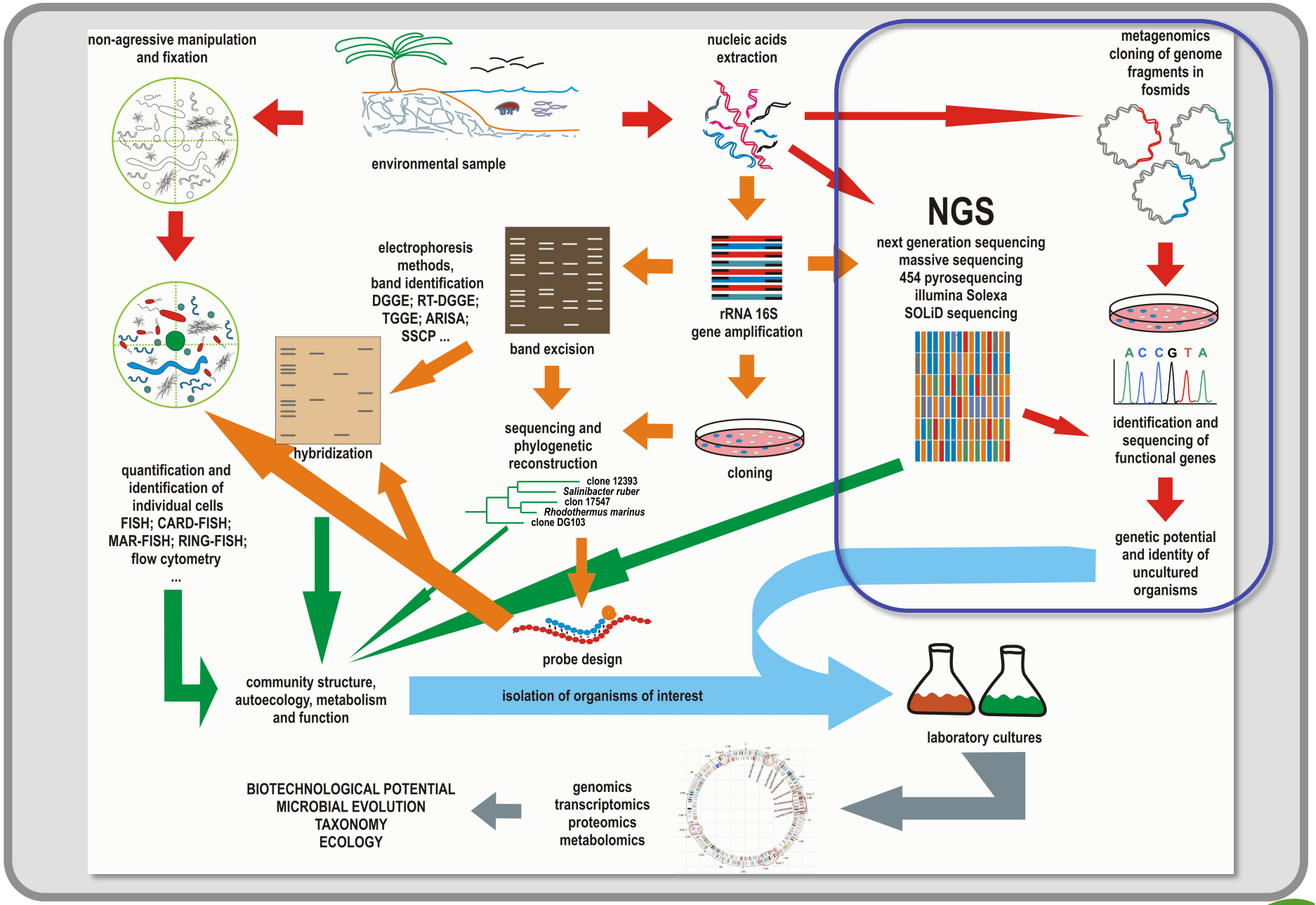
16s, primer selection, tree reconstruction



1- Approaches to understand environmental microbiology

Intra-gene variability

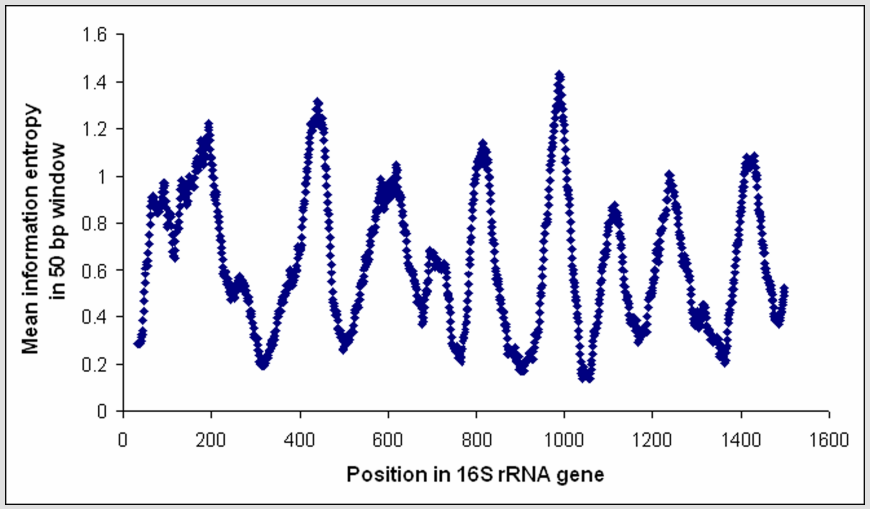
⇨ secondary structure shows differences in the conservation of homologous sites

⇨ highly conserved zones give information on deep-genealogies

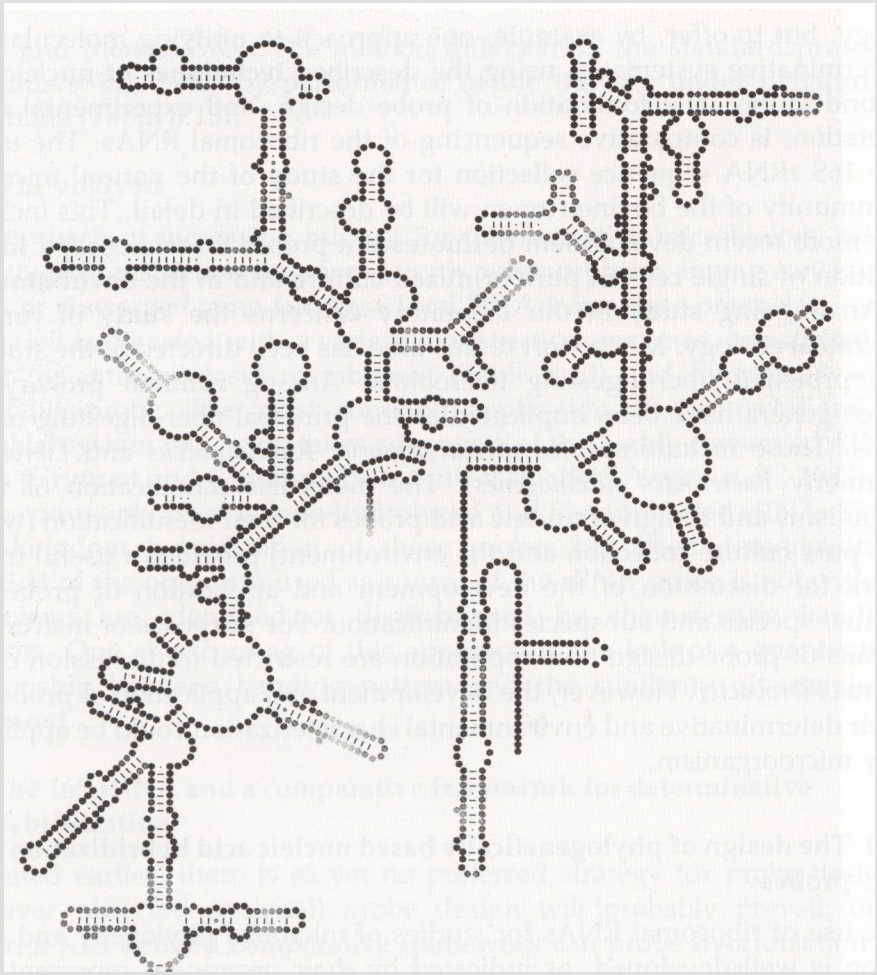
(higher resolution for distantly related)

⇨ hypervariable zones give information on recent events

(higher resolution for close relatives)



Anderson et al., 2008 PLoS ONE, 3: e2836



Stahl and Amann, 1991 John Wiley and Sons

Bias on primer selection ⇔ universality of target sites

- ⇨ Universal primers target highly conserved regions
- ⇨ Universality depends on the known dataset
- ⇨ Different phyla may have differences in the “universal” regions (e.g. EUB 338)
- ⇨ Primers used for rRNA cloning may give biased results
- ⇨ Metagenomics without amplification steps may reveal hidden diversity

EUB338 I	**Most *Bacteria***	**GCTGCCTCCCGTAGGAGT**
EUB338 II	***Planctomycetales***	**GCAGCCACCCGTAGGTGT**
EUB338 III	***Verrucomicrobiales***	**GCTGCCACCCGTAGGTGT**
EUB338 IV	***Isosphaera/annamox***	**GCAGCCTCCCGTAGGAGT**

Daims et al. 1999. System Appl Microbiol 22, 434-444
Schmidt et al., 2005. Appl Environ. Microbiol. 71, 1677- 1684

Bias on PCR length heterogeneity⇔ self-splicing introns

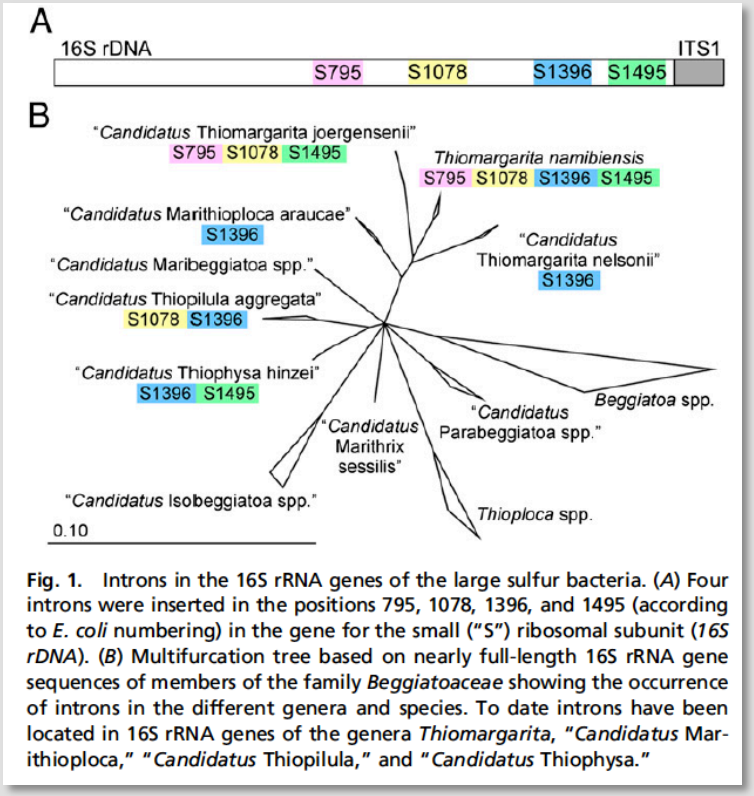


Fig. 1. Introns in the 16S rRNA genes of the large sulfur bacteria. (*A*) Four introns were inserted in the positions 795, 1078, 1396, and 1495 (according to *E. coli* numbering) in the gene for the small ("S") ribosomal subunit (*16S rDNA*). (*B*) Multifurcation tree based on nearly full-length 16S rRNA gene sequences of members of the family *Beggiatoaceae* showing the occurrence of introns in the different genera and species. To date introns have been located in 16S rRNA genes of the genera *Thiomargarita*, "*Candidatus* Marithioploca," "*Candidatus* Thiopilula," and "*Candidatus* Thiophysa."

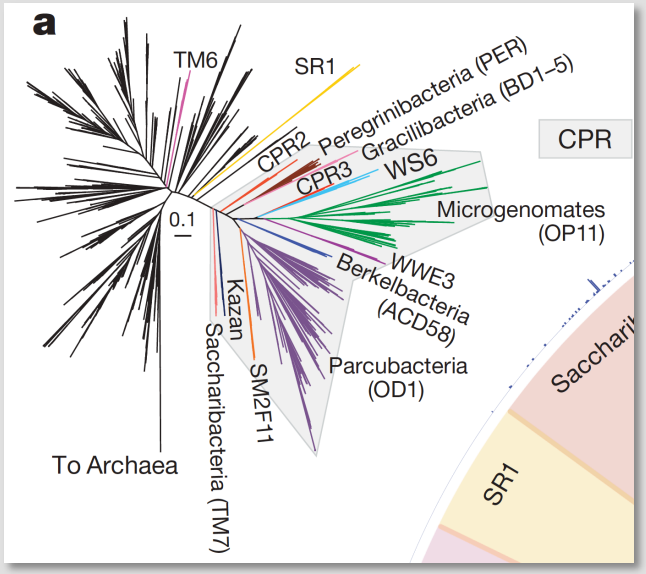
Salman et al. (2012) PNAS 109, 4203-4208

⇨ some organisms as large sulfur bacteria could have 16S rRNA genes with length >3500 nuc

⇨ long genes may escape diversity surveys due to PCR bias or selection of 1.6 Kb amplicons

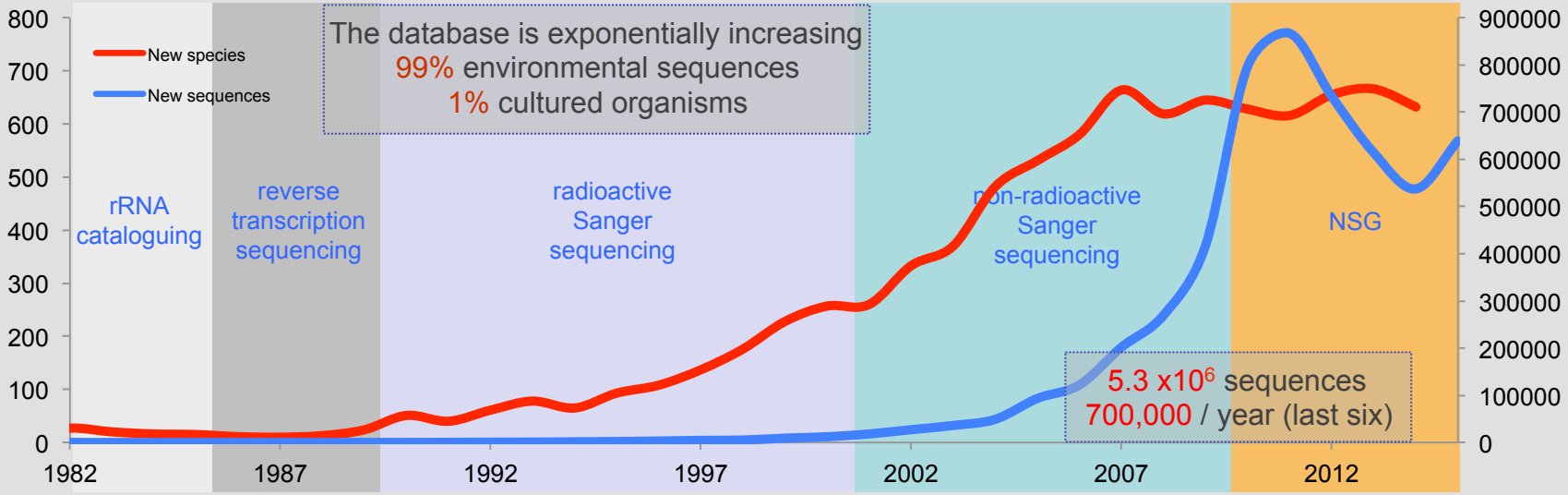
⇨ new lineages as the Candidate Phyla Radiation (CRP) contain self-splicing introns and proteins within rRNA genes

⇨ Metagenomics help in revealing PCR biases



Brown et al. (2012) Nature 523, 208-221

Bias due to sequence quality⇔ different methods render different quality



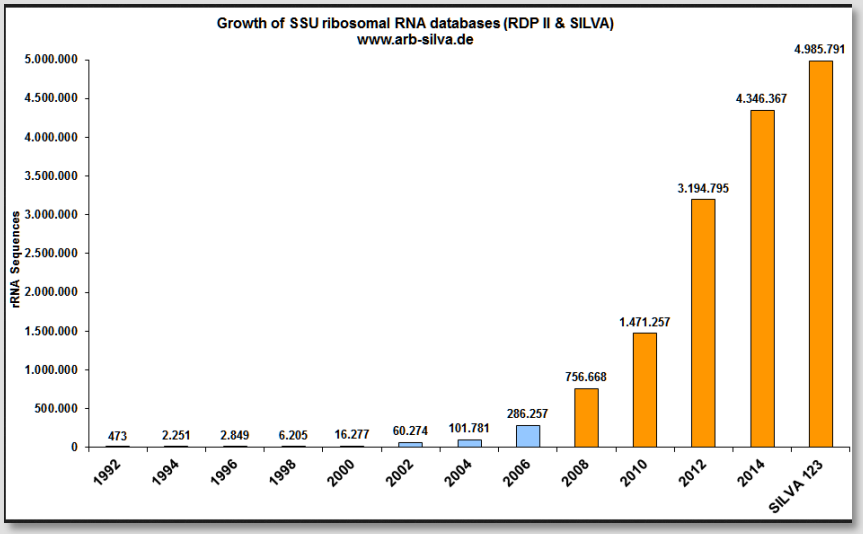
Yarza et al., Nature Revs. 2014. 12: 635-645

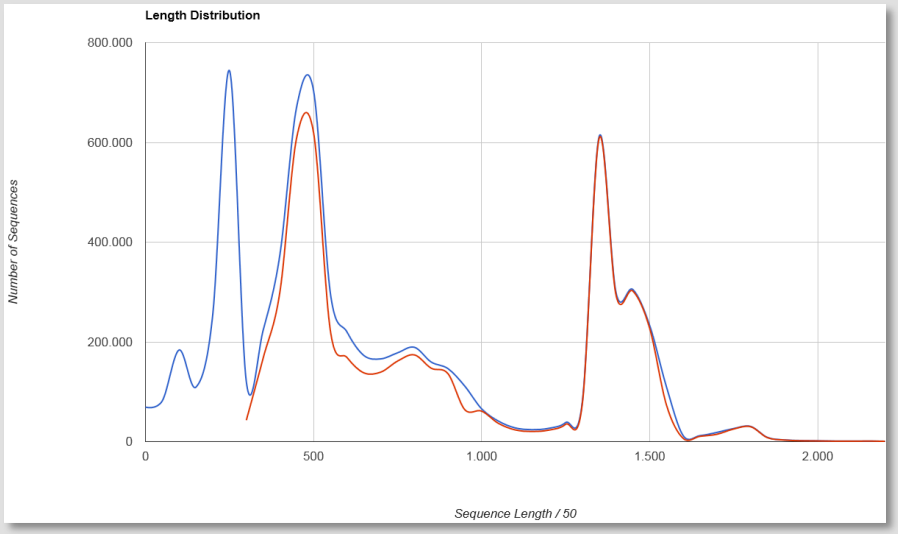
Tamames & Rosselló-Móra 2012 TIM 20:514-516

Sources of sequences and quality

- ⇨ rRNA Cataloguing (up to late 80's), bad quality
- ⇨ reverse transcription sequencing (up to late 90's), bad quality
- ⇨ Sanger methods (radioactive, biotin-labeled, terminal-dye… still in use)
  - ⇨ cloning DNA, good quality
  - ⇨ direct amplification, good quality
  - ⇨ DGGE/TGGE, short sequences, bad quality
- ⇨ NSG, short sequences
  - ⇨ 454 technology (now up to 800nuc, mean of 500nuc), moderate quality
  - ⇨ illumina (now 2x 250nuc), too short

Bias due to sequence quality⇔ short sequences are worse than longer





www.arb-silva.de

SILVA release 123(July 2015)

rate of rejection of about 30% of the existing sequences

short sequences are generally worse than long stretches

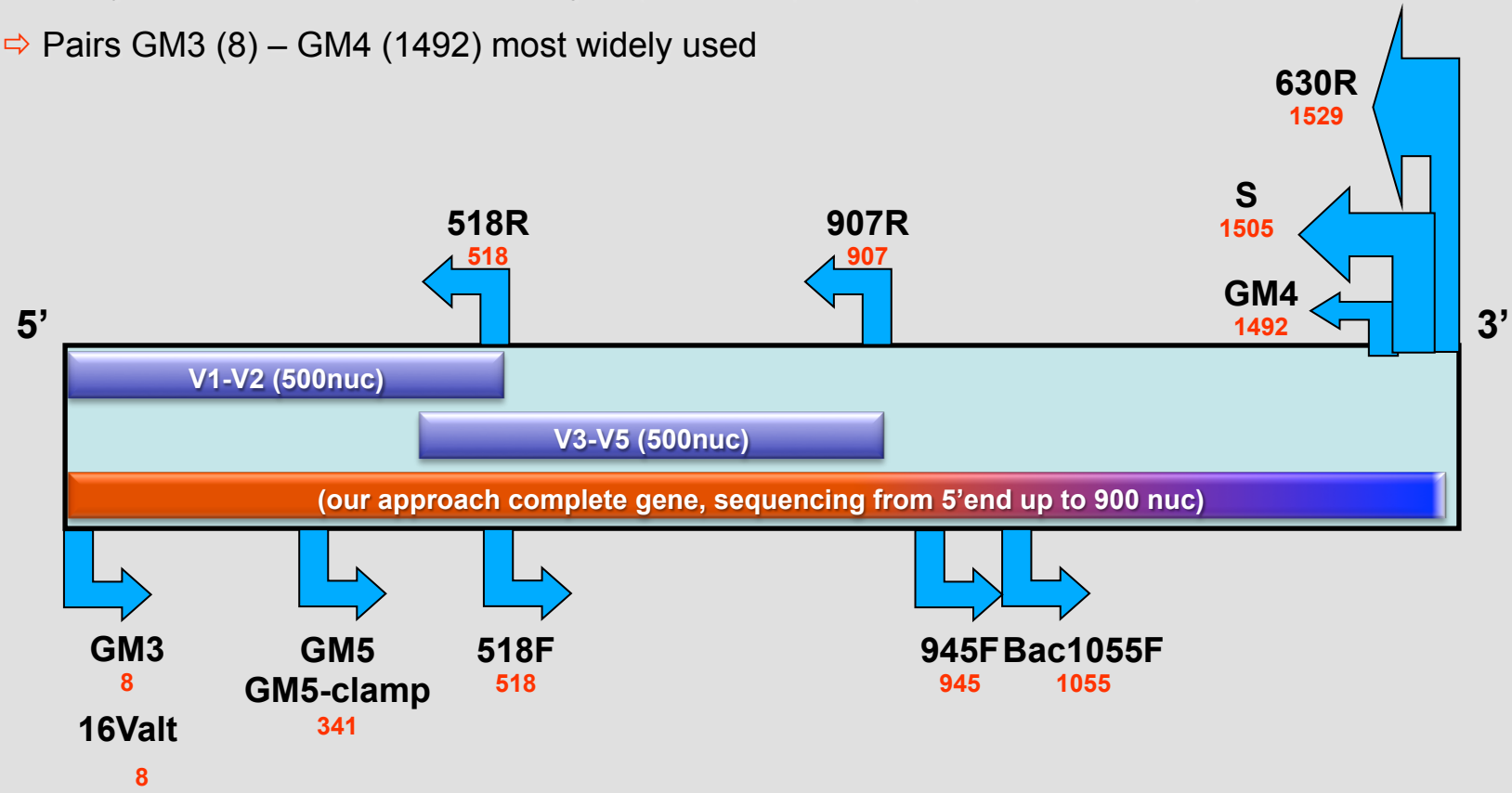
Quast et al., 2013, Nuc Acid Res. 41: D590-D596

Sequence Retrieval and Processing

	SSU 123
candidates (total)	7,168,241
RNAmmer	84,098
< 300 bases	1,650,058
> 2% ambiguities	25,713
> 2% homopolymers	122,621
> 2% vector contamination	2054
low alignment identity	481,981
total rejected by QC	2,182,450

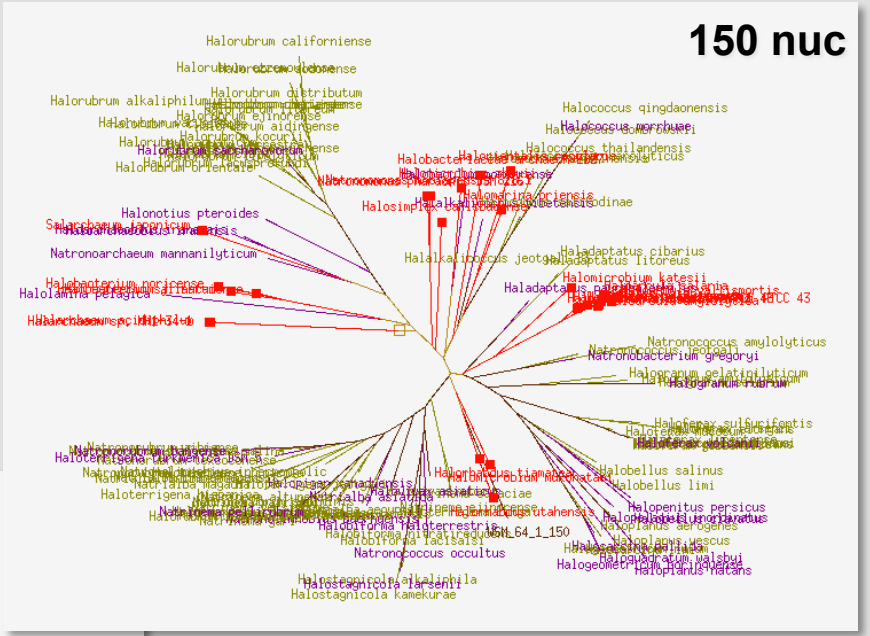
Primer selection ⇔ size of the amplicon

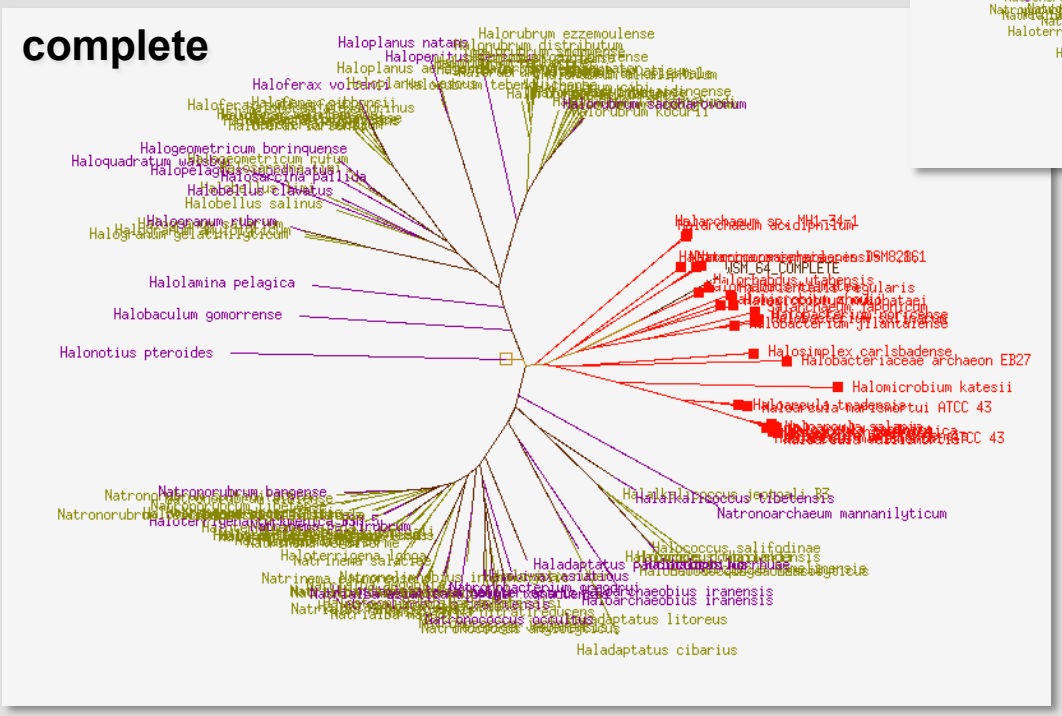
- ideally the almost complete gene (~ 1520 nucleotides) should be sequenced
- many amplifications skip sequencing the helix 50 (~ 1490 nucleotides)
- many clone libraries are based on just partial amplicons (~ 900 nucleotides)
- Pairs GM3 (8) – GM4 (1492) most widely used



Size & information content

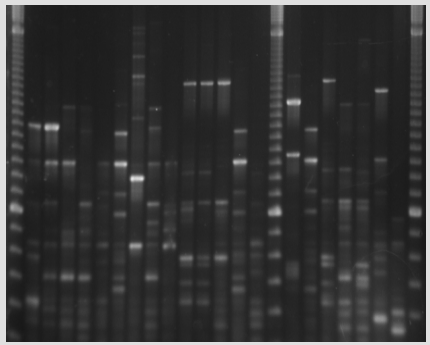
- ⇨ complete sequences give complete information
- ⇨ partial sequences lose phylogenetic signal
- ⇨ short sequences lose resolution





Short sequences do not have enough resolution for reconstructing phylogenies

Phenetics vs Cladistics



⇨ Data can be treated as presence/absence/intensity to generate similarity matrices

⇨ If data is analyzed by their similarity ⇔ PHENETICS

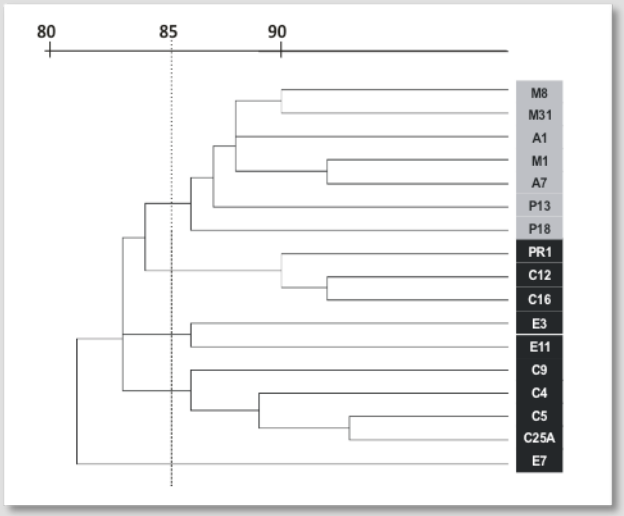
⇨ If data is analyzed in an evolutionary context (i.e. changes in homologous characters are mutations or evolution steps) ⇔ CLADISTICS

⇨ For evolutionary purposes is necessary to recognize HOMOLOGY

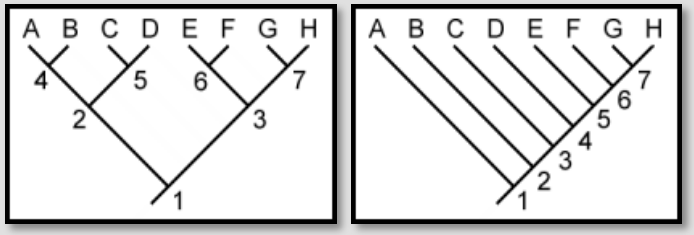
Similarity matrix or alignment

```
OTU A     10100010010010010
OTU B     11010001010001010
OTU C     00010010011110101
OTU D     00111110010101010
OTU E     00010010111001101
...
```

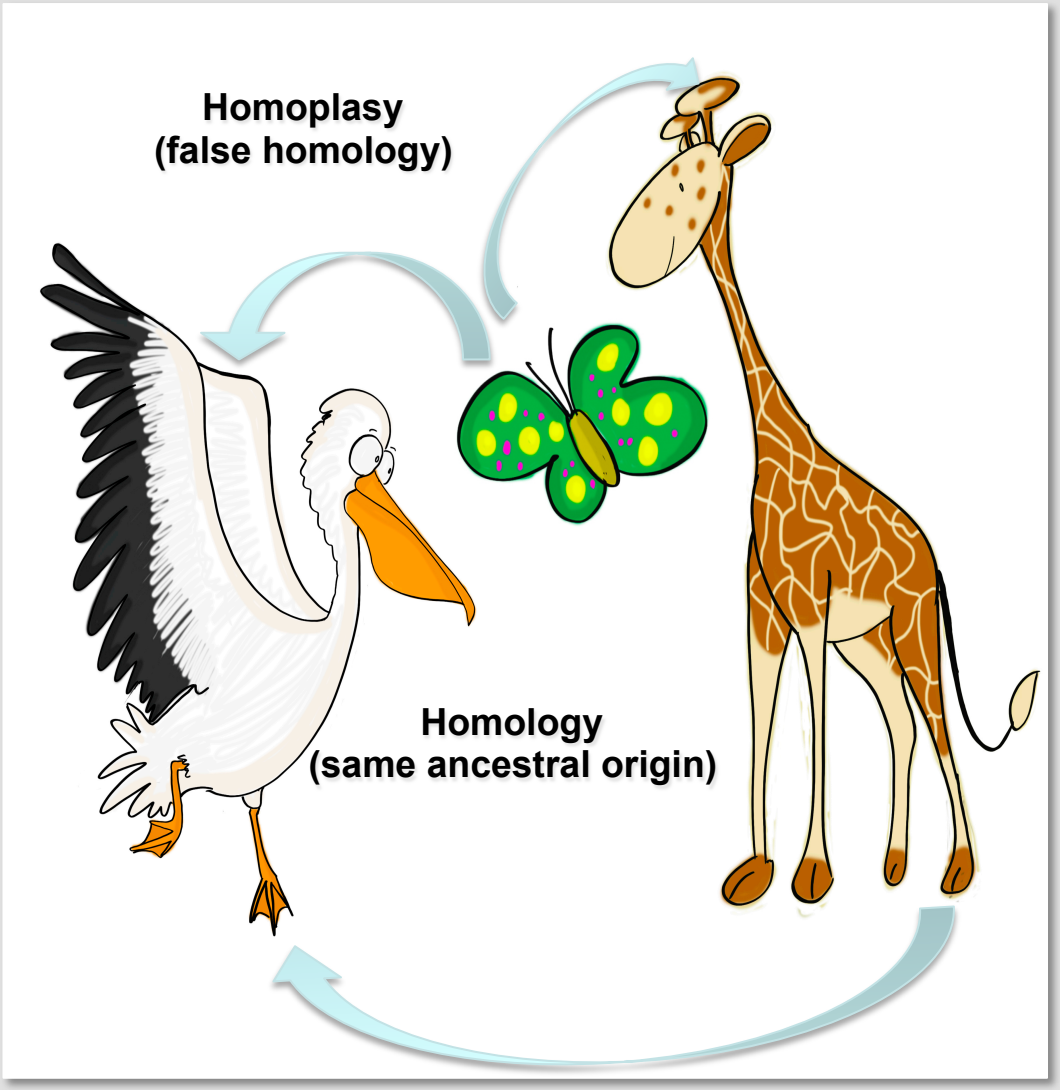
PHENETICS

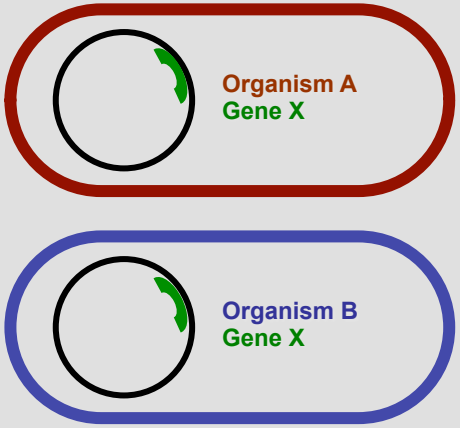


CLADISTICS

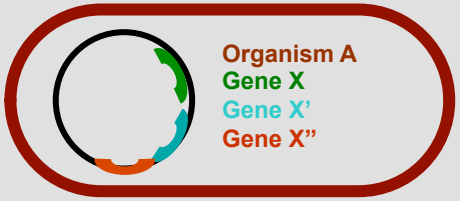


HOMOLOGY ⇔ ORTOLOGY ⇔ PARALOGY ⇔ HOMOPLASY





Orthology ⇔ homologous genes in different organisms



Paralogy ⇔ homologous genes in the same organism, gene duplications with identical or different function

The relevance of the alignment

To perform cladistic analyses we should first align al sequences in order to recognize all homologous positions.

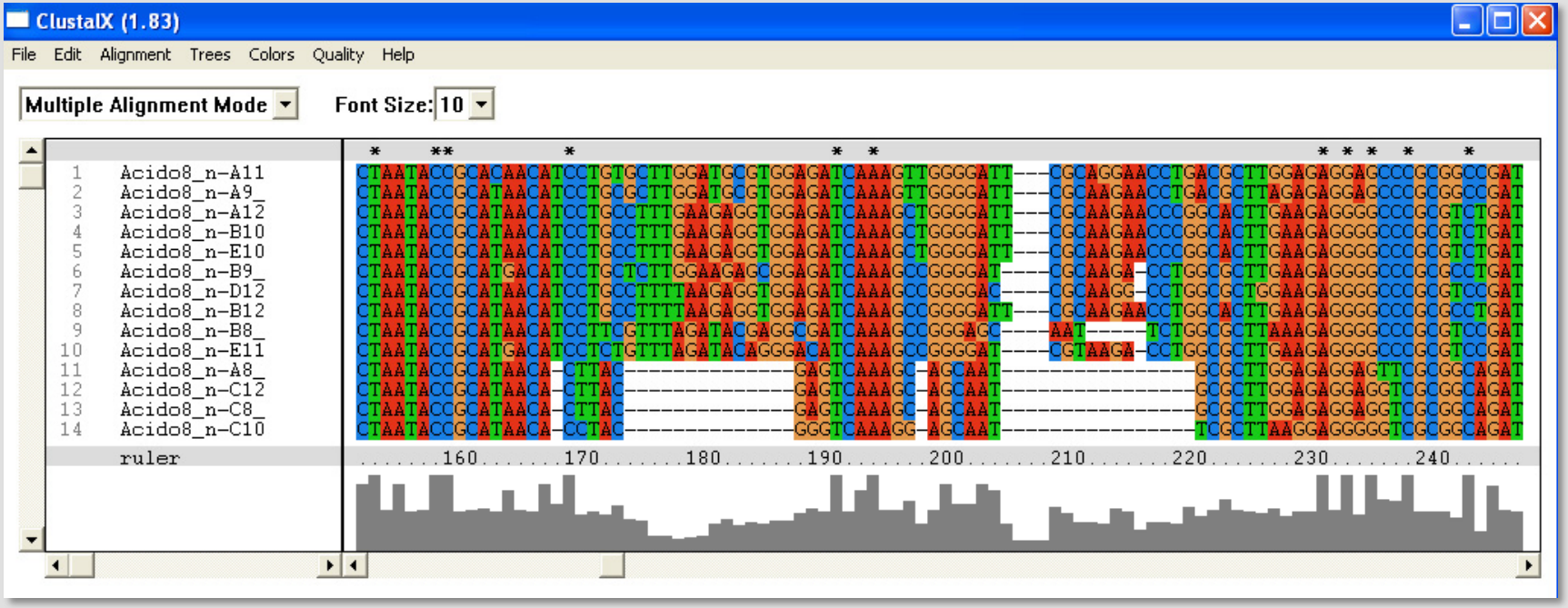
Recognition by:

- **Sequence similarities**
- **Base pairing due secondary structure (helixes for rRNA)**
- **Insertions & deletions**
- **Empirically (subjective)**
- **Minimize homoplasic influences**

There are many alignment programs, all look to common features that may indicate homologous sites:

- **Clustal X**
- **MAFFT**
- **PileUp**
- **…**

The relevance of the alignment



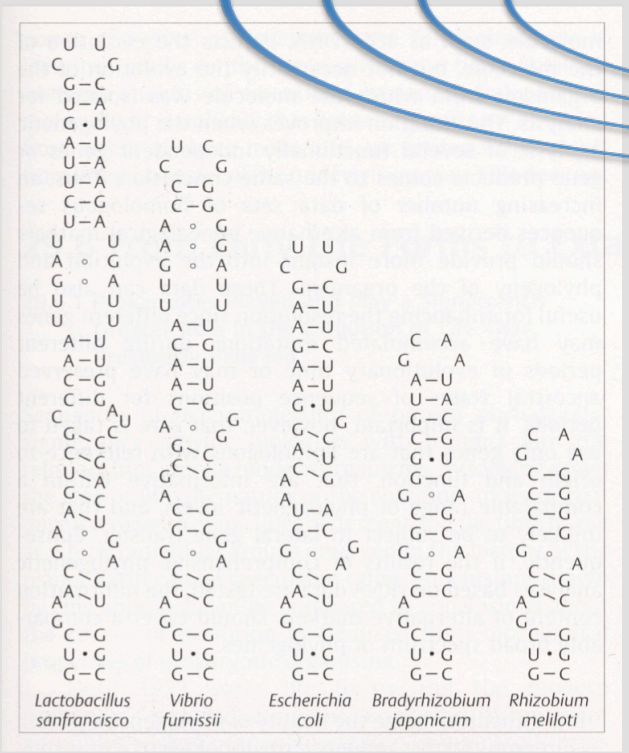
Most of the programs do not take into account secondary structure, just sequence motive similarities

rRNA has a secondary structure with helixes that help in aligning sequences

Functional gene or translated proteins cannot be improved by secondary structure analysis

The relevance of the alignment: usefulness of ARB program package

www.arb-home.de
www.arb-silva.de



ARB does take into account features as helix pairing

By increasing the numbers of sequences, the alignment improves

The algorithms

Neighbor Joining:

```
G C C A T => a
G C A C T => b
G C A C C => c
```

alignment

```
a => 100
b =>  60   100
c =>  40    80   100
      a     b     c
```

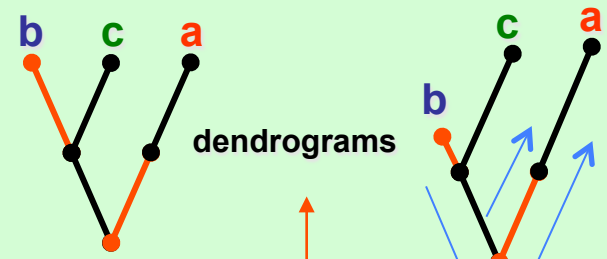
Similarity matrix

Distance transformation

Jukes-Cantor
Kimura
De Soete

```
a =>  0
b =>  40    0
c =>  60   20    0
      a     b     c
```

Distance matrix



(pitfalls: does not take into account multiple mutations)

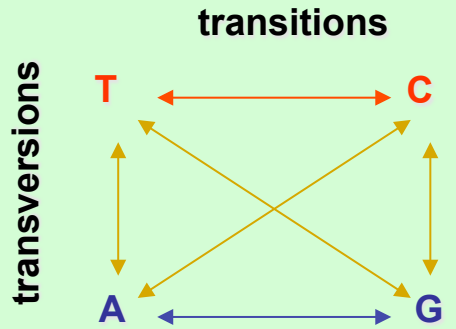
Maximum Likelihood

Like Maximum Parsimony but takes into account

⇒ difficulties in mutation events (transitions vs. transversions)

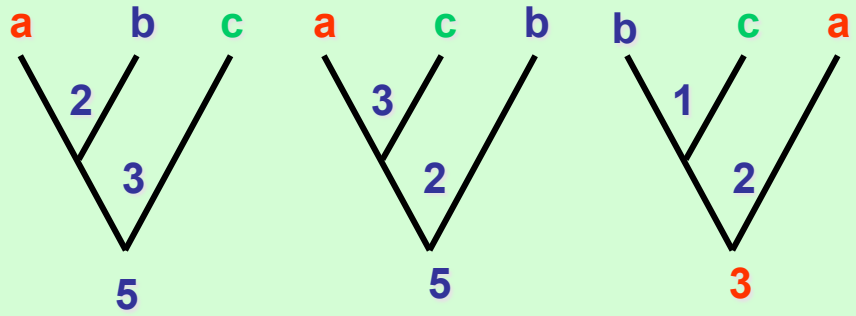
⇒ mutation position

⇒ Slower



Maximum Parsimony

```
G C C A T => a
G C A C T => b
G C A C C => c
```



a – b => 2 mutations

a – c => 3 mutations

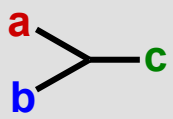
b – c => 1 mutation

(pitfalls: nature may not be parsimonious)

The number of trees (rooted vs unrooted)

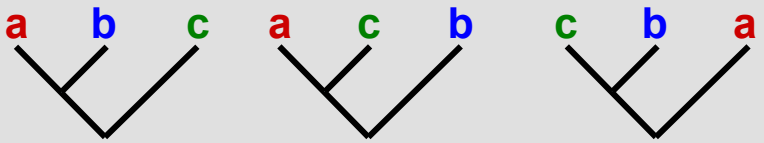
unrooted

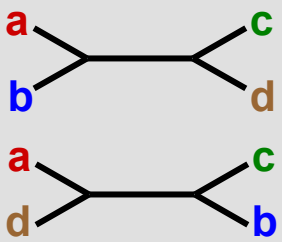
$$\frac{(2s-5)!}{2^{s-3}(s-3)!}$$

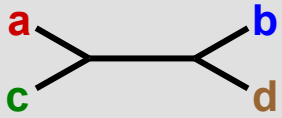


rooted

$$\frac{(2s-3)!}{2^{s-2}(s-2)!}$$

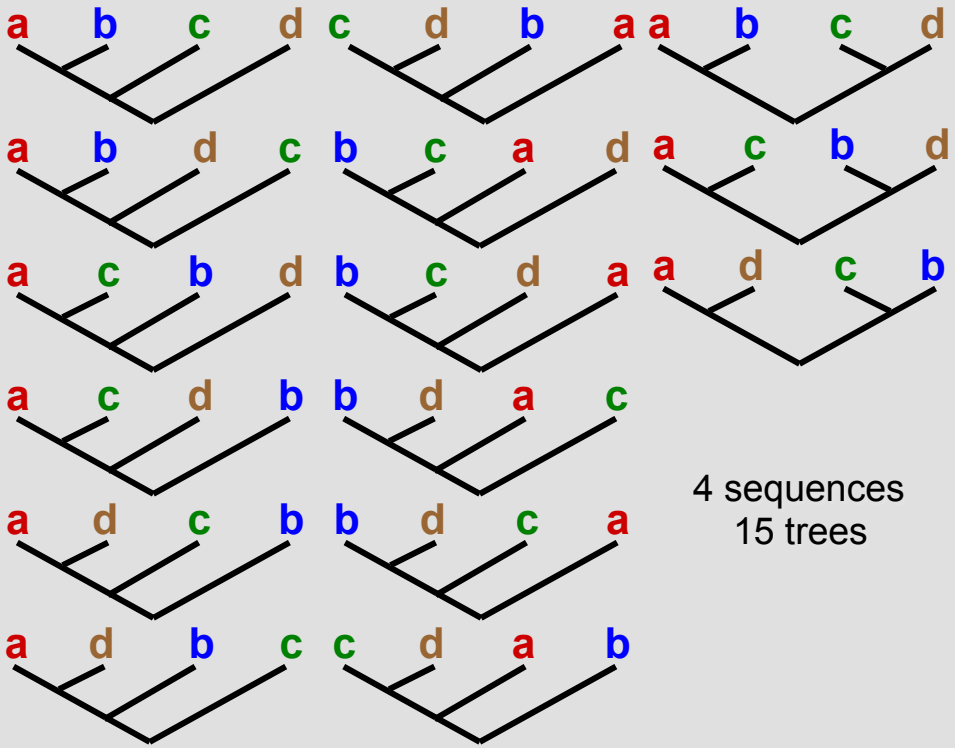






4 sequences
3 trees

taxa	unrooted	rooted	comment
4	3	15	
8	10,395	135,135	
10	2,027,025	34,459,425	
22	$3x10^{23}$		1 mole trees
50	$3x10^{74}$		> trees as atoms in the universe
100	$2x10^{182}$		

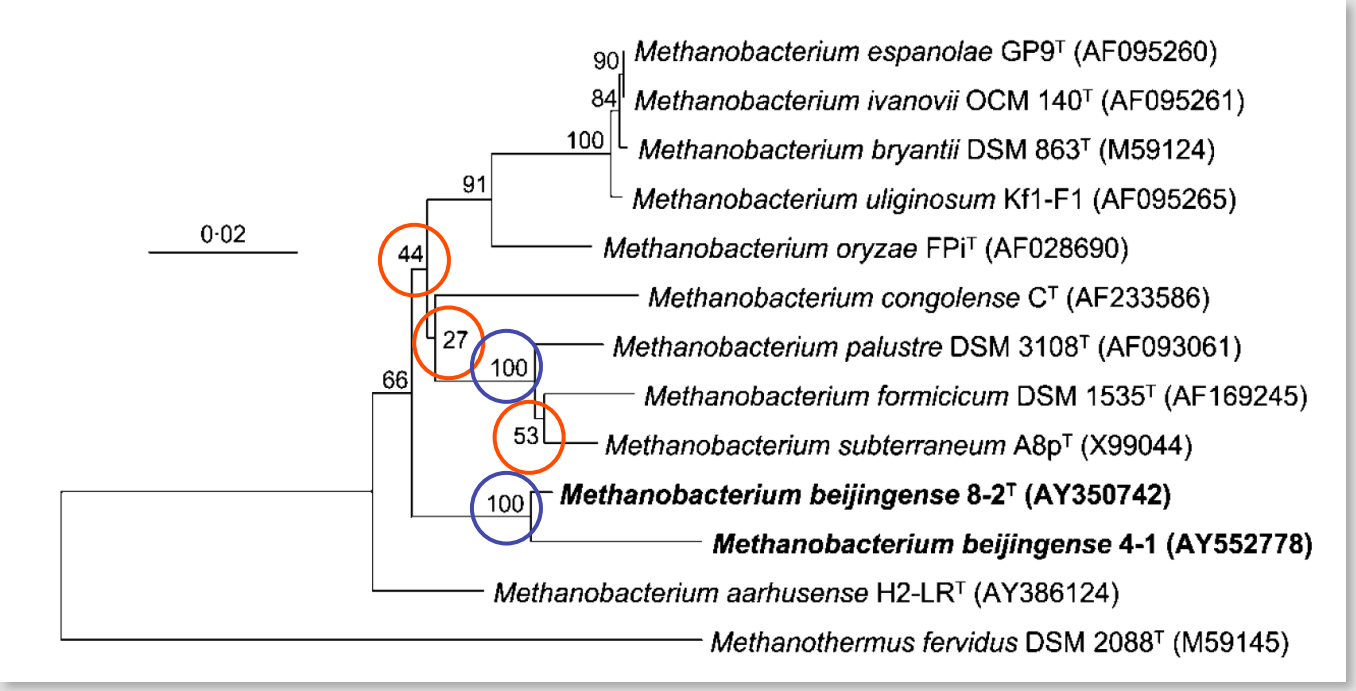


4 sequences
15 trees

Bootstrap

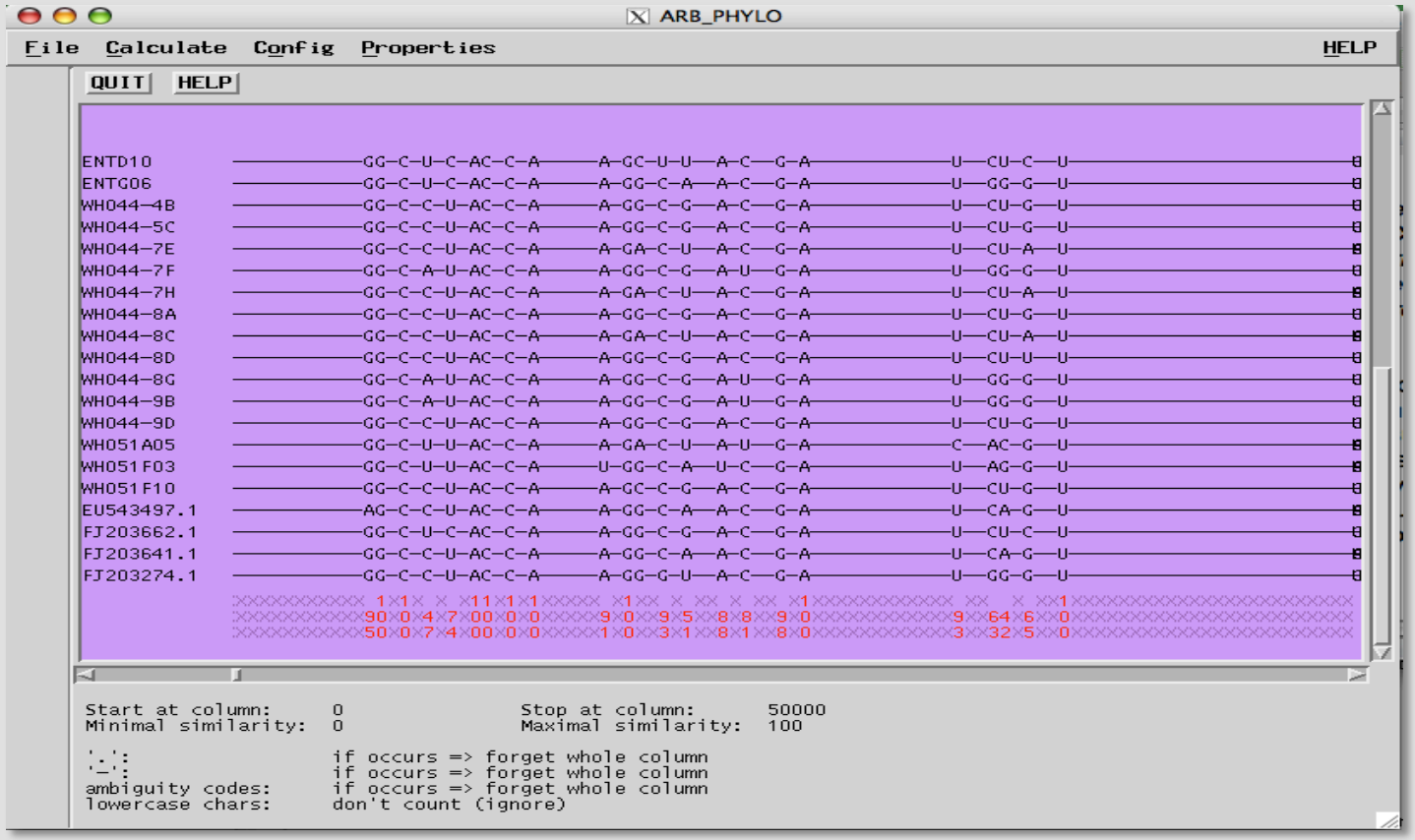
Bootstrap indicates how stable is a branching order when a given dataset is submitted to multiple analysis

Generally short internode branches will have low bootstrap values



PHYLOGENETIC FILTERS

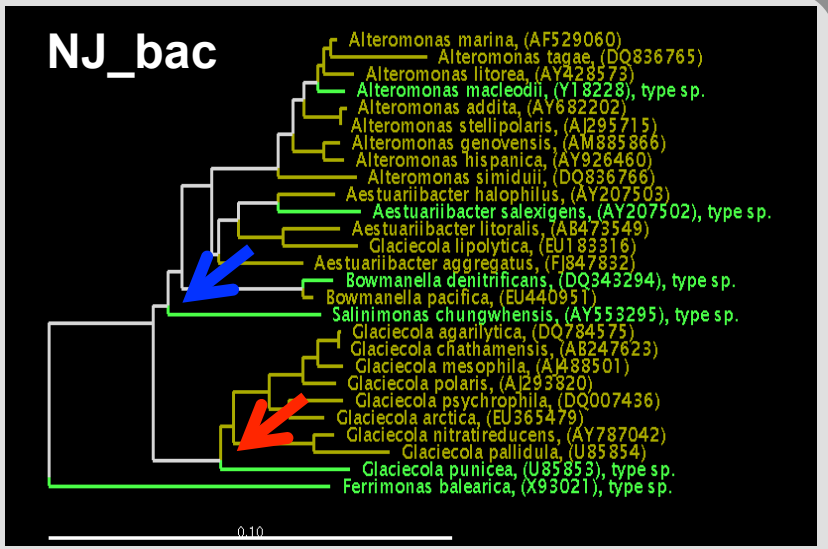
- ⇨ TERMINI ➔ 42,284 homologous positions
- ⇨ BACTERIA ➔ 1,532 homologous positions
- ⇨ 30% ➔ 1,433 homologous positions
- ⇨ 50% ➔ 1,288 homologous positions

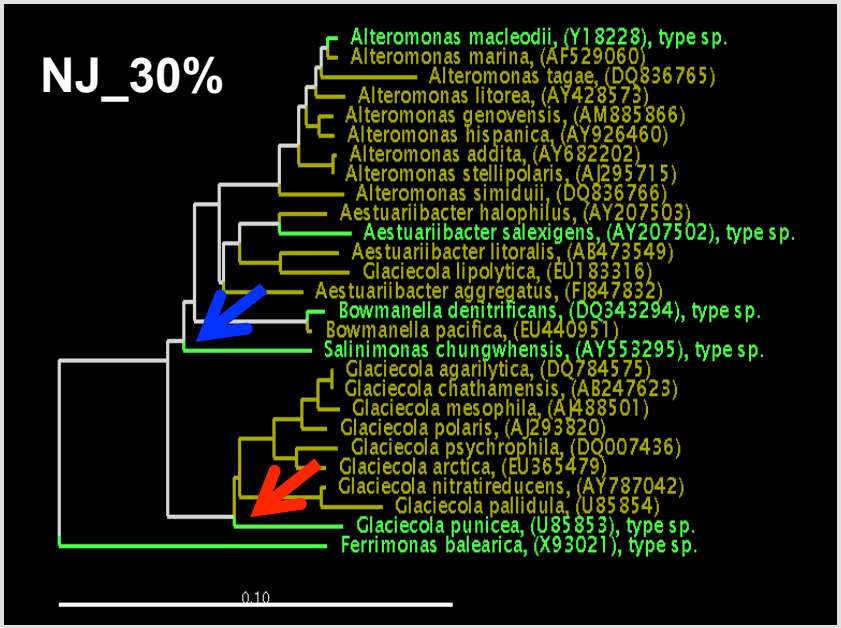


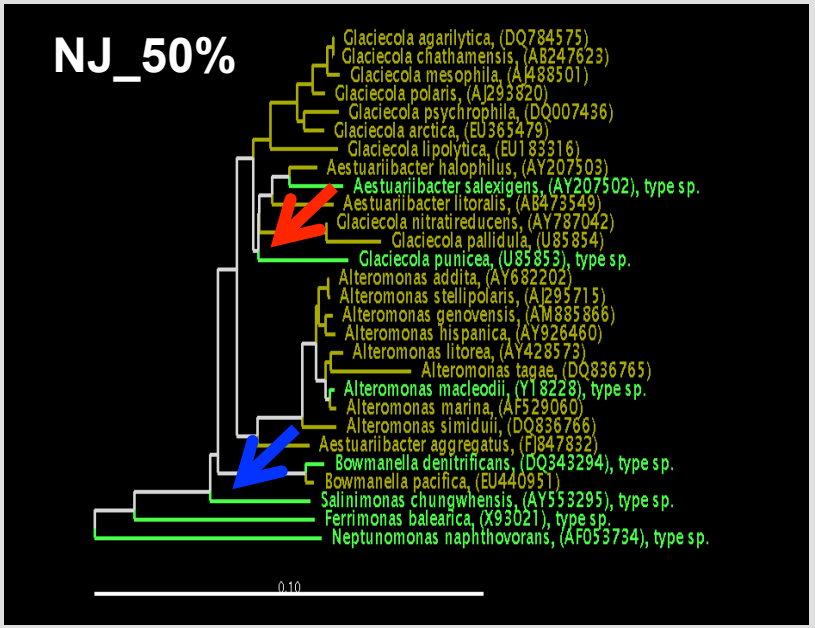
USE OF PHYLOGENETIC FILTERS

⇨ **Conservational filters are useful for deep-branching phylogenies**

⇨ **complete sequences are useful for close relative organisms**

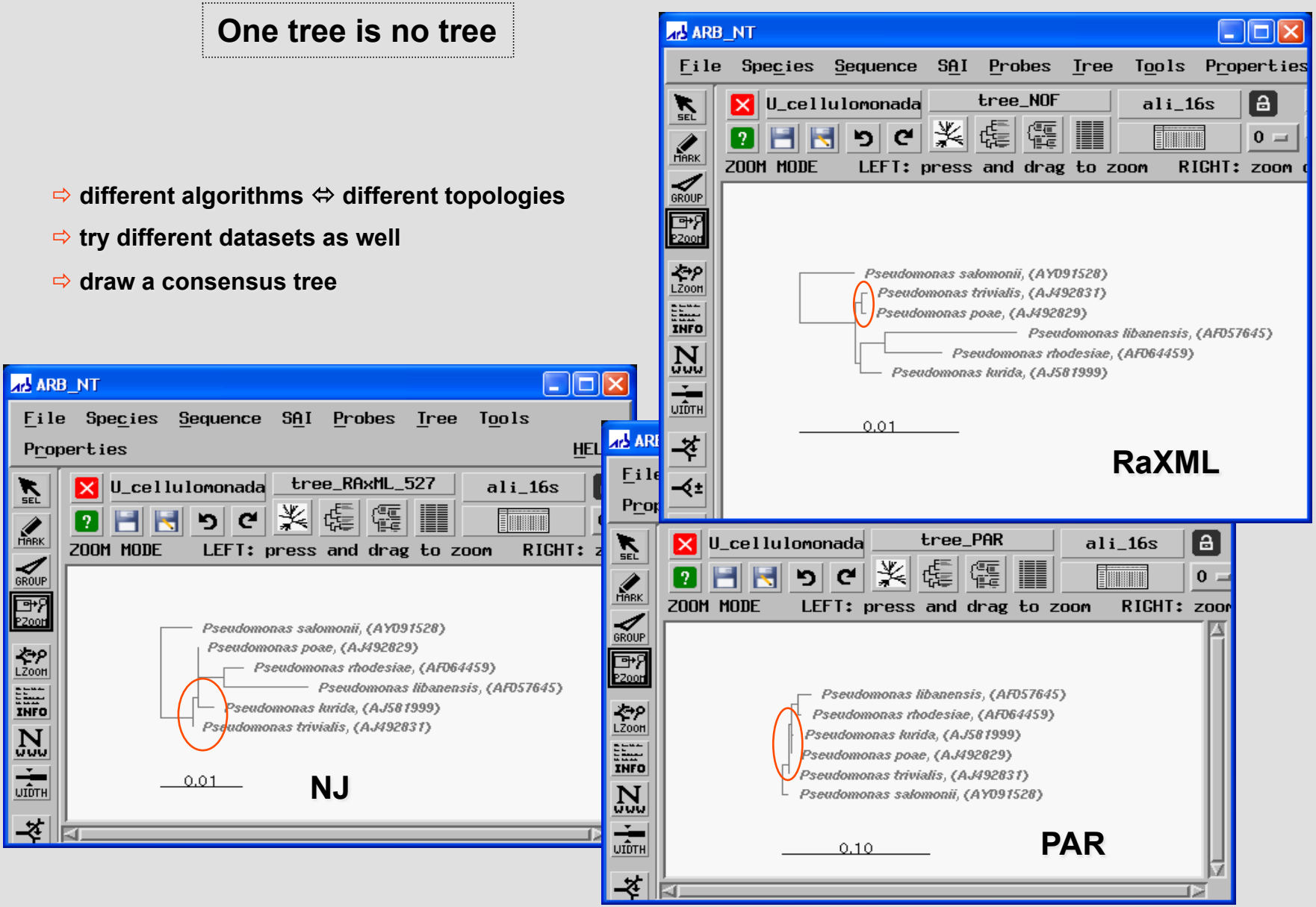






One tree is no tree

- ⇨ different algorithms ⇔ different topologies
- ⇨ try different datasets as well
- ⇨ draw a consensus tree



RECOMMENDATIONS FOR 16S rRNA TREE RECONSTRUCTION

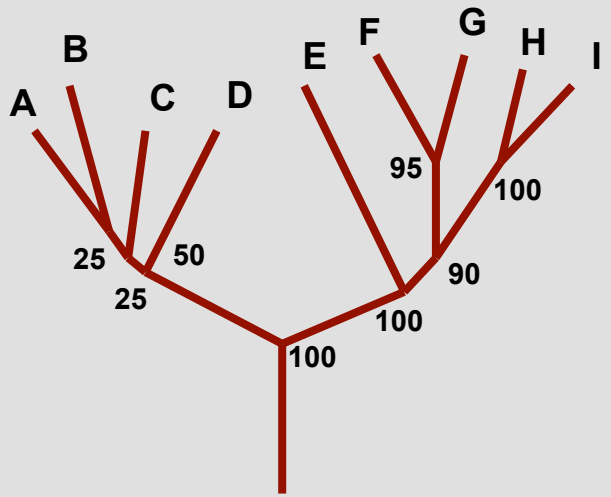
⇨ **SEQUENCE➔** almost complete is better than short partial sequences

⇨ **ALIGNMENT➔** Better take into account secondary structures

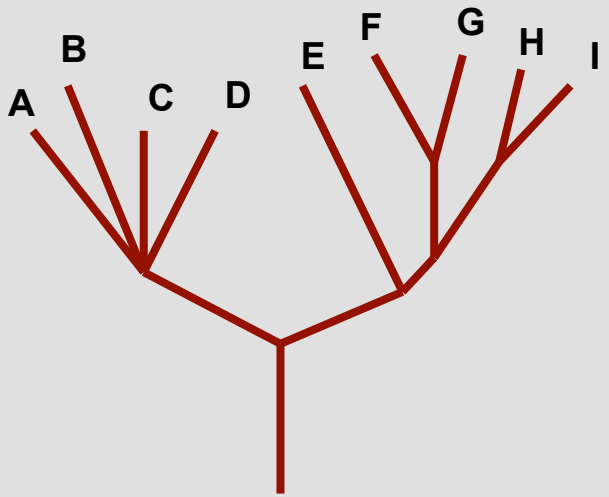
⇨ **ALGORITHM➔** Better maximum likelihood, but compare with other as neighbor joining and maximum parsimony

⇨ **DATASET➔** Never just one dataset, try different sets of data (i.e. different number of sequences; different filters to find the best resolution)

⇨ **FINAL TREE ➔** Either you show all trees, or the best bootstrapped, or a multifurcation showing unresolved branching order.



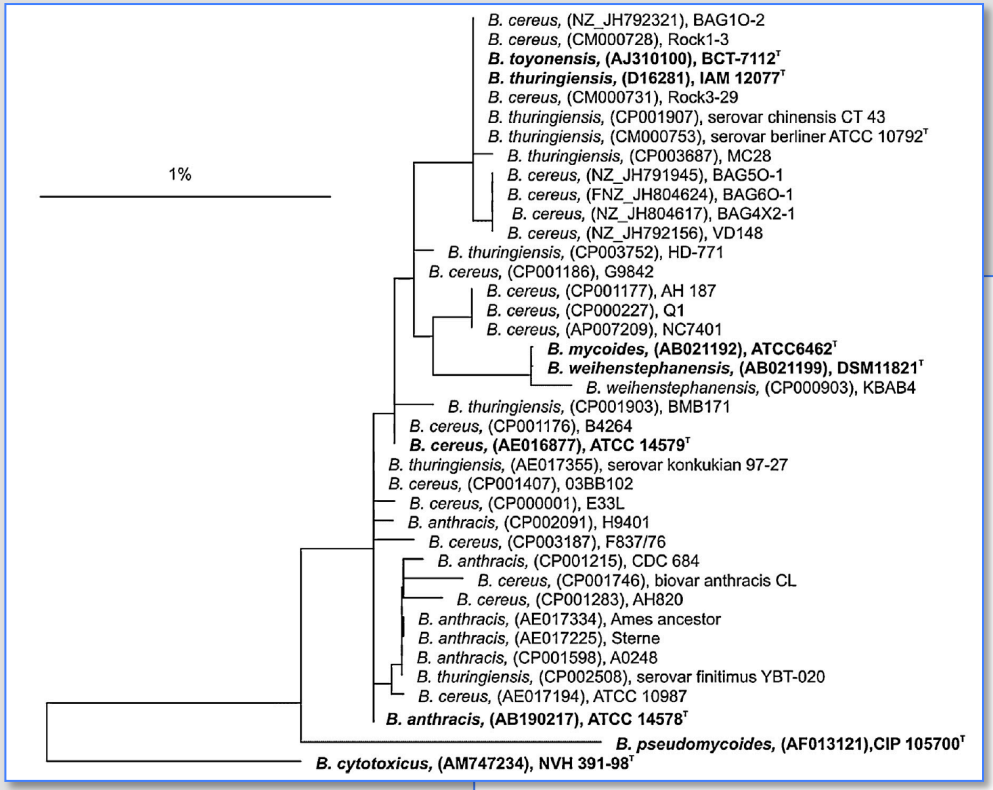
Tree with bootstrap



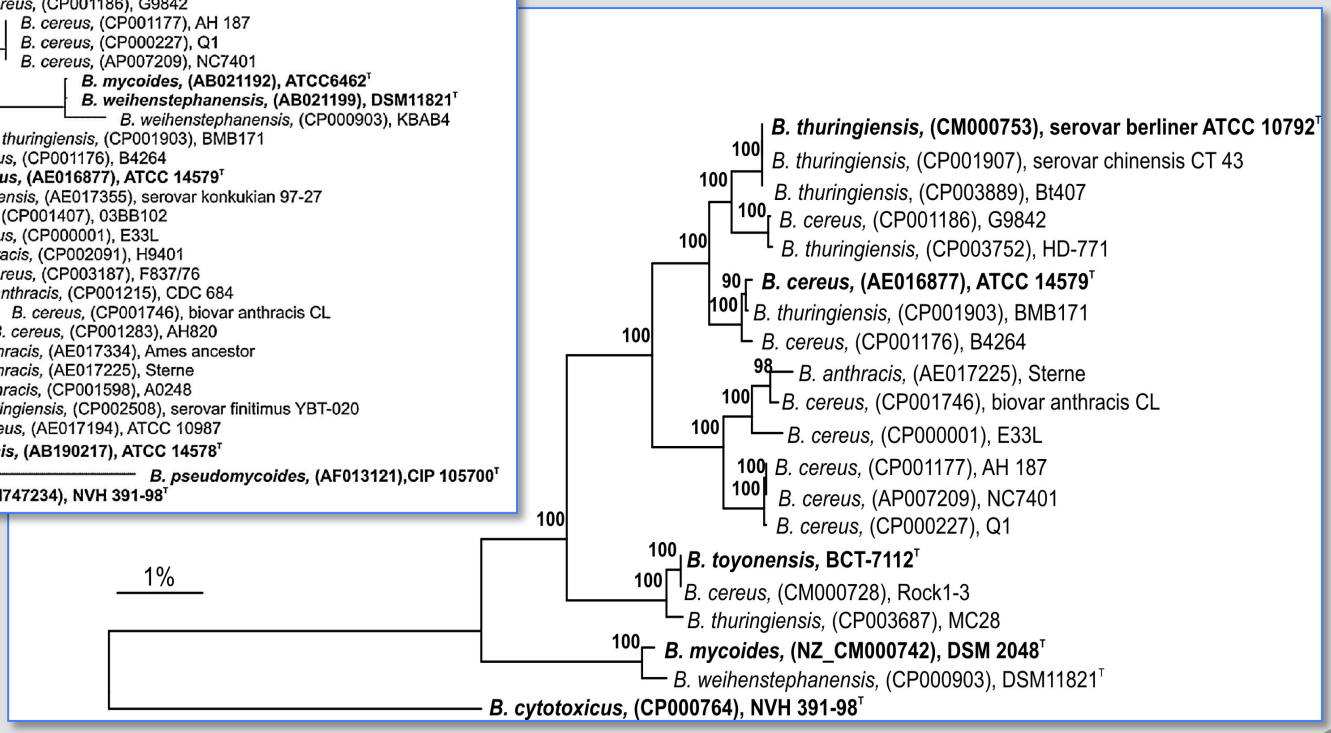
Tree with multifurcation

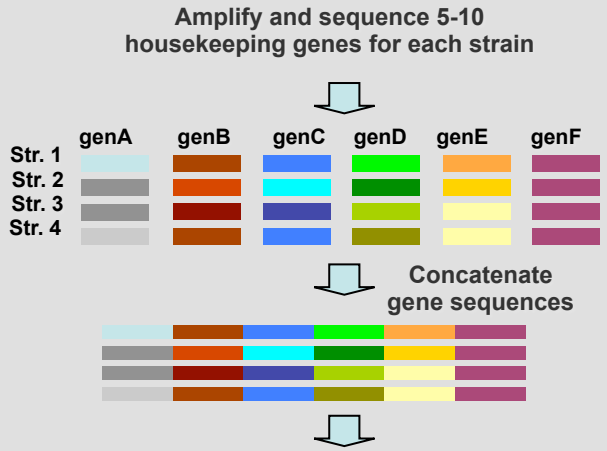
MULTIPLE SEQUENCE ALIGNMENTS

- ➡ sometimes have better resolution than the 16S rRNA gene
- ➡ 16S rRNA gene can have very low resolution



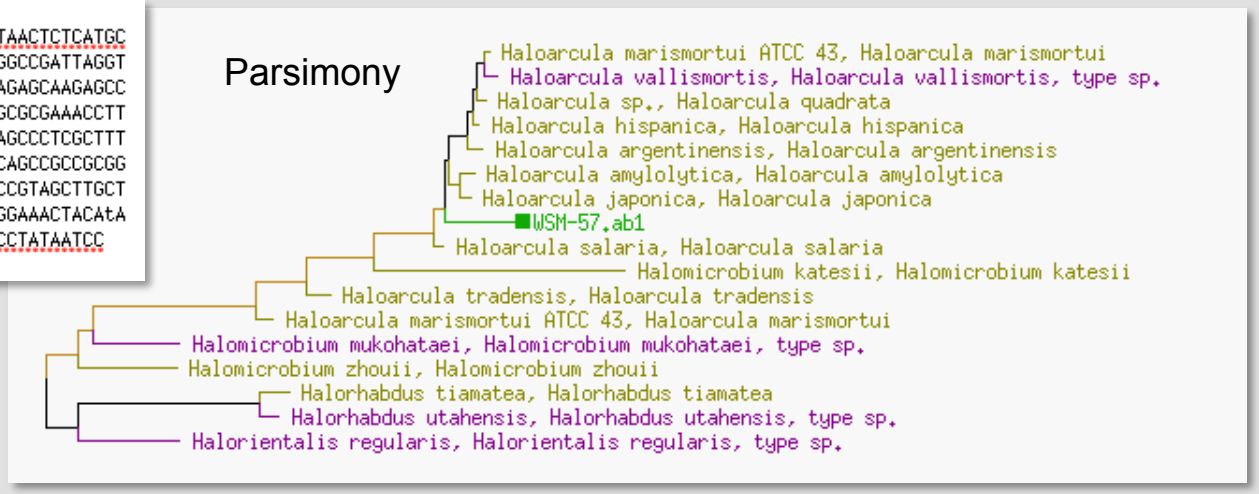
Jiménez et al., 2013, System Appl Microbiol, 36: 383- 391



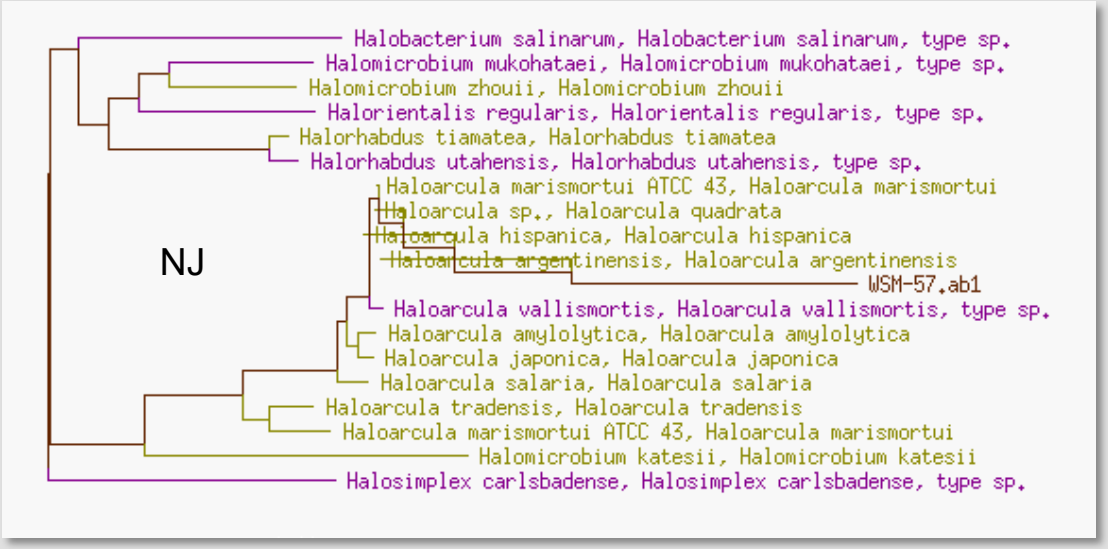


ARB PARSIMONY TOOL ⇔ PARTIAL SEQUENCES

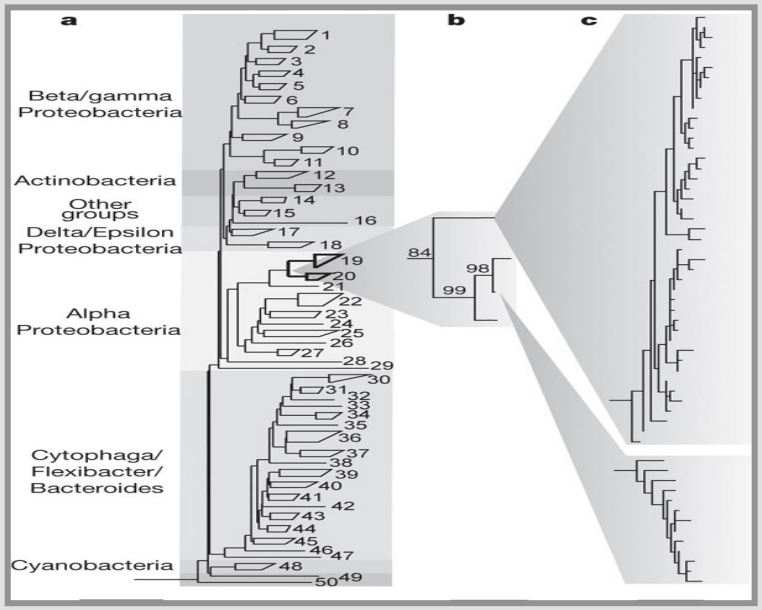
```
>WSM-57.ab1
ACCCTACAGACCGCAATAACCTCGGGAAACTGAGGCCAATAGCGGATATAACTCTCATGC
TGGAGTGCCGAGAGTTAGAAACGTTCCGGCGCTGTAGGATGTGGCTGCGGCCGATTAGGT
AGATGGTGGGGTAACGGCCCACCATGCCGATAATCGGTACAGGTTGTGAGAGCAAGAGCC
TGGAGACGGTATCTGAGACAAGATACCGGGCCCTACGGGGCGCAGCAGGCGCGAAACCTT
TACACTGCACGACAGTGCGATAGGGGGACTCCGAGTGTGAGGGCATATAGCCCTCGCTTT
TCTGTACCGTAAGGTGGTACAAGAATAAGGACTGGGCAAGACCGGTGCCAGCCGCCGCGG
TAATACCGGCAGTCCAAGTGATGGCCGATATTATTGGGCCTAAAGCGTCCGTAGCTTGCT
GTGTAAGTCCATTGGGAAATCGACCAGCTCAACTGGTCGACGTCCGGTGGAAACTACAtA
GCTTGGGGCCGAGAGACTTGACGGGTACGTCCGGGGTAGGAGTGAAATCCTATAATCC
```



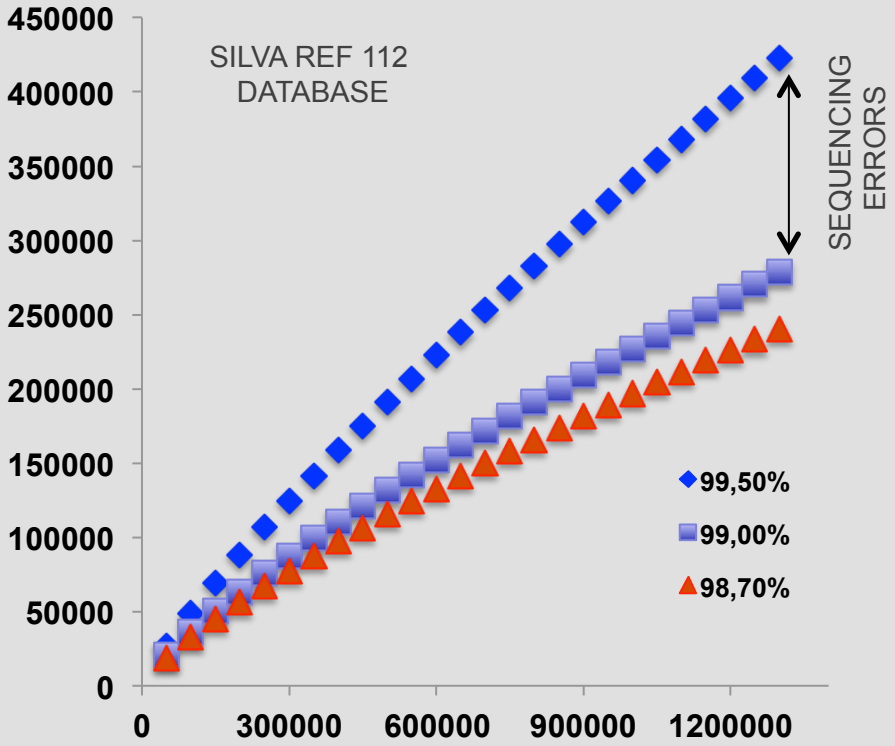
- Partial sequence of 538 bases
- Parsimony adds the sequence in the place where less mutations are necessary
- Neighbor joining reconstructs a tree only with the positions where the partial sequence has information



OTUs OPERATIONAL TAXONOMIC UNITS⇔ 97% sequence identity threshold



Acinas et al., 2004 Nature 430:551-554



Yarza et al., Nature Revs. 2014. 12: 635-645

Clone libraries

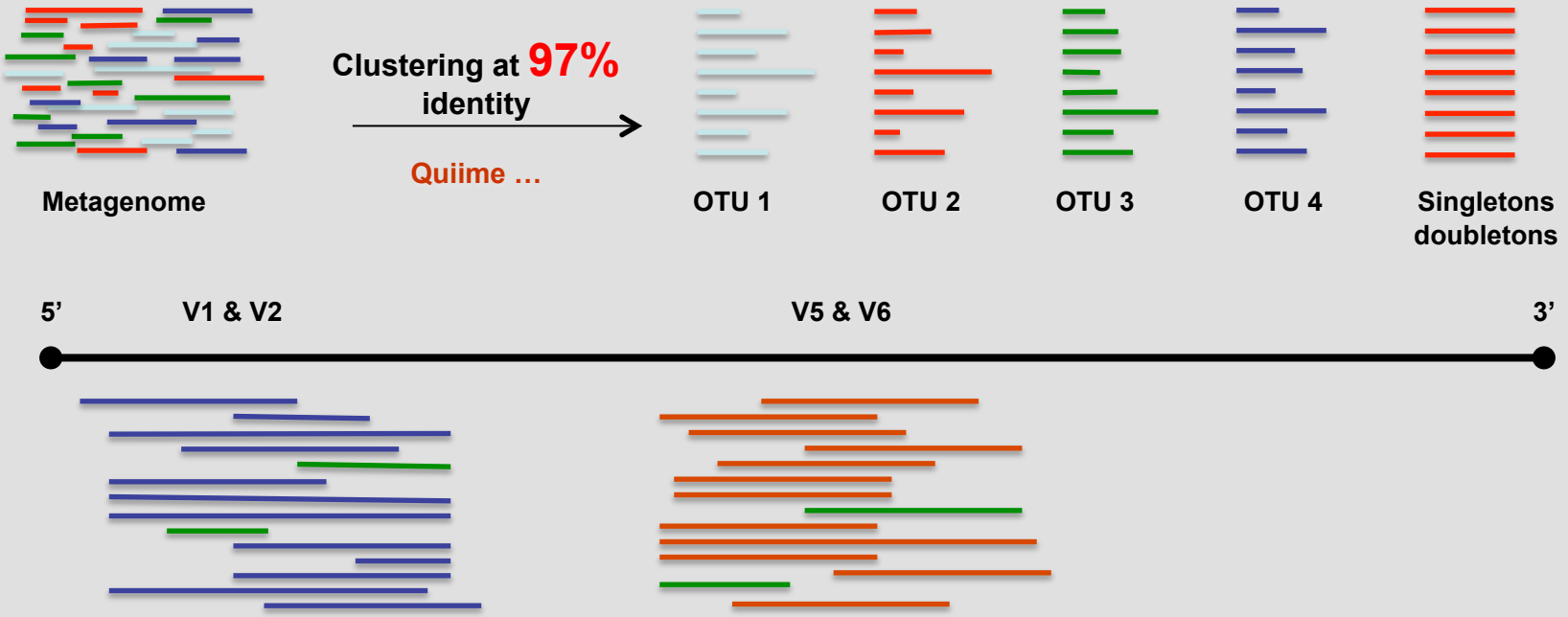
- great phylotype diversity
- PCR errors (reconditioning)
- microdiversity (several operons?)
- grouping through % identity
- **OTU (Operational Taxonomic Unit)**
- **97% one species?**

NGS: Platforms

	ADVANTAGES	DISADVANTAGES
Roche 454 SEQUENCING	Long seq ≈ 800 n (mean >550 n) Low error	Expensive About to disappear Not many labs have it
illumina	Large number of seq. In each run Low error & cheap	Too short fragments for some experiments (max 2 x 250 n)
PACIFIC BIOSCIENCES®	Long seq up to 10 kb (mean >4 Kb)	Too high error for single reads (10 -20%)
Oxford NANOPORE Technologies 	?	?

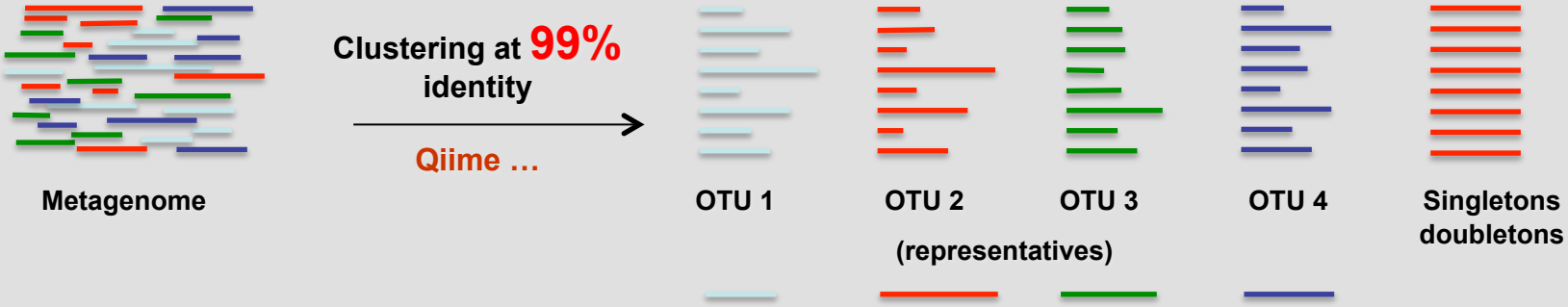
IMEDEA

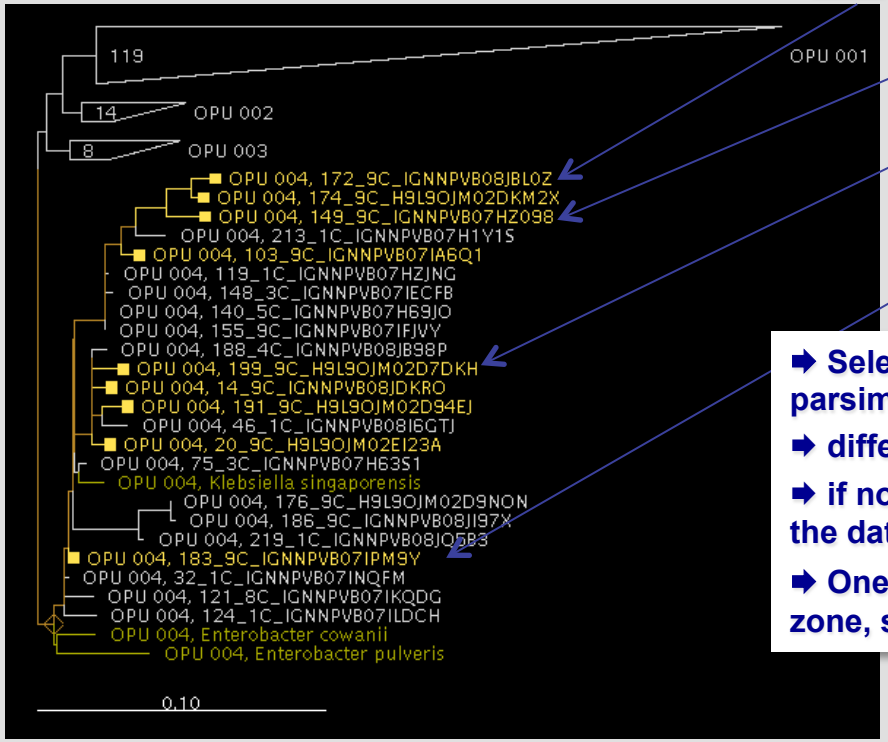
OTUs (Operational Taxonomic Units) ⇔ 454 Pyrotagging (amplicons) ≅ clone libraries



- Different groups use different variable zones (V1- V2 or V5 – V6)
- Sequences of different zones are not comparable
- identical sequences of different stretches may match different OTUs (green highlighted)
- High identity does not mean common ancestry

Phylogenetic inference ⇔ OPUs (clustering at 99% ➔ parsimony insertion)





➔ Selected sequences are inserted in a tree preexisting using parsimony inference

➔ different sizes, zones, may affiliate together

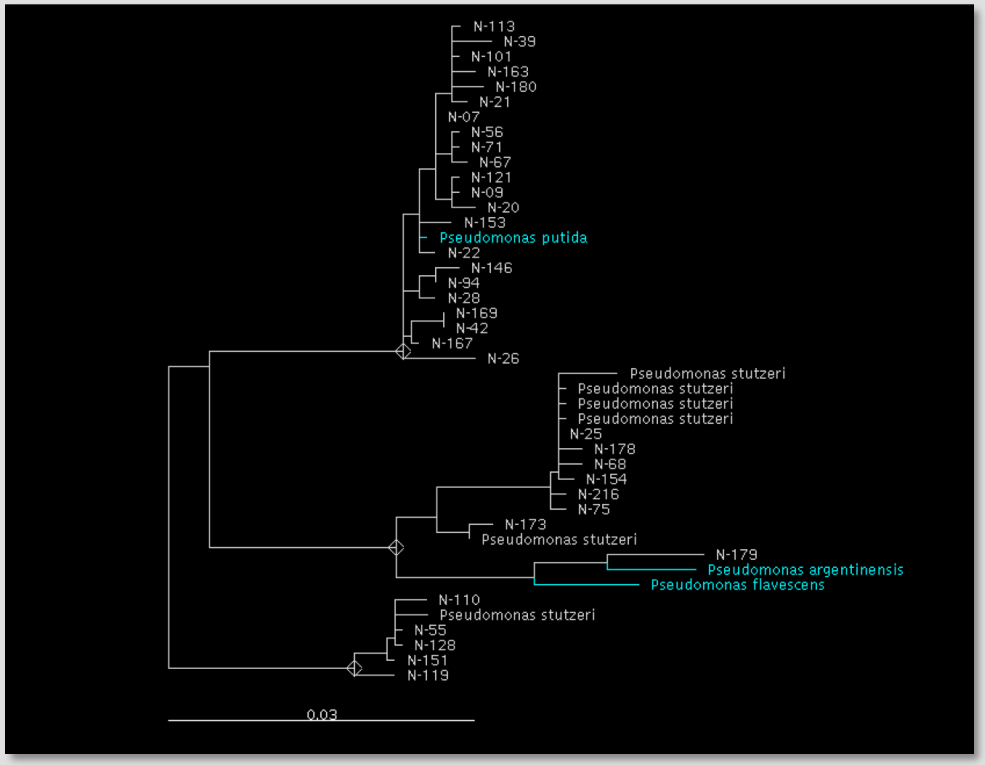
➔ if no use of ARB, one can select the best sequences from the databases matching OTUs and reconstruct properly

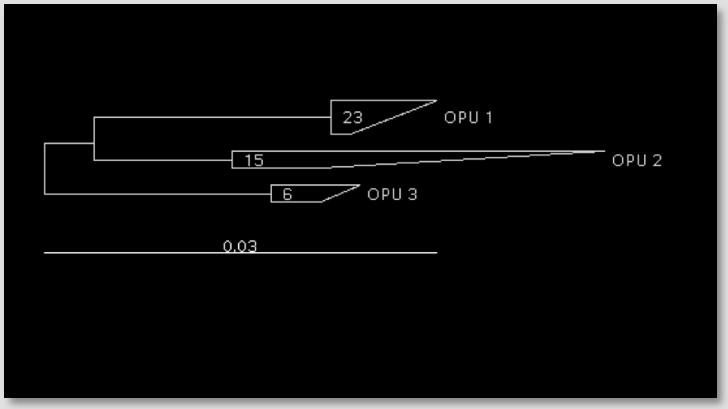
➔ One OPU may contain different OTUs of different length, zone, sample, etc…

OPUs ⇔ subjective ⇔ best solution to measure diversity

OPU ⇔ Operational Phylogenetic Unit

- similar to "Operational Phylogenetic-based Microbial Populations" (Pernthaler & Amann)
- somehow subjective, but may reflect better ecologically relevant populations





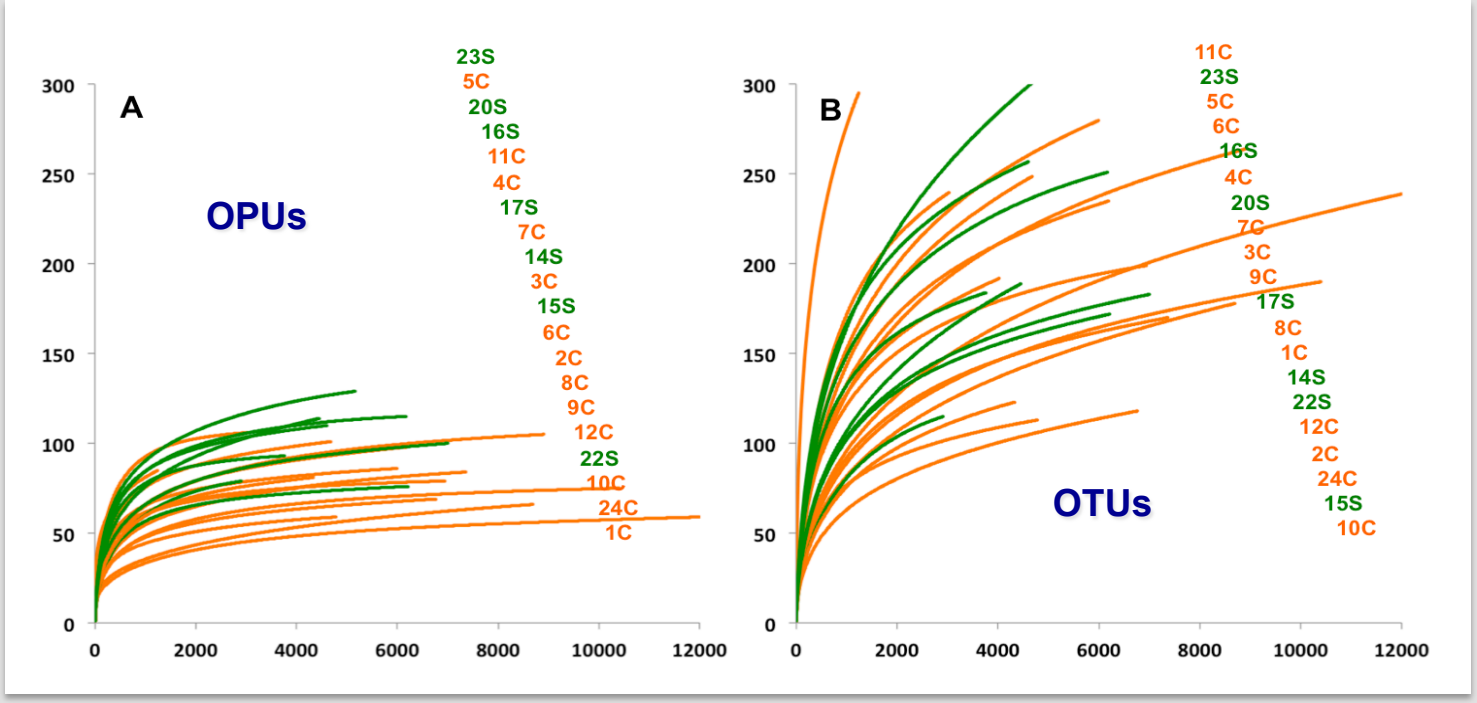
Rosselló-Móra & López-López, 2008. In: Accessing Uncultivated Microorganisms ASM Press

López-López et al., 2010 Environ Microbiol Reports 2:258-271

Pernthaler & Amann. 2005. Microbiol Mol Biol Rev 69:440-461

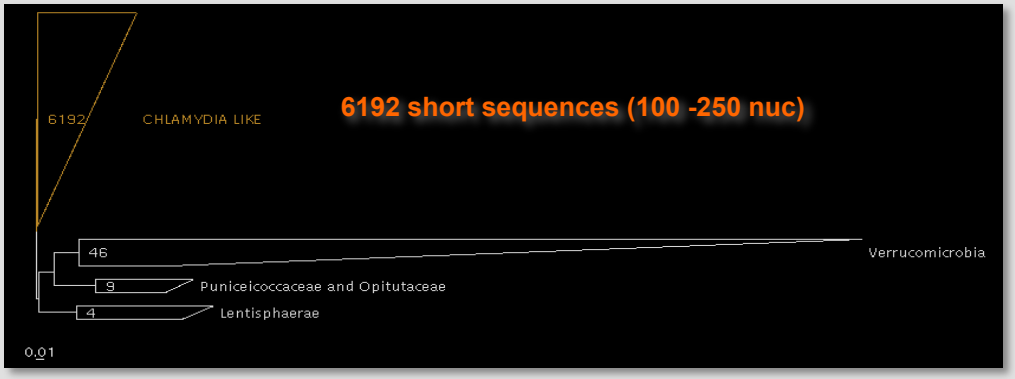
Example case ⇔ biopsies associated to Crohn microbiomes

- 20 samples (7 Healthy ; 13 Crohn)
- 150,000 sequences (mean 550 nuc; <300 nuc discarded)
- Mean of 6,592 ±2,622 in each sample
- 73% with known taxa (up to species identity)

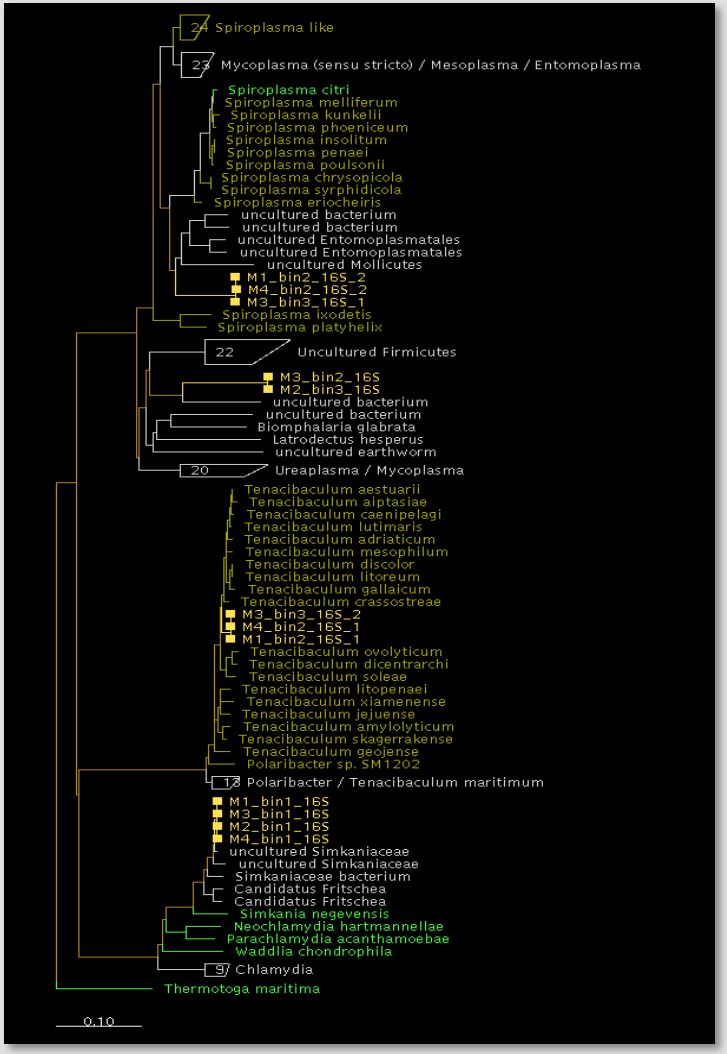


Vidal et al. Syst Appl Microbiol (2015) 38, 442-452

Fragments extracted from metagenome data



16S rRNA (1600 bp)



- direct sequencing (Illumina) renders extraced 16S rRNA sequences of 100 – 250 nuc
- Parsimony approach allows affiliating to the same branch despite cover different parts of the gene
- if the sequencing is deep enough, when binning the almost full gene sequence is retrievable

RECOMMENDATION PIPELINE

OPUs ⇔ tedious, but more robust

Sequences

(trimming & removing chimeras)

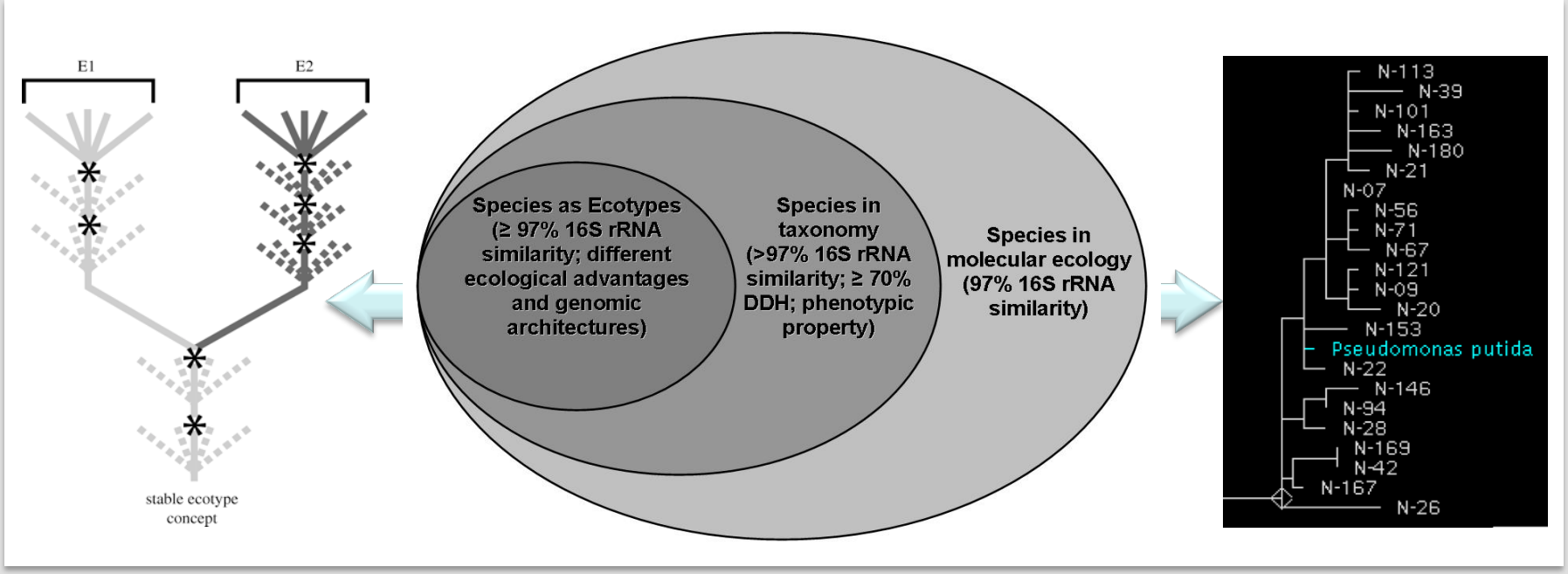
OTUs

(grouping at 99% identity, high restrictive threshold)

OPUs

(alignment, phylogenetic inference [parsimimony])

16S rRNA: distinct disciplines stays in different thresholds



Ecotype; early stage of speciation

A stable framework needs PRAGMATISM

Rosselló-Móra, 2011 Environ Microbiol 14:318-334

Compile microdiversity into OTUs at 97% identity