

<http://rast.nmpdr.org>

0 sec

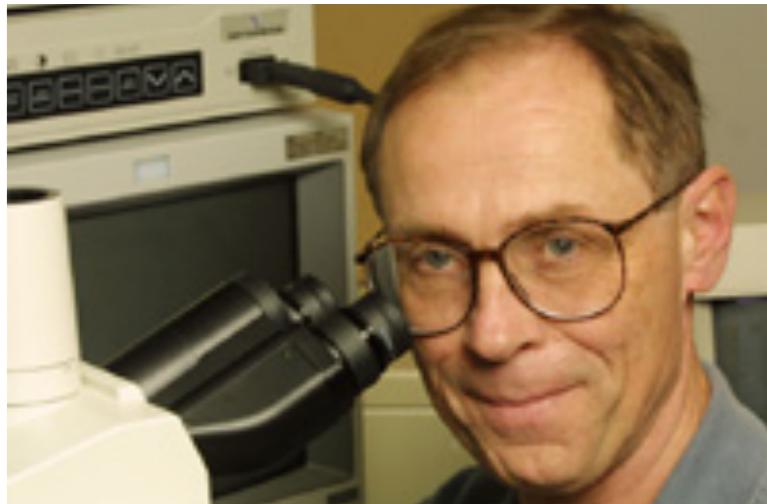
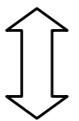
stru

elles



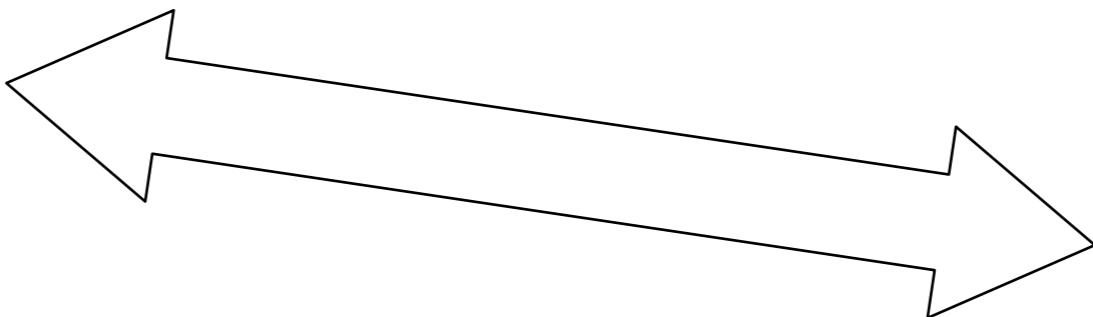


**Carl Woese**  
U. Illinois



**Norm Pace**  
CU Boulder

# My academic lineage



**S. Giovannoni**  
Oregon State

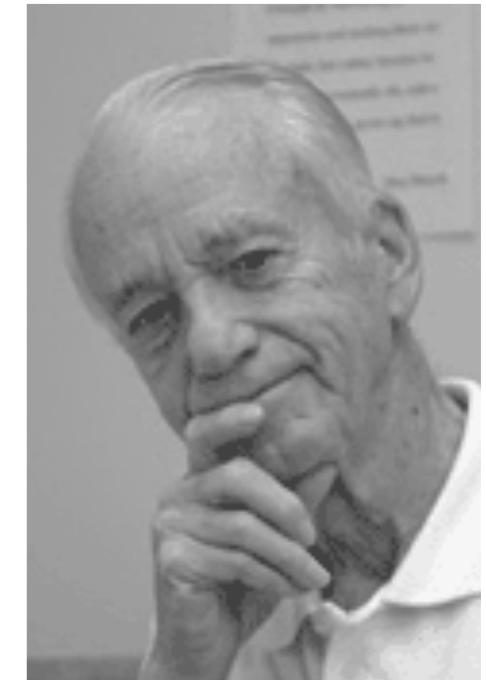
**D. Stahl**  
Univ. of Washington

**E. Engert**  
Cornell

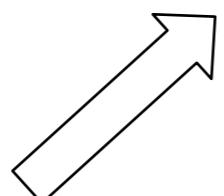
**E. DeLong**  
MIT

**Phil Hugenholtz**  
U Brisbane

**Scott Dawson**  
UC Davis



**Ralph Wolfe**  
U. Illinois



**Tom Schmidt**  
U Mich



Scott Dawson - IMDb

www.imdb.com/name/nm1869993/?ref\_=fn\_al\_nm\_5

WOULD YOU RISK YOUR MARRIAGE TO S...

**IMDb**

Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

 **Scott Dawson** (IX)

Contribute to IMDb. Add a bio, trivia, and more.  
[Update information for Scott Dawson »](#)

[More at IMDbPro »](#)

 [Represent Scott Dawson? Add or change photos](#)

[1 video »](#)



**Filmography** Hide all Show by... Edit

**Self** (1 credit) Hide

**Journey Into Amazing Caves** (Documentary short) 2001  
 Himself - U.C. Berkeley (as Scott Dawson Ph.D.)

**Related Videos**

  
on IMDb 01:24  
Trailer Journey Into Amazing Caves 



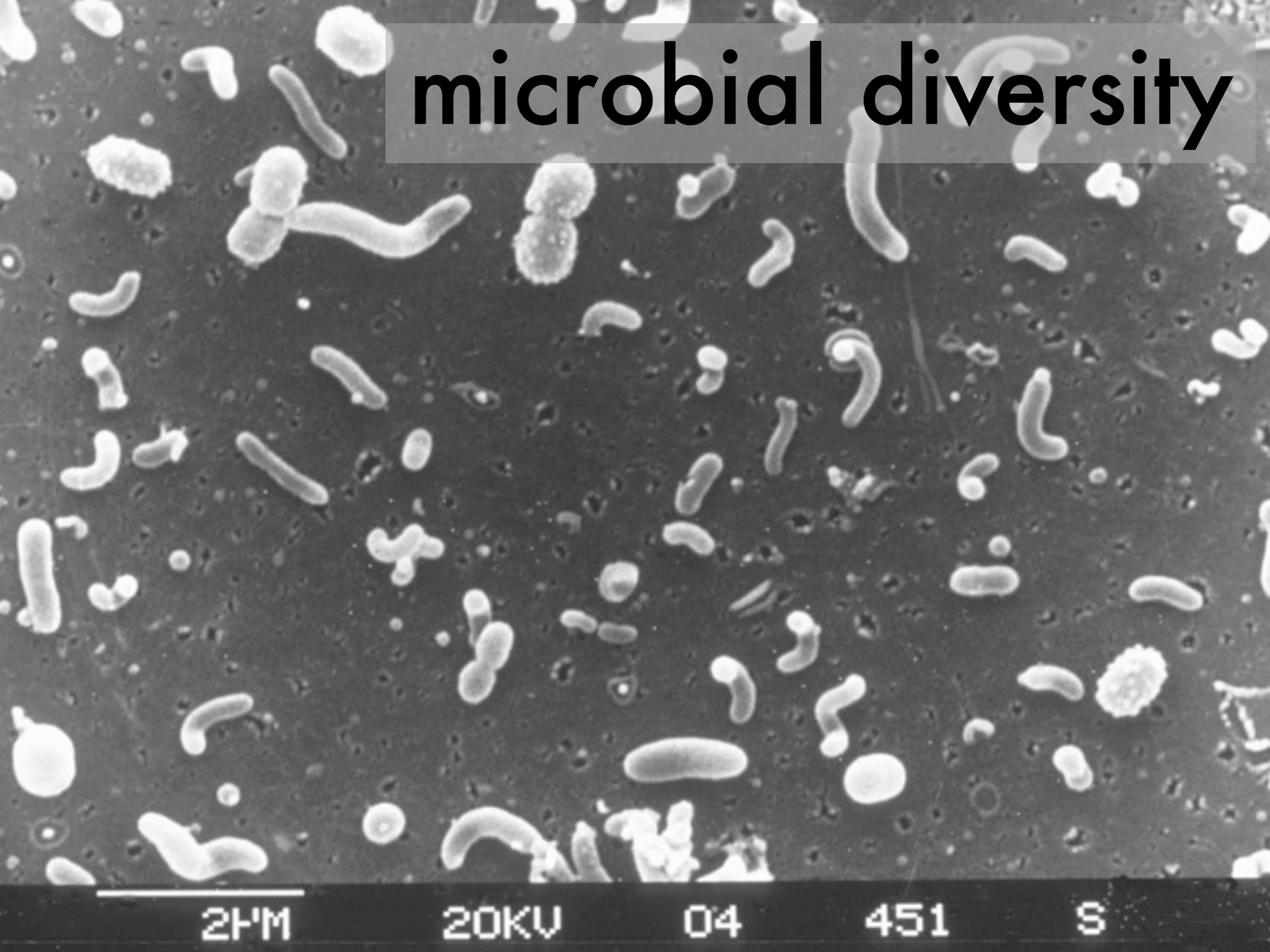
Grand Prismatic Spring, Yellowstone National Park.

JIM URQUHART / REUTERS

# The Man Who Blew The Door Off The Microbial World

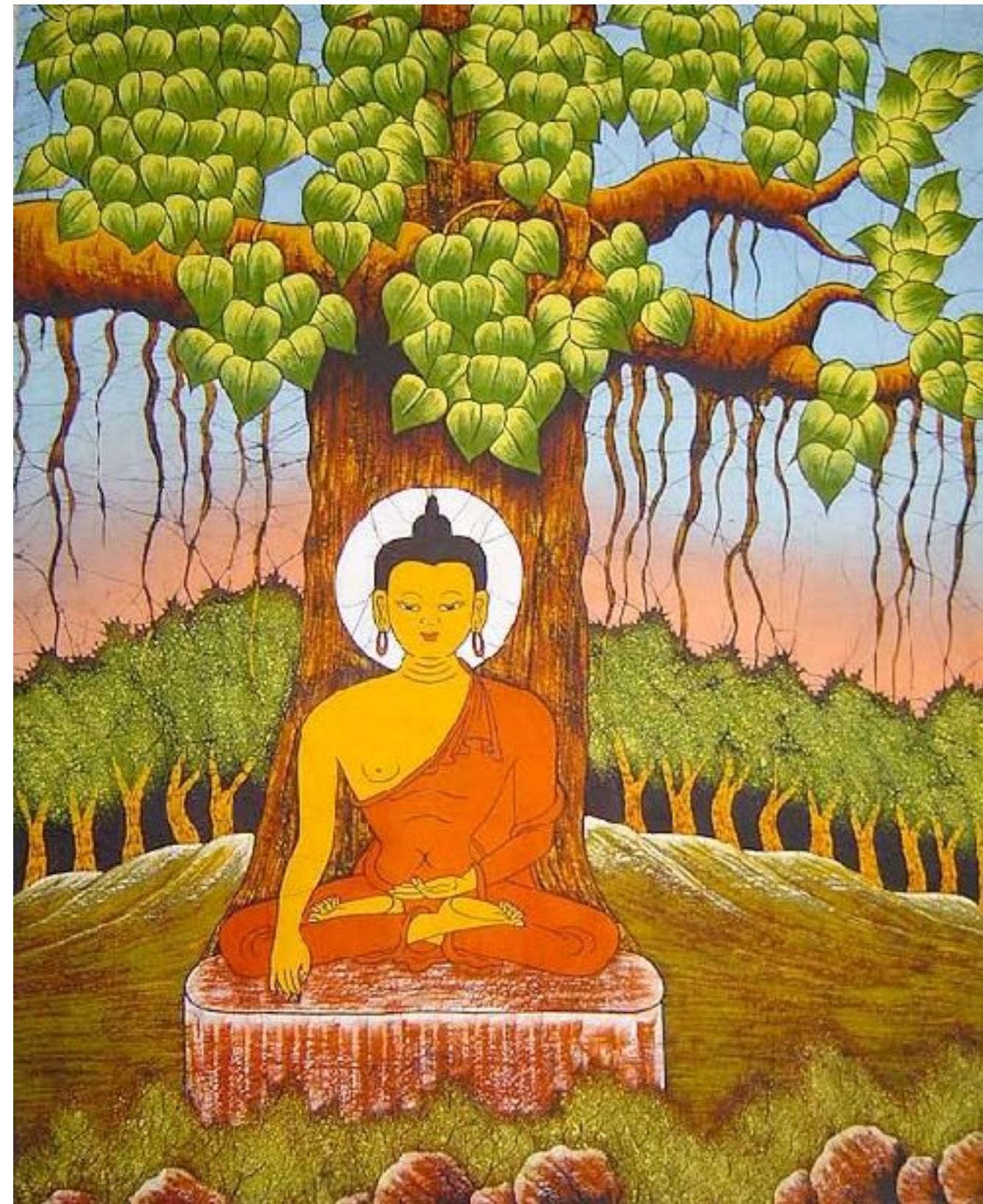
Norm Pace has spent his life as an explorer, charting dangerous caves and ushering in the golden age of microbiology.

# microbial diversity

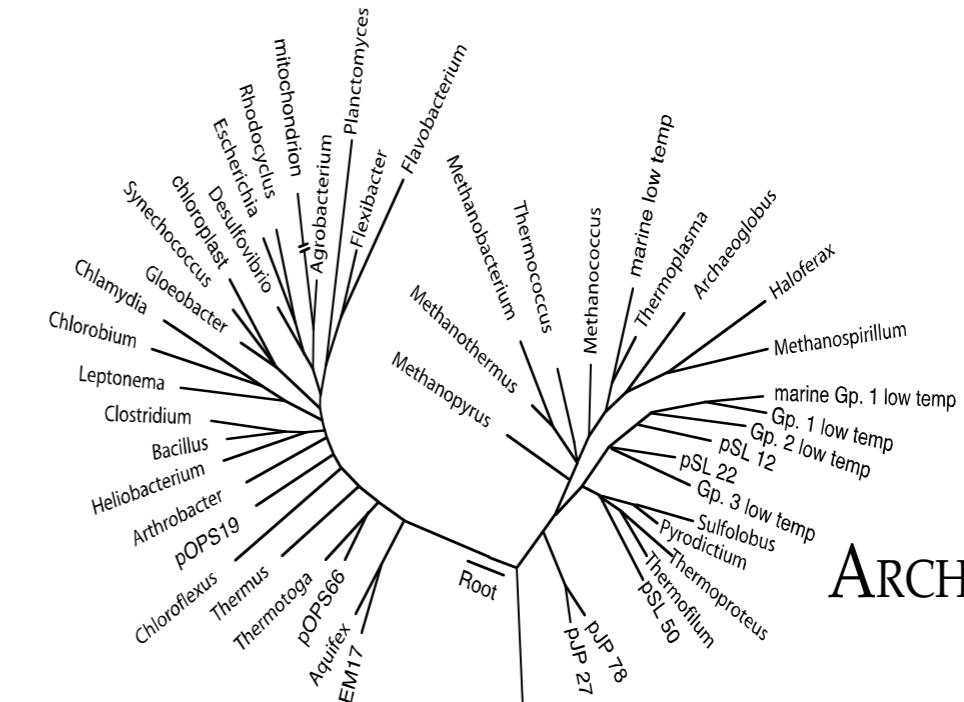


How to define  
microbial diversity?

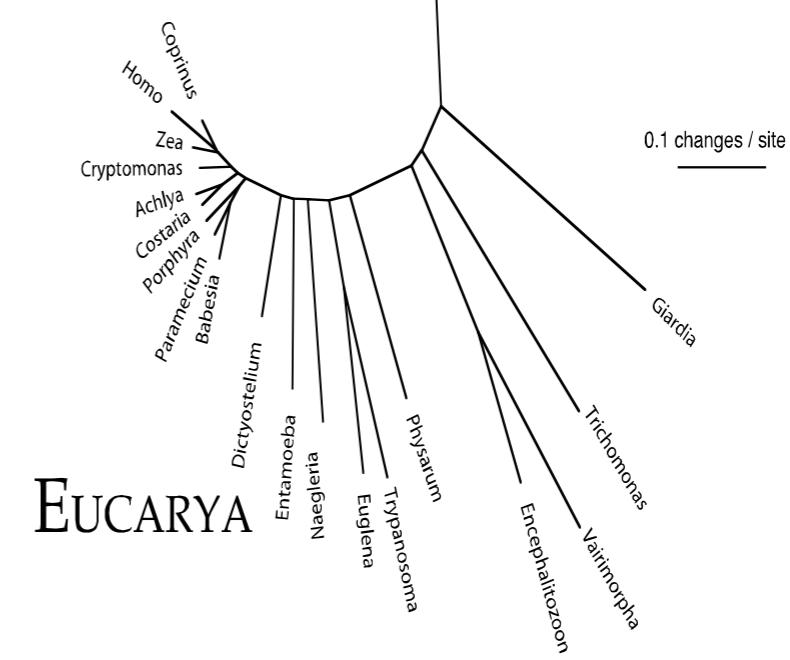
# Ponderings under the Big Tree



BACTERIA

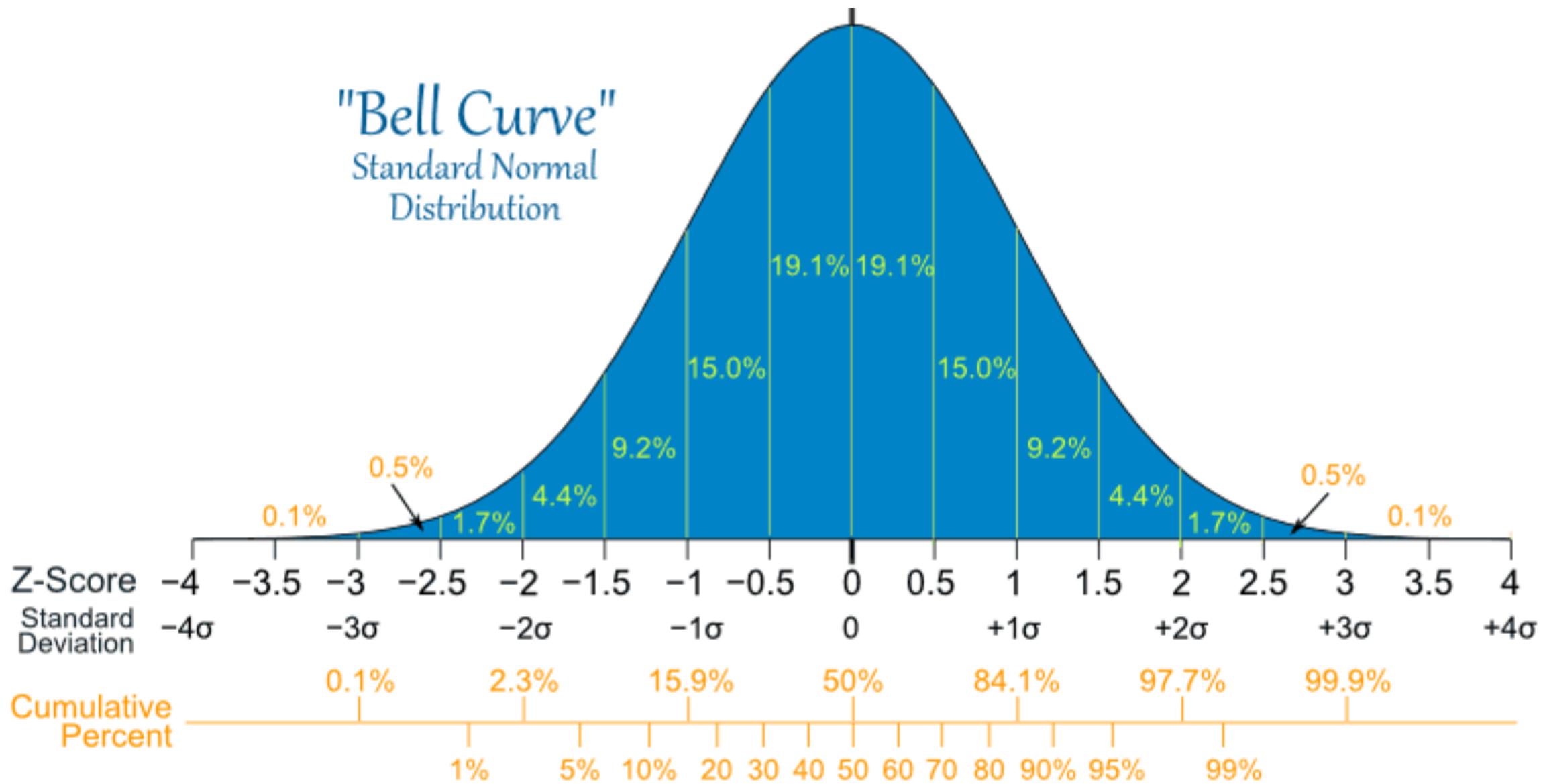


ARCHAEA



EUCARYA

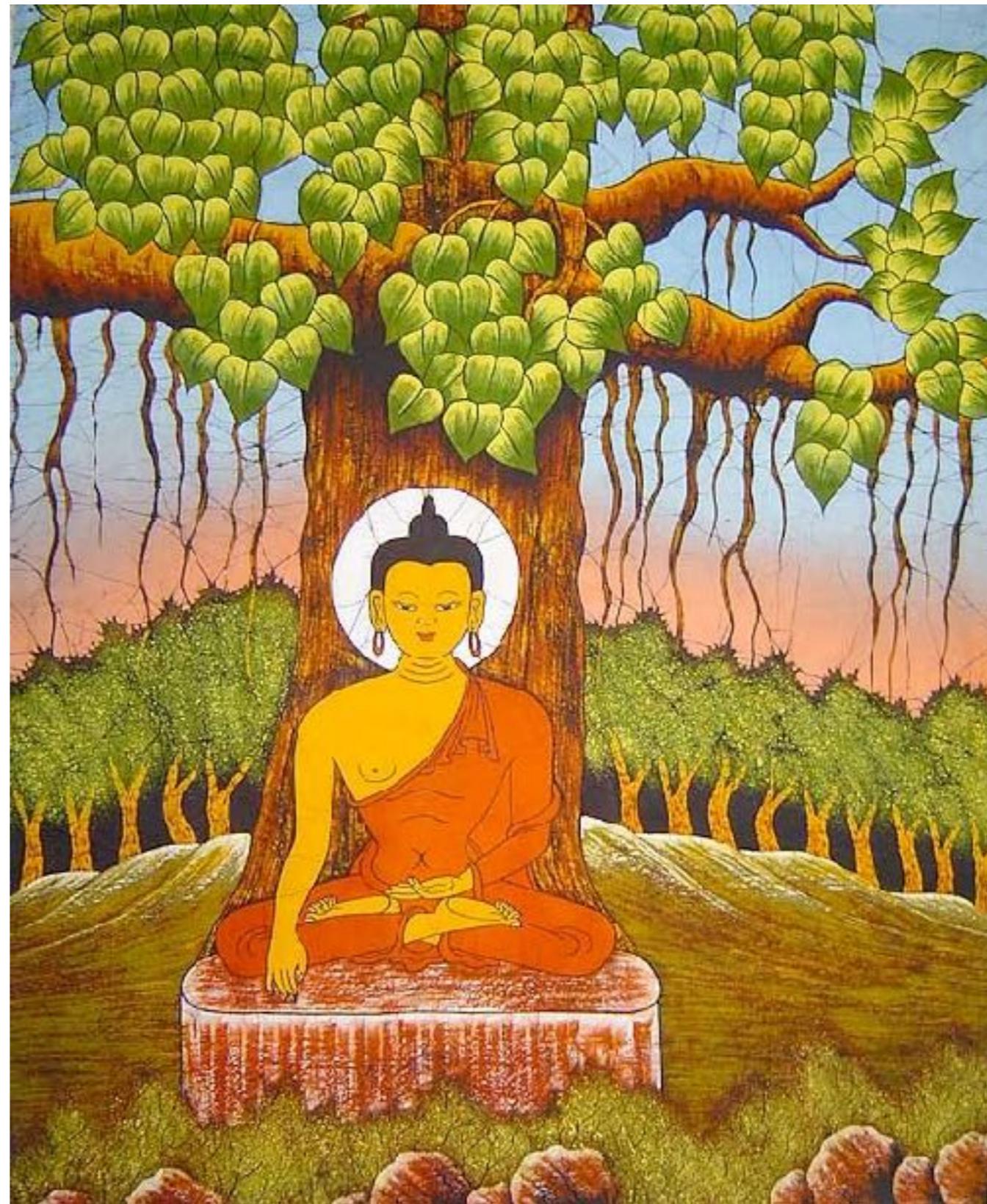
# it's all about the variation...



but what's “normal”?

# *Zen and the Art of Molecular Phylogeny*

**Steps on the Road to  
Evolutionary  
Enlightenment**



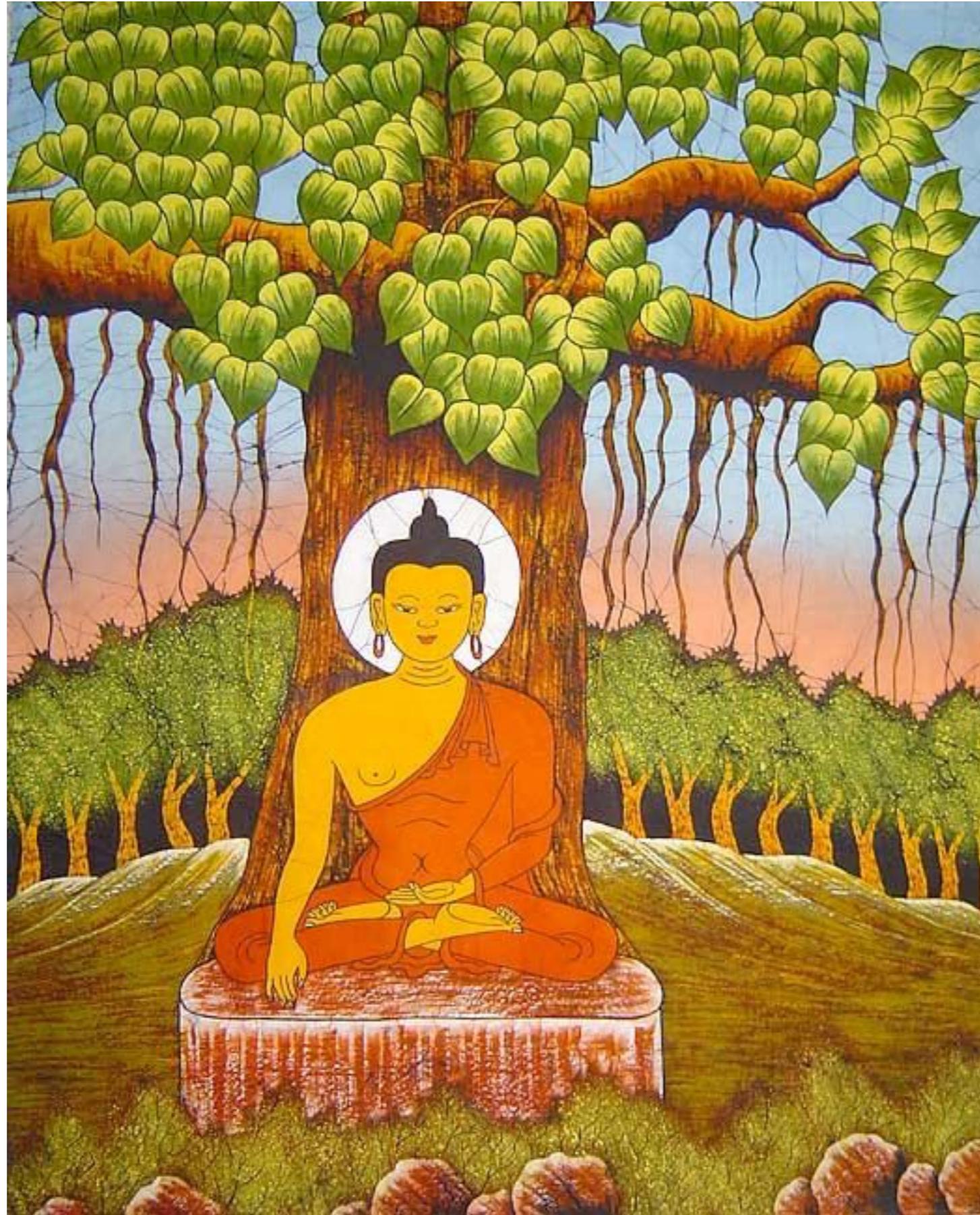
# *Molecular phylogeny is the basis of many aspects of microbial diversity*

- infer organismal evolutionary relationships using one or more genetic sequences
- help to build multiple sequence alignments
- estimate divergence times of lineages using molecular clocks
- classify “families” of genes and detect **orthologs/paralogs/xenologs**
- reconstruct ancestral genomes at nodes
- identify genetic inventories that covary with phenotypic traits (phylogenomics)
- examine evolutionary patterns and/or processes

# Questions we will consider:

- How do we define homologs?
  - How do we align homologous sequences?
  - How can we read/ build phylogenetic trees?
  - How do we define a bacterial species?
- 
- How do we know what we \*know\*?

# *How the hell did I get under this Big Tree?*



# Taxonomy or phylogeny?

**Table 17.2 Taxonomic hierarchy of classification.**

Taxon rank	A long-studied taxon	A less-studied taxon	An uncultivated environmental sample
Domain	Bacteria	Archaea	Bacteria
Division (phylum)	Actinobacteria Filamentous gram-positive	Euryarchaeota Methanogens and halophiles	Proteobacteria Purple bacteria and relatives; gram-negative
Class	Actinobacteria High GC gram-positive	Methanococci Methanogens	Alpha Proteobacteria Gram-negative bacteria
Subclass	Actinobacteridae		
Order	Actinomycetales Filamentous; acid-fast stain	Methanococcales Methanogenic cocci	Rickettsiales Includes intracellular bacteria
Suborder	Streptomycineae		
Family	Streptomycetaceae Filamentous; hyphae produce spores	Methanocaldococcaceae Thermophilic methanogens	SAR11 cluster Nonculturable planktonic marine bacteria
Genus	<i>Streptomyces</i>	<i>Methanocaldococcus</i>	<i>Pelagibacter</i>
Species (date first described)	<i>S. coelicolor</i> (1908)	<i>M. jannaschii</i> (1984)	<i>P. ubique</i> (2002)

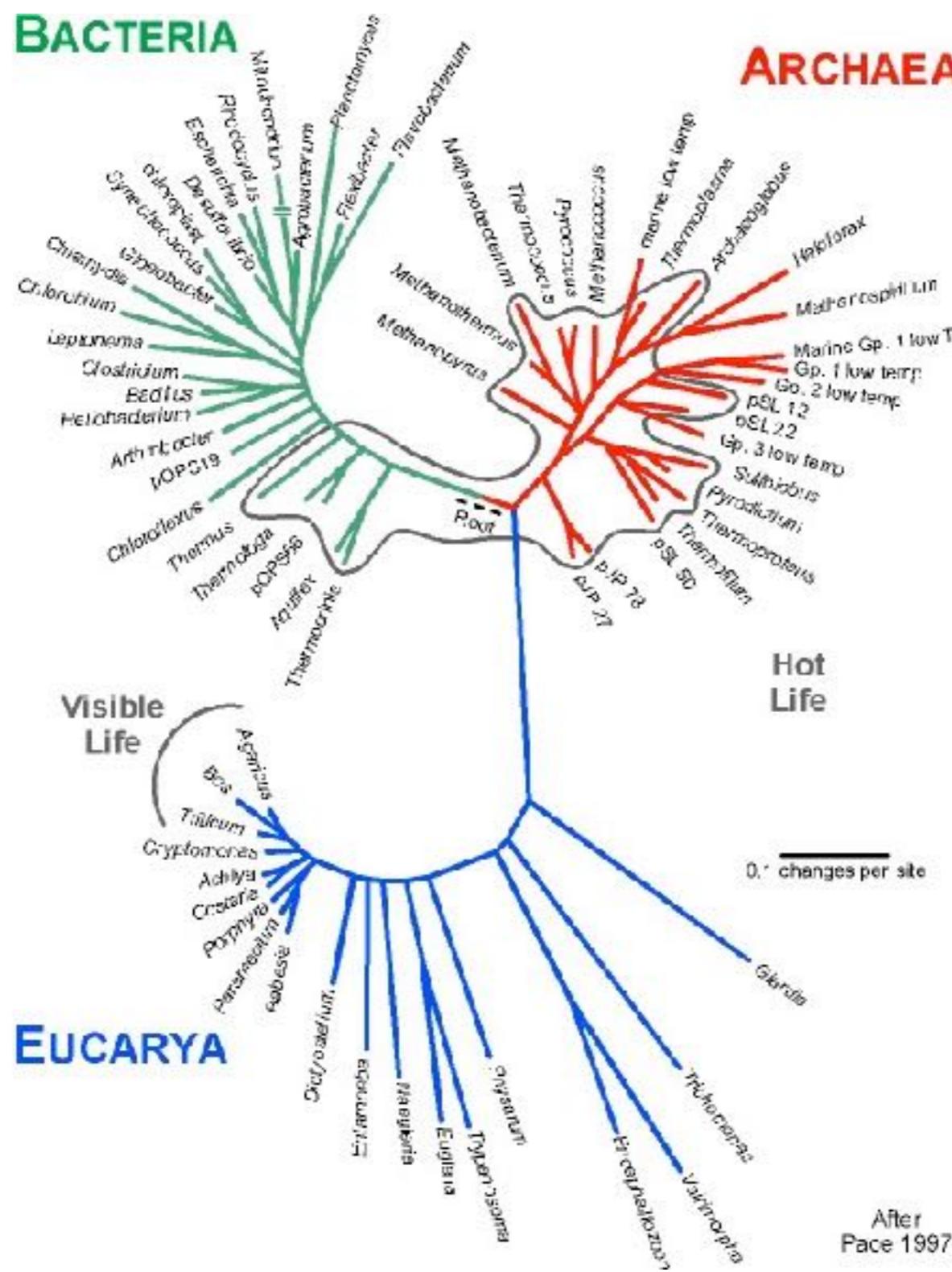
Table 17.2 Microbiology: An Evolving Science

© 2009 W.W. Norton & Company, Inc.

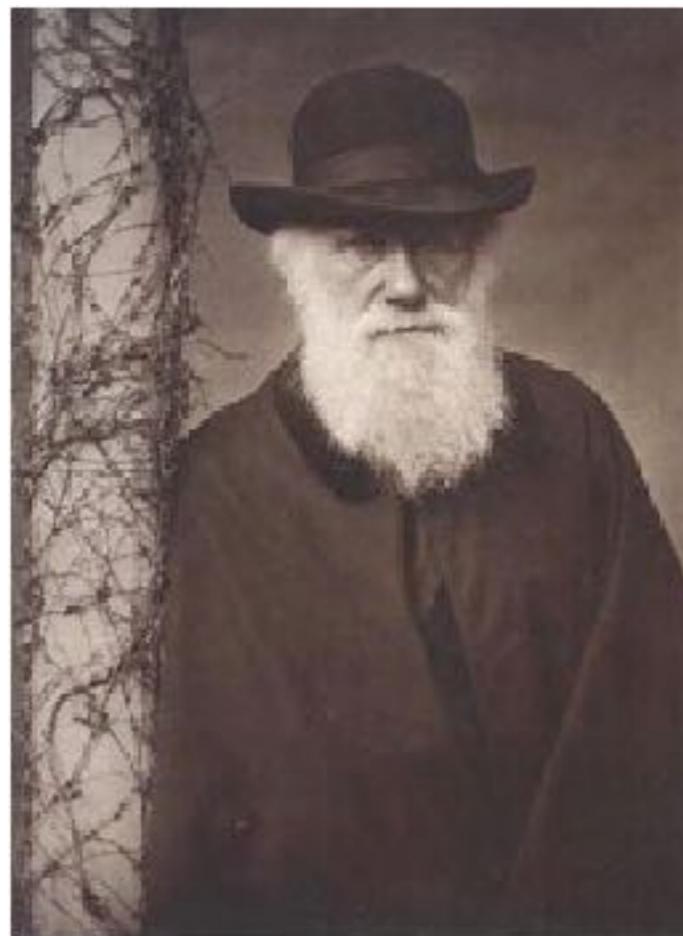
**Taxonomy is based on phenotype,  
molecular is phylogeny based on genotype.**

**Physiological characteristics of microbes  
(e.g., metabolism) rarely seem to track with  
evolutionary relationships. Why?**

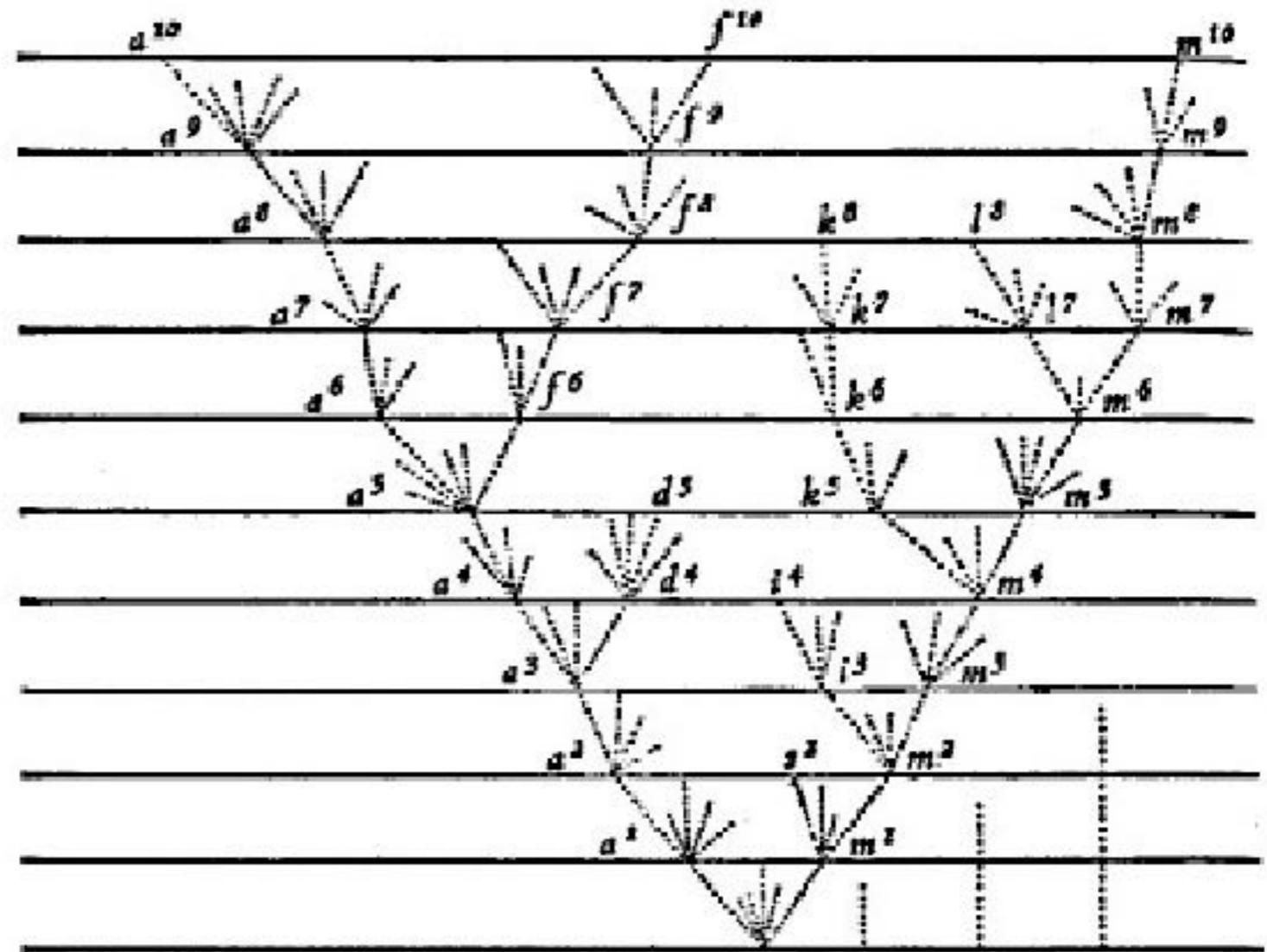
# How to read trees



# Common ancestry gives rise to independent lineages



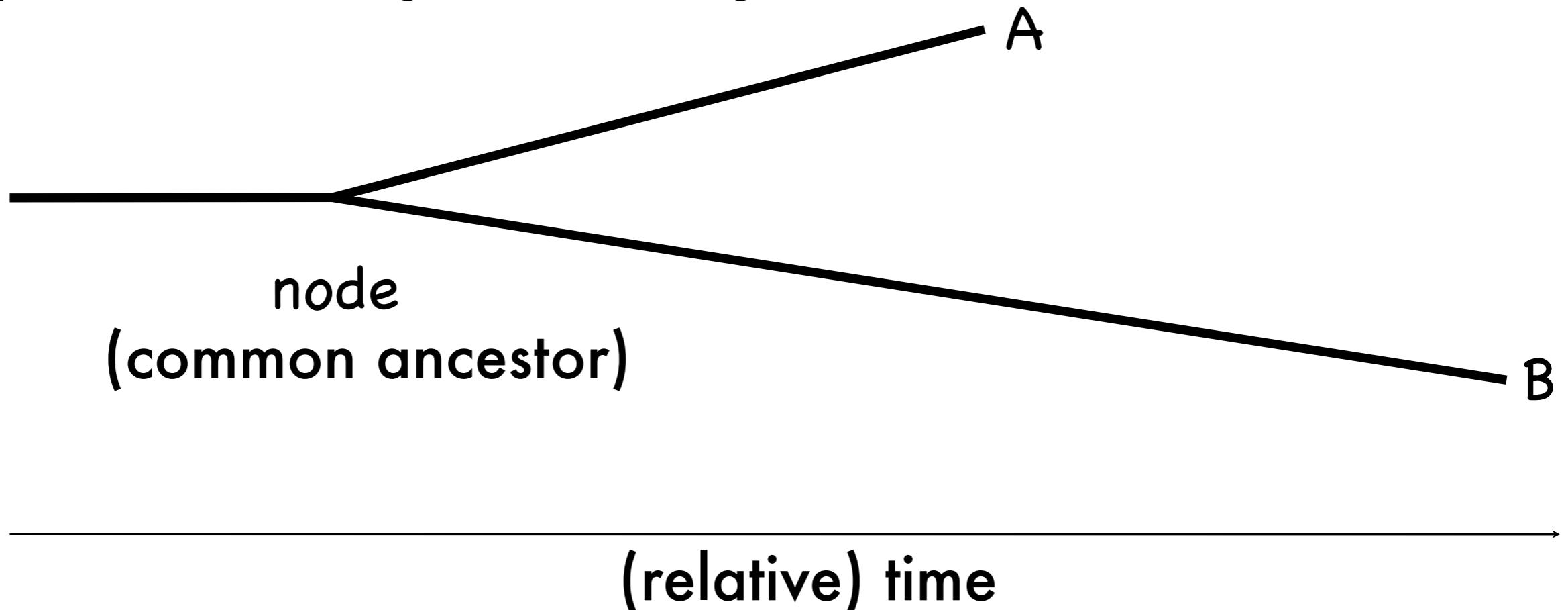
Charles Darwin  
The Origin of Species  
1859



"A structure is similar among related organisms because those organisms have all descended from a common ancestor that had an equivalent trait."

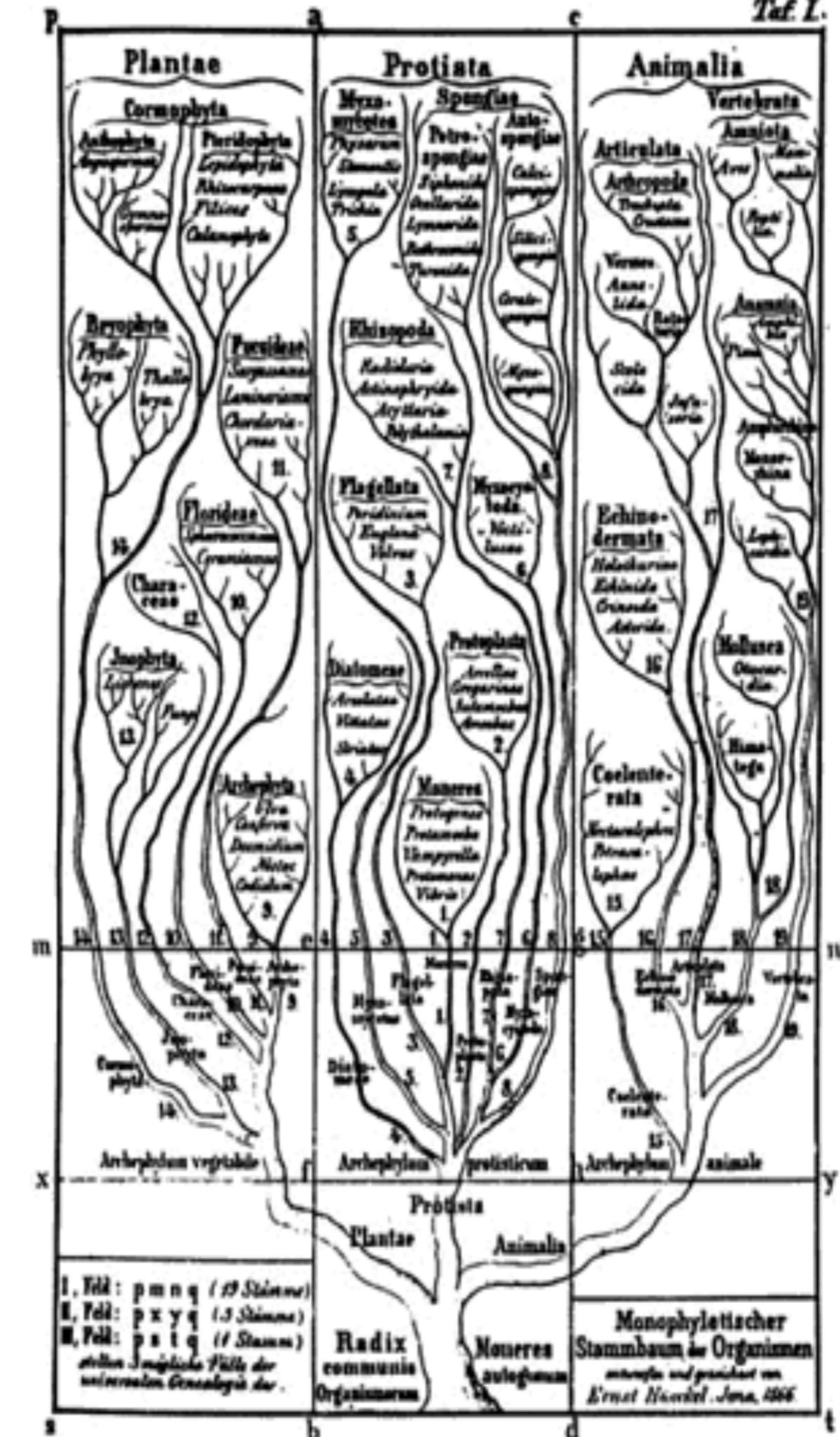
# Common ancestry around 1843

- Ancestral lineage gives rise to 2 independent lineages
- Length of line proportional to amount of change (variable)
- Natural groups, or “clades”, share certain properties (one invention)
- Mode = pattern of change (tree topology)
- Tempo = rate of change (branch length)



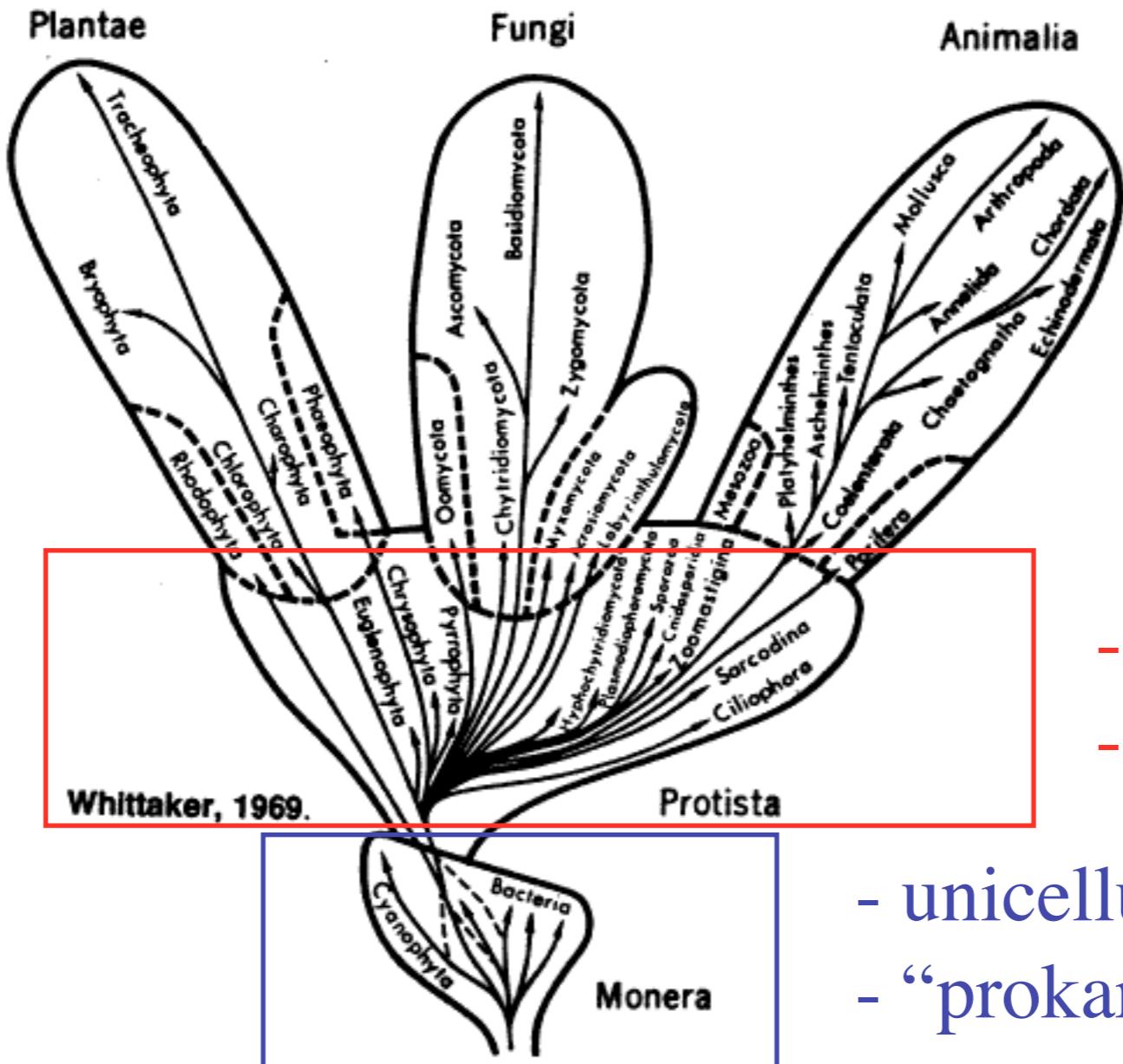


# Ernst Haeckel



\*

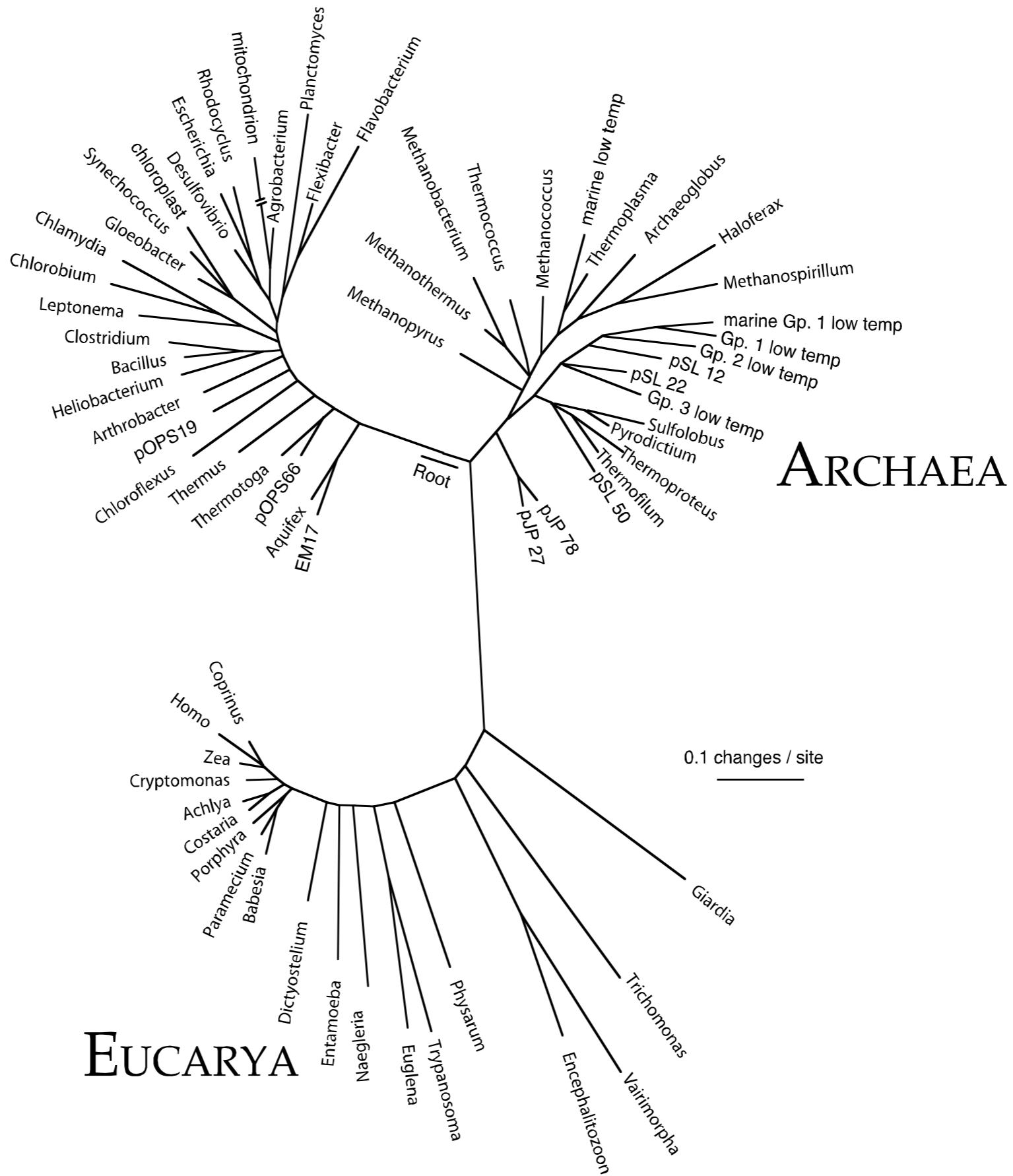
# The 5 Kingdom Scheme



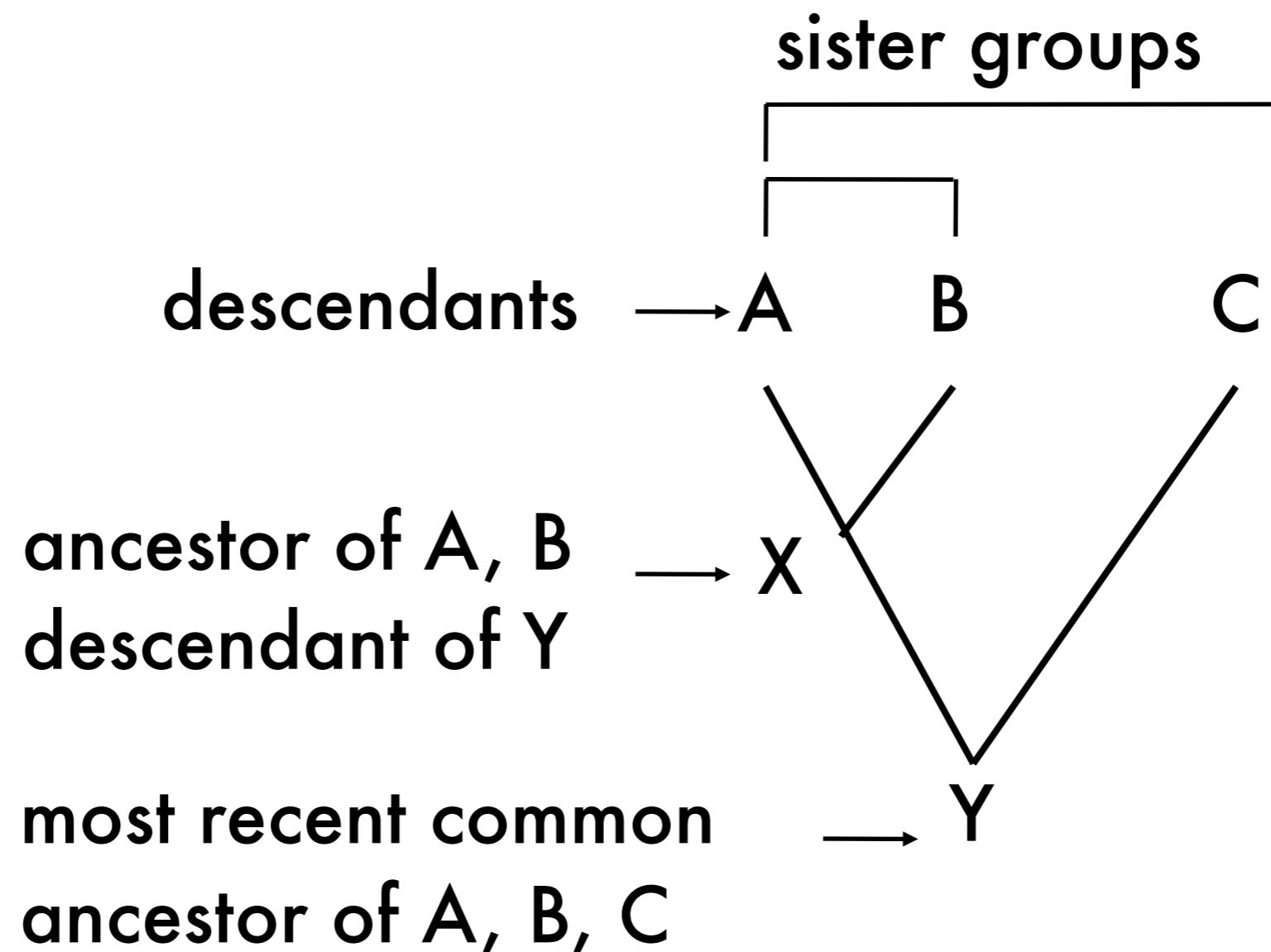
- unicellular
- eukaryotic

- unicellular
- “prokaryotic”

# The Big Tree BACTERIA



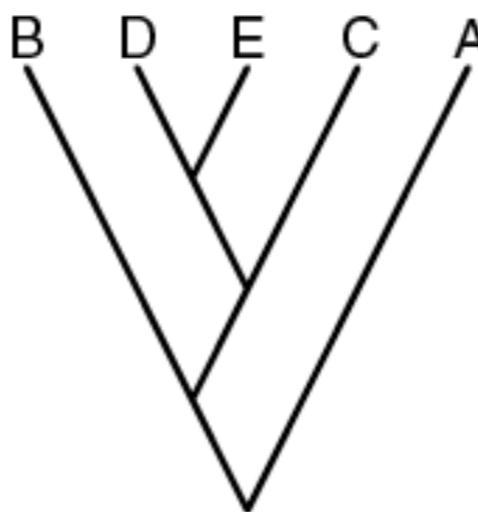
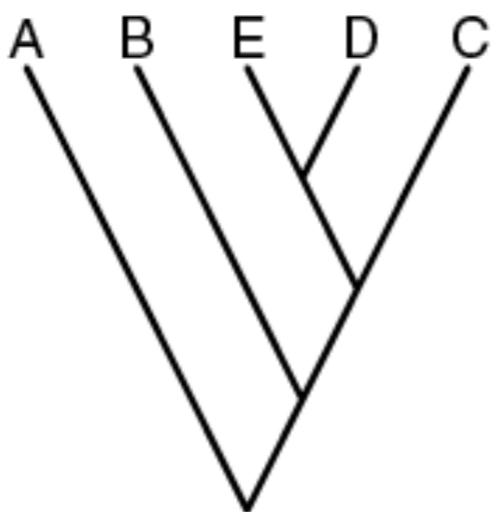
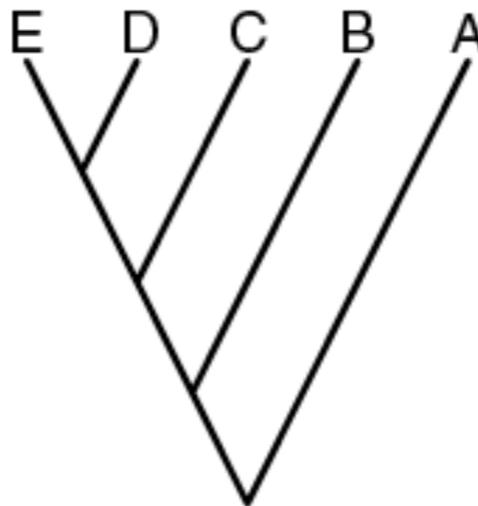
# “natural” groups of organisms (or genes)



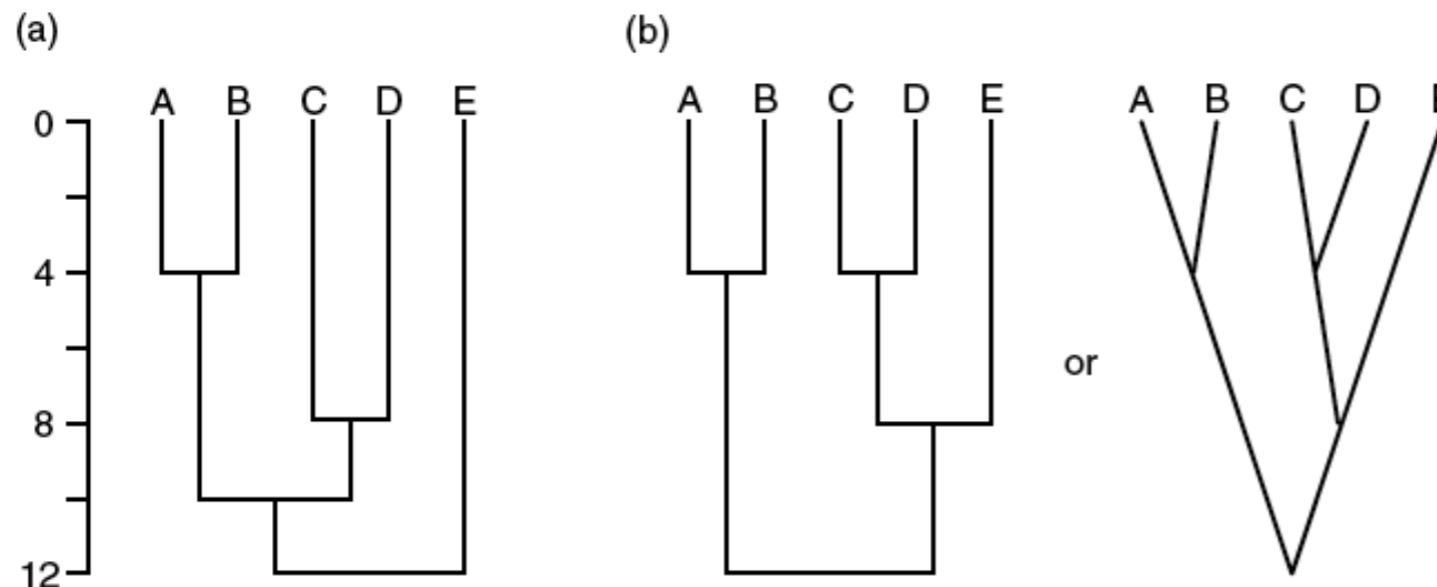
*think about your own genealogy.....*

## Branches can rotate around nodes

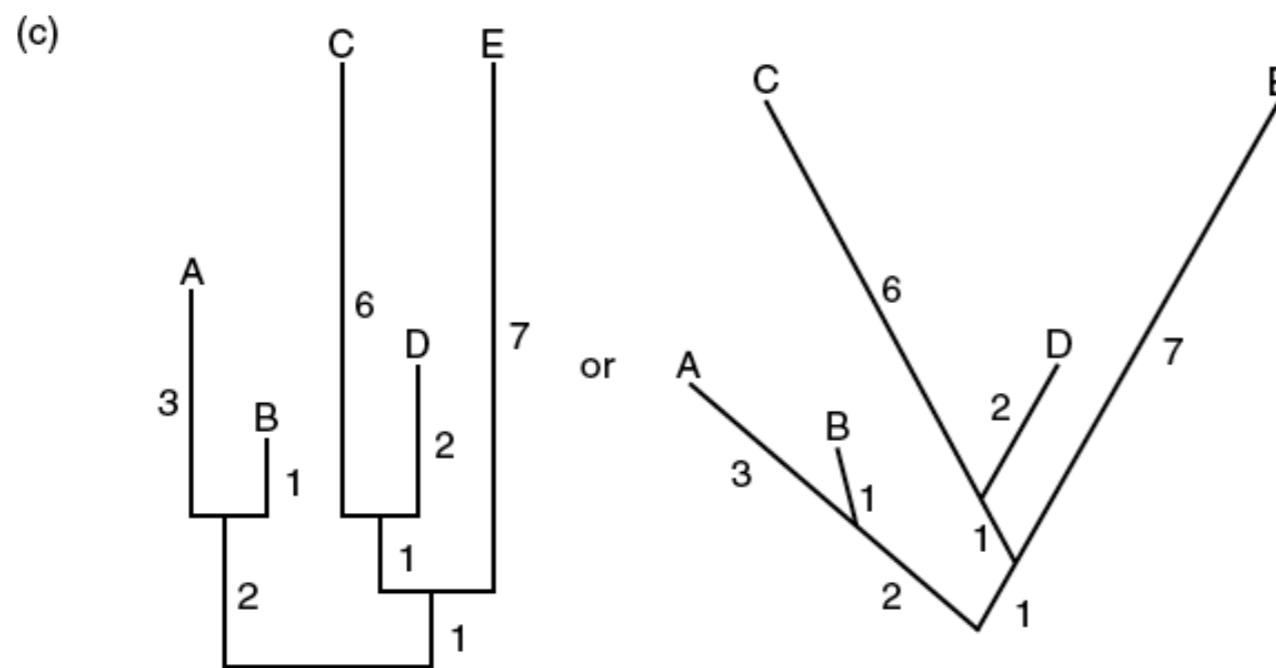
- \* Nodes are can rotate 360° - like a mobile



# Cladograms vs. phylogenograms



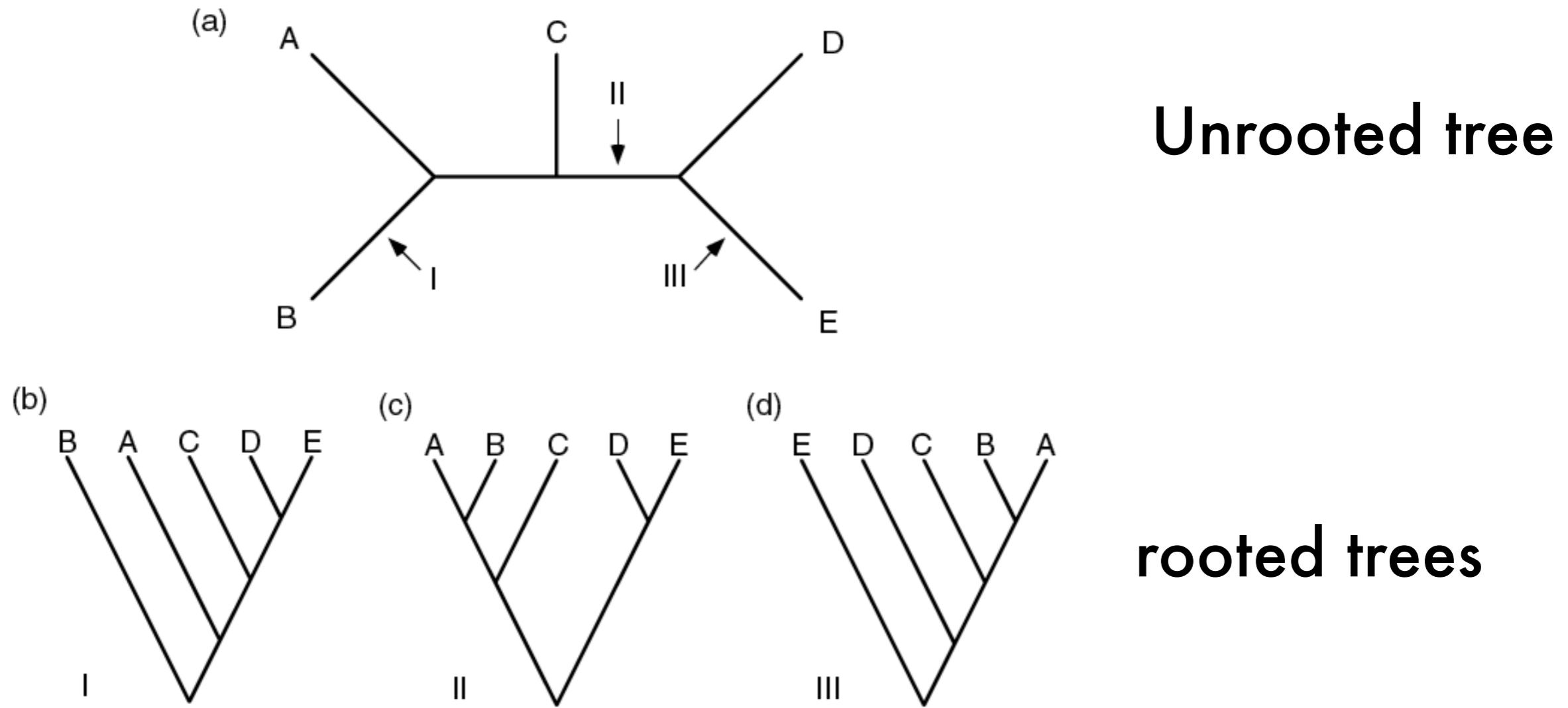
Cladograms show  
branching order -  
branch lengths are  
meaningless



Phylogenograms show  
branch order and  
branch lengths

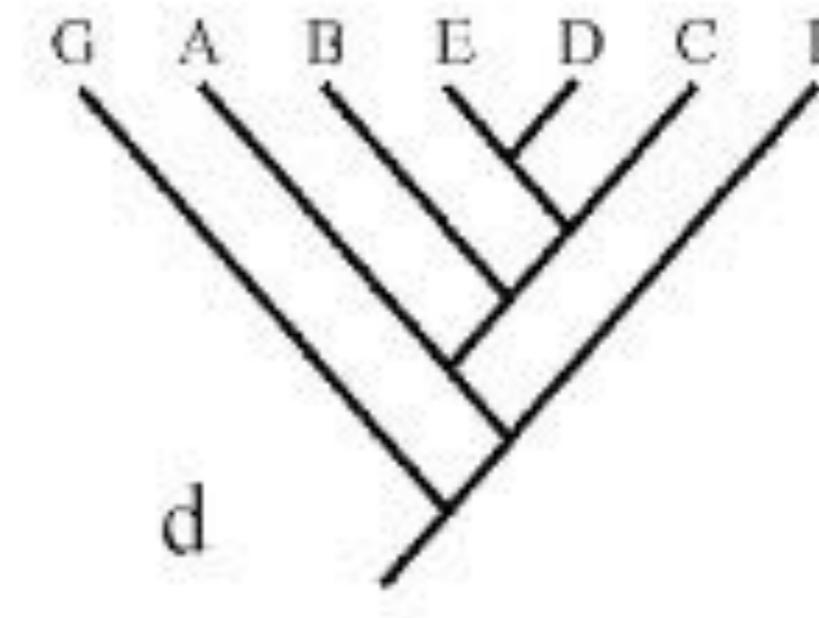
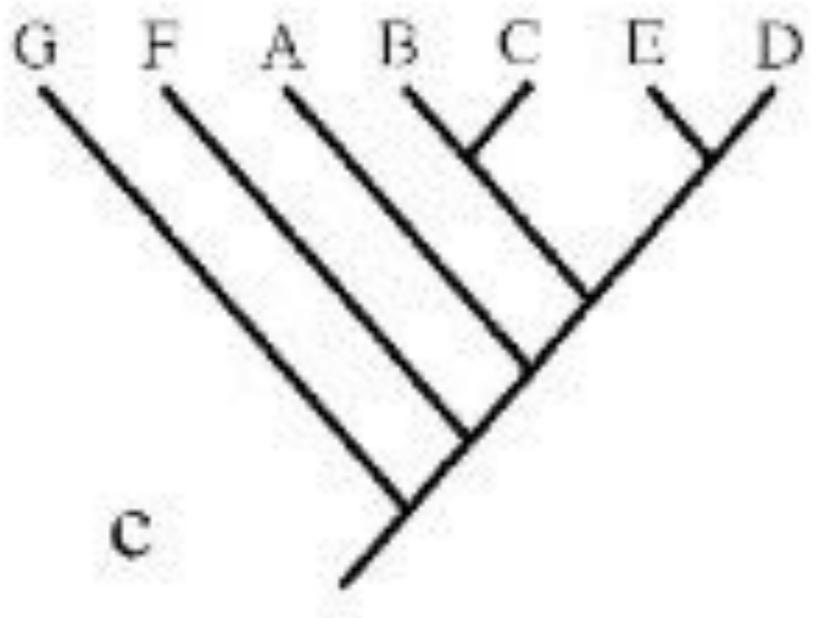
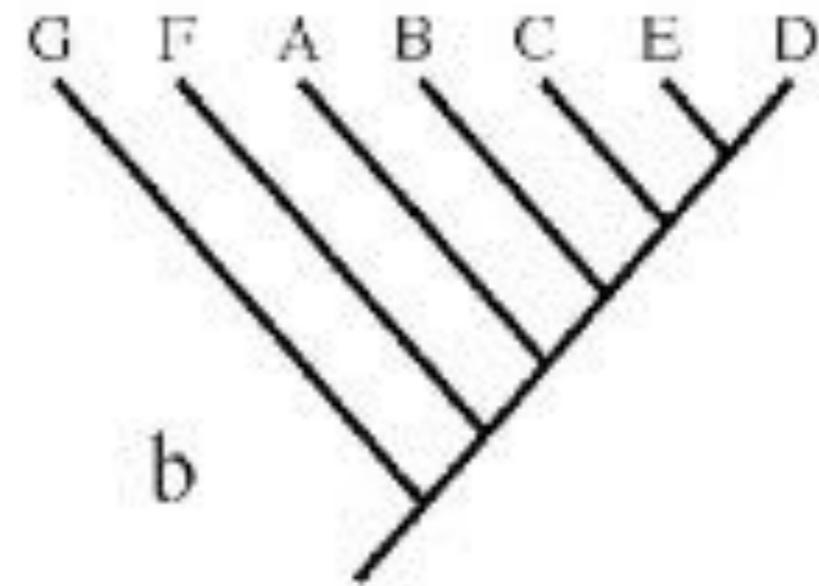
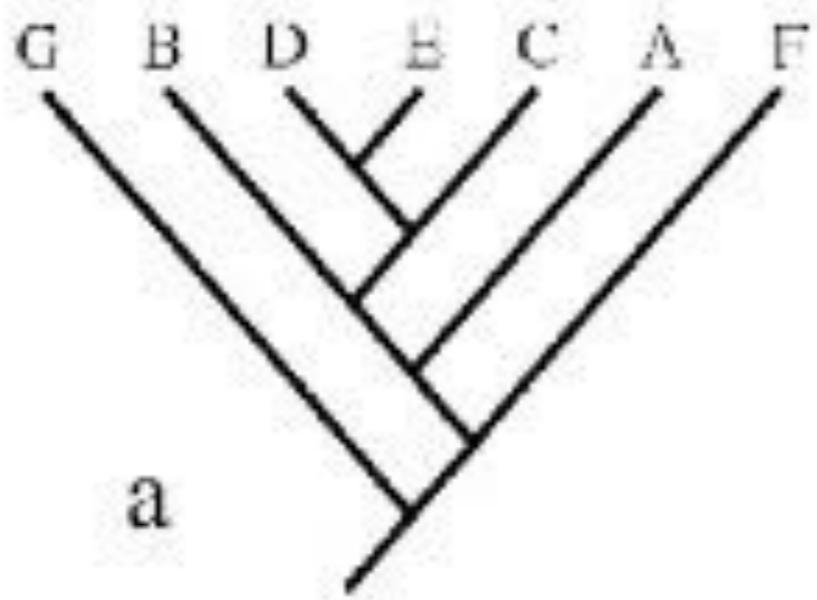
*Branch length is directly proportional to the amount of change*

# Rooting using an outgroup



Unrooted: “direction” of evolution unknown, or not inferred

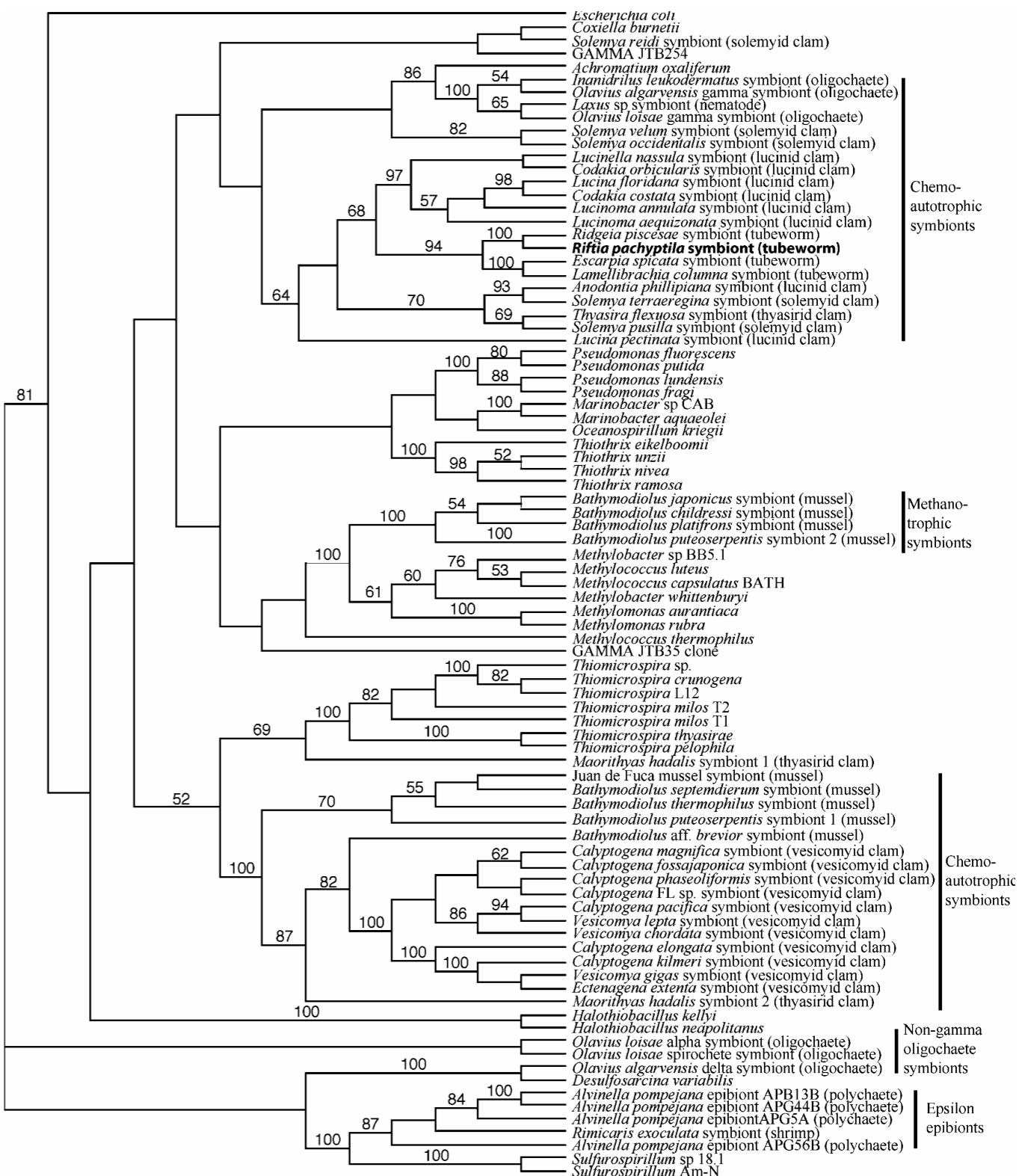
Rooted: “direction” of evolution known, or inferred

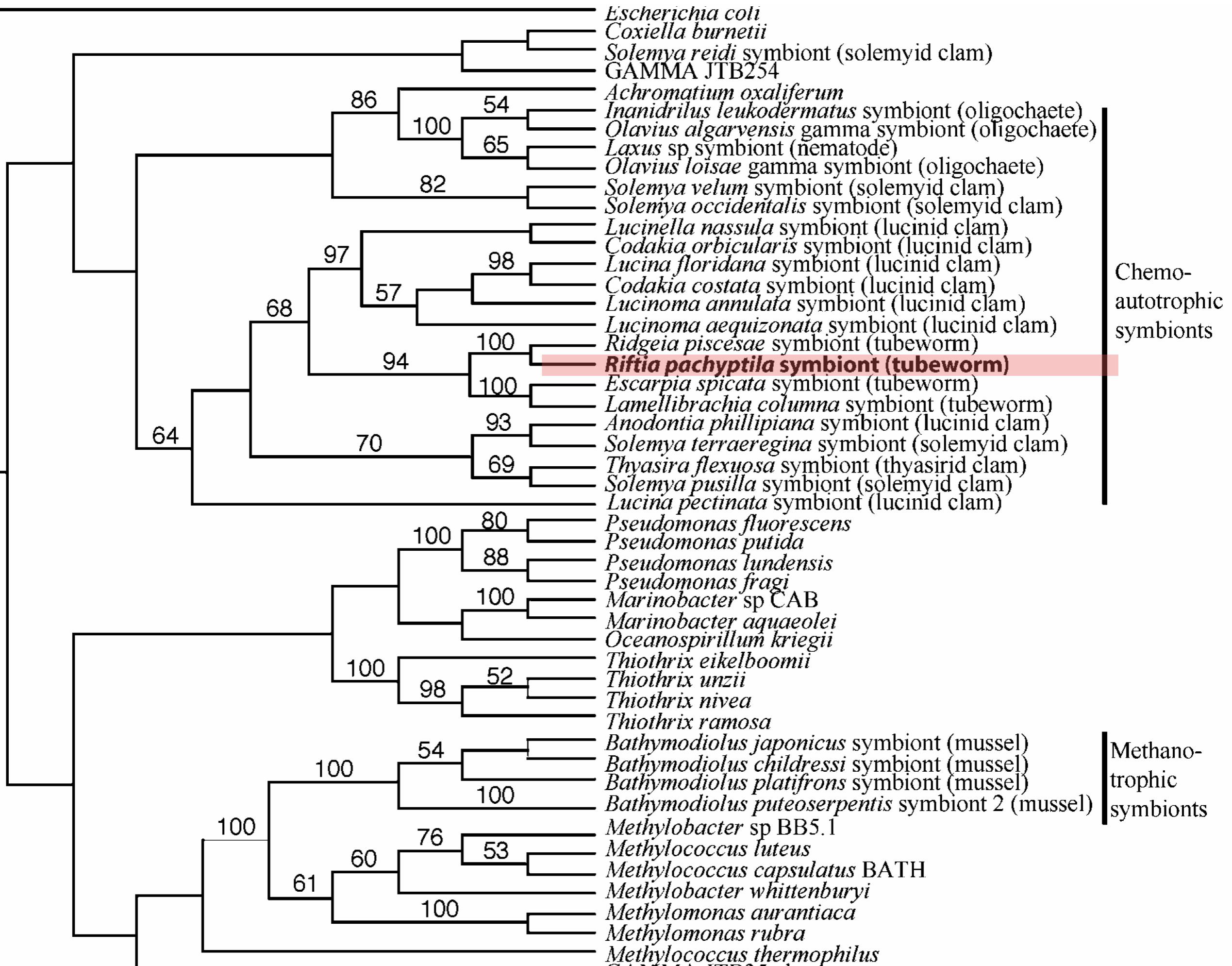


Which of the four trees above depicts a different pattern?

# How do we evaluate the quality of trees?

What parts of the tree are we confident in?





# \* Non-parametric bootstrapping

- Analyze original dataset
- Generate new samples by resampling original data ( $n = \# \text{ columns}$ ,  $n$  times)
- Analyze new datasets like original
- Summarize variation ("consensus")
- Bootstrap values = variation in tree topologies (generally want  $>50$ )

\*

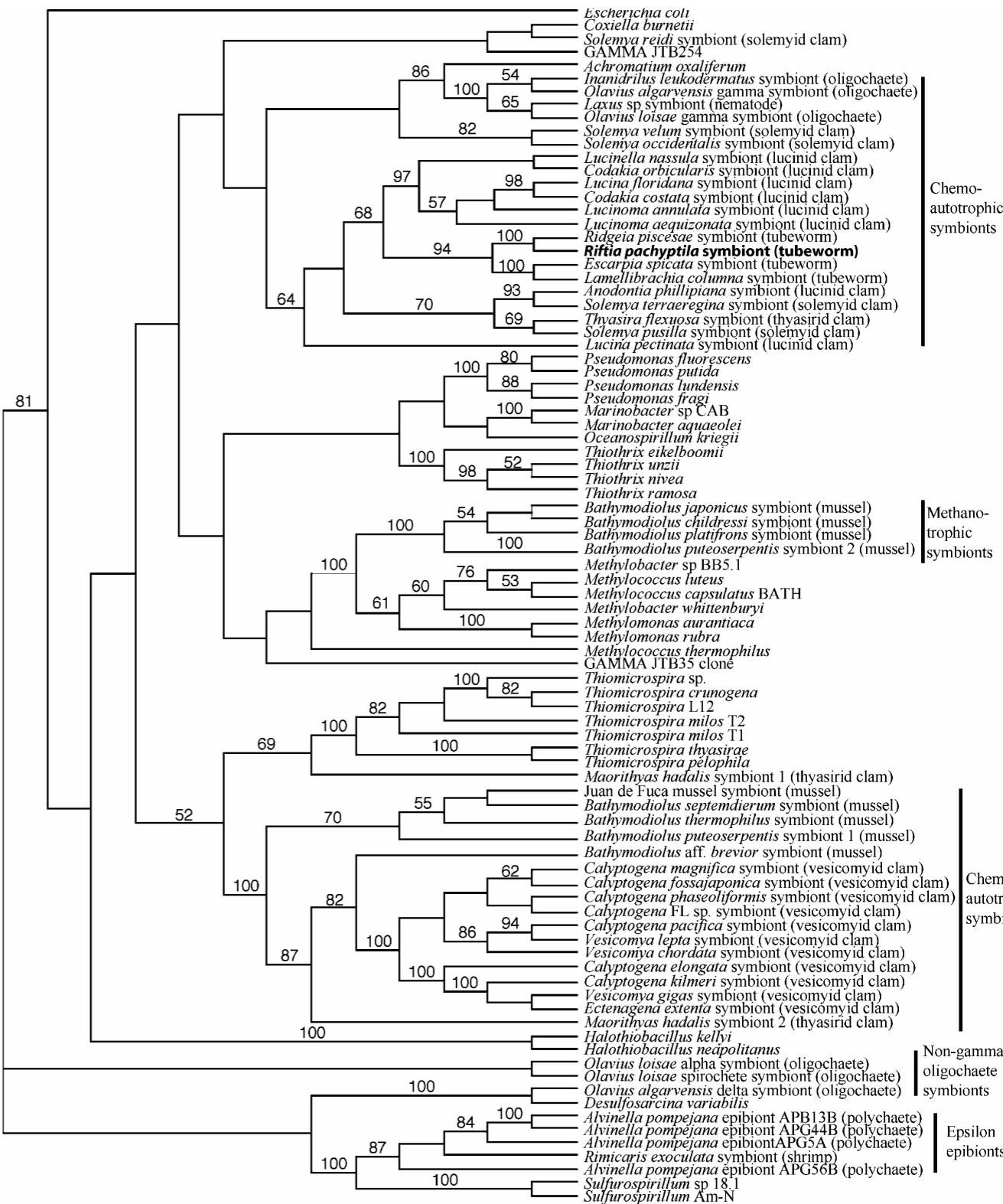
# Bootstrapping

- The number of times a particular branch is formed in the tree

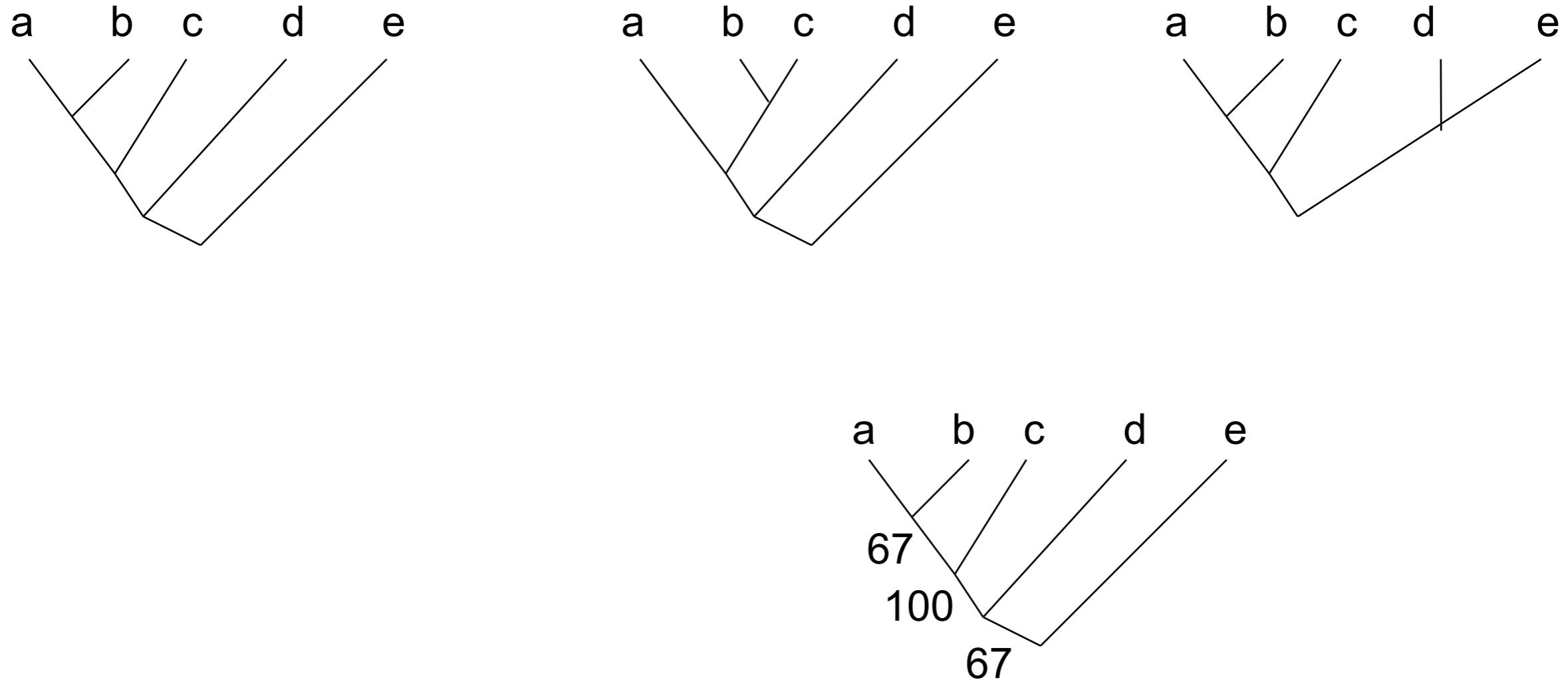
(out of the X times the analysis is done).

- used to estimate probability and indicated on a consensus tree

Alignment and evolutionary assumptions are critical



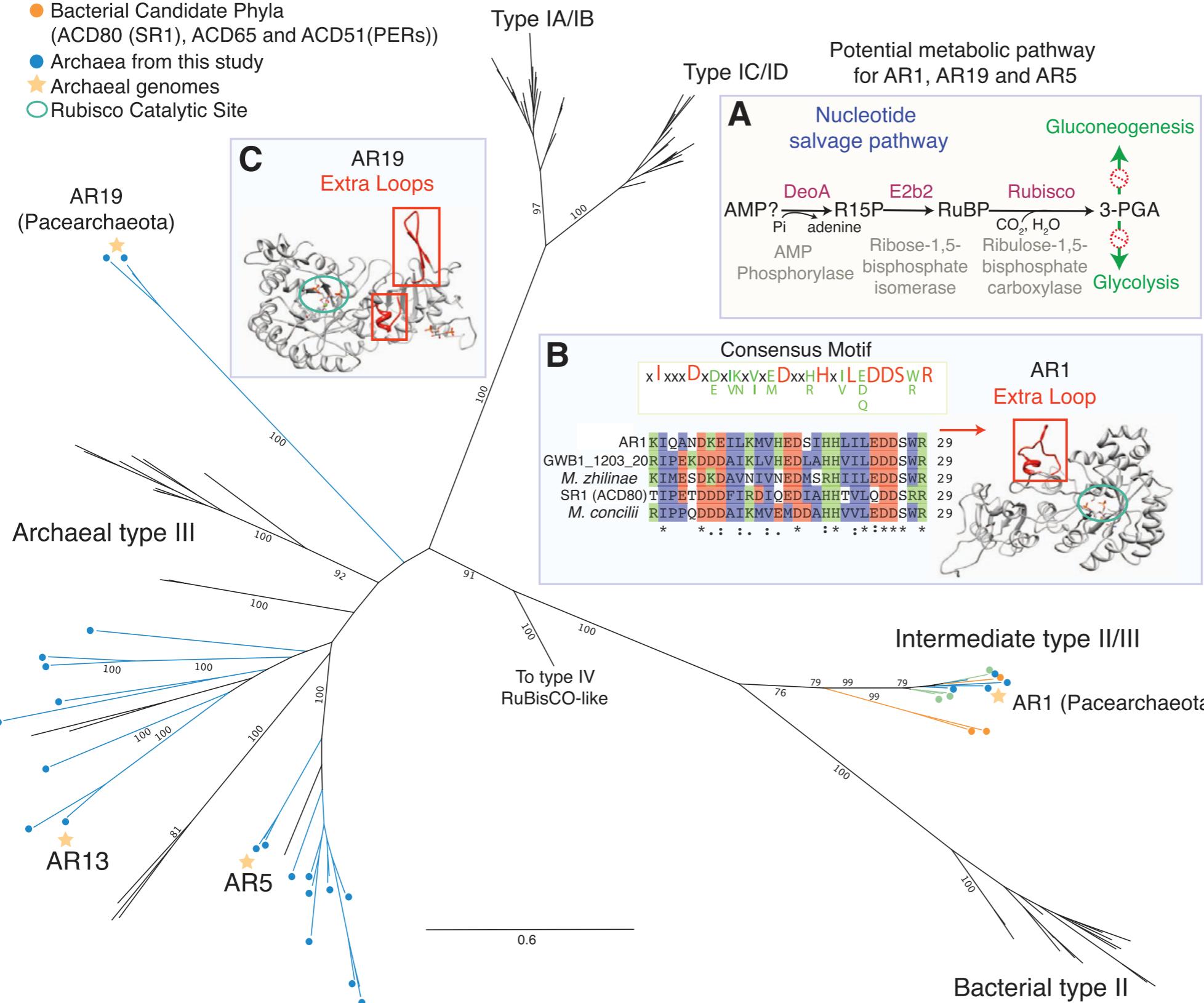
# Majority rule Consensus

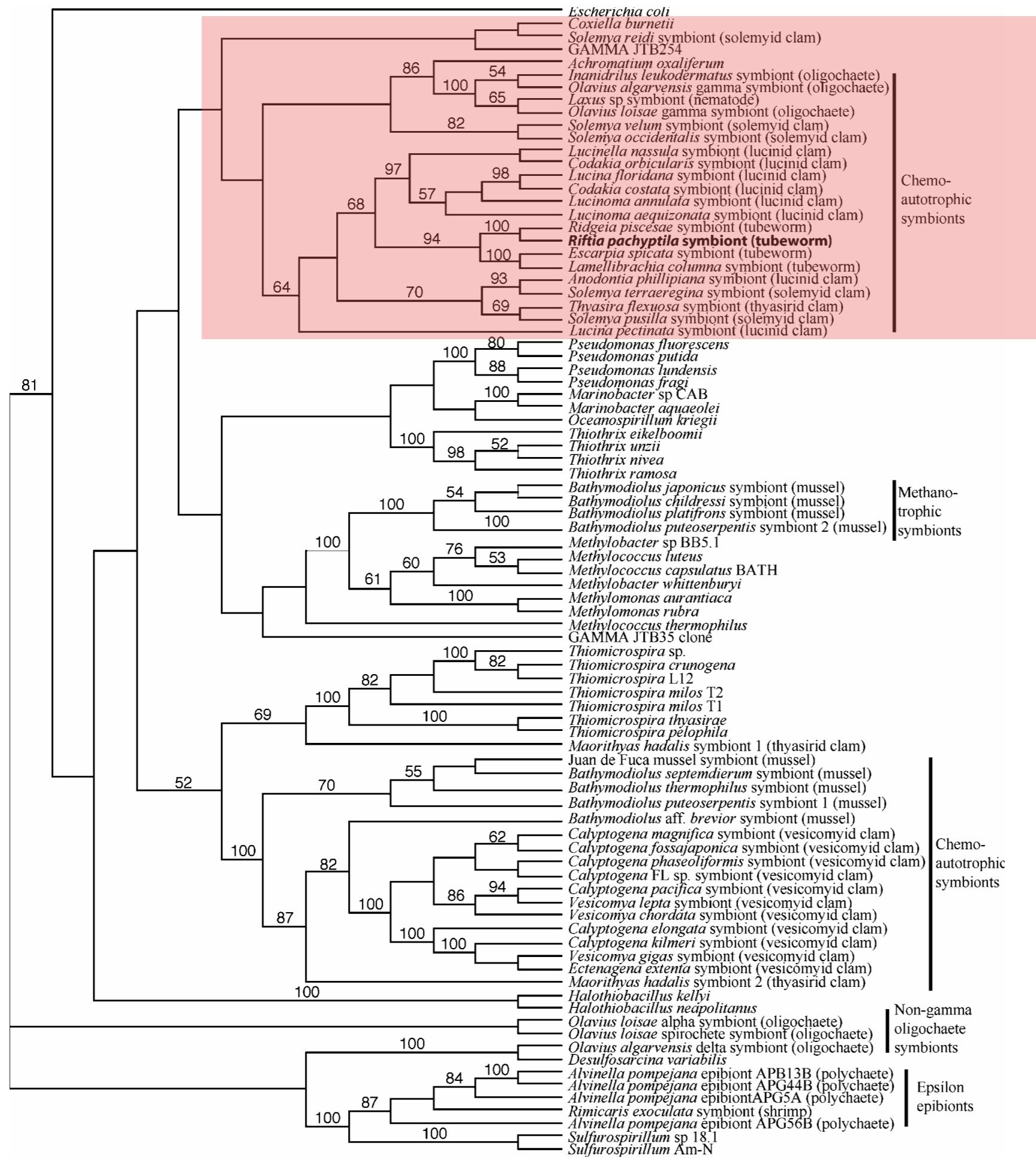


The percentage of the trees supporting each splits are indicated

# Large subunit RuBisCo phylogeny

- Methanoscinciales (Euryarchaeota)
- Bacterial Candidate Phyla (ACD80 (SR1), ACD65 and ACD51(PERs))
- Archaea from this study
- Archaeal genomes
- Rubisco Catalytic Site





# \* Monophyly, Paraphyly, and Polyphyly

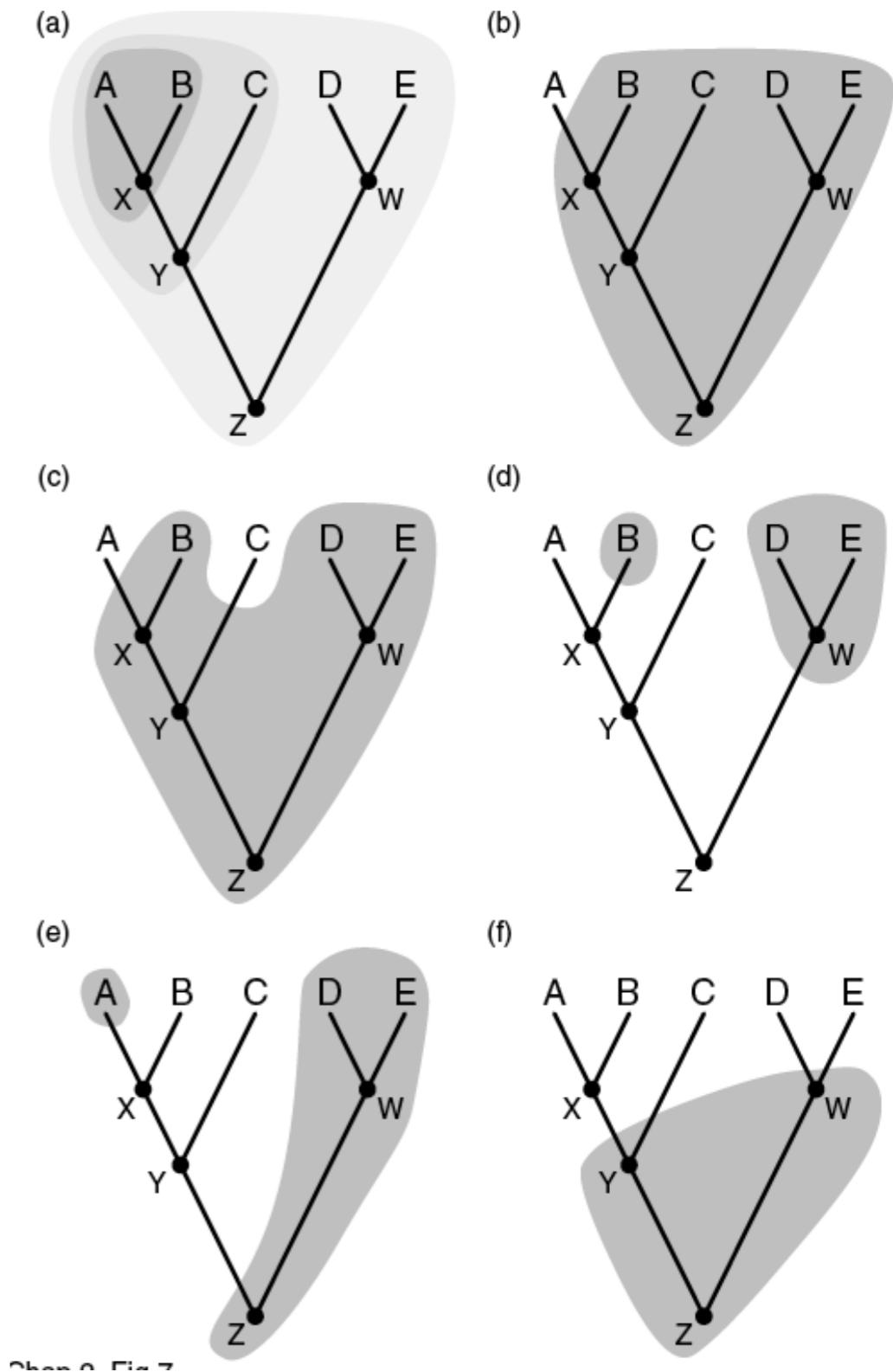
“natural” groups = monophyletic

1. monophyletic: ancestor + all descendants (a.k.a. “clades”)

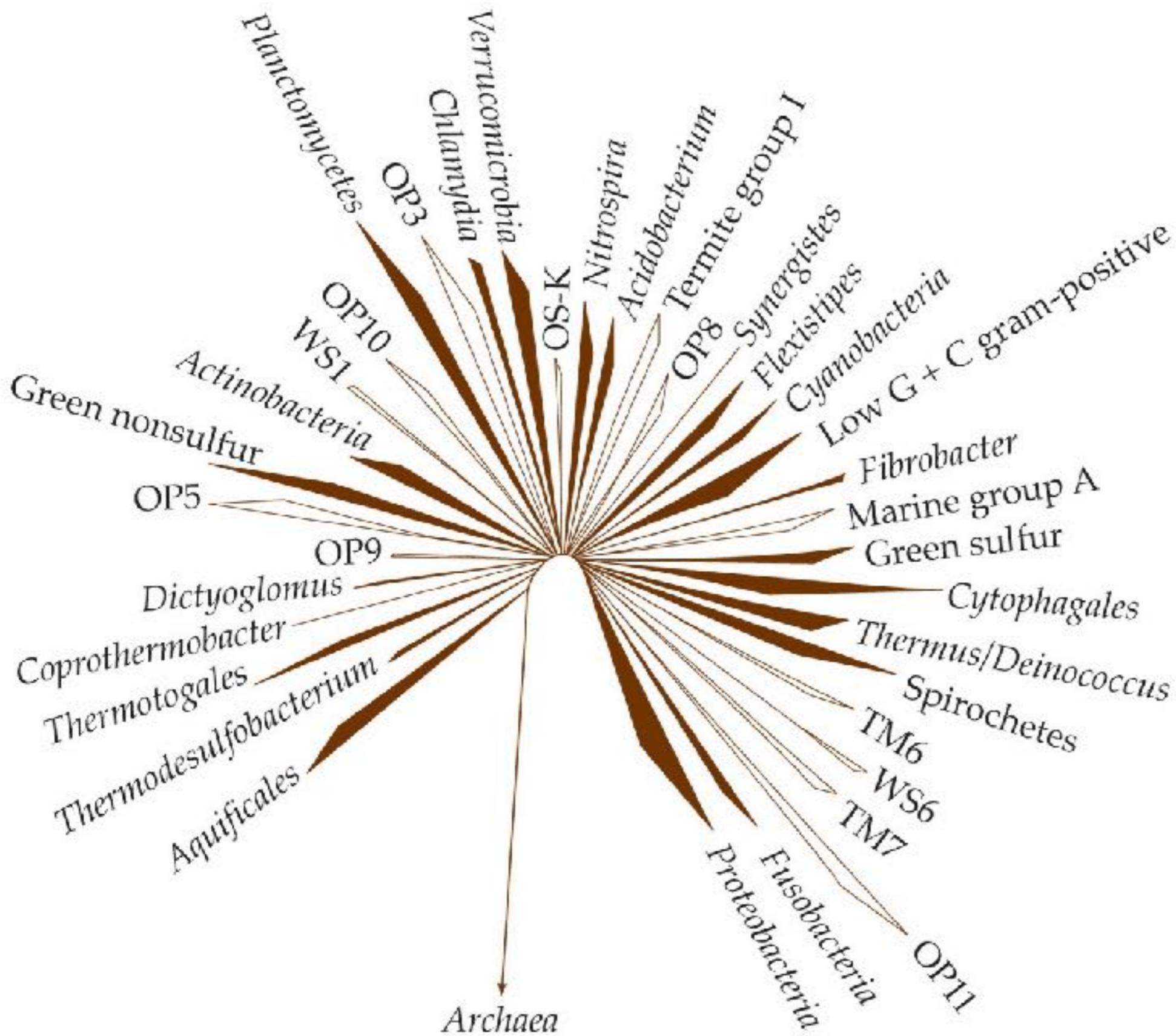
2. polyphyletic: multiple ancestries

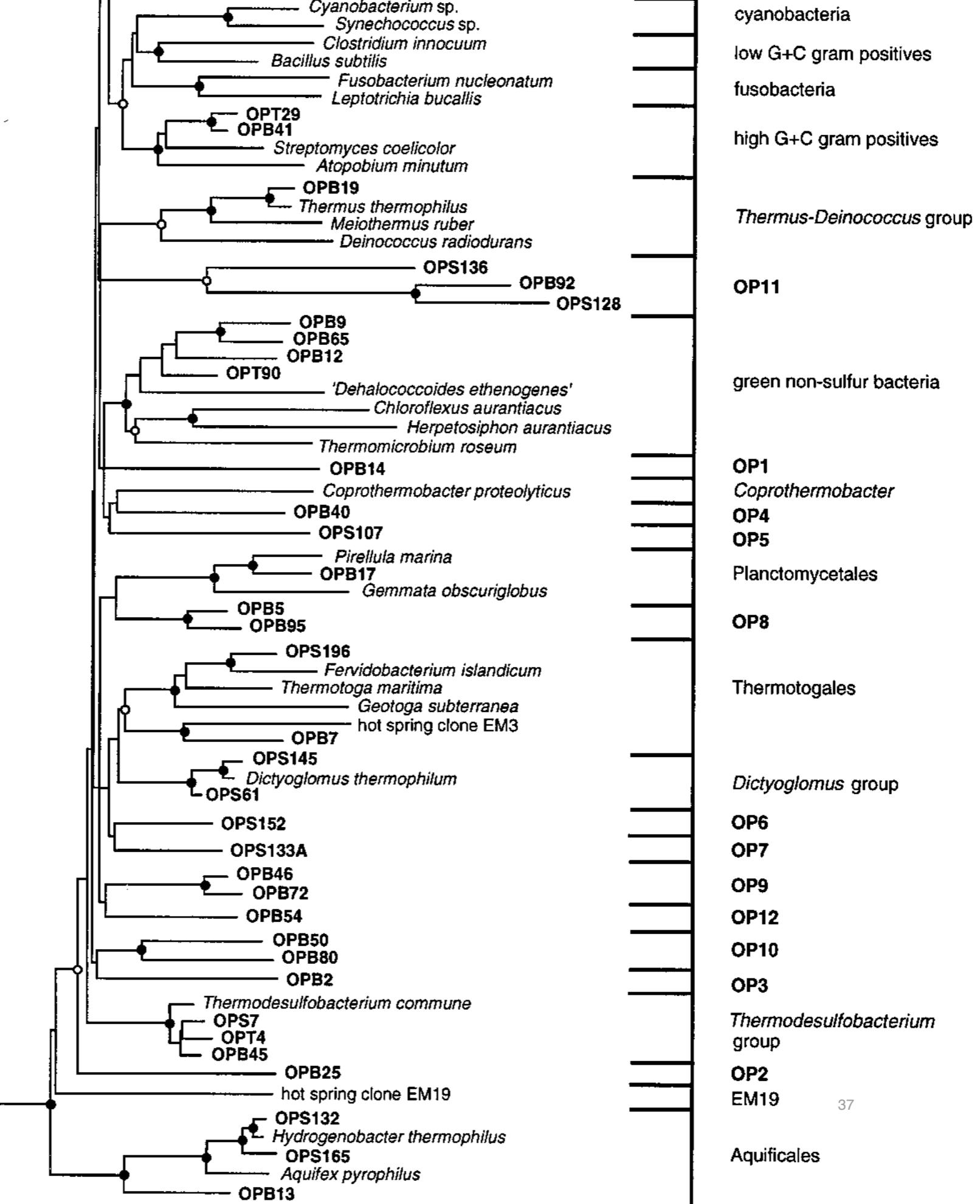
3. paraphyletic: ancestor but not all descendants

What about dark groups to the right?



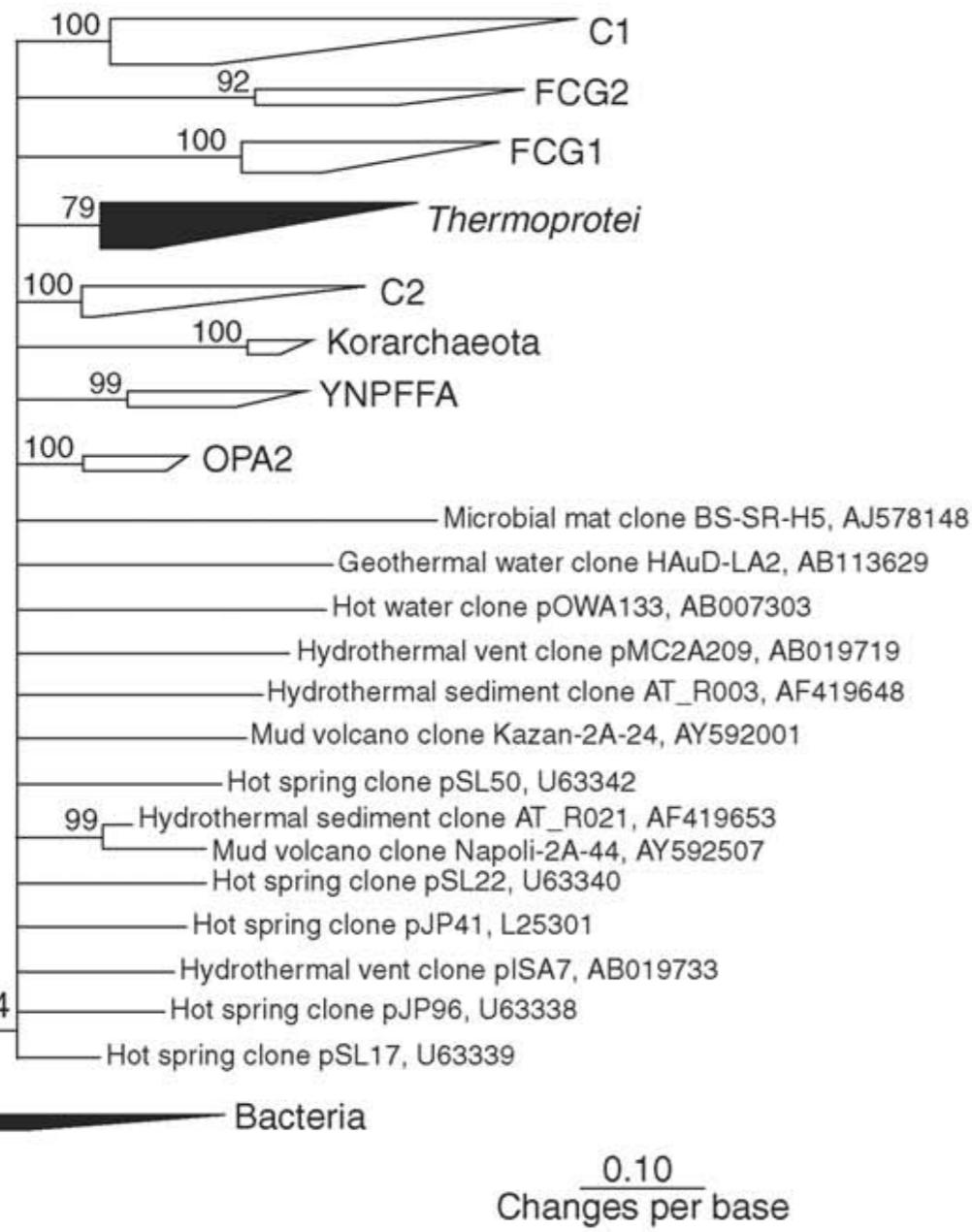
# Bacterial divisions



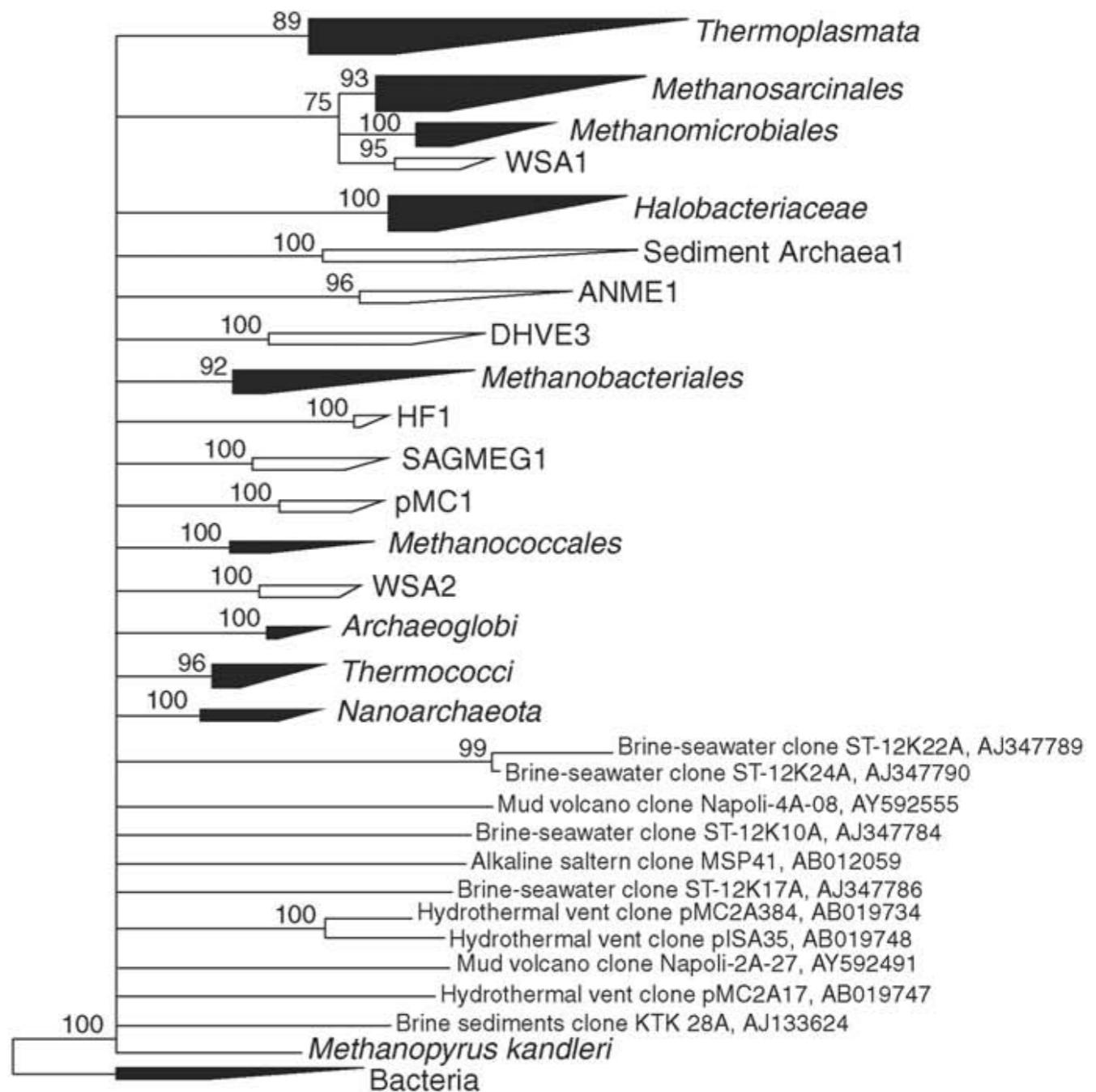


# Archaeal Diversity (ssu rRNA)

**(a) Crenarchaeota**



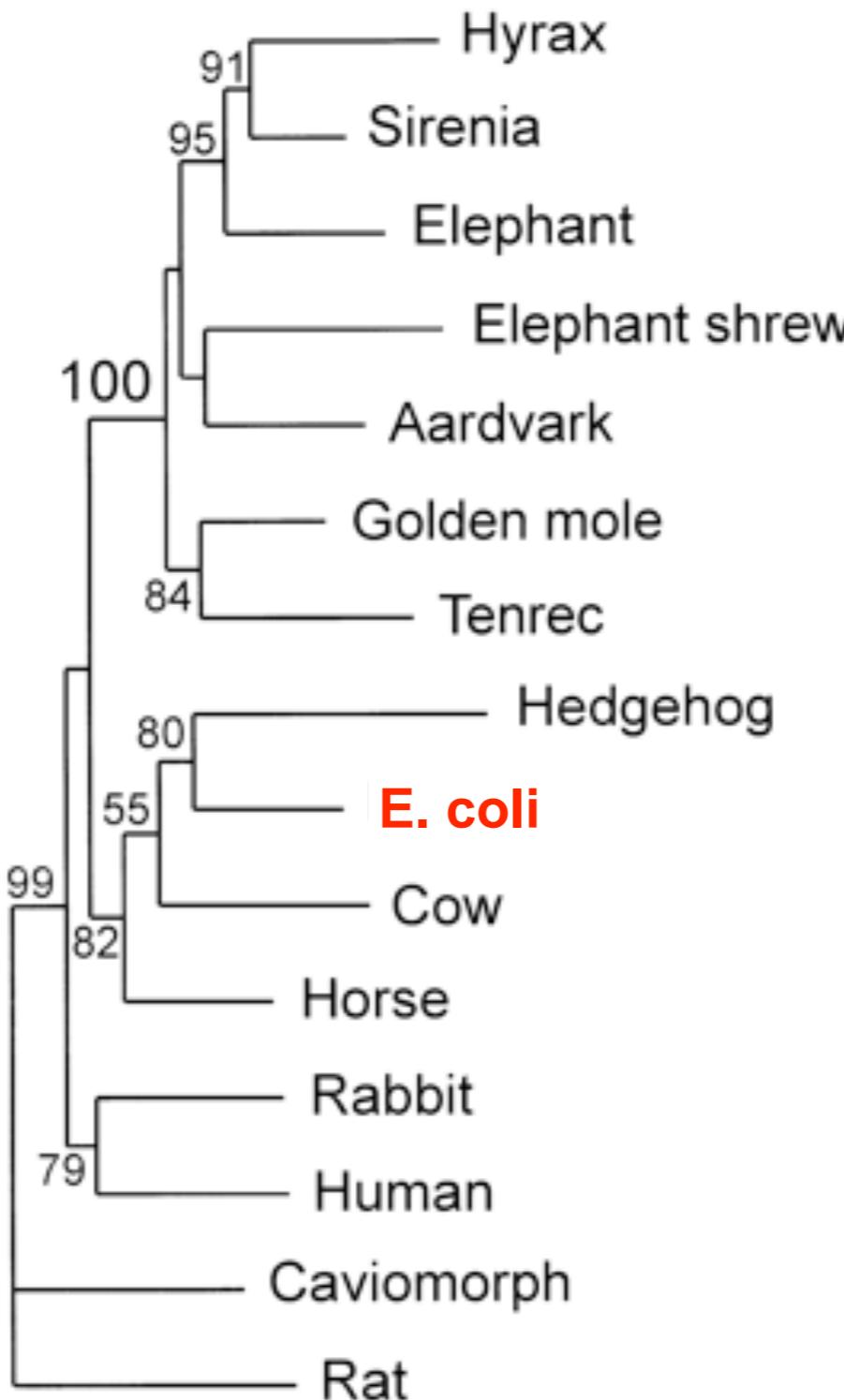
**(b) Euryarchaeota**



# Horizontal Gene Transfer?

A close-up photograph of a tree trunk, likely a eucalyptus, showing its rough, textured bark and numerous white, fibrous root structures (prop roots) growing downwards. The scene is dimly lit, with dappled sunlight filtering through leaves, creating a natural, organic feel.

# Xenologs: Horizontal gene transfer



Horizontal gene transfer: genes acquired not through common “vertical” ancestry

\*

## Monophyletic tree: Limited lateral transfer

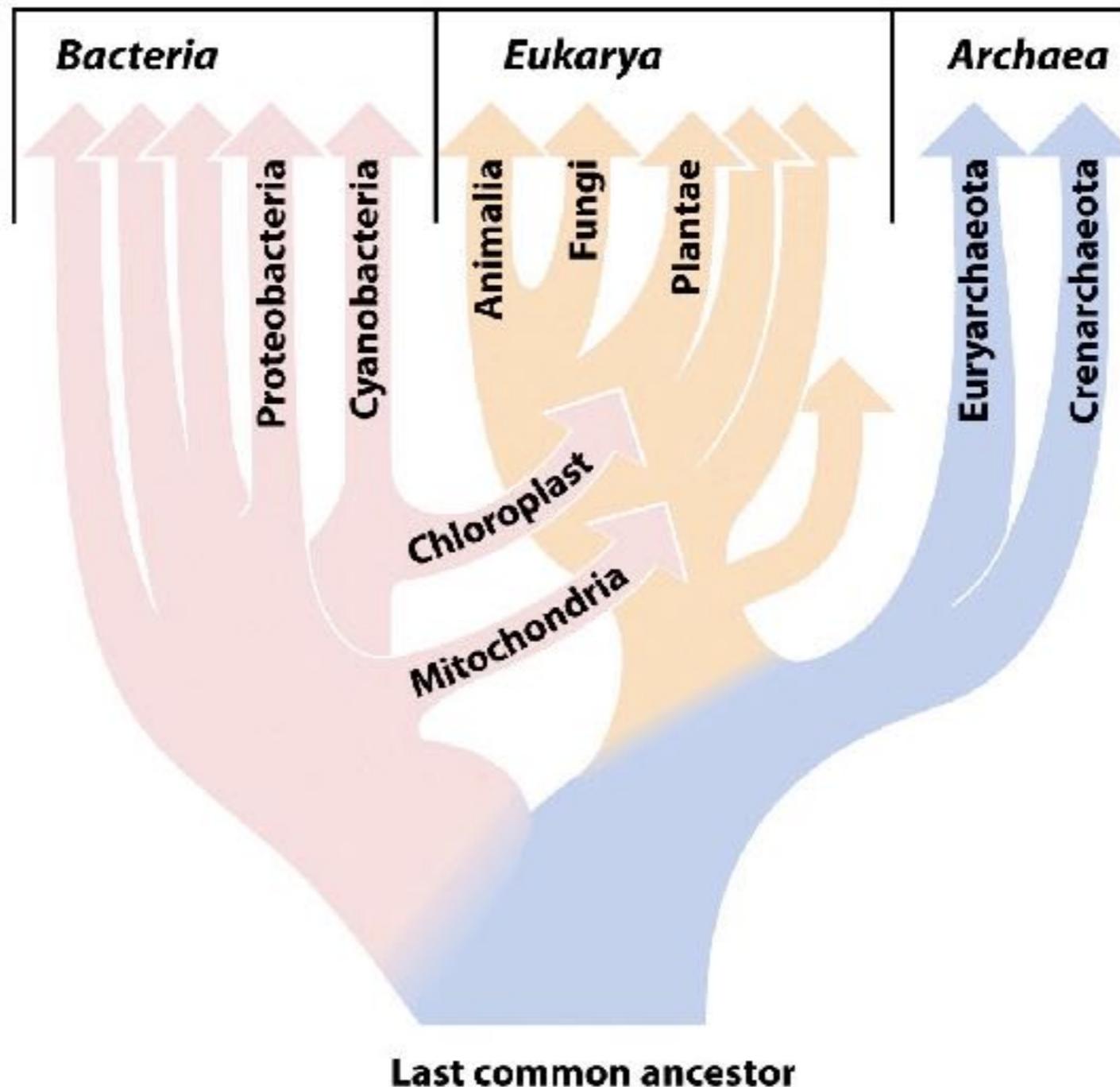
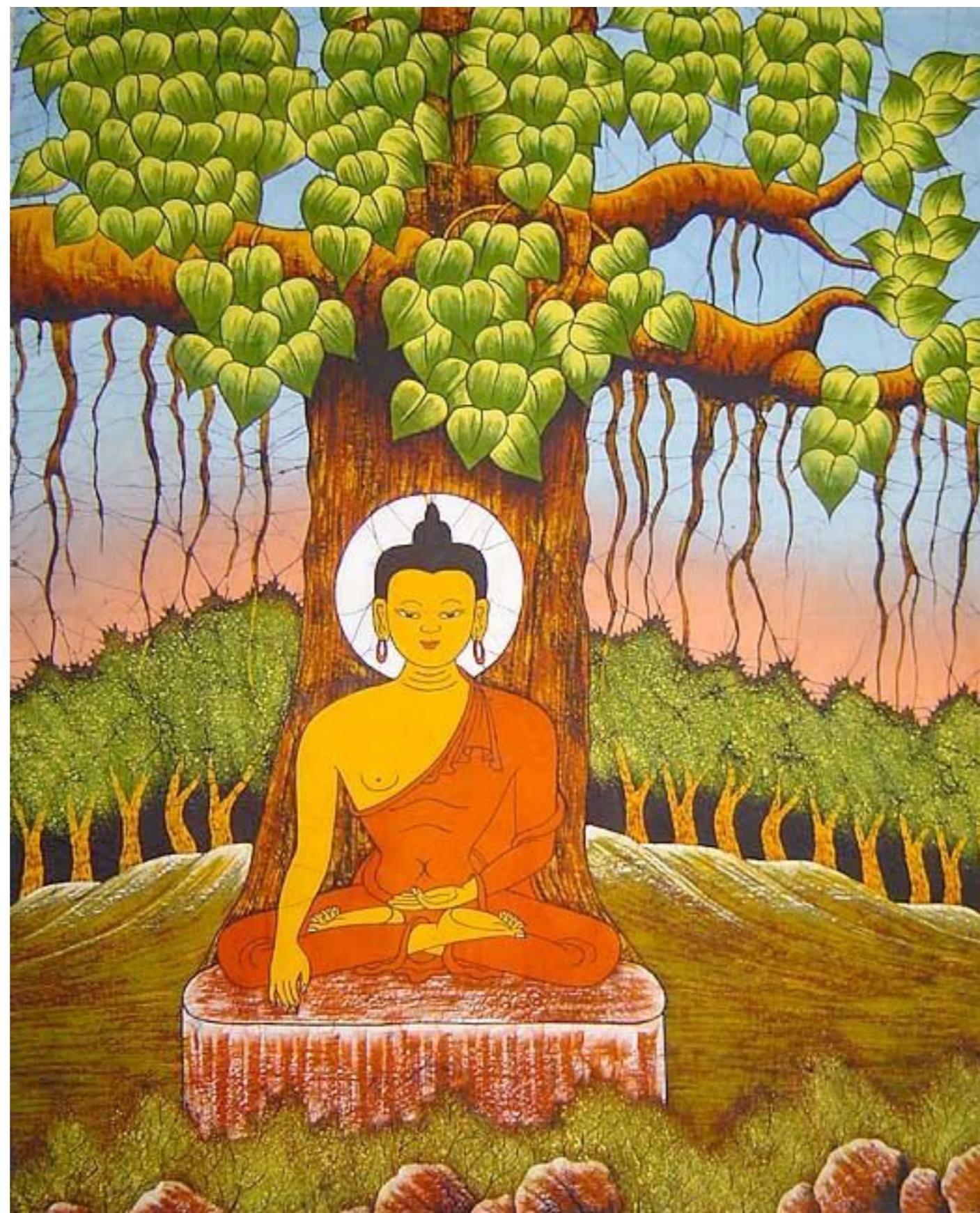


Figure 17.27a Microbiology: An Evolving Science  
Source: W. Ford Doolittle. 1999. *Science* 284:2126.

Alpha proteobacterial genes in eukaryotes?  
Cyanobacterial genes in eukaryotes?

# What's the sound of ten species diverging?

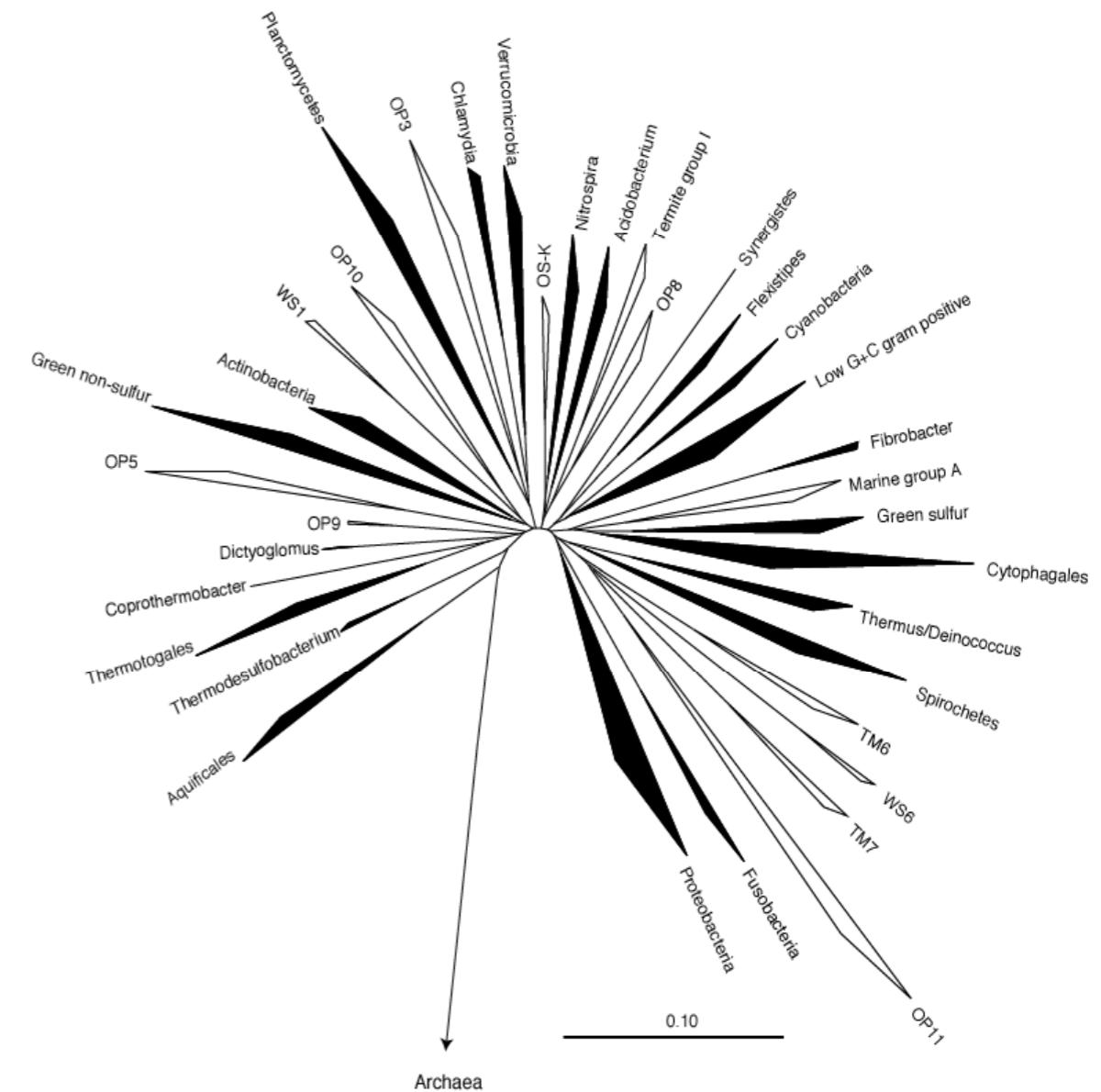


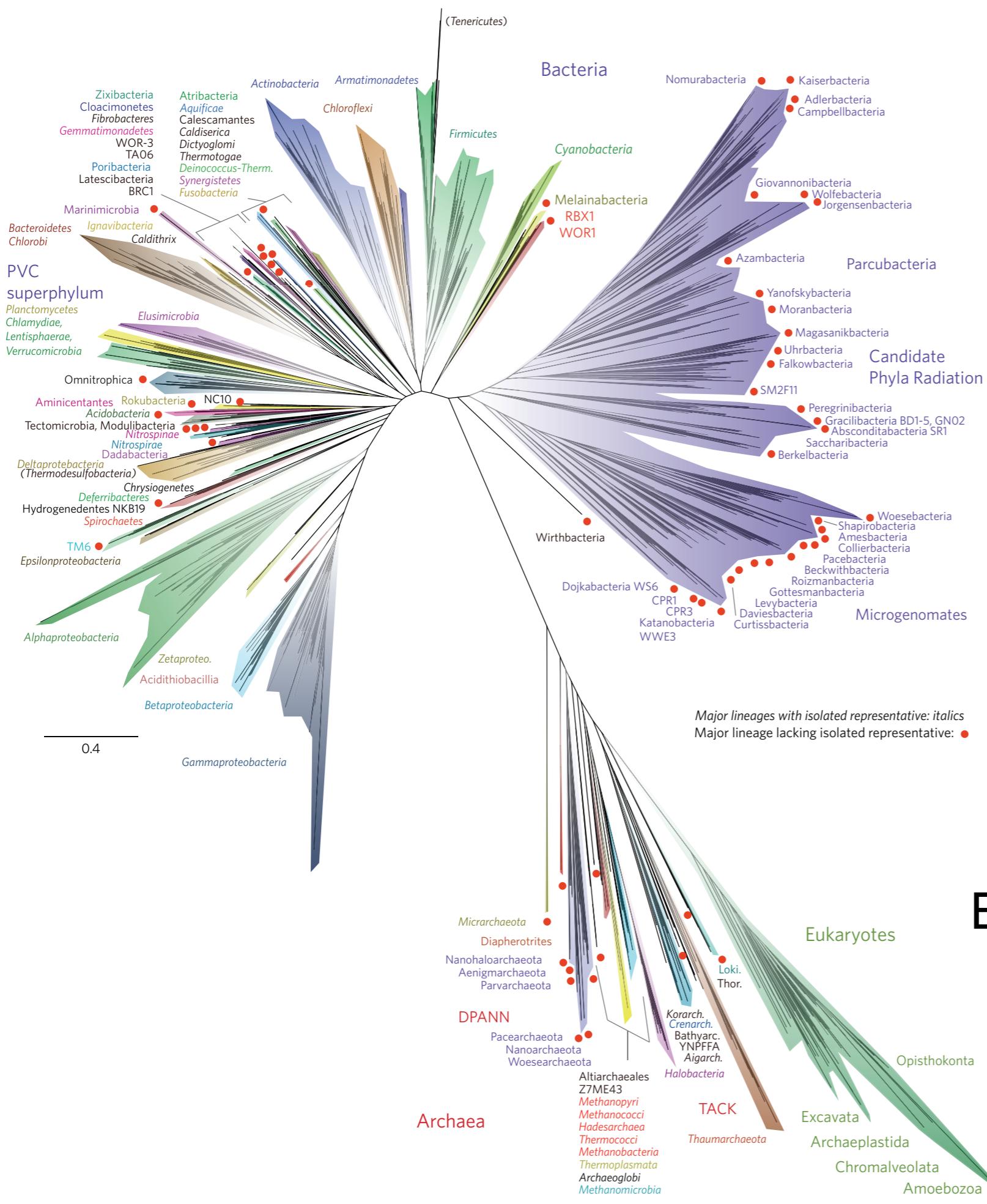
# Polytomy:

unresolved branching order, meaning either:

“unresolvable” using given phylogenetic methods (e.g., low bootstrap values) or multiple species diverged simultaneously

Polytomies are usually represented as more than two branches coming from a single node.

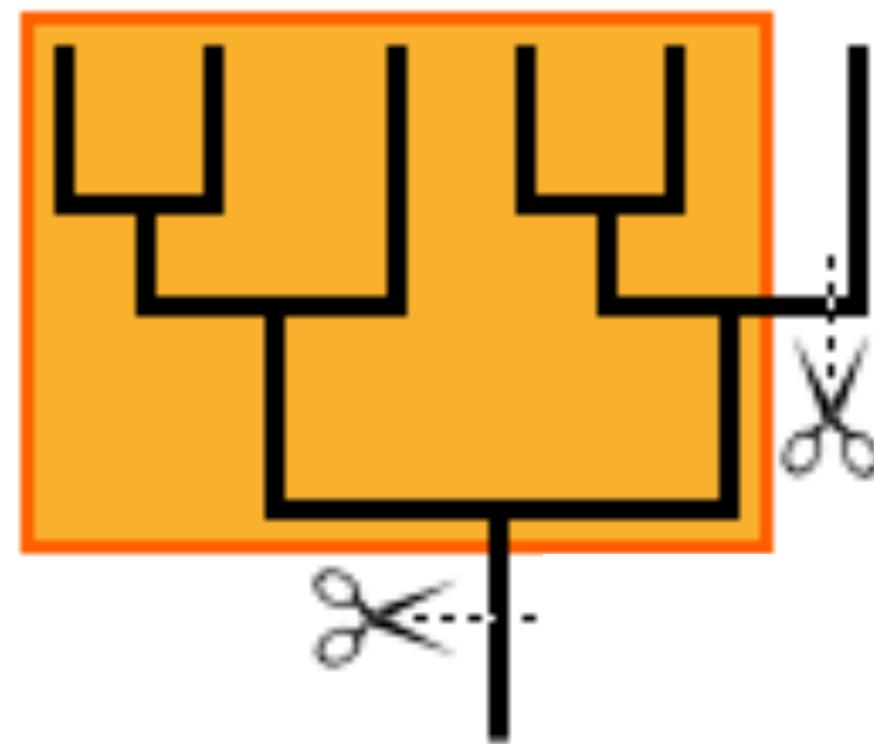
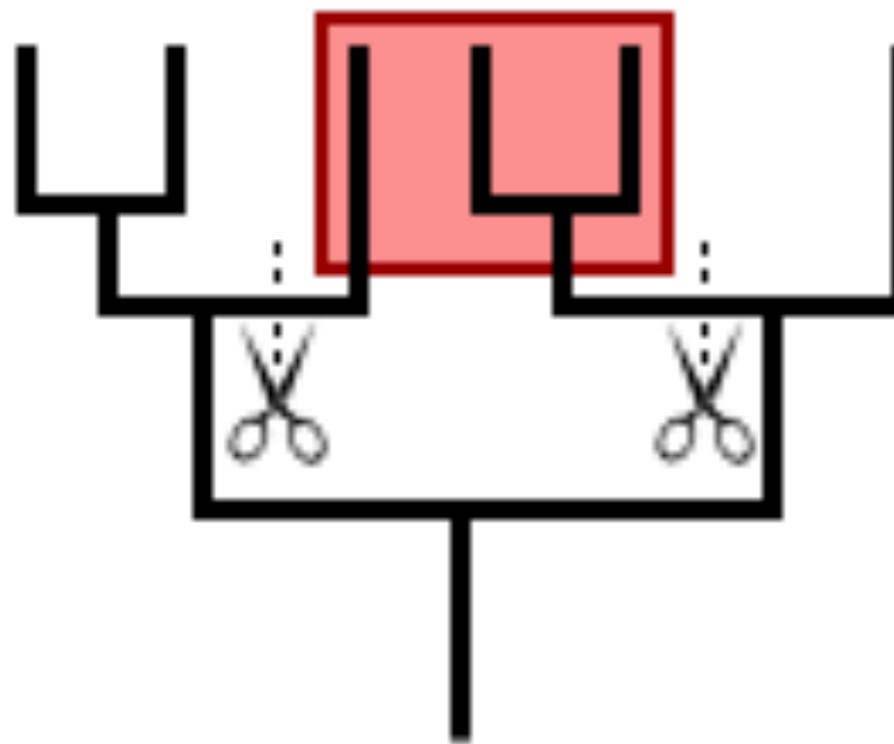
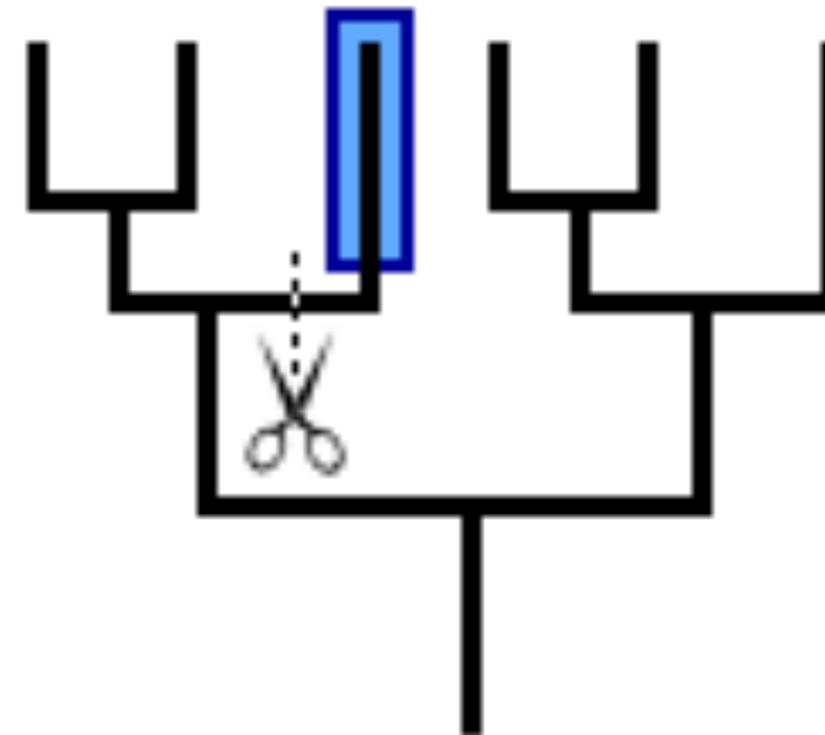
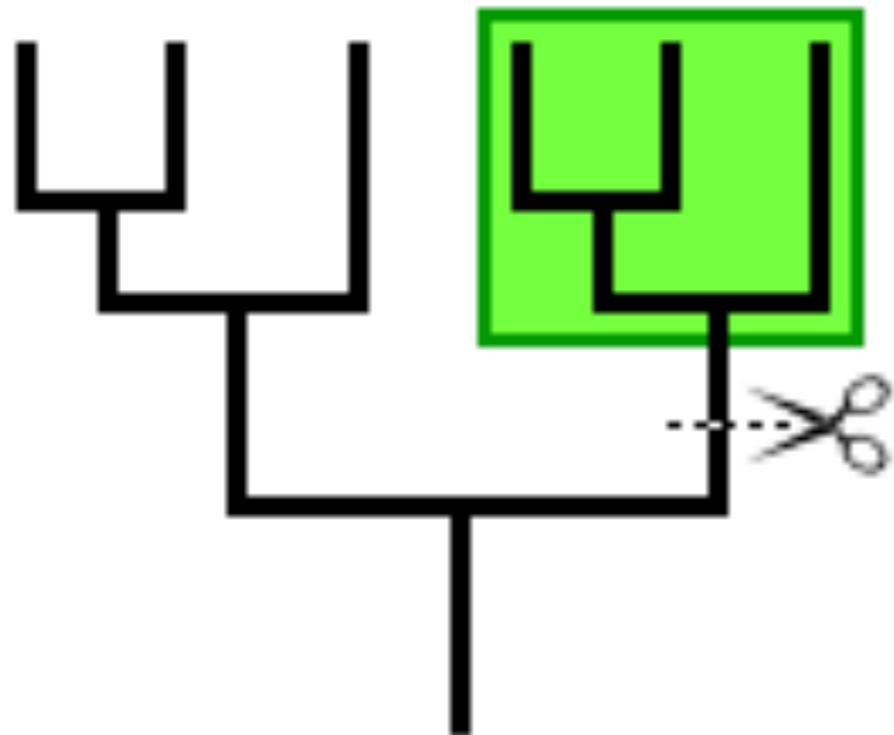




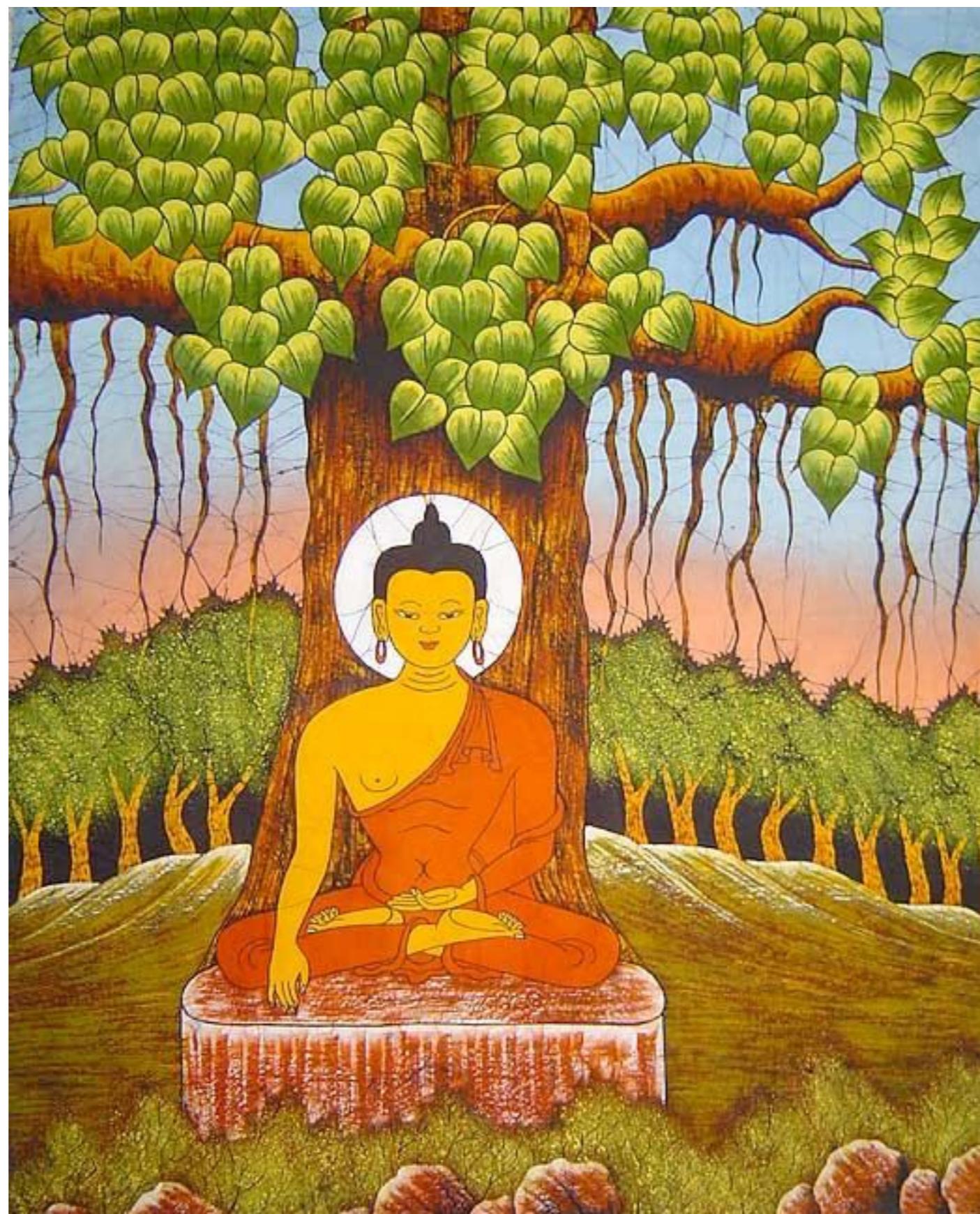
# What's the scale?

# Euks are archaea?

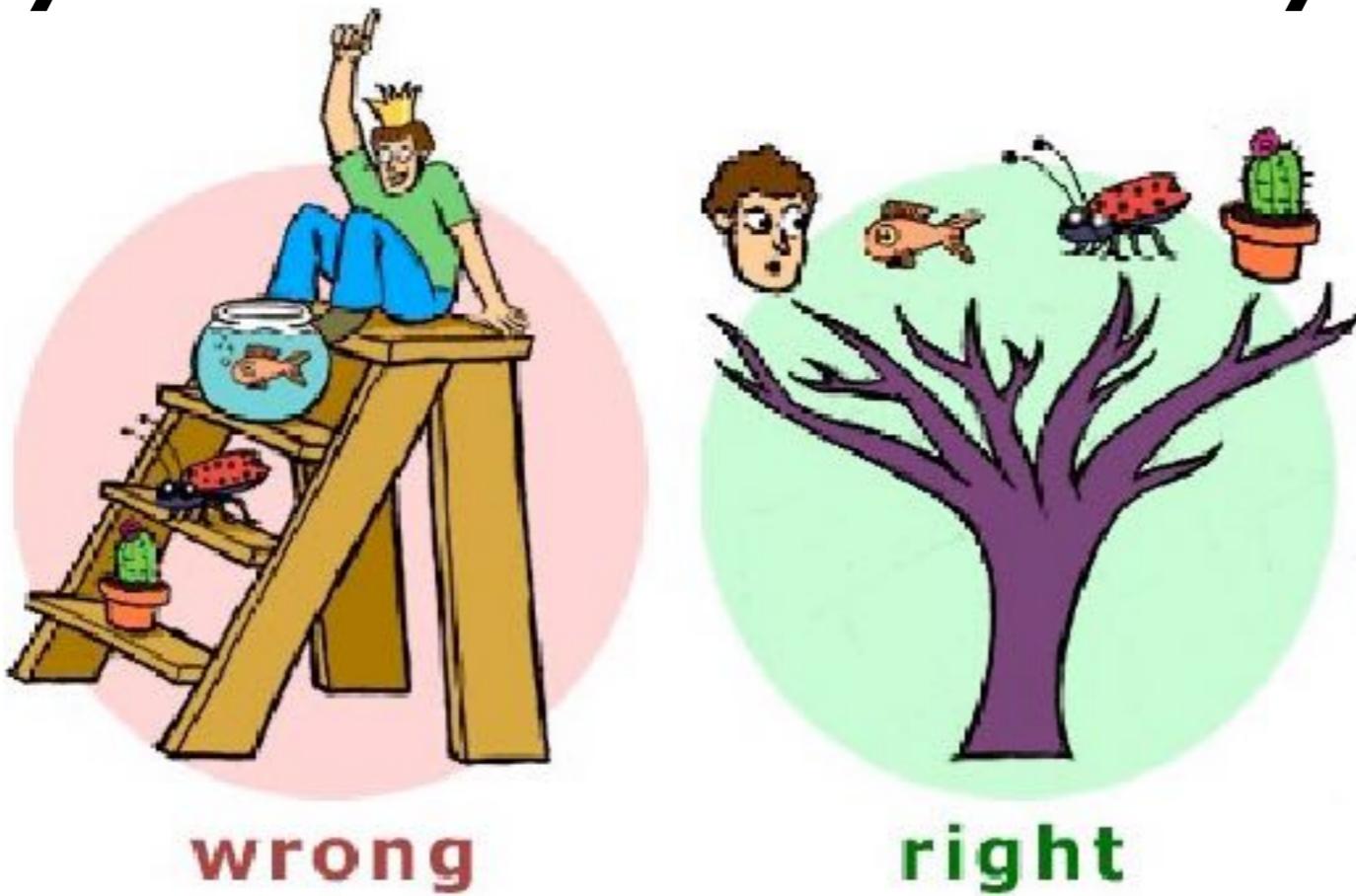
# “scissor test” for monophyly



# Does life really get better and better?

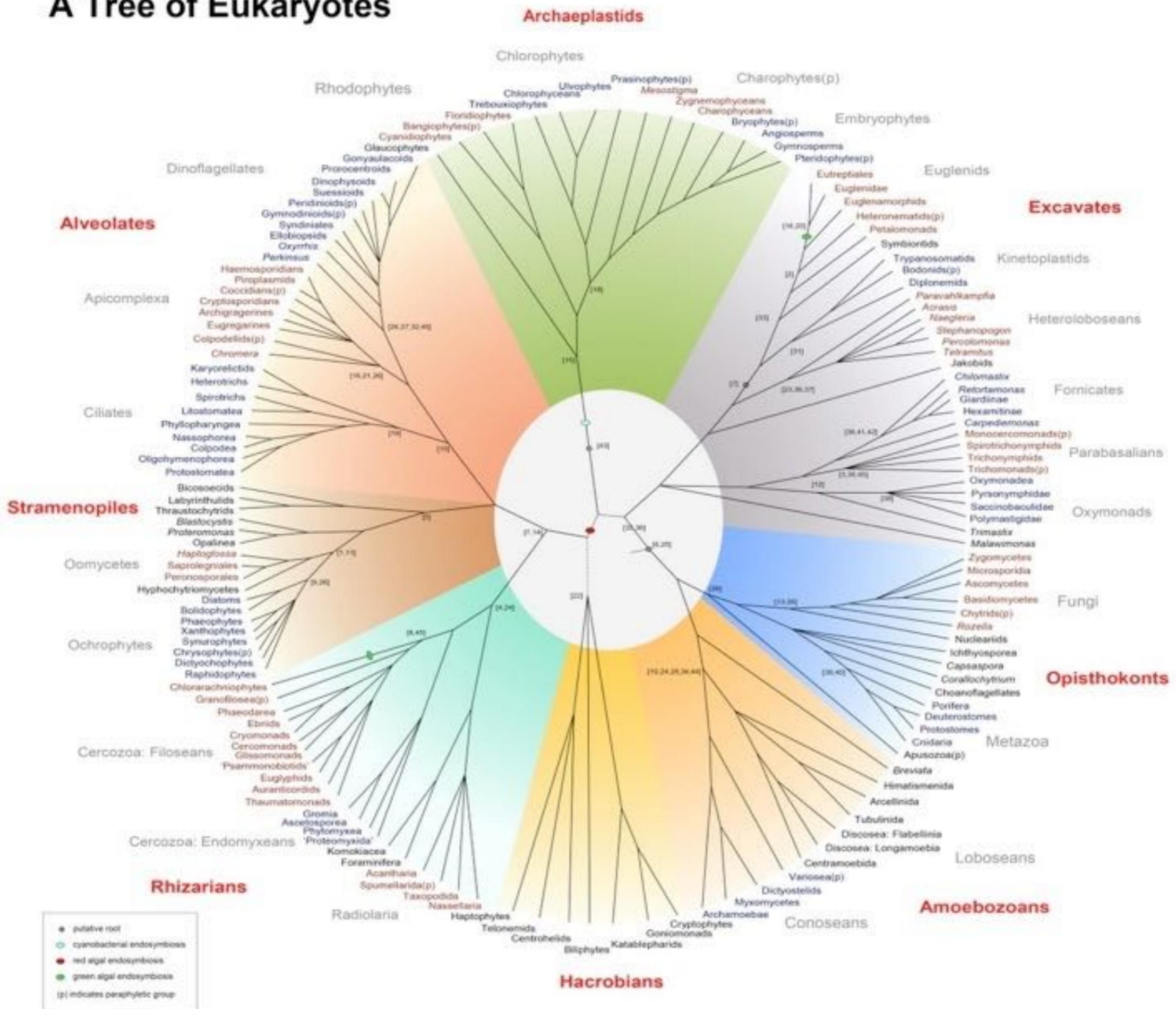


# The myth of the evolutionary ladder



- No “Higher” and “Lower” organisms
- Lineages evolve by radiation/diversification, not progression
- Populations of organisms are “selected” for life in their environment - not for increasing complexity
- No such thing as “primitive” or “missing link”
- “Simple” or “complex” are anthropocentric
- Prokaryote-> Eukaryote

# A Tree of Eukaryotes



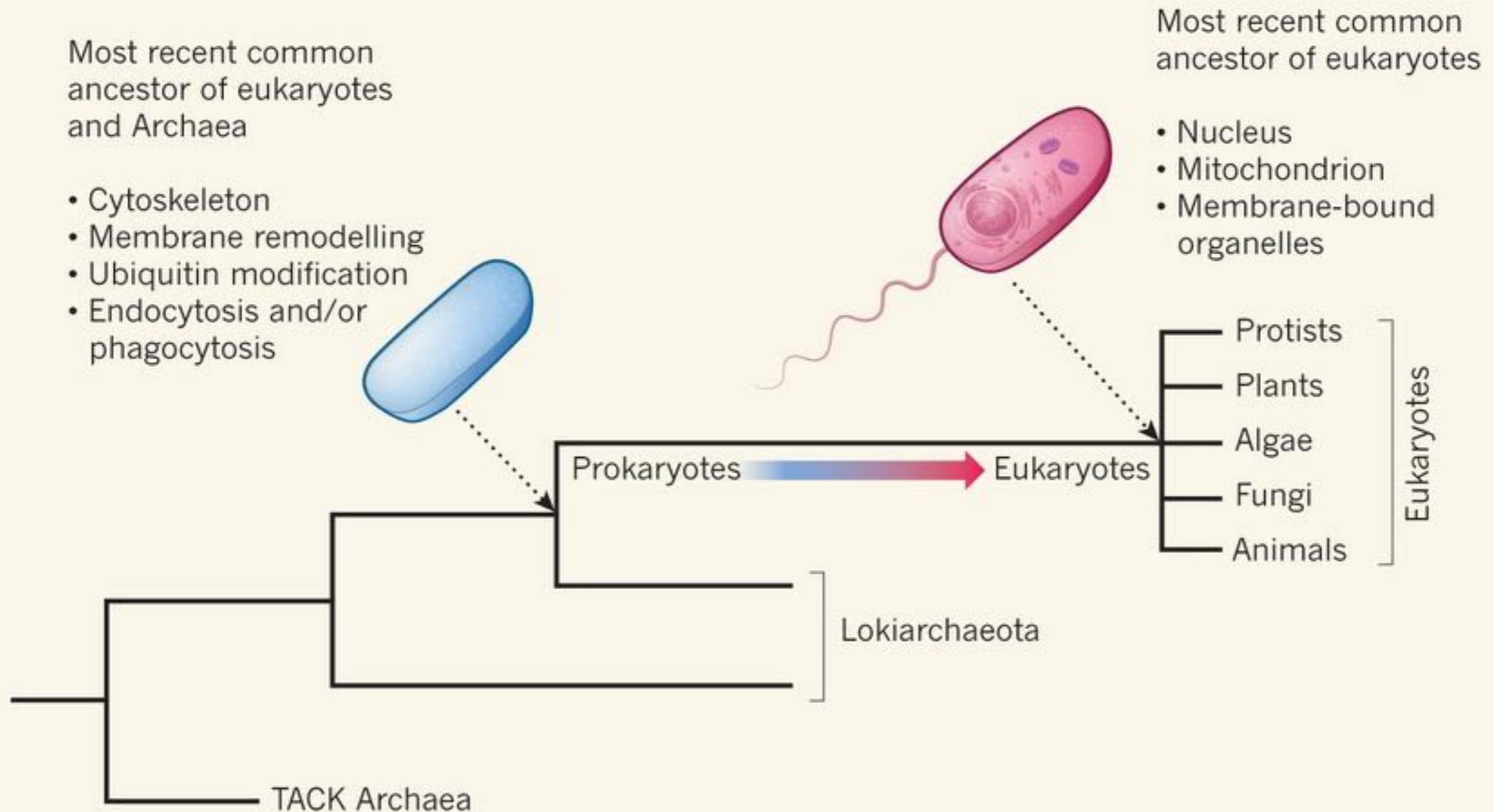
# Complex archaea that bridge the gap between prokaryotes and eukaryotes

Anja Spang<sup>1\*</sup>, Jimmy H. Saw<sup>1\*</sup>, Steffen L. Jørgensen<sup>2\*</sup>, Katarzyna Zaremba-Niedzwiedzka<sup>1\*</sup>, Joran Martijn<sup>1</sup>, Anders E. Lind<sup>1</sup>, Roel van Eijk<sup>1†</sup>, Christa Schleper<sup>2,3</sup>, Lionel Guy<sup>1,4</sup> & Thijs J. G. Ettema<sup>1</sup>

The origin of the eukaryotic cell remains one of the most contentious puzzles in modern biology. Recent studies have provided support for the emergence of the eukaryotic host cell from within the archaeal domain of life, but the identity and nature of the putative archaeal ancestor remain a subject of debate. Here we describe the discovery of ‘Lokiarchaeota’, a novel candidate archaeal phylum, which forms a monophyletic group with eukaryotes in phylogenomic analyses, and whose genomes encode an expanded repertoire of eukaryotic signature proteins that are suggestive of sophisticated membrane remodelling capabilities. Our results provide strong support for hypotheses in which the eukaryotic host evolved from a bona fide archaeon, and demonstrate that many components that underpin eukaryote-specific features were already present in that ancestor. This provided the host with a rich genomic ‘starter-kit’ to support the increase in the cellular and genomic complexity that is characteristic of eukaryotes.

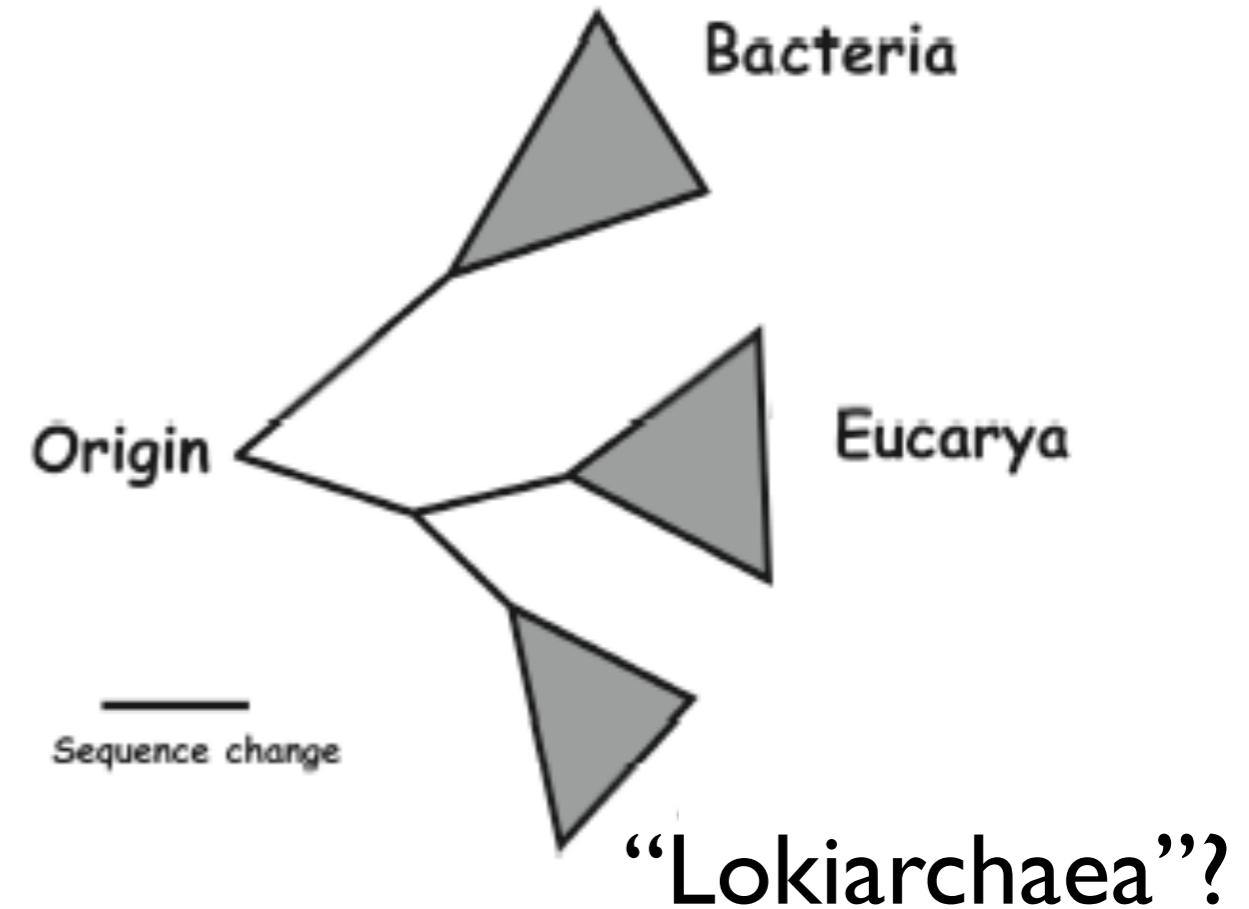
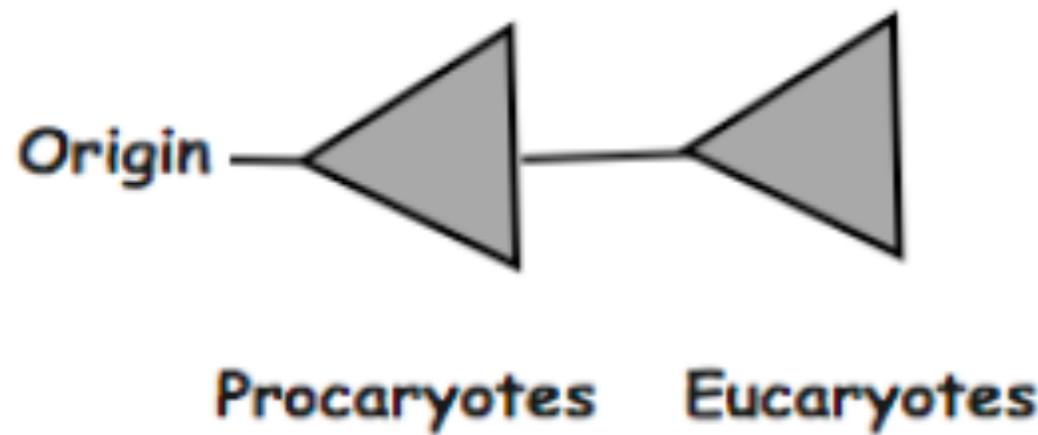
“eukaryotic host evolved from a bona fide archaeon...”

# what's wrong with this picture?



iMessage with **Charles Darwin**

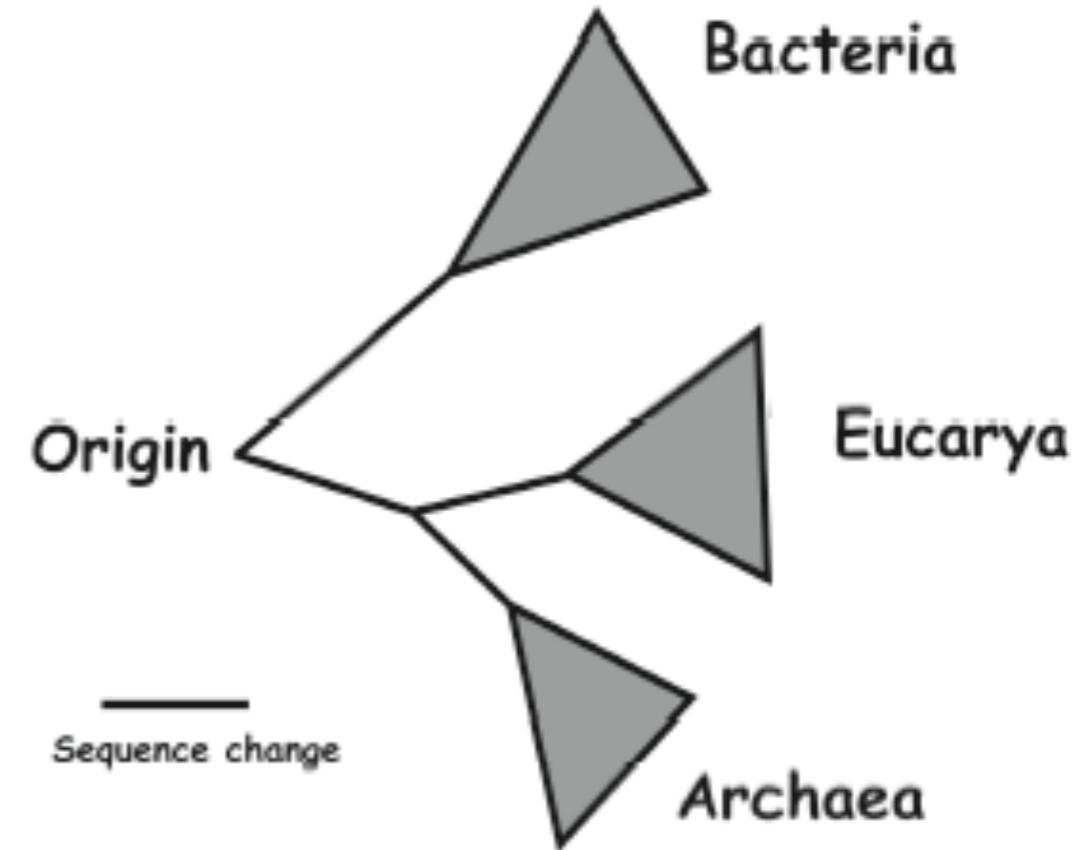
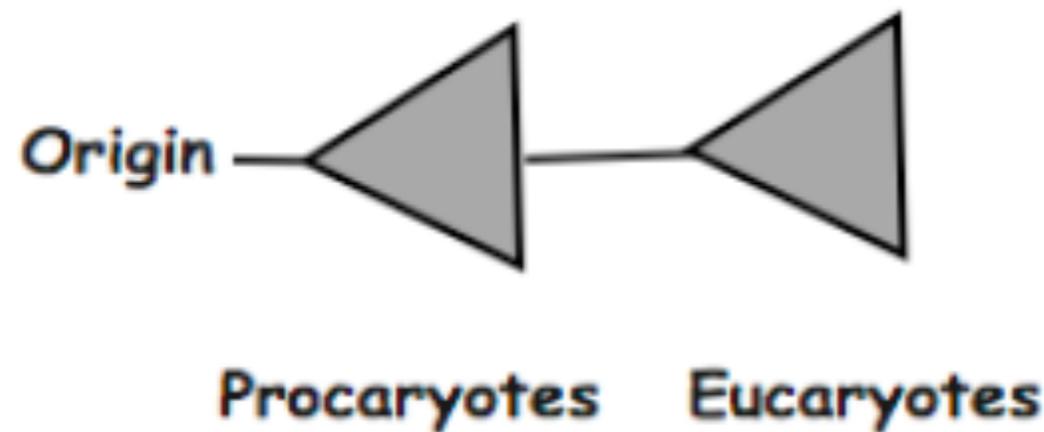
Today, 9:49 PM



“Prokaryote” is still not a monophyletic group

# “Prokaryote”= “pre-eukaryote”

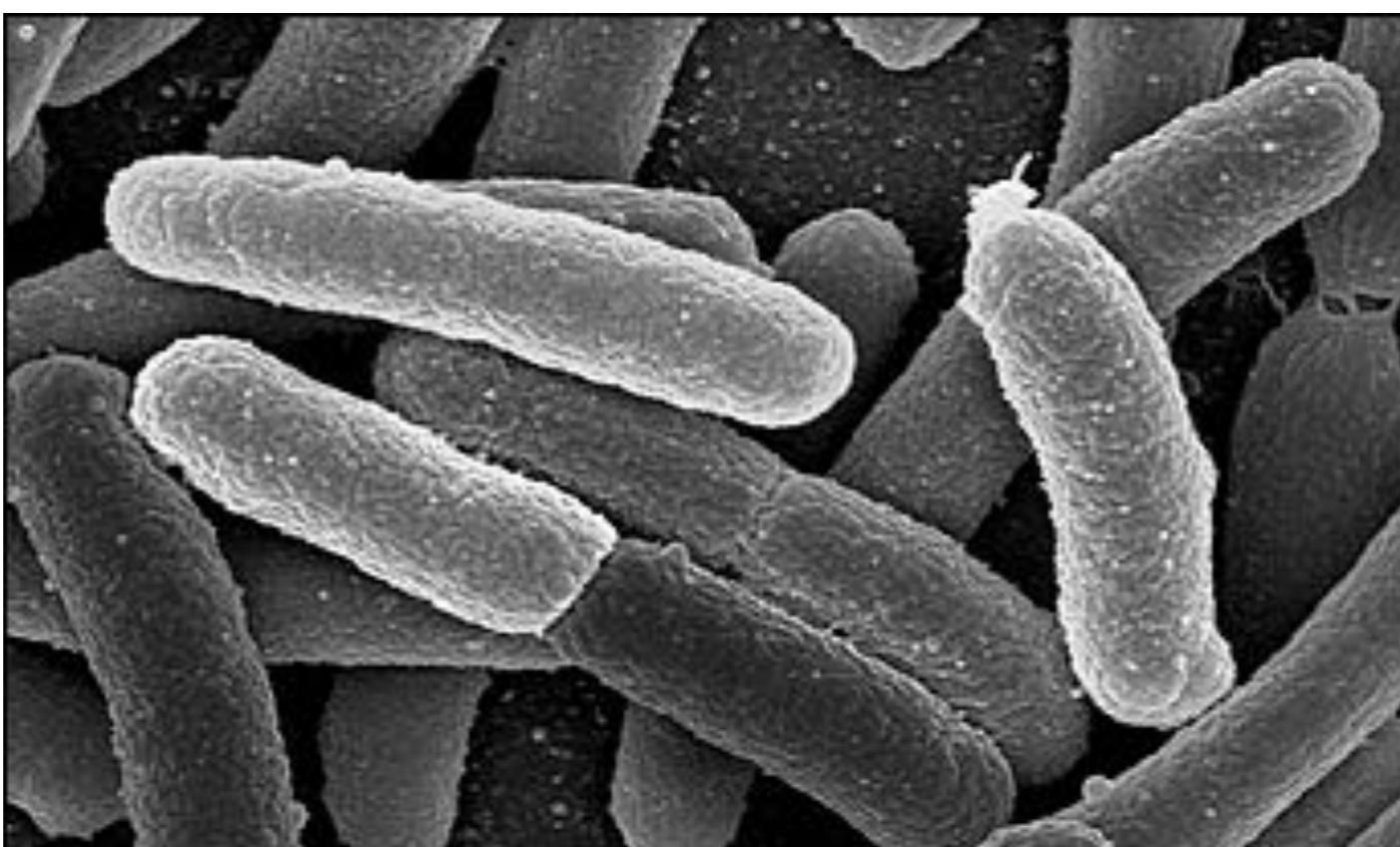
This is a hypothesis, not supported by evidence



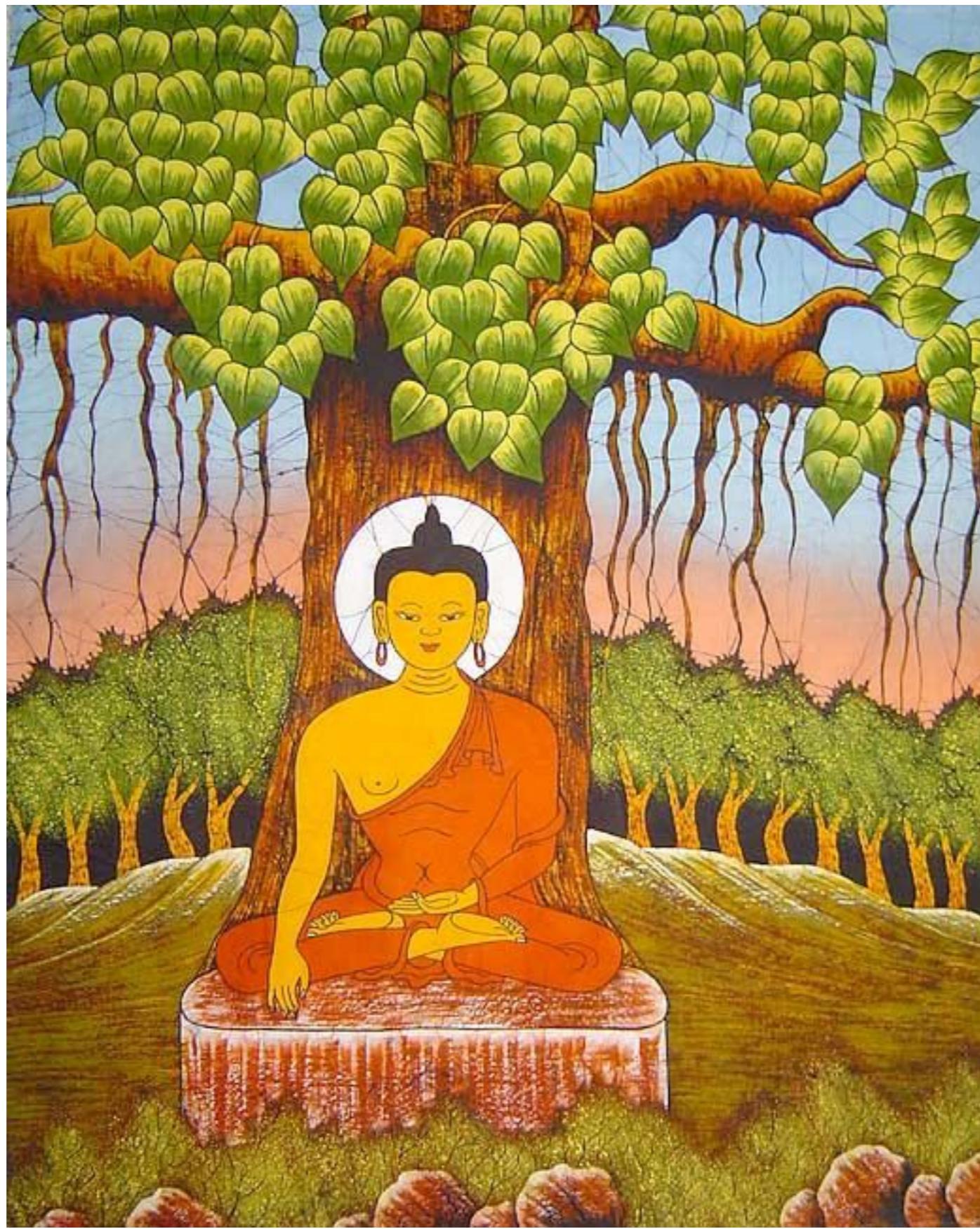
## “Prokaryote” is not a clade

**most recent common  
ancestor**





# What's common to all life?



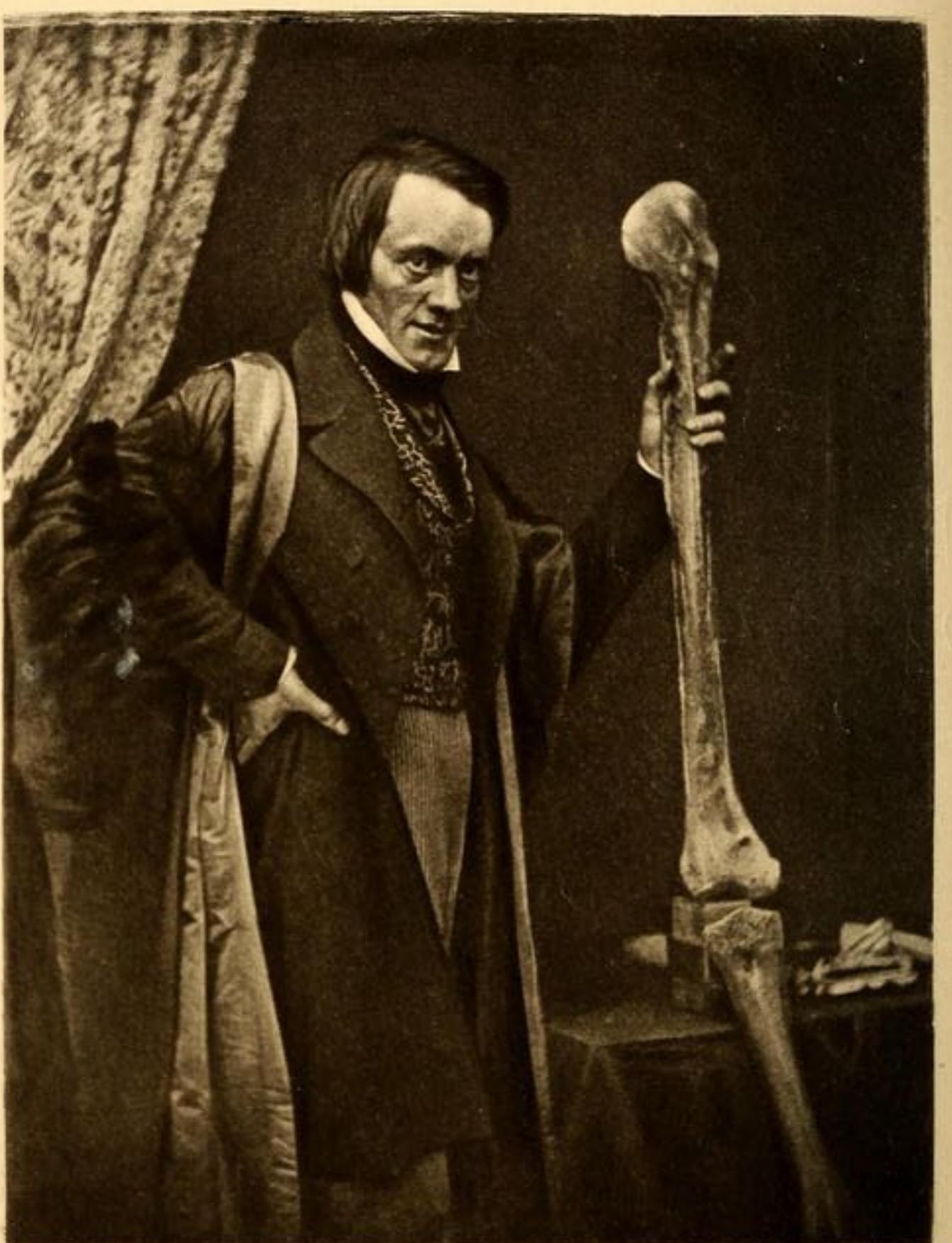


Photo-engraved by Walker & Boutall from a Daguerreotype.

Richard Owen

# Homology

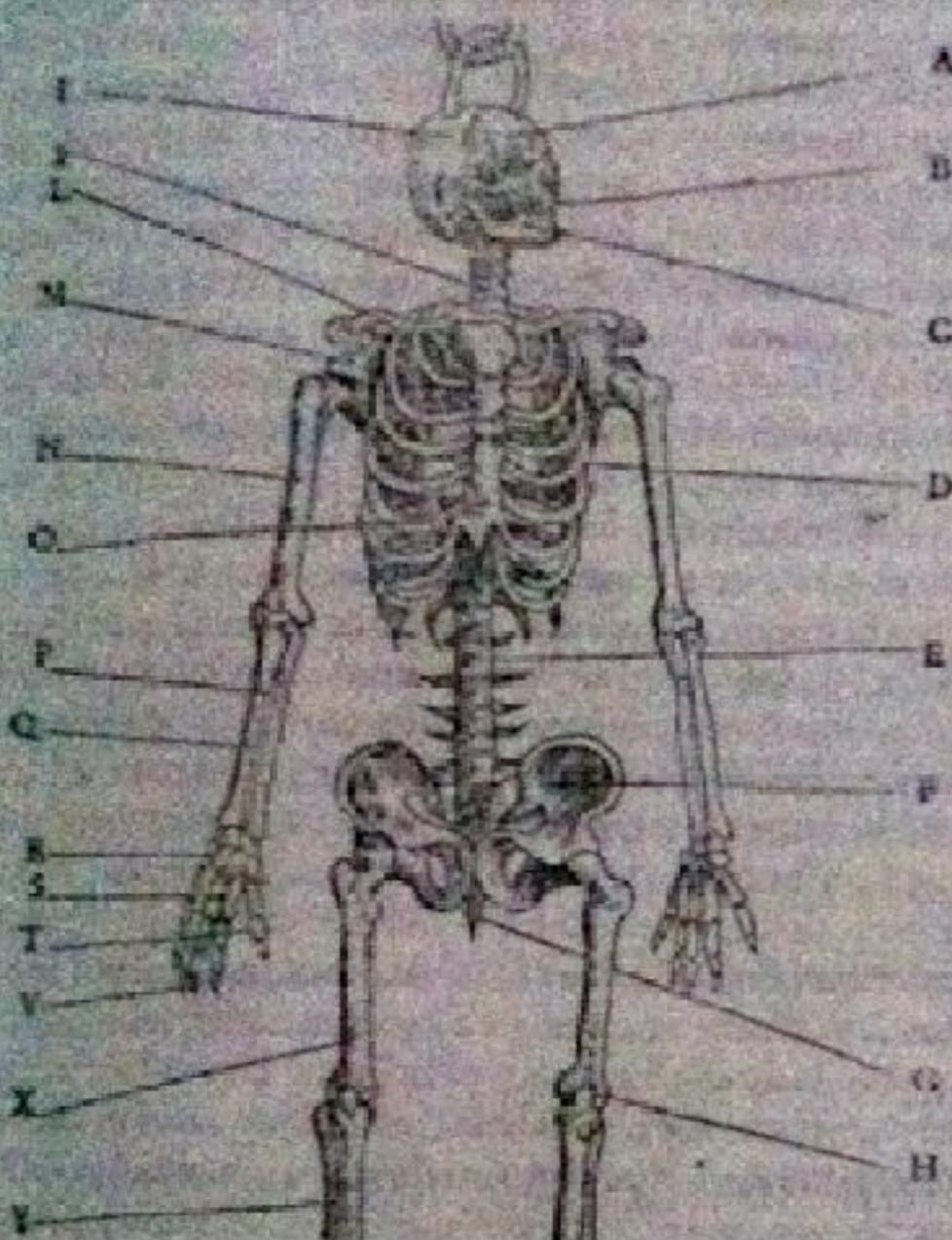
“The same organ in different animals under a variety of form and function”

Sir Richard Owen (1843)

# Homology

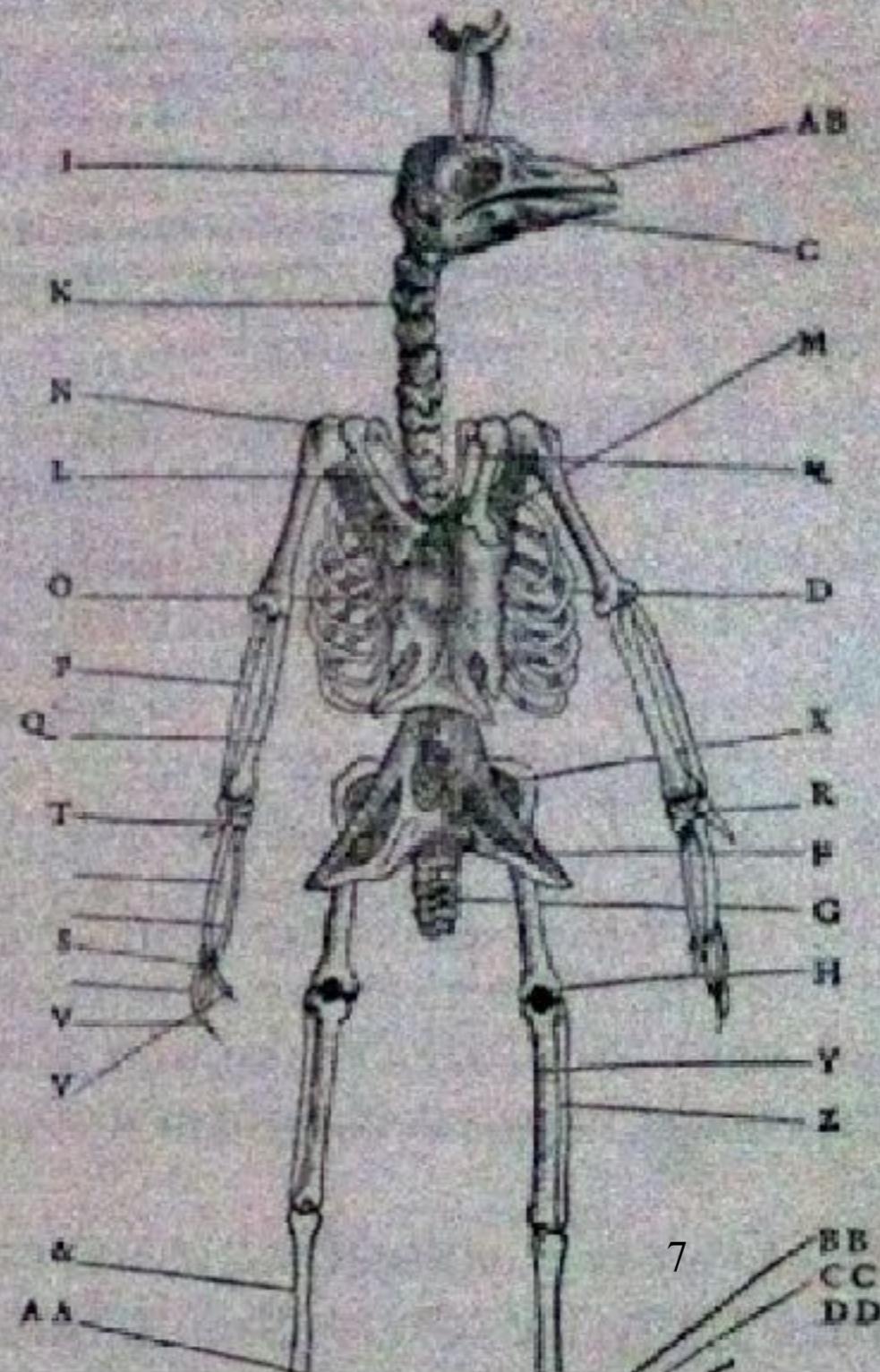
- Because all life shares a common ancestry - so should the “parts” of life
- Homology is similarity due to common descent
- Analogy is due to adaptation to the same environmental condition (convergence)
- “Levels” of homology

Portrait de l'os de l'oreille humaine, pris en comparaison  
de l'os de l'oreille des oiseaux, faisant que les  
lettres d'ordre se rapportent à celle-ci, pour  
faire approuver combien l'affinité est  
grande des vins aux autres.



La comparaison du fonds portant des os humains montre com-  
bien cette oy est qui est d'un oy le plus prochain.

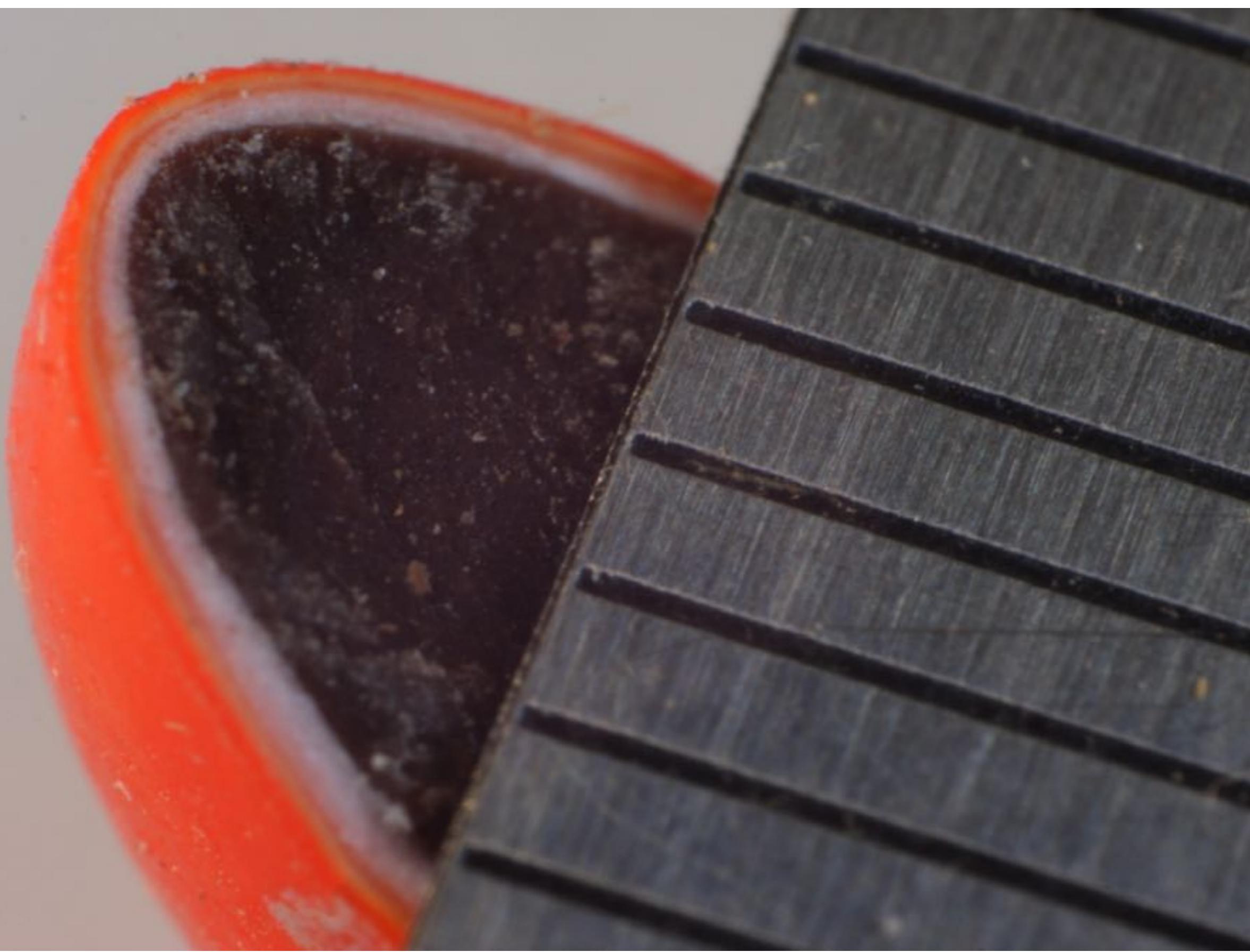
#### Portrait des os de l'oyeau.



# steps in assessing homology

1. Classify features
2. Classify similar in similar positions (e.g. structures)
3. Link features to tree of evolutionary relationships



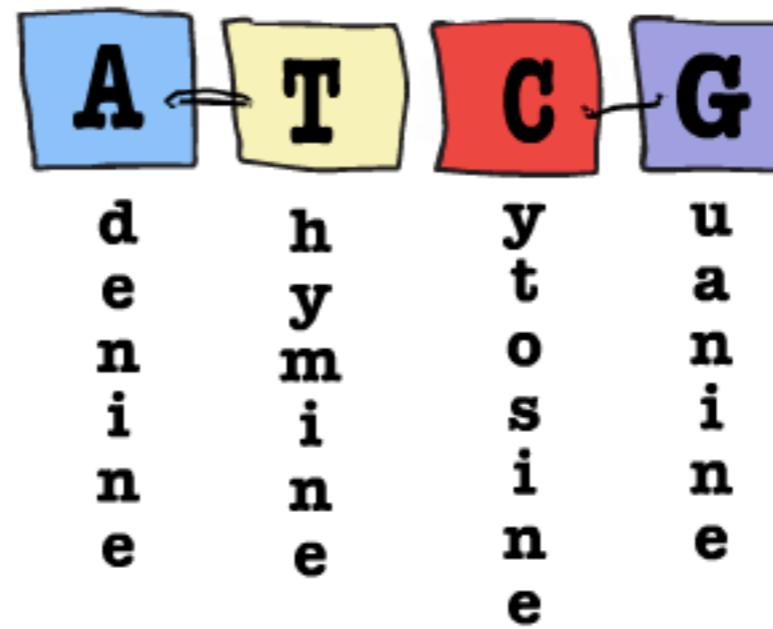
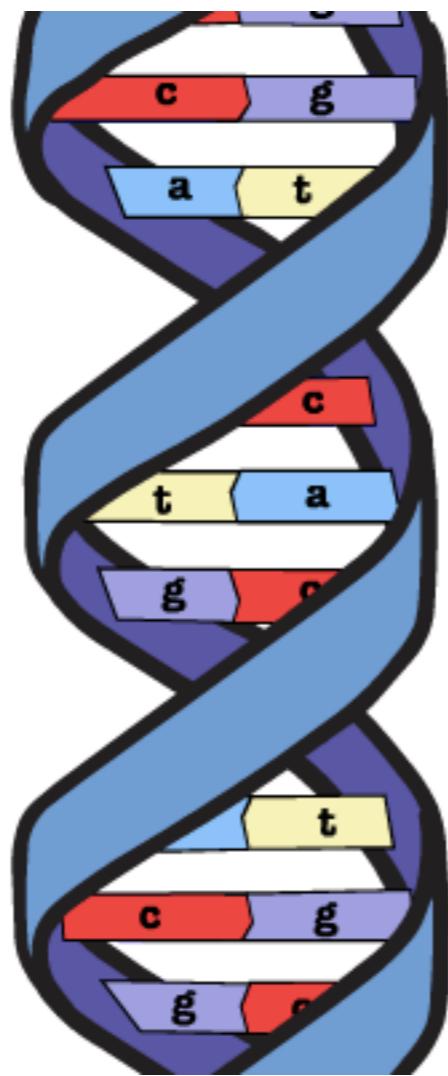


# M & M diversity



# “Molecules as documents of evolutionary history”

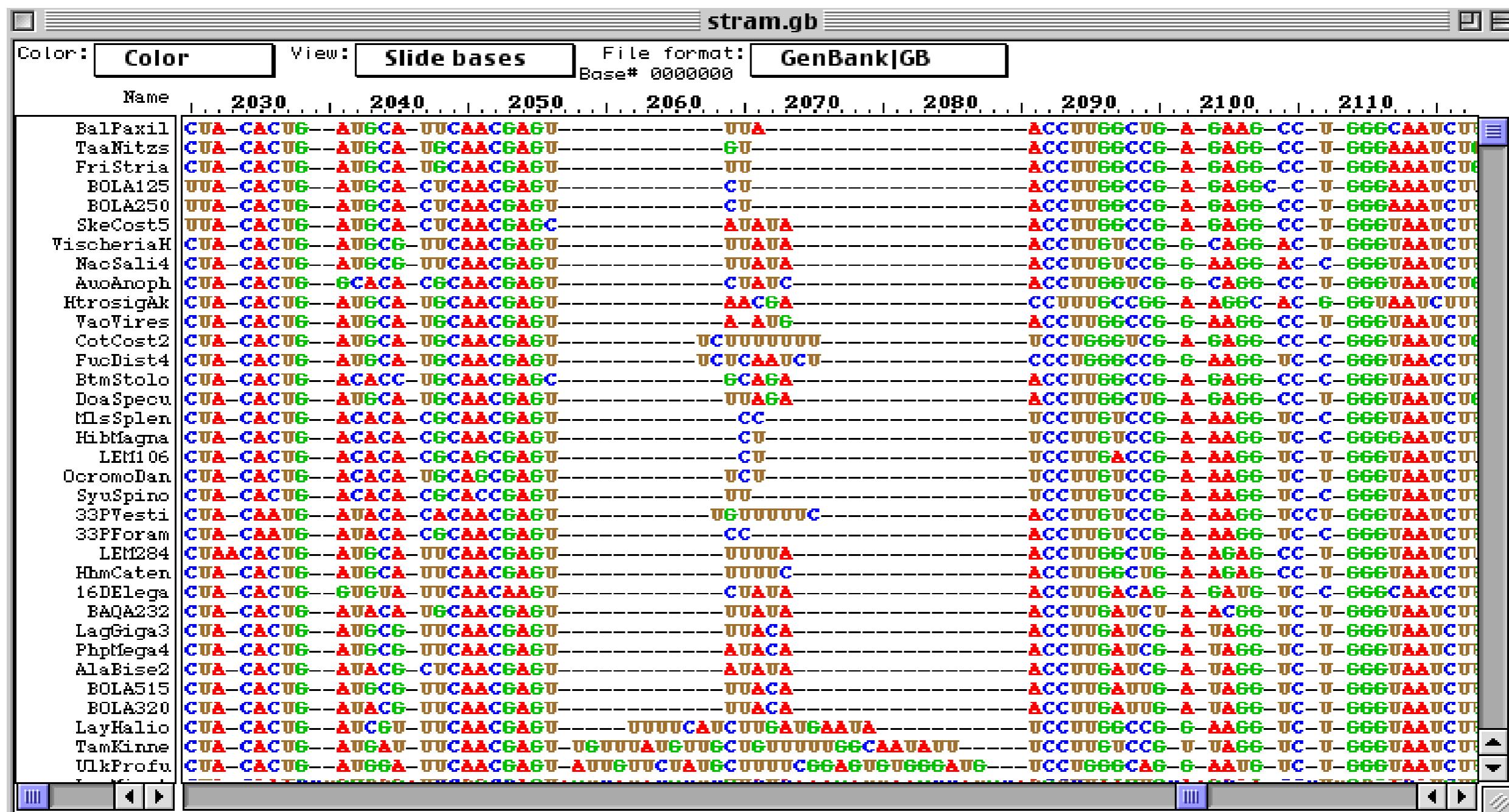
Zuckerkandl and Pauling (1965)



# Homology

- Because all life shares a common ancestry - so should the “parts” of life
- Shares common ancestry (yes or no)
- Does NOT = similarity (not a percentage)
- % identity (nts); % similarity (aa's)
- “Degrees” or “Levels” of homology
- Is sequence similarity better than chance? Tends to be homology
- Use BLAST or HMMer (probalistic method for distant homologs)
  - <http://hmmer.janelia.org/>

# Structure-based sequence alignment

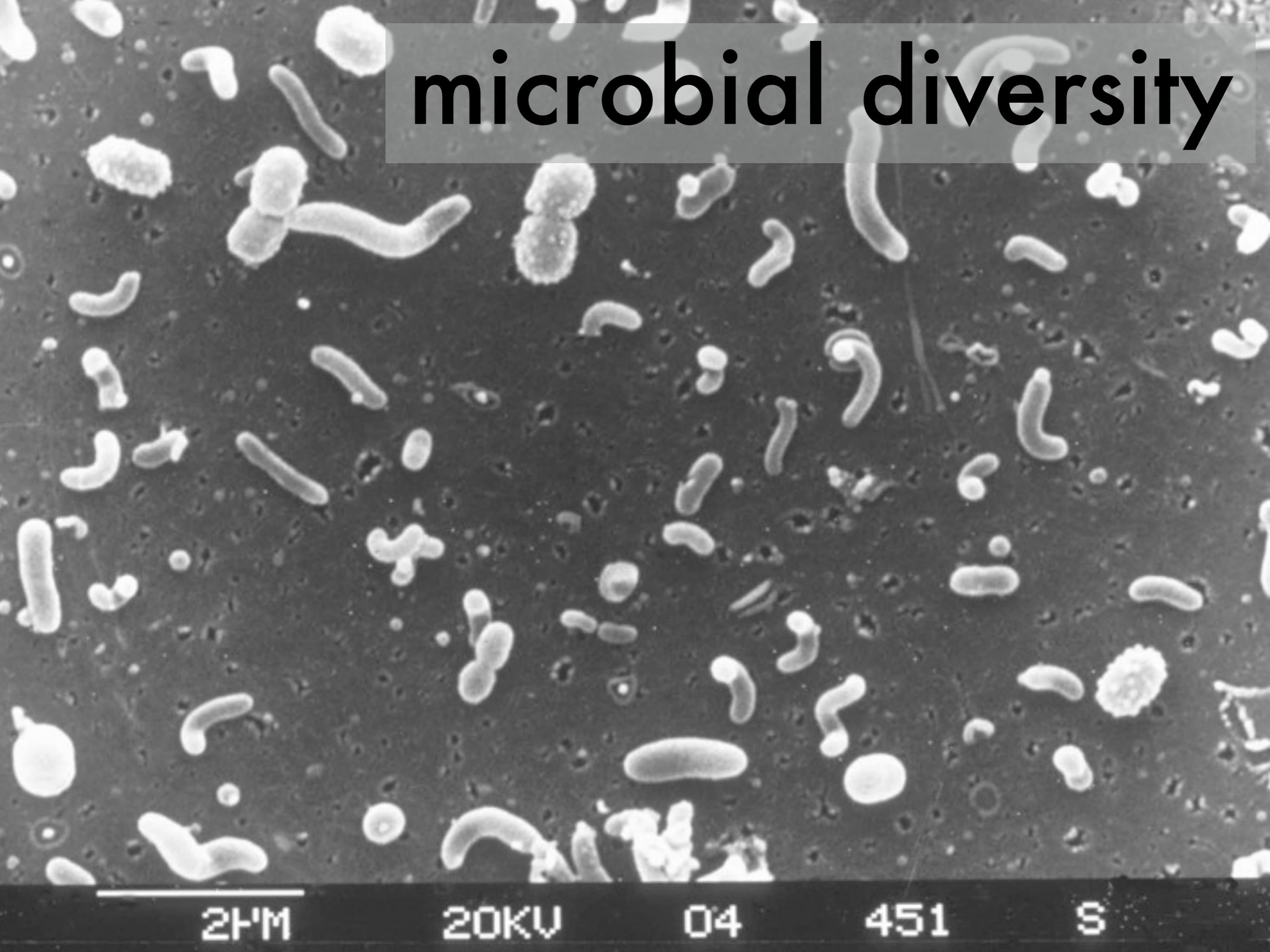


Are different sized stems and loops homologous?

# similarity vs. homology

- **homology**: common ancestor but **not** necessarily common function
- sequence similarity does not imply homology
- homologous sequences ***tend*** to be similar
- homologs are usually > 40% identical at nucleotide (4)  
and usually > 25% at amino acid (20) levels (length matters)
- ***domains may be homologous***, not always entire protein

# microbial diversity



\*

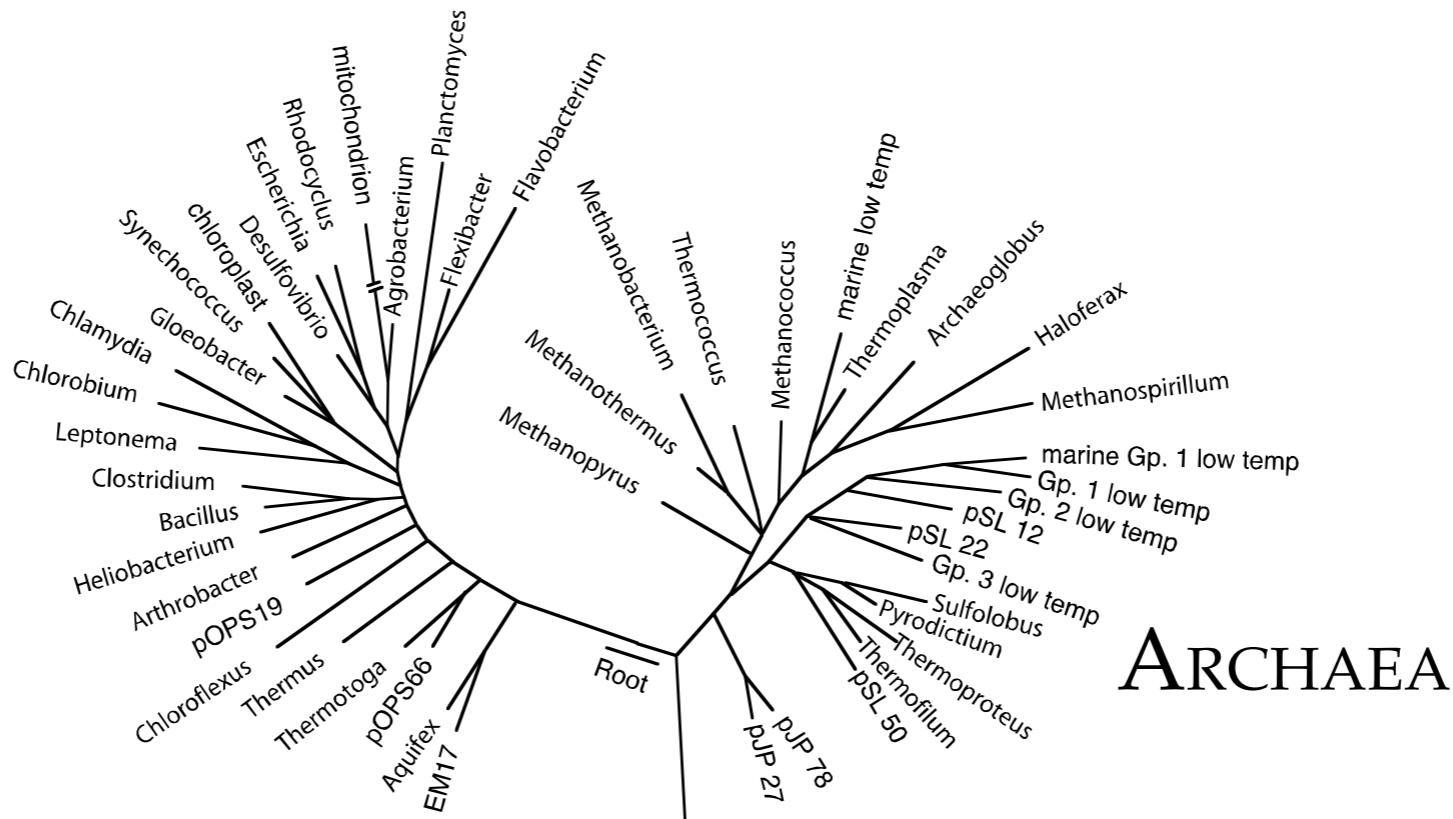
**Carl Woese used rRNA genes to determine “natural” or evolutionary relationships among microbes**



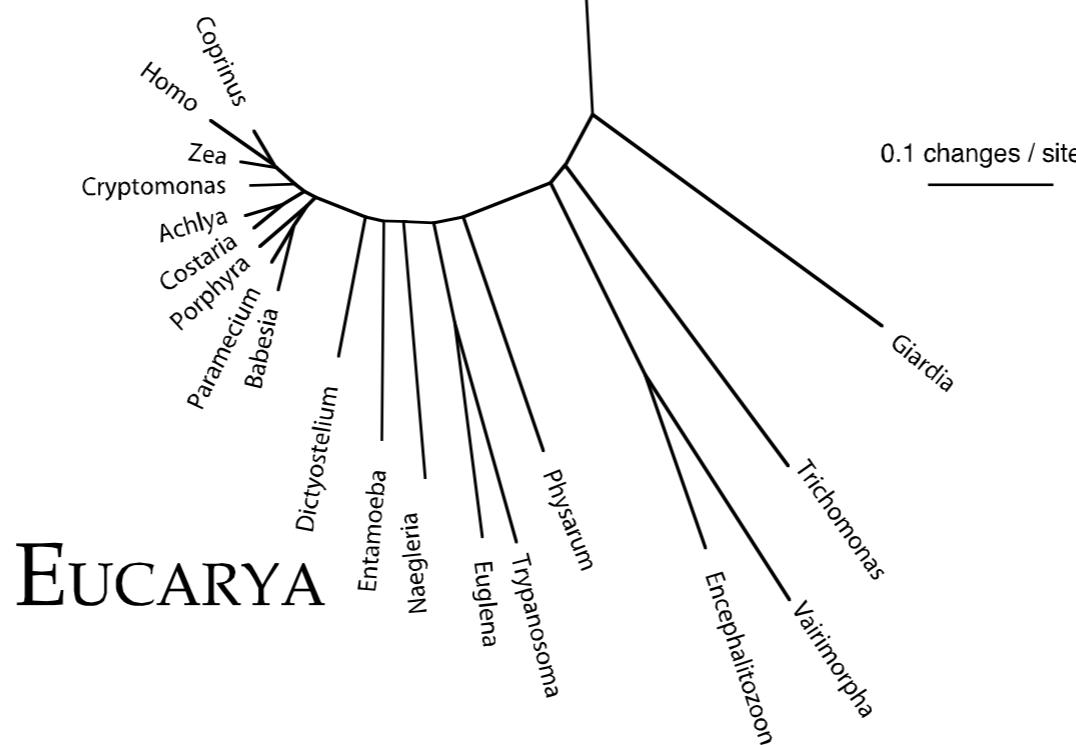
\* all 16S rRNA (small subunit rRNA) genes are homologous

# Gene Trees as Proxies for Organism Trees

## BACTERIA



## ARCHAEA



## EUCARYA

0.1 changes / site

# Requirements of a gene “proxy” for an organismal phylogeny

- Present in all organisms in single copy

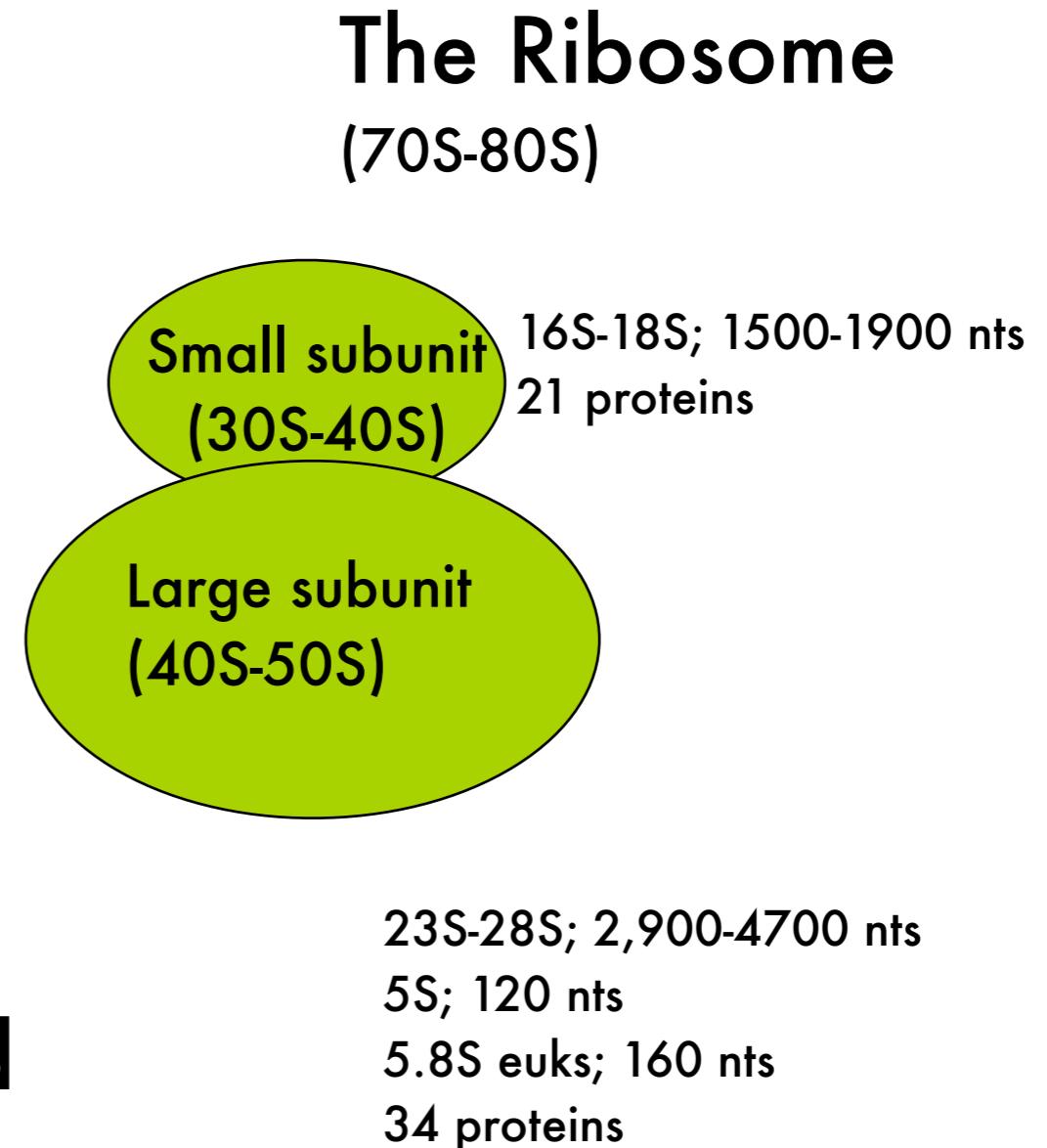
- Homologous  
(common evolutionary origin)

- Robust (conserved) yet  
“evolvable” (varied)

- Lack lateral transfer

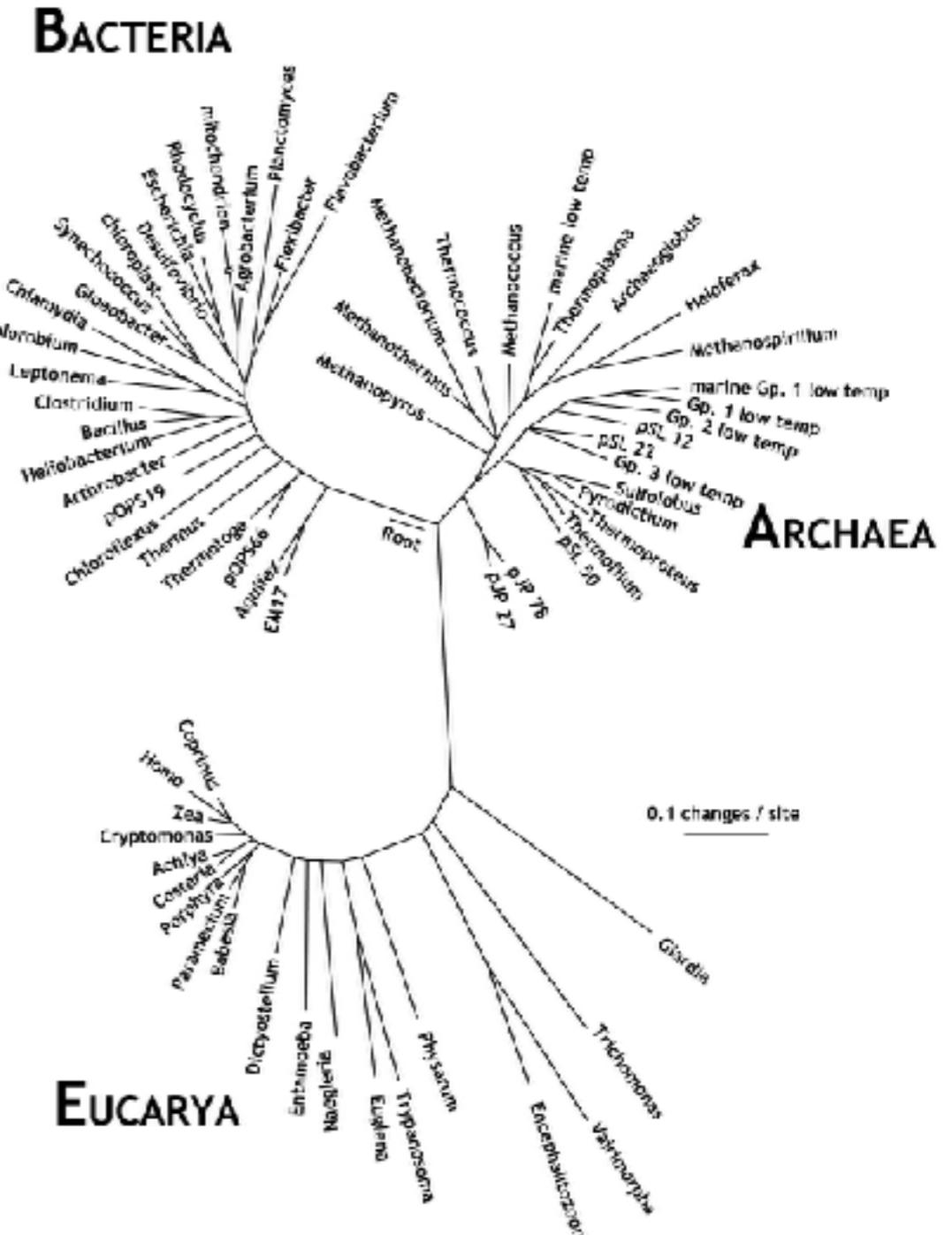
- Readily obtained and sequenced

- ribosomal RNA (rRNA)



# Lessons from the “Big Tree”

- All cellular life is related (one origin)
  - Three domains of life, not two: Bacteria, Archaea, Eukarya
  - Most life is microbial (still)
  - Natural taxonomy: based on evolutionary relatedness  
i.e, allows phylogeny to be predictive
    - Related organisms (or related genes) should have similar properties (not that we always know “which” properties those are)



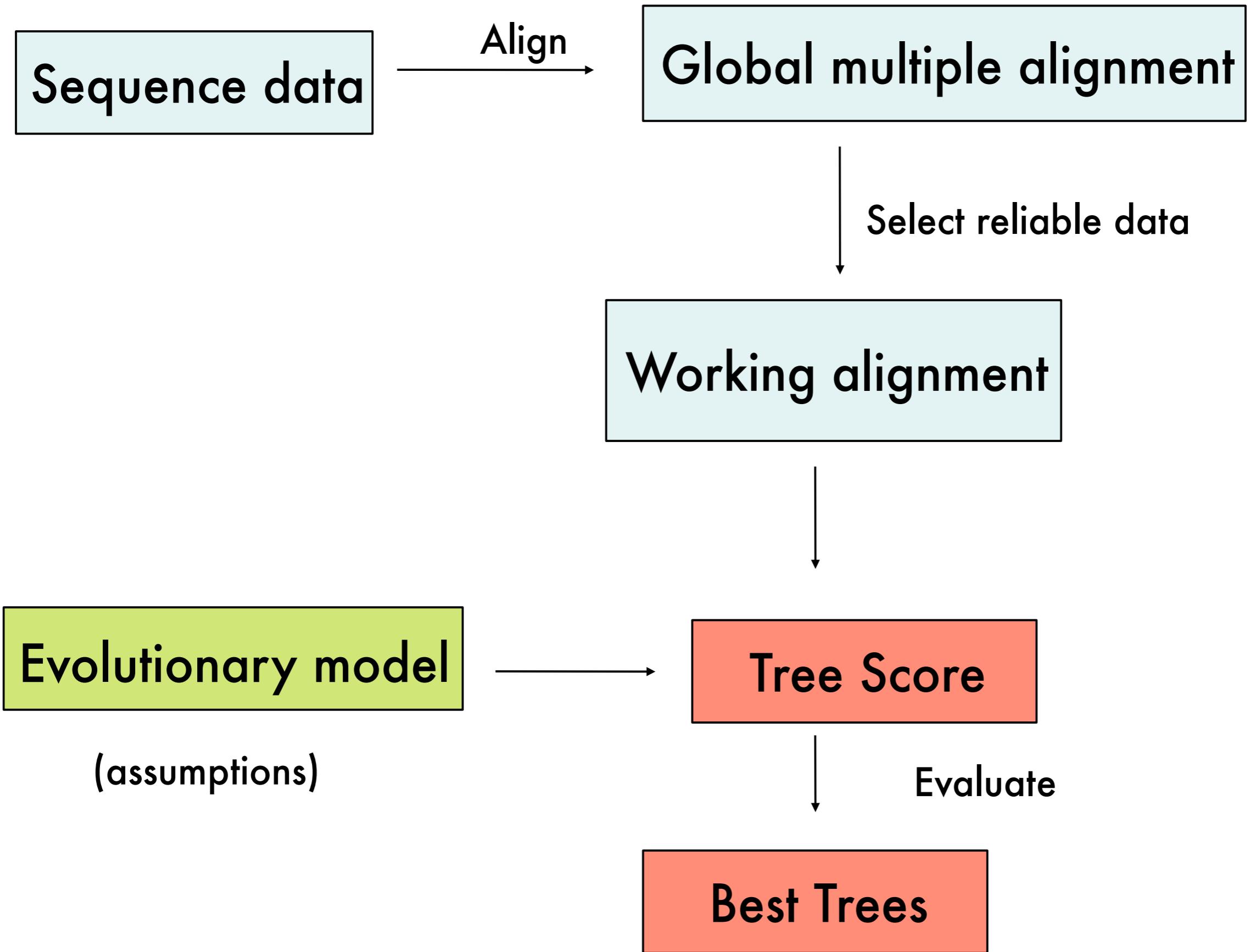
# Zen and the Art of Molecular Phylogeny

## Part 2: Building Trees

The road to  
Phylogenetic Nirvana  
by overcoming the  
Five Hindrances:

Desire  
Anger  
Sloth  
Worry  
Doubt





# **Multiple Sequence Alignment**

## **MSA**

# multiple sequence alignment

- extension of pairwise alignment
- need to know which parts correspond (homology)
- infer “positional” homology
- conserved regions can make “profiles” of protein family
- structure prediction (more conserved than 1° sequence)
- genome assembly - construction of contig maps
- accurate MSA required for phylogenetic analysis

**stram.gb**

Color:	Color	View:	Slide bases	File format:	GenBank GB				
				Base# 0000000					
Name	2030	2040	2050	2060	2070	2080	2090	2100	2110
BalPaxil	CUA-CACUG-AUGCA-UUCAACGAGU				UUU		ACCUUUGGCCUG-A-GAAAG-CC-U-GGGCAAAUCU		
TaaNitzs	CUA-CACUG-AUGCA-UUCAACGAGU				GU		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
FriStria	CUA-CACUG-AUGCA-UUCAACGAGU				UU		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
BOLA125	UUU-CACUG-AUGCA-UUCAACGAGU				CU		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
BOLA250	UUU-CACUG-AUGCA-UUCAACGAGU				CU		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
SkeCost5	UUU-CACUG-AUGCA-UUCAACGAGC				AUAUA		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
VischeriaH	CUA-CACUG-AUGCG-UUCAACGAGU				UUUAU		ACCUUUGGCCUG-G-CAGG-AC-U-GGGAAAUCU		
NacSali4	CUA-CACUG-AUGCG-UUCAACGAGU				UUUAU		ACCUUUGGCCUG-G-AAGG-AC-C-GGGAAAUCU		
AuoAnoph	CUA-CACUG-GCACCA-CBCAACGAGU				CUATC		ACCUUUGGCCUG-G-CAGG-CC-U-GGGAAAUCU		
HtrosigAk	CUA-CACUG-AUGCA-UUCAACGAGU				AACGA		CCUUUGGCCUG-A-AAGC-AC-G-GGUAAAUCU		
TaoTires	CUA-CACUG-AUGCA-UUCAACGAGU				& AUG		ACCUUUGGCCUG-G-AAGG-CC-U-GGGAAAUCU		
CotCost2	CUA-CACUG-AUGCA-UUCAACGAGU				TCUUUUUUU		UCCUUGGCCUG-A-GAGG-CC-C-GGGAAAUCU		
FucDist4	CUA-CACUG-AUGCA-UUCAACGAGU				TCUCAACU		CCCUUGGCCUG-G-AAGG-UC-C-GGGAAAUCU		
BtmStolo	CUA-CACUG-ACACC-UUCAACGAGC				GCAGA		ACCUUUGGCCUG-A-GAGG-CC-C-GGGAAAUCU		
DoaSpecu	CUA-CACUG-AUGCA-UUCAACGAGU				UUAGA		ACCUUUGGCCUG-A-GAGG-CC-U-GGGAAAUCU		
MlsSplen	CUA-CACUG-ACACA-CBCAACGAGU				CC		UCCUUGGCCUG-A-AAGG-UC-C-GGGAAAUCU		
HibMagna	CUA-CACUG-ACACA-CBCAACGAGU				CU		UCCUUGGCCUG-A-AAGG-UC-C-GGGAAAUCU		
LEM106	CUA-CACUG-ACACA-CBCAACGAGU				CU		UCCUUGGCCUG-A-AAGG-UC-U-GGGAAAUCU		
OchromoDan	CUA-CACUG-ACACA-UUCAGCGAGU				UCU		UCCUUGGCCUG-A-AAGG-UC-U-GGGAAAUCU		
SyuSpino	CUA-CACUG-ACACA-CBCAACGAGU				UU		UCCUUGGCCUG-A-AAGG-UC-C-GGGAAAUCU		
33PTesti	CUA-CAAUG-AUACA-CACAACGAGU				UGUUUUC		ACCUUUGGCCUG-A-AAGG-UCCT-GGGAAAUCU		
33PForam	CUA-CAAUG-AUACA-CBCAACGAGU				CC		ACCUUUGGCCUG-A-AAGG-UC-C-GGGAAAUCU		
LEM284	CUAACACUG-AUGCA-UUCAACGAGU				UUUUA		ACCUUUGGCCUG-A-AAGG-CC-U-GGGAAAUCU		
HhmCaten	CUA-CACUG-AUGCA-UUCAACGAGU				UUUUC		ACCUUUGGCCUG-A-AAGG-CC-U-GGGAAAUCU		
16DElega	CUA-CACUG-GUGUA-UUCAACGAGU				CUAUU		ACCUUUGACAG-A-GAU-UC-C-GGGAAAUCU		
BAQA232	CUA-CACUG-AUACA-UUCAACGAGU				UUUAU		ACCUUUGACU-A-ACG-UC-U-GGGAAAUCU		
LagGiga3	CUA-CACUG-AUGCG-UUCAACGAGU				UUACU		ACCUUUGACU-G-AAGG-UC-U-GGGAAAUCU		
PhpMega4	CUA-CACUG-AUGCG-UUCAACGAGU				UUACA		ACCUUUGACU-G-AAGG-UC-U-GGGAAAUCU		
AlaBise2	CUA-CACUG-AUACA-CUCAACGAGU				AUAUA		ACCUUUGACU-G-AAGG-UC-U-GGGAAAUCU		
BOLA515	CUA-CACUG-AUGCG-UUCAACGAGU				UUACA		ACCUUUGACU-G-AAGG-UC-U-GGGAAAUCU		
BOLA320	CUA-CACUG-AUACA-UUCAACGAGU				UUACA		ACCUUUGACU-G-AAGG-UC-U-GGGAAAUCU		
LayHalio	CUA-CACUG-AUCGU-UUCAACGAGU				UUUCAUCUUGAUUAUA		UCCUUGGCCUG-G-AAGG-UC-U-GGGAAAUCU		
TamKinne	CUA-CACUG-AUGAU-UUCAACGAGU				UGUUUAGUUCUUGUUCUUGGGCAAAUU		UCCUUGGCCUG-U-AAGG-UC-U-GGGAAAUCU		
UlkProfu	CUA-CACUG-AUGGA-UUCAACGAGU				AUUGUUCAUCUUGGGCGGAGUUGGGGAU		UCCUUGGCCUG-G-AAGG-UC-U-GGGAAAUCU		

**Root sequence: A T G T T C T T G C A T A A C G**



Figure 17.20 Microbiology: An Evolving Science  
© 2009 W.W. Norton & Company, Inc.

% identity?

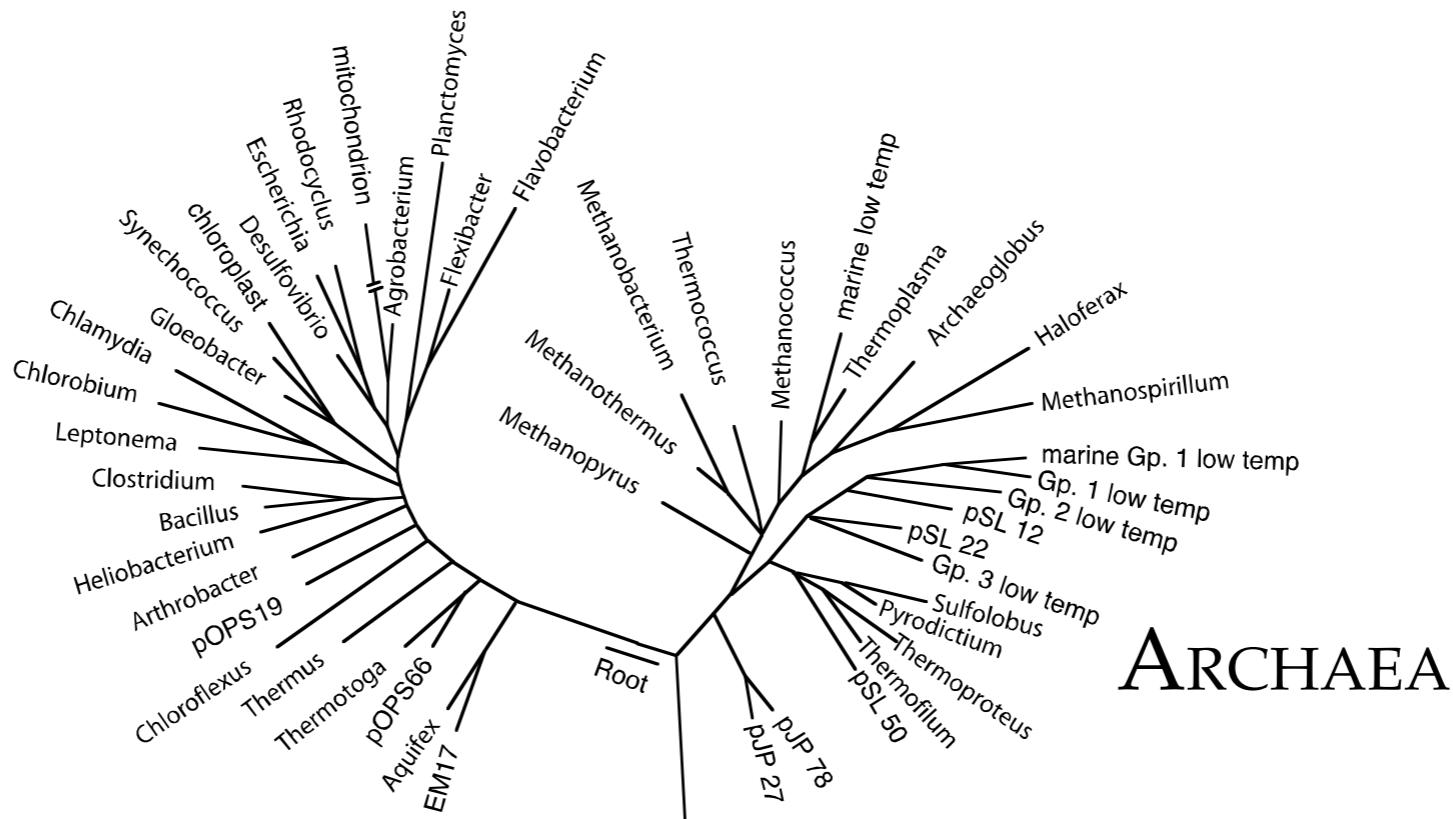
## Multiple sequence alignment:

- taxa in rows
- homologous positions in columns
- use “gaps” or “indels” to bring homologous positions in alignment

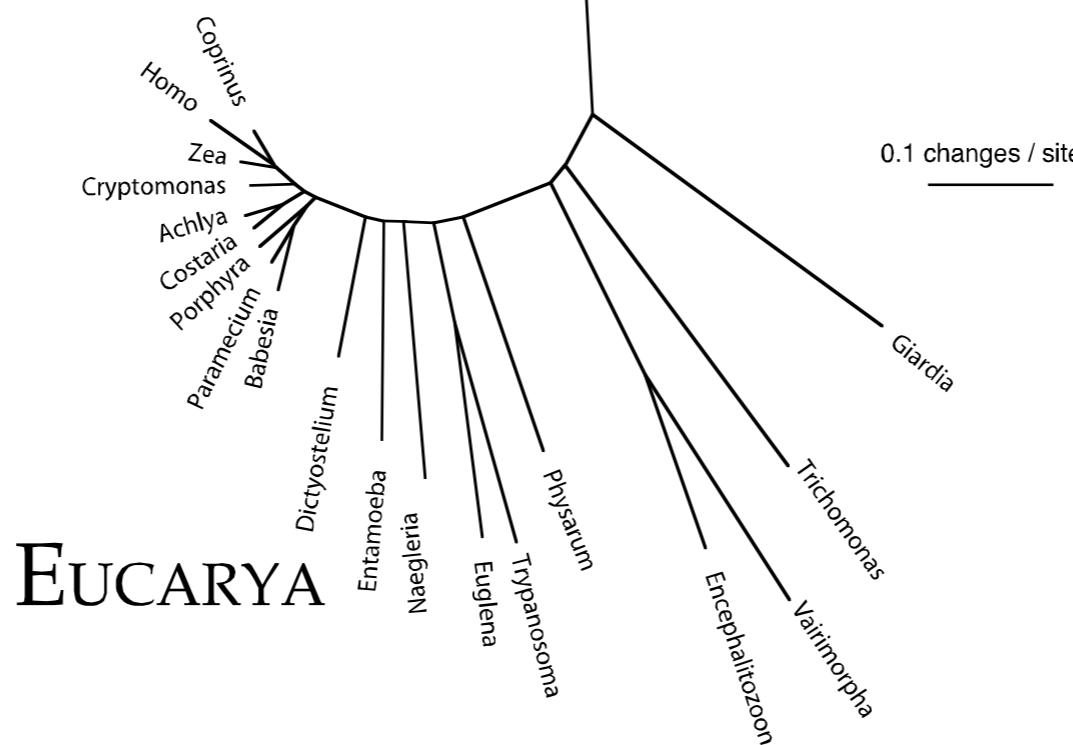
\*

# Gene Trees as Proxies for Organism Trees

## BACTERIA

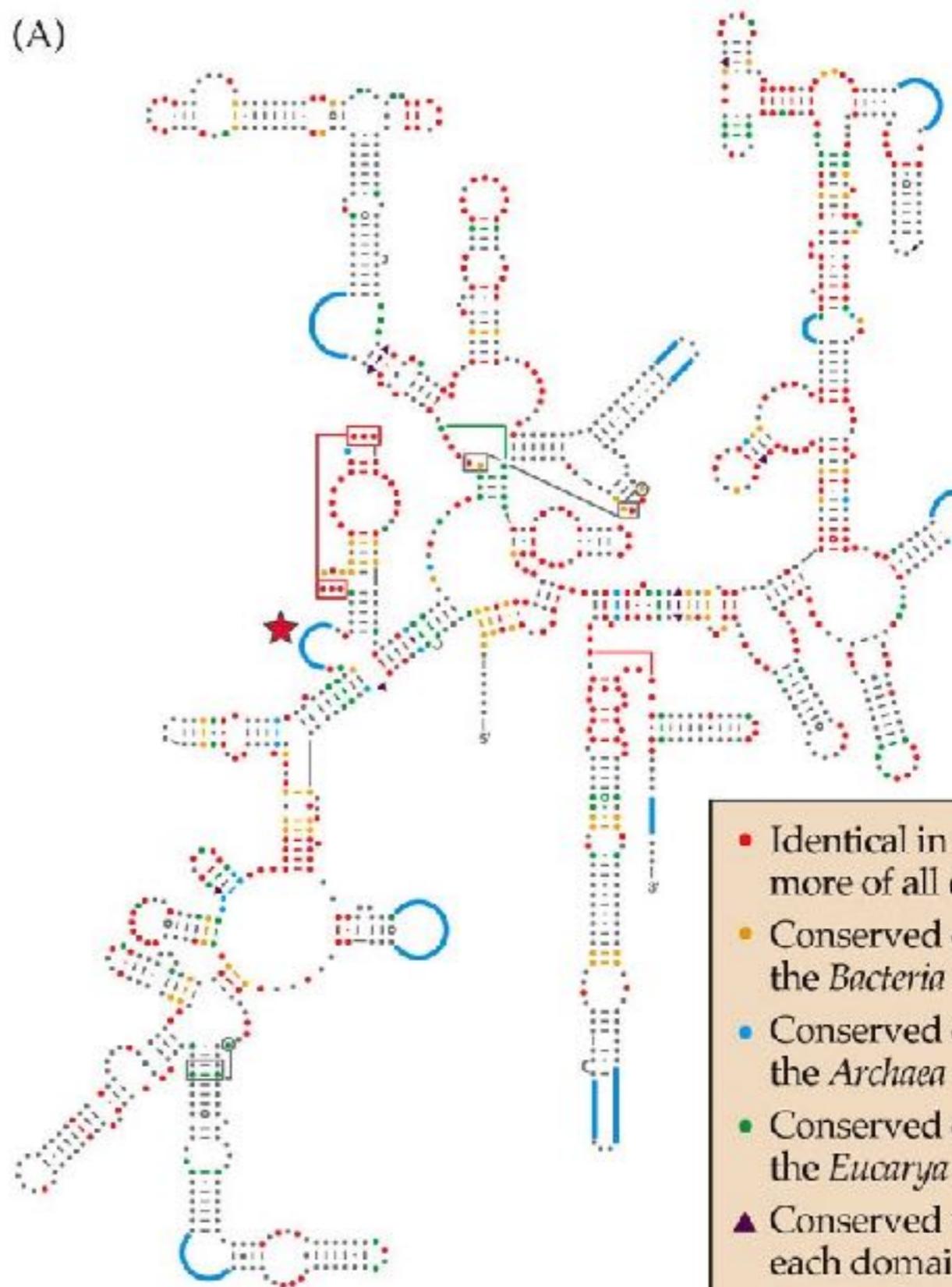


## ARCHAEA



## EUCARYA

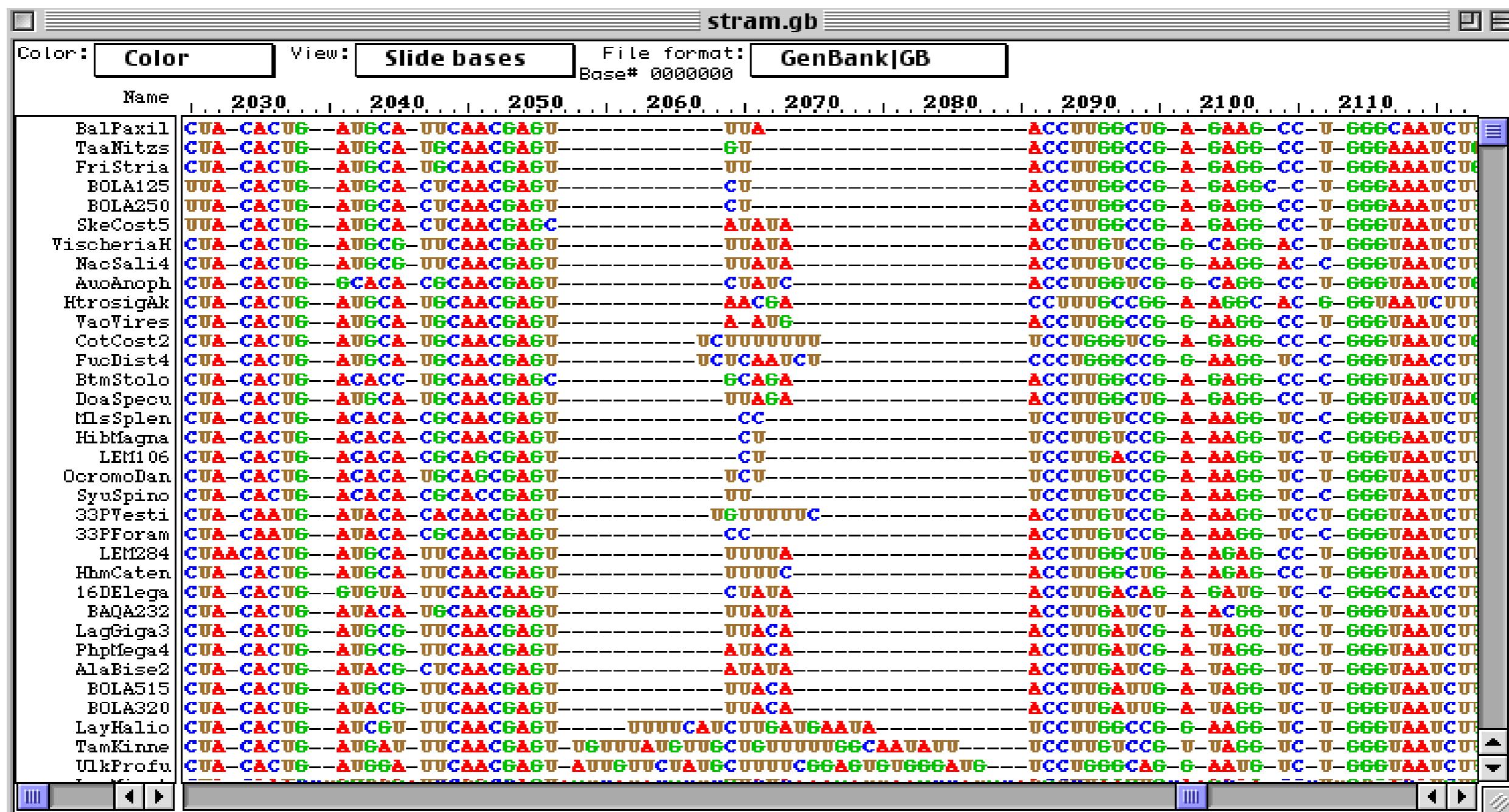
(A)



- Identical in 98% or more of all organisms
- Conserved only in the *Bacteria*
- Conserved only in the *Archaea*
- Conserved only in the *Eucarya*
- ▲ Conserved within each domain, variable among domains
- Regions that vary structurally among domains

similar “phylotypes”= similar seqs

# Structure-based sequence alignment



Are different sized stems and loops homologous?

\*

## Use amino acids to align nucleotides for protein encoding genes

-E---G---S---S---T---L---L---L---G---S--

-E---G---S---S---T---L---L---I---G---S--

-Q---G---S---A---P---L---L---L---G---S--

-Q---G---S---A---T---L---L---A---G---S--

GAA-GGA-A<sup>GC</sup>-T<sup>CC</sup>-T<sup>GG</sup>-TTA-C<sup>T</sup>C-C<sup>T</sup>G-GGA-T<sup>CC</sup>

GAG-GGT-T<sup>CC</sup>-A<sup>GC</sup>-T<sup>AT</sup>-C<sup>T</sup>A-T<sup>TA</sup>-ATT-GGT-A<sup>GC</sup>

GAC-GGC-AGT-GCA-TGG-TTG-CTT-TTG-GGC-AGT

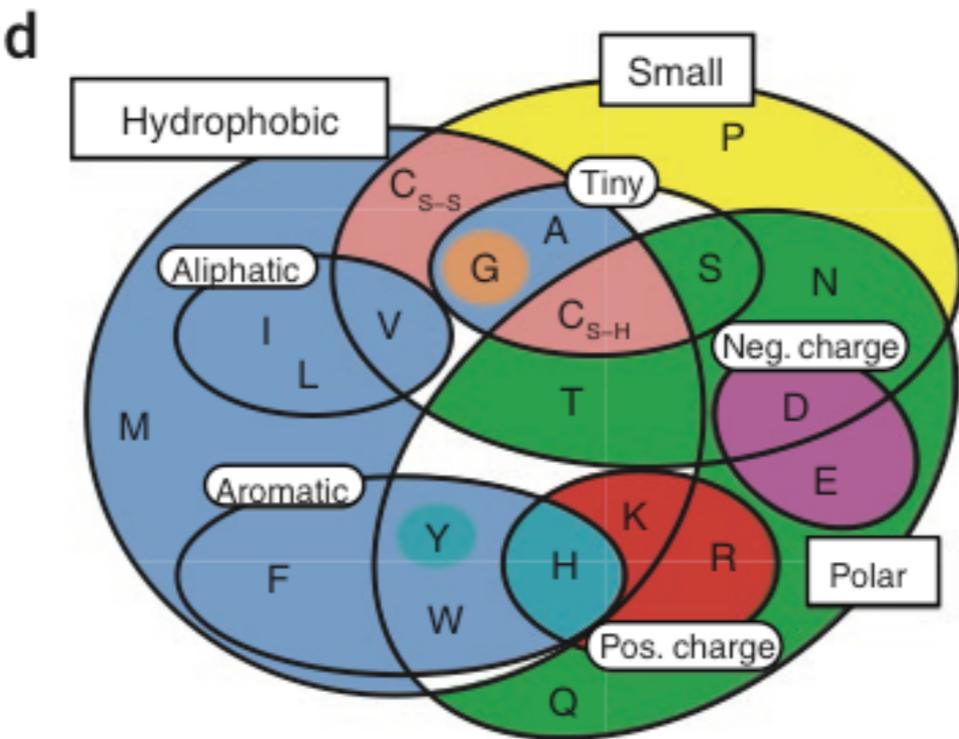
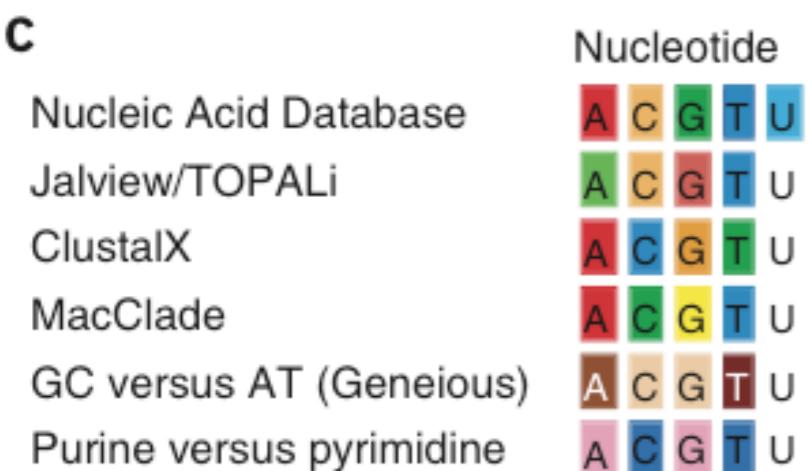
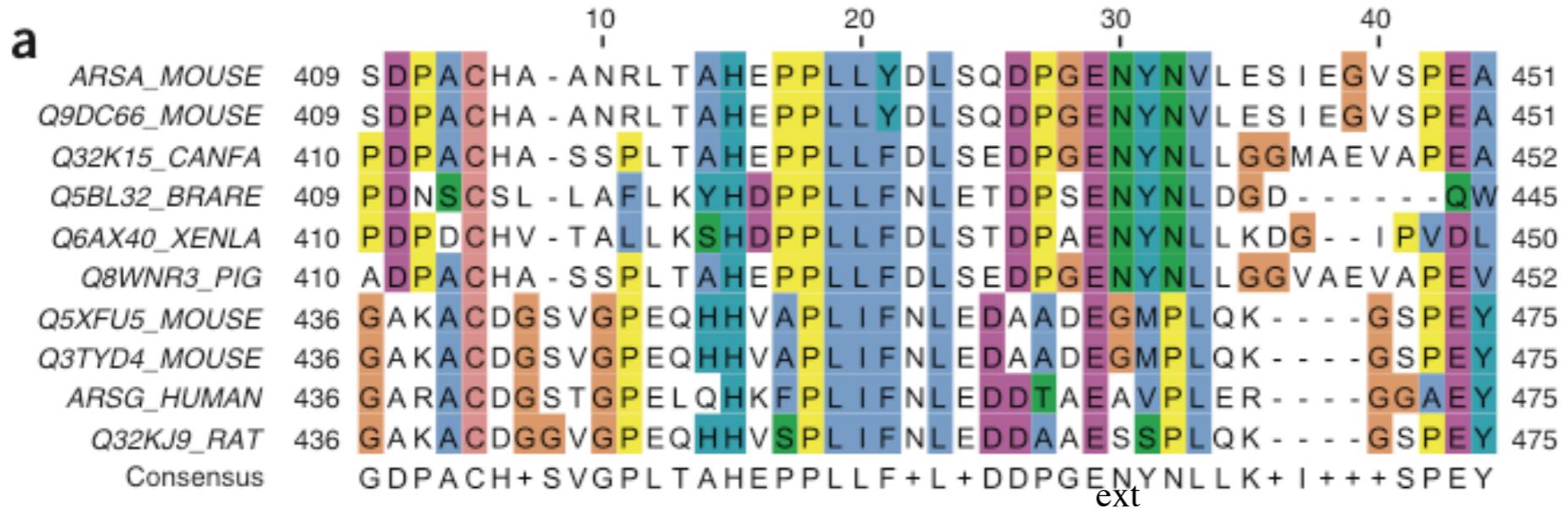
GAT-GGG-TCA-GCT-TAC-CTC-CTG-GCC-GGG-TCA

Calculate % identity both ways

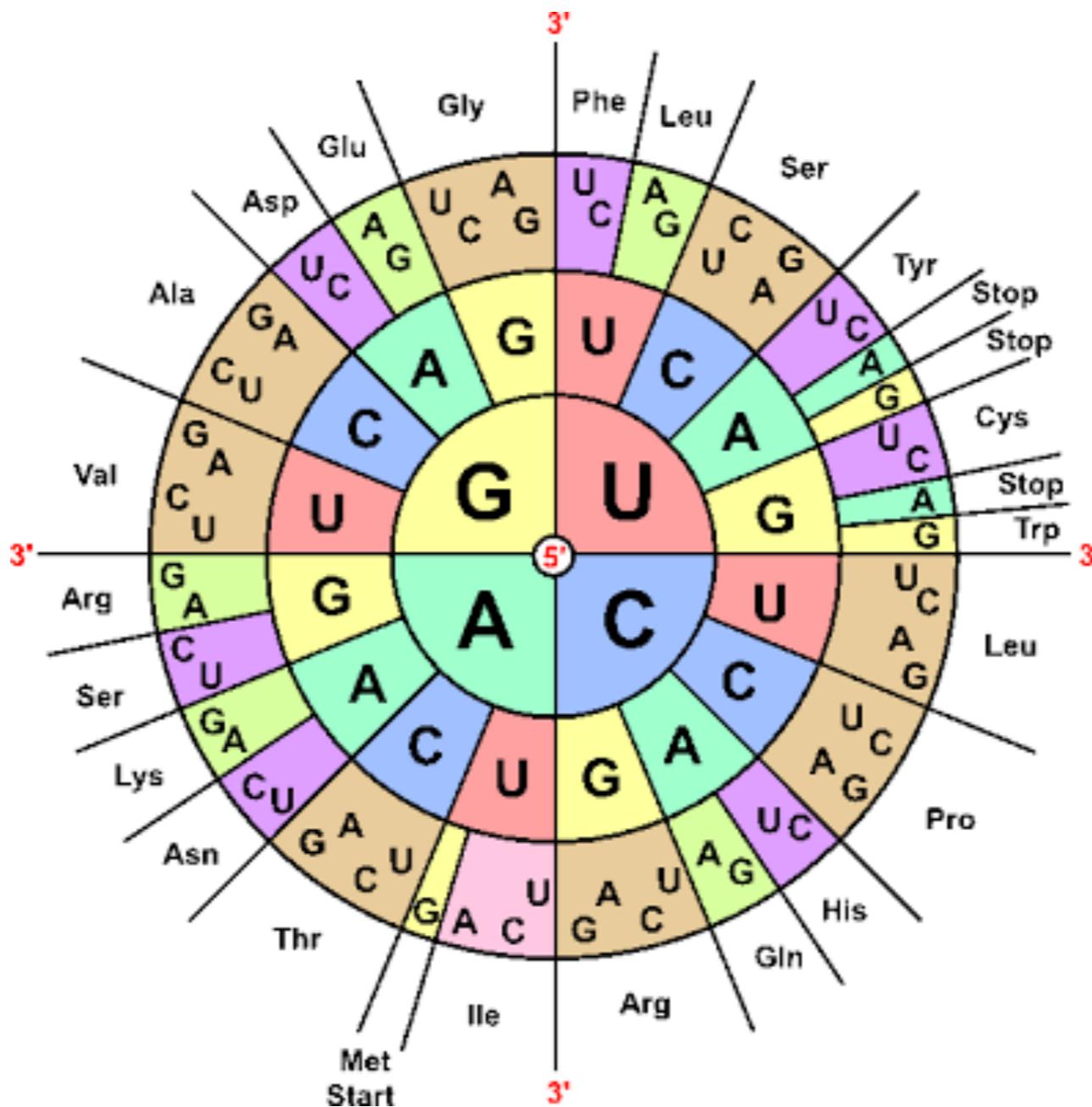
9/10 (aa) : 90%

14/30 (nt) : 47%

# Jalview



# How to design degenerate PCR primers?

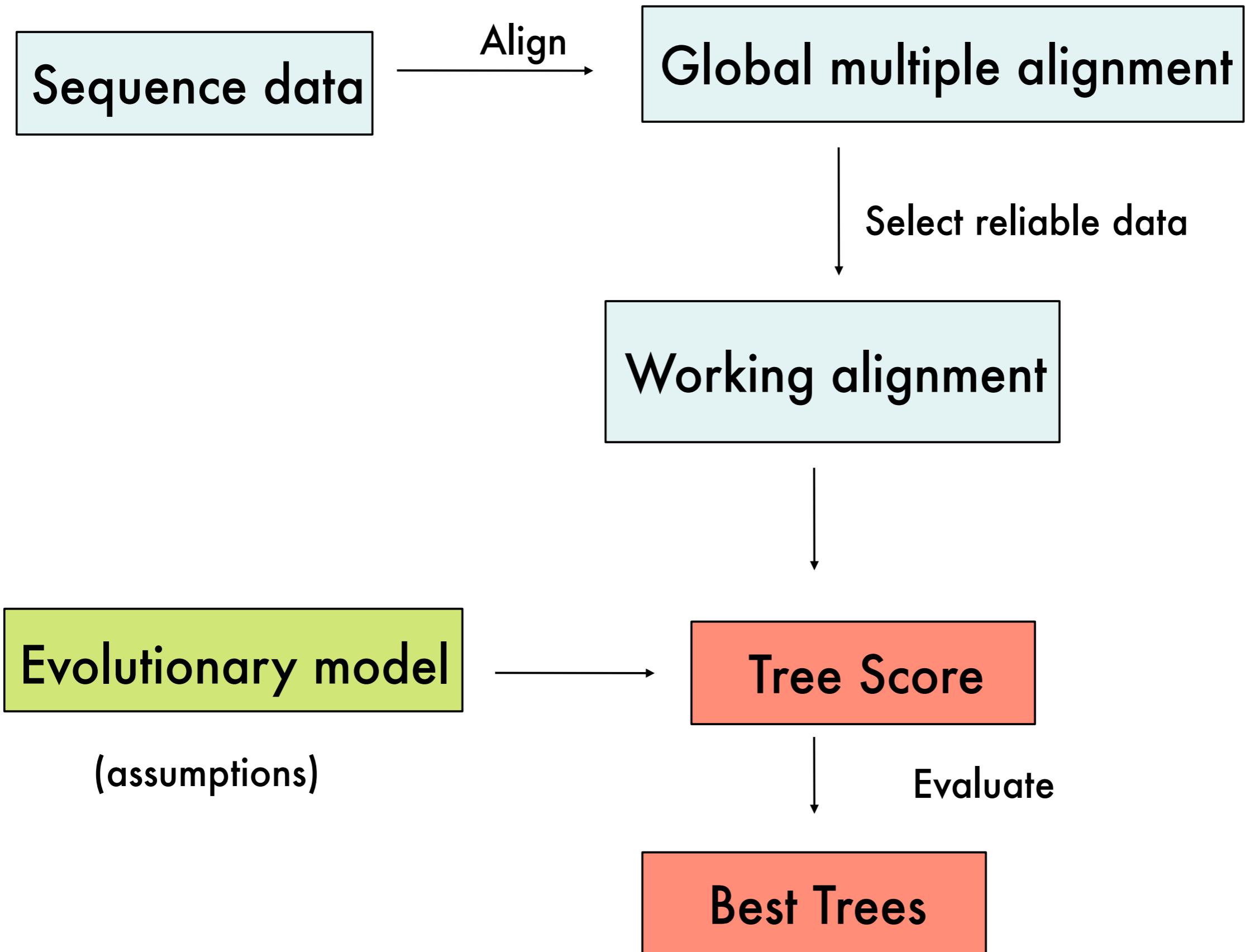


<https://github.com/andand/DEGEPRIME>

# a little more on sequence alignment...

- always good to manually curate alignments
- sequence dependent - obviously harder to align distant things (do you need them?)
- sometimes good to just keep homologous regions (align only conserved domains)
- non-coding regions, repetitive, or regulatory regions hard to align generally

\*



\*

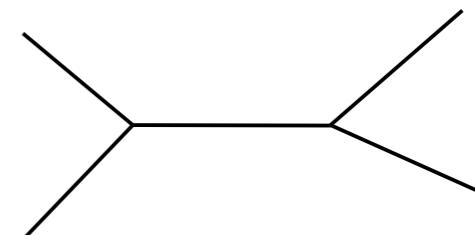
# Parsimony methods:

- Align homologous sequences
- Search for tree topology that requires the smallest number of mutations to explain observed differences among species (fewest changes)
- Hint: count changes to make pairwise comparisons

Position #

2 45 8

Organism A: 5' ...GGTTCGATGT...3'  
Organism B: 5' ...GGATCGAAGT...3'  
Organism C: 5' ...GCTAGGAAGT...3'  
Organism D: 5' ...GCGAGGATGT...3'



\*

# Finding the “best” tree:

- Need to SEARCH for best tree
- Exhaustive
- Branch and bound
- Heuristic - evaluate as many trees as possible, but cannot assure best has been found.
- Use some “optimality criterion” (a way to score best trees)

# Need to do “heuristic” searching

## 1. Exhaustive

Calculate the number of tree possibilities for n taxa=

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

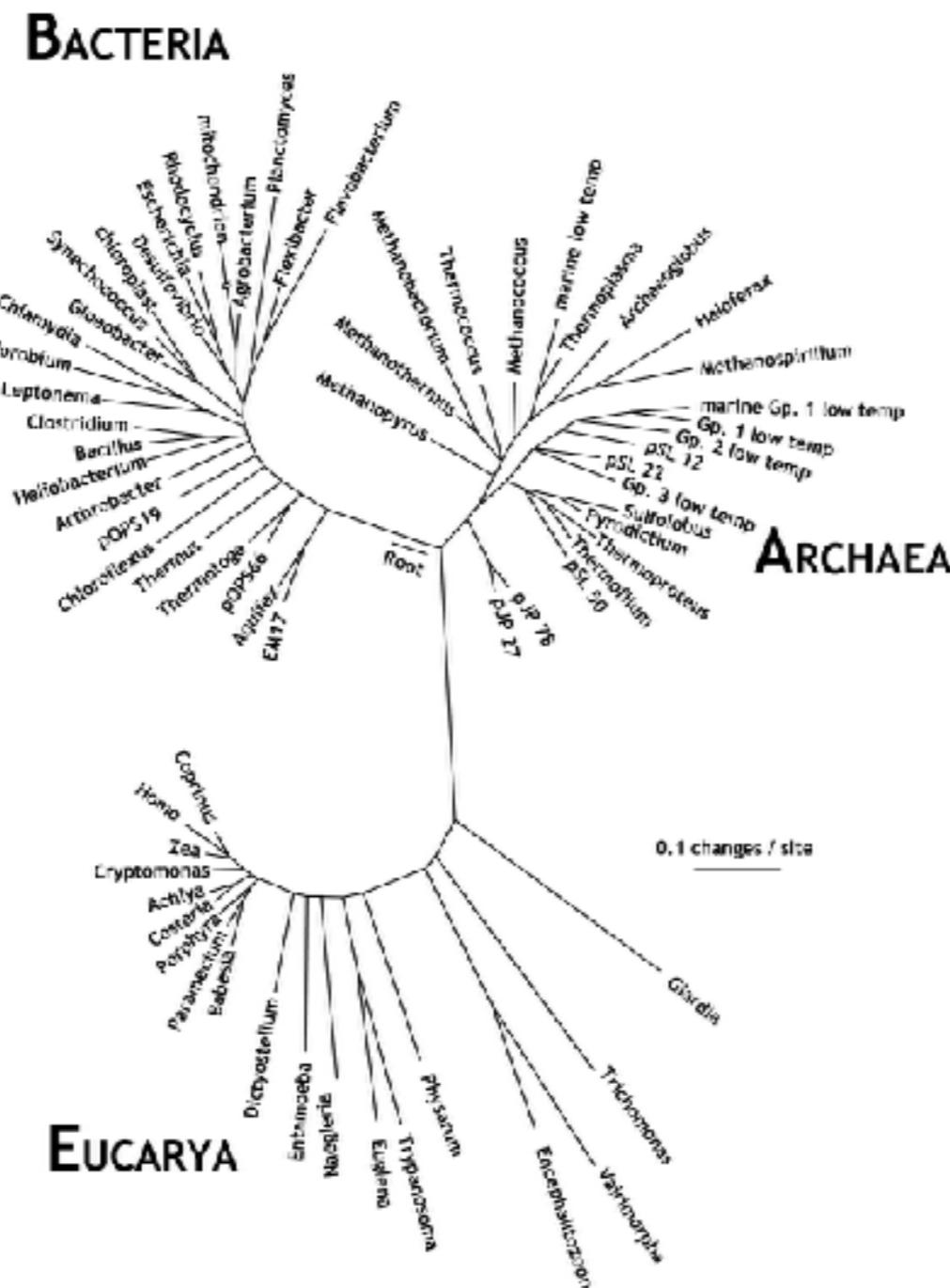
1, 3, 15, 105, 945, 10395....

## 2. Branch and bound: eliminates clearly inferior trees

## 3. Heuristic: Concentrates search “near” previously-found good trees, uses branch swapping, pruning, re-grafting

\*please don't memorize this unless you really want to...

# How to get the best tree?



# It's all about the data...

More nucleotides or amino acids?

More gene sequences from more taxa?

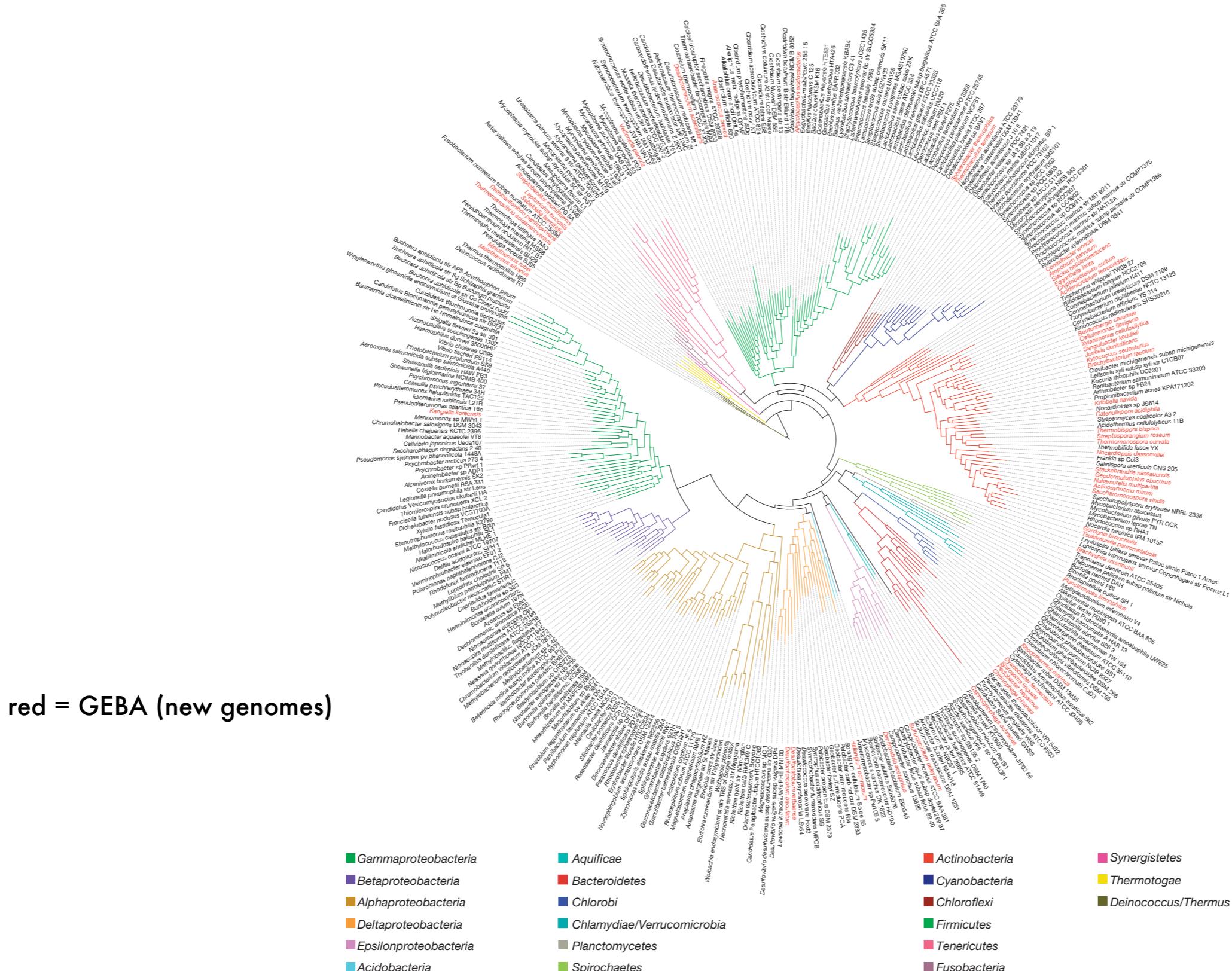
More kinds of genes from diverse taxa?

More genomes from diverse taxa?

# How to improve phylogeny ?

- Increased taxonomic sampling  
(break up long branches)  
\*\* add more organisms
- Concatenated gene phylogenies  
\*\*add more data (meaning more genes)
- Multicharacter data sets (gene phylogenies, morphology, etc.)  
\*\* add more non-molecular data

# Bacterial tree using 31 broadly conserved concatenated protein-coding genes



**Figure 1 | Maximum-likelihood phylogenetic tree of the bacterial domain based on a concatenated alignment of 31 broadly conserved protein-coding genes<sup>16</sup>. Phyla are distinguished by colour of the branch and GEBA genomes are indicated in red in the outer circle of species names.**

# *Genome-enabled strategies for determining evolutionary relationships*

- consensus phylogenies of single genes
- consensus phylogenies of many genes
- concatenated gene phylogenies

do all genes have similar evolutionary histories? (i.e., are all genes good proxies?)

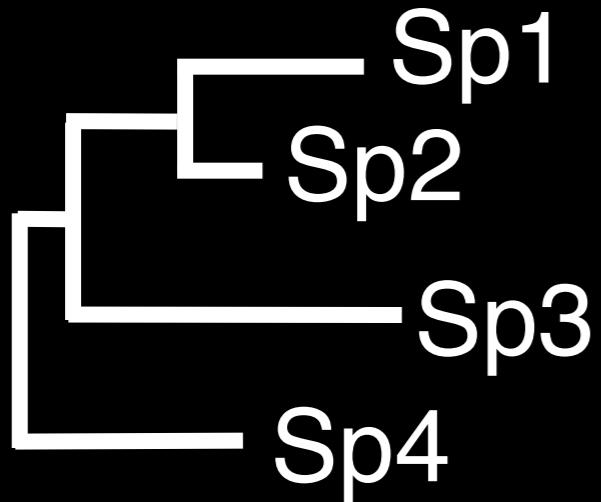
# Multiple genes

# separate analysis

## Gene 1

Sp1 : TCTGT...AAC  
Sp2 : TCTGC...GAC  
Sp3 : CTTAT...GAT  
Sp4 : CCTAT...GAT

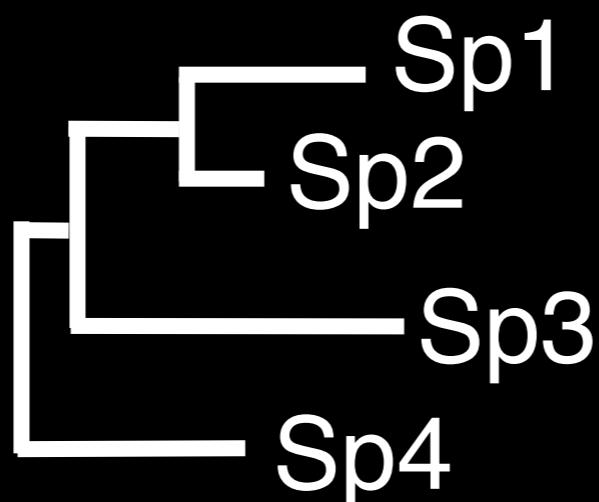
## Evolutionary model



## Gene 2

Sp1 : TCTTT...GAA  
Sp2 : TCGCT...GGA  
Sp3 : CTATT...GGA  
Sp4 : CCATT...GGA

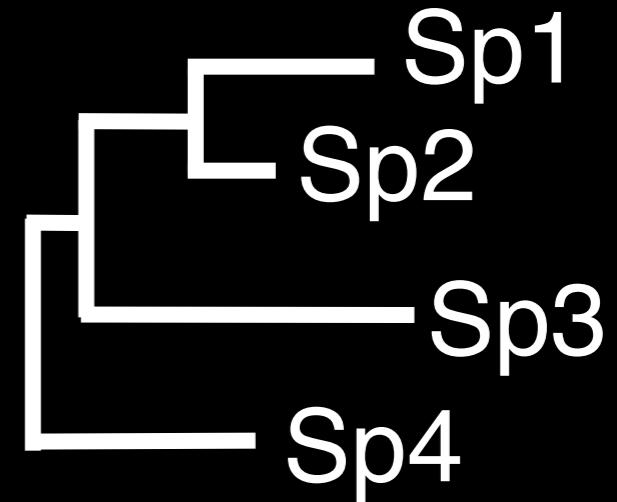
## Evolutionary model



## Gene 3

Sp1 : TCGTT...GCC  
Sp2 : ACGCT...CCC  
Sp3 : ATATT...CGA  
Sp4 : CCATT...CCA

## Evolutionary model



e.g., Murphy *et al.* (2001)

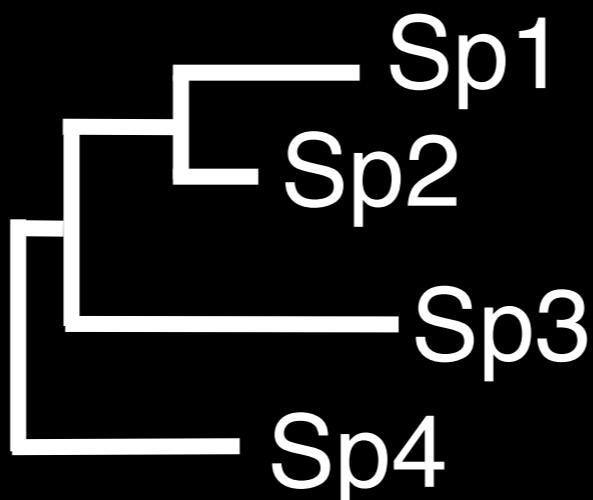
# Multiple genes

# concatenate analysis

## Gene 1 + Gene 2 + Gene 3

Sp1 : TCTGT...AACTCTTT...GAATCGTT...GCC  
Sp2 : TCTGC...GACTCGCT...GGAACGCT...CCC  
Sp3 : CTTAT...GATCTATT...GGAATATT...CGA  
Sp4 : CCTAT...GATCCATT...GGACCATT...CCA

## Evolutionary model



e.g., Murphy *et al.* (2001)

# Current Biology

## Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling

### Highlights

- Two new archaeal phyla were defined within a recently suggested superphylum
- Complete and closed genomes have been reconstructed for uncultured Archaea
- Uncultured Archaea with small genomes have limited metabolic capabilities
- Novel genome compositions suggest primary roles in carbon and/or hydrogen cycling

### Authors

Cindy J. Castelle, Kelly C. Wrighton, ..., Kenneth H. Williams, Jillian F. Banfield

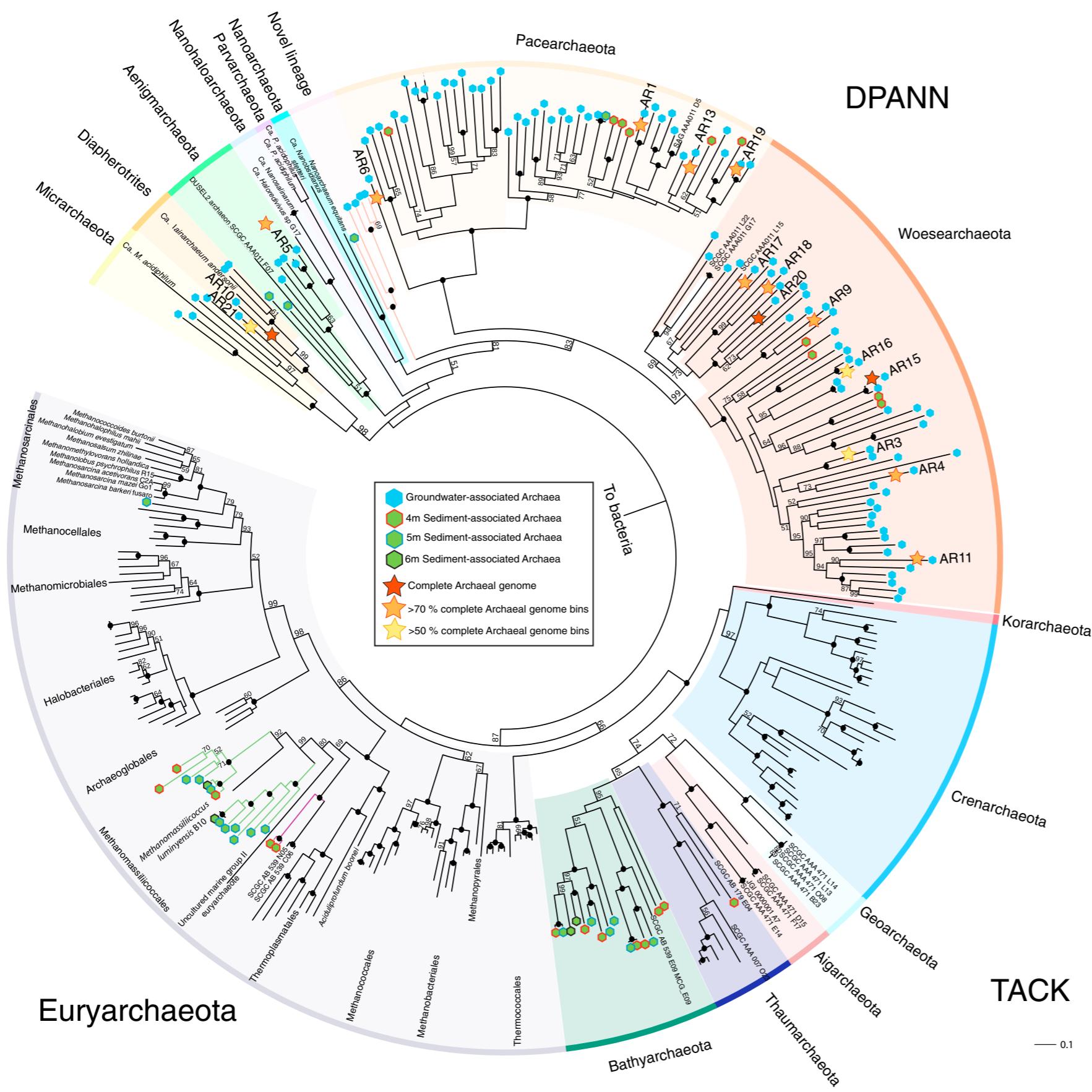
### Correspondence

jbanfield@berkeley.edu

### In Brief

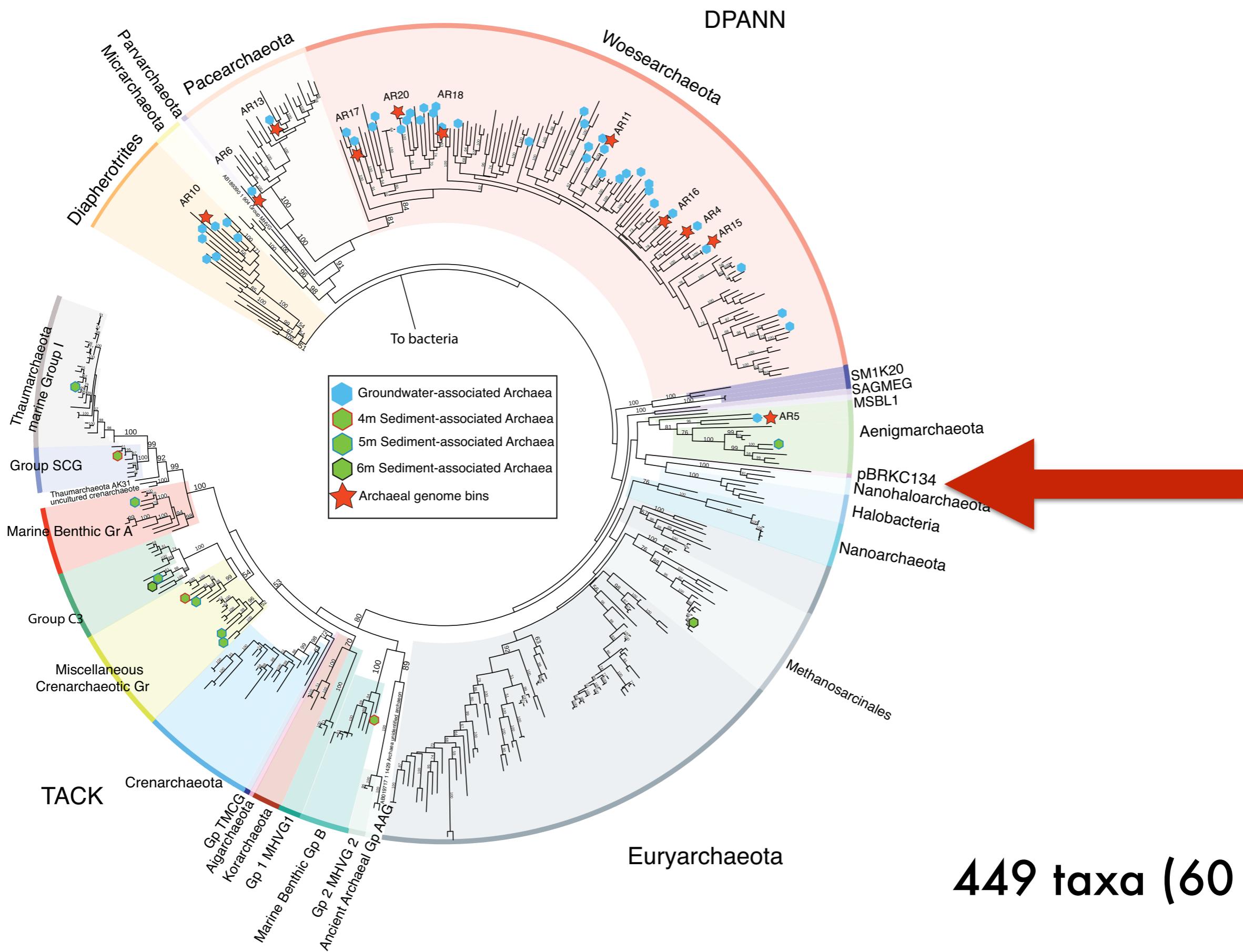
Castelle et al. expand understanding of archaeal diversity by genomically sampling a little-studied branch of the domain. They resolve two new phyla within a major superphylum radiation. The organisms have small genomes and limited metabolic capacities; their primary biogeochemical impact appears to be on anaerobic carbon and hydrogen cycles.

# concatenated ribosomal protein phylogeny



153 archaeal taxa  
15 proteins

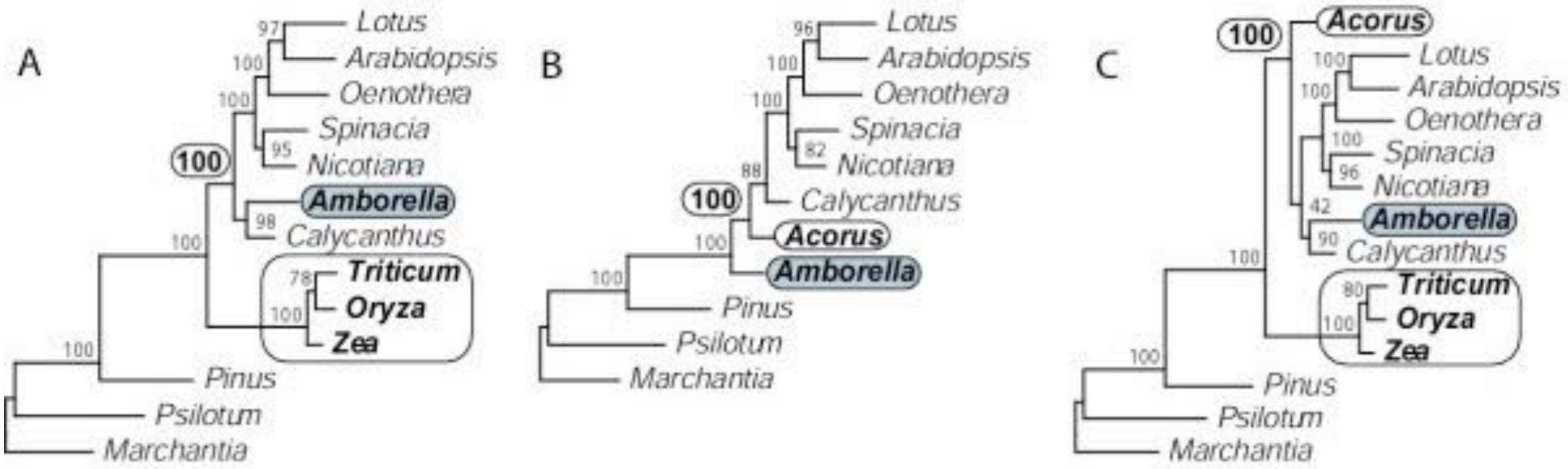
# ssu rRNA phylogeny



LOCUS AF118669 967 bp DNA linear ENV 07-MAY-2004  
DEFINITION Uncultured euryarchaeote pBRKC134 small subunit ribosomal RNA gene, partial sequence.  
ACCESSION AF118669  
VERSION AF118669.1 GI:4580659  
KEYWORDS ENV.  
SOURCE uncultured euryarchaeote pBRKC134  
ORGANISM uncultured euryarchaeote pBRKC134  
Archaea; Euryarchaeota; environmental samples.  
REFERENCE 1 (bases 1 to 967)  
AUTHORS Dawson, S.C. and Pace, N.R.  
TITLE Mesophilic crenarchaeota diversity in a brackish, sulfide-rich marsh sediment  
JOURNAL Unpublished  
REFERENCE 2 (bases 1 to 967)  
AUTHORS Dawson, S.C. and Pace, N.R.  
TITLE Direct Submission  
JOURNAL Submitted (08-JAN-1999) Molecular and Cellular Biology, University of California-Berkeley, 371 Koshland Hall, Berkeley, CA 94720, USA  
FEATURES Location/Qualifiers  
source 1..967  
/organism="uncultured euryarchaeote pBRKC134"  
/mol\_type="genomic DNA"  
/db\_xref="taxon:91312"  
/clone="pBRKC134"  
/environmental\_sample  
/country="USA: California, Berkeley, Berekely Aquatic Park"  
/note="anoxic sediment"  
rRNA <1..>967  
/product="small subunit ribosomal RNA"  
ORIGIN  
1 atagagtgg tgcattgatt atttggctta aagtgtgcgt agcaggaaat tctaagtcc  
61 ctgtgaaatg tcaggtctta agcttgacac gtgcagagga tactggaatc cttggactg  
121 ggataagggt aatgtacggt ttggtaggg gtgaaatccg ataatccaag ccggacaacc  
181 ggtggcgaag gcgttacact ggaacaggtc cgactgtgag gcacgaaagc tacggtcgc  
241 agccggatta catacccgaa tagtcctatg ctgtaaacga tgctcaactaa gtgttaggta  
301 atgcaagaca ttctcttagtg ccgaagcgaa agcgtaagt gagccgcctg ggaagtacgt  
361 tcgcaagaaa tgaaacttaa aggaattggc gggggcctac tacaagaagt ggagcctgcg

\*

# Taxon sampling is a big deal

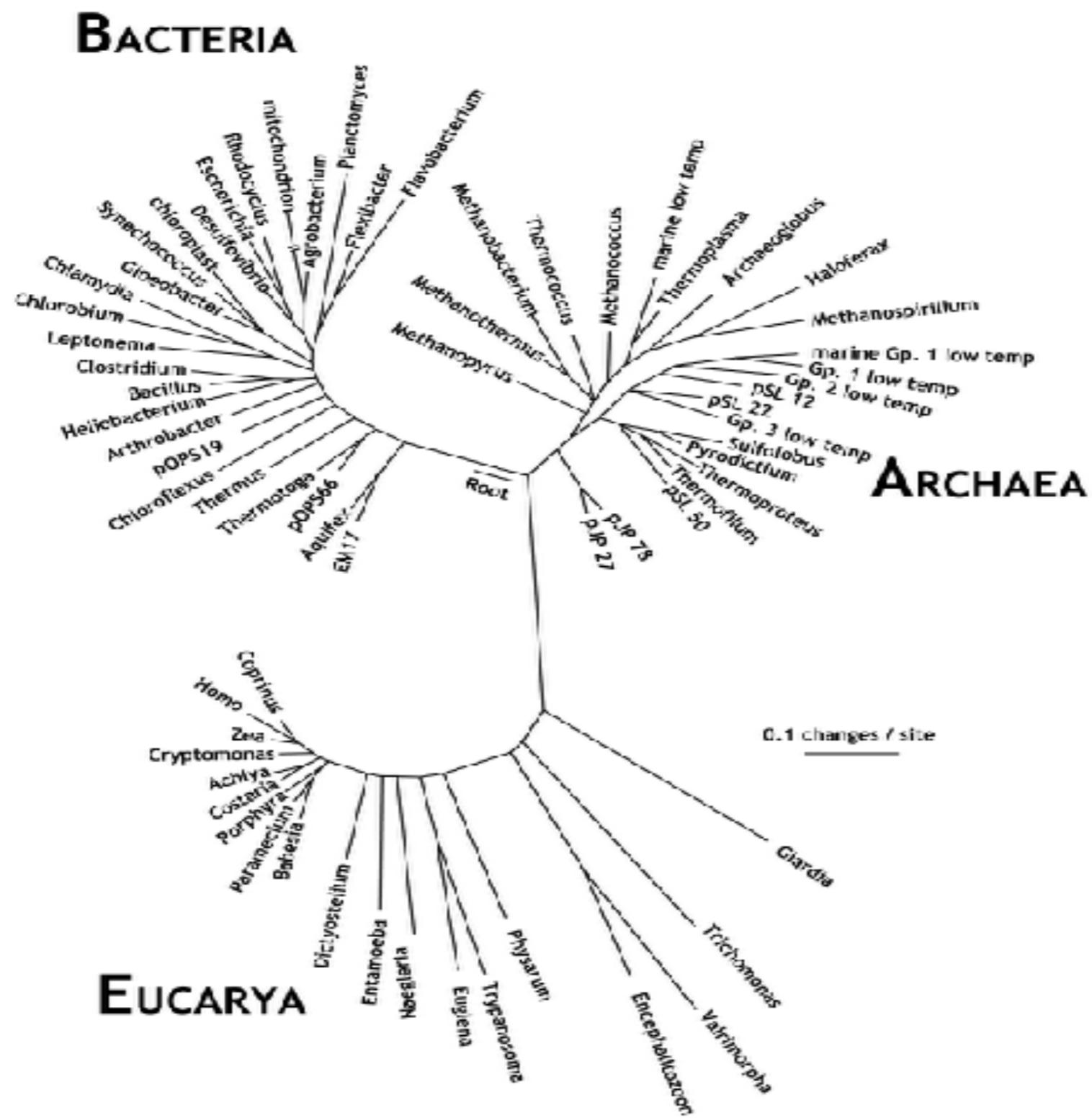


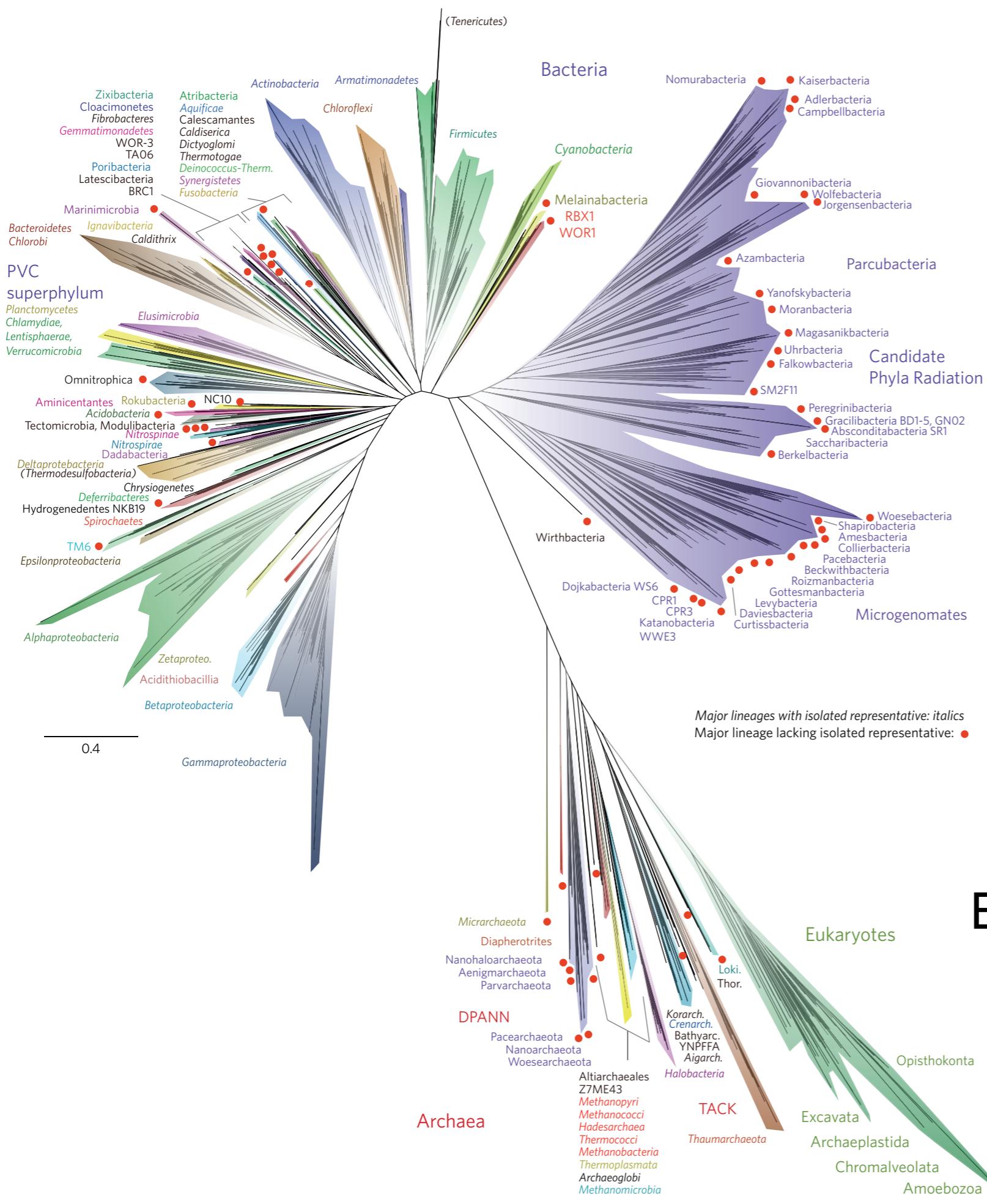
What is the closest relative of Amborella in each tree?  
(each is Maximum parsimony, different taxa)

# Taxonomic sampling= more variation

- In general, more taxa (diverse) better trees
- less taxa has negative impact on the robustness of phylogenies
- affects the number of groups/relationships among groups
- causes trouble in sorting out gene families

# Is this the best map of life?

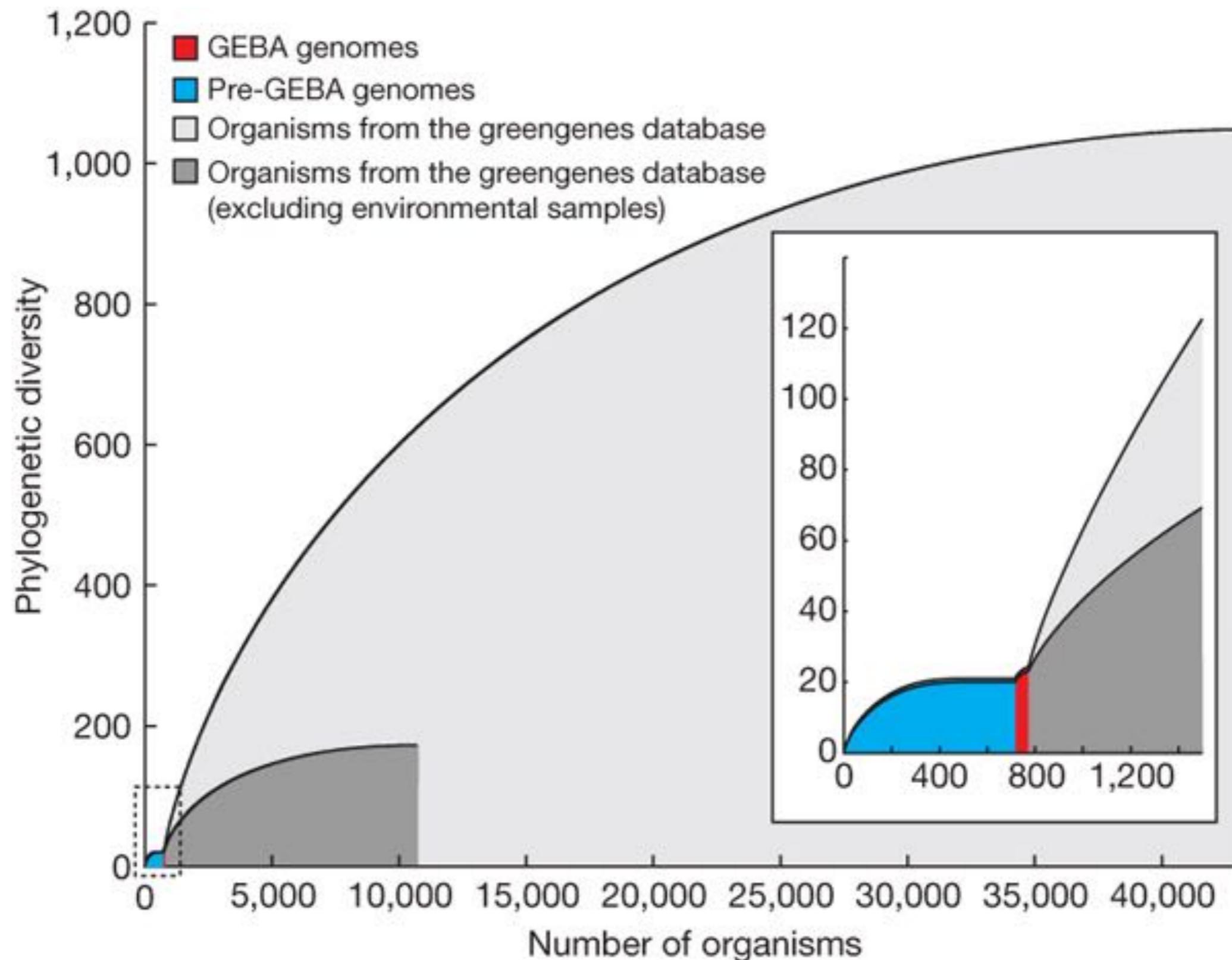




# What's the scale?

# Euks are archaea?

# Phylogenetic diversity of bacteria and archaea on the basis of SSU rRNA genes.





<https://www.arb-silva.de>



## SILVA SSU / LSU 123 - full release

	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref	LSU Ref NR
Minimal length	300	1200/900	1200/900	300	1200/900	1200/900
Quality filtering	basic	strong	strong	basic	strong	strong
Guide Tree	no	no	yes	no	no	yes
Release date	23.07.15	23.07.15	23.07.15	23.07.15	23.07.15	23.07.15
Aligned rRNA sequences	4,985,791	1,756,783	597,607	4,985,791	1,756,783	597,607

C <https://rdp.cme.msu.edu>

[ABOUT RDP](#) | [ASSIGNMENT GENERATOR](#) | [CITATION](#) | [CONTACTS](#) | [RELATED SITES](#) | [RESOURCES](#) | [TUTORIALS](#) | [USER WIKI](#)





**ANNOUNCEMENTS**

RDP News

**06/30/2016 RDP Classifier Updates**  
The Classifier 16S training set and Fungal ITS Warcup set have been updated

**06/03/2016 RDP staff on the road!**  
Teaching in China, Genomic Standards Consortium meeting in Crete, special ASM Microbe events in Boston

**10/07/2015 Xander assembler article is published.**  
Xander: Employing a Novel Method for Efficient Gene-Targeted Metagenomic Assembly

**10/07/2015 Warcup Fungal ITS article is accepted!**  
Fungal identification using a Bayesian Classifier and the 'Warcup' training set of Internal Transcribed Spacer sequences.

**07/08/2015 \*\*\* Pyro Job Submission up \*\*\***  
Hardware issues causing pyro issues now fixed

**05/28/2015 RDP Staff attending ASM Meeting in New Orleans**  
RDP staff will be attending the ASM General Meeting in New Orleans in the coming week. Two RDP posters will be presented: first on Tuesday morning....

**05/26/2015 RDP Release 11.4 available**  
Updated 16S rRNA hierarchy model to training set No. 14.

**03/27/2015 FrameBot: new option Add de novo to references available**  
Unique abundant query sequences will be added to the starting reference set if qualifications are met.

**02/23/2015 WARNING -- RDP unavailable Sat., March 7th**  
Building network infrastructure upgrades planned 8 A.M. through 6 P.M.

**RDP Release 11, Update 4 :: May 26, 2015**

**3,224,600 16S rRNAs :: 108,901 Fungal 28S rRNAs**  
Find out what's new in RDP Release 11.4 [here](#).

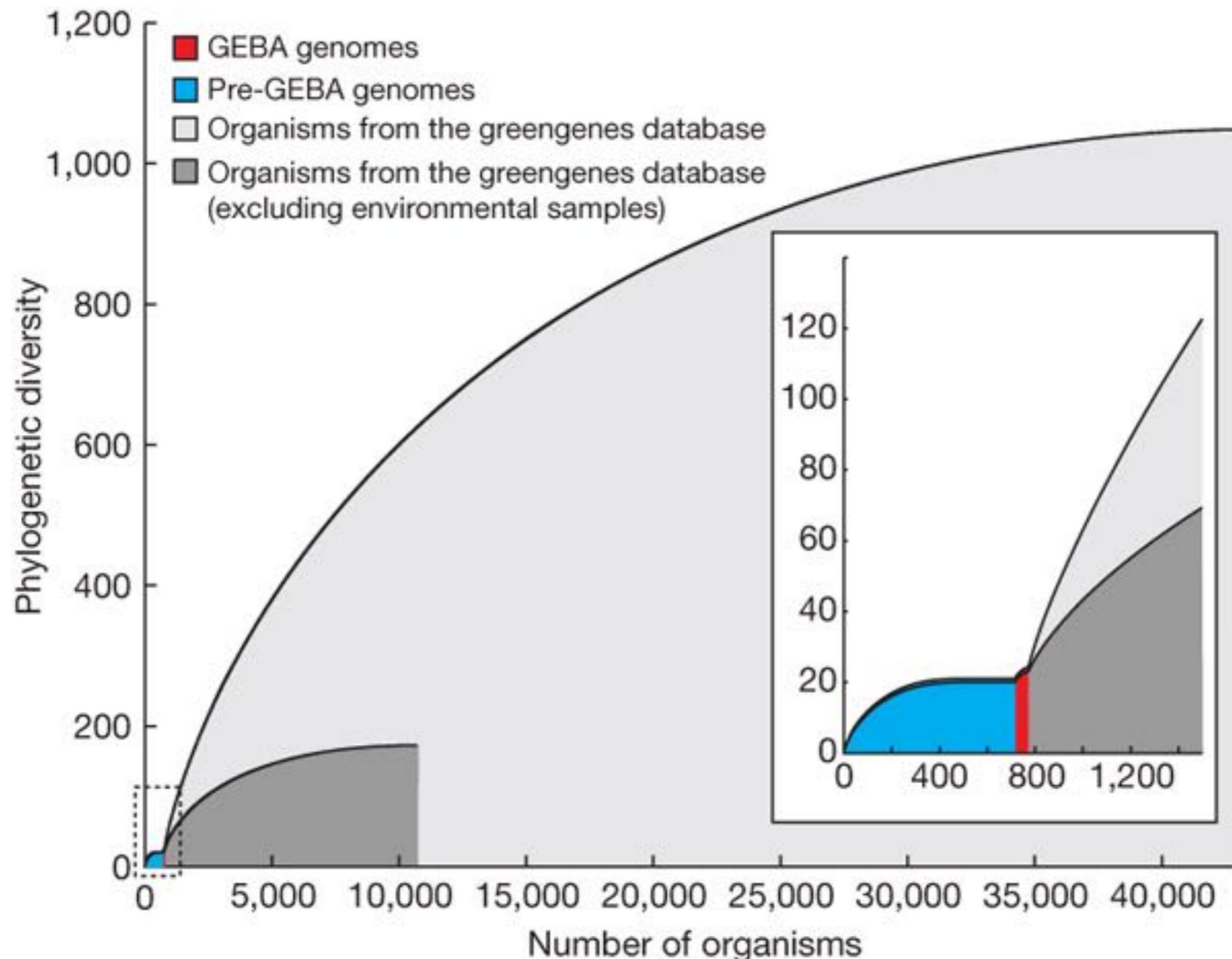
[Cite RDP's latest tool articles.](#)

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPIPeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.



# Phylogenetic diversity of bacteria and archaea on the basis of SSU rRNA genes.





It's all about the bass.

-Megan Trainor (2014)

It's all about that  
(data)'base.

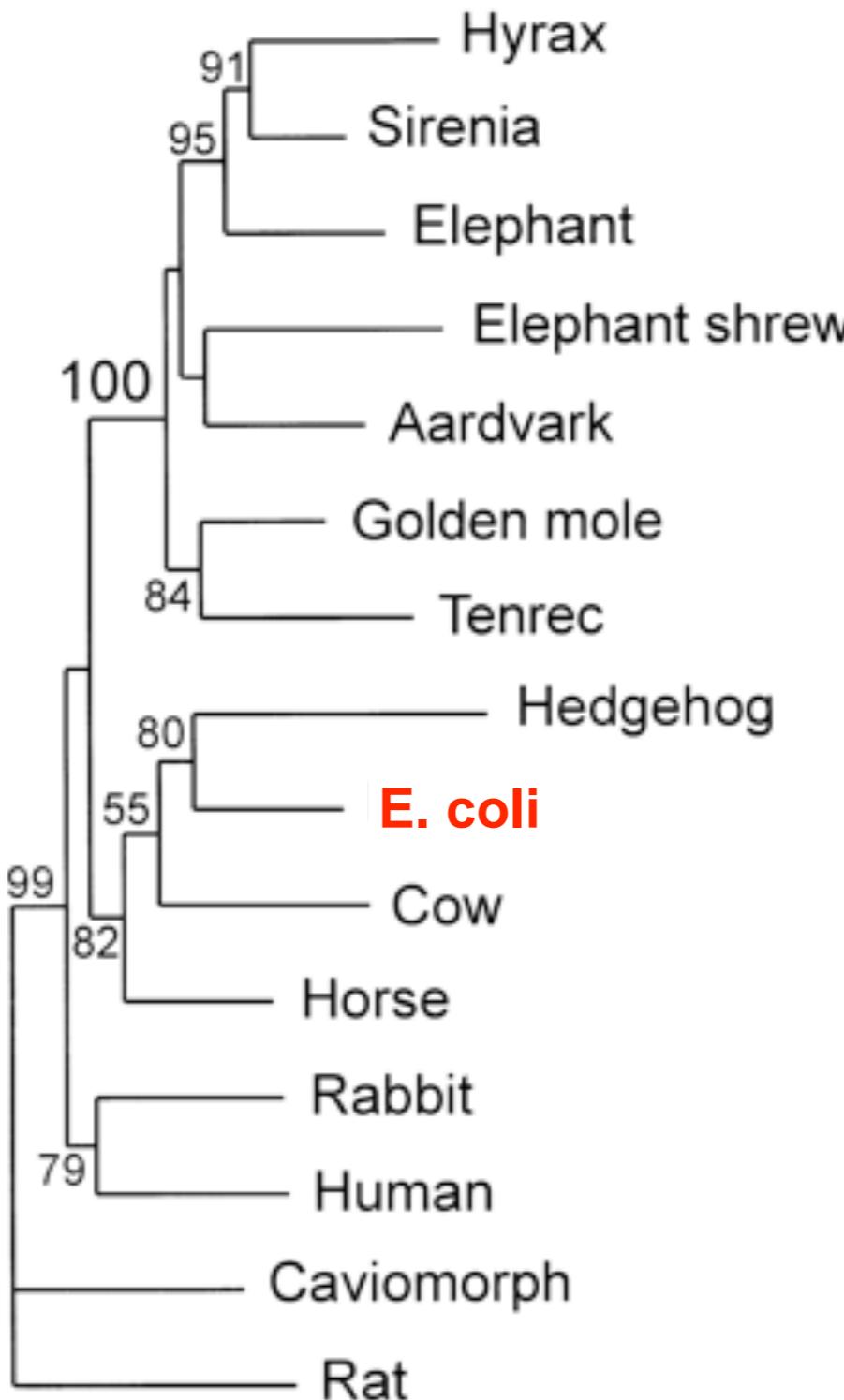
-*Scott Dawson (2015)*



# Horizontal Gene Transfer?

A close-up photograph of a tree trunk, likely a eucalyptus, showing its rough, textured bark and numerous white, fibrous root structures (prop roots) growing downwards. The scene is dimly lit, with dappled sunlight filtering through leaves, creating a natural, organic feel.

# Xenologs: Horizontal gene transfer



Horizontal gene transfer: genes acquired not through common “vertical” ancestry

\*

## Monophyletic tree: Limited lateral transfer

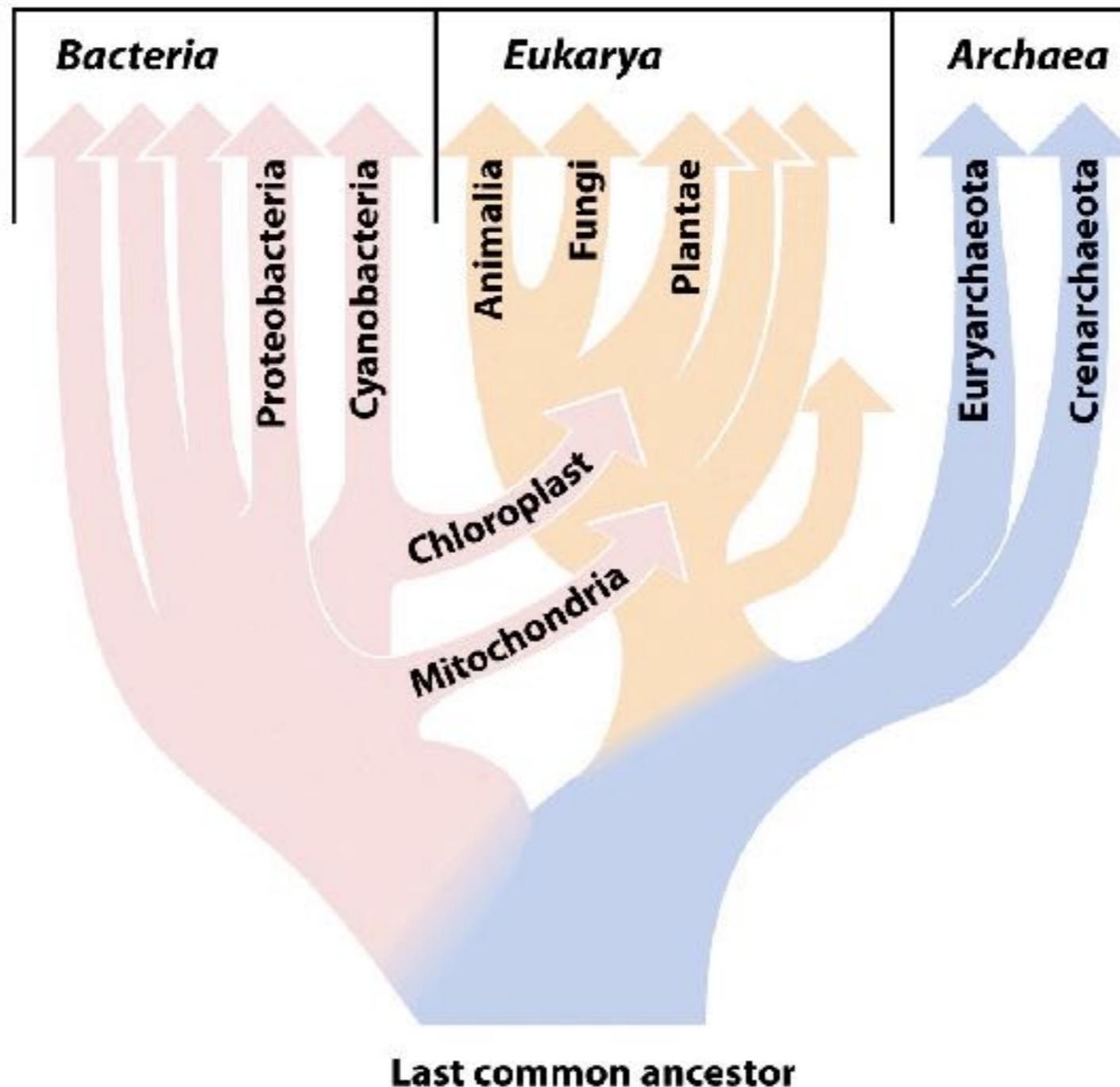
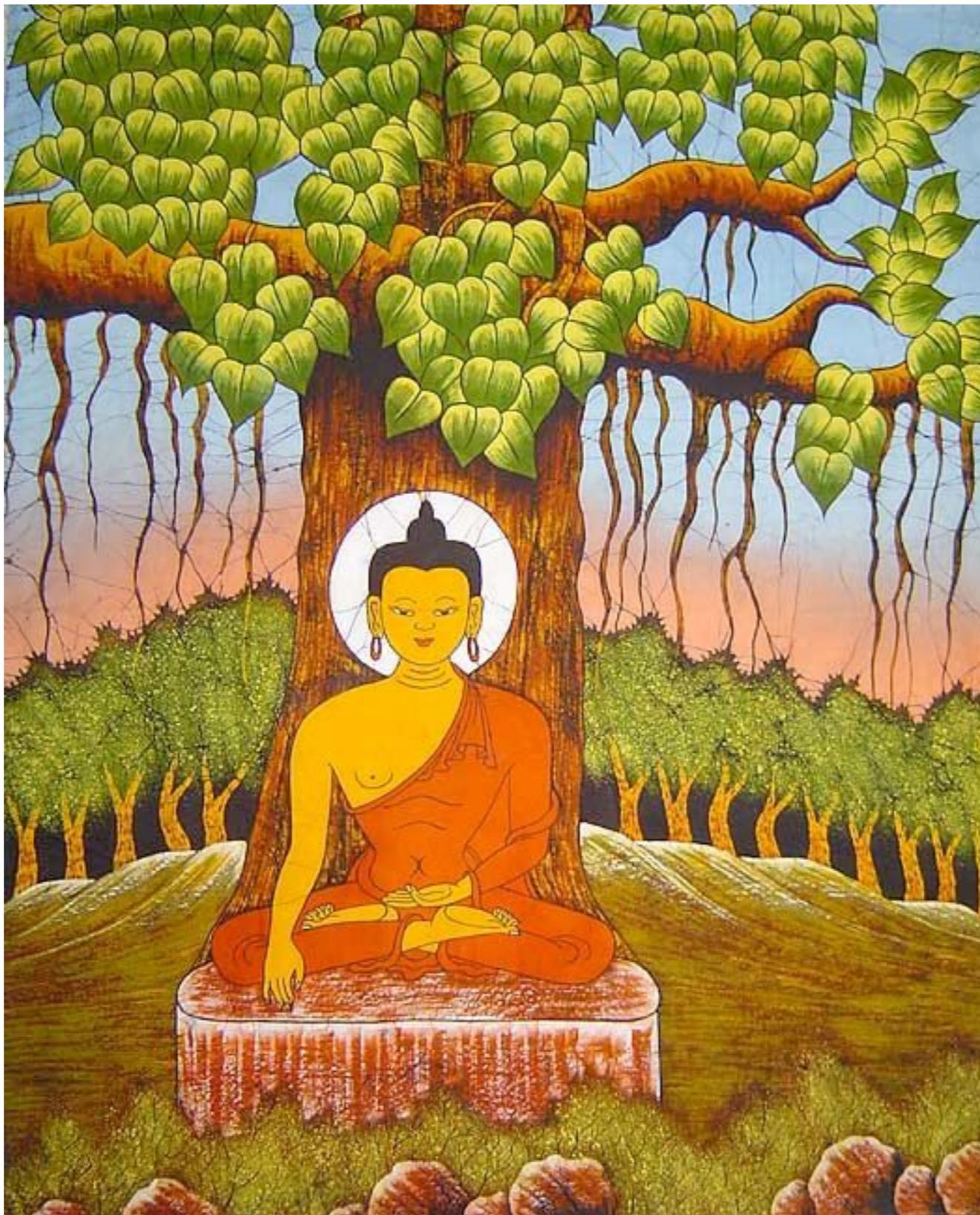


Figure 17.27a Microbiology: An Evolving Science  
Source: W. Ford Doolittle. 1999. *Science* 284:2126.

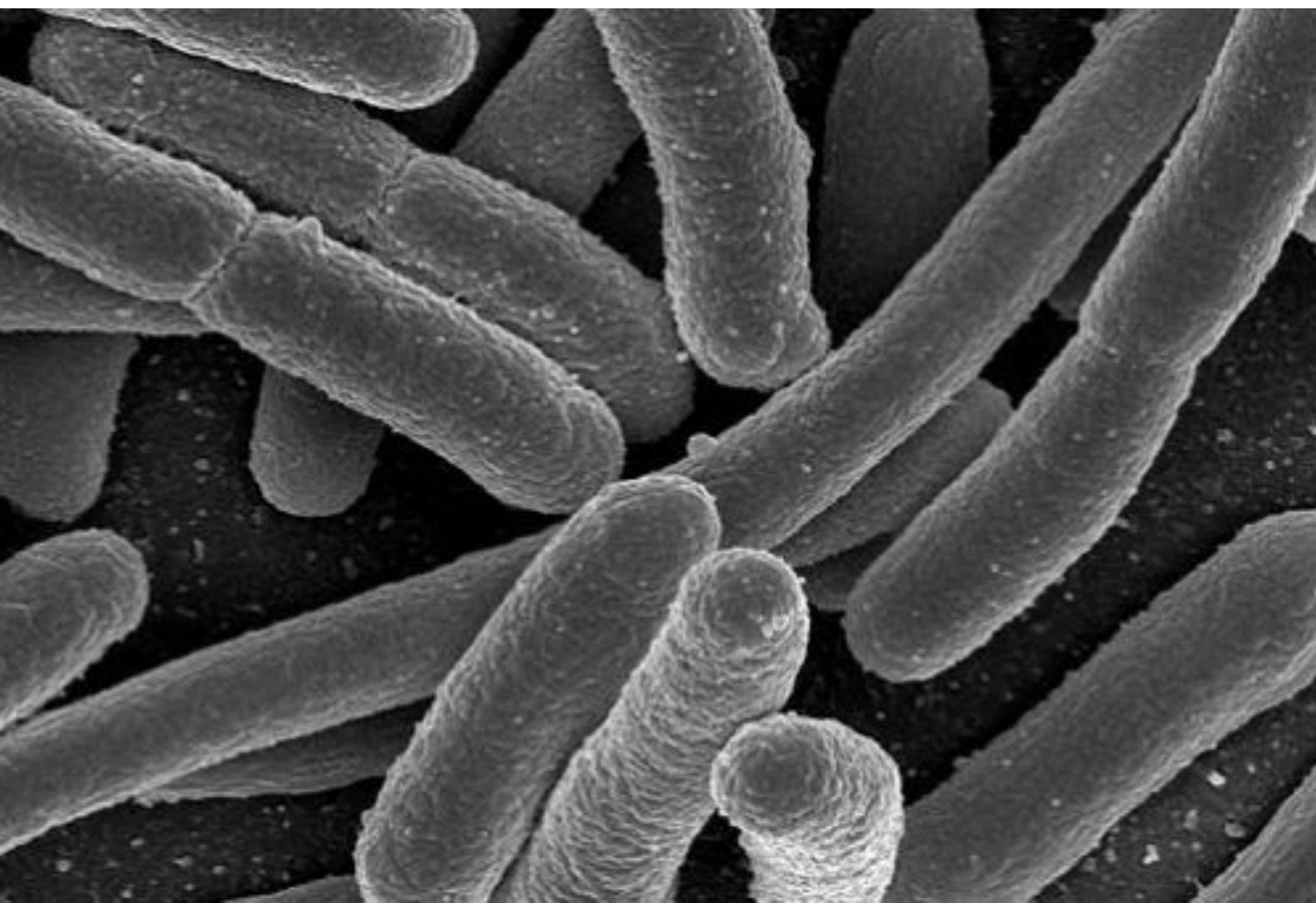
Alpha proteobacterial genes in eukaryotes?  
Cyanobacterial genes in eukaryotes?

# What's a bacterial "species"?

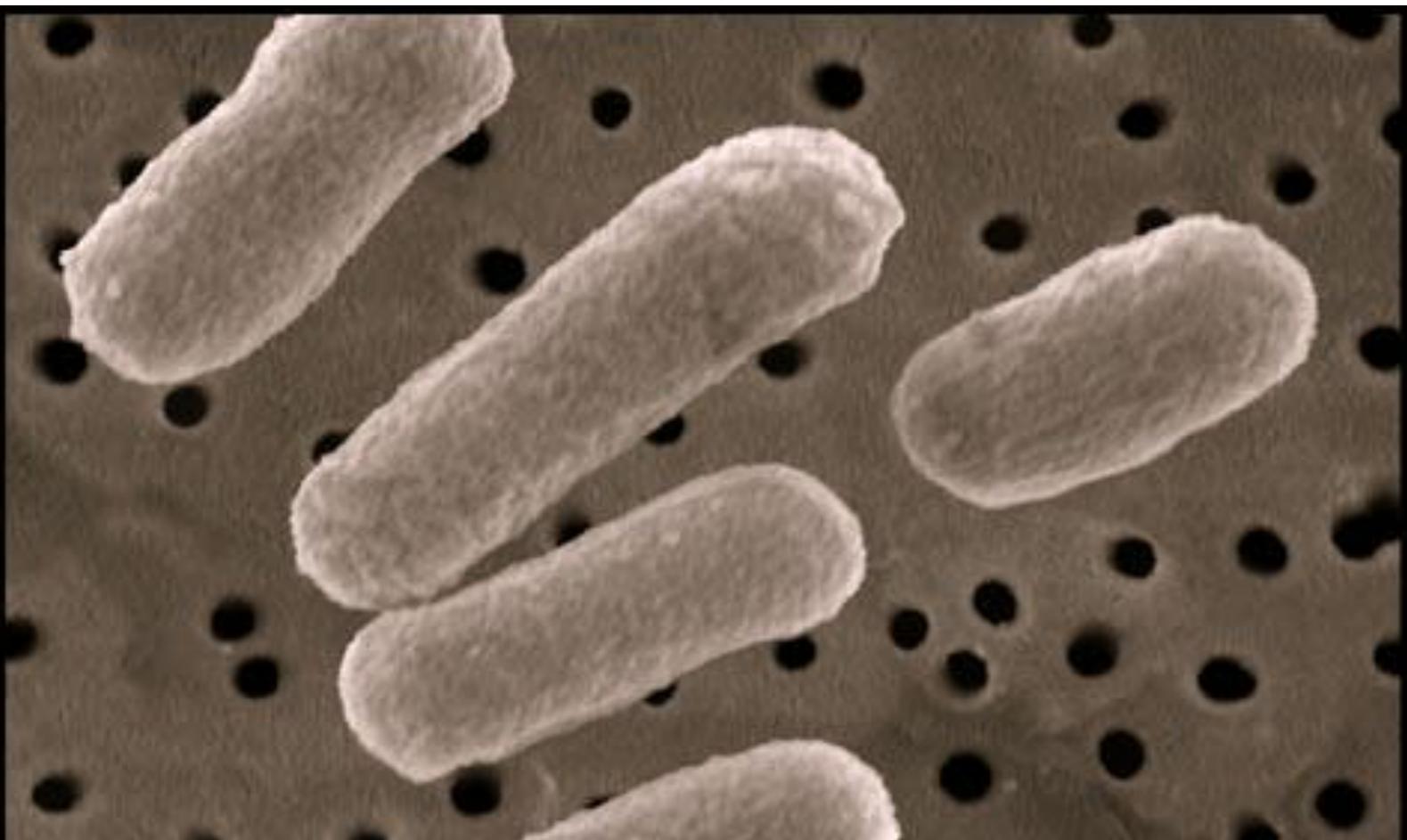


**genotype or phenotype?**

**OTU = 97% identical ssu rRNA**



*Escherichia coli*



*Shigella sonii*

# *E. coli* and *Shigella* bacterial strain characteristics

**Table 1.** Principal characteristics of the 20 *Escherichia coli*/*Shigella* strains and 1 *E. fergusonii* strain.

Strains	Host	Sample	Clinical condition (Pathotype <sup>a</sup> )	Phylogenetic group <sup>b</sup>	Extraintestinal mouse model phenotype <sup>c</sup> (Number of mice killed out of 10)	Genome sequence reference
K-12 MG1655	Human	Faeces	Commensal	A	NK (0)	[115]
K-12 W3110	Human	Faeces	Commensal	A	NK (0)	Nara Institute of Science and Technology
<b>IAI1</b>	<b>Human</b>	<b>Faeces</b>	<b>Commensal</b>	<b>B1</b>	<b>NK (0)</b>	<b>This work</b>
<b>55989</b>	<b>Human</b>	<b>Faeces</b>	<b>Diarrhoea (EAEC)</b>	<b>B1</b>	<b>K (10)</b>	<b>This work</b>
<i>S. boydii</i> 4 227 (Sb 227)	Human	Faeces	Shigellosis	S1	ND <sup>d</sup>	[116]
<i>S. sonnei</i> 046 (Ss 046)	Human	Faeces	Shigellosis	SS	ND	[116]
<i>S. flexneri</i> 2a 301 (Sf 301)	Human	Faeces	Shigellosis	S3	ND	[117]
<i>S. flexneri</i> 2a 2457T (Sf 2457T)	Human	Faeces	Shigellosis	S3	NK (0)	[118]
<i>S. flexneri</i> 5b 8401 (Sf 8401)	Human	Faeces	Shigellosis	S3	ND	[119]
<i>S. dysenteriae</i> 1 197 (Sd 197)	Human	Faeces	Shigellosis	SD1	ND	[116]
O157:H7 EDL933	Human	Faeces	Diarrhoea (EHEC)	E	NK (1)	[120]
O157:H7 Sakai	Human	Faeces	Diarrhoea (EHEC)	E	NK (1)	[121]
<b>UMN026</b>	<b>Human</b>	<b>Urine</b>	<b>Cystitis (ExPEC)</b>	<b>D</b>	<b>K (10)</b>	<b>This work</b>
<b>IAI39</b>	<b>Human</b>	<b>Urine</b>	<b>Pyelonephritis (ExPEC)</b>	<b>D</b>	<b>K (8)</b>	<b>This work</b>
UTI89	Human	Urine	Cystitis (ExPEC)	B2	K (10)	[122]
APEC O1	Chicken	Lung	Colisepticemia (ExPEC)	B2	K (10)	[123]
<b>S88</b>	<b>Human</b>	<b>Cerebro- spinal fluid</b>	<b>New born meningitis (ExPEC)</b>	<b>B2</b>	<b>K (10)</b>	<b>This work</b>
CFT073	Human	Blood	Pyelonephritis (ExPEC)	B2	K (10)	[30]
<b>ED1A</b>	<b>Human</b>	<b>Faeces</b>	<b>Healthy subject</b>	<b>B2</b>	<b>NK (0)</b>	<b>This work</b>
536	Human	Urine	Pyelonephritis (ExPEC)	B2	K (10)	[124]
<i>E. fergusonii</i>	Human	Faeces	Unknown	Outgroup	NK (1)	This work

# *E coli* strain genome characteristics

**Table 2.** General features of the *Escherichia coli* and *E. fergusonii* genomes sequenced in this work with *E. coli* K-12 MG1655 as reference (chromosome features).

Chromosome features	<i>E. coli</i> K-12 MG1655	<i>E. coli</i> strains					<i>E. fergusonii</i> ATCC
		55989	IAI1	ED1a	S88	IAI39	UMN026
Genome Size (bp)	4 639 675	5 154 862	4 700 560	5 209 548	5 032 268	5 132 068	5 202 090
G+C content (%)	50.8	50.7	50.8	50.7	50.7	50.6	50.7
rRNA operons	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)
tRNA genes	86	94	86	91	91	88	88
Total Protein-coding genes <sup>a</sup>	4306	4969	4491	5129	4859	4906	4918
Pseudogenes <sup>b</sup> (nb)	81	79	51	95	90	80	45
Protein coding density <sup>c</sup>	85.7	87.4	87.6	86.2	87	86.1	87.8
Assigned function <sup>d</sup> (%)	80	74	77	74	77	78	76.5
Conserved hypothetical (%)	12.5	23	21.5	23	22	20	22
Orphans (%)	7.5	3	1.5	3	1	2	1.5
IS-like genes (nb)	66	150	42	118	47	224	92
Phage-associated genes (nb)	231	406	201	657	507	393	429

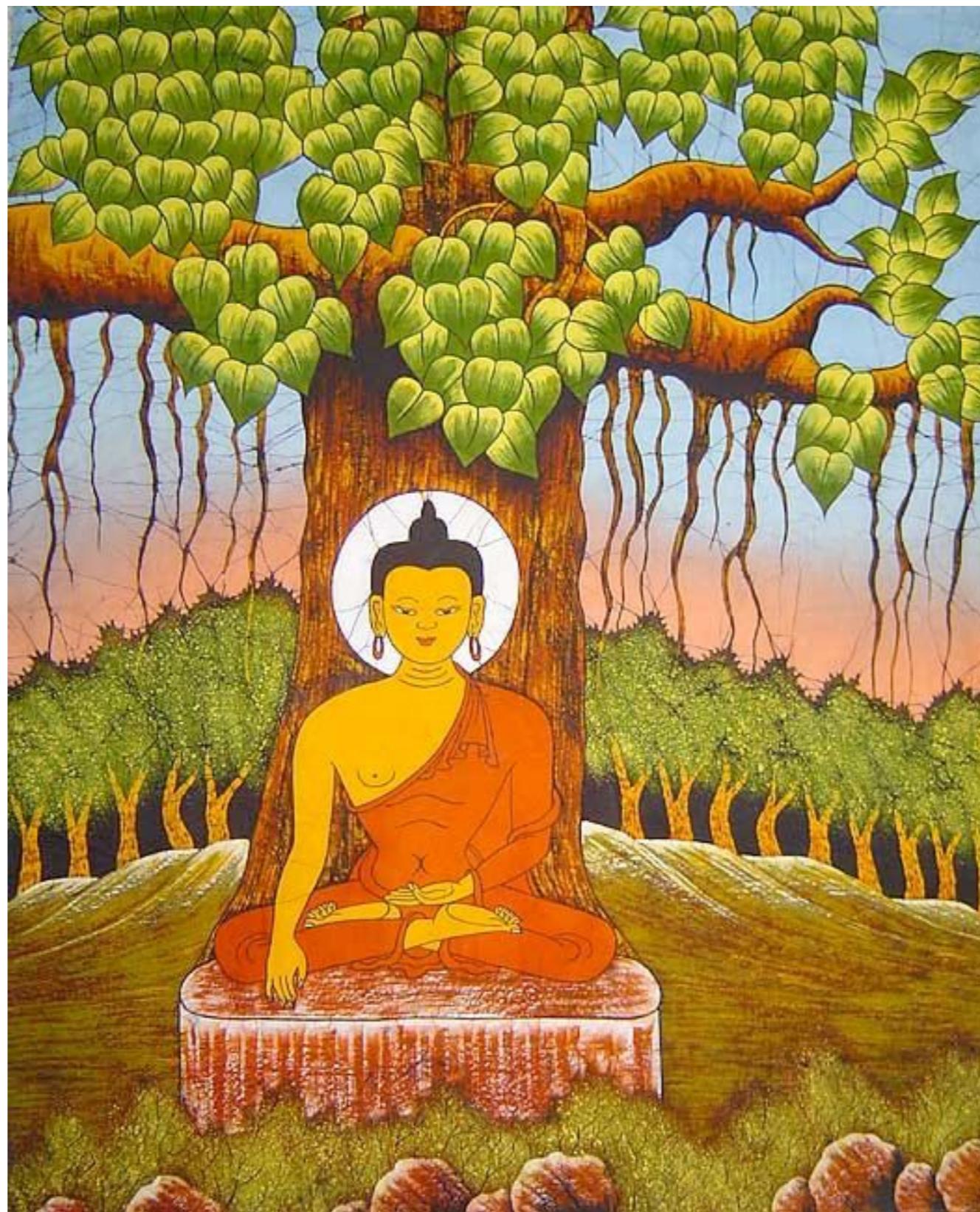
<sup>a</sup>The number of protein-coding genes is given without the number of coding sequences annotated as artificial genes (Supplementary Table 2A).

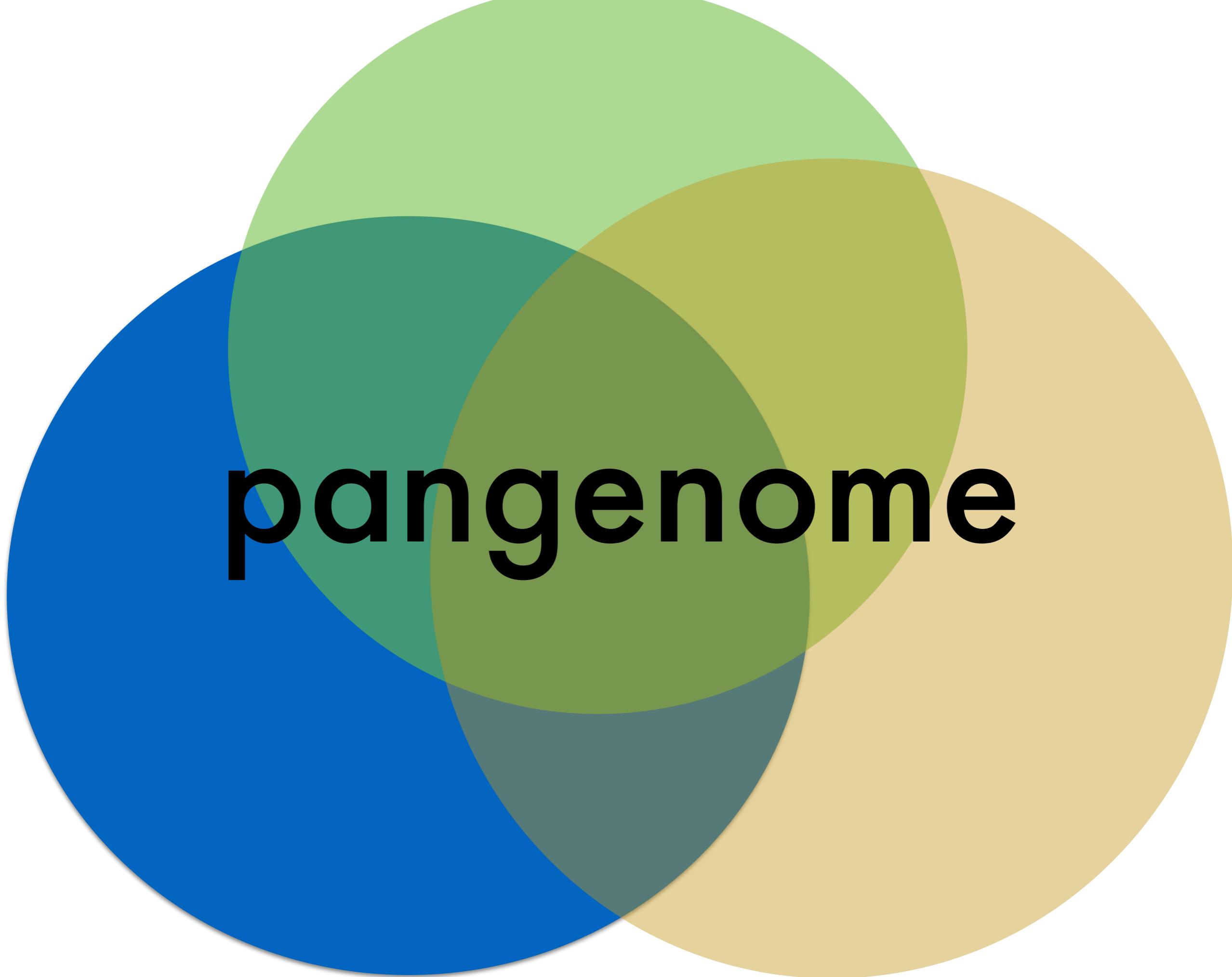
<sup>b</sup>The number of pseudogenes computed for each genome corresponds to the real number of genes that are pseudogenes: one pseudogene can be made of only one CDS (in this case the gene is partial compared to the wild type form in other *E. coli* strains) or of several CDSs (generally two or three CDSs corresponding to the different fragments of the wild type form in other *E. coli* strains). These lists of pseudogenes are available in Supplementary Table 1.

<sup>c</sup>The computed protein coding density takes into account the total length of protein genes excluding overlaps between genes, artifacts, and RNA genes.

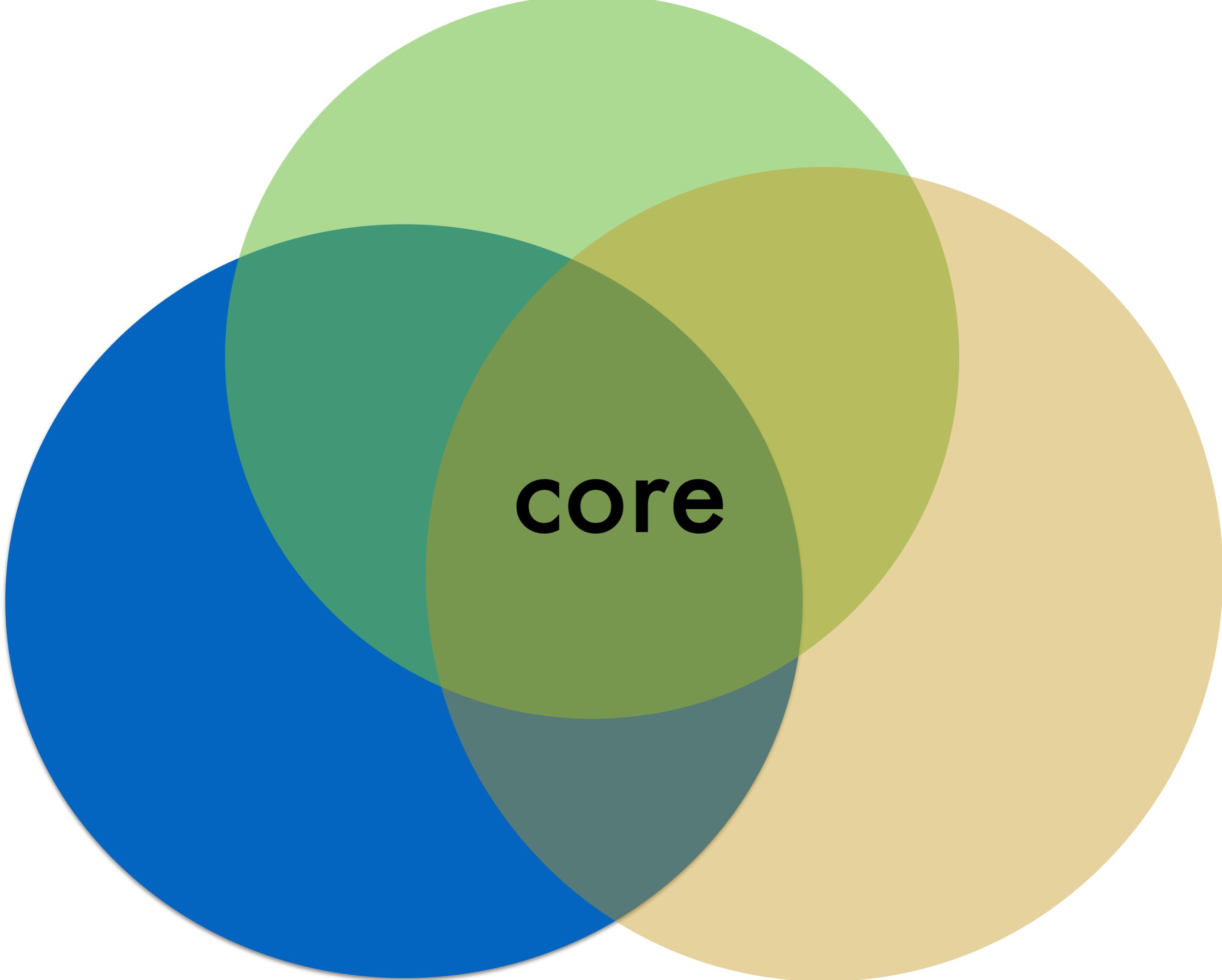
<sup>d</sup>Protein genes with assigned function include the total number of definitive and putative functional assignments.

# Please discuss.





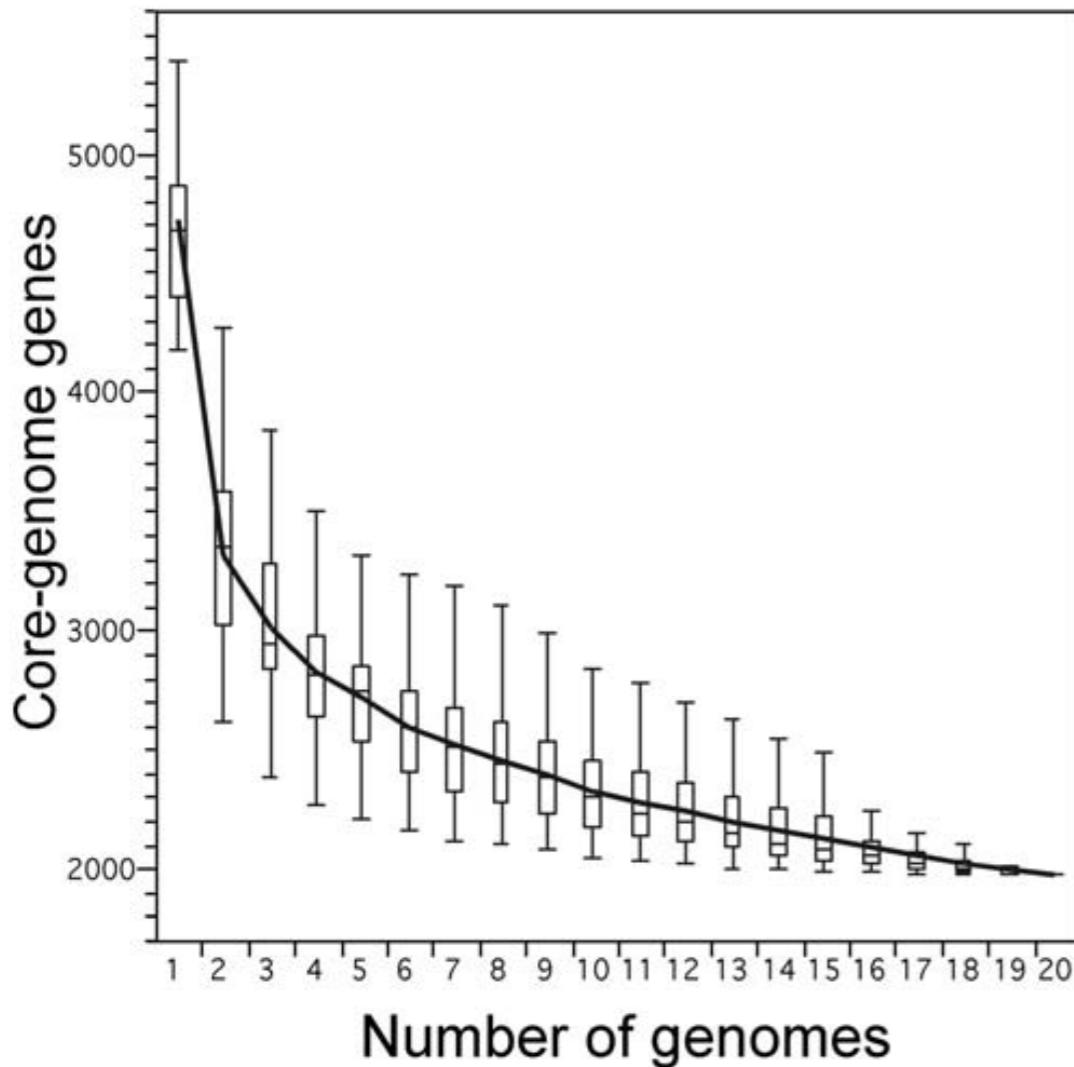
**pangenome**



core

# *Escherichia coli*

## Core genome



**Figure 1. *Escherichia coli* core and pan-genome evolution according to the number of sequenced genomes.** Number of genes in common (left) and total number of non-orthologous genes (right) for a given number of genomes analysed for the different strains of *E. coli*. The upper and lower edges of the boxes indicate the first quartile (25th percentile of the data) and third quartile (75th percentile), respectively, of 1000 random different input orders of the genomes. The central horizontal line indicates the sample median (50th percentile). The central vertical lines extend from each box as far as the data extend, to a distance of at most 1.5 interquartile ranges (i.e., the distance between the first and third quartile values). At 20 sequenced genomes, the core-genome had 1976 genes (11% of the pan-genome), whereas the pan-genome had (i) 17 838 total genes (black), (ii) 11 432 genes (red) with no strong relation of homology (<80% similarity in sequence), and (iii) 10 131 genes (blue) after removing insertion sequence-like elements (3834, 21% of all genes) and prophage-like elements (3873, 22% of all genes).

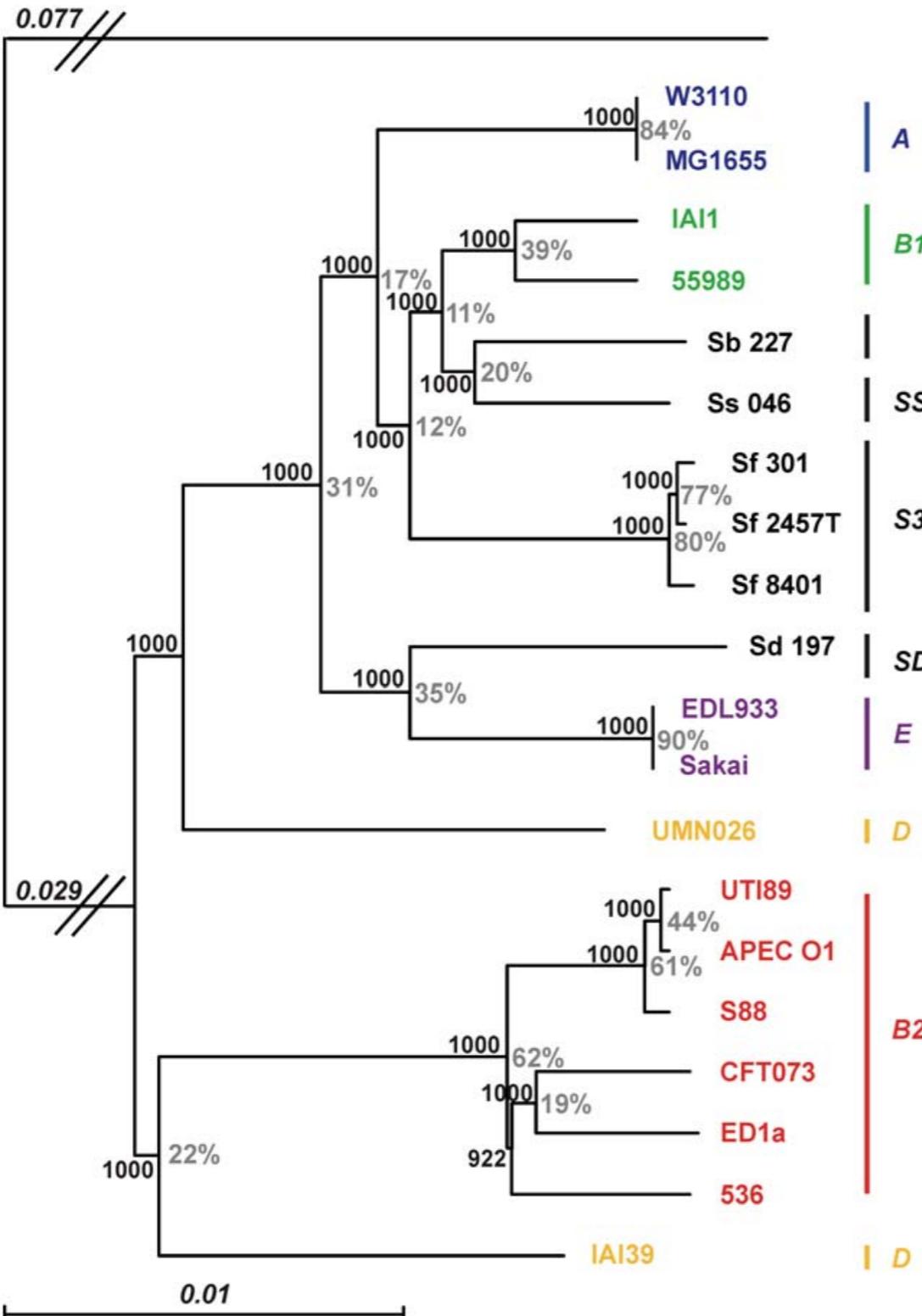
doi:10.1371/journal.pgen.1000344.g001

**core genome** = set of genes shared by all  
the strains of the same bacterial species

(tend to not be horizontally transferred)

# *Escherichia coli*

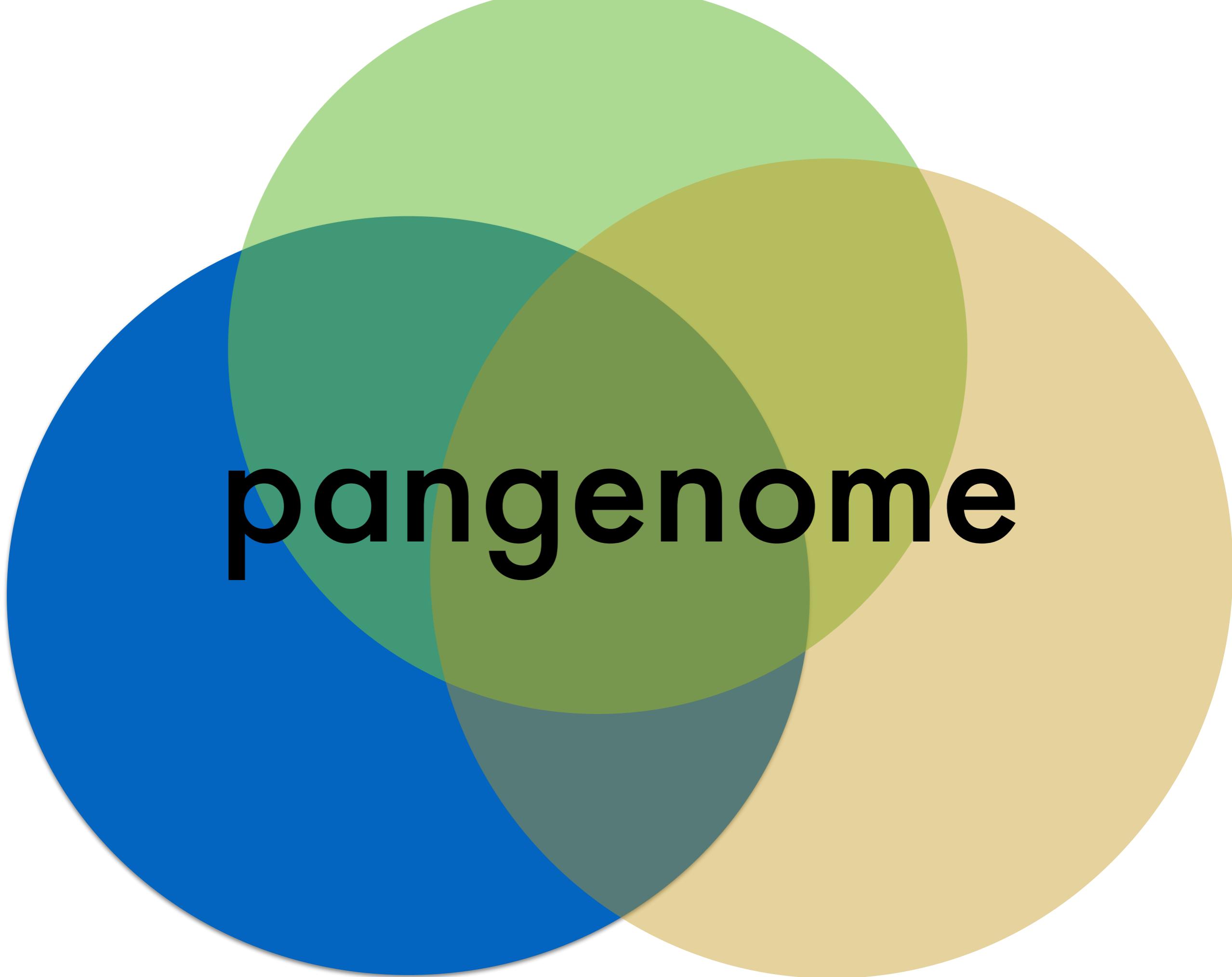
1878 gene  
Core genome phylogeny



bootstrap values (1000)

% = consensus strength  
(number of genes confirming that node)

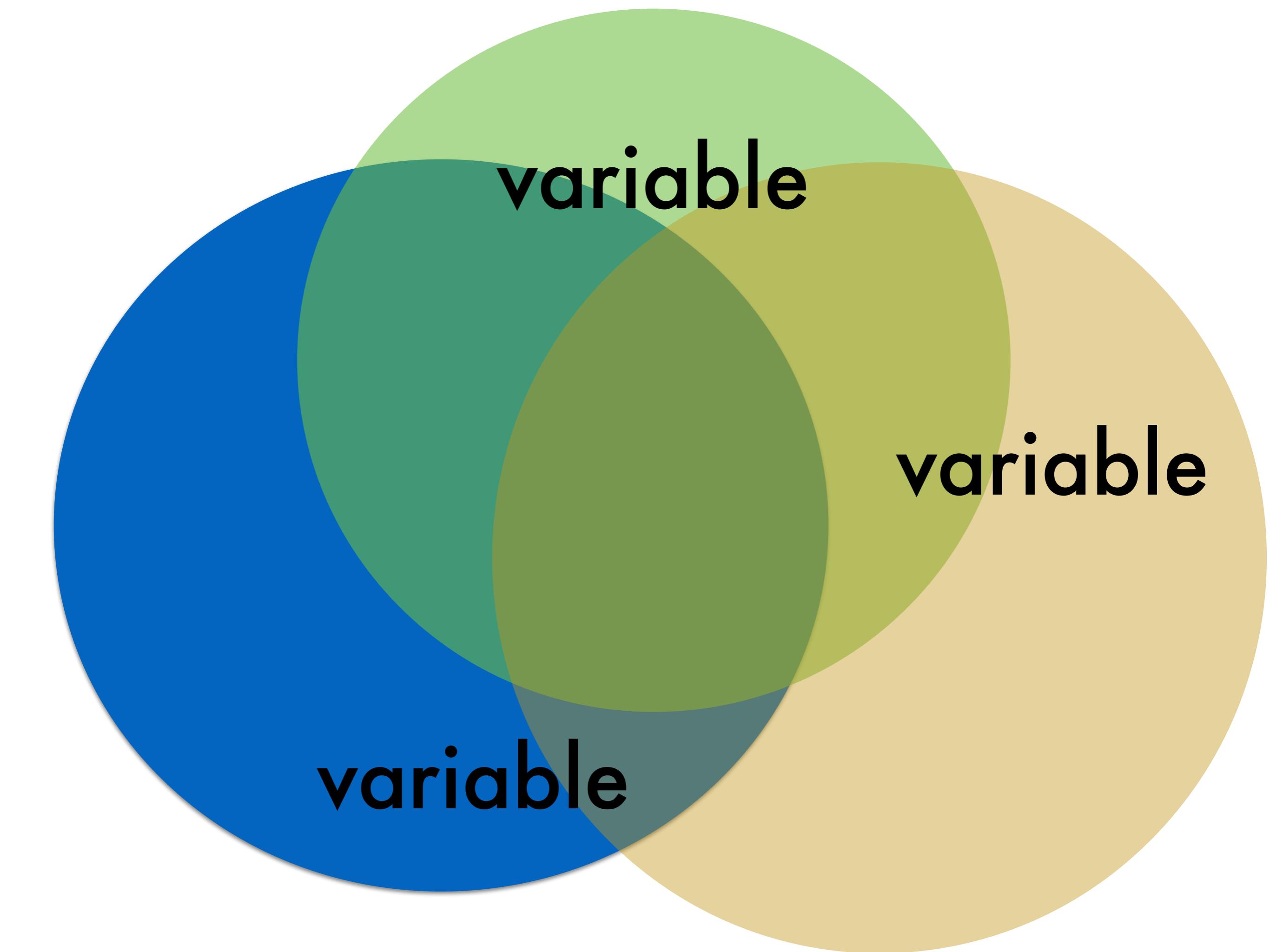
**Figure 4. Maximum likelihood phylogenetic tree of the 20 *Escherichia coli* and *Shigella* strains as reconstructed from the sequences of the 1878 genes of the *Escherichia* core genome.** The earliest diverging species, *E. fergusonii*, was chosen to root the tree. The numbers at the nodes correspond, in black, to the bootstrap values (1000 bootstraps) and, in grey, to a “consensus strength”, which is the number of genes that confirms the bipartition (see Materials and Methods). The latter value is displayed only in instances where consensus and tested trees correspond. The branch length separating *E. fergusonii* from the *E. coli* strains is not to scale; the numbers above the branch indicate its length. Phylogenetic group membership of the strains is indicated with bars at the right of the figure.  
doi:10.1371/journal.pgen.1000344.g004



**pangenome**

**pangenome** = the 'union' of the 'gene sets'

The determination of the pan-genome for a species depends on choice of strains and on the detection of homologs (orthologs and paralogs)



**variable**

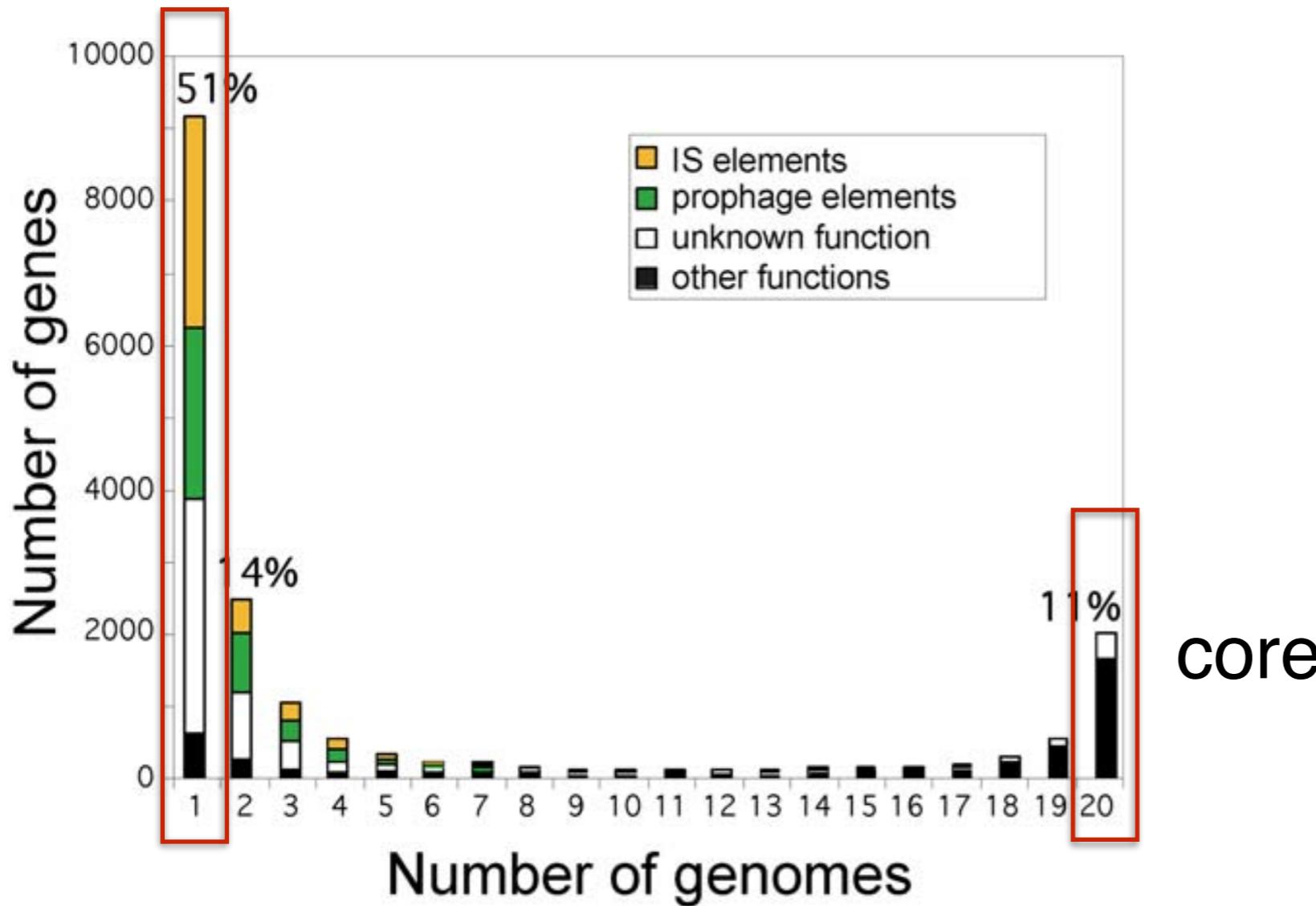
**variable**

**variable**

**variable genome** = set of genes present in single strains – or a subset of strains – of a bacterial species.

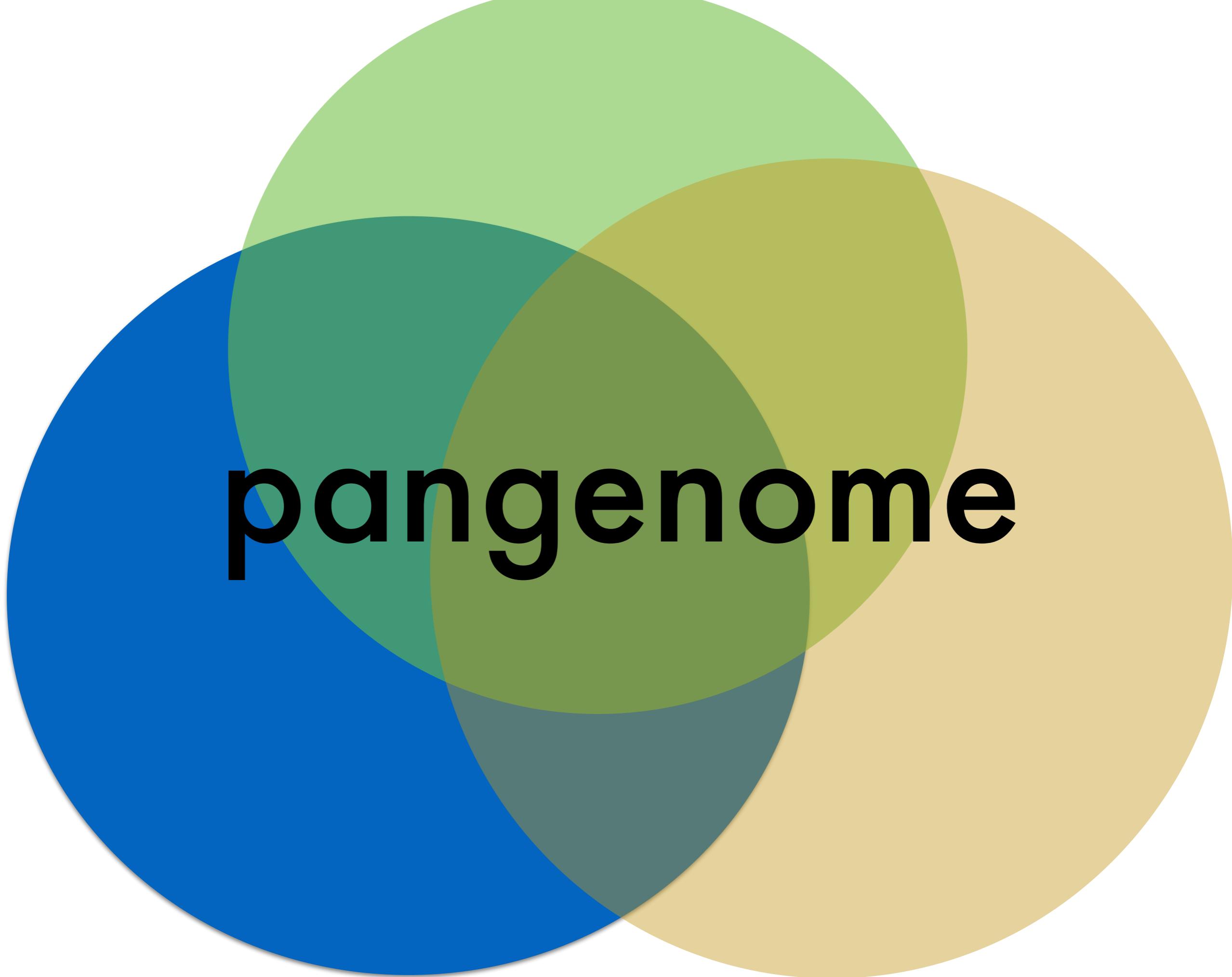
(tend to be hypothetical or unknown function)

# strain-specific



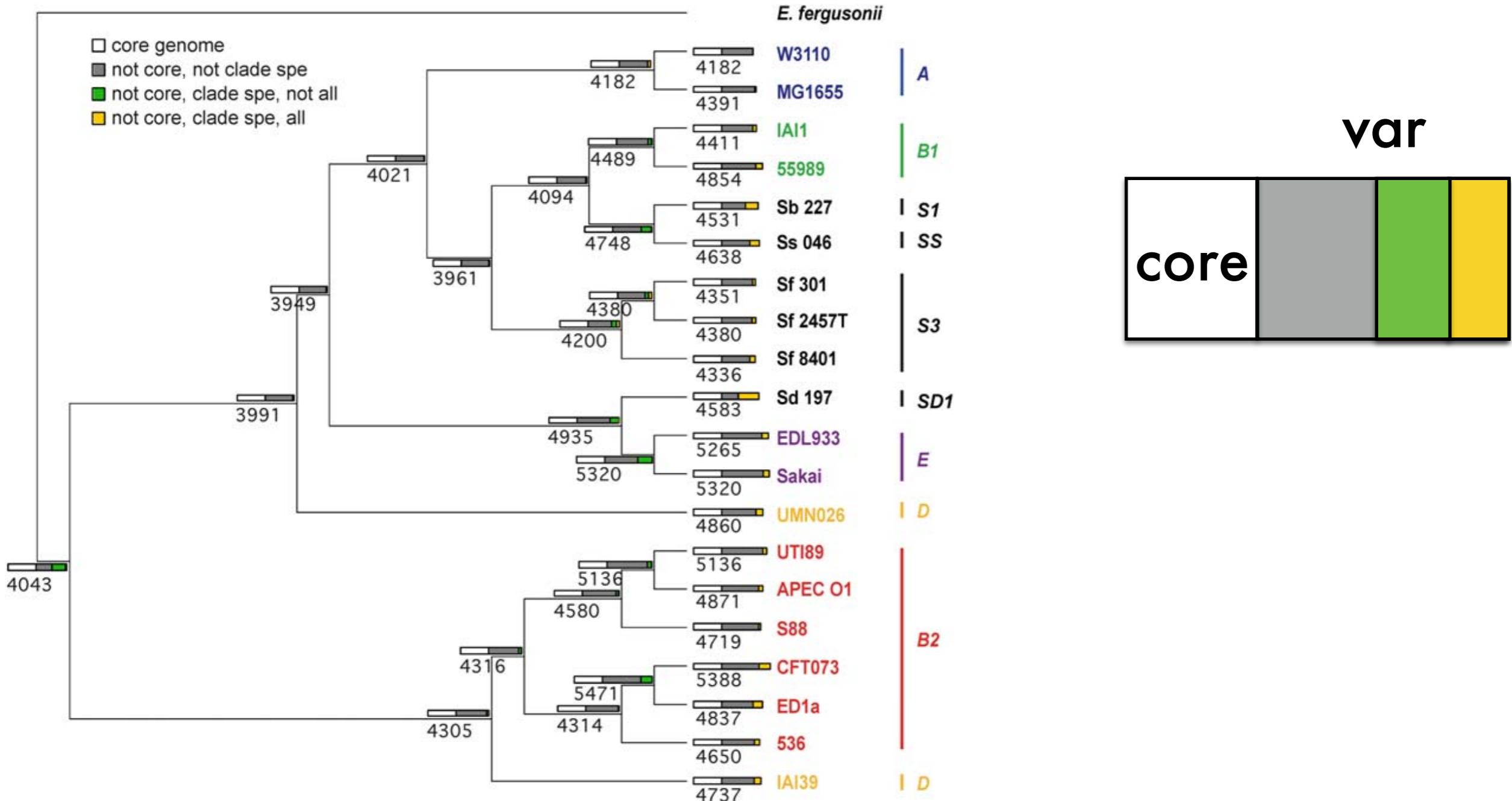
**Figure 2. Frequency of genes within the 20 analysed *Escherichia coli* genomes.** At one extreme of the x-axis are the genes present in a single genome which are regarded as strain specific genes (9054 genes: 51% of the pan-genome), while at the opposite end of the scale are situated the genes found in all 20 genomes, which represent the *E. coli* core-genome (1976 genes: 11% of the pan-genome). Coloured rectangles represent the proportion of insertion sequence (IS)-like elements (yellow), prophage-like elements (green), and genes of unknown/unclassified function (white). Black rectangles represent genes for which a function can be assigned. Strain-specific genes correspond to 2885 IS-like elements (32%), 2352 prophage-like elements (26%), and 3220 genes of unknown/unclassified function (35%).

doi:10.1371/journal.pgen.1000344.g002



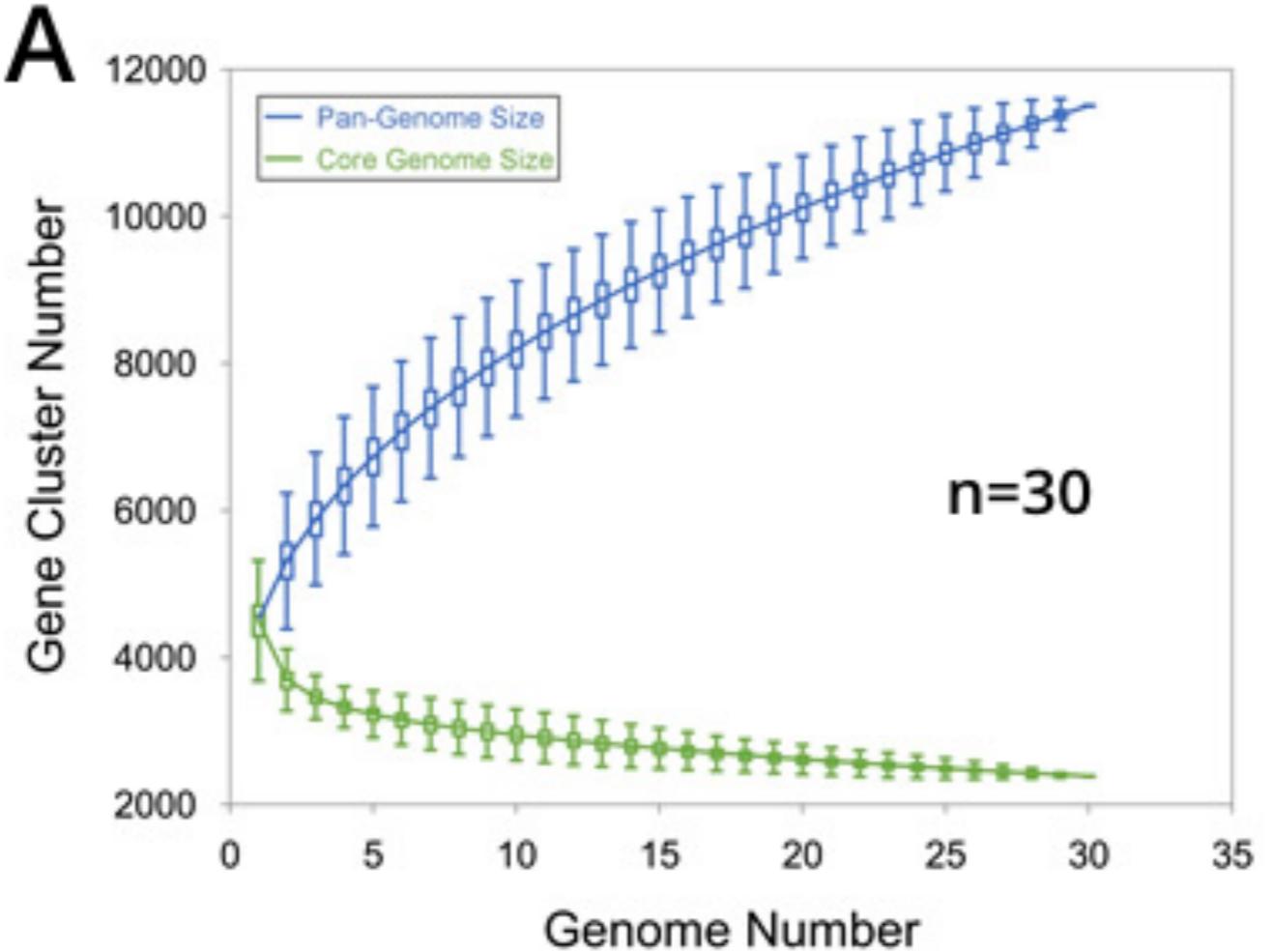
**pangenome**

# Inferred gene content in *E coli* lineage

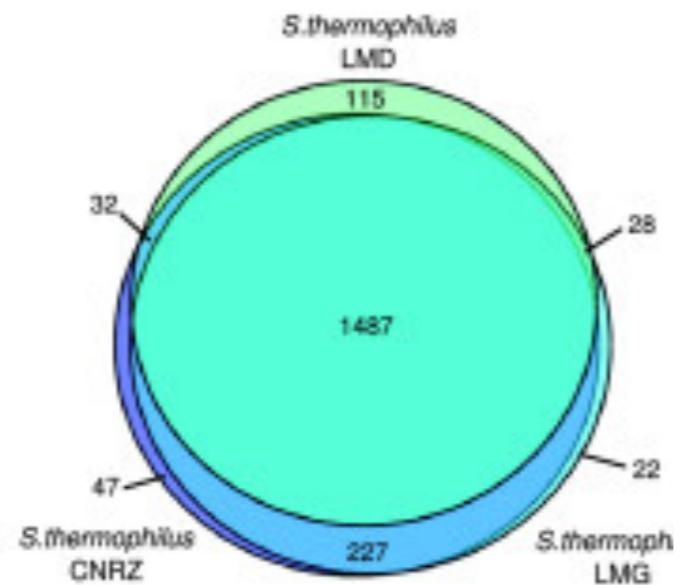


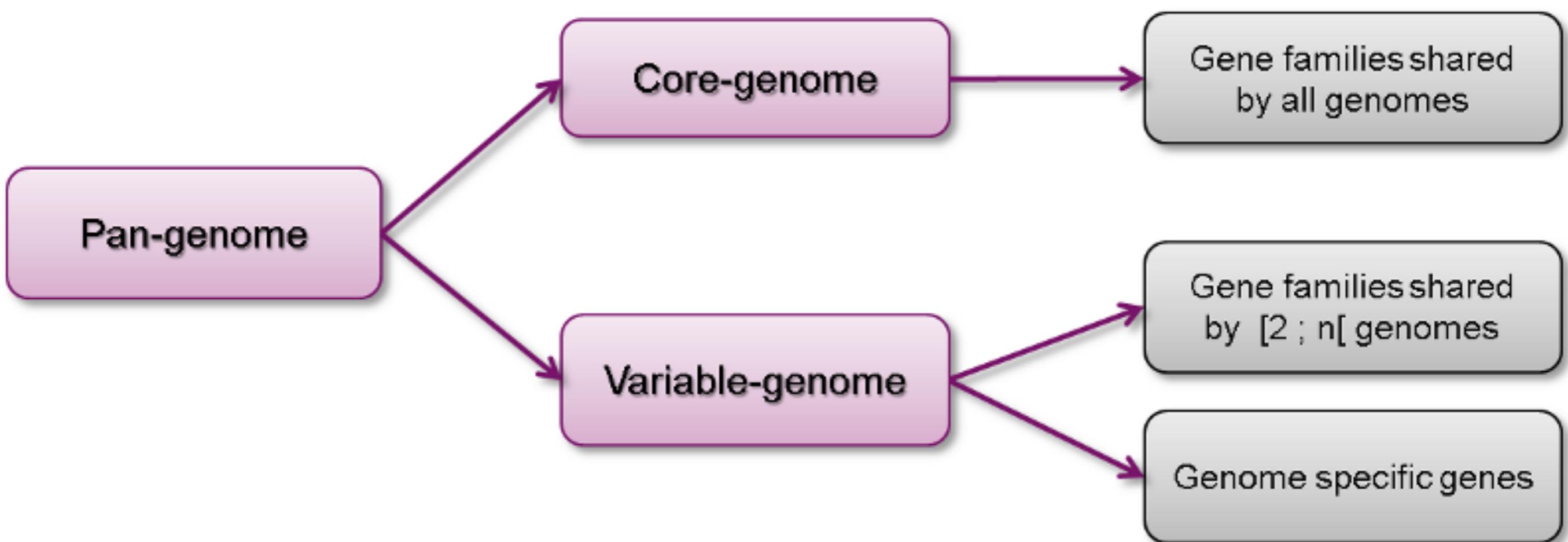
**Figure 6. Inferred gene content evolution in the lineage of *Escherichia coli*.** The cladogram shows the phylogenetic relationships among the 20 *E. coli/Shigella* genomes rooted on the *E. fergusonii* genome, as in Figure 4, but ignoring branch lengths. The major phylogenetic groups are indicated by the vertical lines. Each strain and internal node of the tree is labelled with the number of genes present (as inferred by maximum likelihood; see Materials and Methods). Coloured rectangles represent different gene classes within the gene repertoires of ancestral and modern *E. coli*. Rectangle widths are proportional to the number of genes. The four different gene classes (by colour) include genes that are: in the core genome (white), not clade-specific (grey), clade-specific but not ubiquitous in the clade (green) and both clade-specific and ubiquitous in the clade (yellow). A clade-specific gene is one that is inferred to be present only in the node and its descendent nodes.

doi:10.1371/journal.pgen.1000344.g006

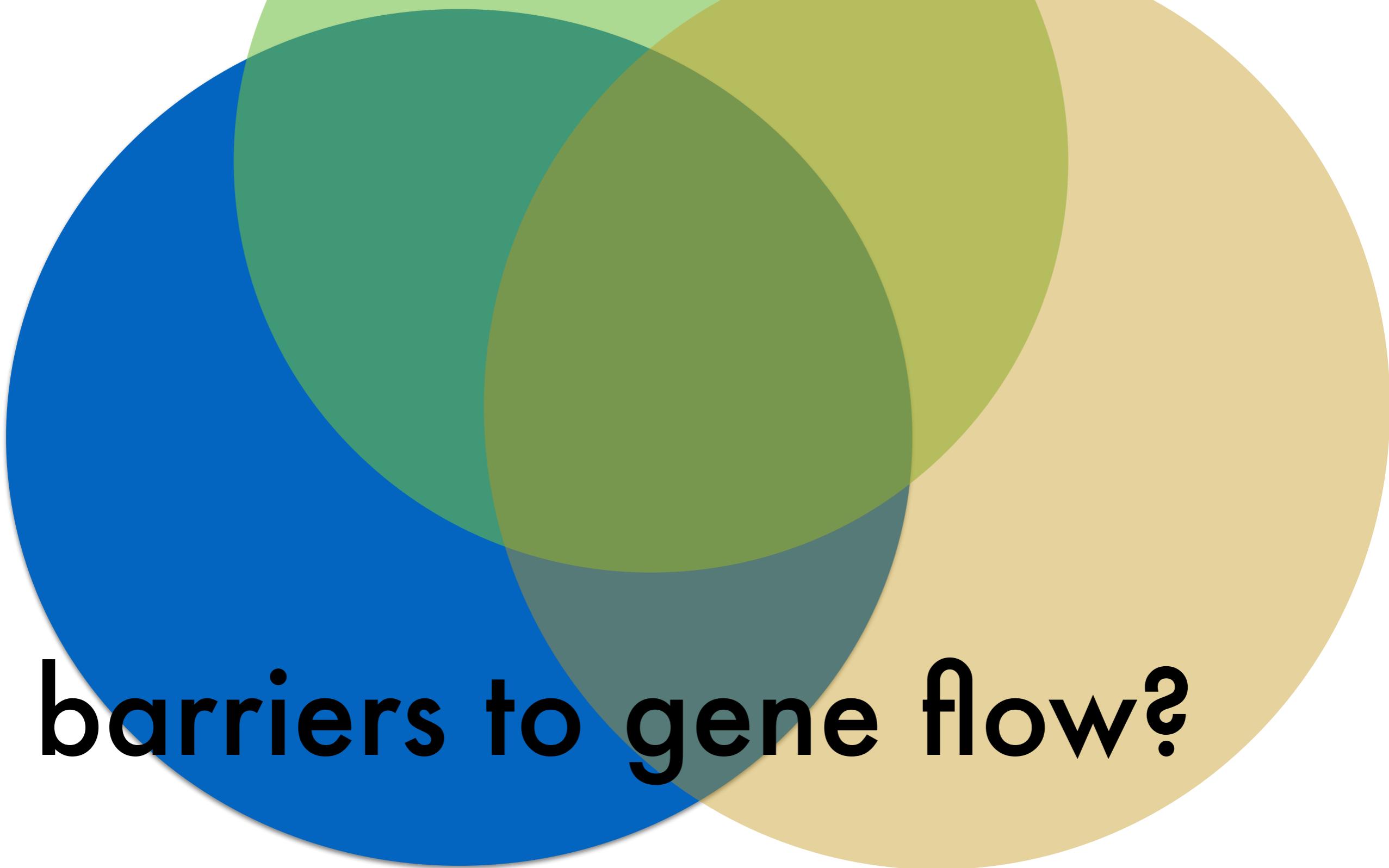


**same  
species**





# Genomes are dynamic



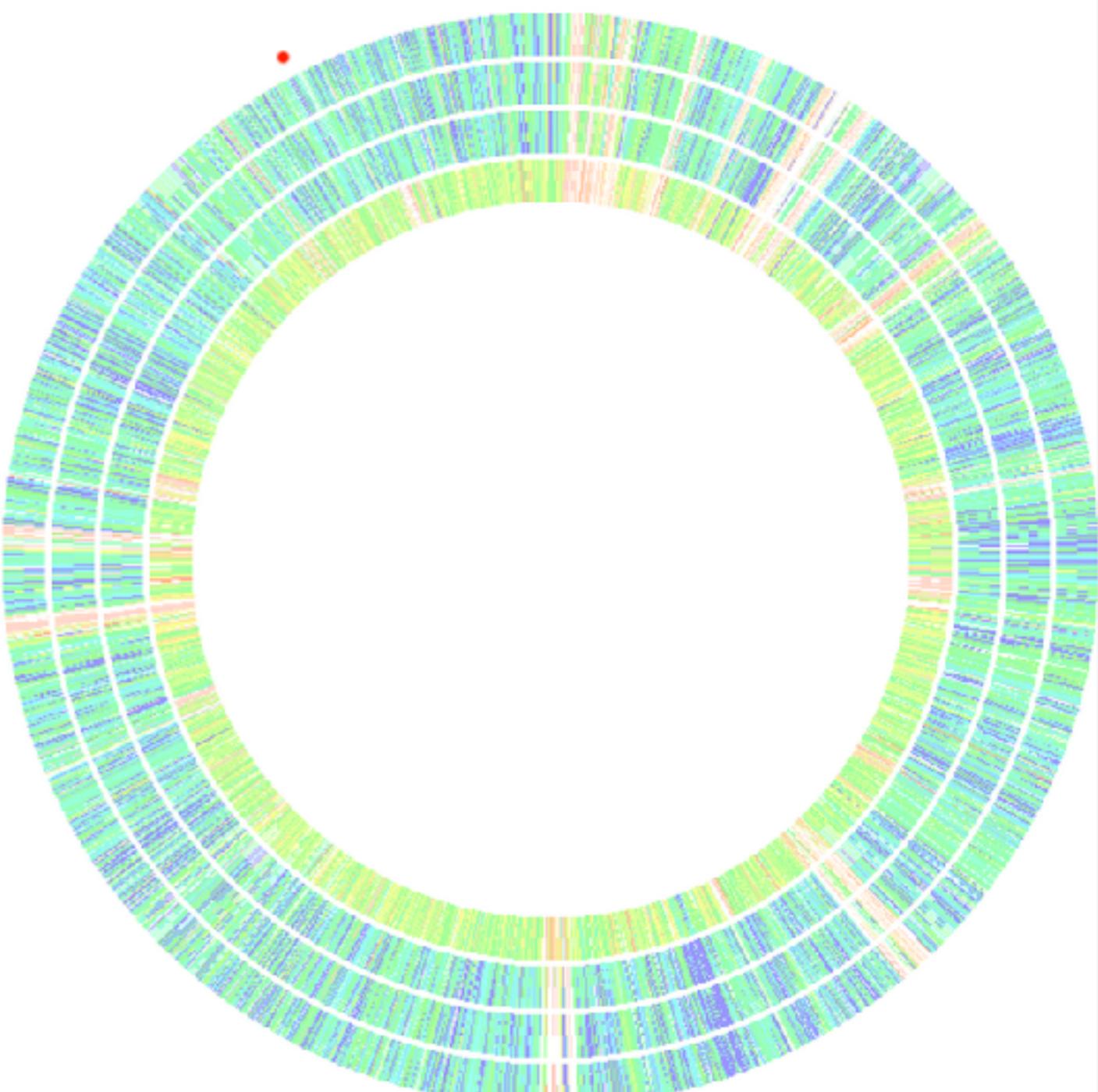
display 30 items per page

first < prev displaying 4089 - 4118 of 4446 next > last >

percent Identity  
408.16

percent identity  
408.19

419610.8				408.16				408.18				408.19				408.17			
Contig	Gene	Length	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene		
1			all	1		all	1		all	1		all	1		all	1			
<a href="#">4110</a>	535	bi	23	<a href="#">3273</a>	bi	45	<a href="#">4256</a>	bi	23	<a href="#">3313</a>	bi	34	<a href="#">3176</a>						
<a href="#">4111</a>	602	bi	23	<a href="#">3274</a>	bi	45	<a href="#">4257</a>	bi	23	<a href="#">3314</a>	bi	34	<a href="#">3177</a>						
<a href="#">4112</a>	414	bi	23	<a href="#">3275</a>	bi	45	<a href="#">4258</a>	bi	23	<a href="#">3315</a>	bi	34	<a href="#">3178</a>						
<a href="#">4113</a>	265	bi	23	<a href="#">3276</a>	bi	45	<a href="#">4259</a>	bi	23	<a href="#">3316</a>	bi	34	<a href="#">3179</a>						
<a href="#">4114</a>	287	bi	23	<a href="#">3277</a>	bi	45	<a href="#">4260</a>	bi	23	<a href="#">3317</a>	bi	34	<a href="#">3180</a>						
<a href="#">4115</a>	488	bi	23	<a href="#">3278</a>	bi	45	<a href="#">4261</a>	bi	23	<a href="#">3318</a>	bi	34	<a href="#">3181</a>						
<a href="#">4116</a>	339	bi	23	<a href="#">3279</a>	bi	45	<a href="#">4262</a>	bi	23	<a href="#">3319</a>	bi	34	<a href="#">3182</a>						
<a href="#">4117</a>	622	bi	23	<a href="#">3280</a>	bi	45	<a href="#">4263</a>	bi	23	<a href="#">3320</a>	bi	34	<a href="#">3183</a>						
<a href="#">4118</a>	520	bi	23	<a href="#">3282</a>	bi	45	<a href="#">4265</a>	bi	23	<a href="#">3322</a>	bi	34	<a href="#">3185</a>						
<a href="#">4119</a>	692	bi	23	<a href="#">3283</a>	bi	45	<a href="#">4266</a>	bi	23	<a href="#">3323</a>	bi	34	<a href="#">3186</a>						
<a href="#">4120</a>	942	bi	23	<a href="#">3284</a>	bi	45	<a href="#">4267</a>	bi	23	<a href="#">3324</a>	bi	34	<a href="#">3187</a>						
<a href="#">4121</a>	215	bi	23	<a href="#">3285</a>	bi	45	<a href="#">4268</a>	bi	23	<a href="#">3325</a>	bi	34	<a href="#">3188</a>						
<a href="#">4122</a>	500	bi	23	<a href="#">3287</a>	bi	45	<a href="#">4270</a>	bi	23	<a href="#">3327</a>	bi	34	<a href="#">3190</a>						
<a href="#">4123</a>	293	bi	23	<a href="#">3289</a>	bi	45	<a href="#">4272</a>	bi	23	<a href="#">3329</a>	bi	34	<a href="#">3192</a>						
<a href="#">4124</a>	412	bi	23	<a href="#">3290</a>	bi	45	<a href="#">4273</a>	bi	23	<a href="#">3330</a>	uni	41	<a href="#">4687</a>						
<a href="#">4125</a>	1070	bi	23	<a href="#">3291</a>	bi	45	<a href="#">4274</a>	bi	23	<a href="#">3331</a>	bi	34	<a href="#">3193</a>						
<a href="#">4126</a>	382	bi	23	<a href="#">3292</a>	bi	45	<a href="#">4275</a>	bi	23	<a href="#">3332</a>	bi	34	<a href="#">3194</a>						
<a href="#">4127</a>	114	bi	23	<a href="#">3293</a>	bi	45	<a href="#">4276</a>	bi	23	<a href="#">3333</a>	bi	34	<a href="#">3196</a>						
<a href="#">4128</a>	457	bi	23	<a href="#">3294</a>	bi	45	<a href="#">4277</a>	bi	23	<a href="#">3334</a>	bi	34	<a href="#">3197</a>						
<a href="#">4129</a>	209	bi	23	<a href="#">3295</a>	bi	45	<a href="#">4278</a>	bi	23	<a href="#">3335</a>	bi	34	<a href="#">3198</a>						
<a href="#">4130</a>	212	bi	23	<a href="#">3296</a>	bi	45	<a href="#">4279</a>	bi	23	<a href="#">3336</a>	bi	34	<a href="#">3199</a>						
<a href="#">4131</a>	244	bi	23	<a href="#">3298</a>	bi	45	<a href="#">4281</a>	bi	23	<a href="#">3338</a>	bi	34	<a href="#">3200</a>						
<a href="#">4132</a>	497	-		-		-	-		-	-	bi	12	<a href="#">1819</a>						
<a href="#">4133</a>	86	-		-		-	-		-	-	-	-							
<a href="#">4134</a>	320	bi	23	<a href="#">3300</a>	bi	45	<a href="#">4283</a>	bi	23	<a href="#">3340</a>	bi	34	<a href="#">3274</a>						
<a href="#">4135</a>	123	bi	23	<a href="#">3301</a>	bi	45	<a href="#">4284</a>	bi	23	<a href="#">3341</a>	bi	34	<a href="#">3275</a>						
<a href="#">4136</a>	1067	bi	23	<a href="#">3302</a>	bi	45	<a href="#">4285</a>	bi	23	<a href="#">3342</a>	bi	34	<a href="#">3276</a>						
<a href="#">4137</a>	525	-		-		-	-		-	-	-	-							
<a href="#">4138</a>	174	bi	23	<a href="#">3305</a>	bi	45	<a href="#">4288</a>	bi	23	<a href="#">3345</a>	bi	34	<a href="#">3280</a>						
<a href="#">4139</a>	639	bi	23	<a href="#">3306</a>	bi	45	<a href="#">4289</a>	bi	23	<a href="#">3346</a>	bi	34	<a href="#">3281</a>						



darker colors = higher % similarity