DTU

DTU Health Technology
Bioinformatics

Metagenomics and 16S

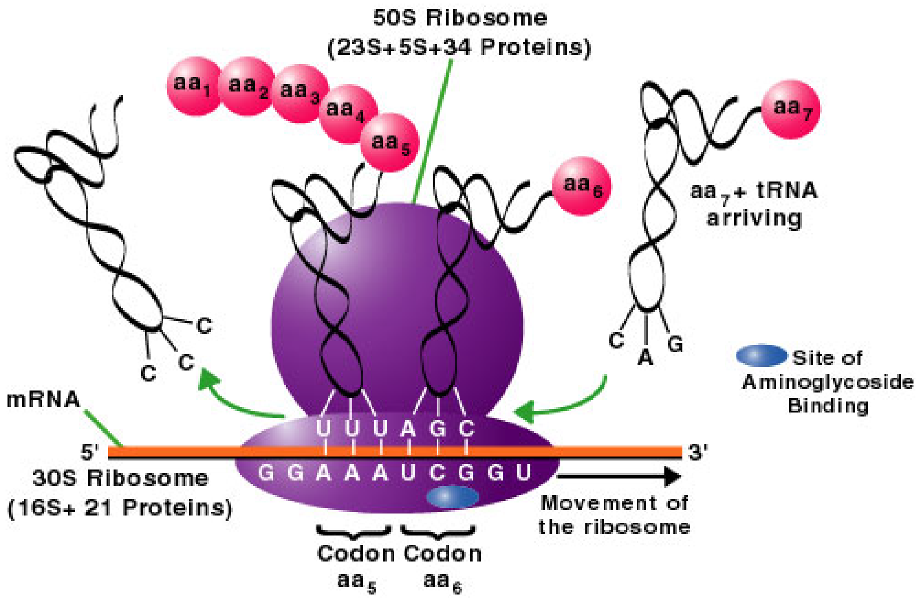
Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

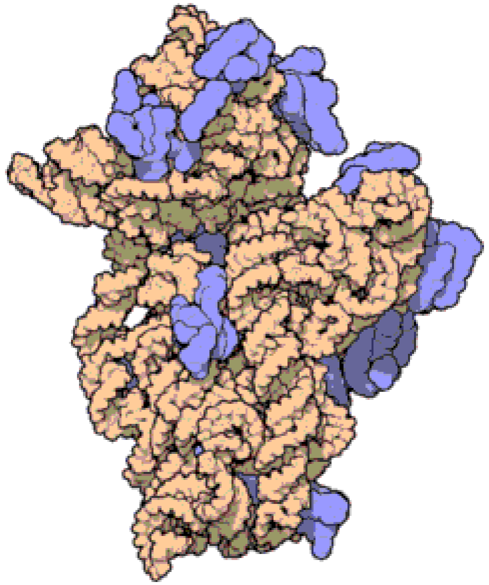
Menu

- 16s rRNA
- Amplicon sequencing
- Quality control
- Clustering
- Taxonomic annotation
- OTU table

The 16S rRNA gene

- Part of the ribosome
- Translation mRNA to proteins
- 16S rRNA is an integral part of the 30S subunit
- Very conserved

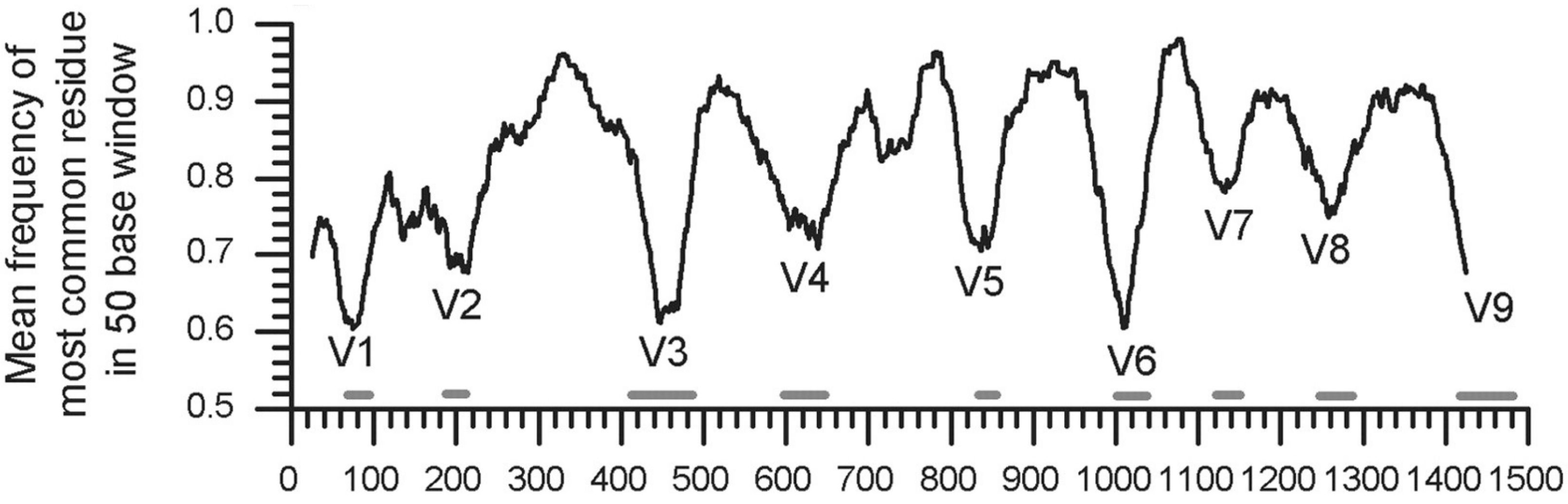


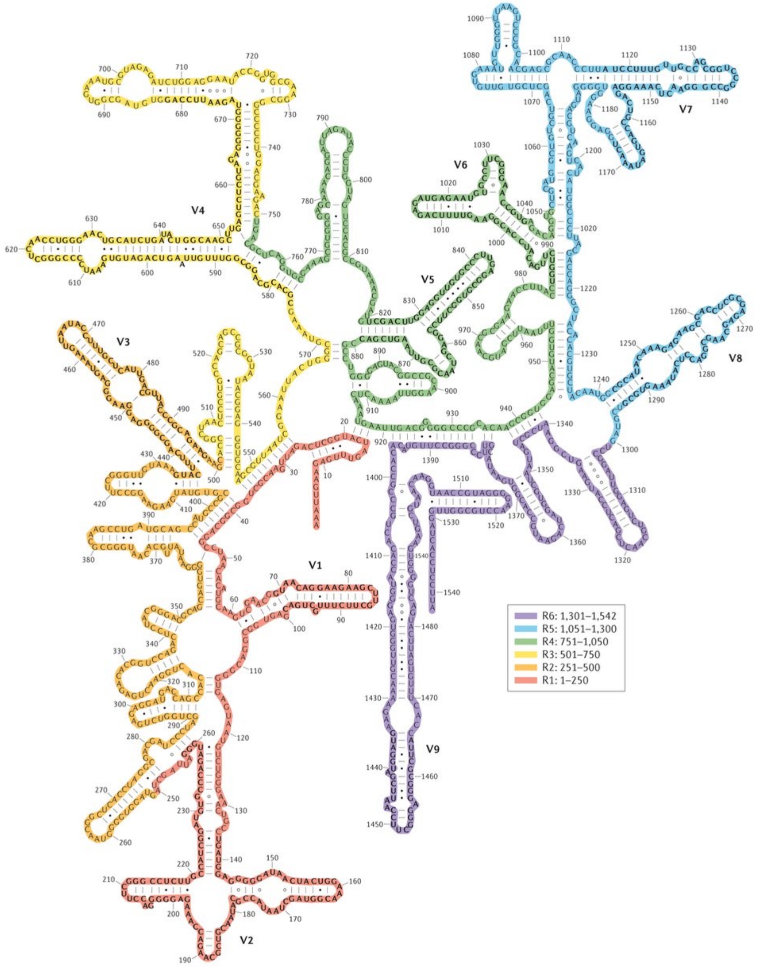


16S rRNA is the orange part of this 30s ribosomal subunit

Conserved and then again

- The 16S rRNA gene
- Coding for RNA – non-protein coding
- Conserved regions
- 9 variable regions

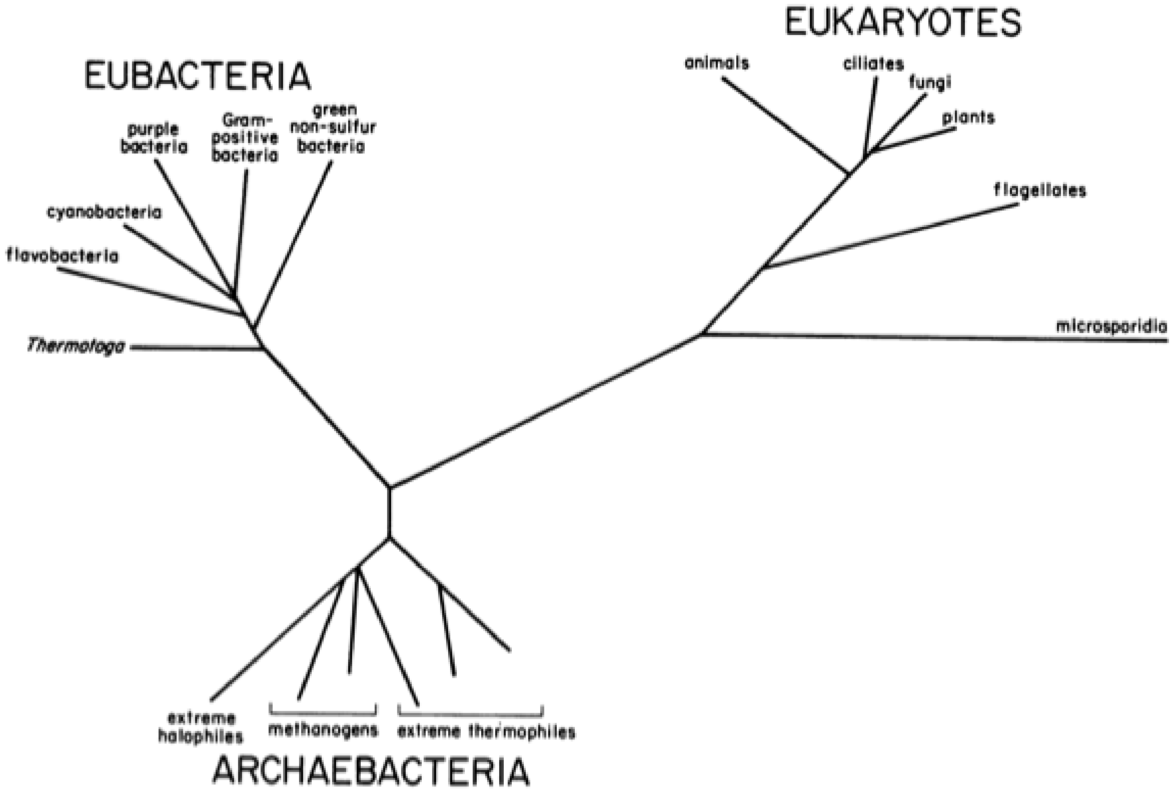


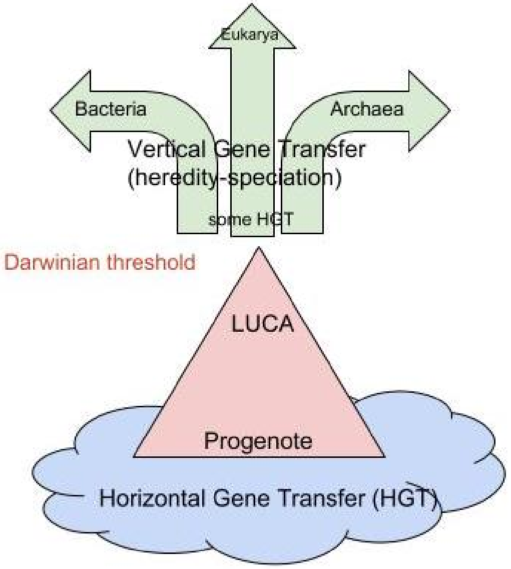


Nature Reviews | Microbiology

Carl Woese

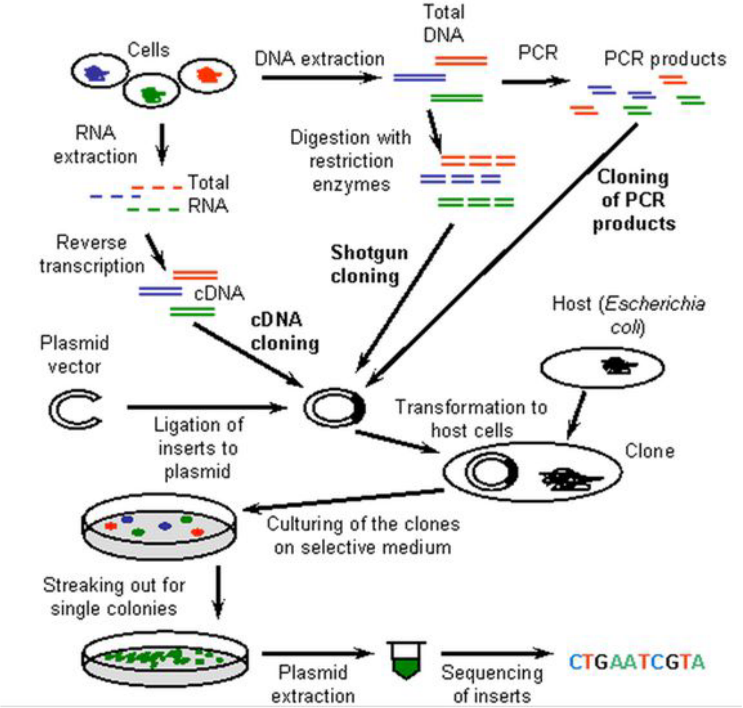
- Defined Archaea using 16s rRNA
- Invented the concept of a pre-Darwinian threshold





16S rRNA gene amplicon sequencing

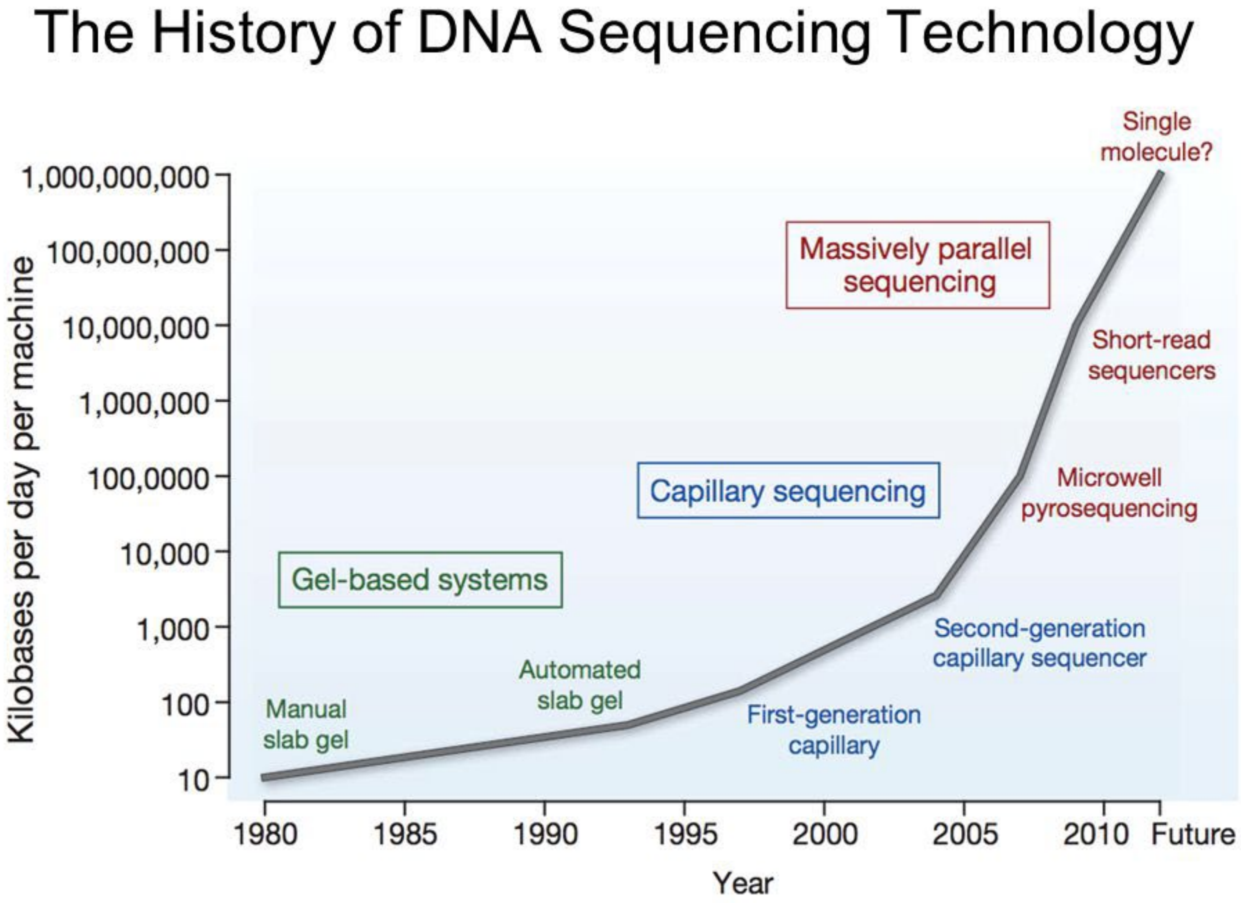
- Norman R. Pace in 1985
- Highly sensitive!



Strategies and steps in cloning.

16S rRNA amplicon sequencing and NGS

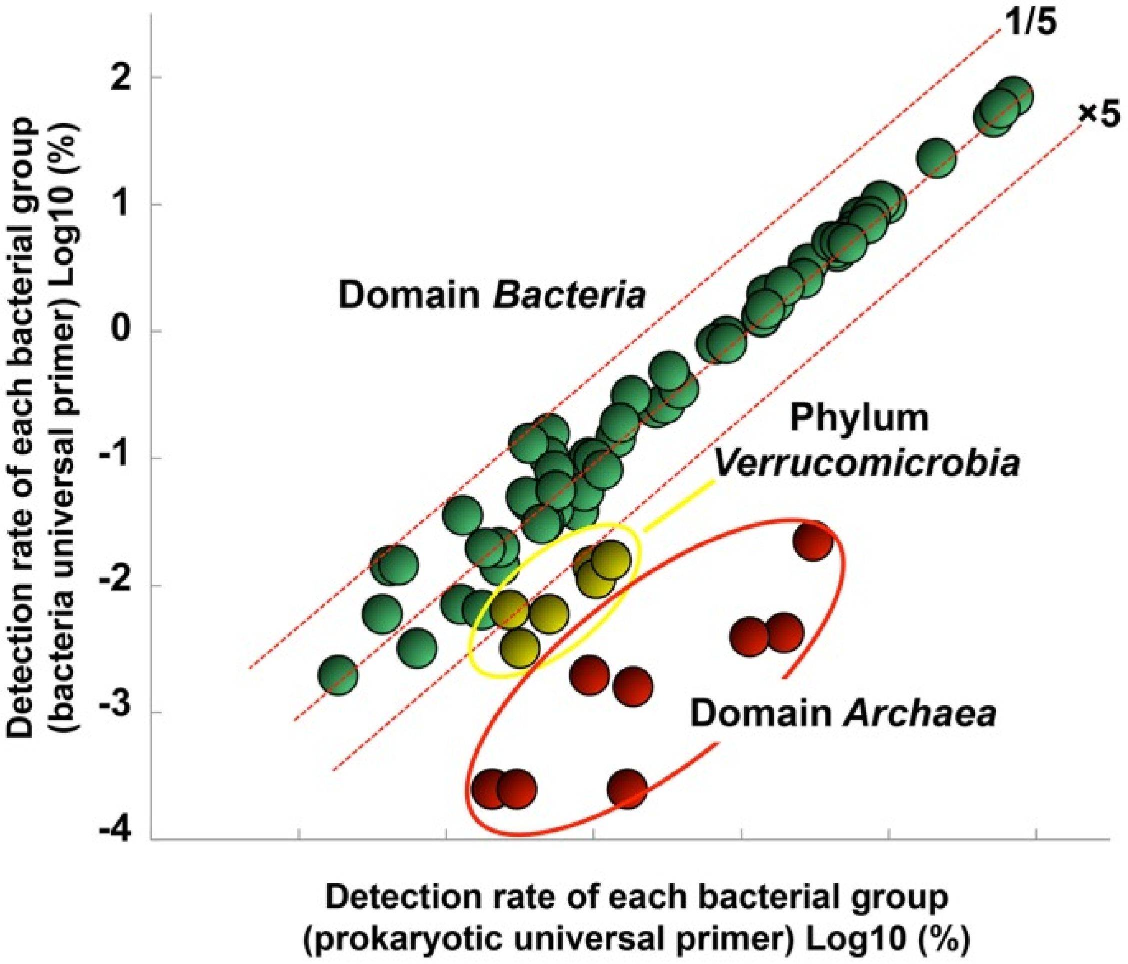
- Short read-length
- We cannot amplify whole 16S rRNA



Primers matter!

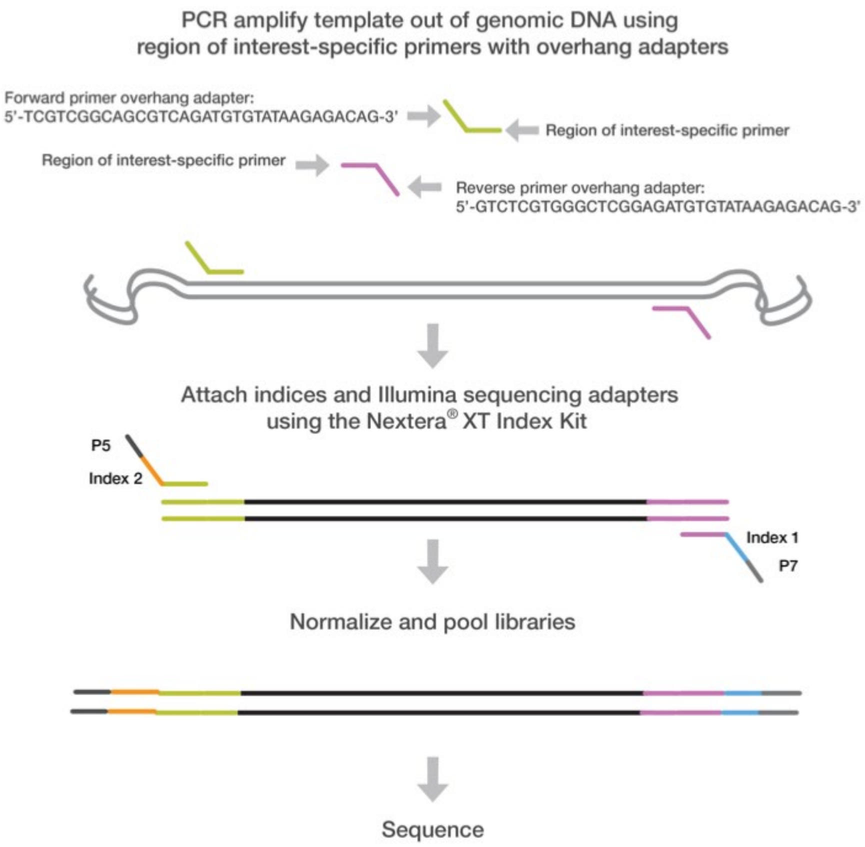
- Nothing is 100% conserved
- Primer design will affect observations
- Stick to community standard?

  OR
- Utilize latest technology?

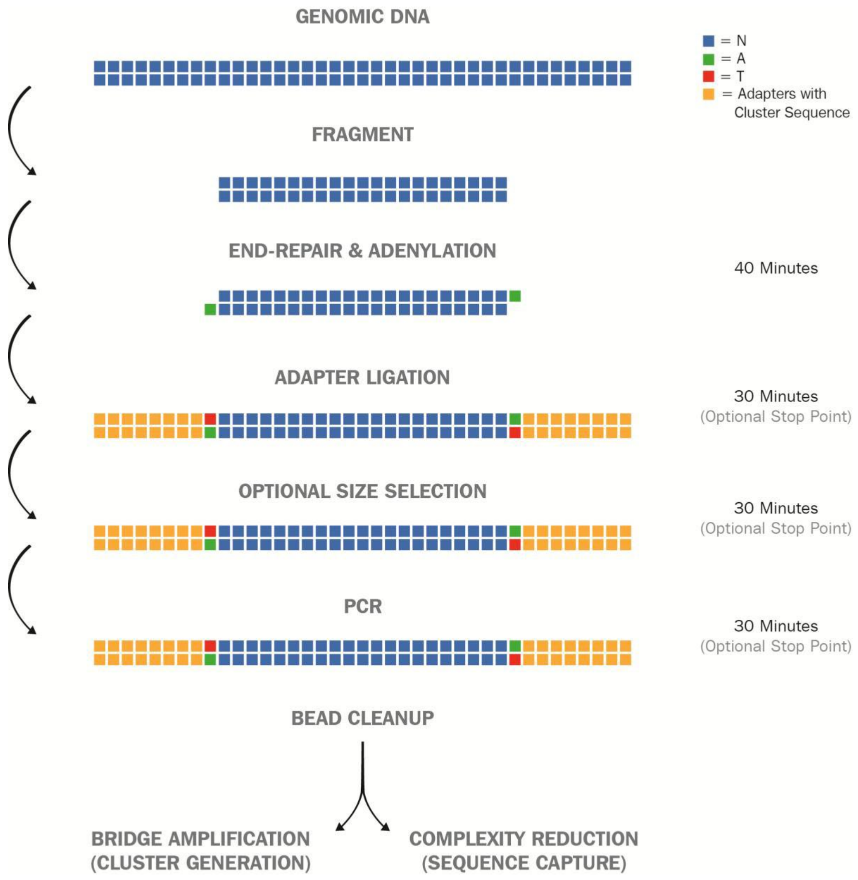


Laboratory protocol is shake & bake

16S rRNA gene amplicons

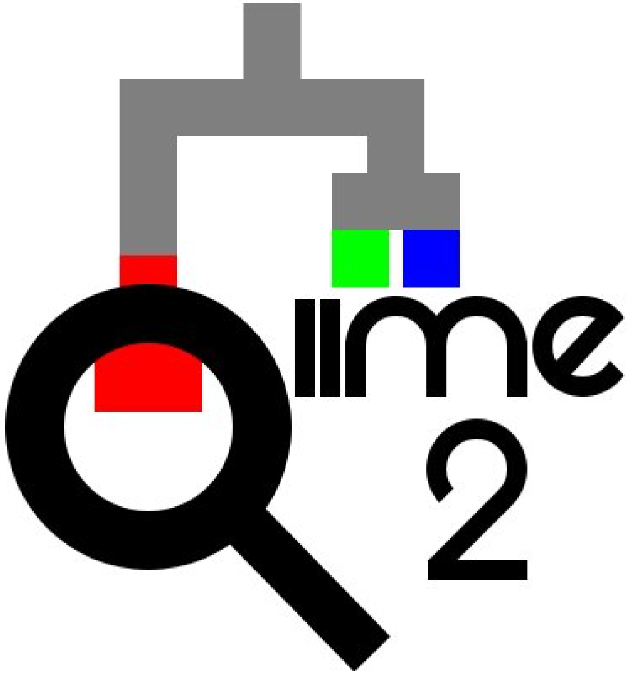

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters
Forward primer overhang adapter:
5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3'
Region of interest-specific primer
Region of interest-specific primer
Reverse primer overhang adapter:
5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'
Attach indices and Illumina sequencing adapters using the Nextera® XT Index Kit
P5
Index 2
Index 1
P7
Normalize and pool libraries
Sequence

Whole metagenome sequencing

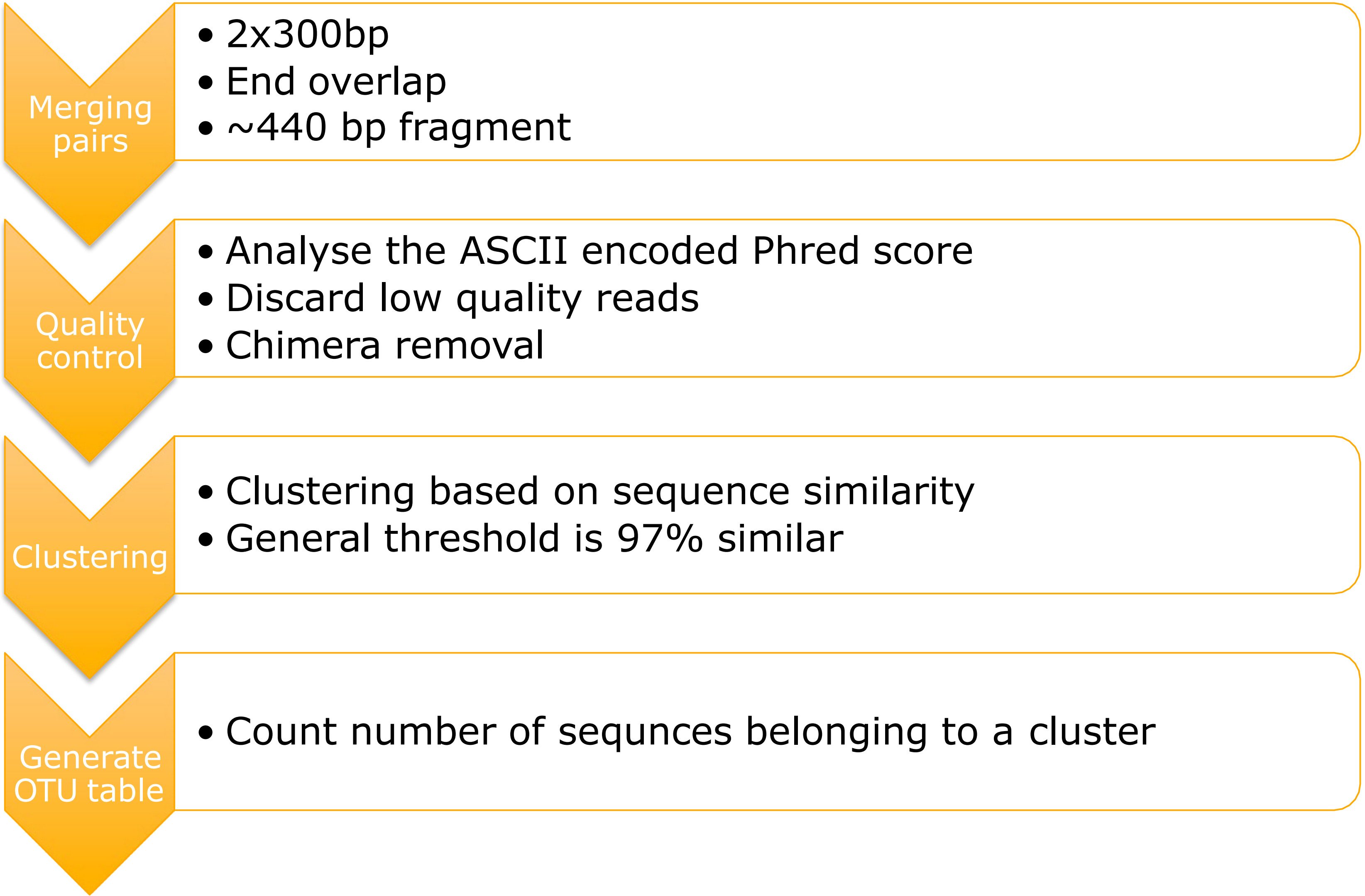


Several different analysis pipelines

- Qiime 2 is easy to use, state-of-the-art and has a large community
- UPARSE is also very popular
- Mothur use to be popular but lacks denoising (we will get back to that)



Basic data analysis



Merging pairs

- 2x300bp
- End overlap
- ~440 bp fragment

Quality control

- Analyse the ASCII encoded Phred score
- Discard low quality reads
- Chimera removal

Clustering

- Clustering based on sequence similarity
- General threshold is 97% similar

Generate OTU table

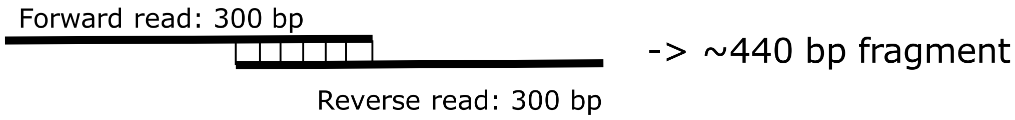
- Count number of sequnces belonging to a cluster

Merging pairs

- This way we utilize all information

Merging pairs

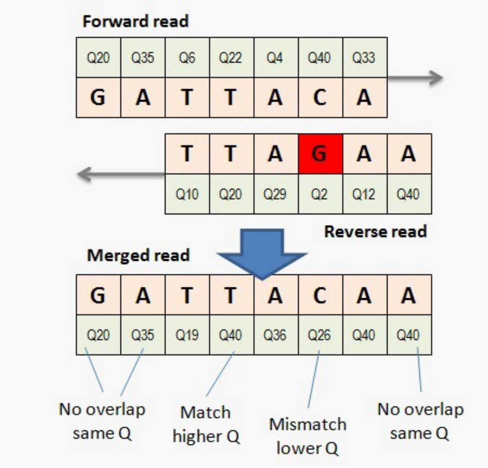
- 2x300bp
- End overlap
- ~440 bp fragment



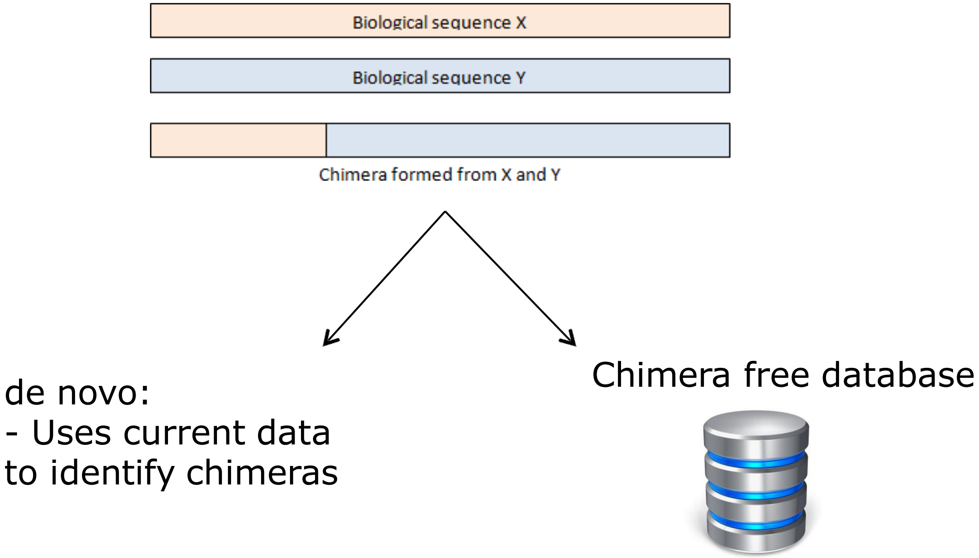
Forward read: 300 bp

-> ~440 bp fragment

Reverse read: 300 bp



Chimeric sequences



Biological sequence X

Biological sequence Y

Chimera formed from X and Y

de novo:
- Uses current data
to identify chimeras

Chimera free database

Even low error levels leads to problems

Expected error rate:

- Q=2 -> 63% error probability (Q = -10*log(p))
- Q=20 -> 1% error probability
- Sum of errors across read = error probability

Very simple example:
100bp read where all bp Q=20

0.01 + 0.01 …. + 0.01 = 1

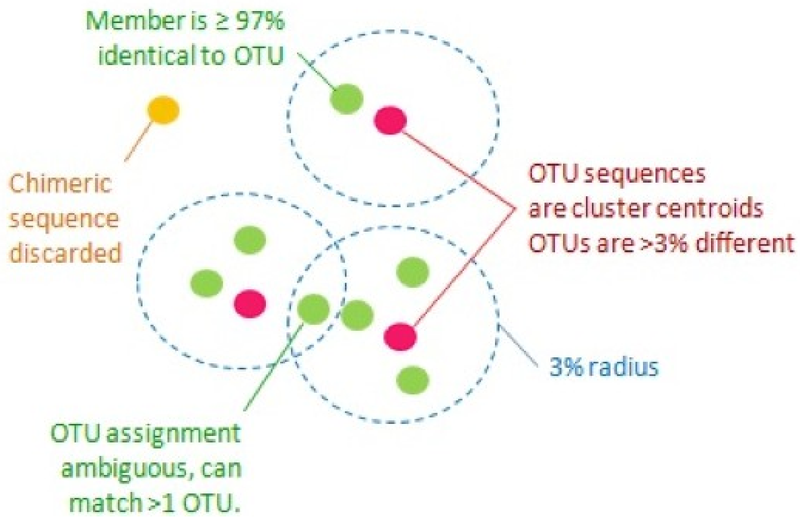
Clustering

Clustering

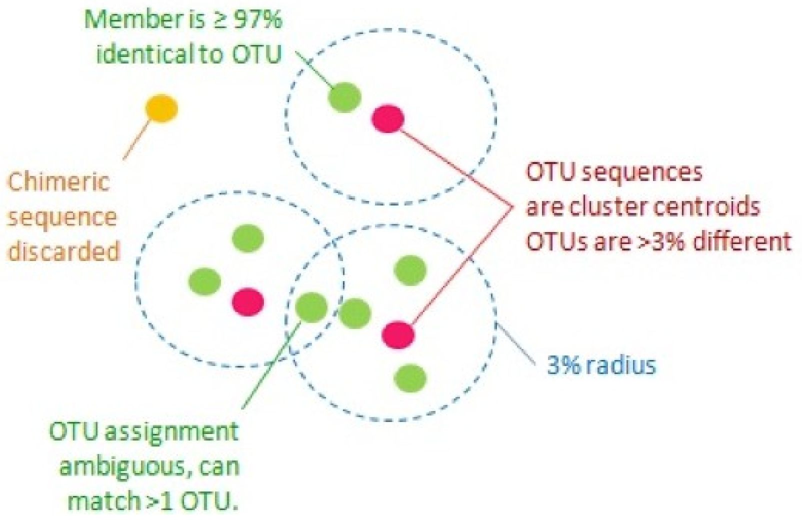
- Clustering based on sequence similarity
- General threshold is 97% similar

Sequence similarity

```
AAAGGAAAAACTCCGCCACCTTGAAGACGGCTGCCGCCAAAACGGCCTCGCCATGACCGTCCAGCGCCGCGTCGTCCTGGACGCCCTTGCGG
AGAATCAAAGCTTCAGGCCCTCGAAGCGGGGTGCCGCAAAAACGGGTTCGCCATGACCGTCCAGCGTCGGGTCATCATGGAGGCACTGGCGG
* *   *** ** *    *** ****  ** ****** *******  ******************* ** *** ** **** ** ** ****
```



OTU: Operational Taxonomic Unit



Member is ≥ 97% identical to OTU
Chimeric sequence discarded
OTU sequences are cluster centroids OTUs are >3% different
3% radius
OTU assignment ambiguous, can match >1 OTU.

OTU: Operational Taxonomic Unit

- Approach a cluster of species including subspecies

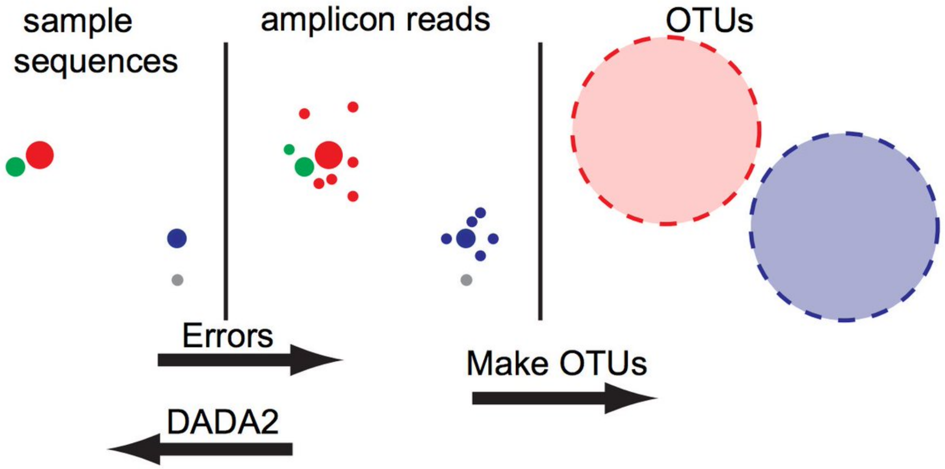
Take into account

- Might contain multiple species, with > 97% similarity on 16S rRNA gene level
- Species split due to subspecies with

  < 97% similarity
- Artifacts created by read errors and chimeras

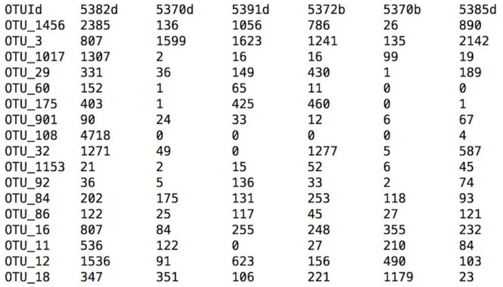
Denoising

- Identify more real variants with high resolution
- CPU intensive
- DADA2 and Deblur
- Produces ASVs or amplicon sequence variants



Callahan, et al. Nature Methods, 2016.

OTU table



OTUId	5382d	5370d	5391d	5372b	5370b	5385d
OTU_1456	2385	136	1056	786	26	890
OTU_3	807	1599	1623	1241	135	2142
OTU_1017	1307	2	16	16	99	19
OTU_29	331	36	149	430	1	189
OTU_60	152	1	65	11	0	0
OTU_175	403	1	425	460	0	1
OTU_901	90	24	33	12	6	67
OTU_108	4718	0	0	0	0	4
OTU_32	1271	49	0	1277	5	587
OTU_1153	21	2	15	52	6	45
OTU_92	36	5	136	33	2	74
OTU_84	202	175	131	253	118	93
OTU_86	122	25	117	45	27	121
OTU_16	807	84	255	248	355	232
OTU_11	536	122	0	27	210	84
OTU_12	1536	91	623	156	490	103
OTU_18	347	351	106	221	1179	23

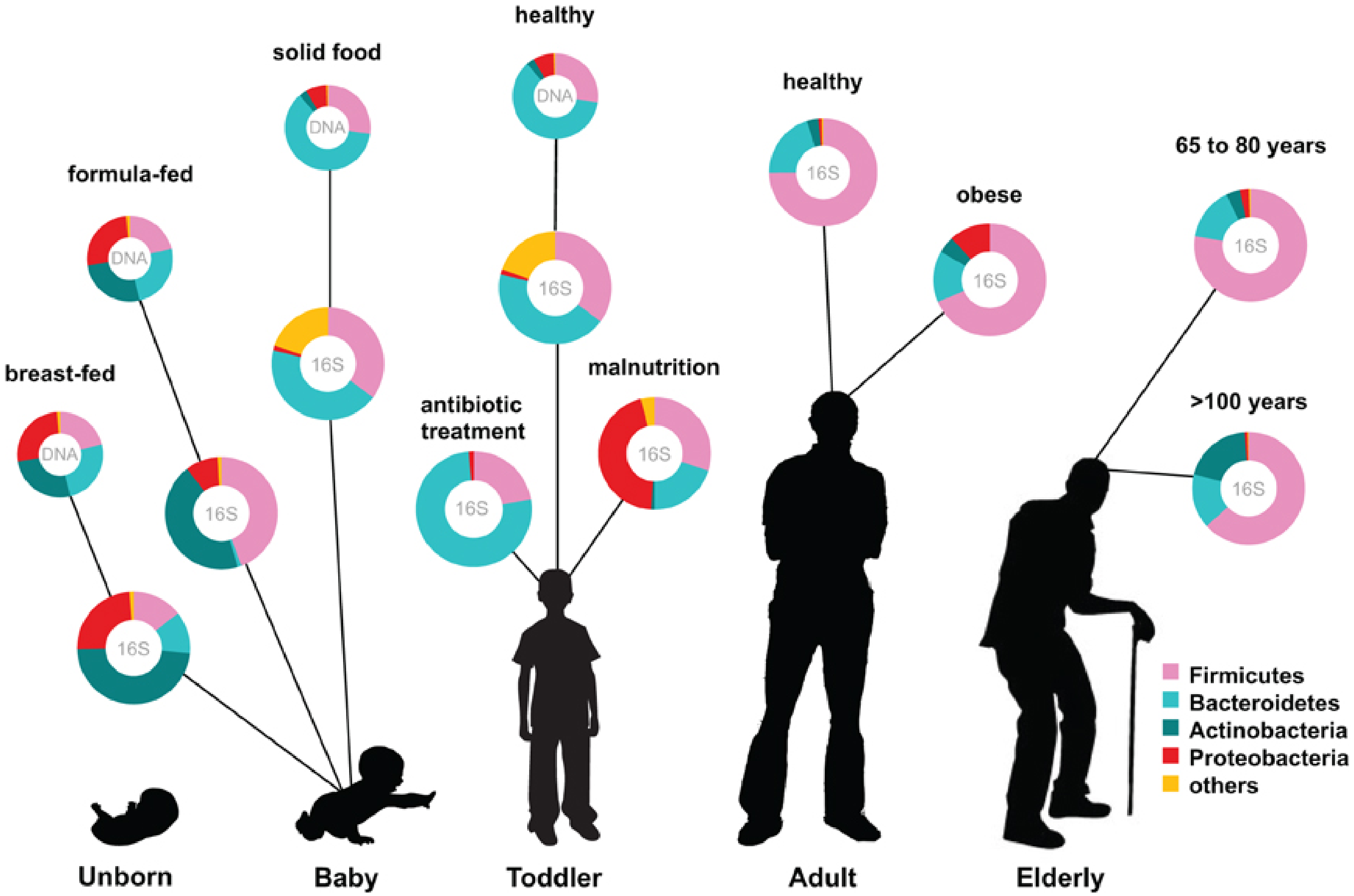
```
>OTU_1
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGC
GAAAGCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTA
>OTU_2
CCTACGGGAGGCAGCAGTGGgggATATTGCACAATGGggg
AAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTT
>OTU_3
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGggg
AAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTC

>OTU_4
AAACCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTATT
AAAGCGTGGGGAGCAAACAGGATTAGATACCCTTGTAGTC
```

Taxonomic annotation
Quantative analyses

Taxonomical Classification

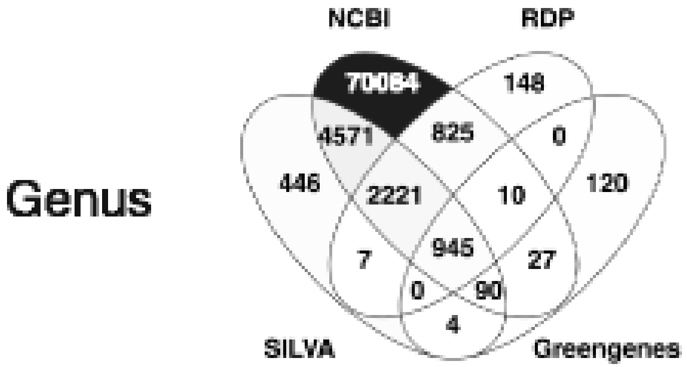
- Adding biological information to our data



Taxonomical Databases

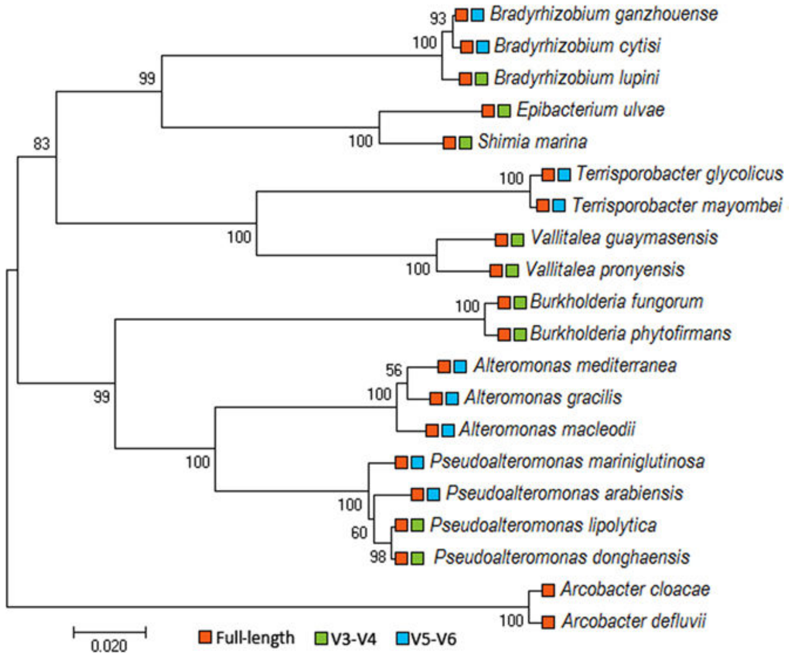
Taxonomical Databases

- Greengenes had latest update in 2013
- RDP had latest update in 2016
- SILVA continously updated



Taxonomical Resolution

- Are you looking at the right variable region?
- Larger databases compound the problem



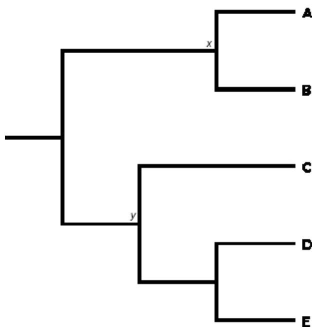
Taxonomical Assignment

- Closed-reference, *de novo* or both
- *de novo* comparison can be done by various methods
  - Blast
  - RDP classifier
  - UCLUST
  - q2-feature-classifier
- naive Bayes methods are fast and precise

Phylogeny

```
>OTU_1
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGC
GAAAGCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTA
>OTU_2
CCTACGGGAGGCAGCAGTGGgggATATTGCACAATGGggg
AAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTT
>OTU_3
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGggg
AAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTC

>OTU_4
AAACCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTATT
AAAGCGTGGGGAGCAAACAGGATTAGATACCCTTGTAGTC
```

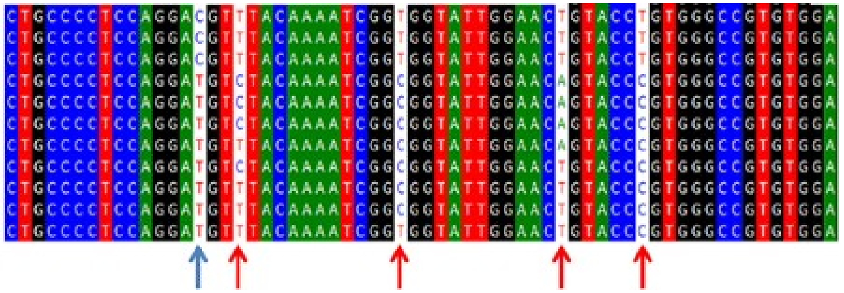


Phylogeny

Making phylogenetic trees is a course in itself!

Multiple alignment:

- PyNAST
- Infernal (secondary structure)
- Muscle



```
CTGCCCCTCCAGGACGTTTACAAAATCGGTGGTATTGGAACTGTACCTGTGGGCCGTGTGGA
CTGCCCCTCCAGGACGTTTACAAAATCGGTGGTATTGGAACTGTACCTGTGGGCCGTGTGGA
CTGCCCCTCCAGGACGTTTACAAAATCGGTGGTATTGGAACTGTACCTGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTCTACAAAATCGGCGGTATTGGAACAGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTCTACAAAATCGGCGGTATTGGAACAGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTCTACAAAATCGGCGGTATTGGAACAGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTTTACAAAATCGGCGGTATTGGAACAGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTCTACAAAATCGGCGGTATTGGAACTGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTTTACAAAATCGGCGGTATTGGAACTGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTTTACAAAATCGGCGGTATTGGAACTGTACCCGTGGGCCGTGTGGA
CTGCCCCTCCAGGATGTTTACAAAATCGGTGGTATTGGAACTGTACCCGTGGGCCGTGTGGA
```

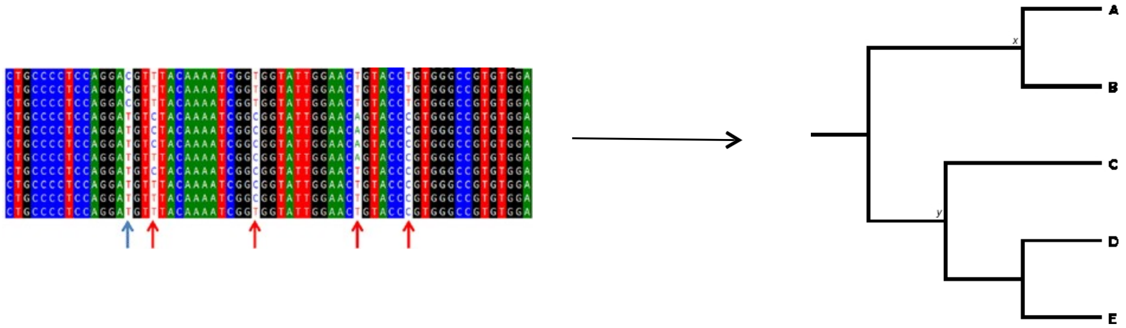
Filtering: conserved vs non conserved positions

Phylogeny

Making phylogenetic trees is a course in itself!

Building a tree:

- Converting the multiple alignment into distances
- MUSCLE
- RAxML
- Fasttree



Summary

- 16S rRNA is great because it contains conserved areas perfect for primers and hypervariable regions to distinguish bacteria
- 16s rRNA amplicon sequencing is great for looking at microbiome composition
- Primers matter!
- Denoising matters BUT does not change everything