MIMM4750G

Metabarcoding

What's metabarcoding?

- So far we've talked about applications of NGS that focus on a single species (bacteria or virus).
- The massive throughput of NGS platforms allows us to target multiple species, or perform untargeted sequencing.
- A key challenge to untargeted sequencing is amplification.
- Amplification requires primers, but primers are usually specific to a particular sequence.

Untargeted sequencing

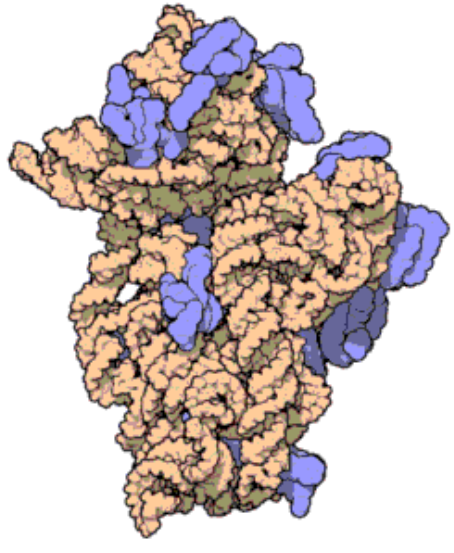
- If we can't use template-specific primers, what *can* we do?
- Find a conserved region — shared across all, if not most — species, for amplification: but the region in between should be non-conserved!
- Random priming - suppliers will generate short oligos of arbitrary sequence.
- Both amplification and priming can lead to biases.

Metabarcoding

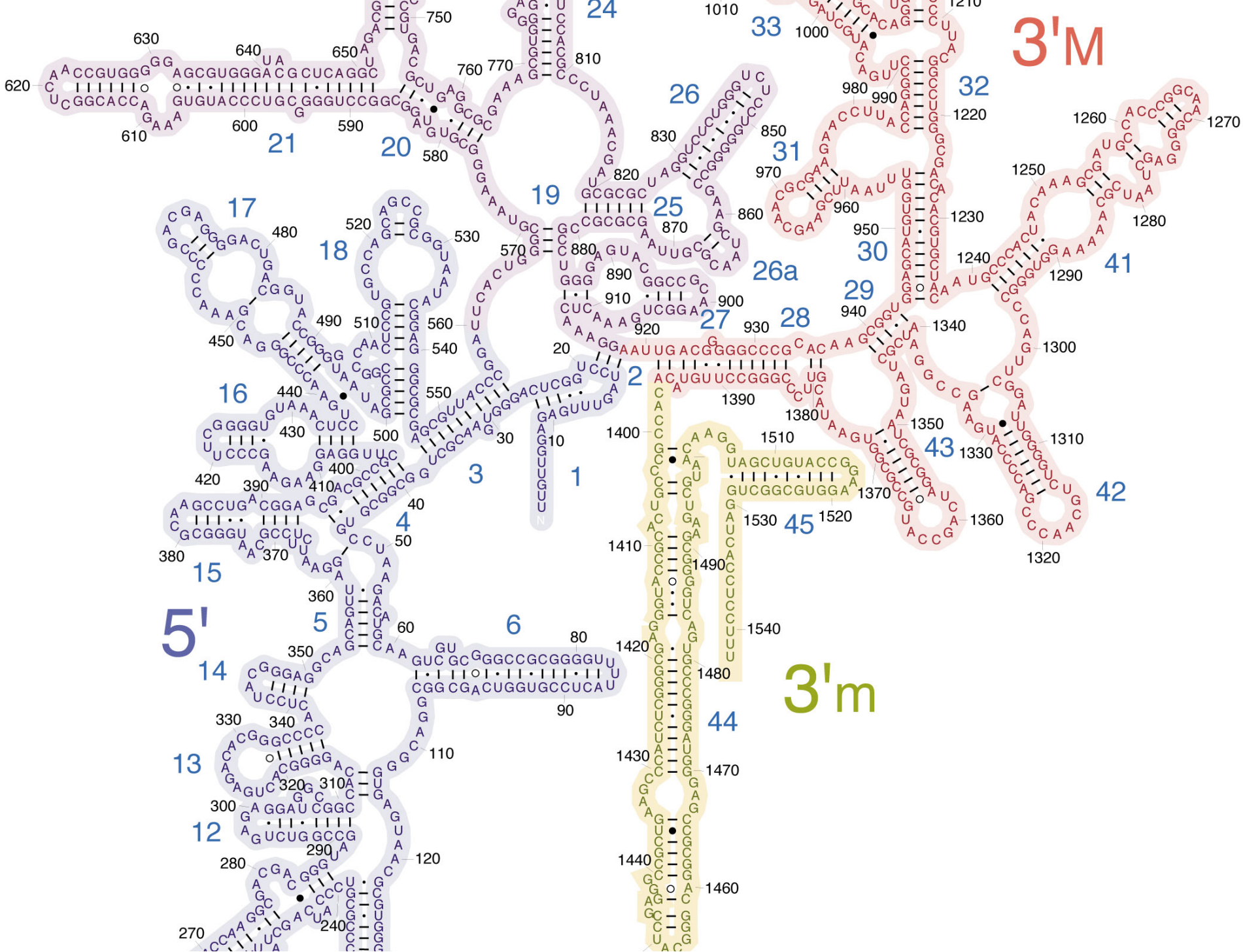
- Not the same as "barcoding" as a step in next-generation sequencing to identify templates from different samples.
- *AKA* targeted metagenomics, amplicon-based metagenomics, and metagenetics.
- A genetic marker (short region of the genome) that should uniquely identify a species.
- It should be universal: found in all organisms in a large taxonomic group.
- There should be alternating regions of high and low sequence conservation.

16S RNA

- 16S ribosomal RNA is a component of 30S small subunit of prokaryotic ribosome
- Use of 16S for phylogenetics proposed by Carl Woese and George Fox in 1977.
- Presently the gold standard for metabarcoding studies of bacterial microbiomes.



3d animation of 30S small ribosomal subunit (PDB 1FKA).


5'
3'M
3'm
1
2
3
4
5
6
10
20
30
40
50
60
80
90
110
120
240
270
290
300
310
320
330
340
350
360
370
380
390
400
410
420
430
440
450
490
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
750
760
770
810
820
830
850
860
870
880
890
900
910
920
930
940
960
970
980
990
1000
1010
1210
1220
1230
1240
1250
1260
1270
1280
1290
1300
1310
1320
1330
1340
1350
1360
1370
1380
1390
1400
1410
1420
1430
1440
1460
1470
1480
1490
1510
1520
1530
1540
11
12
13
14
15
16
17
18
19
20
21
24
25
26
26a
27
28
29
30
31
32
33
41
42
43
44
45

Variable regions of 16S

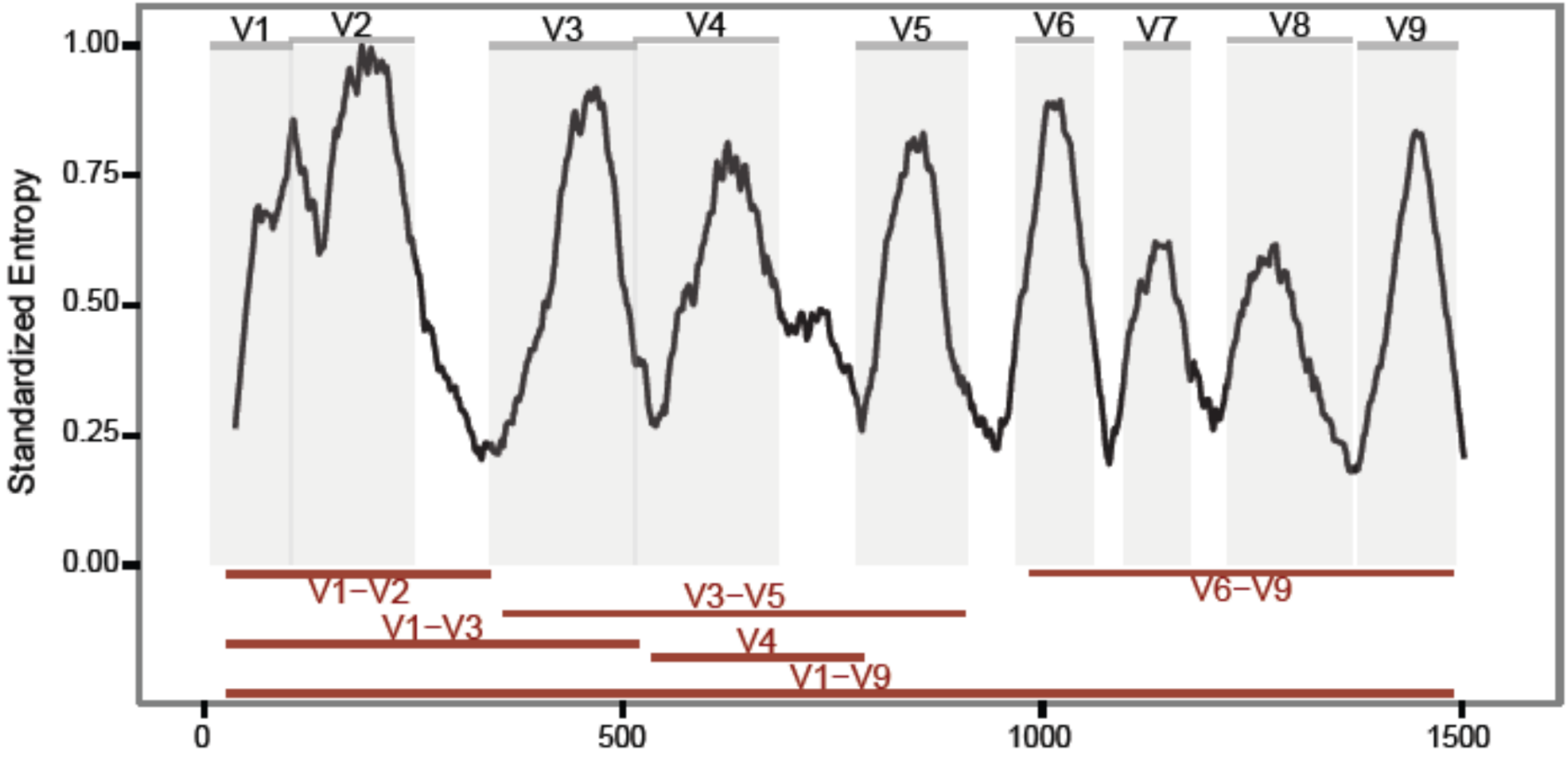


Image credit: Jethro Johnson, https://www.pacb.com/wp-content/uploads/Jethro-Johnson-PacBio-East-Coast-UGM-2017.pdf

Taxonomic binning

- With the sequence barcode, you now need to look up the species (or higher taxonomic level).
- Not all 16S sequences in Genbank are classified.
- Some sequences are chimeric (PCR artefacts that combine two or more templates).
- Multiple databases maintained by different groups might not agree on the same sequence.

Taxonomic databases

Database	URL	# Records	Last update
NCBI RefSeq	https://www.ncbi.nlm.nih.gov/refseq/	21,073	2019-03-12
SILVA	https://www.arb-silva.de/	1,928,733	2017-12
RDP	https://rdp.cme.msu.edu/	3,356,809	2016-09
Greengenes	http://greengenes.lbl.gov	1,262,986	2015-05

Who do you believe?

> *I found 249,490 identical sequences with conflicting annotations in SILVA v128 and Greengenes v13.5 at ranks up to phylum (7,804 conflicts), indicating that the annotation error rate in these databases is ~17%.*

Robert C. Edgar (developer of *MUSCLE*), 2018. *Taxonomy annotation and guide tree errors in 16S rRNA databases*, PeerJ: 5030.

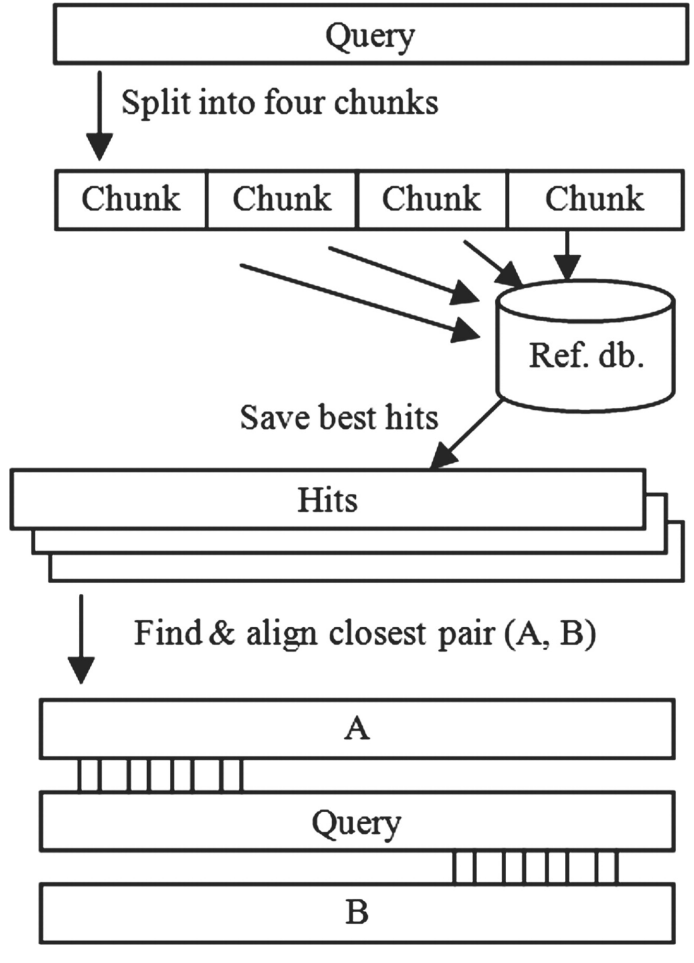
- Some sequences actually from organelles (*e.g.*, mitochondria) or eukaryotes (*e.g.*, protists; Lesack and Birol 2018).

Denoising

- DADA2 uses a probability model to assign reads to cluster given the sequencing error rate.
    1. For every pair of sequences, check k-mer distance
    2. If close enough, do sequence alignment.
    3. Calibrates error model for different nucleotide substitutions, given quality score.
    4. Calculates probability of read abundance given error model.
- See also Deblur

Chimera detection

- UCHIME, Robert Edgar
- Splits query sequence into four parts
- Each part is mapped to a reference database assumed to be chimera-free.
- Retrieves two potential "parent" sequences.



OTU clustering

- Operational taxonomic units
- Separate within-species variation from among-species variation.
- 97% sequence identity clustering has become standard.

Taxonomic assignment (binning)

- BLAST - remember BLAST?
- USEARCH (R.Edgar again) - highly optimized k-mer search for a small number of good hits.
- Closed source, feature-limited distribution.
- Measure the intersection for the sets of all 8-mers in query sequence and database seqeunce:

$$U_i = |W(Q) \cap W(T_i)|$$

Naive Bayes classifier

- RDP Classifier
- Open source, implemented in Java
- Also extracts 8-mers from query and database sequences.
- The probability that a member of genus G contains word w_i is

$$P(w_i|G) = \frac{n_G(w_i) + P(w_i)}{N_G + 1}$$

where n_G is the number of sequences in G with word w_i, N is the total size of G, and $P(w_i)$ is the prior probability (the overall frequency of w_i).

Phylogenetic placement

- Place a new sequence in a fixed reference tree
- Much faster than regenerating the entire tree!
- Uses a model of evolution, better measure of distance.
- Used to estimate that about 1% of 16S rRNA databases are misclassified.
- pplacer
- EPA - part of RAxML

Pipelines

- A pipeline or workflow is a collection of programs and scripts used to process data in a fixed sequence.
- Pipelines for metabarcoding have become extremely popular.
- mothur - open source C++, all in-house programs
- QIIME - Python scripts that wrap around a collection of third-party programs.

Further reading

- Taxonomy annotation and guide tree errors in 16S rRNA databases
- Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era
- Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches