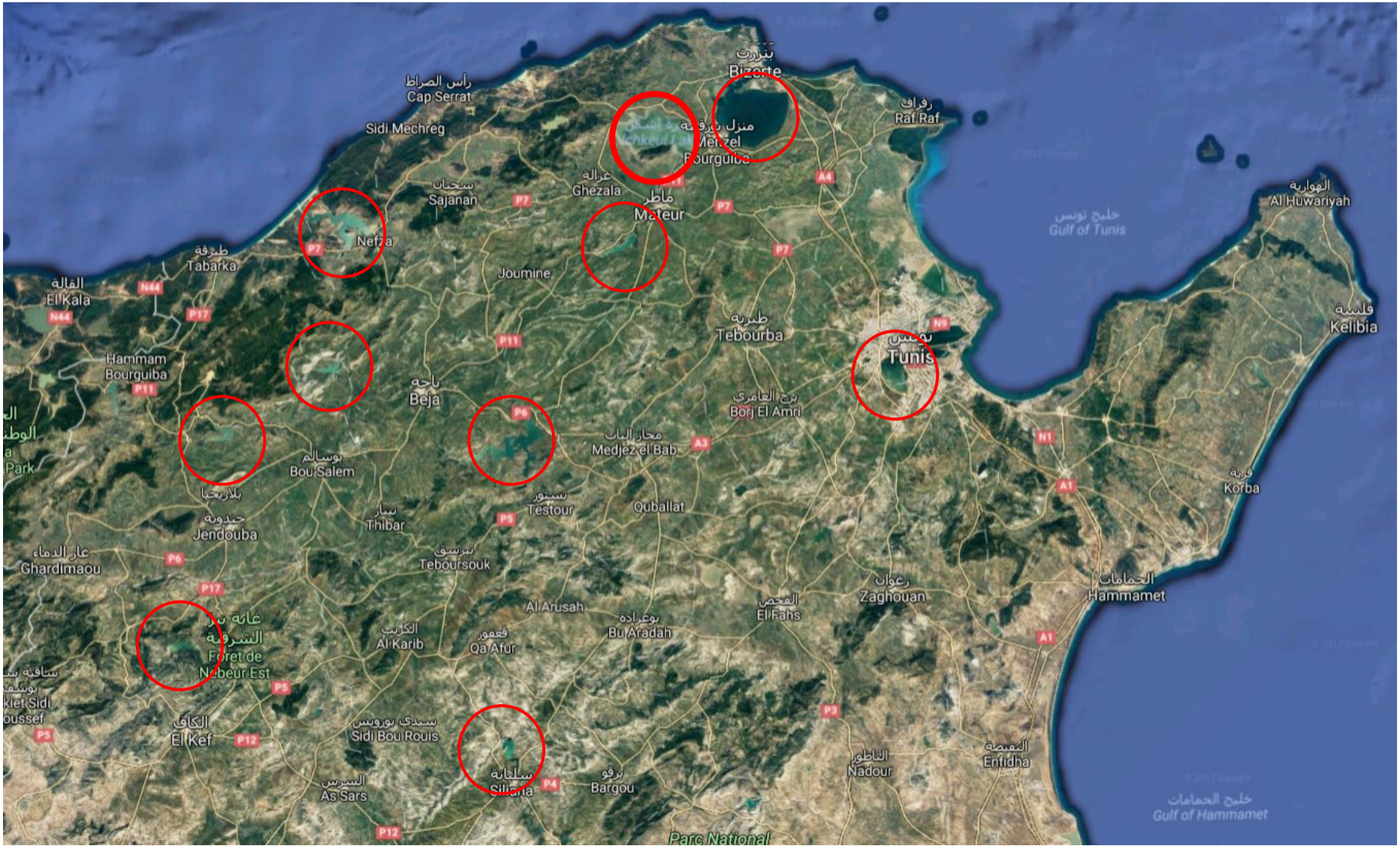
Handling amplicon sequences
“Where did the counts come from?”

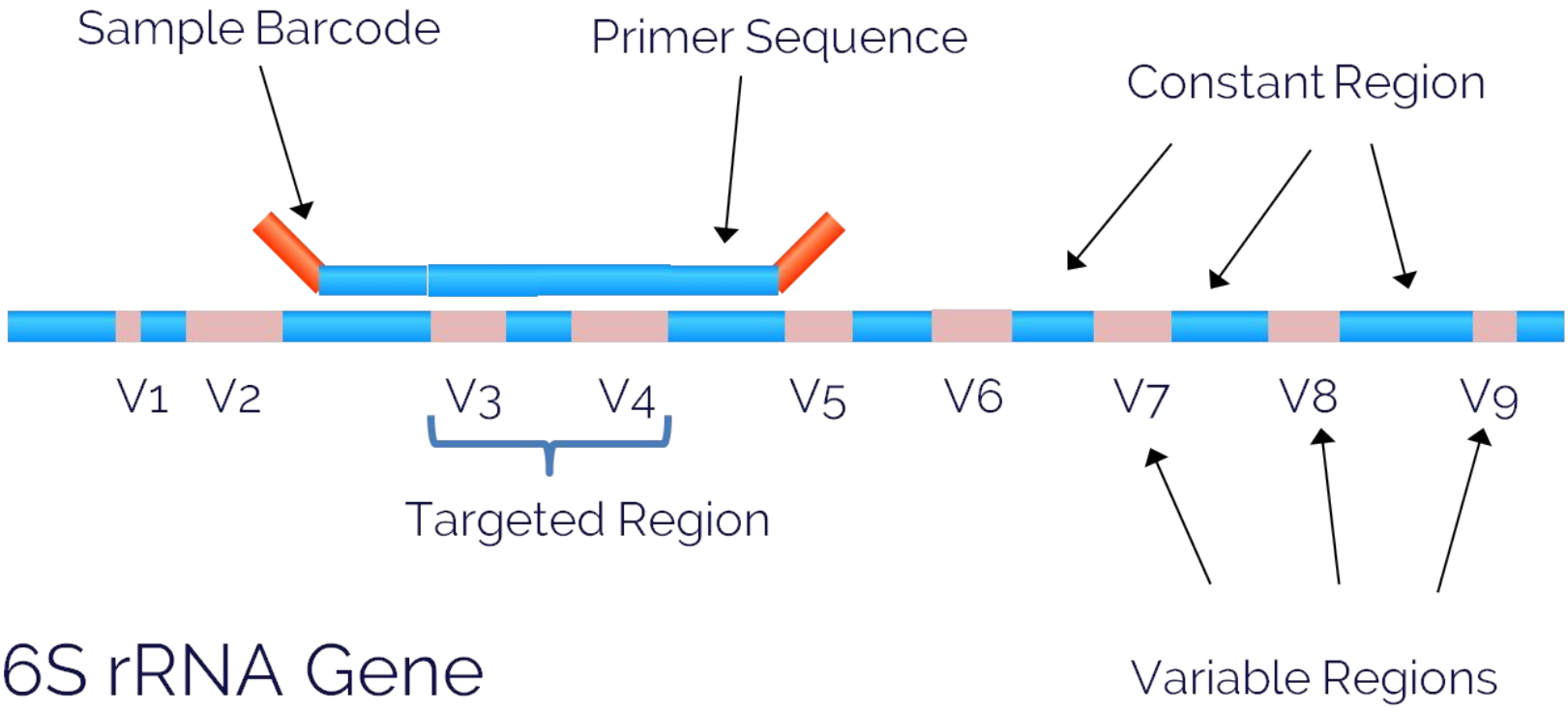
Daniel Lundin

Linnæus University | Centre
Ecology and Evolution in Microbial model Systems



Linnæus University | Centre
Ecology and Evolution in Microbial model Systems

Amplicon sequencing



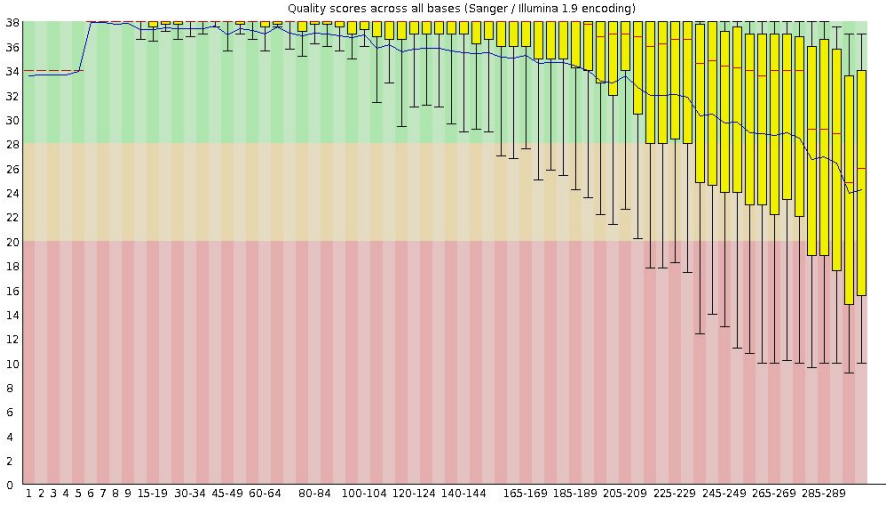
http://www.lcsciences.com

Illumina MiSeq sequencing

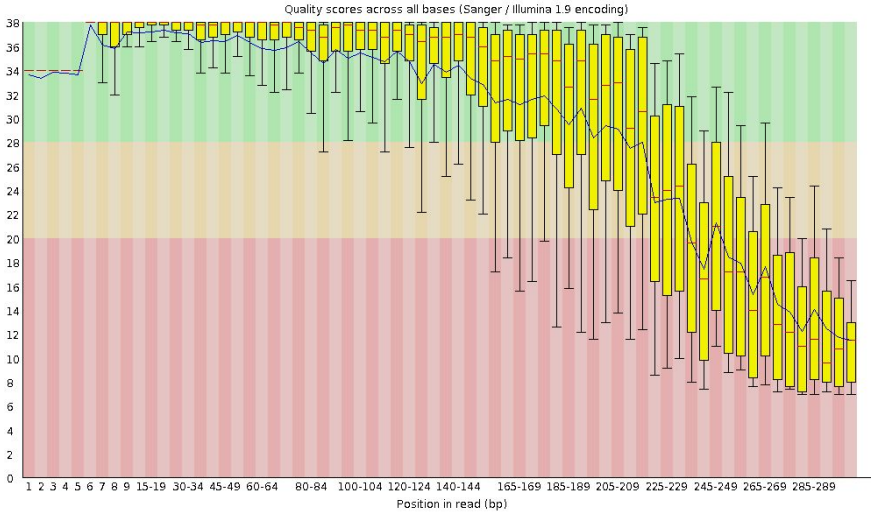
http://www.illumina.com

Up to 2x300 bp

Sequence quality



Forward reads



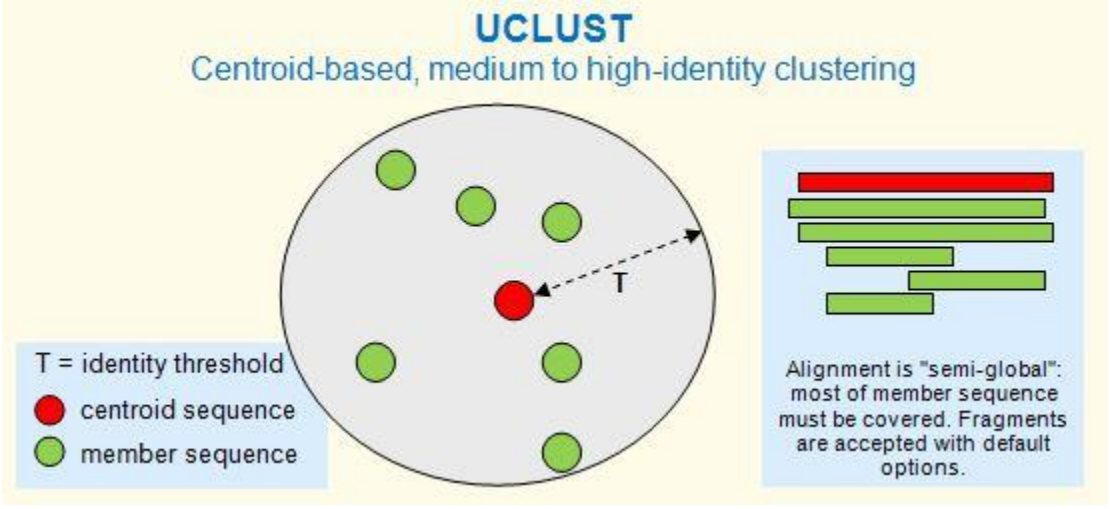
Reverse reads

Sequences are not born perfect...

OTU clustering

Error correction

Operational taxonomic unit (OTU) clustering



http://www.drive5.com

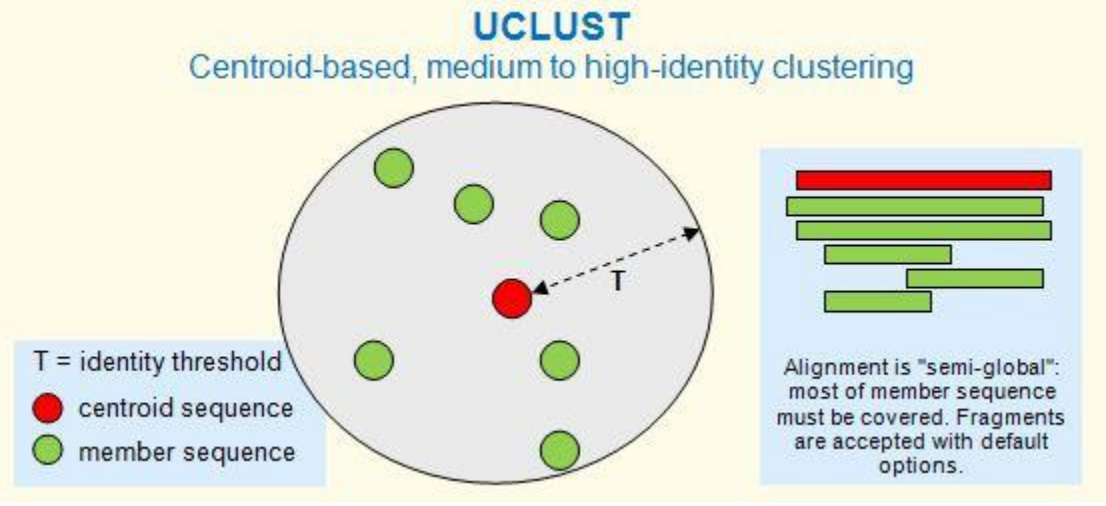
Pragmatic species concepts

70% DNA-DNA hybridization

95% average nucleotide identity

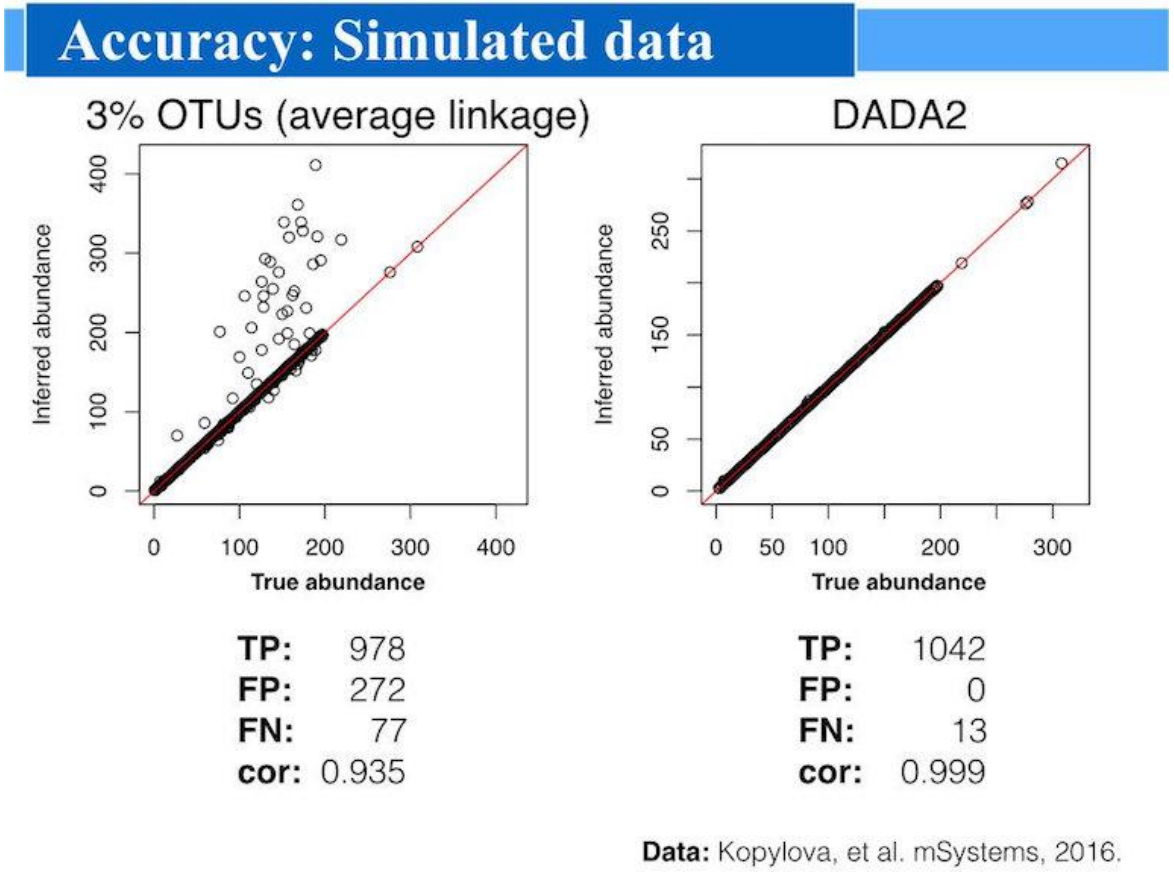
97% 16S rRNA nucleotide identity

How to interpret this for a cluster?

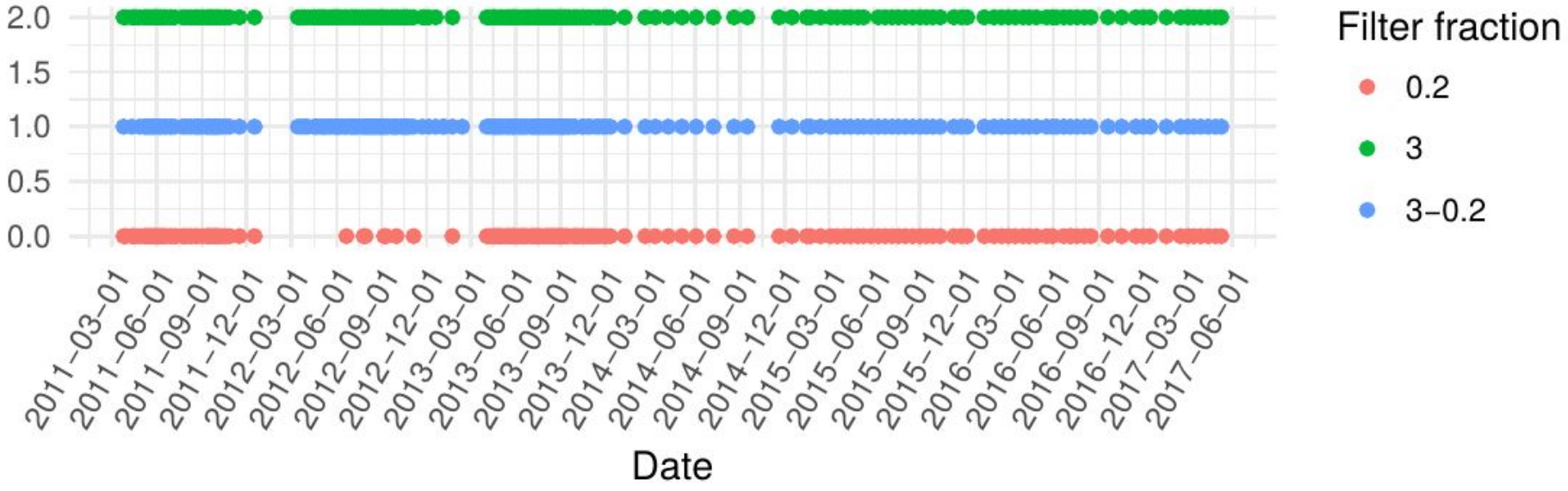


http://www.drive5.com

Statistical read correction

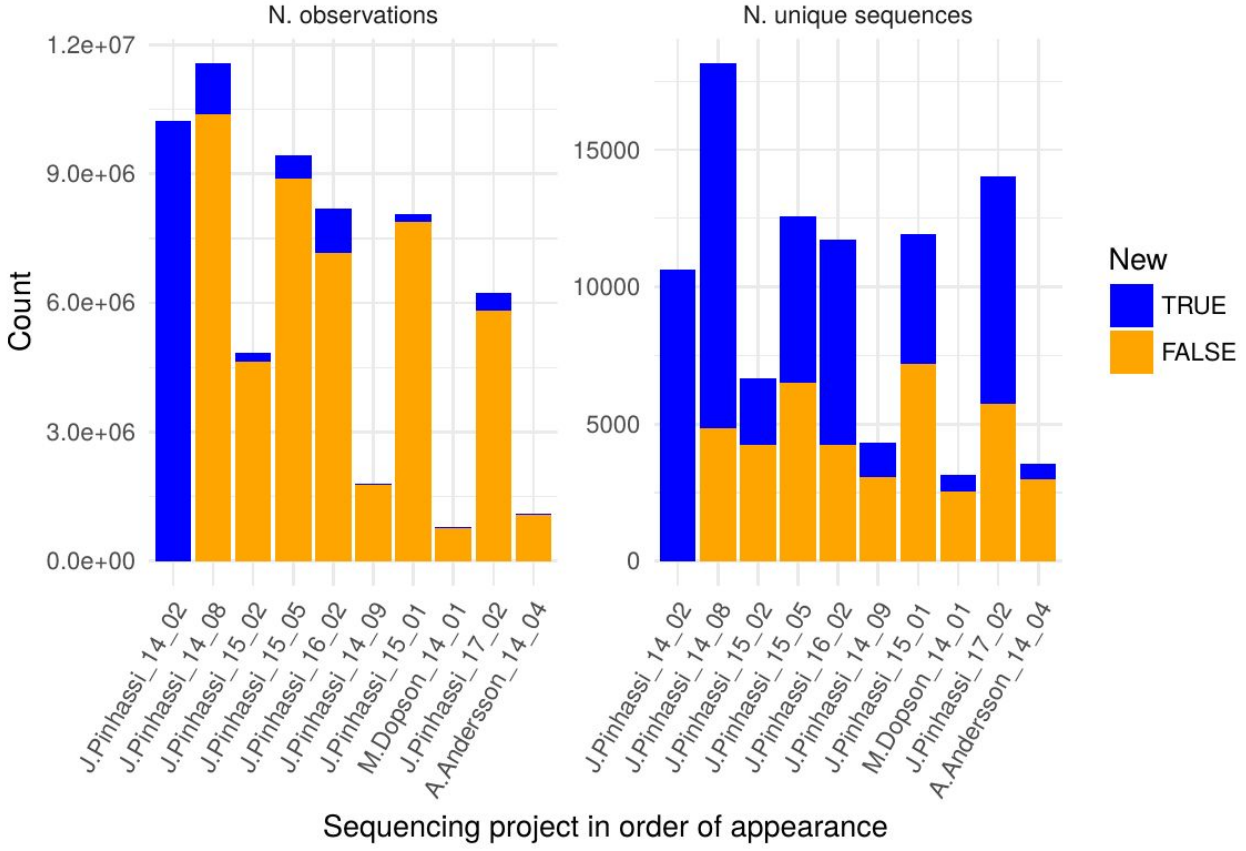


Long time series difficult with OTU clustering



Sampling at Linnaeus Microbial Observatory (LMO)

Correcting reads discovers the same sequences again and again



The DADA2 algorithm

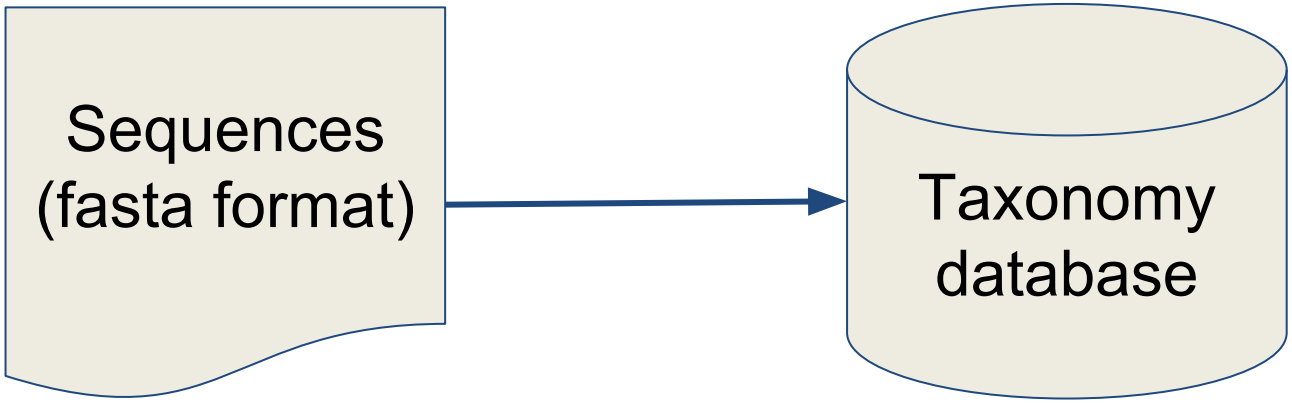
1. Trim sequences to same length and discard too short sequences (forward and reverse separately)
2. Calculate error profiles (using a subset of samples) (forward and reverse separately)
3. Correct reads using the error models (forward and reverse separately)
4. Merge forward and reverse reads into one long sequence
5. Detect and delete “bimeras”, artefactual PCR products (mixes of nucleotides from different organisms)

Taxonomy

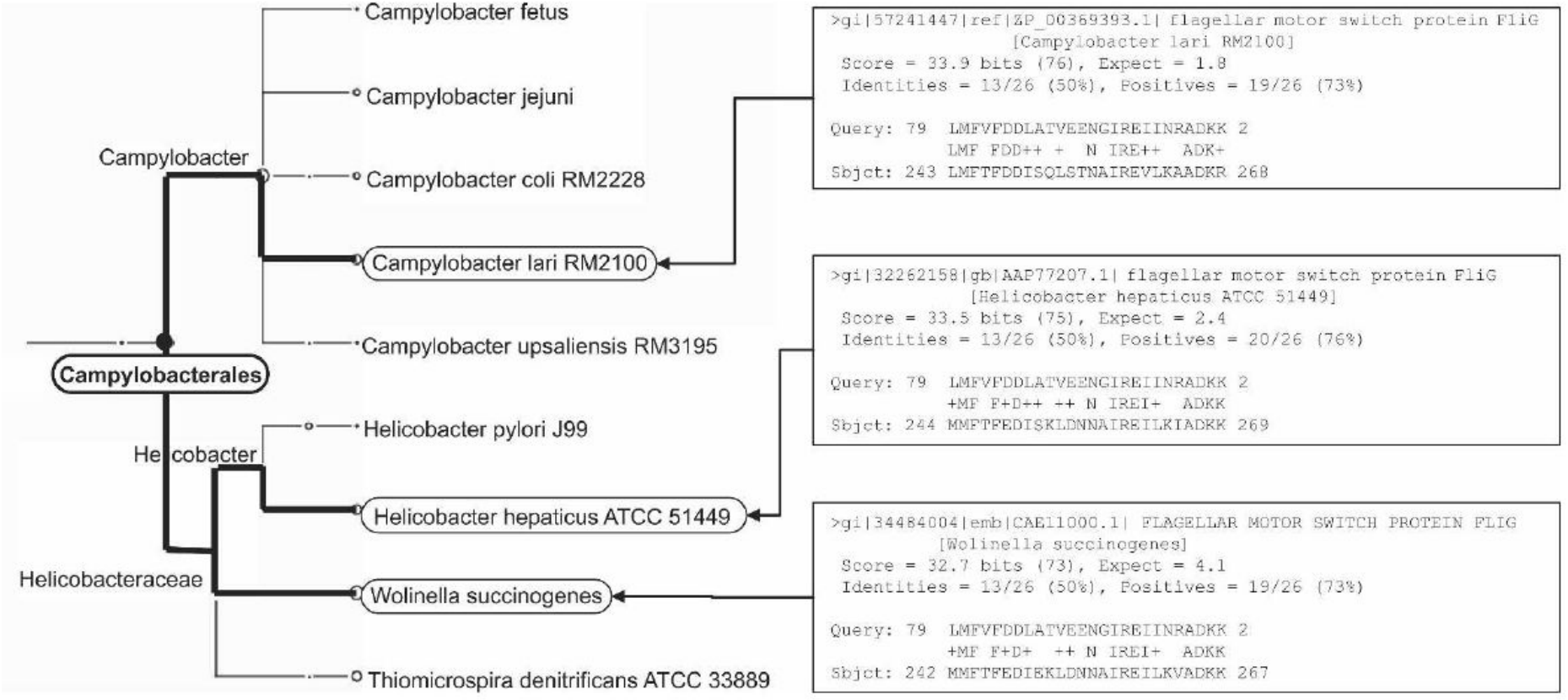
“If you don't know the names of things, the knowledge of them is lost too.”
(“Nomina si nescis, perit et cognitio rerum.”)

Linnaeus, Carl. 1751. *Philosophia botanica*.

Determining taxonomy



“Last Common Ancestor” (LCA)



Huson et al. Genome Research, 2007

The common alternatives

https://www.mothur.org/

http://qiime.org/