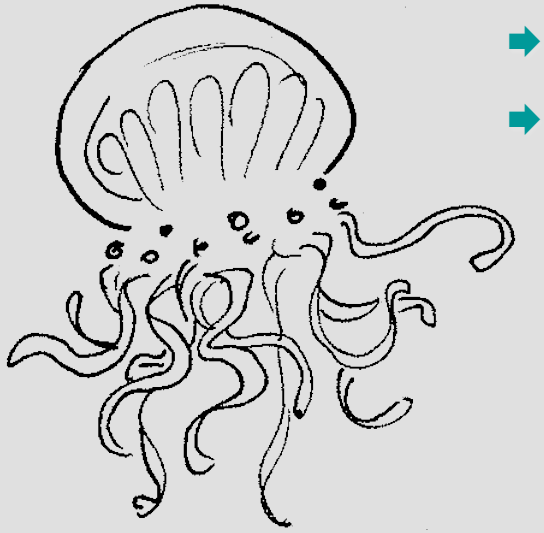
Application of taxonomic thresholds to manage diversity based on 16S RNA gene sequences

Fishing with my Father

- he does not know what is the real essence of the species category
- for him "different species" are organisms:
  - sufficiently different to not belong to the same unit.
  - sufficiently similar to belong to a unique group.

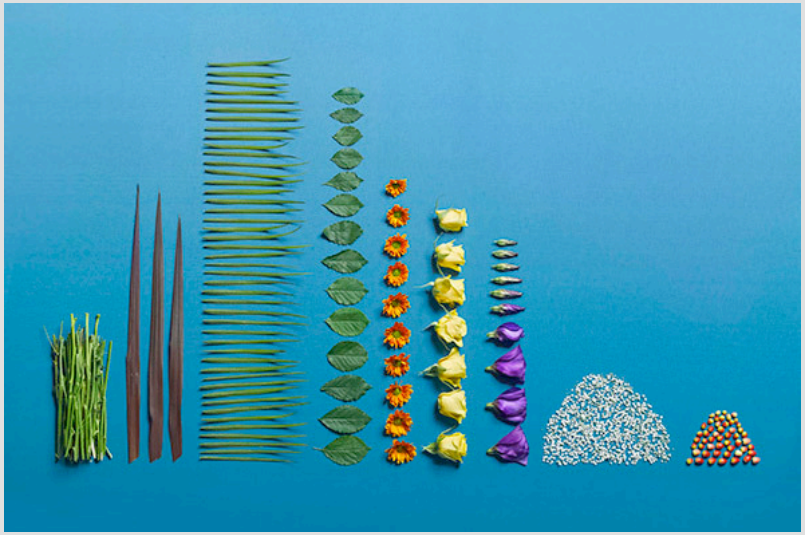


how is this possible?

All intellectual activities have as common denominator the introduction of some kind of order

Structuralism (Levi Strauss)

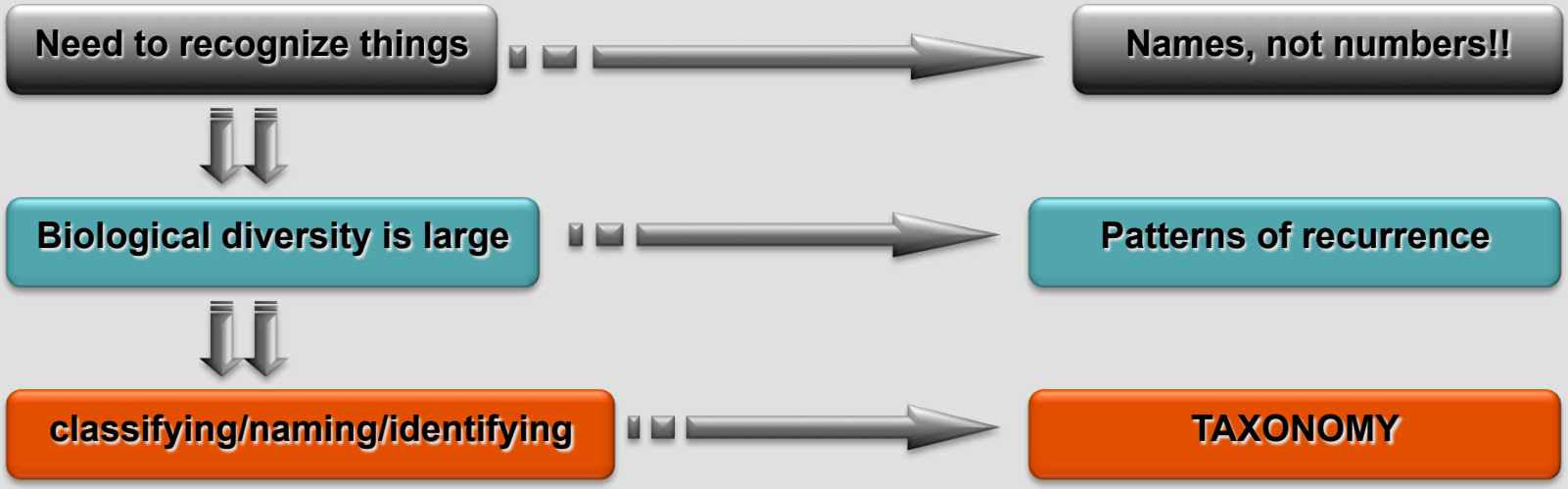
Nature does not appear ordered to our eyes



humans need to generate some kind of order to understand nature

There is a basic need of order in the human mind.

Any kind of thinking, ‘primitive’ or not, develops its own taxonomies, i.e. classification of biological things.



We <u>assume</u> that:

- there is an order in the Nature
- discontinuities exist ⇔ thus, we can recognize units
- recognition is limited by the observational methods

SYSTEMATICS

- any purpose
- choice ⇔ needs
- no necessarily scientific background

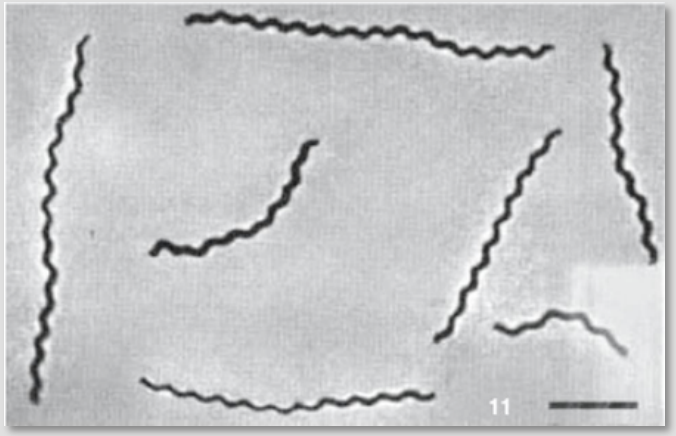
TAXONOMY

- general purpose classification
- comprehensive DB
- system:

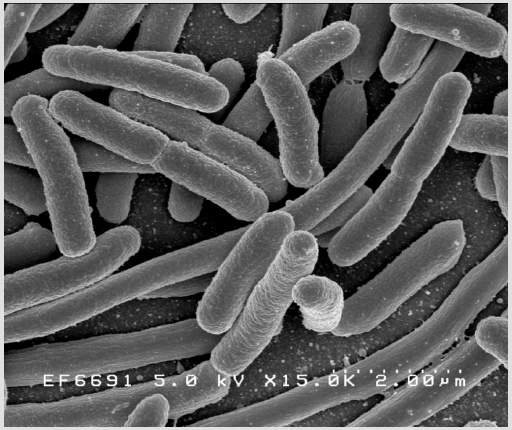
- ➡ operational
- ➡ predictive
- ➡ universal

PRAGMATISM: end users do not want to deal with the theoretical background, but to know if they have something new in their hands

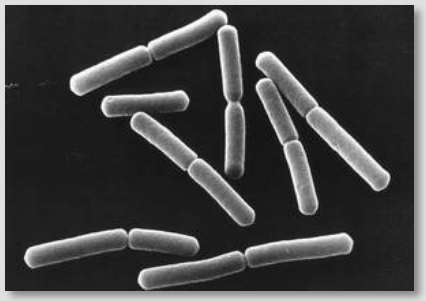
?


11

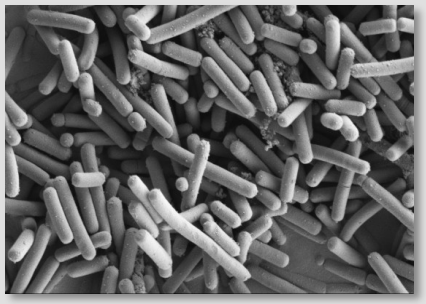
?


EF6691 5.0 kV X15.0K 2.00µm

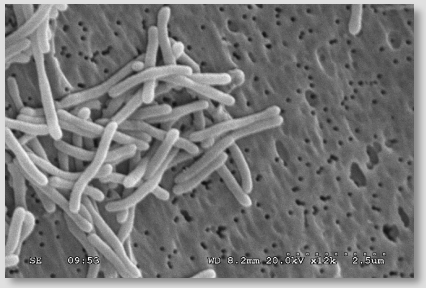
SPECIES patterns of recurrence



Bacillus subtilis



Lactobacillus delbrueckii



Salinibacter ruber

recurrence patterns:

- **depends on the morphological complexity**
- **for different kind of organisms recurrence may respond to different evolutionary constrains**
- **prokaryotes do not exhibit a morphology that can be recognized as a pattern**
- **we identify each pattern as a unit ⇔ species**

Pelagia noctiluca

Xyrichtys novacula

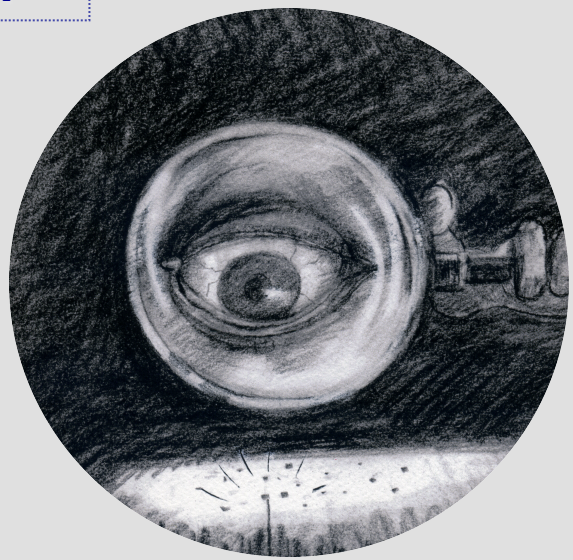
CONCEPT versus DEFINITION

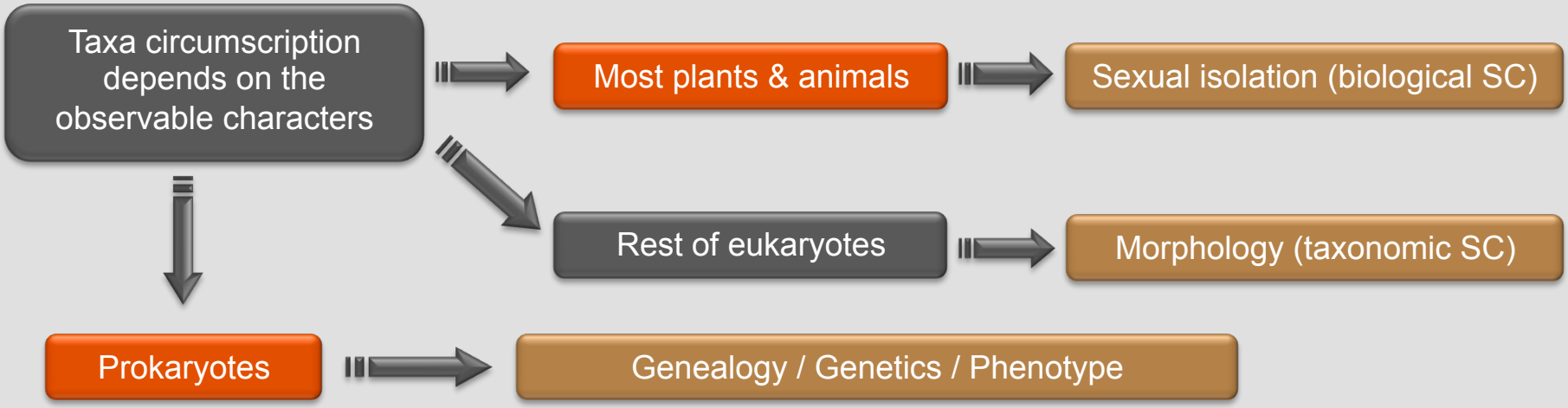
The CONCEPT is the IDEA, what embraces a unit

(generally immutable)

The DEFINITION is the WAY to embrace a unit

(changes with technical developments)
(depends on the observable characters)





Rosselló-Mora & Amann 2001, FEMS Rev. 25:39-67

CONCEPT versus DEFINITION

The CONCEPT is the IDEA, what embraces a unit (generally immutable)

- monophyletic group of isolates
- genomically coherent
- sharing high similarity in many independent phenotypic features

The DEFINITION is the WAY to embrace a unit (changes with technical developments)

- monophyly ⇔ gene sequence analysis (i.e. 16S rRNA)
- genomic coherence ⇔ DDH
- phenotype (biochemical tests, chemotaxonomy…)

estimates on the abundance of prokaryotes in the biosphere

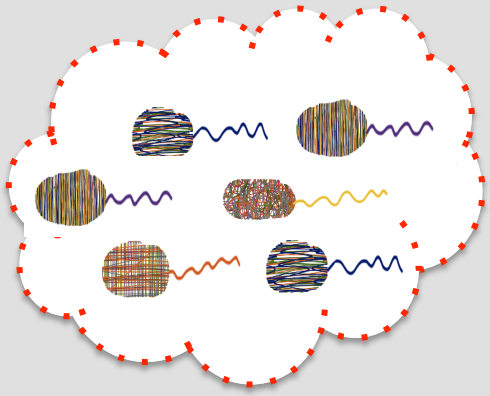
(6x10^{30} prokaryotes)	6,000,000,000,000,000,000,000,000,000,000
(1x10^{12} expected SP optimistic)	1,000,000,000,000
(3x10^{4} expected SP pessimistic)	30,000
(1x10^{4} species)	12,000

Whitman et al., 1998. PNAS 95, 6578-6583 / Mora et al., 2011. *PLOS Biol*. 9, e1001127 / Dykhuizen, 1998. *A. Van Leeuw*. 73, 25-33

- Species is the basic unit to measure biodiversity
- Different disciplines (taxonomy, ecology, evolutionary biology) use different definitions

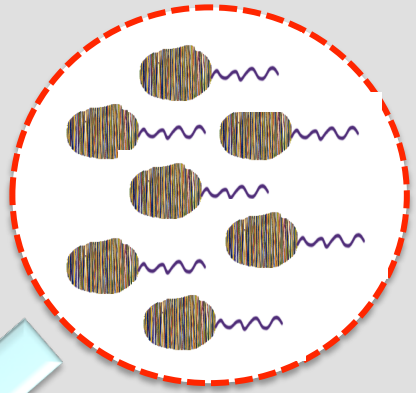
SPECIES

Environmental sample



isolation

Pure culture



definition

MONOPHYLY:
- rRNA genes
- HKG

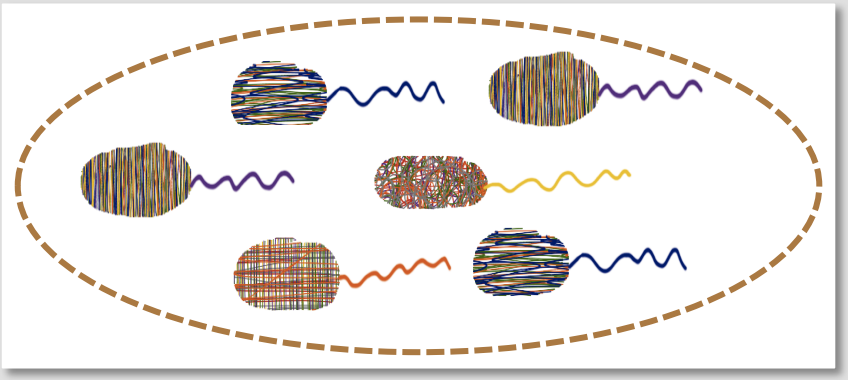
PHENOTYPIC COHERENCE:
- Biochemical tests
- Chemotaxonomy
- …

GENOMIC COHERENCE:
- DDH
- ANI
- MLSA

The effort of a taxonomist:

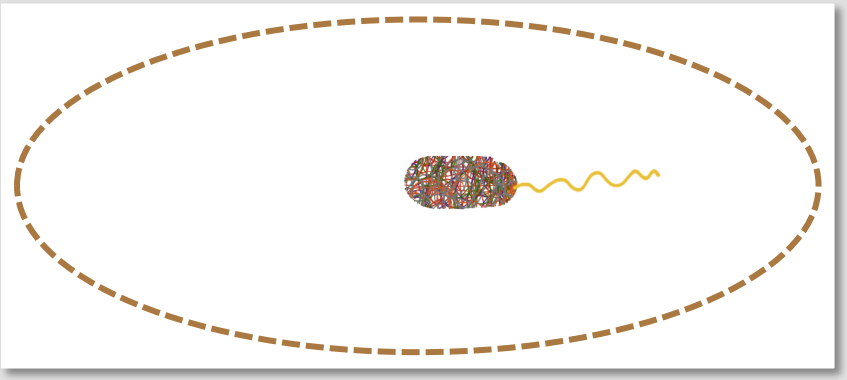
- ➔ Isolate (several strains)
- ➔ Clarify genealogy (phylogenetic inference)
- ➔ Exhaustive analysis of genetic traits (Genomics)
- ➔ Exhaustive analysis of phenotype (metabolism + chemical properties)

Several strains



Intraspecific diversity
Ensure valid common characters
Solid diagnostic criteria

One strain



Single strain valid characters
No clue about the validity of the diagnostic criteria

Increase of the Single Strain Species Descriptions (SSSD)

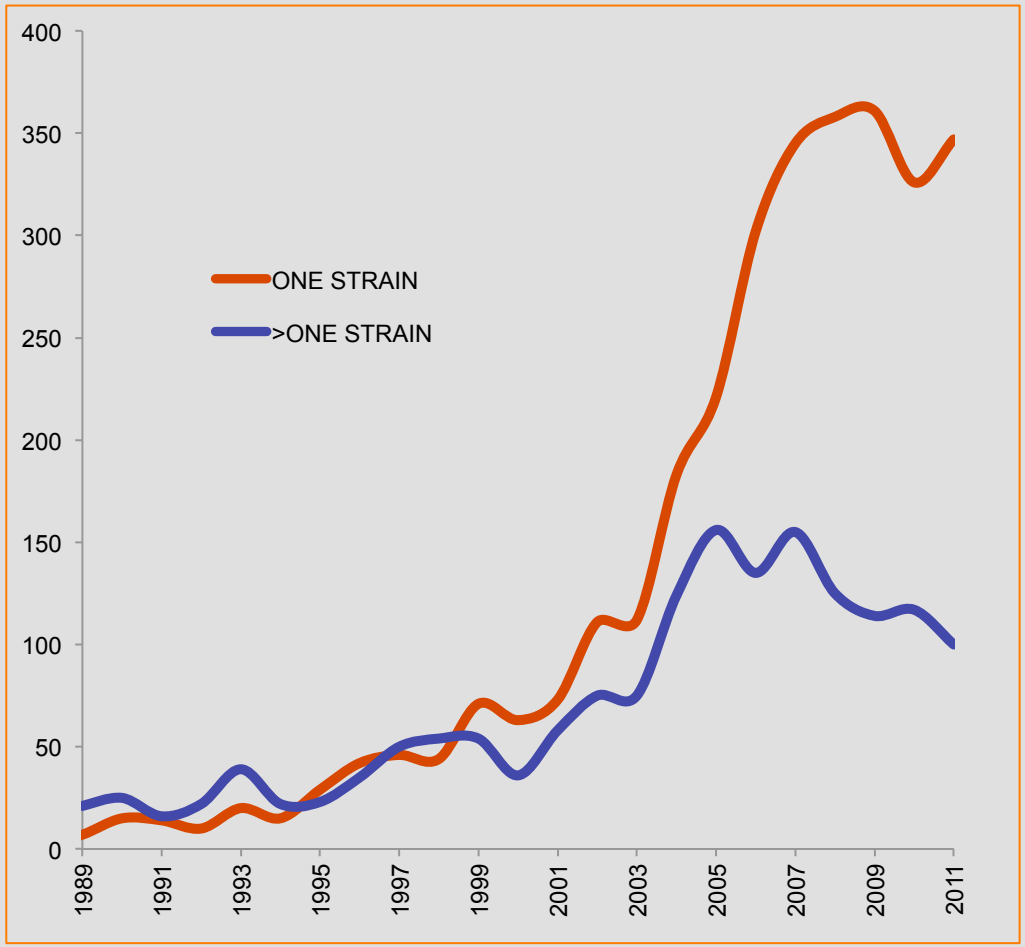
The effort of taxonomists:

>1 strain (about 20%)
1 strain (about 80%)

allowing new species descriptions by means of only 1 isolate:

⇒ **facilitate the increase in descriptions**

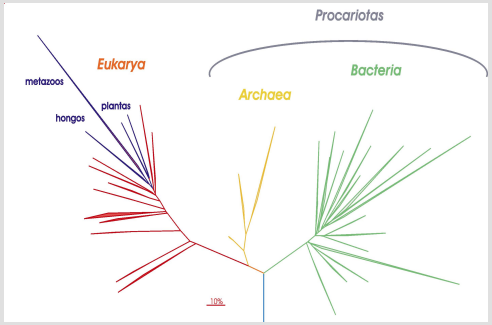
⇒ **preclude the understanding of the biological diversity**



Tamames & Rosselló-Móra 2012 Trends in Microbiol 20:514-516

How do we define / circumscribe species for prokaryotes

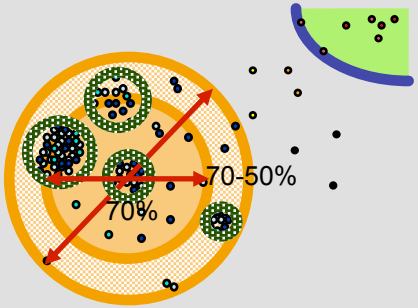
phylogenetic coherence



RNAr 16S
Functional genes (MLSA)
Genomic analyses

- ⇨ 16S rRNA gene sequence (gold std)
- ⇨ 97% - 98.7% identity threshold
- ⇨ all organisms must be monophyletic

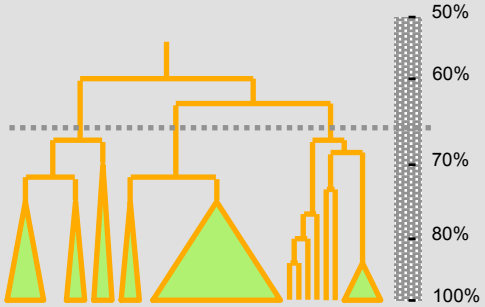
genomic coherence



Reasociación DNA-DNA
G+C, AFLP, MLSA
Genomic comparisons
(ANI; AAI)

- ⇨ DDH (gold standard)
- ⇨ 70% similarity threshold
- ⇨ 96% ANI

phenotypic coherence

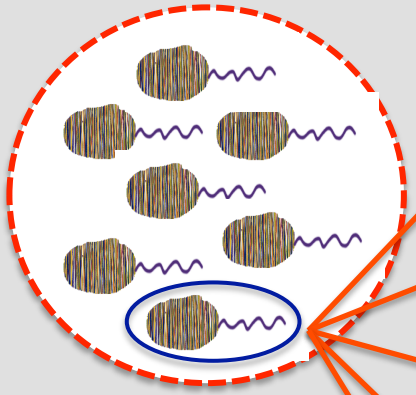


metabolism
chemotaxonomy
spectrometry
(Maldi-Tof; ICR-FT/MS)

- ⇨ identificative phenotypic property
- ⇨ chemotaxonomic markers
- ⇨ metabolic homogeneity

Tindall et al., 2010 IJSEM 60:249-266

TYPE STRAIN



DEPOSIT (compulsory):
LIVING CULTURE IN 2 INTERNATIONAL CULTURE COLLECTIONS (CECT, DSMZ, ATCC, …)

DEPOSIT (compulsory):
ALMOST COMPLETE 16S rRNA GENE IN PUBLIC REPOSITORIES (EMBL, GenBank, …)

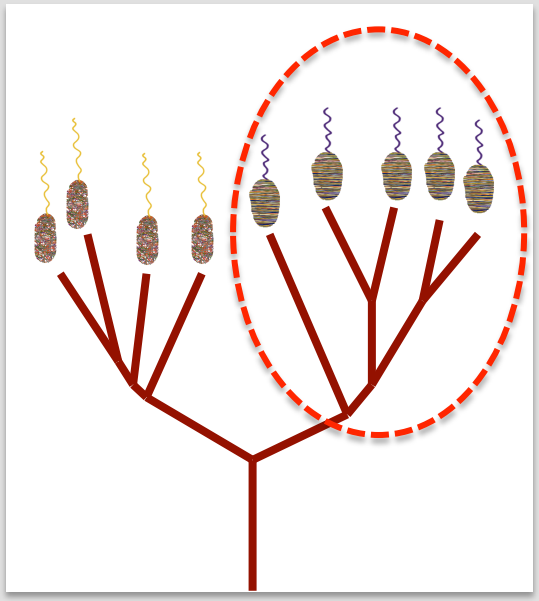
DEPOSIT (strongly recommended):
ALMOST COMPLETE GENOME IN PUBLIC REPOSITORIES (NCBI, …)

DEPOSIT (future recommendation):
DNA AS TYPE MATERIAL IN 2 INTERNATIONAL CULTURE COLLECTIONS (CECT, DSMZ, ATCC, …)

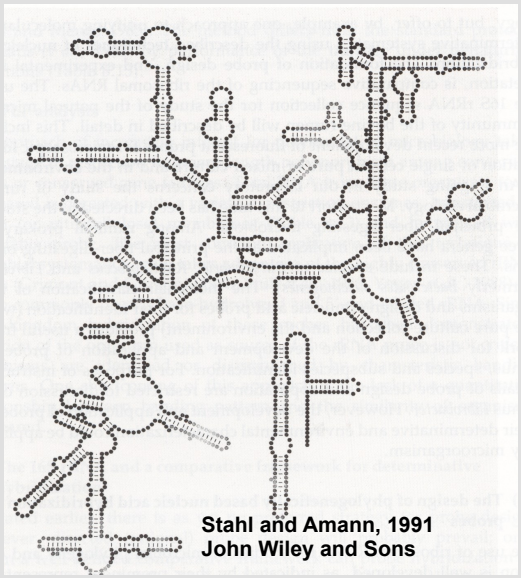
DEPOSIT (future recommendation):
PHENOTYPIC DATA IN PUBLIC REPOSITORIES (MALDI-TOF; METABOLOMICS, …)

➡ **Ensure that a name is what it should be**

SPECIES: monophyletic group of organisms



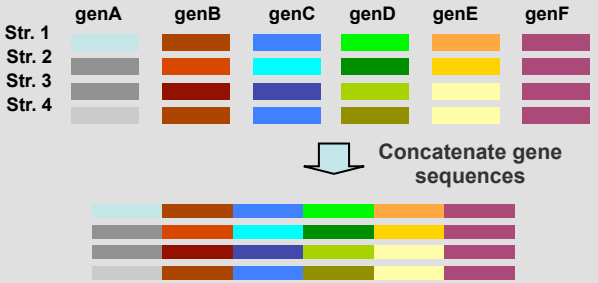
rRNA genes



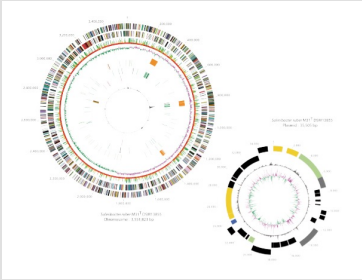
PHYLOGENY

HK genes

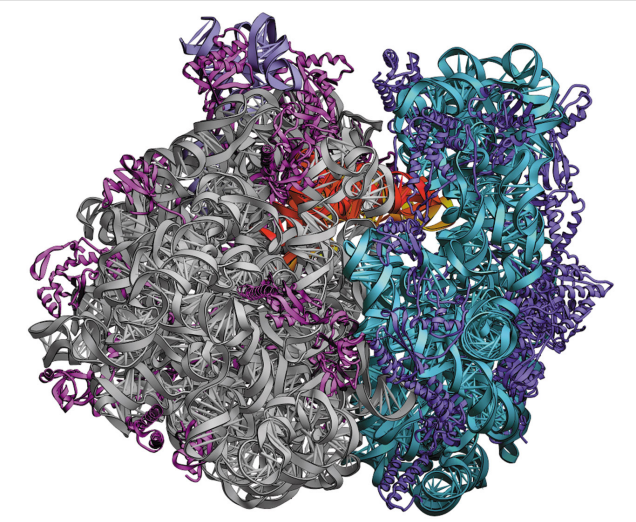
(MLSA)



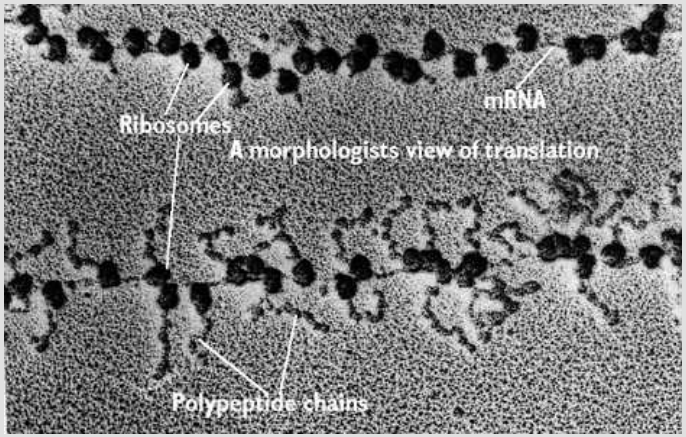
Genomes (phylogenomics)



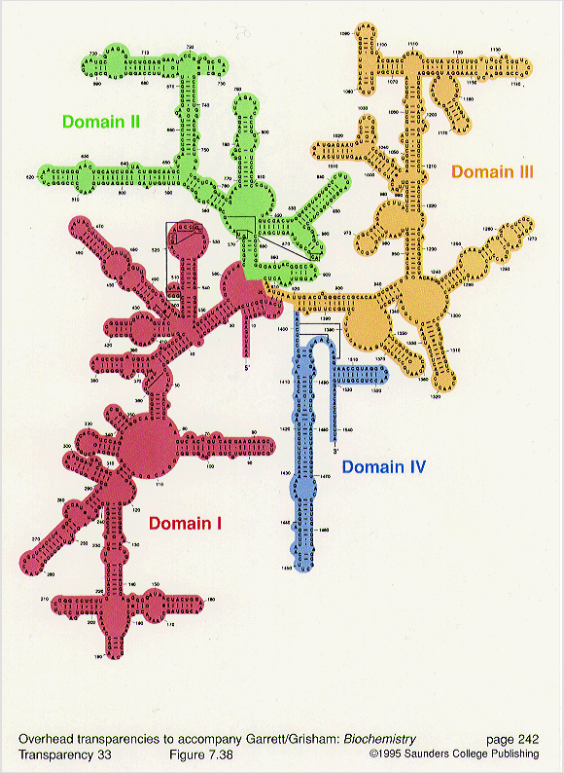
16s rRNA gene marker



http://rna.ucsc.edu/rnacenter/ribosome_images.html



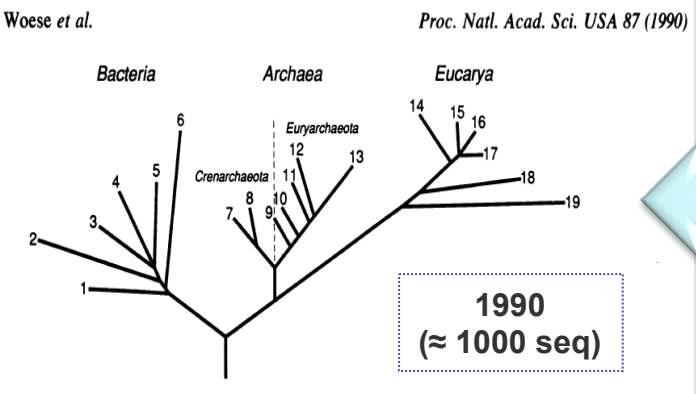
Natural gene amplification



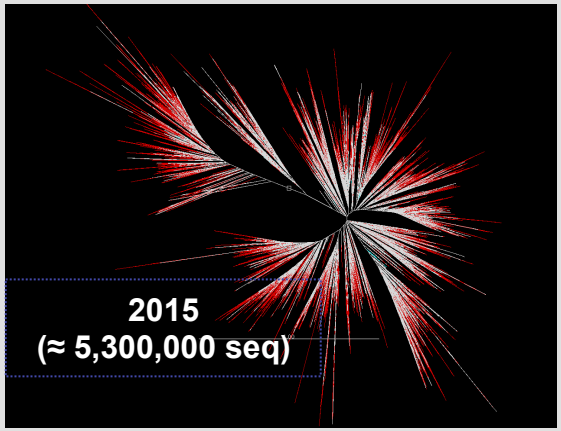
Conserved 2º structure

Ludwig and Schleifer, 1994 FEMS Rev 15:155-173

16S rRNA: 25 years of history

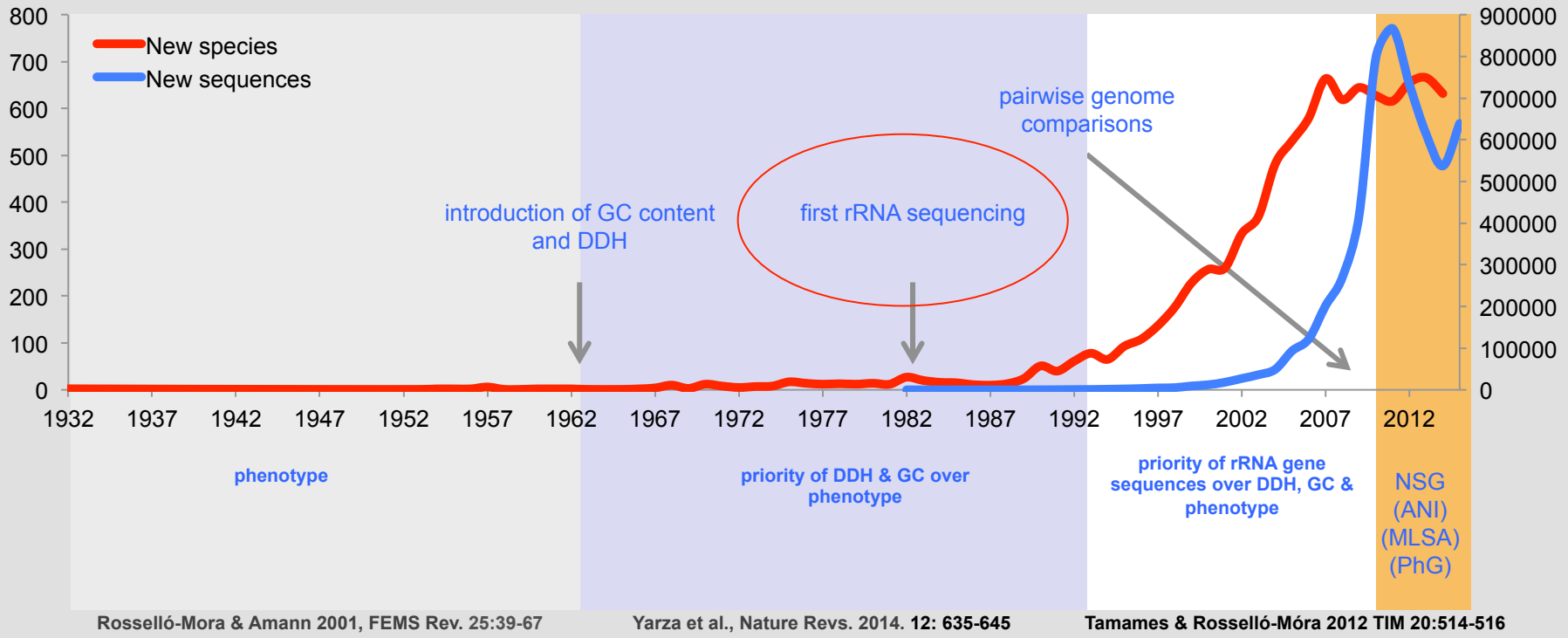


25 years



➡ **In almost ¼ century the database increased exponentially**

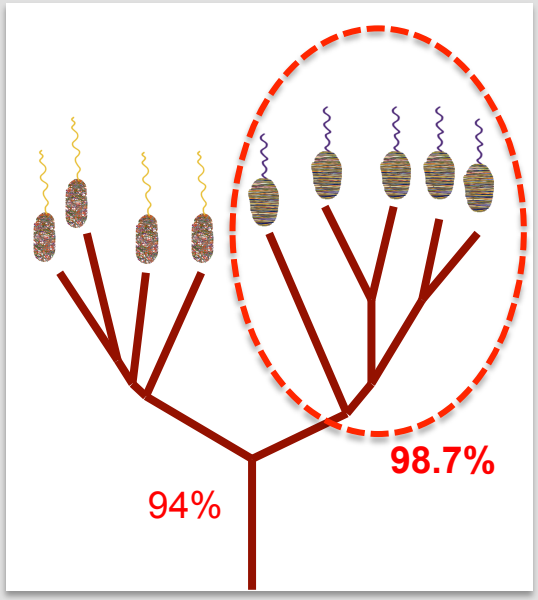
16S rRNA: exponential increase



Rosselló-Mora & Amann 2001, FEMS Rev. 25:39-67 Yarza et al., Nature Revs. 2014. **12: 635-645** **Tamames & Rosselló-Móra 2012 TIM 20:514-516**

- **99% environmental (700,000 / year)**
- **<1% cultured (700 species /year)**

16S rRNA: calculating taxonomic thresholds



Evaluation of 16S rRNA intraspecific identity

Stackebrandt & Goebel (1994); IJSB 44: 846
(97% species threshold)

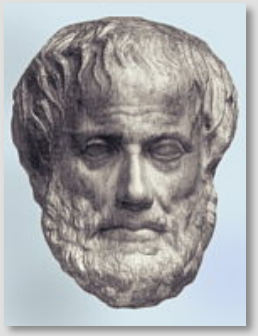
Stackebrandt & Goebel (2006); MT 33: 152
(98.7% species threshold)

Nothing evaluated for high taxa

Very low resolution for species

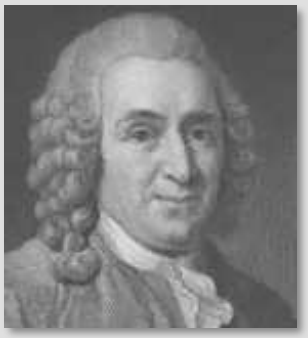
For genus unwritten rule of 94%

Classification has 2400 years of exercise ⇔ hierarchical structure



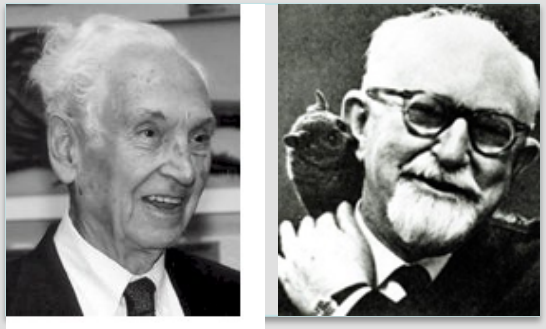
Aristotle (-2400 years)

genus
species



Linné (-266 years)

kingdom
class
order
genus
species



Mayr & Simpson (-45 years)

kingdom
phylum
subphylum
superclass
class
infraclass
cohort
superorder
order
infraorder
superfamily
family
subfamily
tribe
subtribe
genus
subgenus
species
subspecies

- Species are real entities
- higher taxa are considered abstract
- the system IS artificial

Ereshefsky 1994. Phyl. Sci. 61:186-205; Rosselló-Móra 2005. JBac. 187:6255-6257

16S rRNA: the Living Tree Project (LTP)

SEPTEMBER 2014

CONSORTIUM

SAM (ELSEVIER) - Ramon Rosselló-Móra, Rudolf Amann, Karl-Heinz Schleifer: leadership. **IMEDEA** - Raul Munoz: curator. **SILVA** - Frank Oliver Glöckner: sequence databases, computational resources and web hosting. **LPSN** - Jean Euzéby: support on taxonomic nomenclature. **ARB** - Wolfgang Ludwig: support on phylogenetic analysis and taxonomic classification. **RIBOCON** - Jörg peplies and Pablo Yarza: training of curators and support on database management.

INSDC International Nucleotide Sequence Database Collaboration — **± 360,000,000 entries**

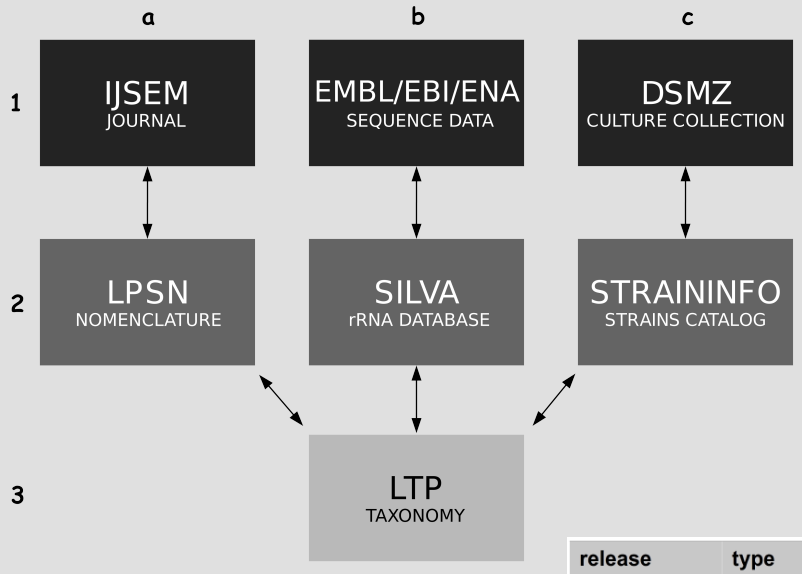
16S rRNA genes (PARC 123) — silva comprehensive ribosomal RNA databases — **± 5,300,884 entries**

Type strain sequences — LTP — **± 12,000 entries**

- Easily select the ADEQUATE sequences for TAXONOMIC purposes
- The BEST sequence for each TYPE strain (length, quality, …)

Yarza et al., 2010, System Appl Microbiol. 33, 291-299

Yarza and Munoz. Methods in Microbiol. in press



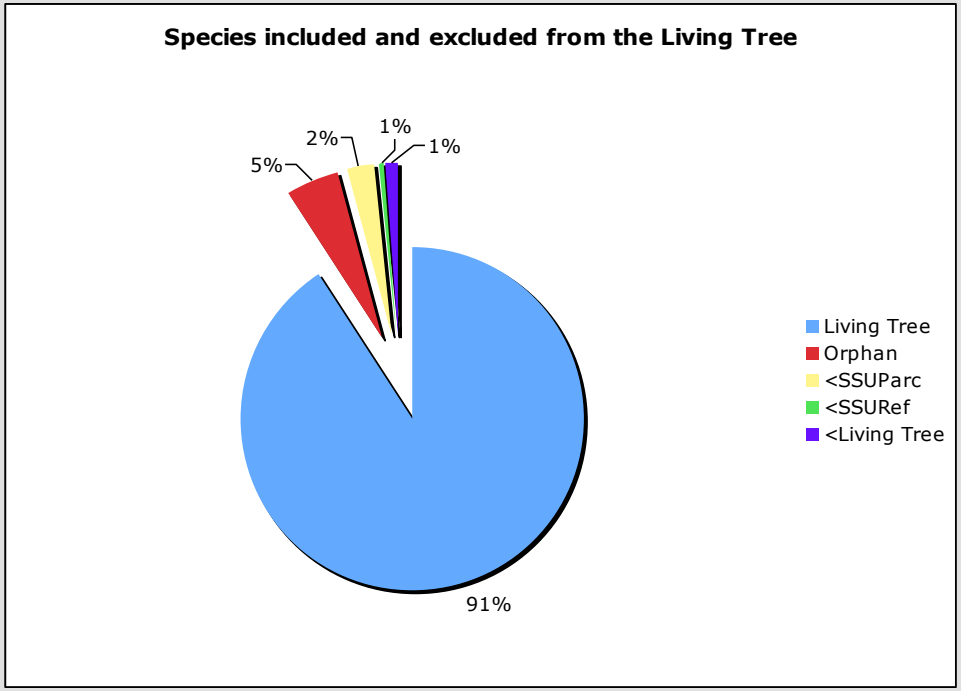
Rationale & Releases

⇒ **merging information of DB**
⇒ **12% => 45% => 75% wrong info**
⇒ **638 species not included (orphan)**
⇒ **9 releases up to now since 2008**

release	type	date	total entries	new entries	deleted entries	net increase	% incorrect
LTPs93	SSU	Aug.2008	6728	6728	0	6728	22
LTPs95	SSU	Oct.2008	7006	299	21	278	45
LTPs100	SSU	Sep.2009	7710	750	46	704	50
LTPs102	SSU	Sep.2010	8029	363	44	319	58
LTPs102	LSU	Sep.2010	792	792	0	792	6
LTPs104	SSU	Mar.2011	8545	545	29	516	74
LTPs106	SSU	Aug.2011	8815	279	9	270	77
LTPs108	SSU	Jul.2012	9279	490	26	464	60
LTPs111	SSU	Feb.2013	9701	415	7	422	73
LTPs115	SSU	Mar.2014	10271	597	27	570	75

Yarza et al., 2010, System Appl Microbiol. 33, 291-299 Yarza and Munoz. Methods in Microbiol. in press

Output of the 1st release in 2008



IT SEEMS TRIVIAL BUT IT IS NOT!

12% of the sequences carried mistakes in the fields "strain" or/and "full name"!

For the 2nd release 45% of the validly published species had a wrong entry in the EMBL in their "full name"!

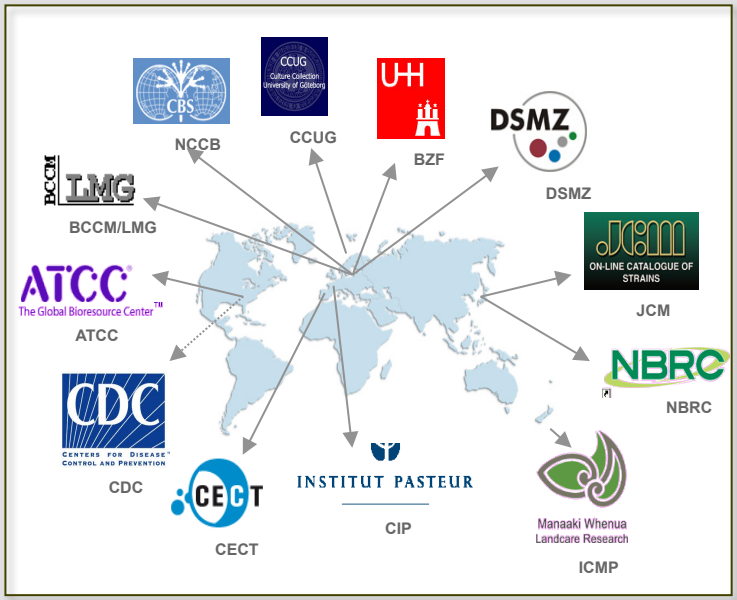
For the 3rd release 75% had a wrong entry!

- 6728 (91%) in the LTP
- 639 not included (9%)
  - 363 (5%) orphan (no sequence entry)
  - 177 (2%) very bad quality (no SSUParc sift)
  - 45 (1%) bad quality (no SSURef sift)
  - 54 (1%) bad quality for the project (ambiguities, homopolymers, chimera, …)

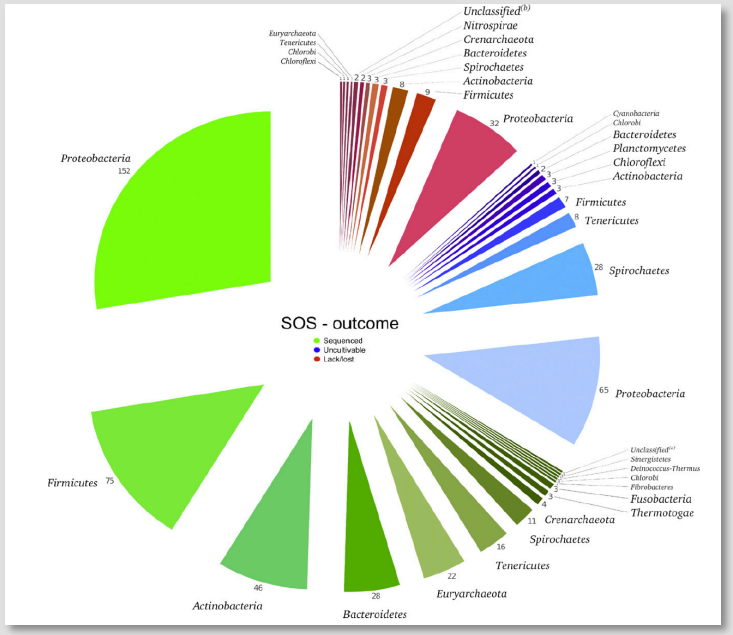
Yarza et al., 2008, System Appl Microbiol. 31:241-250

Sequencing Orphan Species

- 638 (9%) not included in the LTP
- 362 (5%) orphan (no sequence entry)
- 276 (3,5%) (very) bad quality



Yarza et al., 2013, System Appl Microbiol. 36, 69-73

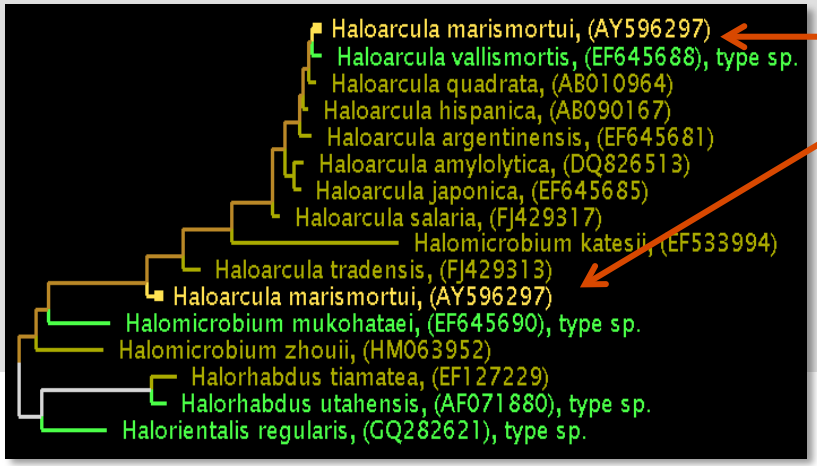


- 390 SOS
- 115 uncultivable
- 59 never deposited or lost

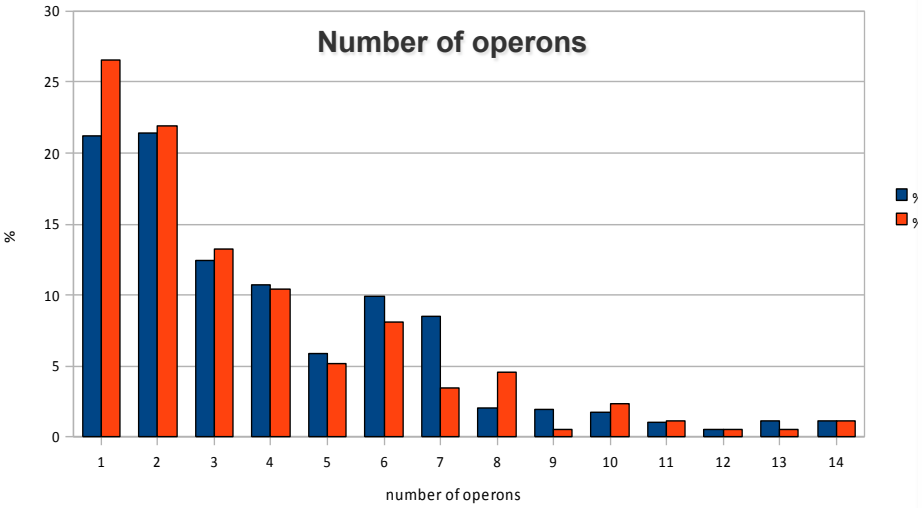
Sequencing orphan species initiative (SOS): Filling the gaps in the 16S rRNA gene sequence database for all species with validly published names

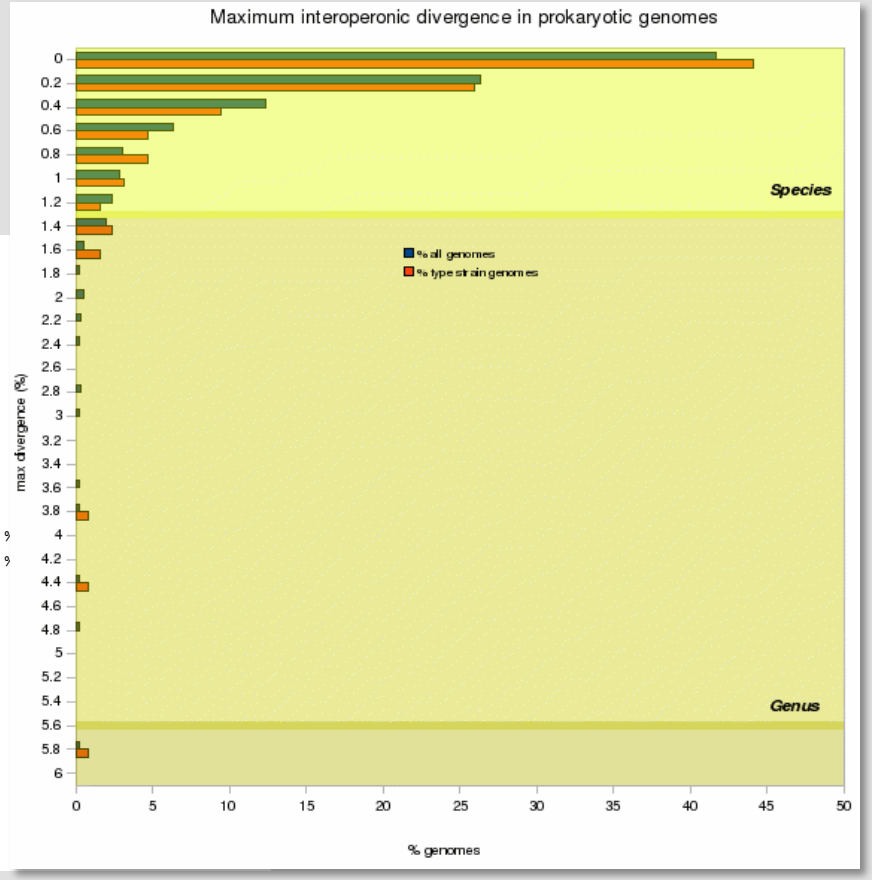
Pablo Yarza[a,r,*], Cathrin Spröer[b], Jolantha Swiderski[b], Nicole Mrotzek[b], Stefan Spring[b], Brian J. Tindall[b], Sabine Gronow[b], Rüdiger Pukall[b], Hans-Peter Klenk[b], Elke Lang[b], Susanne Verbarg[b], Audra Crouch[c], Timothy Lilburn[c], Brian Beck[c], Christel Unosson[d], Sofia Cardew[d], Edward R.B. Moore[d], Margarita Gomila[e], Yasuyoshi Nakagawa[f], Danielle Janssens[g], Paul De Vos[g], Jindrich Peiren[g], Timo Suttels[g], Dominique Clermont[h], Chantal Bizet[h], Mitsuo Sakamoto[i], Toshiya Iida[i], Takuji Kudo[i], Yoshimasa Kosako[i], Yumi Oshida[i], Moriya Ohkuma[i], David R. Arahal[j], Eva Spieck[k], Andreas Pommerening Roeser[k], Marian Figge[l], Duckchul Park[m], Peter Buchanan[m], Ana Cifuentes[a], Raul Munoz[a], Jean P. Euzéby[n], Karl-Heinz Schleifer[o], Wolfgang Ludwig[o], Rudolf Amann[p], Frank Oliver Glöckner[p,q], Ramon Rosselló-Móra[a]

16S rRNA: paralogy in the database



One strain ◄► one sequence?



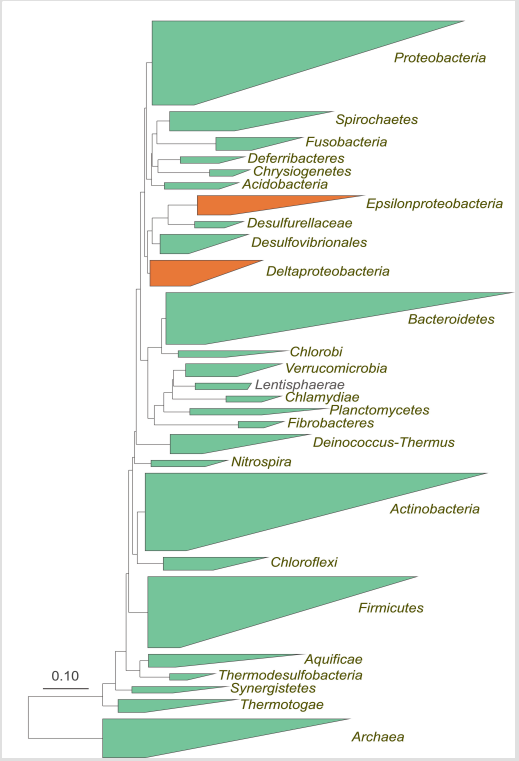


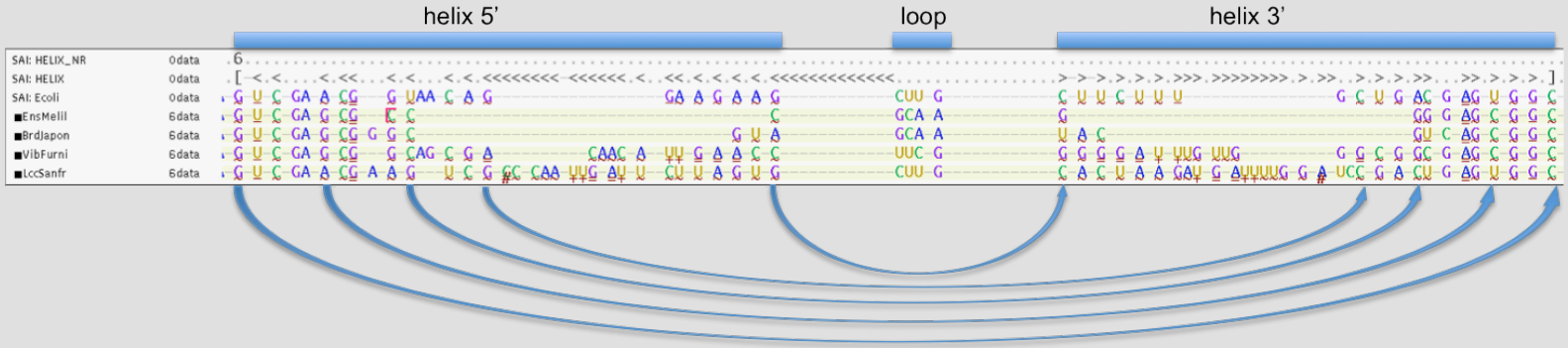
➡ 70% ◄►1-4 operons
➡ 99% <1.6% divergence (240 nucleotides)

Yarza et al., 2010, Syst Appl Microbiol 33:291-299

benefits of the releases:

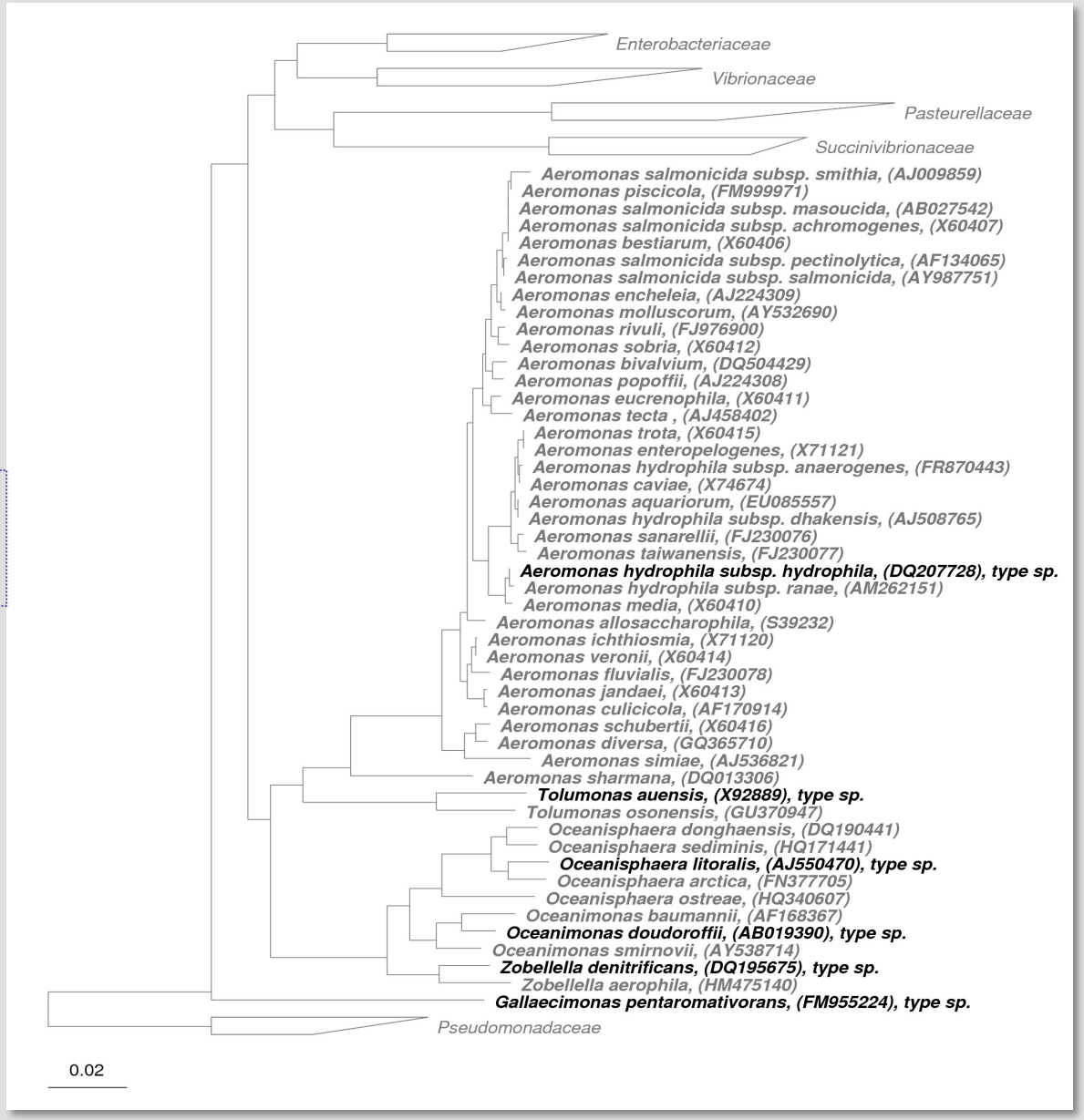
- data is released in ARB and CSV formats
- sequences are pre-aligned taking into account secondary structures
- conservational filters (domain; 10% ⇔ 50%) for reconstruction purposes
- special dataset of 750 sequences of high quality for neighbor joining reconstructions
- a single tree with all type strains (RaxML)
  - facilitates insertion by ARB parsimony tool
  - facilitates selection of close relatives
  - the tree topology IS NOT EXACT!
    - Olsen et al., (1994) Jbac 176: 1-6 (2 pages)
    - Our PDF 98 pages





you will find the releases:

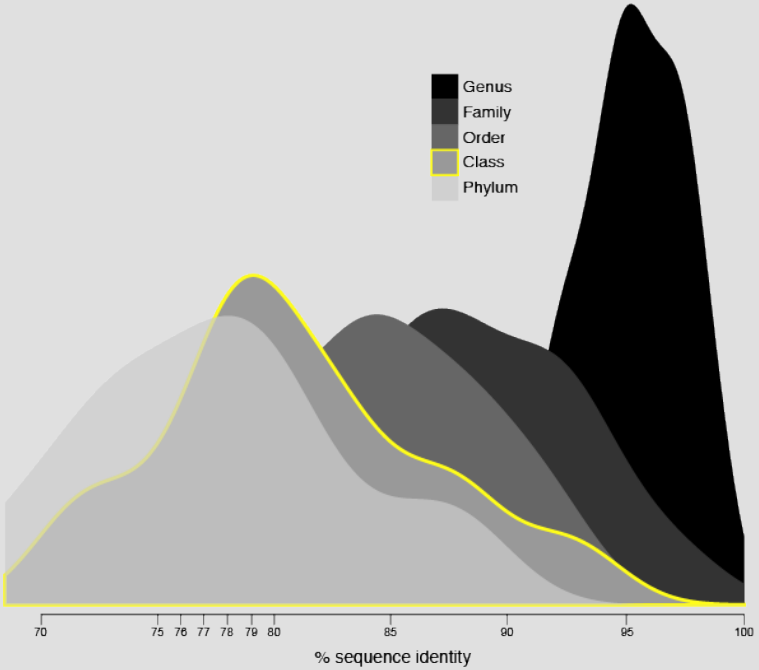
www.arb-silva.de/projects/living-tree



Thresholds calculated with the LTP 102

- Species threshold at 98.7% (Stackebrandt & Ebers, 2006, Microbiol. Today 33: 152-155)
- All taxa with ≥ 3 representatives
- Removing outliers (wrongly classified)
- Removing taxa pending of being classified
- historical consensus among taxonomists when circumscribing new genera and families

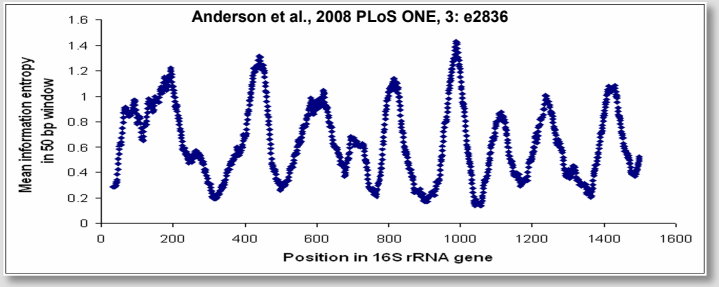
(8,602 species, 1,779 genera, 285 families, 115 orders, 52 classes, 29 phyla and 2 domains)

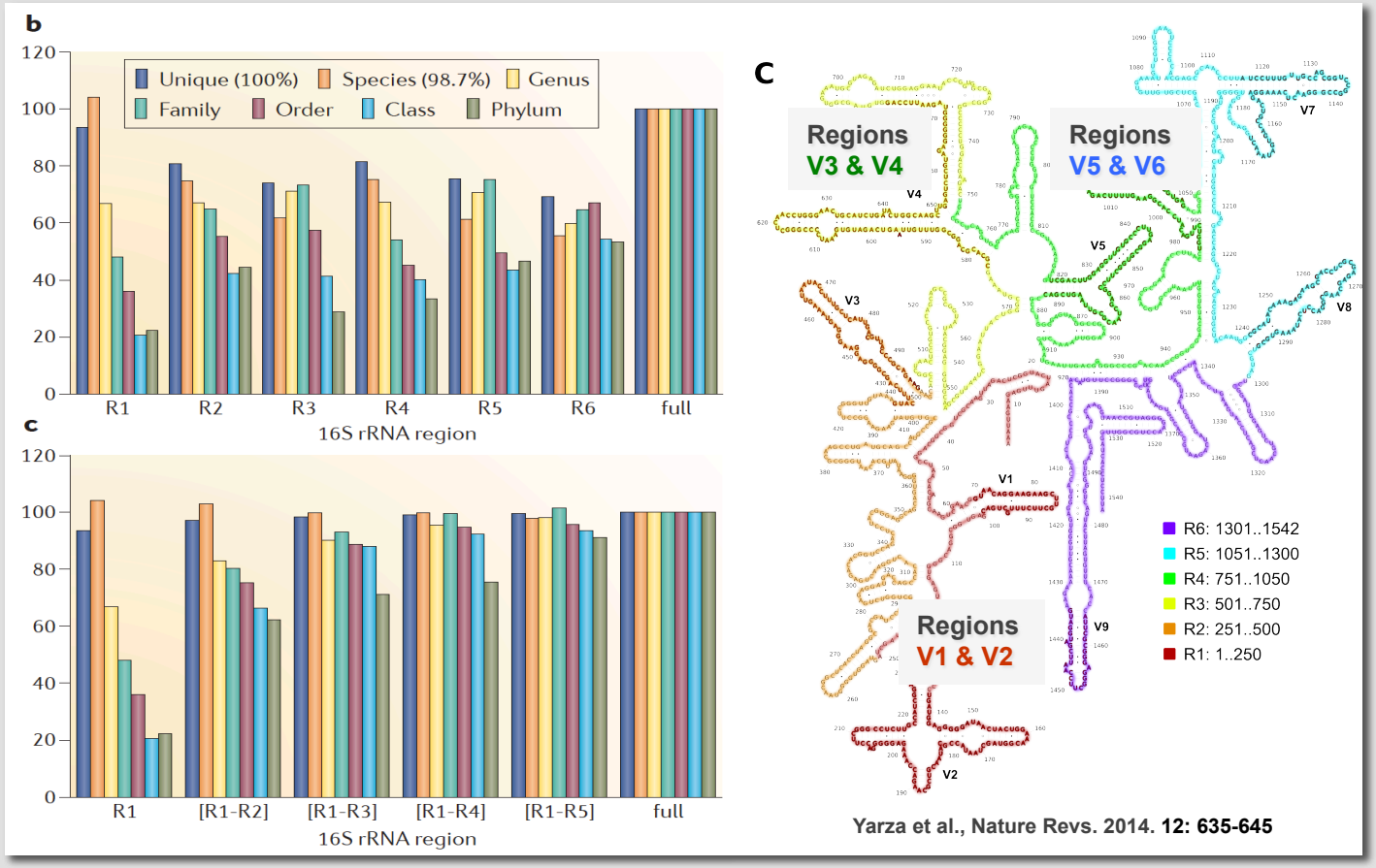


Category	minimum	mean
Species	98.7%	
Genus	94.5%	96.4 ± 0.25
Family	86.5%	91.7 ± 0.63
Order	82.0%	89.2 ± 0.93
Class	78.5%	86.4 ± 1,63
Phylum	75.0%	83.7 ± 2.15

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

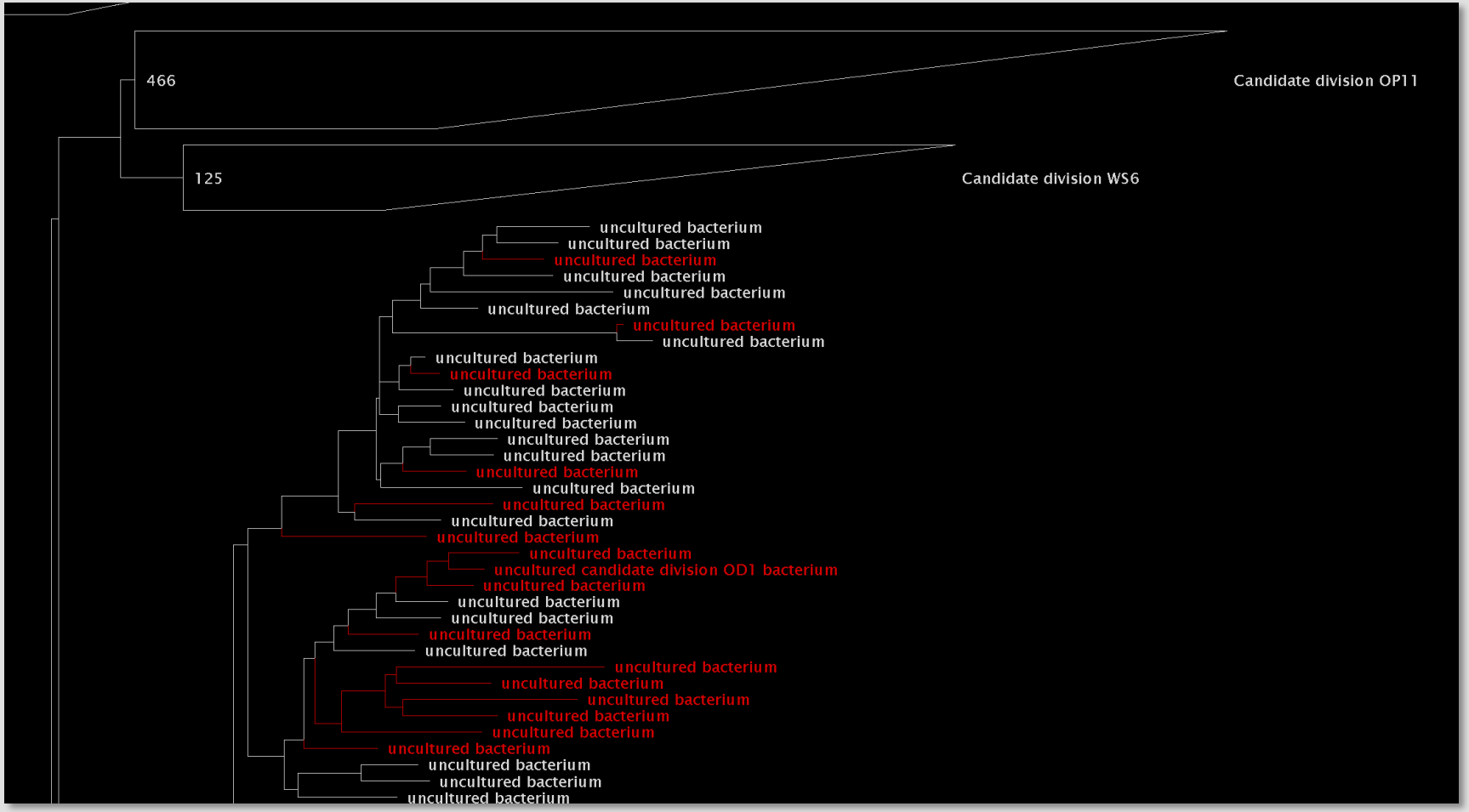
- The 5 ‘region (V1-V2) overestimates species
- The remaining regions tend to underestimate all taxa
- Increases in length tend to mirror full sequence





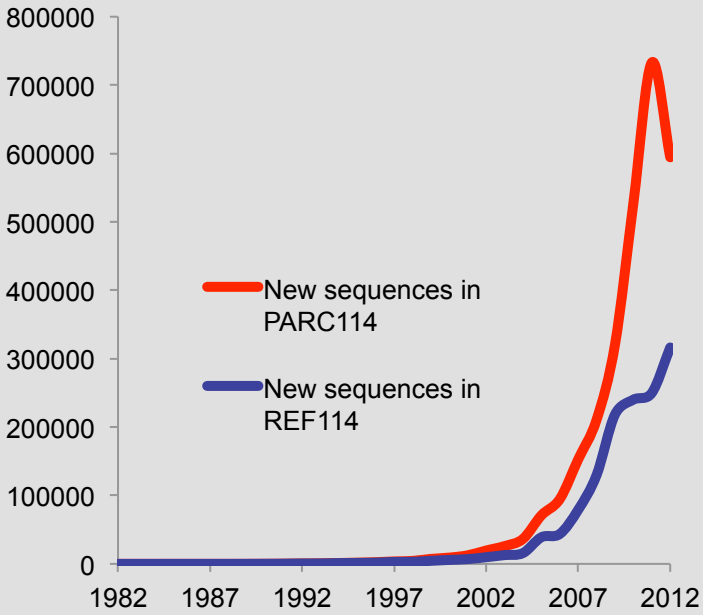
16S rRNA: classifying uncultured taxa ⇔ Candidate Taxonomic Units

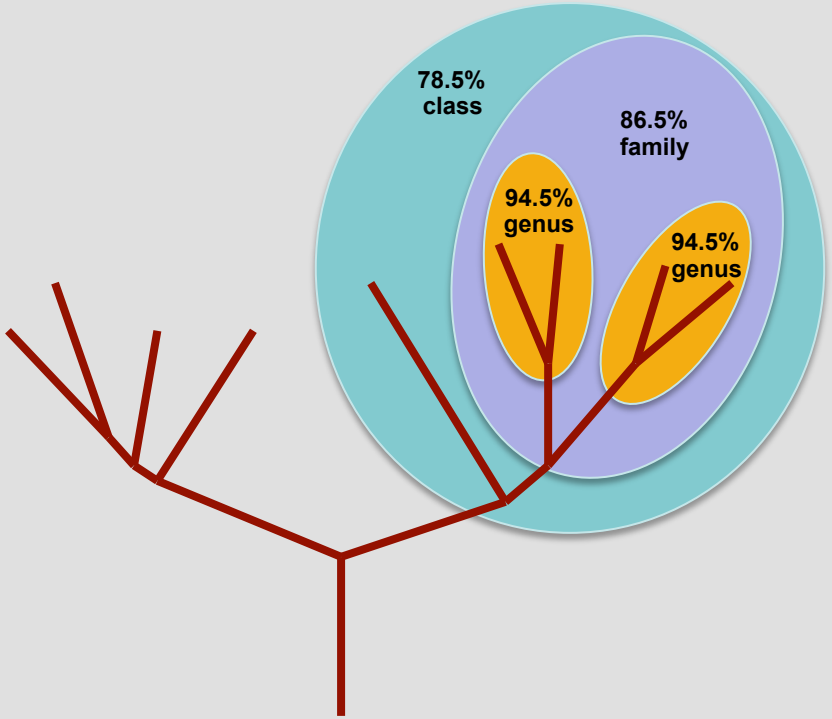
- Taking into account that the sequences may correspond to real organisms
- Need to be classified



16S rRNA: classifying uncultured taxa ⇔ Candidate Taxonomic Units

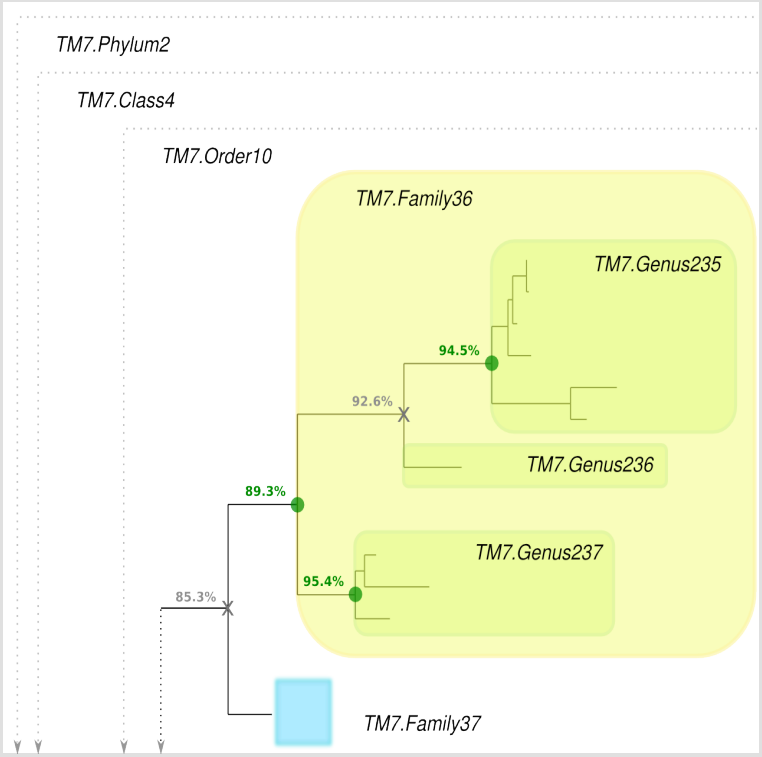
- SILVA_PARC (4.3 million)
- SILVA_REF (1.6 million)





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

CANDIDATE TAXONOMIC UNITS (CTU)



Nomenclature follows the format

X.ABC

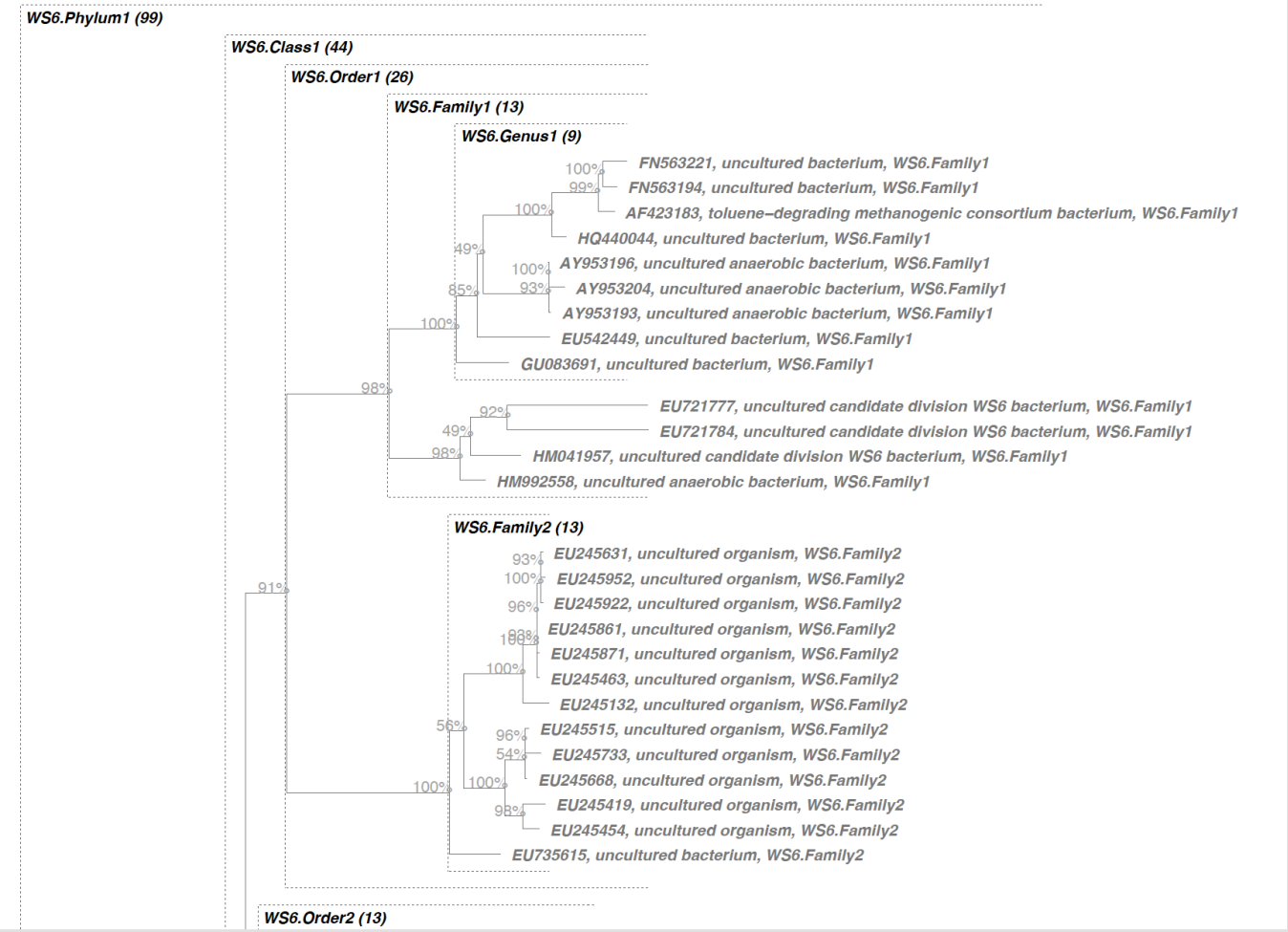
(eg. OP11.Family29-1)

- *X* is the candidate division or clade
- *A* is the name of the taxonomic rank, i.e. Species, Genus, Family, Order, Class …
- *B* is the taxon number
- *C* is an optional element that can be used to differentiate several CTUs of the same OTU

- ► **OP11.Phylum2**
- ► **OP11.Class3**
- ► **OP11.Order1**
- ► **OP11.Family4**
- ► **OP11.Genus20-1**
- ► **OP11.Species30**

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

We have classified all the environmental clades and implemented in the ARB database

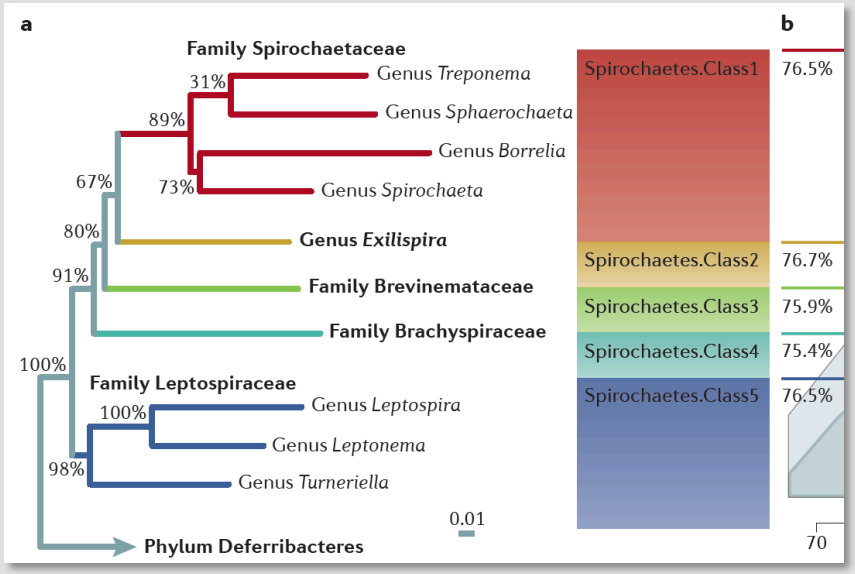


Yarza et al., Nature Revs. 2014. 12: 635-645

Harmonizing categories helps in taxonomic **re-evaluations**: e.g. phylum ***Spirochaetes***

- **Currently:**
- **1 Phylum**
  - **1 Class**
    - **1 Order**
      - **4 Families**
        - **16 Genera**
          - **112 species (82 informative)**

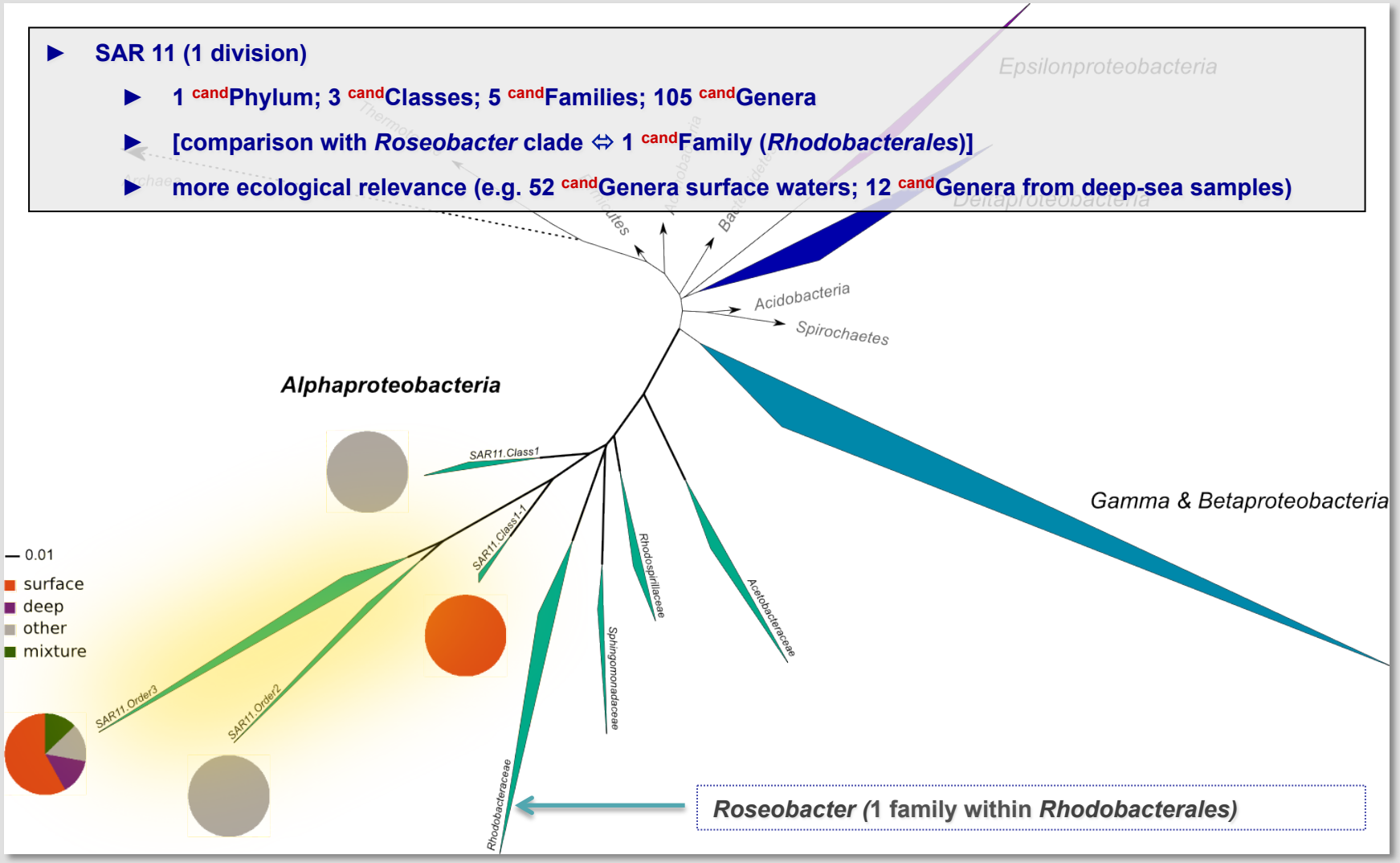
- **re-evaluating:**
- **1 Phylum (5 Classes)**
  - **Class 1 (*Borrelia, Treponema, Sphaerochaeta, Spirochaeta*)**
  - **Class 2 (*Exilispira*)**
  - **Class 3 (*Brevinema*)**
  - **Class 4 (*Brachyspira*)**
  - **Class 5 (*Leptospira, Leptonema, Turneriella*)**



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

Harmonizing categories helps in devising comparable units for ecological studies: e.g. **SAR 11**

- **SAR 11 (1 division)**
  - **1 candPhylum; 3 candClasses; 5 candFamilies; 105 candGenera**
  - **[comparison with *Roseobacter* clade ⇔ 1 candFamily (*Rhodobacterales*)]**
  - **more ecological relevance (e.g. 52 candGenera surface waters; 12 candGenera from deep-sea samples)**



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

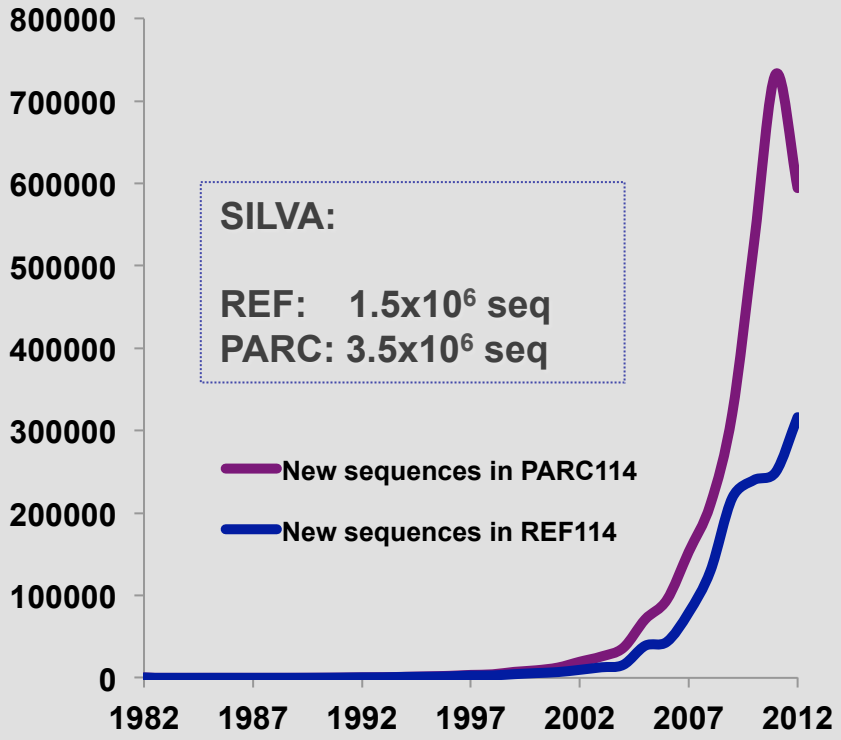
16S rRNA: counting number of detected taxa (2012)

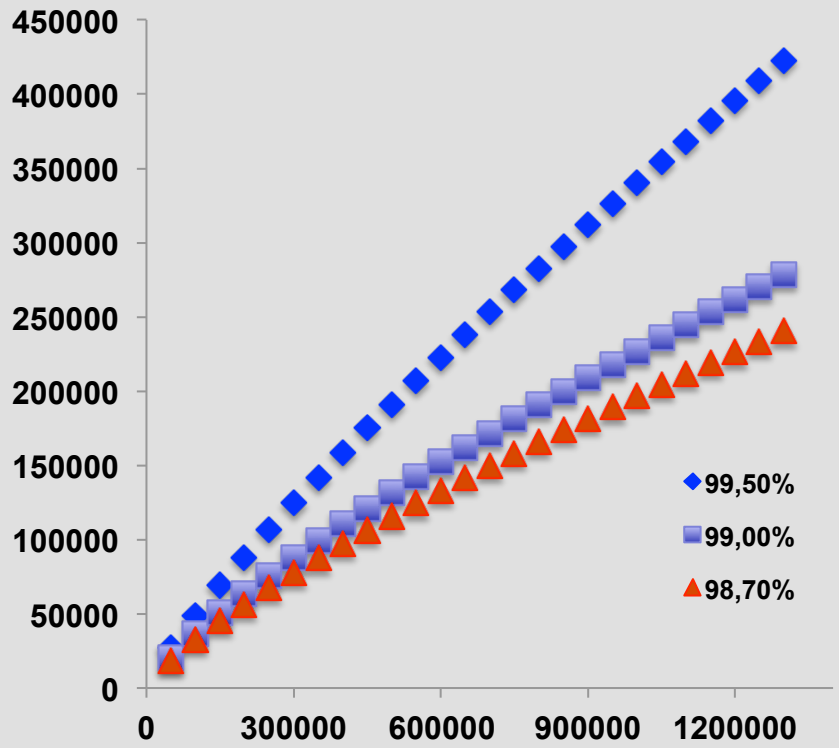
	LPSN 2012		SILVA REF 114		
	Bacteria	*Archaea*	*Bacteria*	*Archaea*	Ambiguous* *Bact + Arch*
Sequences	--	--	1,265,442	41,228	31,469
Species	**9,624**	**391**	**246,841**	**13,159**	28,983
Genera	1,916	113	90,068	4,383	19,607
Families	290	27	16,719	1,239	4,803
Orders	122	15	6,230	682	1,148
Classes	79	9	2,969	431	408
Phyla	**27**	**2**	**1,481**	**287**	84

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

The increase in deposited sequences:

- Rarefaction: no saturation yet
- 99.5% (error ⇔ noise)



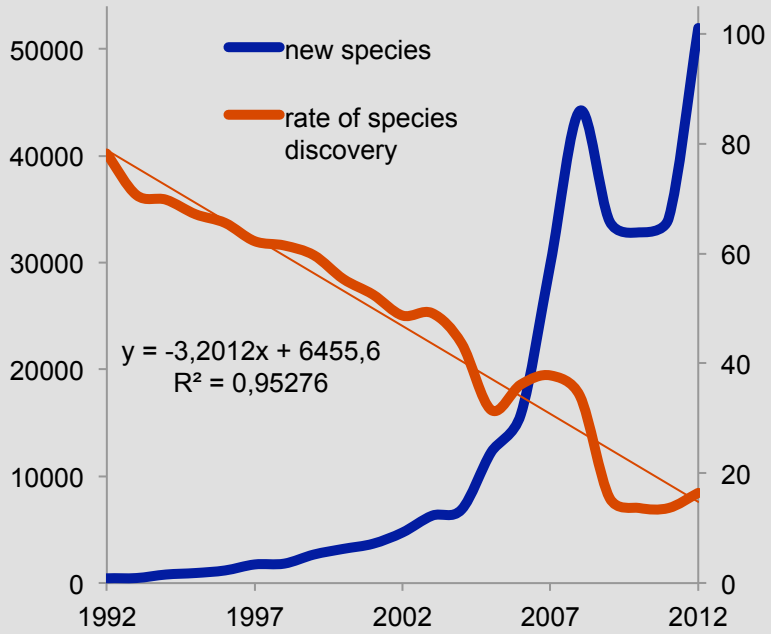


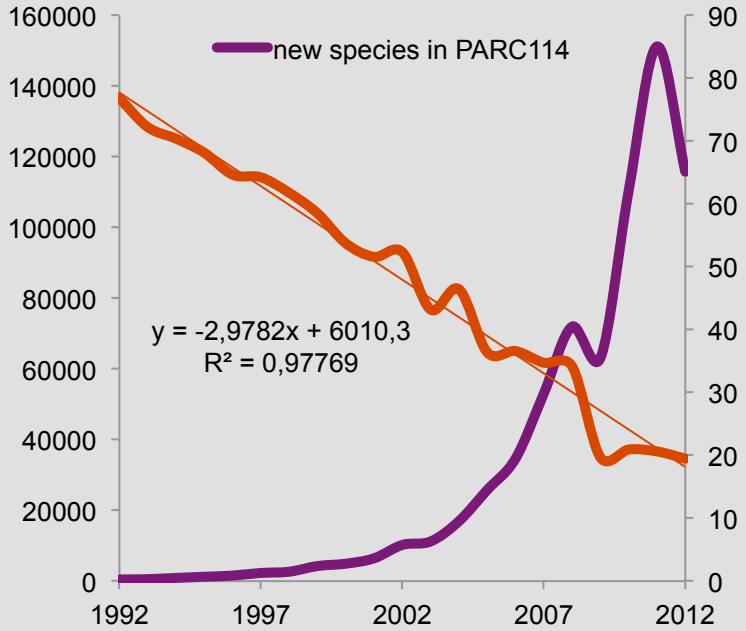
Yarza et al., Nature Revs. 2014. 12: 635-645

At the end of this decade

⇒ NO MORE NEW SPECIES will be discovered (saturation)

⇒ With a rate of **50.000**/yearly ⇔ **500.000** species (saturation)





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

- WE CURRENTLY HAVE
  - 11,000 SPECIES DESCRIBED
  - 250,000 SPECIES DETECTED IN THE SILVA REF 111 (2012)
  - 500,000 EXPECTED AT THE END OF THE DECADE (2020)
  - DO WE EXPECT
    - 2x (1,000,000)?
    - 4x (2,000,000)?

REDUNDANCY IN THE ENVIRONMENTS STUDIED, PERHAPS SEARCHING IN UNEXPLORED ECOSYSTEMS WILL IMPROVE THE NUMBERS

1,000,000 SPECIES IS AN ACHIEVABLE NUMBER TO BE CLASSIFIED