Metagenomics: Building a taxonomic markers collection

Jean-François Taly, Christine Oger, Jean-Pierre Flandrois and Guy Perrière
France Génomique: WP Métagénomique
30/06/2015

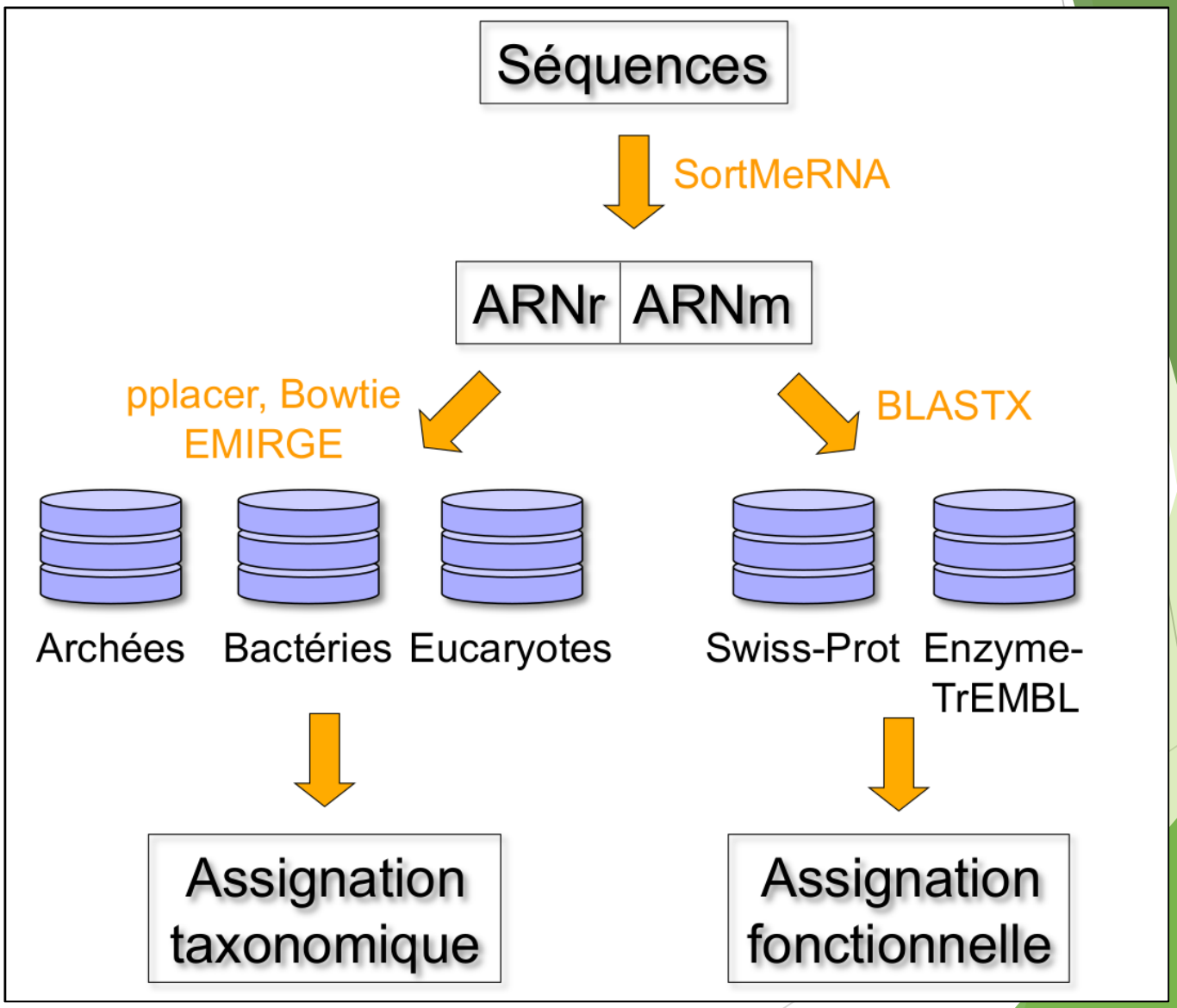
LBBE
BIOMETRIE ET BIOLOGIE EVOLUTIVE

prabi
Pôle Rhône-Alpes de Bioinformatique

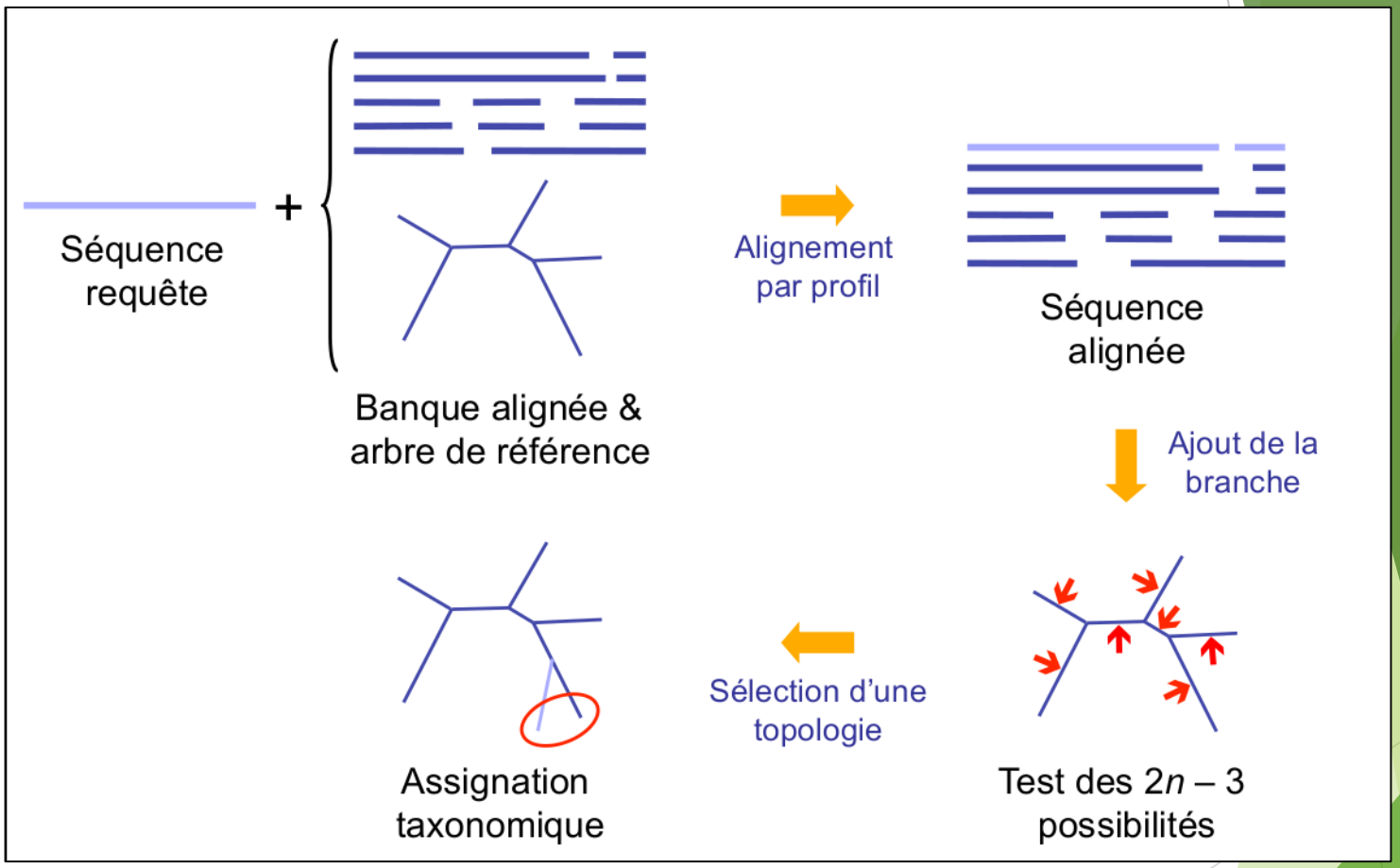
FRANCE GÉNOMIQUE

Previous work @ PRABI

François Bartolo & Clément Lionnet 2012-2014



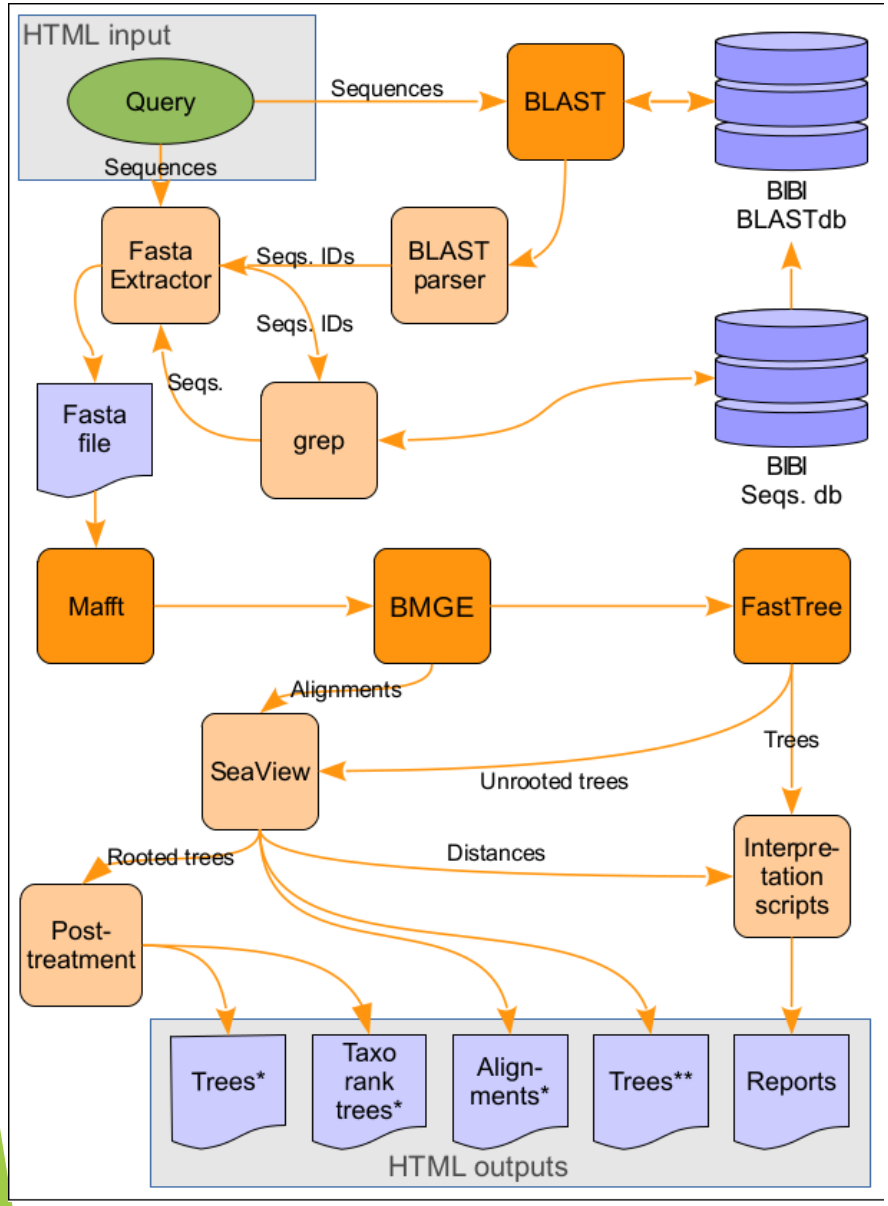
Phylogenetic Replacement



Must have reference trees

- Large trees are difficult to build

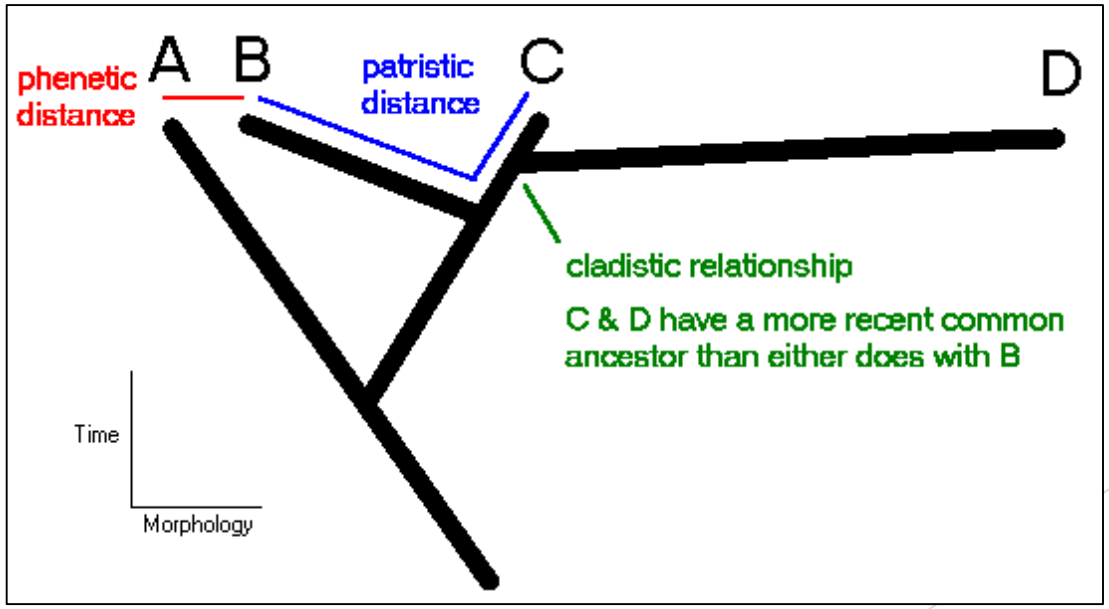
LeBiBi^{QBPP} identification engine



- Phylogeny on the fly
  - Recruit ref. seq. with Blast
  - MSA with MAFFT
  - Clean the MSA with BMGE
  - Build a FastTree
  - Report the closest ref. in term of patristic distance

Patristic Distance

- Sum of branch lengths
- Branch lengths are the number of expected mutations per site.



http://www.mun.ca/biology/scarr/Phenetic_Patristic_Cladistic.html

LeBiBi server

https://umr5558-bibiserv/lebibi/lebibi.cgi

Results

Query data

Database: **procaryota_SSU-rDNA-16S_stringent**
Note: The use of the stringent database (all the GenBank sequences for a given species) may lead to errors due to the low quality identification of some sequences, a careful analysis is n
Original name of the query (from fasta) : **QRY_water34_16S.** leBiBI ref : pqpU72lhp

Sequence composition

Length of Query sequence is : 1274
Nucleotide Composition 318 A - 247 T - 302 C - 407 G - **0 N 0.0 % Gaps % =0.0**
GC%=0.55

Quality of the BLAST analysis

The number of Blast hits is **OK**
All is OK, the best BLAST hit length **(1274)** is =100.% of the query length

Biodiversity level

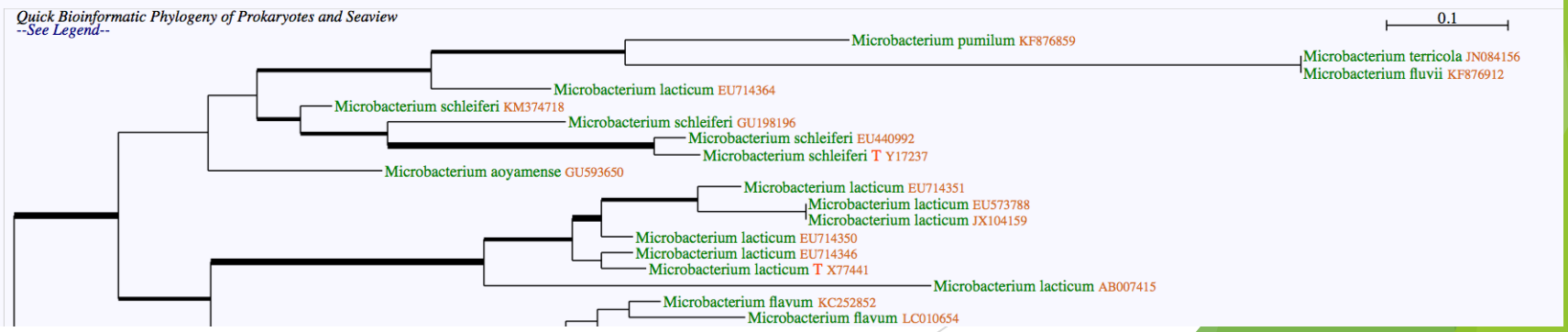
All extracted sequences appear to belong to Bacteria and their lowest shared taxon is: Bacteria-Actinobacteria-Actinobacteridae-Actinomycetales-Micrococcineae-Microbacteriaceae-M
CAUTION ! One genus only in the extracted sequences. No outgroup *may* be pertinent enough for a good phylogeny. You may need to increase the number of desired sequences or use

Phylogenetic tree analysis

Proximal Cluster **['Microbacterium_aurum EU373396', 'Microbacterium_aurum T Y17229', 'Microbacterium_aurum KJ127516', 'Microbacterium_aurum EU714355']**

ANALYSIS OF PATRISTIC DISTANCES

Microbacterium_aurum T Y17229 is the closest sequence based on patristic distances
On the basis of patristic Distances, the **closest** sequence is: Microbacterium_aurum T Y17229 AND is in the proximal cluster
Remark: patristic distance to Microbacterium_aurum T Y17229 is in the 75th percentile of the Microbacterium_aurum intra-species patristic distances (but this is dependant on the conf
Remark: the patristic distance to Microbacterium_aurum T Y17229 is in the 75th percentile of the Microbacterium genus inter-species patristic distances : the sequence may correspond
The closest sequence based on patristic distances (Microbacterium_aurum T Y17229) is NOT the FIRST BLAST hit.



Recipe for an efficient taxonomic assignment

1. A Taxonomic Classification
   - Species delimitation is crucial
2. Types representative of each taxon
   - Reduce biological and methodological variability
   - Access to a physical specimen
3. Biological sequences
   - In agreement with the taxonomic relationship
4. A Method of assignment
   - Efficient way to map reads to the references

Taxonomy

- Rank-based classification of Organisms
  - Originally popularized by Carl Linnaeus
- Codes of nomenclature
  - formalized the protocols and requirements for publication of scientific names
  - Judicial commissions decision for new species
  - Several codes maintain different lineages

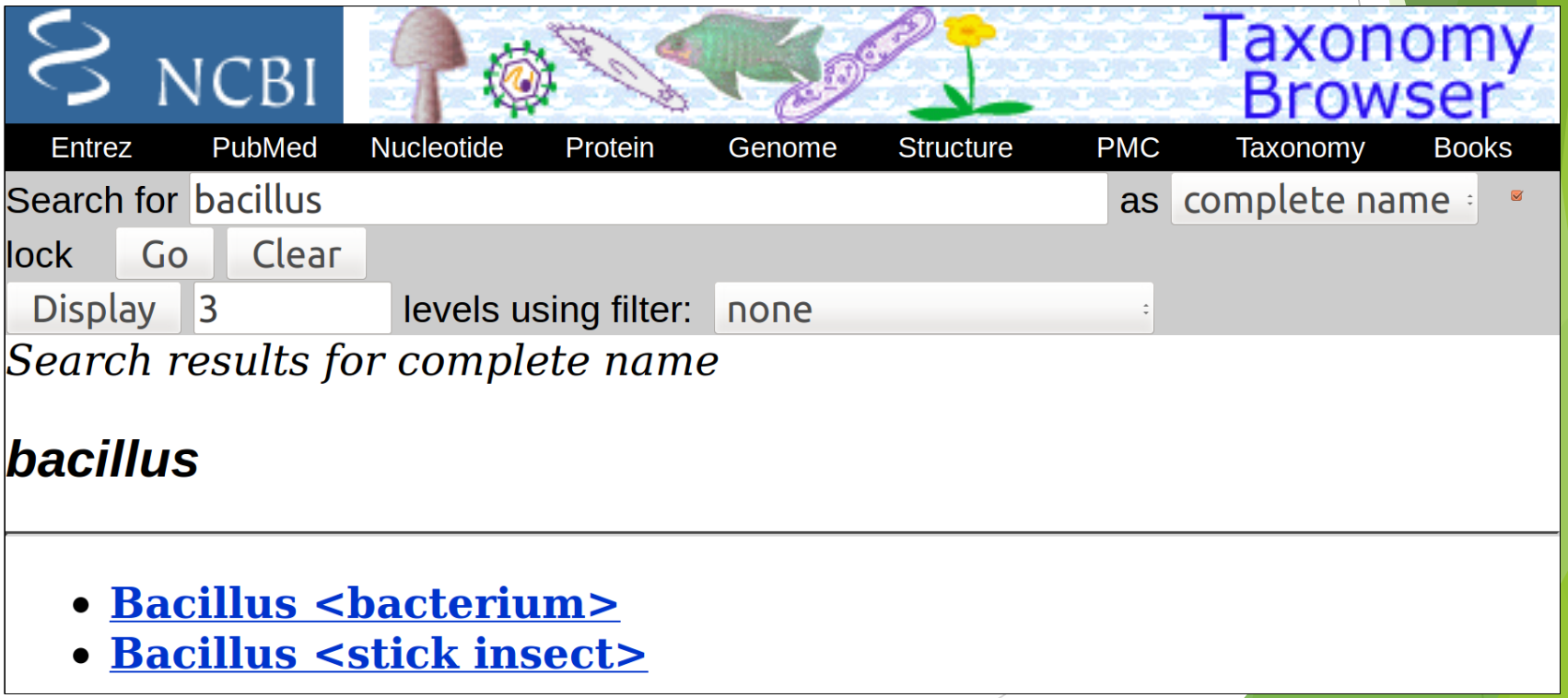
Table 1. The Codes of nomenclature

ICZN	International Code of Zoological Nomenclature http://iczn.org/
ICN	International Code of Nomenclature for algae, fungi, and plants http://www.iapt-taxon.org/nomen/main.php
ICNB	International Code of Nomenclature of Bacteria http://www.ncbi.nlm.nih.gov/books/NBK8808/
ICTV	International Committee on Taxonomy of Viruses http://www.ictvonline.org/
ICNCP	International Code of Nomenclature for Cultivated Plants http://www.ishs.org/sci/icracpco.htm
CTPPB	Committee on the Taxonomy of Plant Pathogenic Bacteria http://www.isppweb.org/about_tppb.asp
BioCode	an alternative universal code http://www.bionomenclature.net/
PhyloCode	a newer alternative universal code http://www.ohio.edu/phylocode/

Federhen, NAR 2015

Code: fastidious but necessary

- The name “Bacillus” has been given to
  - A gender of bacteria with the identifier 1386
  - A gender of insect with identifier 55087



NCBI Taxonomy

- Pragmatic base gathering all taxonomies in one
- Serves as the standard nomenclature and classification for the International Sequence Database (INSD)
- Not generated automatically from sequences
- Incorporate phylogenetic and taxonomic knowledge from a variety of sources

Ranks:	higher taxa	genus	species	lower taxa	total
Archaea	145	144	533	0	822
Bacteria	1393	2685	13625	841	18544
Eukaryota	20866	69896	312932	23405	427099
Fungi	1570	4767	30696	1147	38180
Metazoa	14972	47430	154977	11728	229107
Viridiplantae	2702	14836	117701	10256	145495
Viruses	634	481	2363	0	3478
All taxa	23068	73213	329484	24246	450011

NCBI Taxonomy statistics 04/2015
No informal names, No uncultured

Types and Specimen

- For each species have been defined a type specimen:
  - Carl Linnaeus is the type of *Homo sapiens*

- For databases, it is interesting to search for the type sequence of the type strain
  - Example with Escherichia coli
    - 362 strains (2162 sequences)
    - Type strain is DSM 30083
    - Type sequence is X80725

Biological Sequences

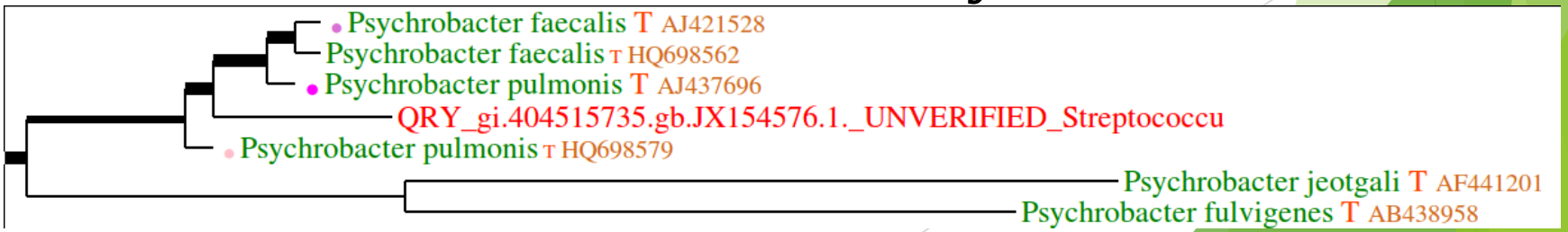
- International Sequence Database:
  - Genbank @ NCBI
  - ENA @ EBI/EMBL
  - DDBJ @ NIG
- "*Genbank relies on the submitters for the correct original taxonomic identification of their specimen*"
  - Sequences may be not or misannotaded
- RefSeq is a manually curated Genbank
  - Representative sequences for each type strain
  - Only a fraction of Genbank is covered

JX154576: A case of wrong annotation

Genbank Card

```
LOCUS       JX154576                868 bp    DNA     linear   BCT 22-SEP-2012
DEFINITION  UNVERIFIED: Streptococcus agalactiae strain IR6 16S ribosomal RNA
            gene, partial sequence.
ACCESSION   JX154576
VERSION     JX154576.1  GI:404515735
KEYWORDS    UNVERIFIED.
SOURCE      Streptococcus agalactiae
  ORGANISM  Streptococcus agalactiae
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 868)
  AUTHORS   Pourgholam,R., Laloei,F., Saeidi,A. and Zahedi,A.
  TITLE     Comparative genomics for identification of Streptococcus sp. by
            nucleotide sequence analysis in Iran
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 868)
  AUTHORS   Pourgholam,R., Laloei,F., Saeidi,A. and Zahedi,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JUN-2012) Biotechnology, Ecology Research Center of
            the Caspian Sea, Farah Abad, Sari, Mazandaran 4847158948, Iran
COMMENT     GenBank staff is unable to verify source organism provided by the
            submitter.
```

LeBiBi^{QBPP} classifies it as a Psychrobacter



rDNA Databases

	RDP	SILVA	LeBiBi QBPP
Markers	SSU (16S) and LSU (Fungal 28S)	SSU (16/18S) and LSU (23/28S)	Prokaryotes: SSU, LSU, rpoB, groEL2 … 12 house-keeping genes
Reference	Cole *et al.* NAR 2014	Yilmaz *et al.* NAR 2014	Flandrois *et al.* under revision
Update periodicity	6 months	12 months	6 months
# 16S sequences	3,019,928	1,454,141	1,341,288

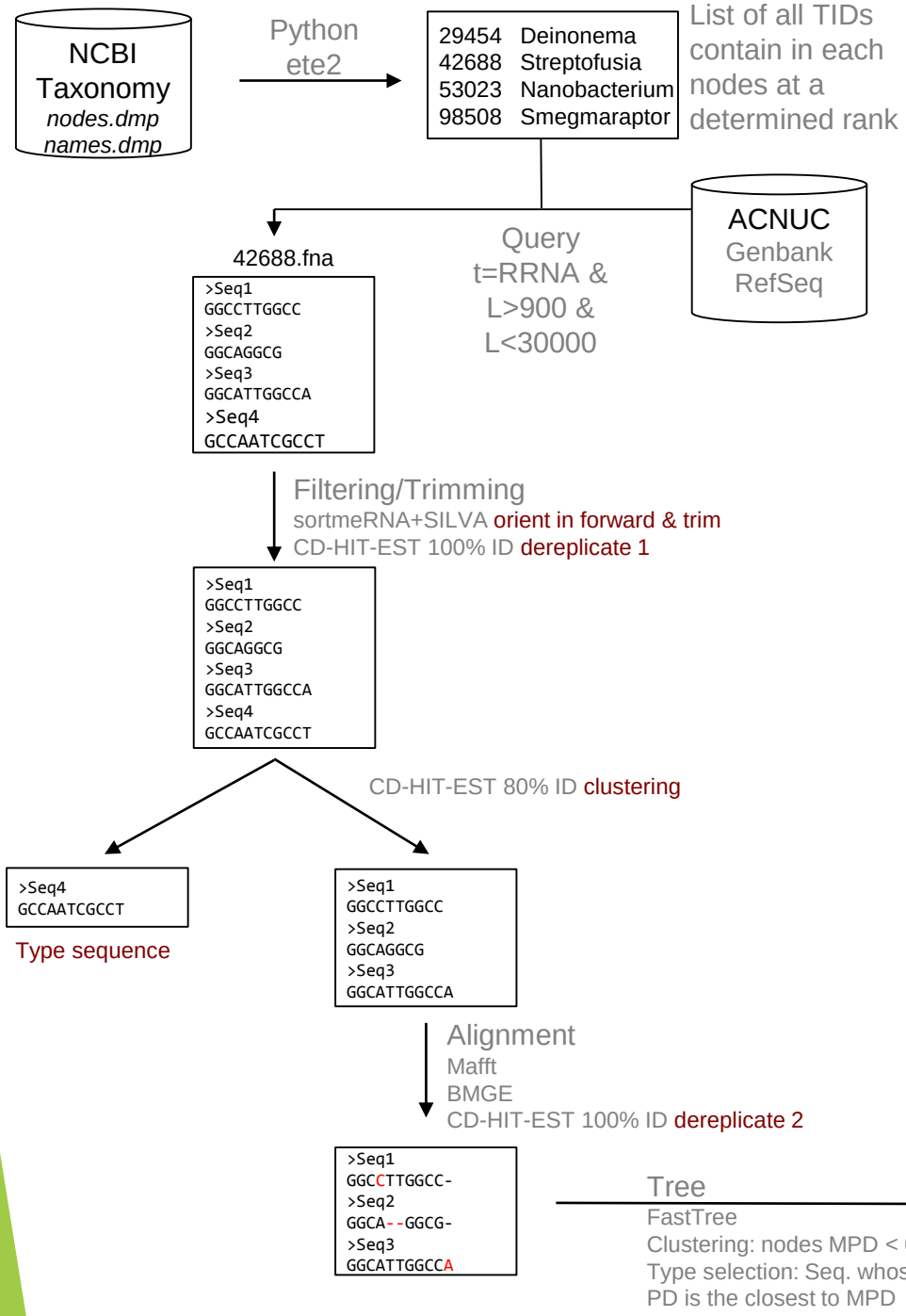
LeBiBi^{QBPP} blastDB

Quick BioInformatic Phylogeny of Prokaryotes

Stringency	Description	# Sequences	# Species
Lax	All 16S rRNA >300bp from Genbank	1,341,288	43,959
Stringent	Sequences with a correct species name from Lax	240,655	12,326
Stringent TS	One type strain per species	21,950	12,322
Super Stringent	One or more type sequences for each type strain	11,631	11,312

Objectives of the automated DB

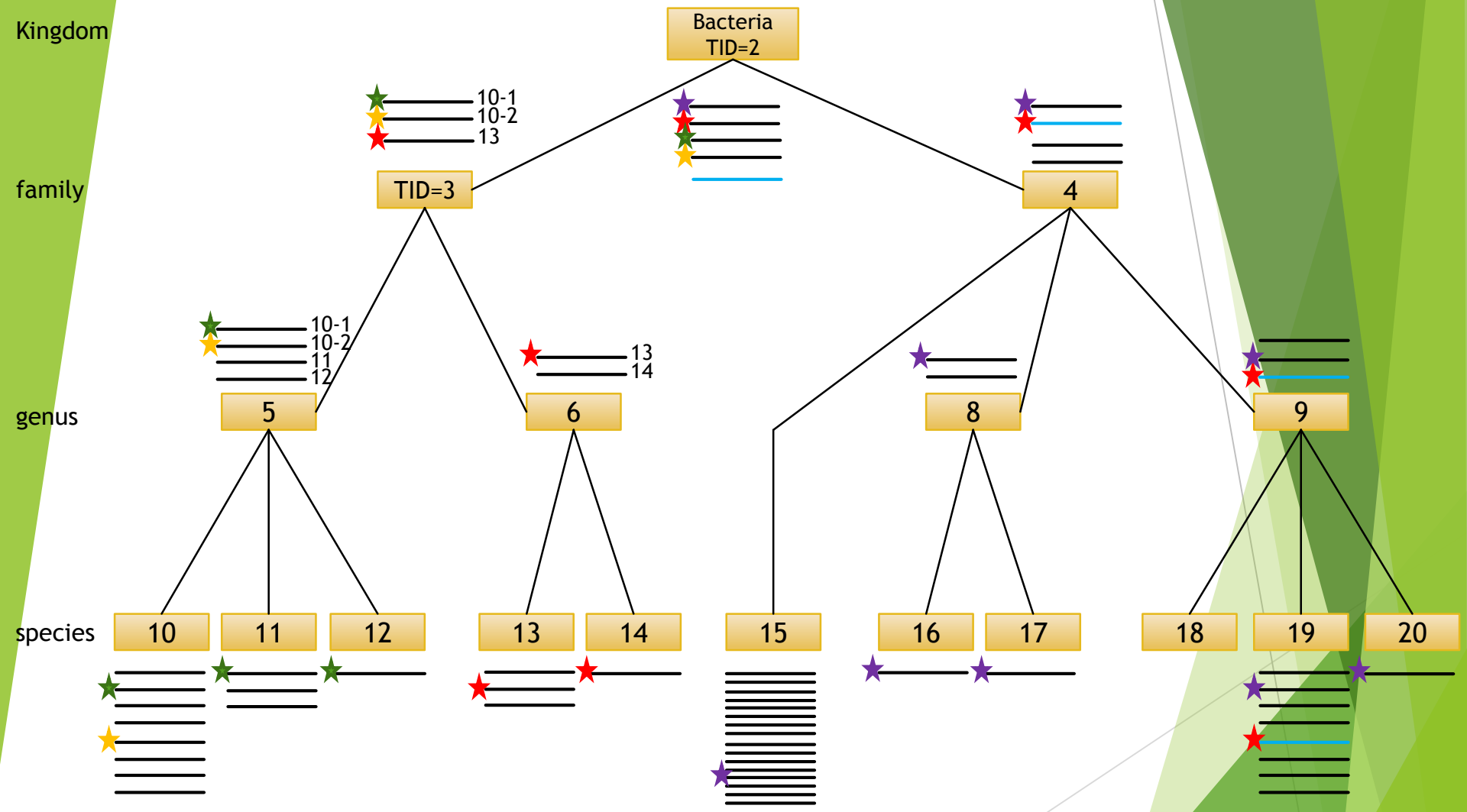
- Frequent Updates:
  - NCBI taxonomy changes every day
- Reannotation:
  - Using the NCBI taxonomy as a guide
- Automatic types selection:
  - Median patristic distance of a taxon
- Eliminate sequencing errors:
  - Singletons



For every species

- Get seq. with ACNUC
- Select 16S rRNA
- Remove duplicates
- Build clusters 80% Id
- MSA with Mafft
- Build a tree: gtr + gamma
- Parse the tree from root and create a group if node:
  - median patristic dist < 0.01
  - Bootstrap >90
- Types = seq. with mean PD the closest to MPD

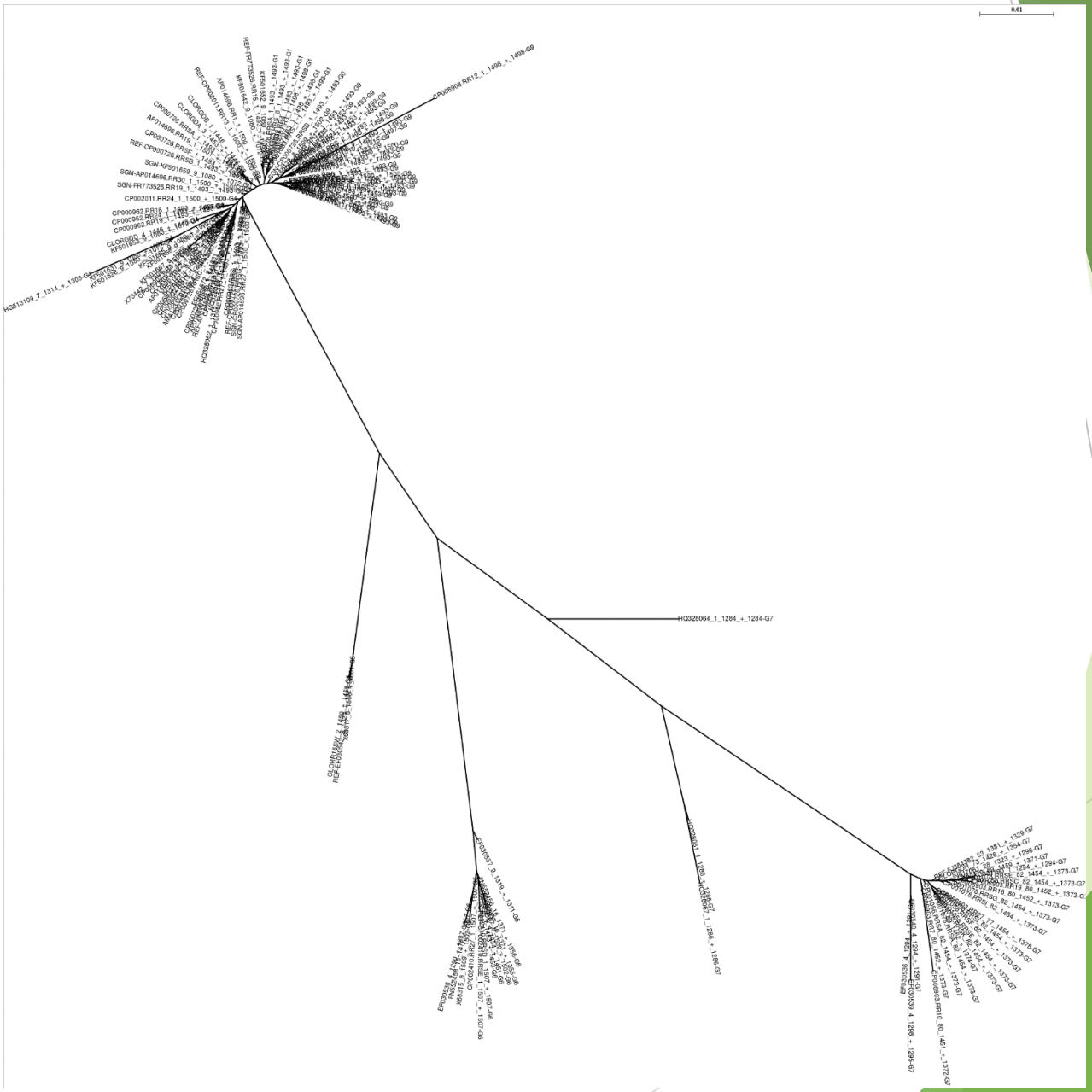
Guided reannotation



Example: Genus Clostridium

- 1910 different Taxonomic Identifiers (TIDs) at the species level
  - Do not mean 1910 different species!
- Clostridium botulinum
  - All species producing a toxin provoking a flaccid paralysis
  - 320 16S-rRNA sequences >900nt
  - 140 non identical sequences
  - 17 representative sequences after processing

Clostridium Botulinum (TID:1491)



http://www.vetbact.org

No. of matching species/subspecies etc: 5	Id	Antal baspar	16S rRNA acc-nr	rRNA op.
Clostridium botulinum, group I	202	3 863 450	L37585	8
Clostridium botulinum, group II	203	3 659 644	L37592	11
Clostridium botulinum, group III	24	2 773 157	L37590	10
Clostridium botulinum, group IV	204		X68316	

16S rRNA Seq.:

Acc-no	Strain	Number of NT	Operon
L37585	ATCC 25763^T	1 453	8

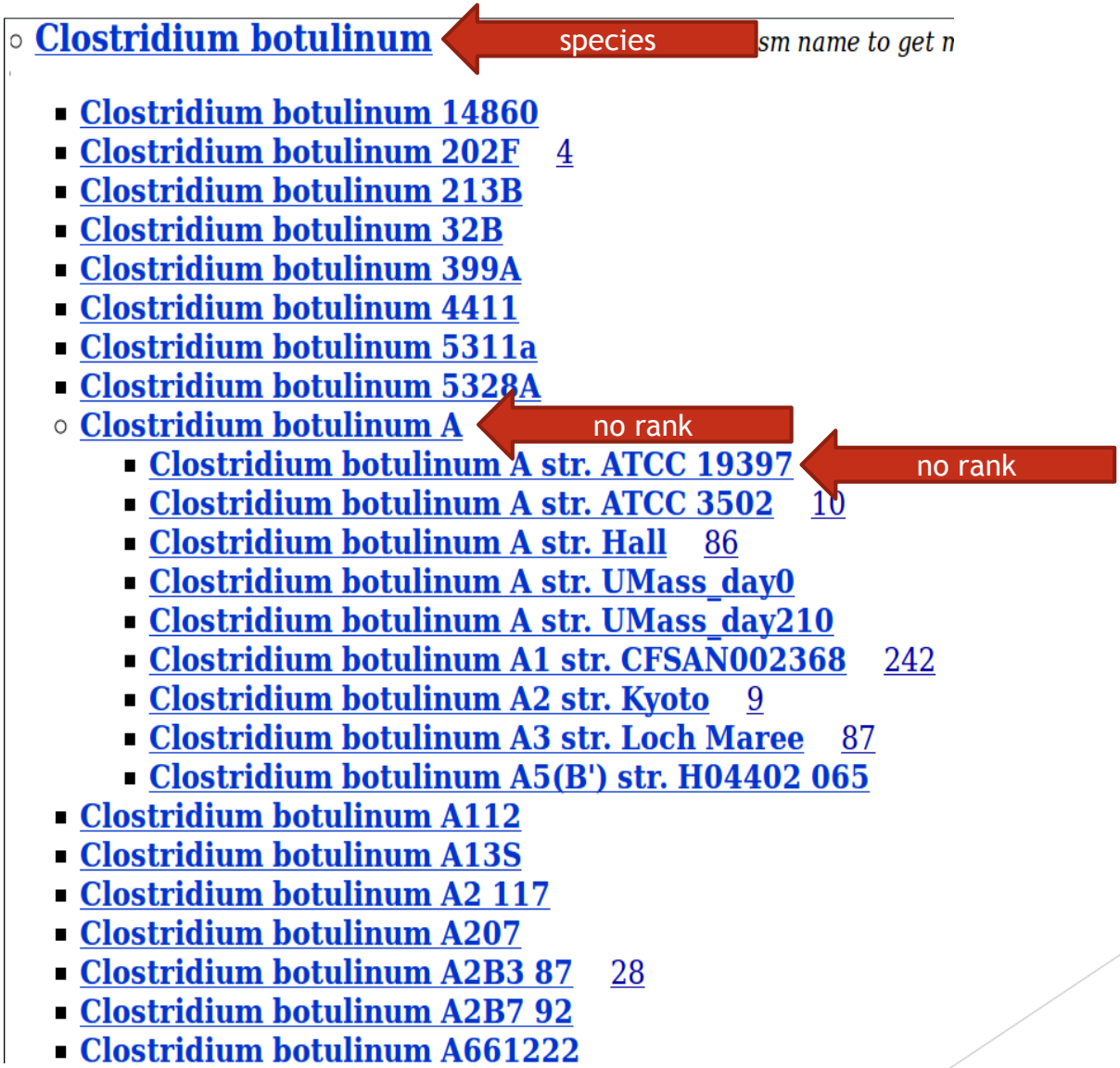
Taxonomy/phylogeny: ***C. botulinum* can be classified into four different phenotypic groups: I-IV.** *C. botulinum*-strains within group I are most closely related to *Clostridium sporogenes* and *Clostridium putrificum* and not to any of the other three phenotypic groups of *C. botulinum*.

Virulence Factors: *C. botulinum* strains within group I produce **botulinum toxin type A, B or F**

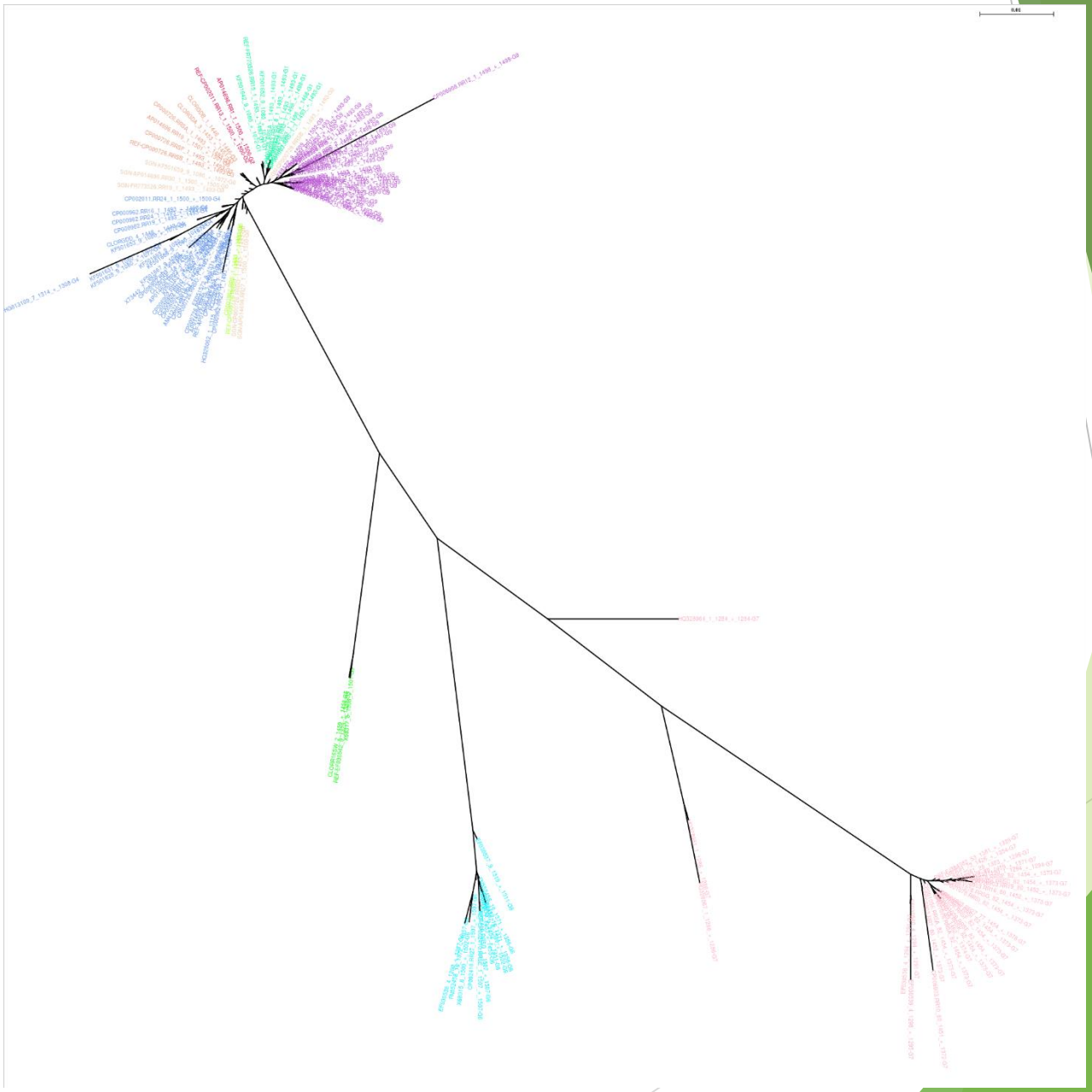
Virulence Factors: *C. botulinum* strains within group II produce **botulinum toxin type B, E or F**

Clostridium Botulinum - NCBI taxonomy

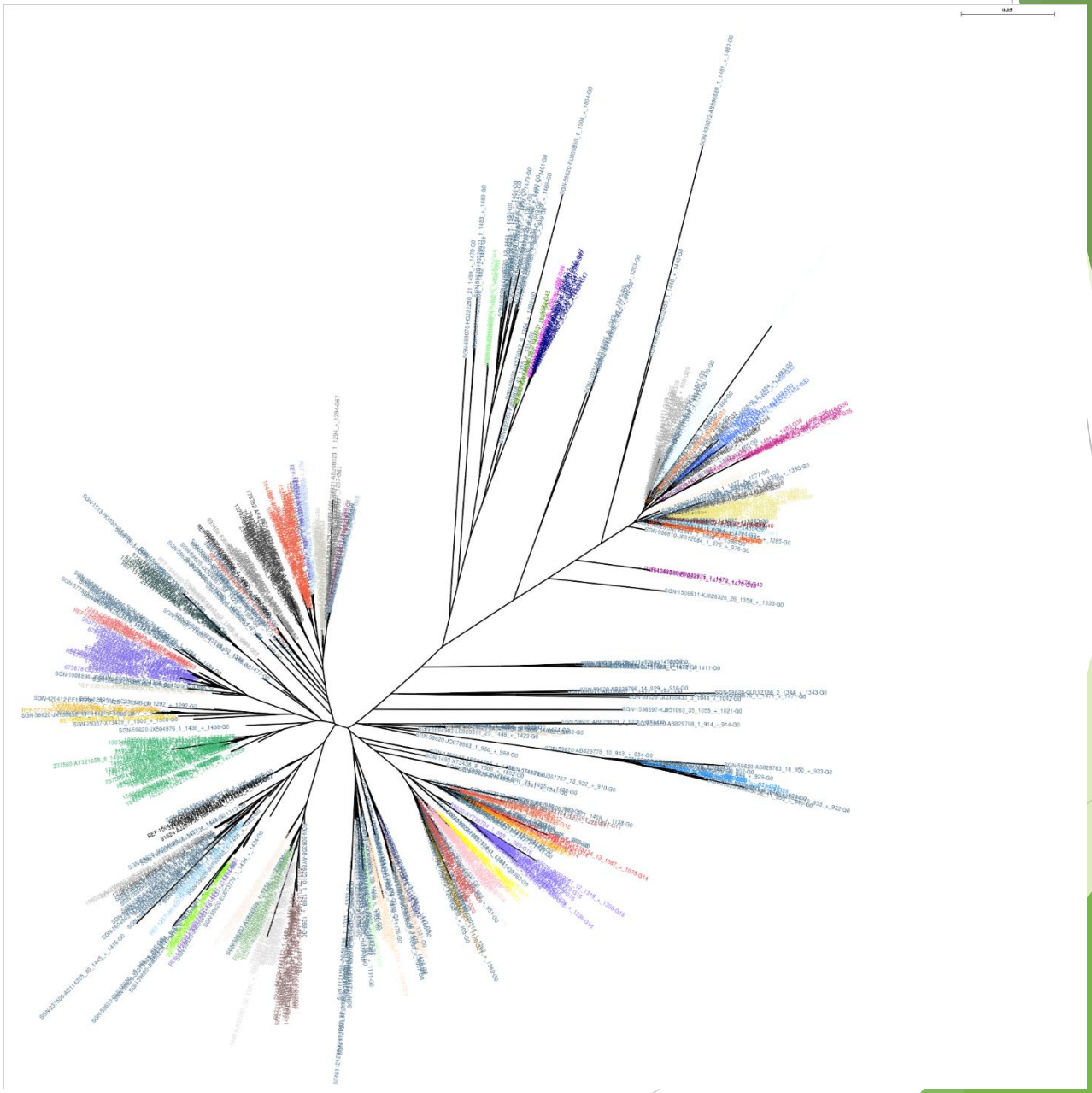
http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi



Clostridium Botulinum (TID:1491)



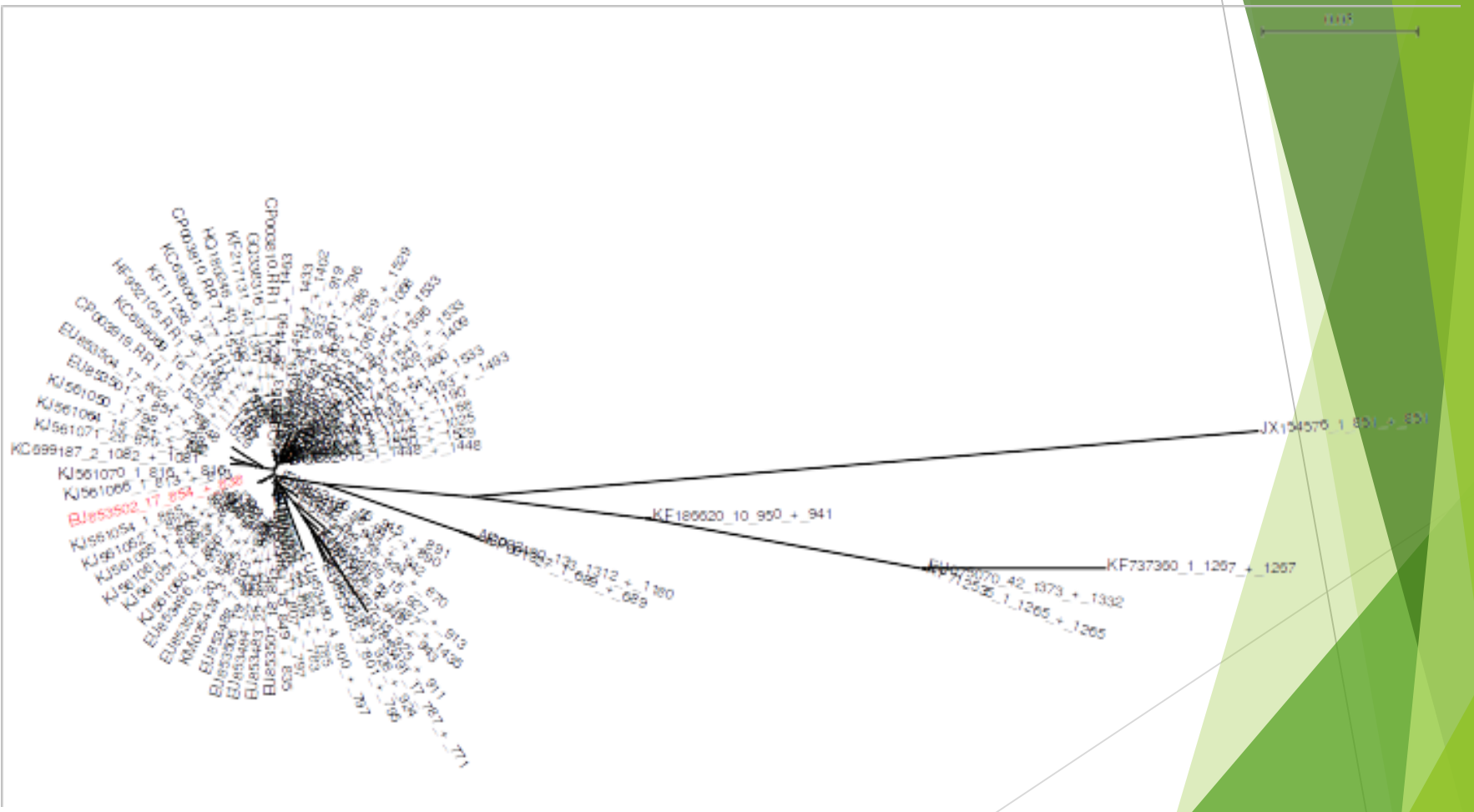
Genus Clostridium (TID:1485)



Singletons

- Sequences not clustering
  - at 80% sequence identity
  - in any nodes of the tree
- What are they ?
  - Dramatic sequencing errors
    - Chimeras
  - Real sequences
    - Misclassified sequences
    - Properly annotated but rare isoforms

Streptococcus Agalactiae (TID:1311)
JX154576 Average Patristic Distance = 0.3271

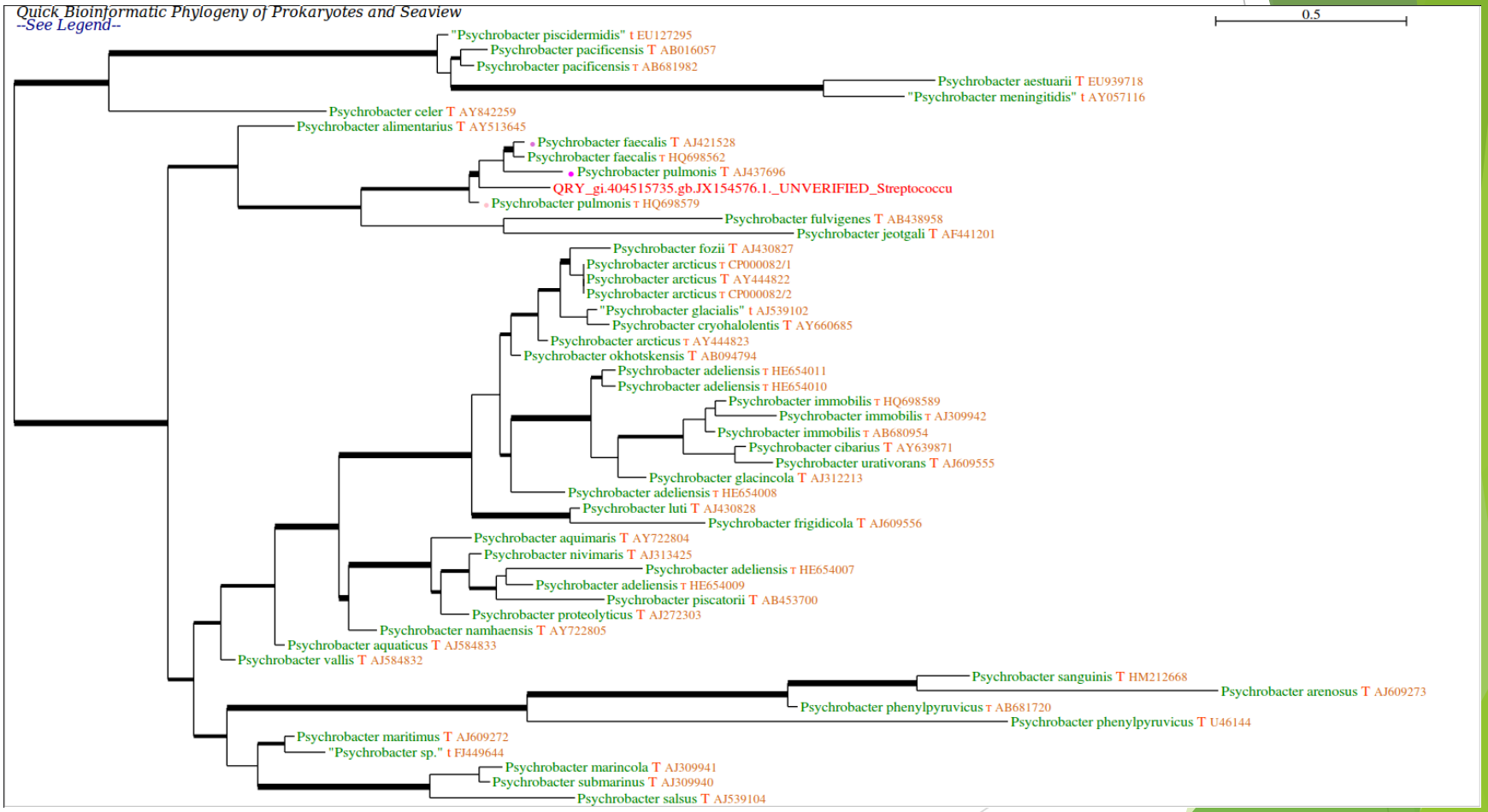


Streptococcus Agalactiae (TID:1311)

JX154576 Average Patristic Distance = 0.3271

BiBi classified it as a Psychrobacter

Common taxon is Bacteria



Conclusions

- Ribosomal genes DBs covers the largest species space
  - 16S rRNA => ~95% of prokaryotes
  - Complete genomes => ~50% of prokaryotes
- Noise in Data must be filtered
  - Millions of 16S sequences but only 13,000 species
  - Taxonomies evolves but annotations are not always updated accordingly
- Need of an automatic process for taking advantage of sequence/taxonomy daily updates
  - LeBiBi^{QBPP} with weekly update

Perspectives

- End-up the iterative process
  - Can we reclassify?
  - What will be the discriminative power?
- Computing on IFB cloud
  - A cluster is not appropriate
- Singletons?
  - How likely it is to have a true separated seq at high taxonomic level?
  - How efficient are chimeras detection software?
- New taxons from environmental samples?
  - Evolving from week to week
- Eukaryotes

Acknowledgements

- François Bartolo
- Clément Lionnet
- Jean-Pierre Flandrois
- Christine Oger
- Guy Perrière

LBBE
BIOMETRIE ET BIOLOGIE EVOLUTIVE

prabi
Pôle Rhône-Alpes de Bioinformatique

FRANCE GÉNOMIQUE

Assignment methods

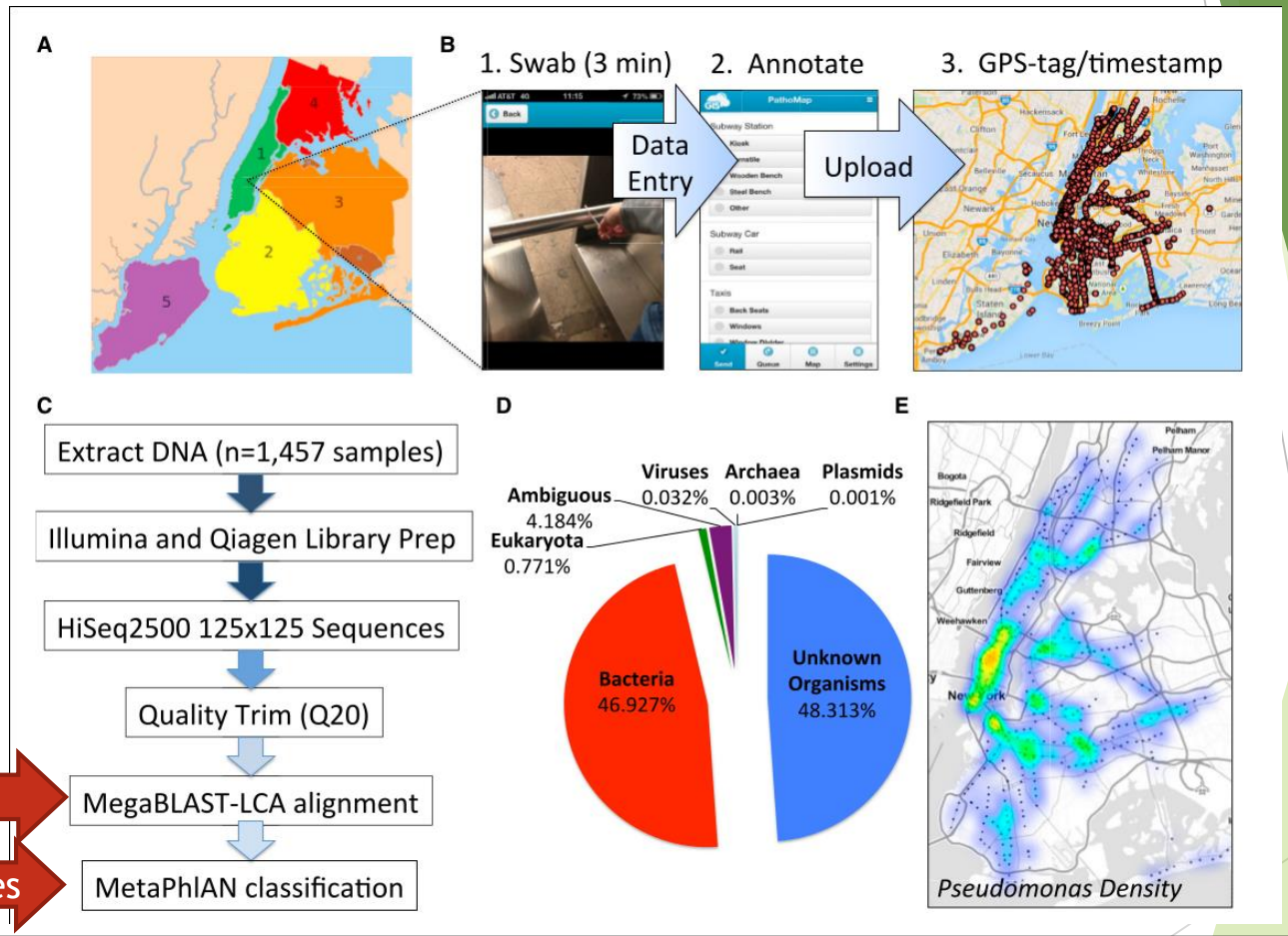
	MOTHUR	QIIME	LeBiBi QBPP
Clustering method	Hierarchical sequence clustering	Greedy heuristic sequence clustering	BLAST + Phylogenetic replacement
Reference	Schloss *et al.* Appl. Env. Microbiol. 2009	Caporaso *et al.* Nature Meth. 2010	Flandrois *et al.* under revision

Most clustering methods do not insure monophyletic OTUs

- SortMeRNA (Kopylova *et al.* Bioinformatics 2012)
- SWARM (Mahé *et al.* PeerJ 2014)

New York Metagenomics

Afshinnekoo *et al.* Cels 2015



NCBI NT

~17,000 genomes

2010). A total of 21,885 and 1,688 taxa were assigned with MegaBLAST and MetaPhlAn, respectively, with 15,152 and 637 specific to the species level (Data Tables 1 and 2), respec-

Discovery of Lokiarchaeota from Metagenomics

Spang *et al.* Nature 2015

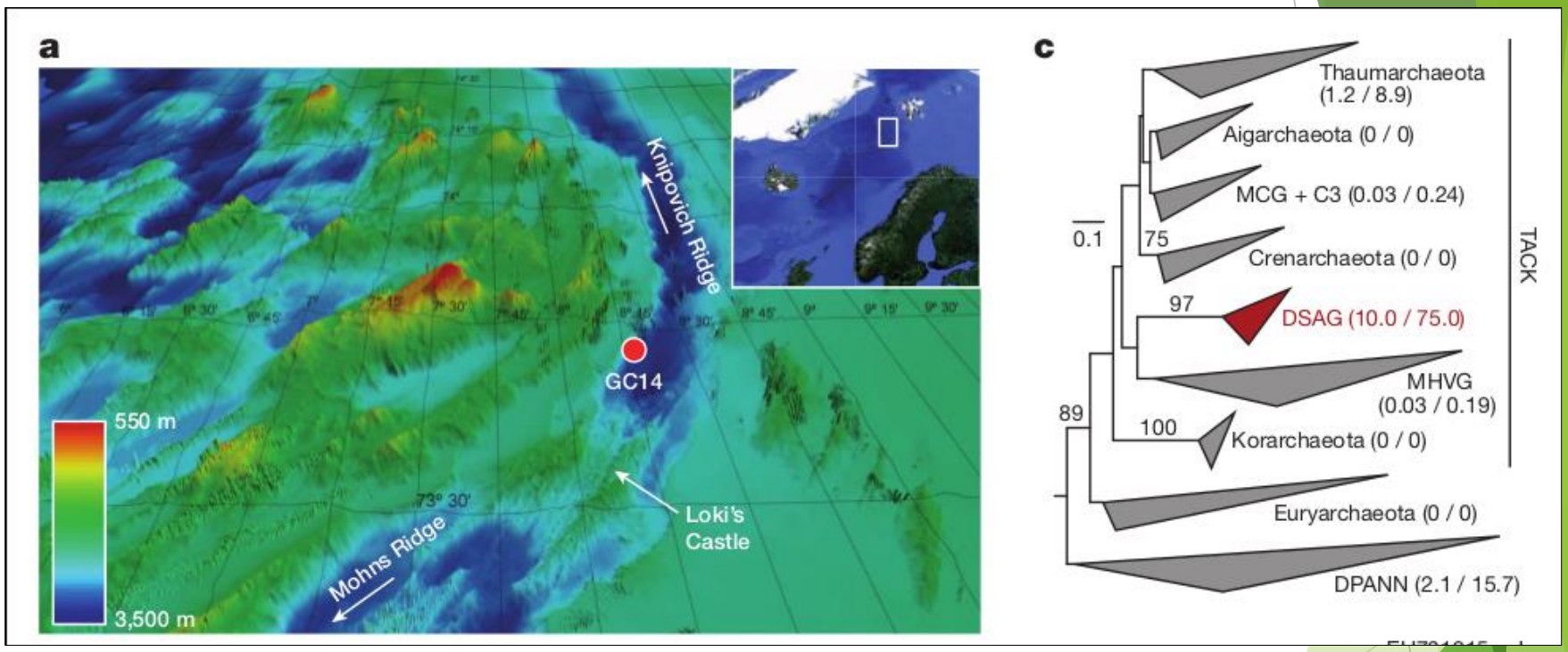


Figure 1: Phylogenetic analysis based on PCR amplification of 16S rDNA sequences

Lokiarchaeota from Metagenomics

Spang *et al.* Nature 2015

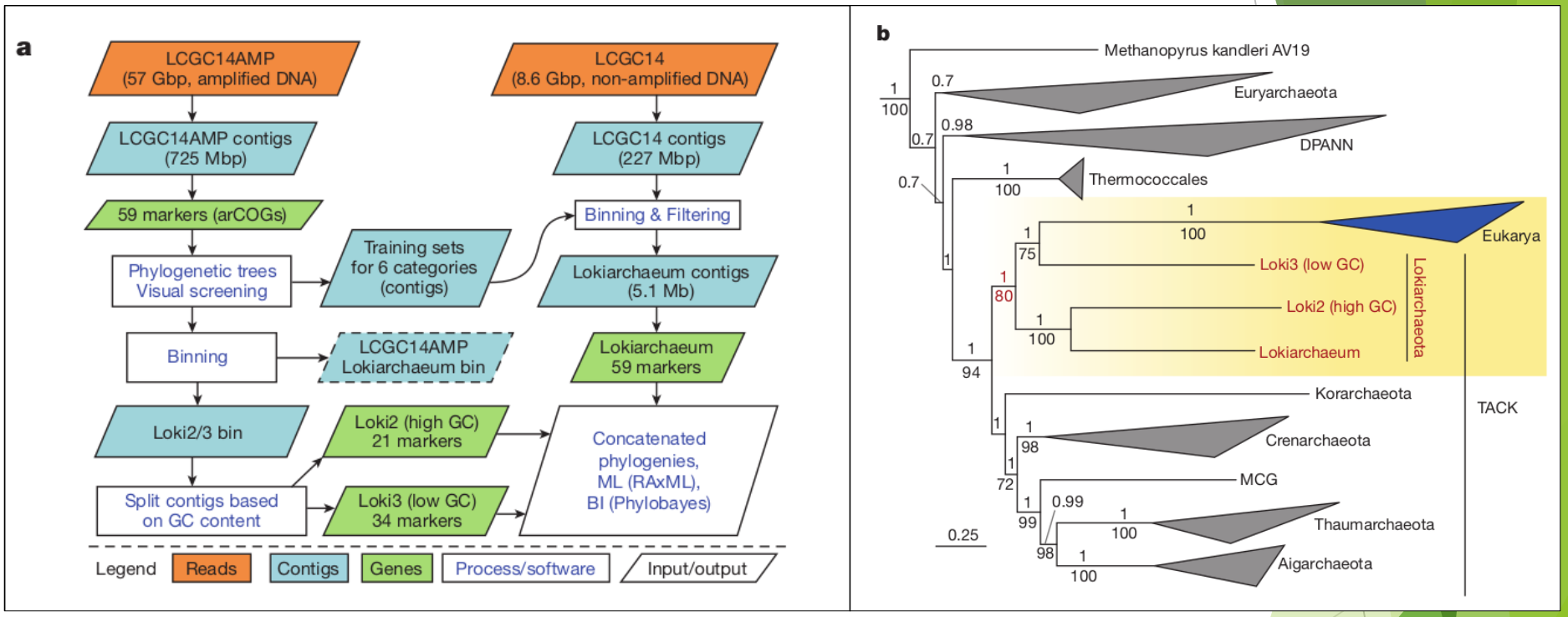
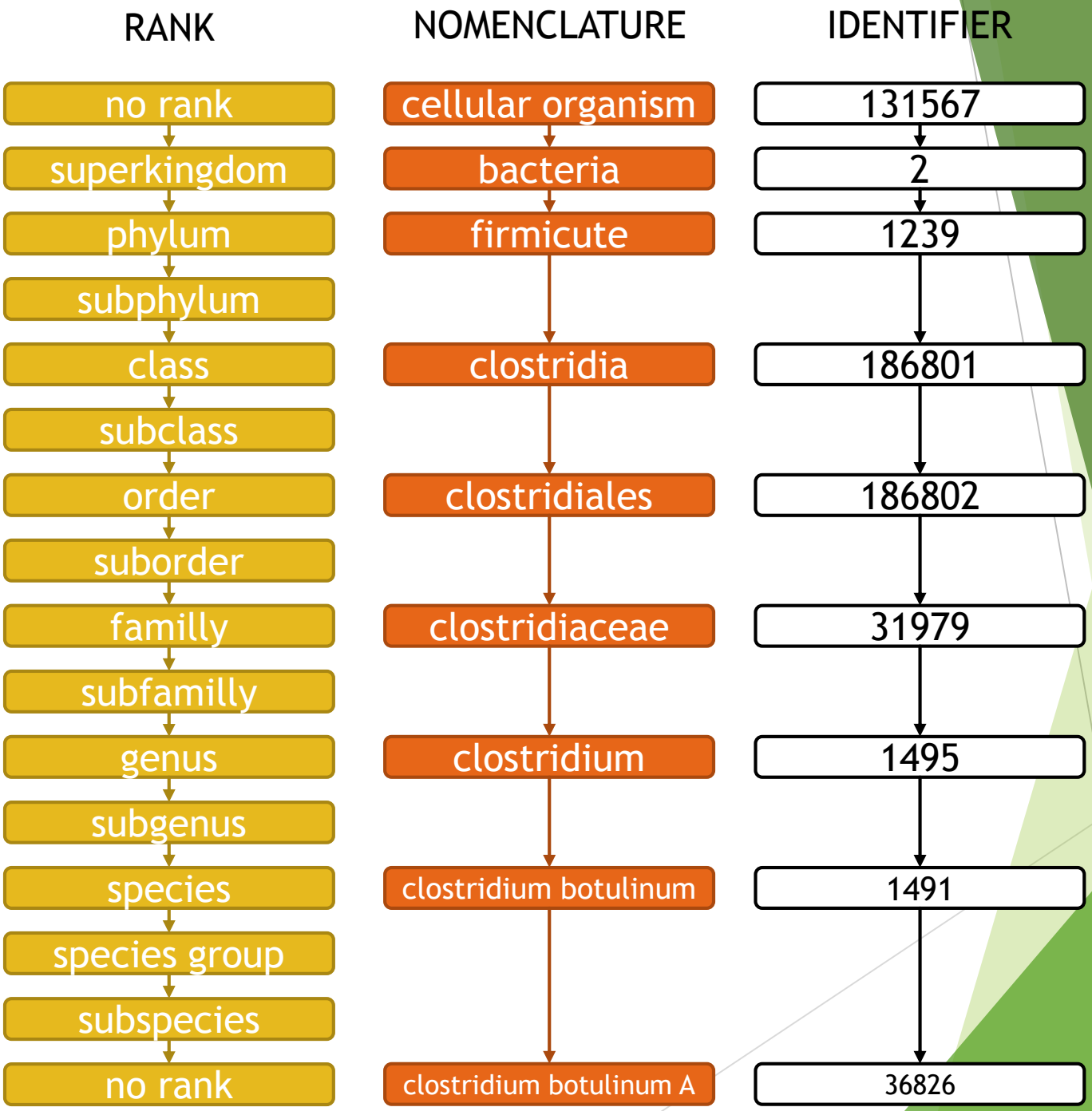


Figure 2 a and b: Multilocus analysis


RANK
NOMENCLATURE
IDENTIFIER
no rank
cellular organism
131567
superkingdom
bacteria
2
phylum
firmicute
1239
subphylum
class
clostridia
186801
subclass
order
clostridiales
186802
suborder
family
clostridiaceae
31979
subfamily
genus
clostridium
1495
subgenus
species
clostridium botulinum
1491
species group
subspecies
no rank
clostridium botulinum A
36826