR feature-table filter-seqs
FeatureData [Sequence]

Data resources

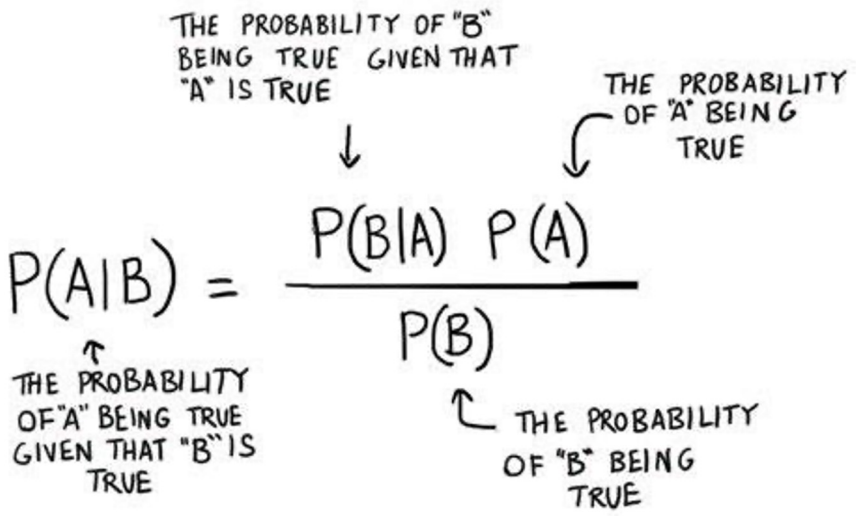
- Reference databases:
  - SILVA (updated 2021) – classifier in QIIME2
  - Greengenes (updated 2013) – classifier in QIIME2
  - RDP (updated 2016; taxonomy updated 2020)
  - EZBioCloud (updated 2021) – offers taxonomic profiling
  - NCBI (updated continuously)



RESCRIPt: a tool for customized database retrieval and management:
https://www.biorxiv.org/content/10.1101/2020.10.05.326504v1.full

Taxonomic inference

- BLAST assigns taxonomic label of closest match
- Supervised learning algorithm : Naïve Bayes classifier



Train your own, or use a pre-trained classifier

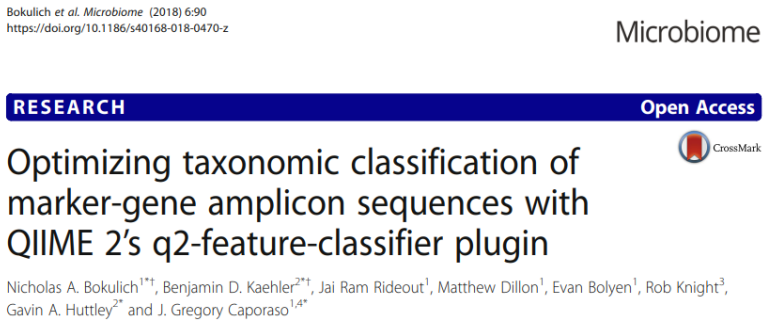
- The classifier's accuracy increases if the classifier is trained on the region of the rRNA that was amplified (eg. V3-V4 or V5-V6 of the 16S rRNA)
- But! If you use common marker genes (16S, 18S, ITS) There are a lot of pre-trained classifiers online trained on the full length gene, or on specific primers

Naive Bayes classifiers trained on:

- Silva 138 99% OTUs full-length sequences (MD5: `b8609f23e9b17bd4a1321a8971303310`)
- Silva 138 99% OTUs from 515F/806R region of sequences (MD5: `e05afad0fe87542704be96ff483824d4`)
- Greengenes 13_8 99% OTUs full-length sequences (MD5: `6bbc9b3f2f9b51d663063a7979dd95f1`)
- Greengenes 13_8 99% OTUs from 515F/806R region of sequences (MD5: `9e82e8969303b3a86ac941ceafeeac86`)

Filtering on “confidence”

- Each taxonomic classification from the feature table comes with an associated “confidence” score between 0 and 1.0
- In the default taxonomic assignment scheme, this is a posterior probability – don’t confuse “confidence” with any formal statistical notion such as “confidence interval”
- Anything with confidence below a threshold (default = 0.7) is not classified
  - Not classified at species level? Try classifying at genus instead, etc.



Bokulich *et al. Microbiome* (2018) 6:90
https://doi.org/10.1186/s40168-018-0470-z

Microbiome

RESEARCH Open Access

Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin

Nicholas A. Bokulich[1*†], Benjamin D. Kaehler[2*†], Jai Ram Rideout[1], Matthew Dillon[1], Evan Bolyen[1], Rob Knight[3], Gavin A. Huttley[2*] and J. Gregory Caporaso[1,4*]

So?

- Be aware that there are many tools, many classifiers, many databases!
- Different feature tables can potentially give you very different outcomes in taxonomy, diversity, statistics, and ML


End of
Part II