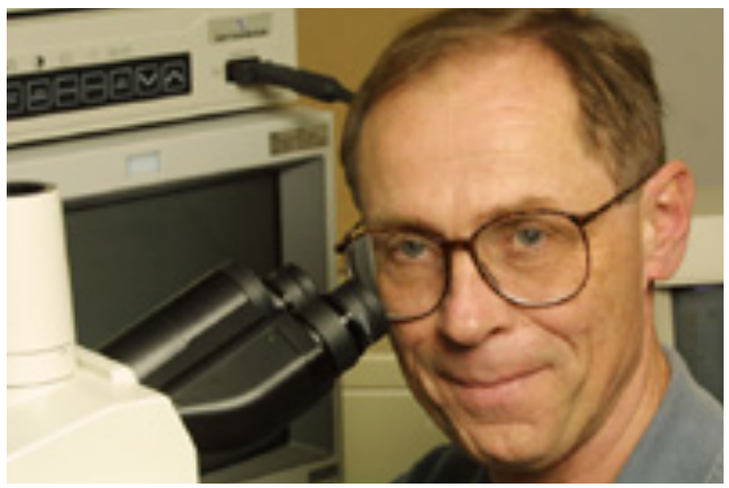
http://rast.nmpdr.org

stru *elles*

http

My academic lineage

Carl Woese
U. Illinois



Norm Pace
CU Boulder

S. Giovannoni
Oregon State

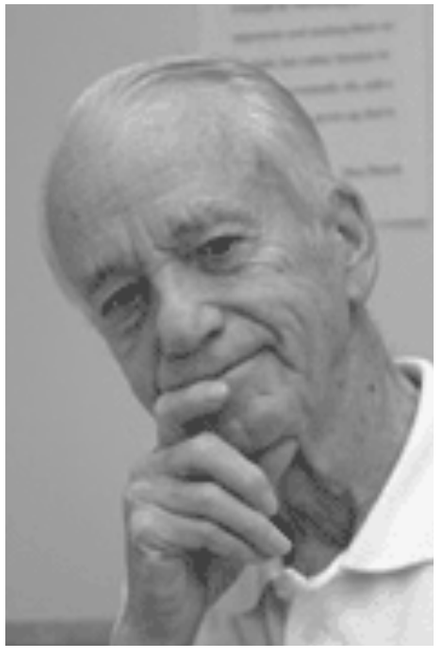
D. Stahl
Univ. of Washington

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

IMDb | Find Movies, TV shows, Celebrities and more... | All | Movies, TV & Showtimes | Celebs, Events & Photos | News & Community | Watchlist

Scott Dawson (IX)

SEE RANK

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

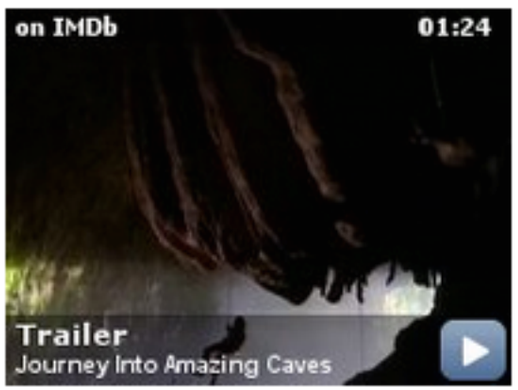
Hide all | Show by... | Edit

Self (1 credit)

Hide

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

Related Videos



Grand Prismatic Spring, Yellowstone National Park.

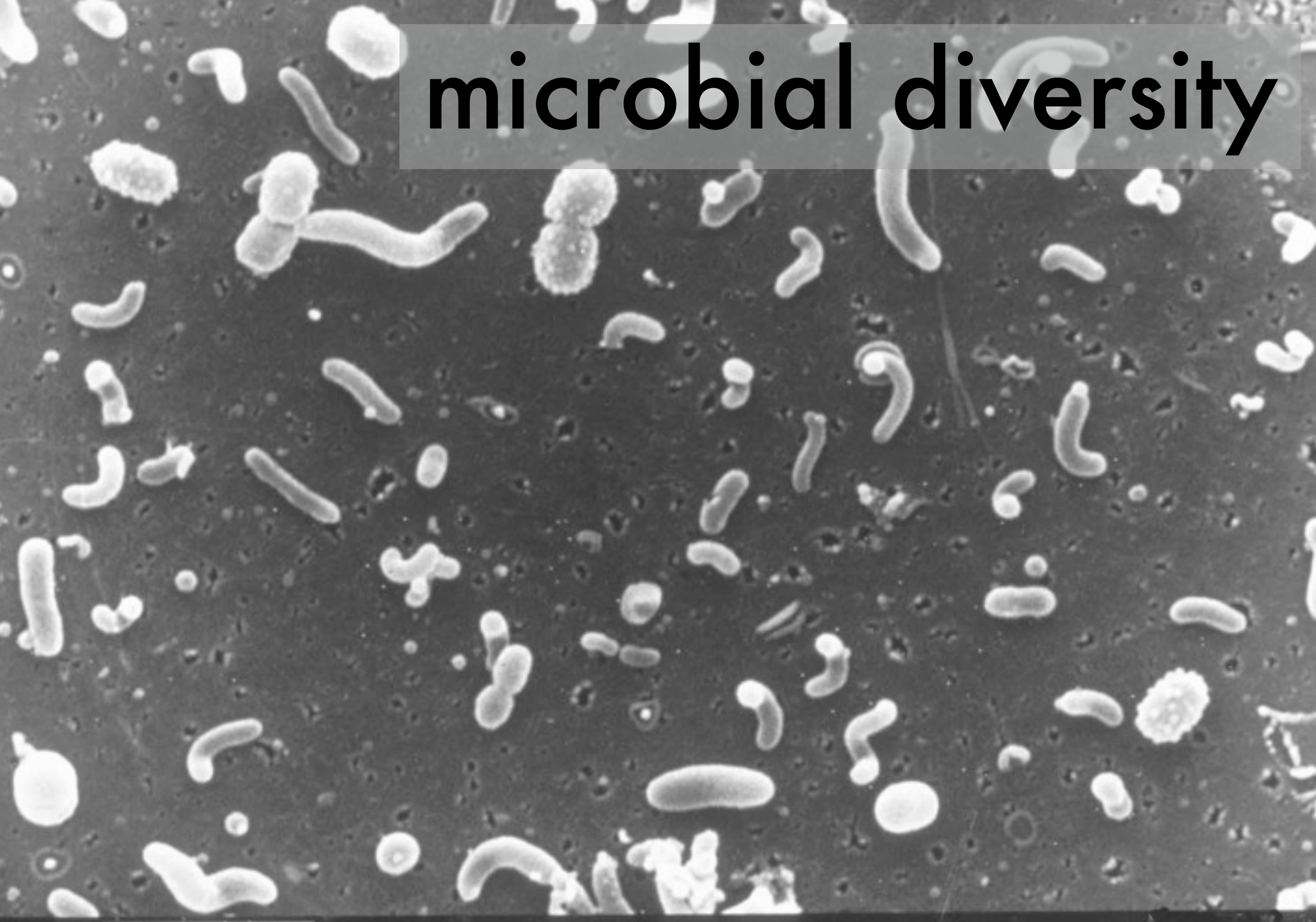
Jim Urquhart / Reuters

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.

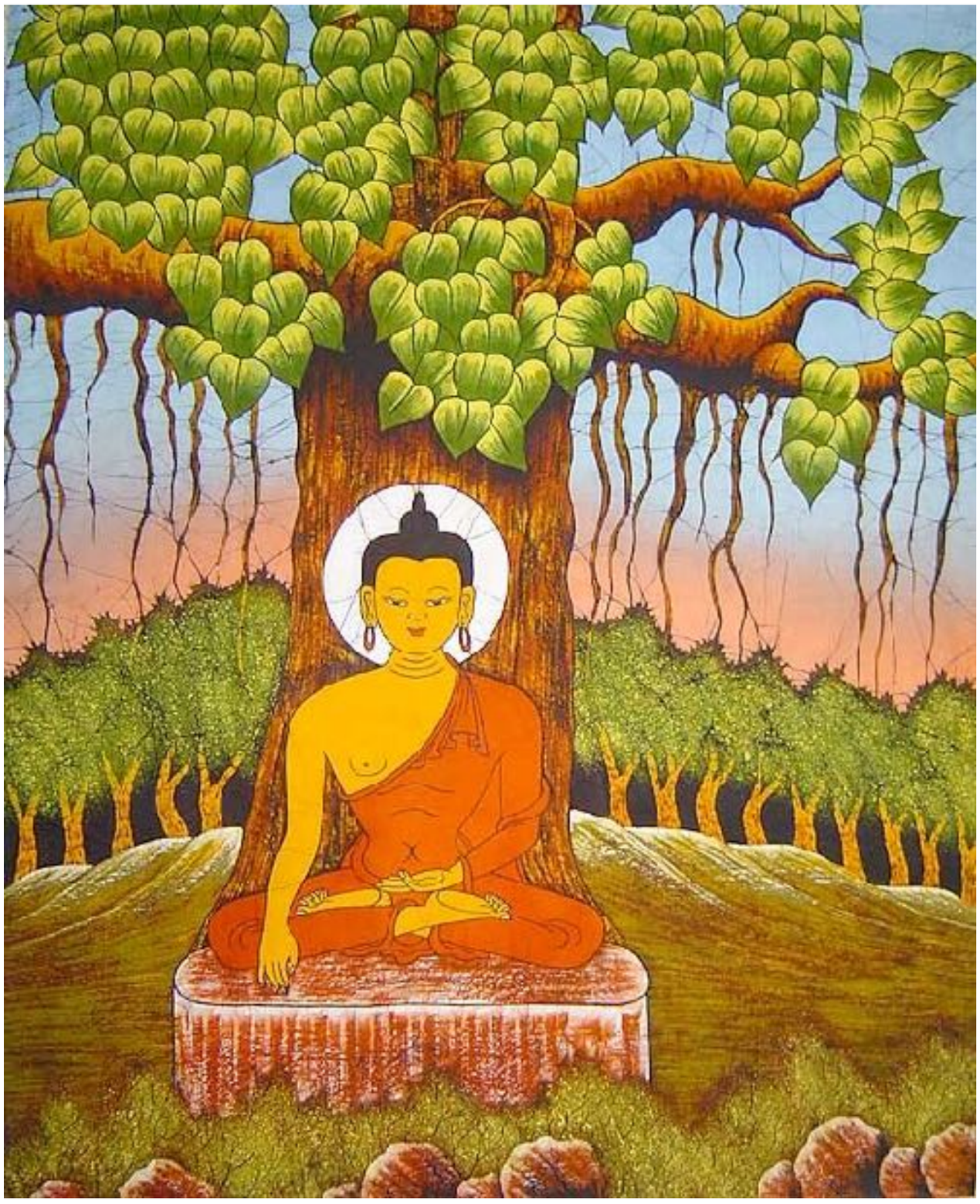
https://www.theatlantic.com/science/archive/2017/07/the-man-who-blew-the-door-off-the-microbial-world/534246/

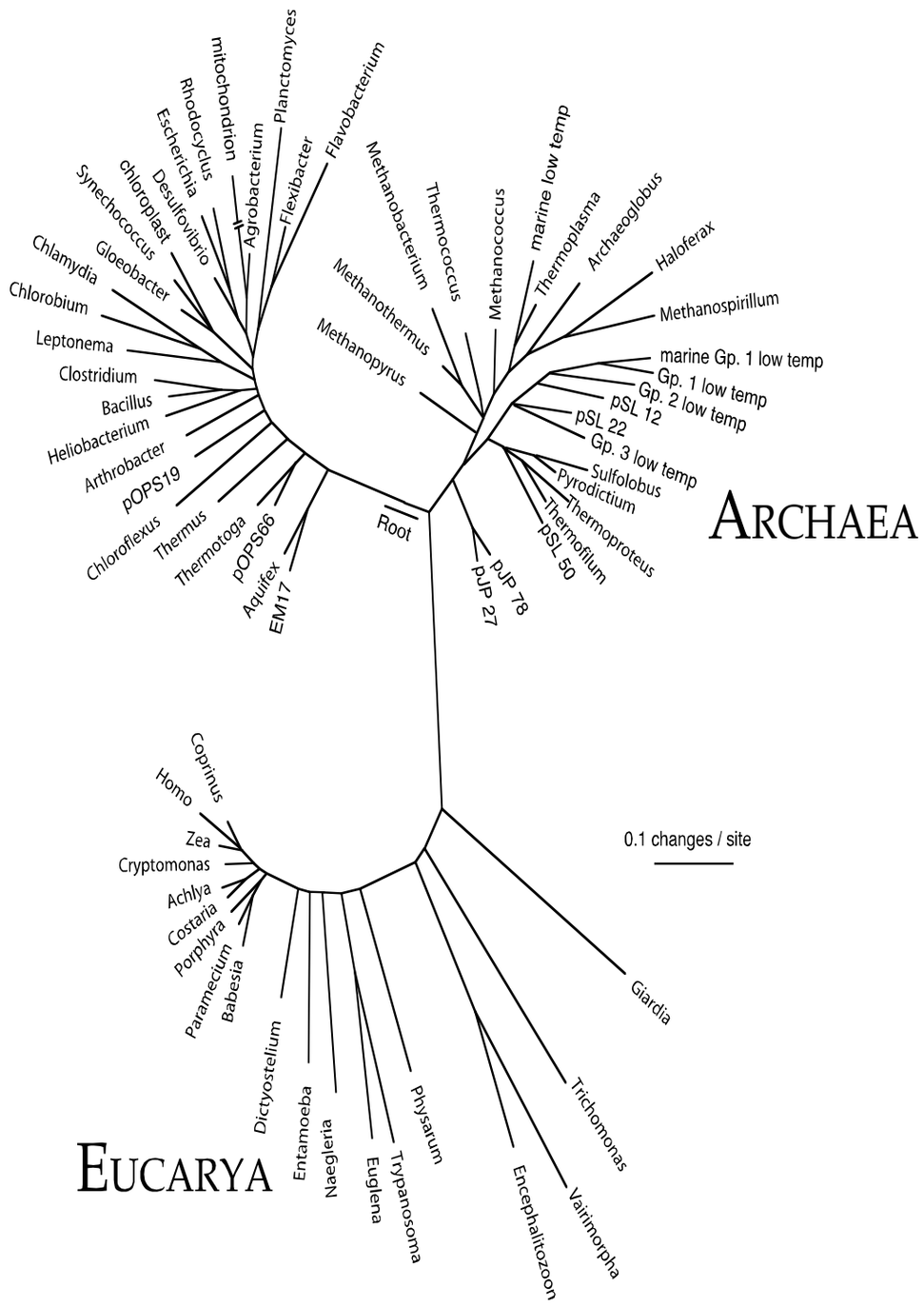
microbial diversity



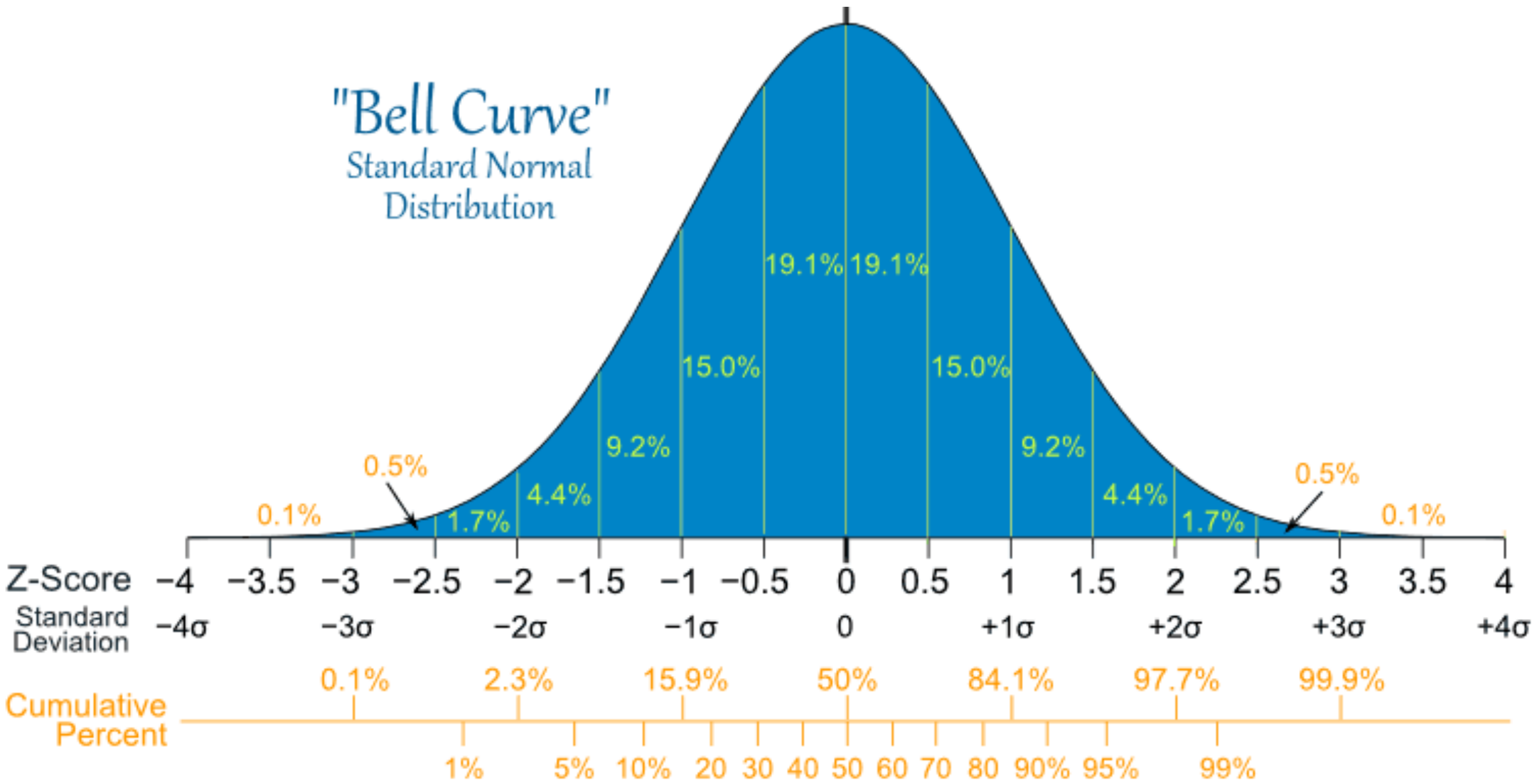
How to define microbial diversity?

Ponderings under the Big Tree





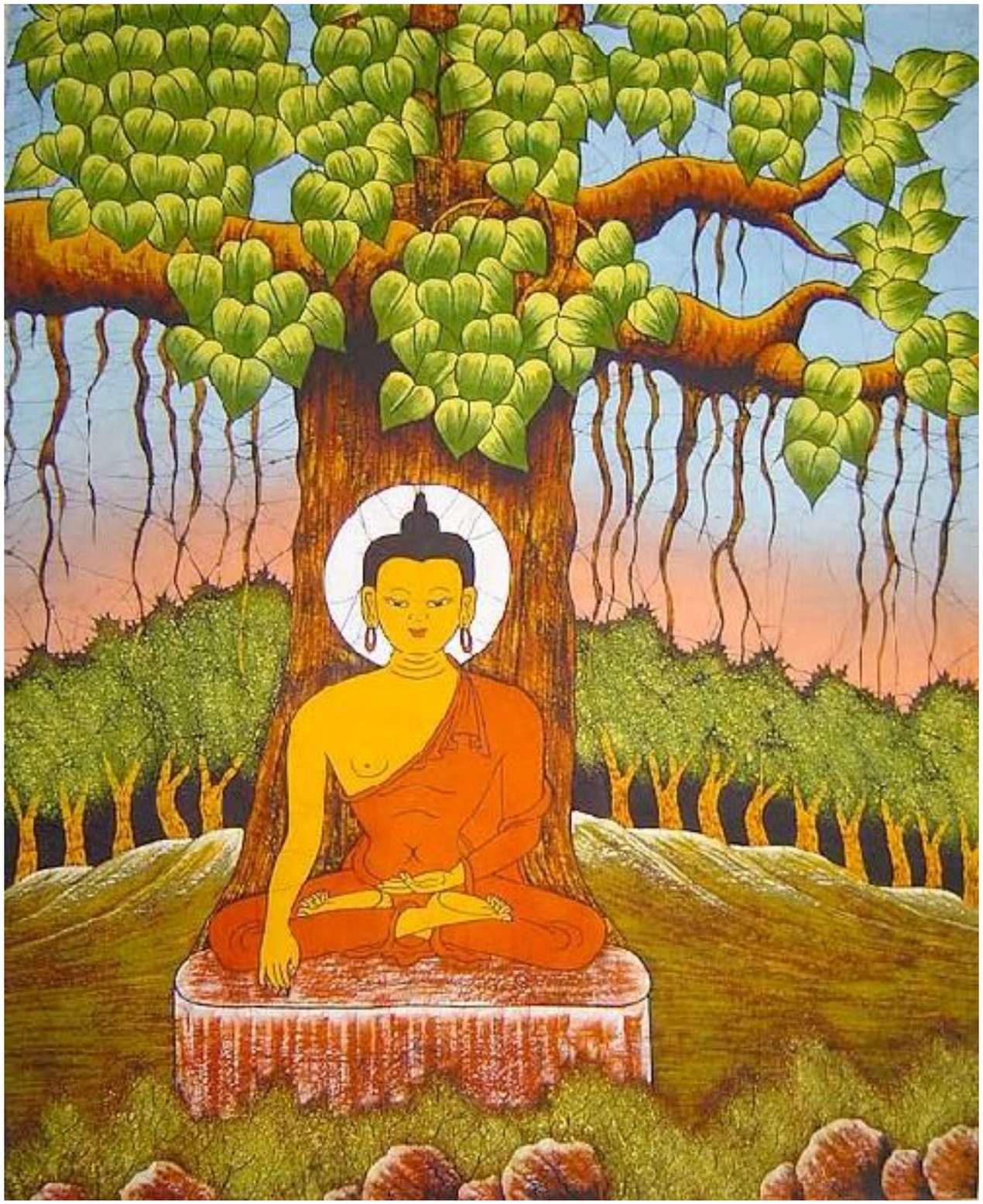
it's all about the variation…



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

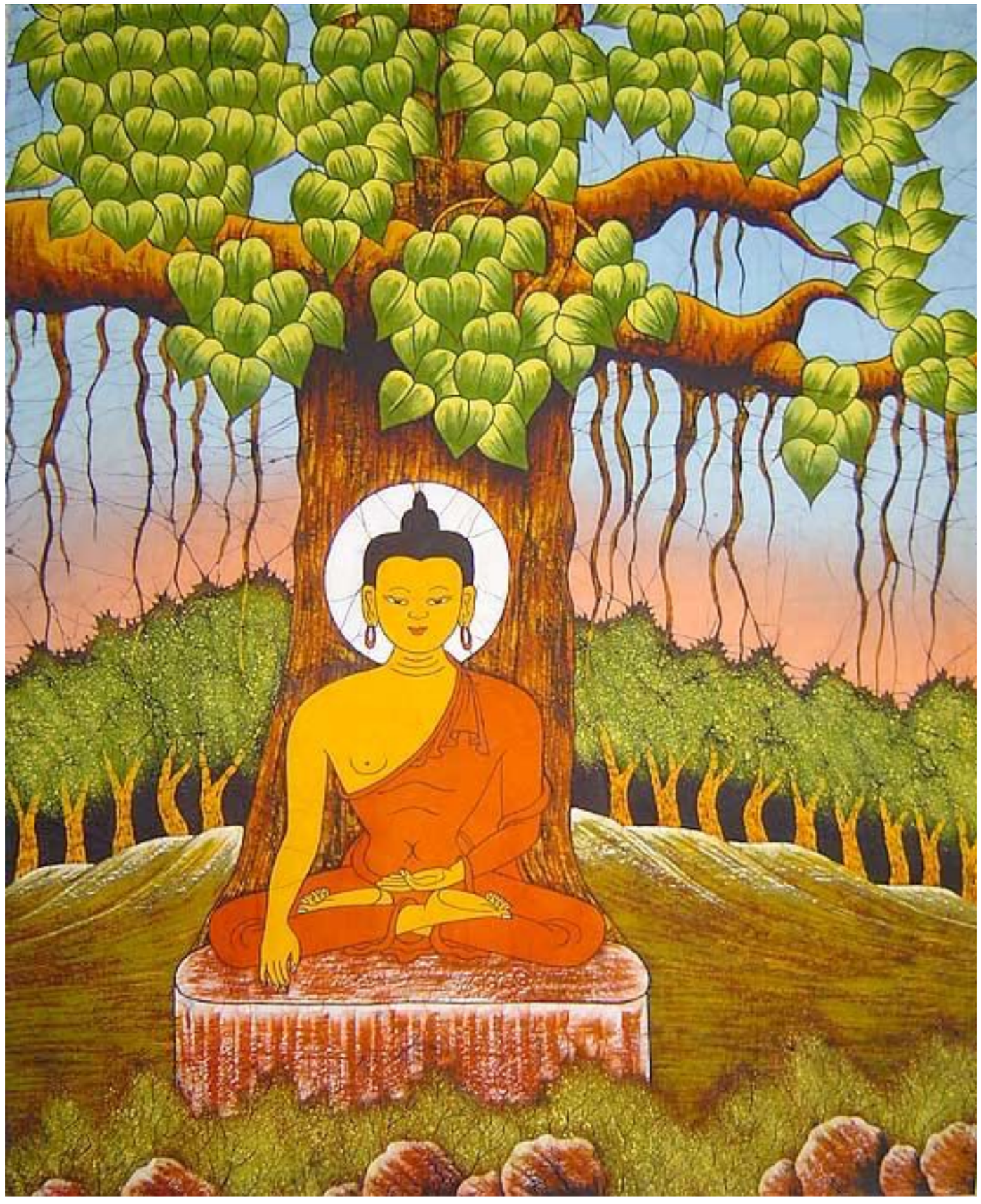
- infer organismal evolutionary relationships using one or more genetic sequences
- help to build multiple sequence alignments
- estimate divergence times of lineages using molecular clocks
- classify "families" of genes and detect **orthologs/ paralogs/xenologs**
- reconstruct ancestral genomes at nodes
- 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.

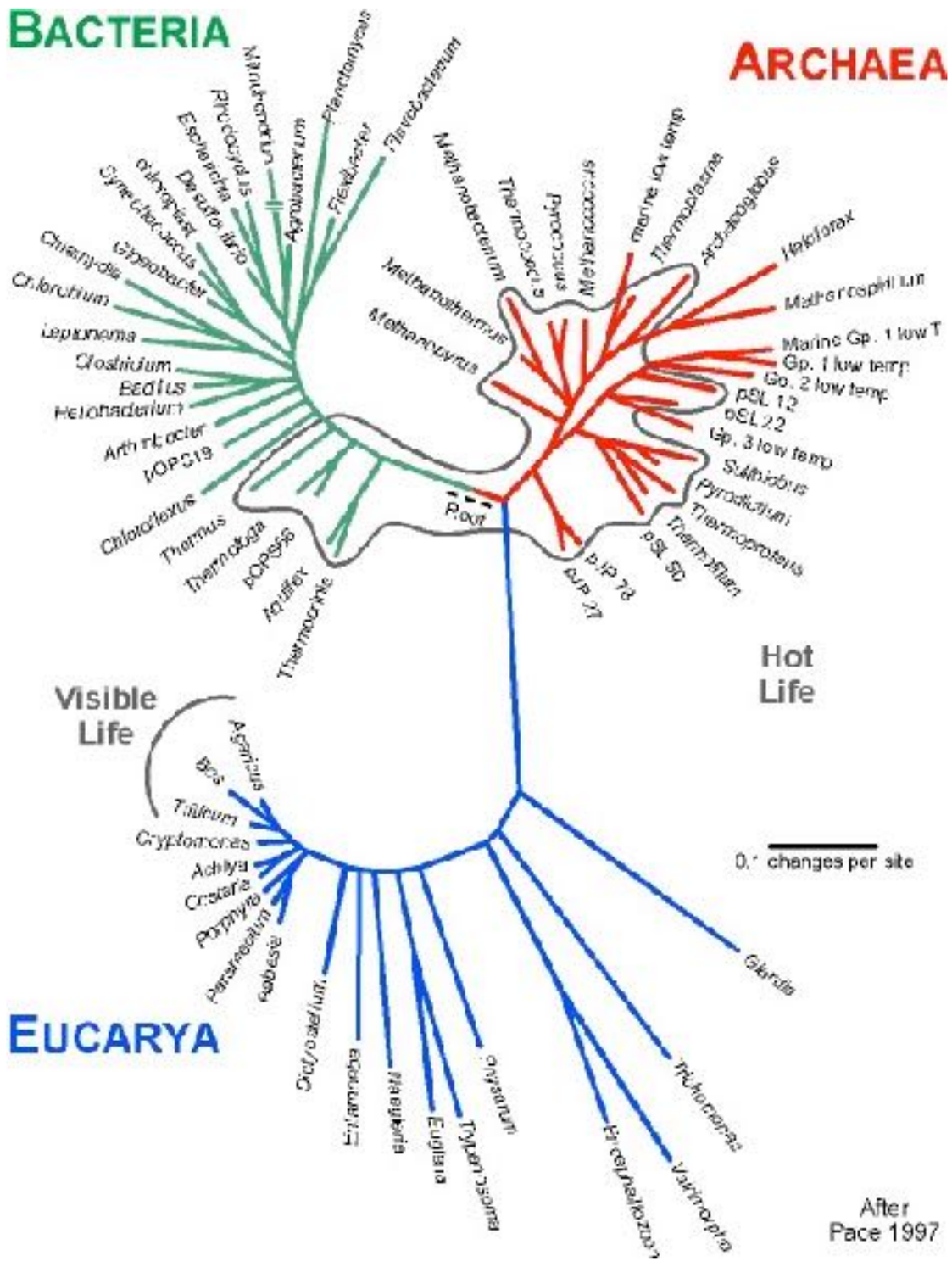
Taxon rank	A long-studied taxon	A less-studied taxon	An uncultivated environmental sample
Domain	**Bacteria**	**Archaea**	**Bacteria**
Division (phylum)	**Actinobacteria** Filamentous gram-positive	**Euryarchaeota** Methanogens and halophiles	**Proteobacteria** Purple bacteria and relatives; gram-negative
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	***Streptomyces***	***Methanocaldococcus***	***Pelagibacter***
Species	***S. coelicolor***	***M. jannaschii***	***P. ubique***
(date first described)	**(1908)**	**(1984)**	**(2002)**

Table 17.2 Microbiology: An Evolving Science

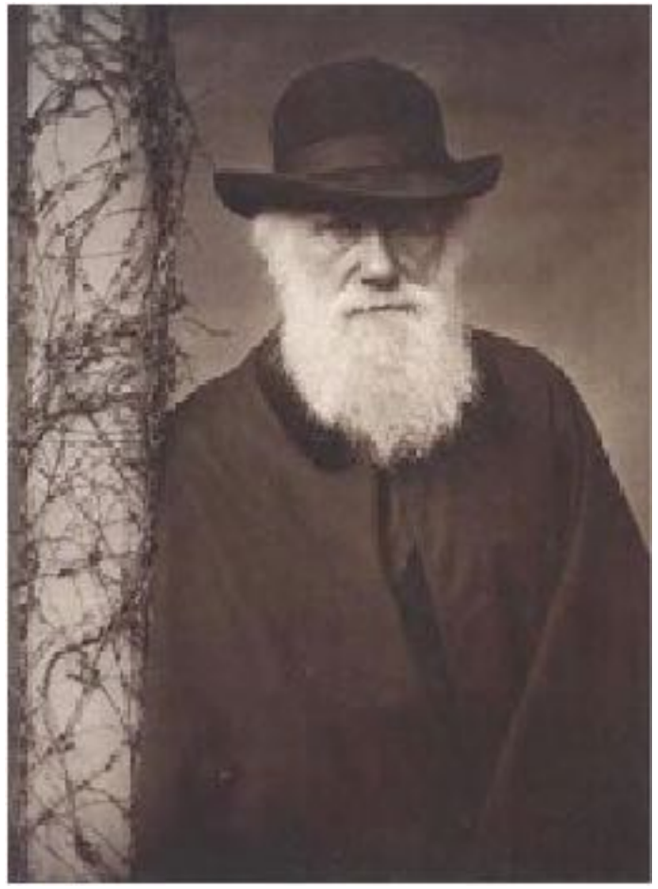
Taxonomy is based on phenotype, molecular is phylogeny based on genotype.

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

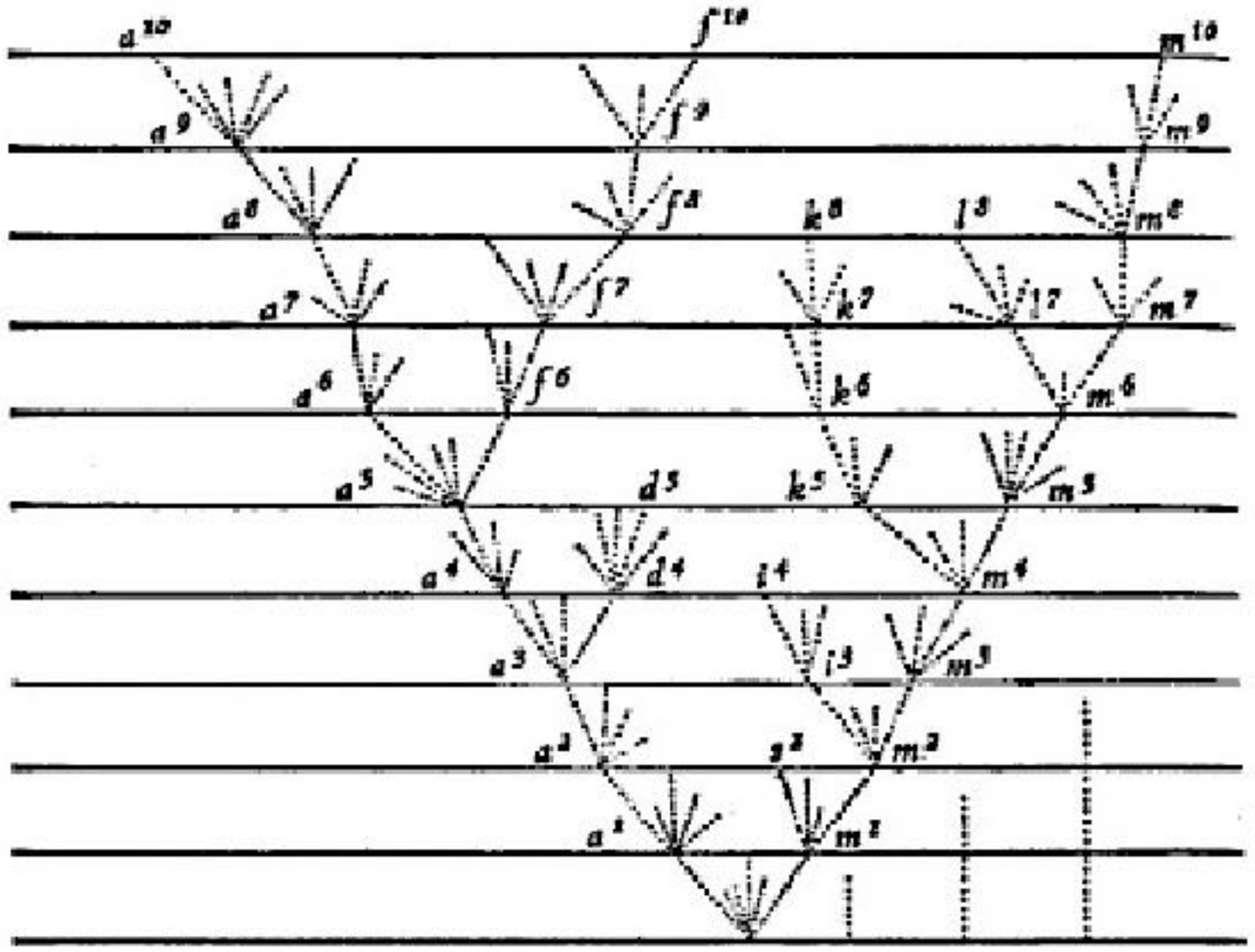
How to read trees



Common ancestry gives rise to independent lineages



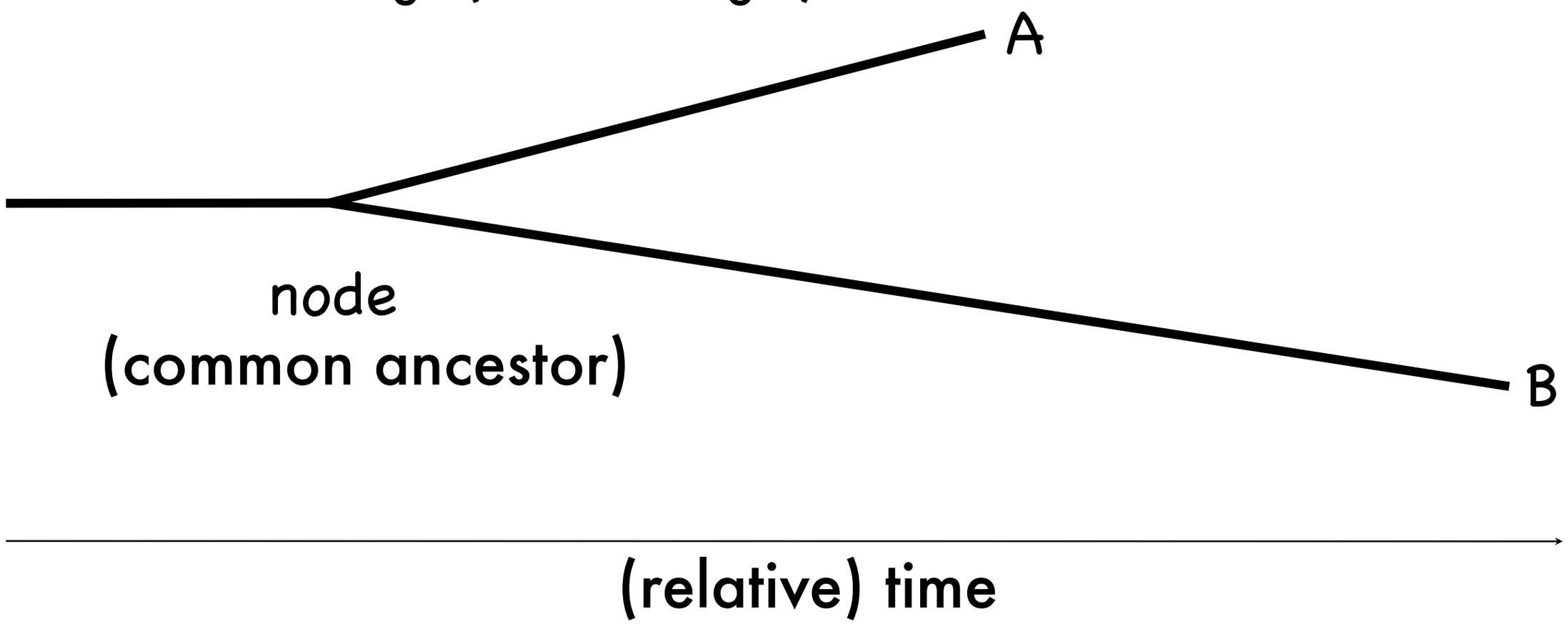
Charles Darwin
The Origin of Species
1859



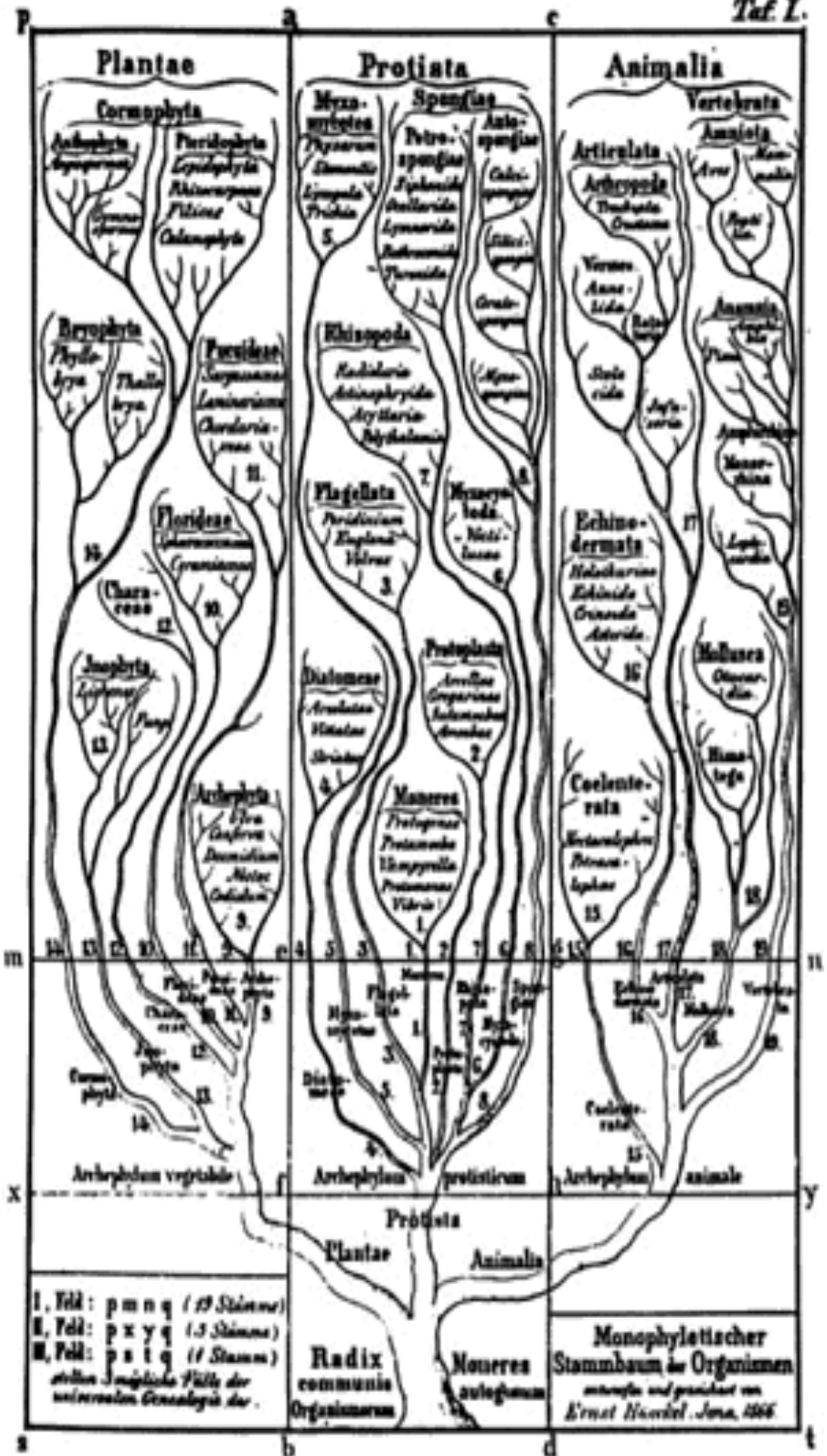
"A structure is similar among related organisms because those organisms h
all descended from a common ancestor that had an equivalent trait."

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)

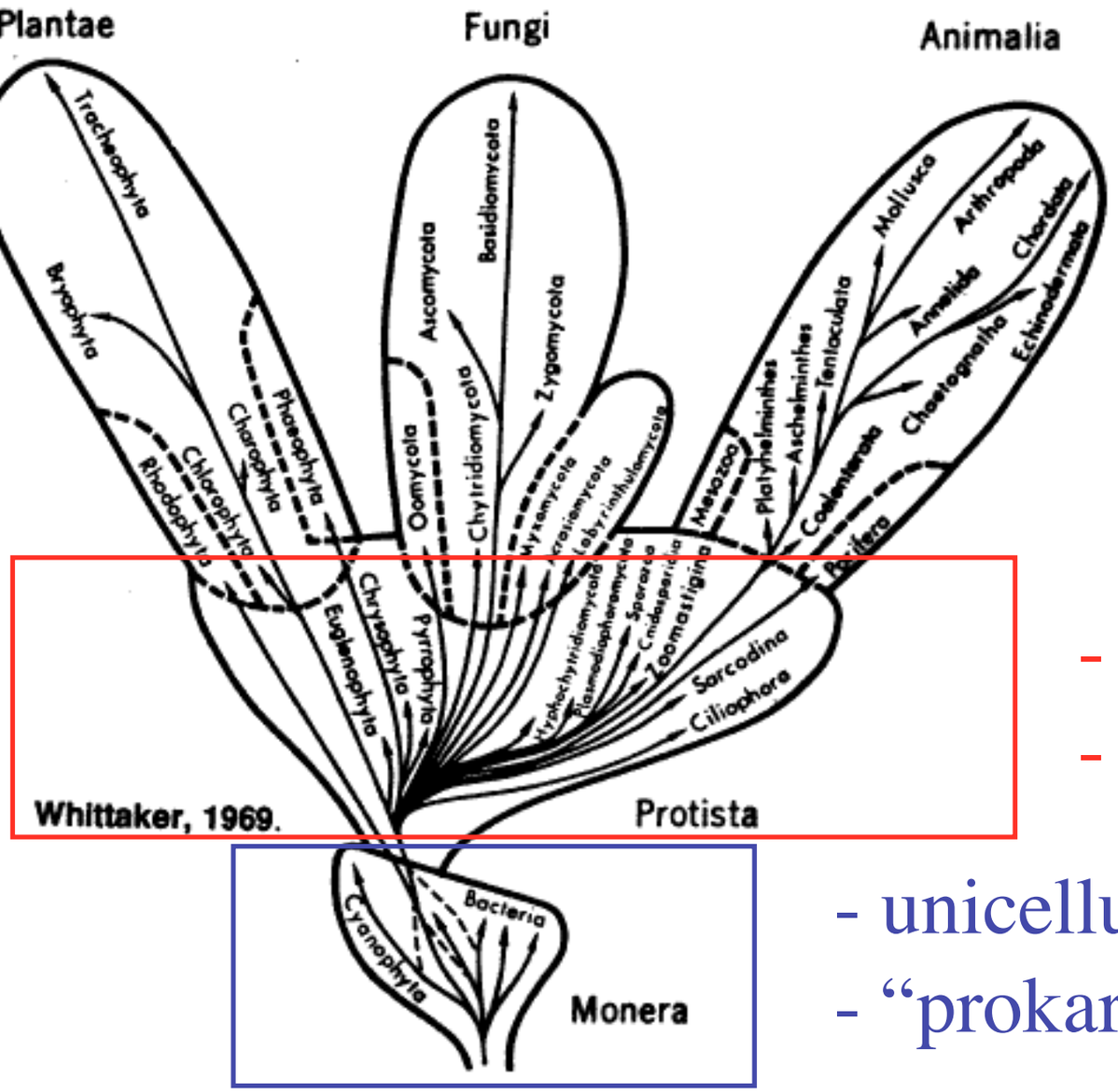


Ernst Haeckel



*

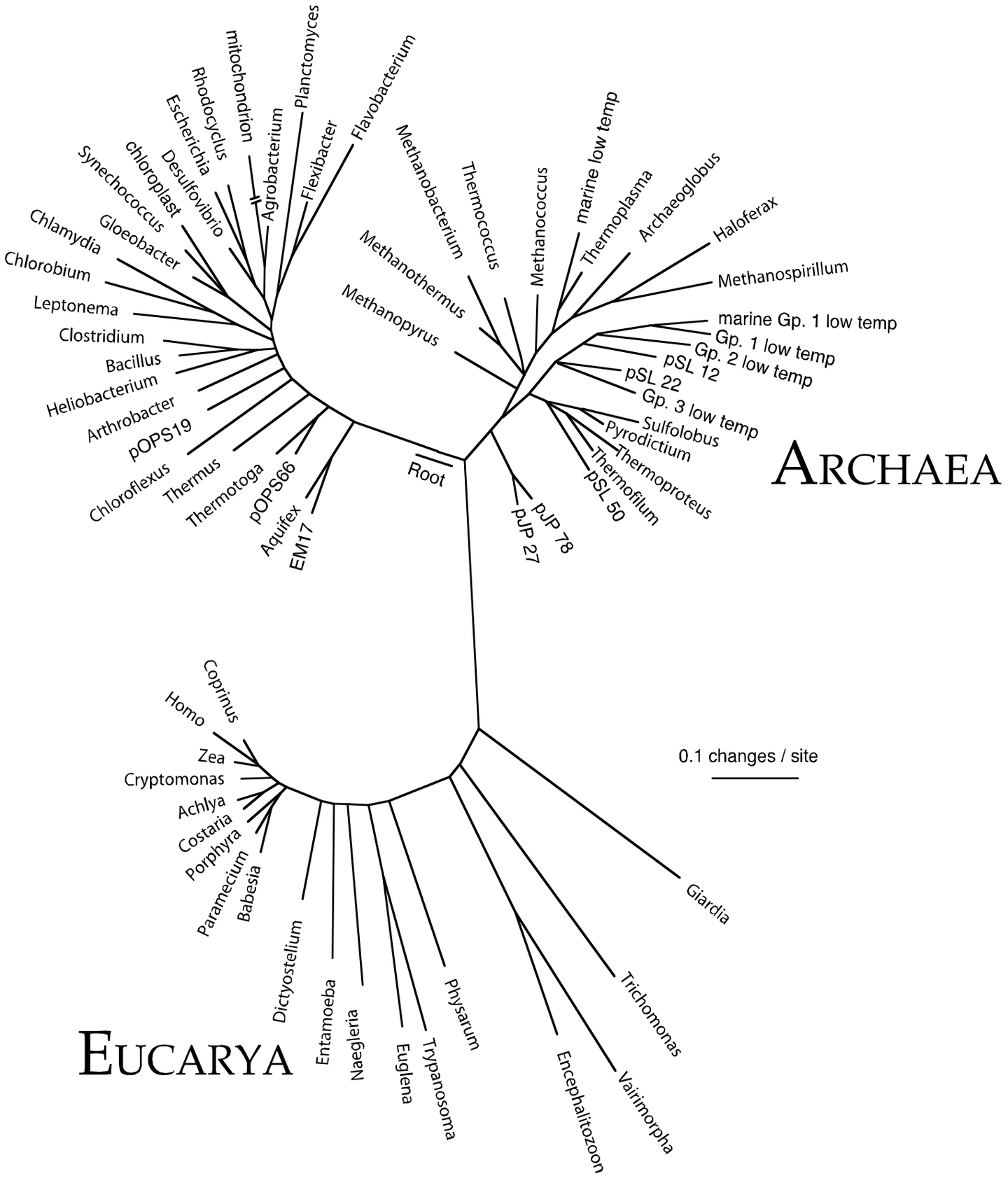
The 5 Kingdom Scheme



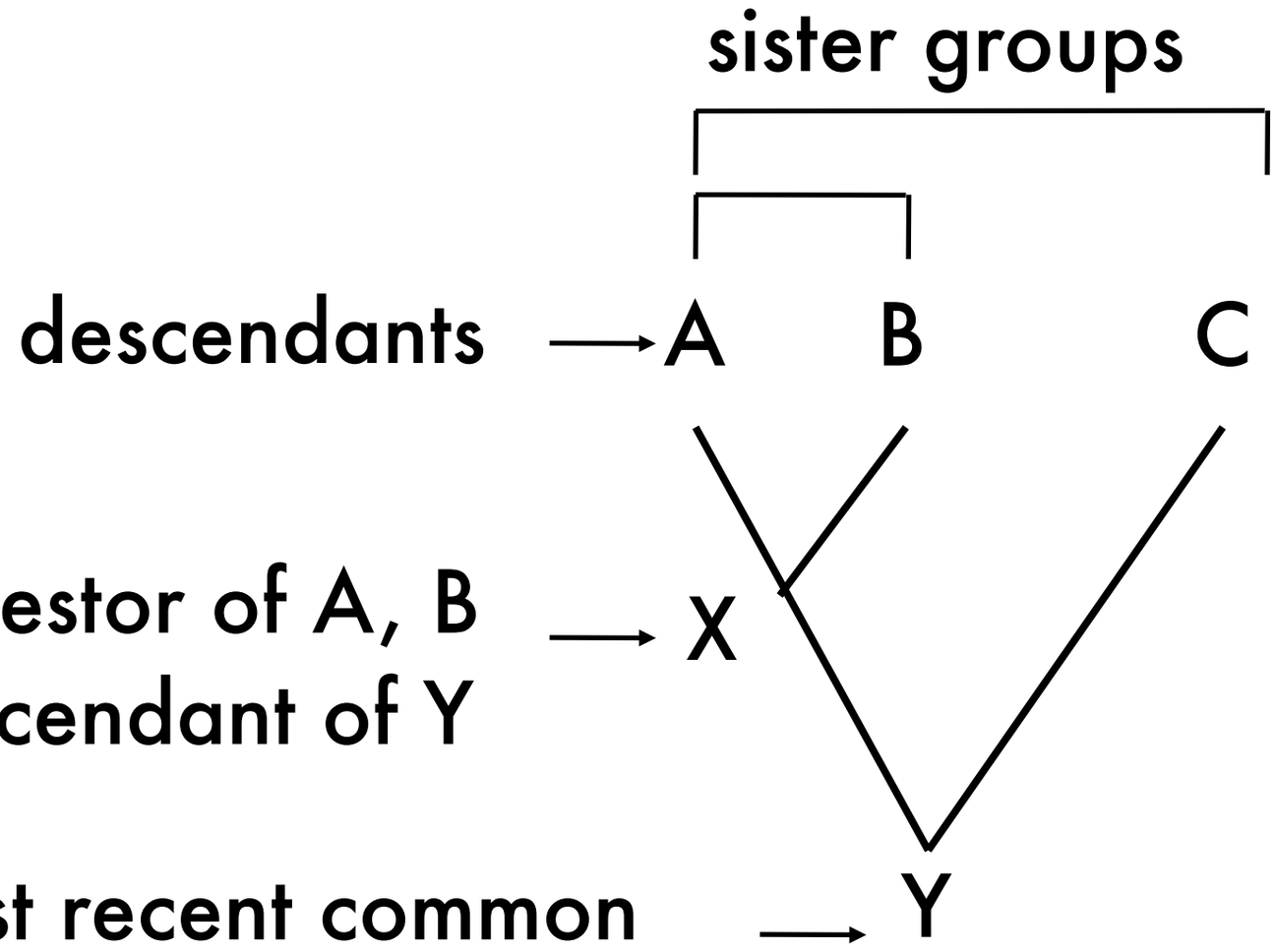
- unicellular
- eukaryotic

- unicellular
- “prokaryotic”

The Big Tree


BACTERIA
mitochondrion
Rhodocyclus
Escherichia
Desulfovibrio
chloroplast
Synechococcus
Chlamydia
Gloeobacter
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Agrobacterium
Planctomyces
Flexibacter
Flavobacterium
Methanobacterium
Thermococcus
Methanococcus
marine low temp
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
Methanothermus
Methanopyrus
marine Gp. 1 low temp
Gp. 1 low temp
Gp. 2 low temp
pSL 12
pSL 22
Gp. 3 low temp
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
Root
ARCHAEA
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum
Encephalitozoon
Vairimorpha
Trichomonas
Giardia
0.1 changes / site
EUCARYA

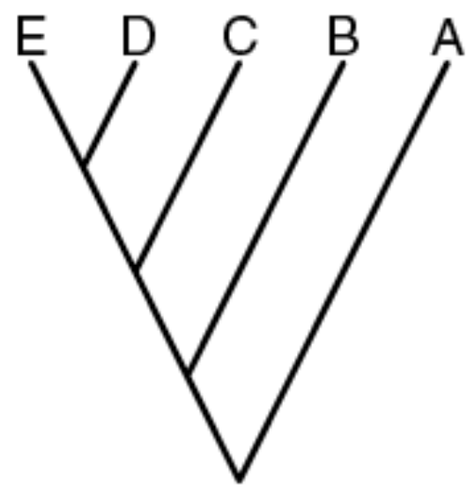
"natural" groups of organisms (or genes)

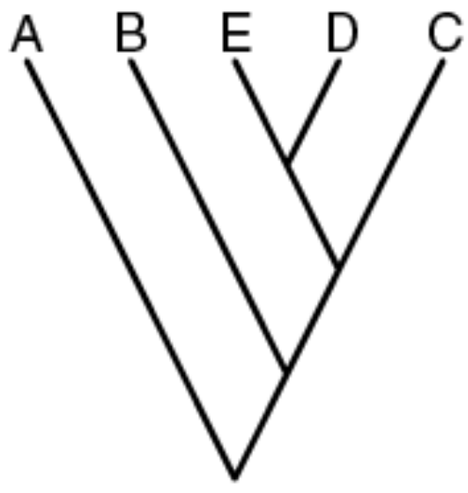


think about your own geneology.....

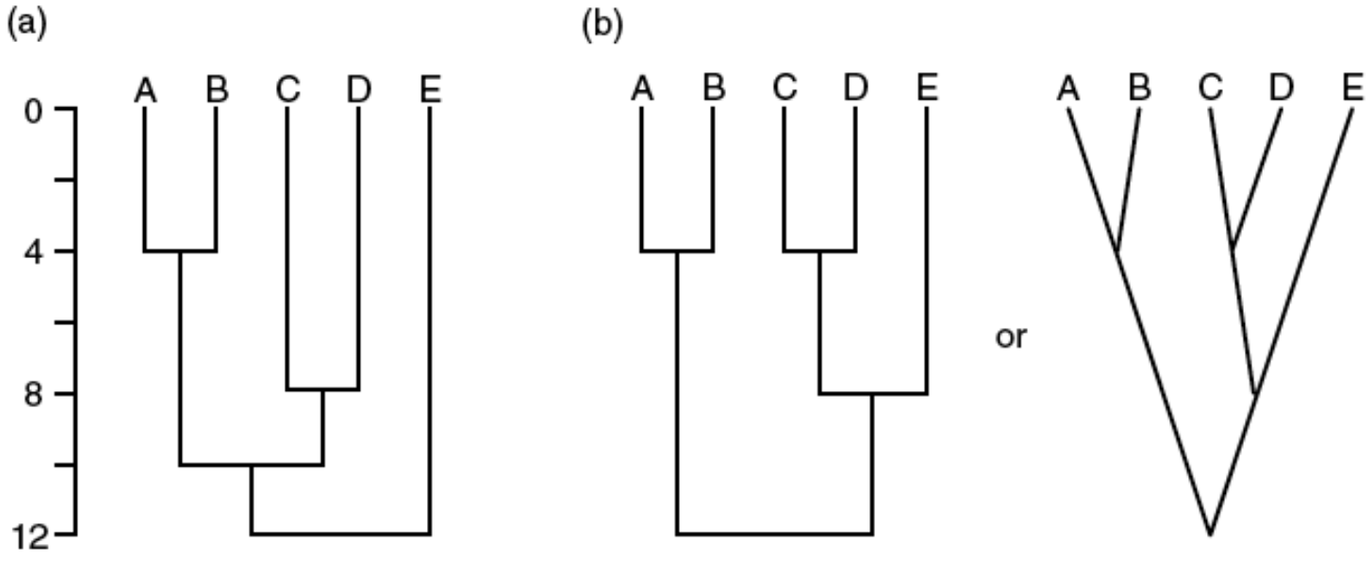
Branches can rotate around nodes

* Nodes are can rotate 360° - like a mobile

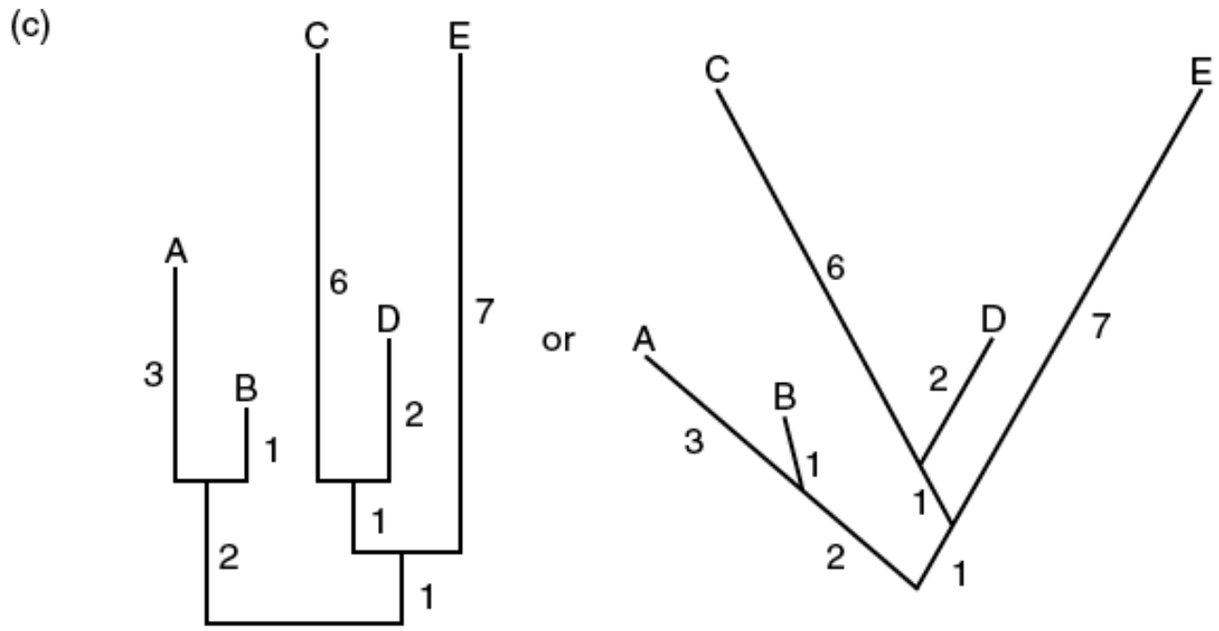




Cladograms vs. phylograms



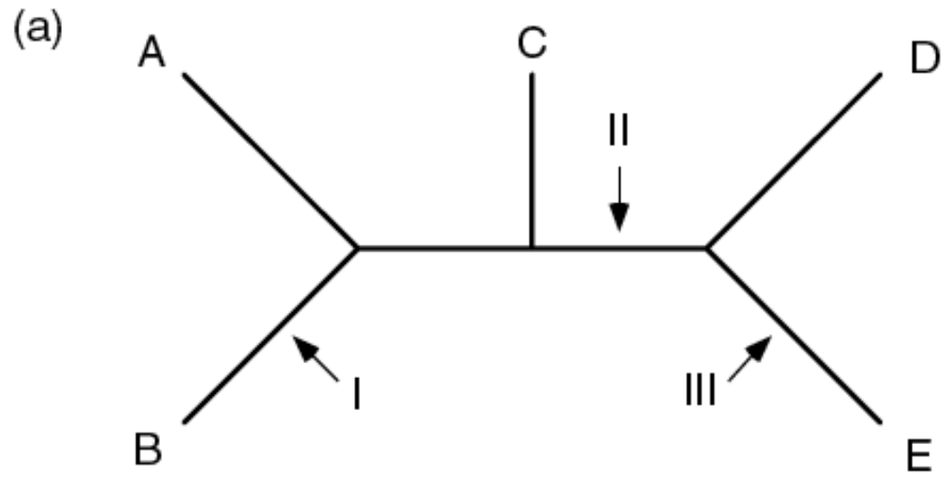
Cladograms show branching order - branch lengths are meaningless



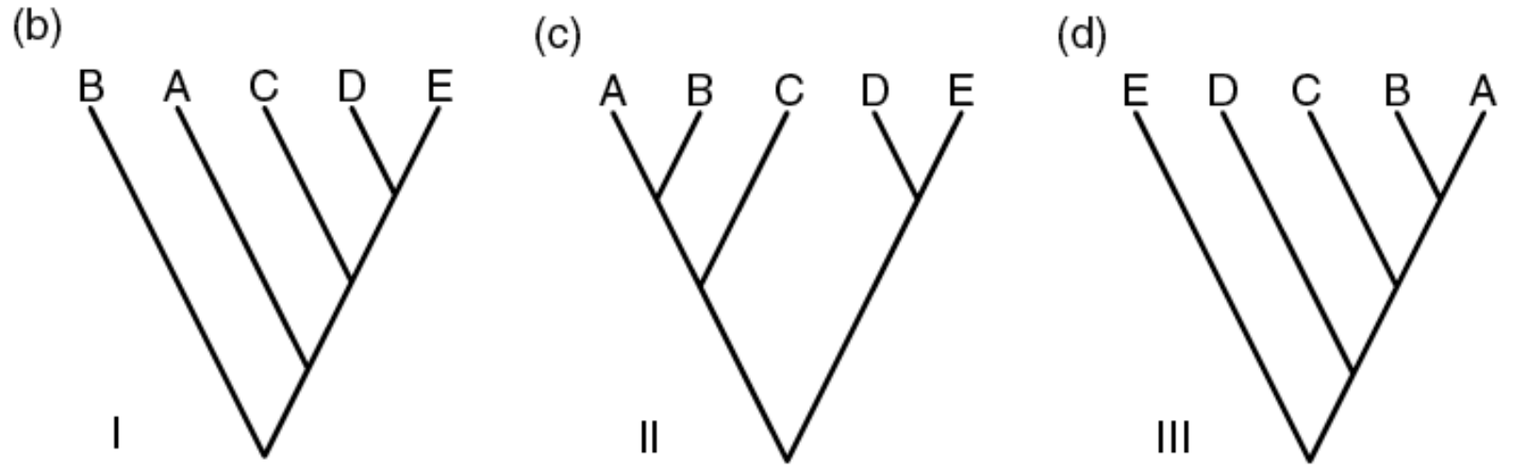
Phylograms show branch order <u>and</u> branch lengths

Branch length is directly proportional to the amount of change

Rooting using an outgroup



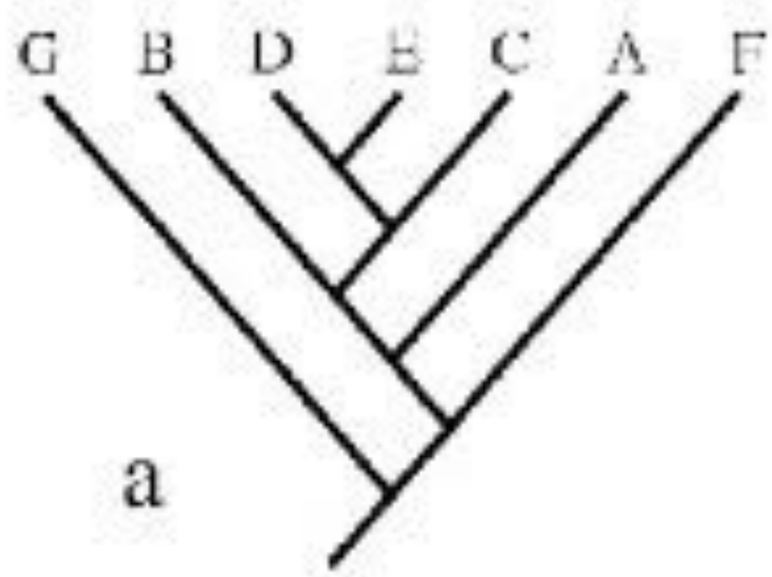
Unrooted tree

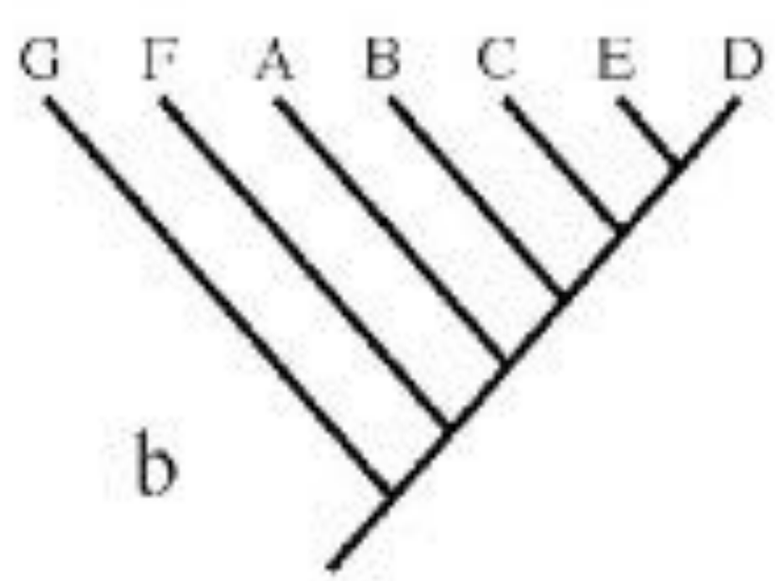


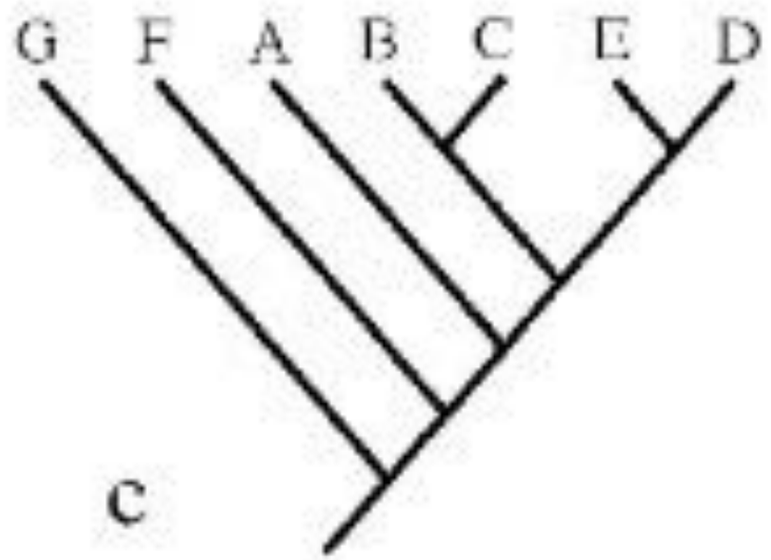
rooted trees

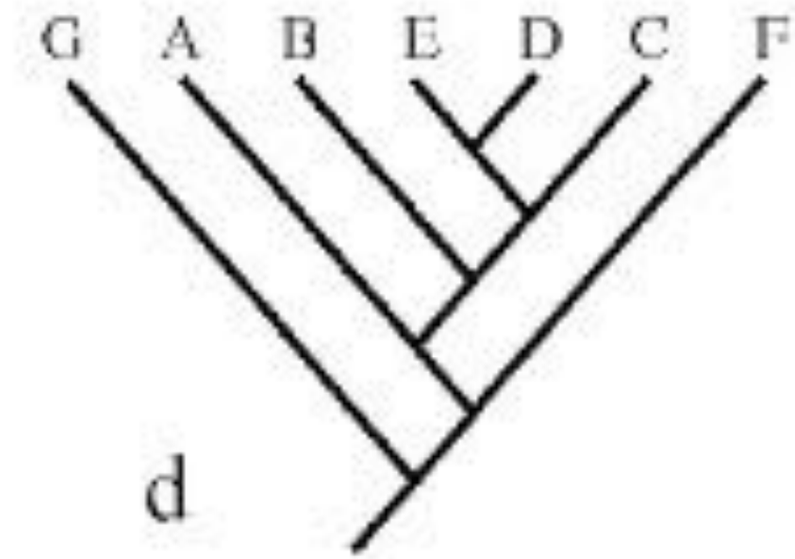
<u>Unrooted</u>:"direction" of evolution unknown, or not inferred

<u>Rooted</u>:"direction" of evolution known, or inferred





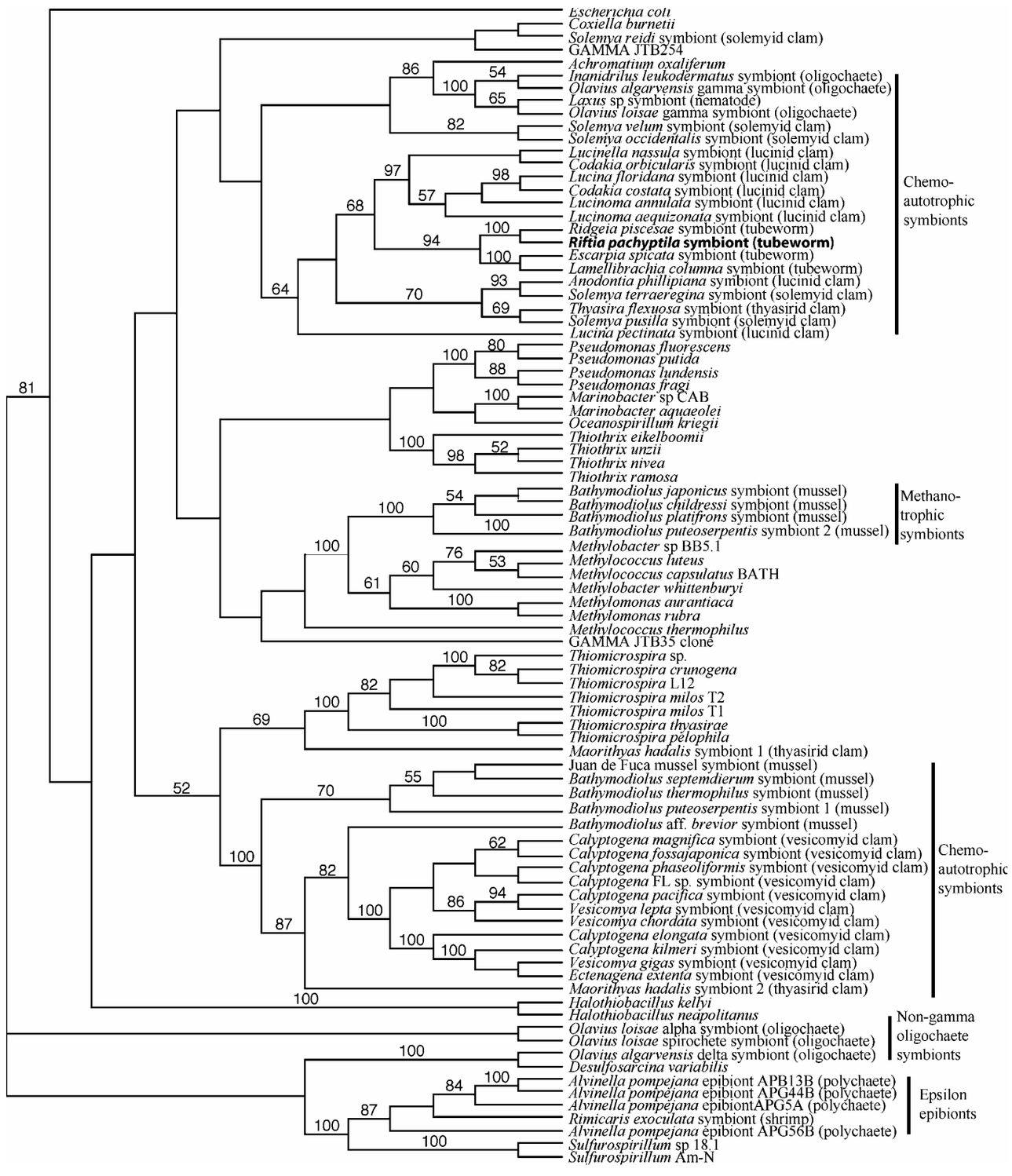


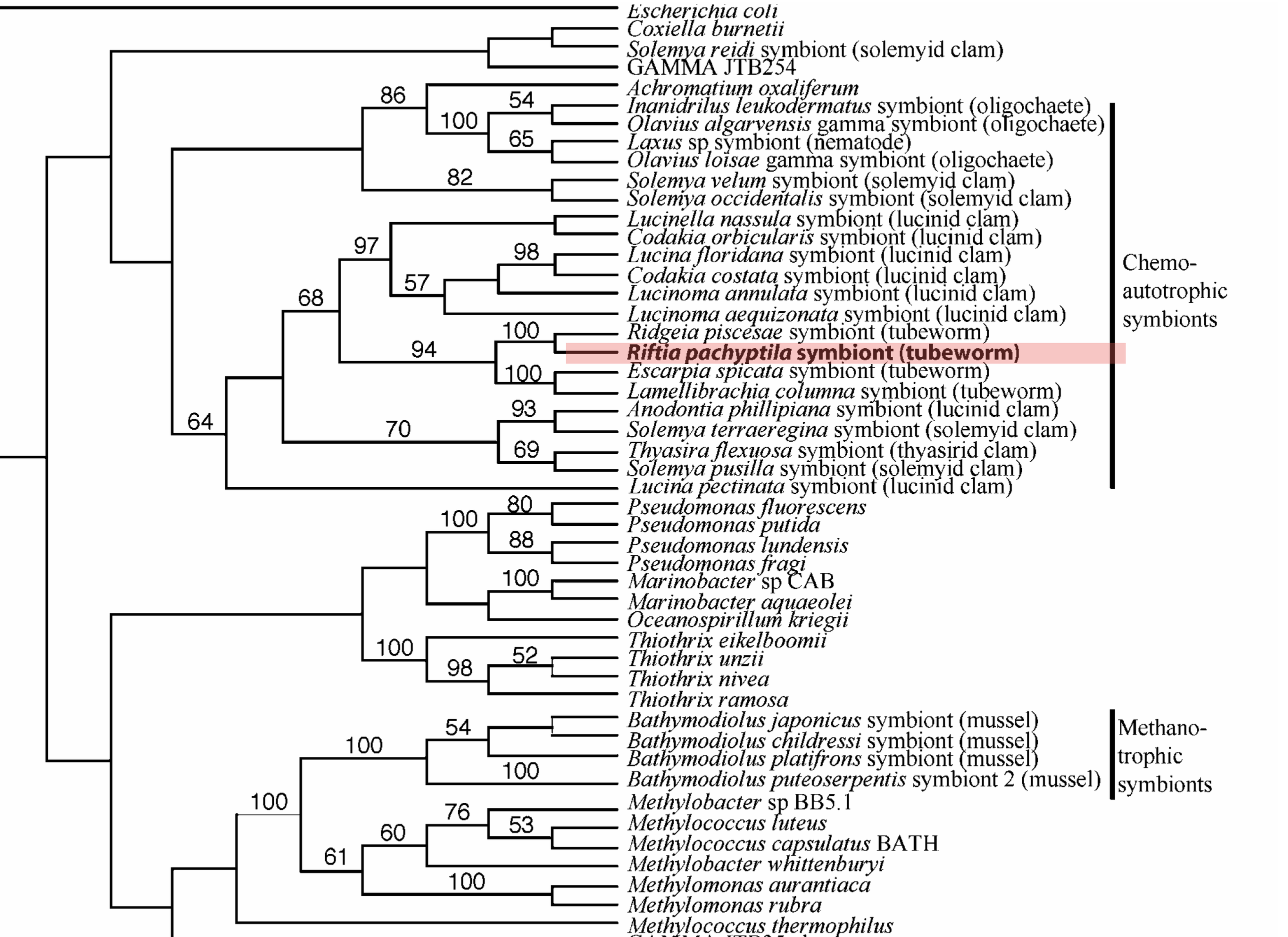


Which of the four trees above depicts a different pattern?

How do we evaluate the quality of trees?

What parts of the tree are we confident in?




Escherichia coli
Coxiella burnetii
Solemya reidi symbiont (solemyid clam)
GAMMA JTB254
Achromatium oxaliferum
Inanidrilus leukodermatus symbiont (oligochaete)
Olavius algarvensis gamma symbiont (oligochaete)
Laxus sp symbiont (nematode)
Olavius loisae gamma symbiont (oligochaete)
Solemya velum symbiont (solemyid clam)
Solemya occidentalis symbiont (solemyid clam)
Lucinella nassula symbiont (lucinid clam)
Codakia orbicularis symbiont (lucinid clam)
Lucina floridana symbiont (lucinid clam)
Codakia costata symbiont (lucinid clam)
Lucinoma annulata symbiont (lucinid clam)
Lucinoma aequizonata symbiont (lucinid clam)
Ridgeia piscesae symbiont (tubeworm)
Riftia pachyptila symbiont (tubeworm)
Escarpia spicata symbiont (tubeworm)
Lamellibrachia columna symbiont (tubeworm)
Anodontia phillipiana symbiont (lucinid clam)
Solemya terraeregina symbiont (solemyid clam)
Thyasira flexuosa symbiont (thyasirid clam)
Solemya pusilla symbiont (solemyid clam)
Lucina pectinata symbiont (lucinid clam)
Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas lundensis
Pseudomonas fragi
Marinobacter sp CAB
Marinobacter aquaeolei
Oceanospirillum kriegii
Thiothrix eikelboomii
Thiothrix unzii
Thiothrix nivea
Thiothrix ramosa
Bathymodiolus japonicus symbiont (mussel)
Bathymodiolus childressi symbiont (mussel)
Bathymodiolus platifrons symbiont (mussel)
Bathymodiolus puteoserpentis symbiont 2 (mussel)
Methylobacter sp BB5.1
Methylococcus luteus
Methylococcus capsulatus BATH
Methylobacter whittenburyi
Methylomonas aurantiaca
Methylomonas rubra
Methylococcus thermophilus
86
54
100
65
82
97
98
57
68
100
94
100
64
93
70
69
80
100
88
100
100
52
98
54
100
100
100
76
53
60
61
100
Chemo-autotrophic symbionts
Methano-trophic symbionts

* Non-parametric bootstrapping

- Analyze original dataset
- Generate new samples by resampling original data (n=# columns, n times)
- Analyze new datasets like original
- Summarize variation ("consensus")

- Bootstrap values = variation in tree topologies (generally want >50)

*

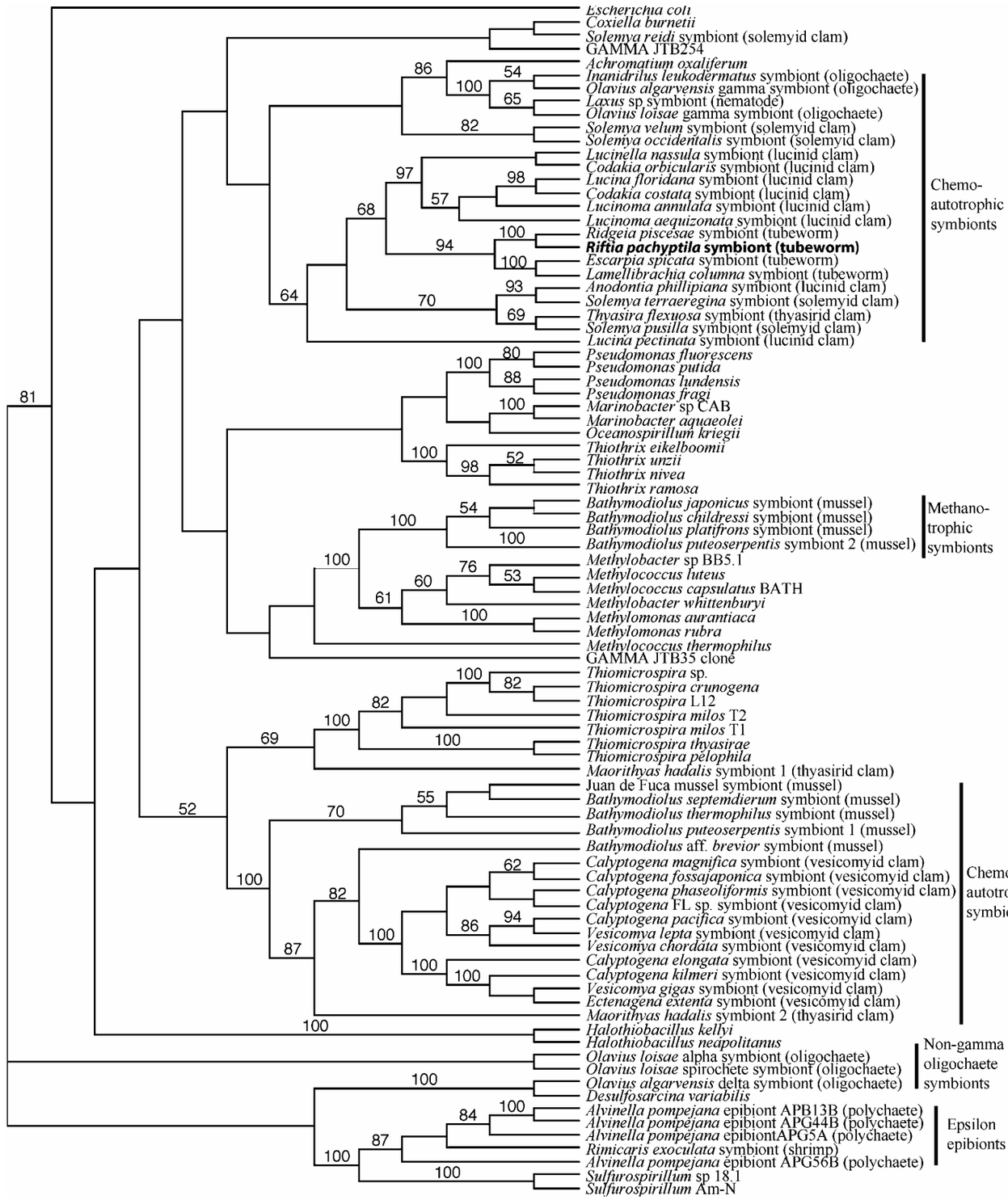
Bootstrapping

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

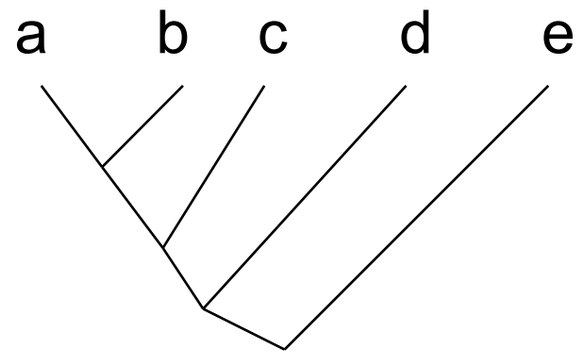
  (out of the X times the analysis is done).

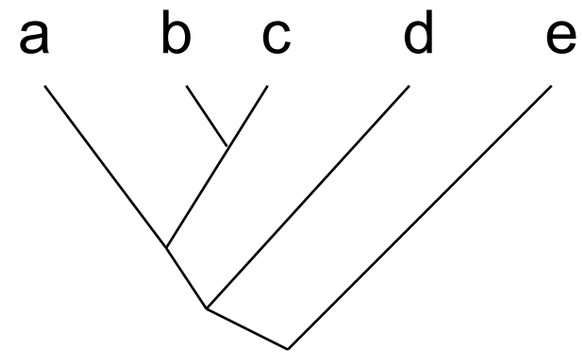
- used to estimate probability and indicated on a consensus tree

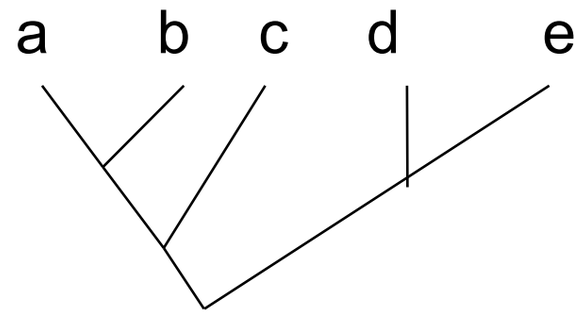
Alignment and evolutionary assumptions are critical

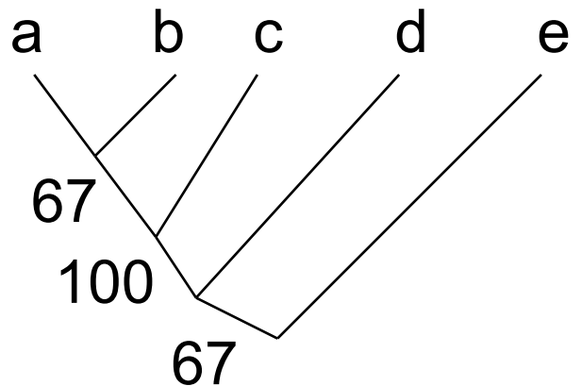


Majority rule Consensus



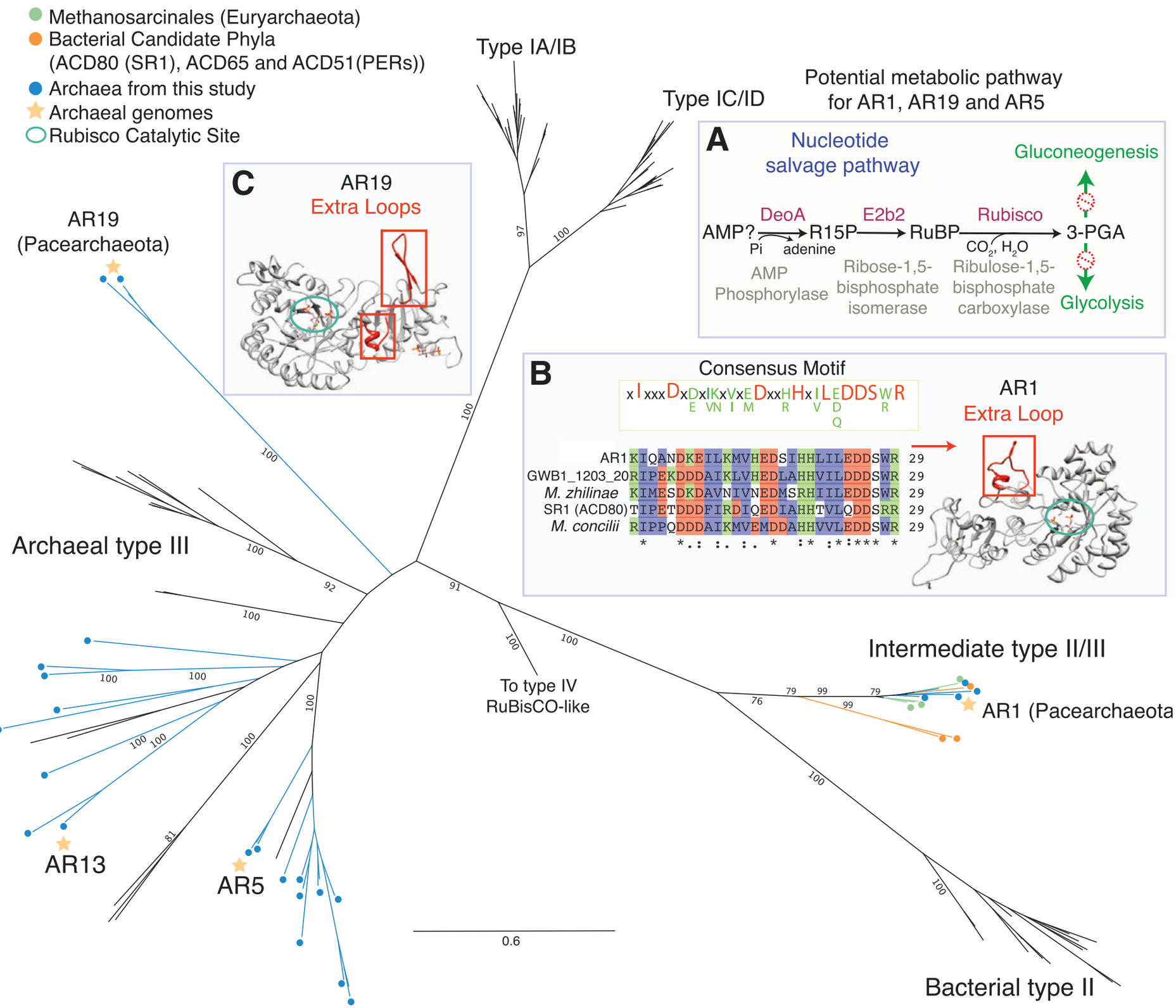


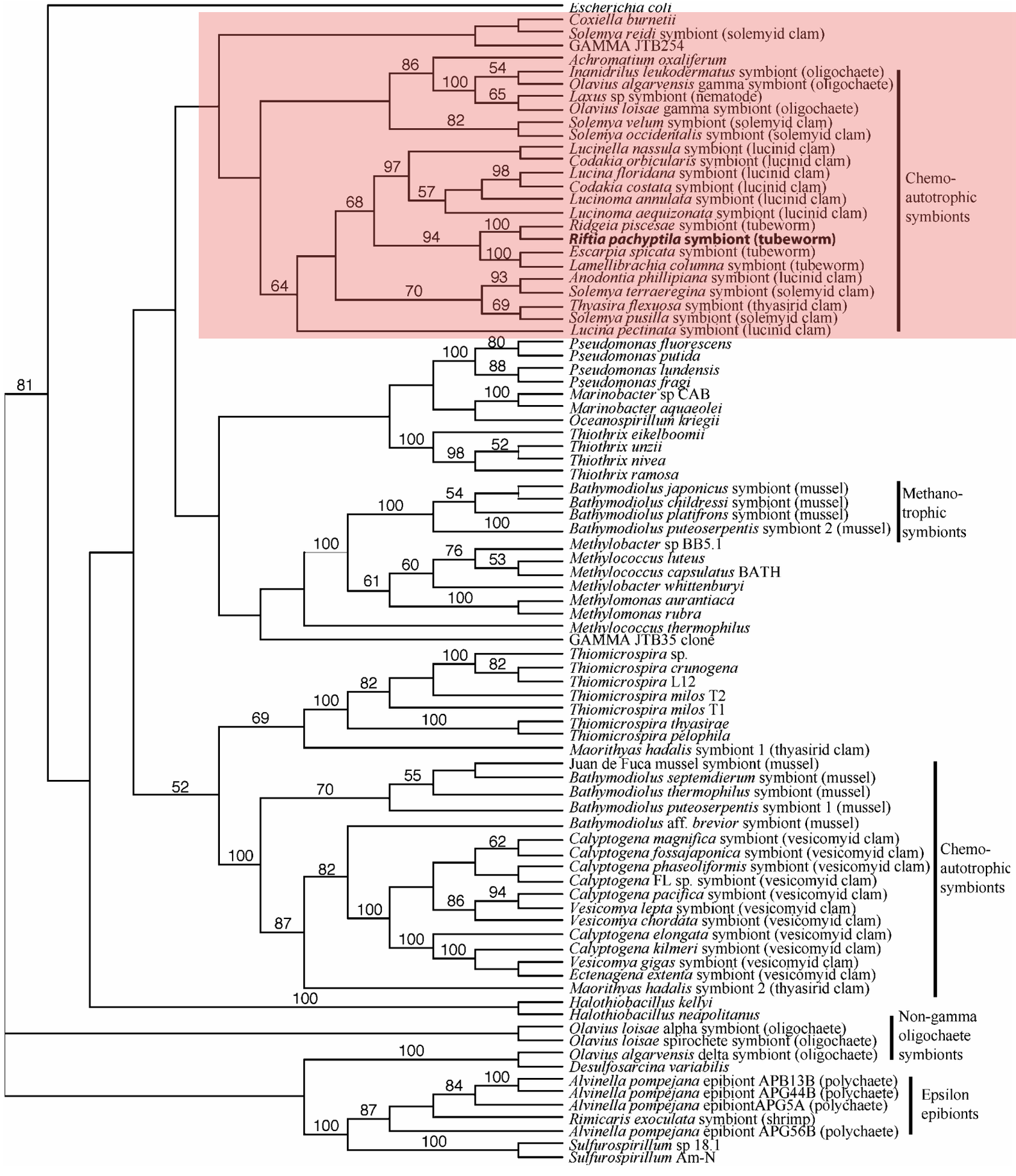




The percentage of the trees supporting each splits are indicated

Large subunit RuBisCo phylogeny




Escherichia coli
Coxiella burnetii
Solemya reidi symbiont (solemyid clam)
GAMMA JTB254
Achromatium oxaliferum
Inanidrilus leukodermatus symbiont (oligochaete)
Olavius algarvensis gamma symbiont (oligochaete)
Laxus sp symbiont (nematode)
Olavius loisae gamma symbiont (oligochaete)
Solemya velum symbiont (solemyid clam)
Solemya occidentalis symbiont (solemyid clam)
Lucinella nassula symbiont (lucinid clam)
Codakia orbicularis symbiont (lucinid clam)
Lucina floridana symbiont (lucinid clam)
Codakia costata symbiont (lucinid clam)
Lucinoma annulata symbiont (lucinid clam)
Lucinoma aequizonata symbiont (lucinid clam)
Ridgeia piscesae symbiont (tubeworm)
Riftia pachyptila symbiont (tubeworm)
Escarpia spicata symbiont (tubeworm)
Lamellibrachia columna symbiont (tubeworm)
Anodontia phillipiana symbiont (lucinid clam)
Solemya terraeregina symbiont (solemyid clam)
Thyasira flexuosa symbiont (thyasirid clam)
Solemya pusilla symbiont (solemyid clam)
Lucina pectinata symbiont (lucinid clam)
Chemo-autotrophic symbionts
Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas lundensis
Pseudomonas fragi
Marinobacter sp CAB
Marinobacter aquaeolei
Oceanospirillum kriegii
Thiothrix eikelboomii
Thiothrix unzii
Thiothrix nivea
Thiothrix ramosa
Bathymodiolus japonicus symbiont (mussel)
Bathymodiolus childressi symbiont (mussel)
Bathymodiolus platifrons symbiont (mussel)
Bathymodiolus puteoserpentis symbiont 2 (mussel)
Methylobacter sp BB5.1
Methylococcus luteus
Methylococcus capsulatus BATH
Methylobacter whittenburyi
Methylomonas aurantiaca
Methylomonas rubra
Methylococcus thermophilus
GAMMA JTB35 clone
Methano-trophic symbionts
Thiomicrospira sp.
Thiomicrospira crunogena
Thiomicrospira L12
Thiomicrospira milos T2
Thiomicrospira milos T1
Thiomicrospira thyasirae
Thiomicrospira pelophila
Maorithyas hadalis symbiont 1 (thyasirid clam)
Juan de Fuca mussel symbiont (mussel)
Bathymodiolus septemdierum symbiont (mussel)
Bathymodiolus thermophilus symbiont (mussel)
Bathymodiolus puteoserpentis symbiont 1 (mussel)
Bathymodiolus aff. brevior symbiont (mussel)
Calyptogena magnifica symbiont (vesicomyid clam)
Calyptogena fossajaponica symbiont (vesicomyid clam)
Calyptogena phaseoliformis symbiont (vesicomyid clam)
Calyptogena FL sp. symbiont (vesicomyid clam)
Calyptogena pacifica symbiont (vesicomyid clam)
Vesicomya lepta symbiont (vesicomyid clam)
Vesicomya chordata symbiont (vesicomyid clam)
Calyptogena elongata symbiont (vesicomyid clam)
Calyptogena kilmeri symbiont (vesicomyid clam)
Vesicomya gigas symbiont (vesicomyid clam)
Ectenagena extenta symbiont (vesicomyid clam)
Maorithyas hadalis symbiont 2 (thyasirid clam)
Chemo-autotrophic symbionts
Halothiobacillus kellyi
Halothiobacillus neapolitanus
Olavius loisae alpha symbiont (oligochaete)
Olavius loisae spirochete symbiont (oligochaete)
Olavius algarvensis delta symbiont (oligochaete)
Desulfosarcina variabilis
Non-gamma oligochaete symbionts
Alvinella pompejana epibiont APB13B (polychaete)
Alvinella pompejana epibiont APG44B (polychaete)
Alvinella pompejana epibiontAPG5A (polychaete)
Rimicaris exoculata symbiont (shrimp)
Alvinella pompejana epibiont APG56B (polychaete)
Epsilon epibionts
Sulfurospirillum sp 18.1
Sulfurospirillum Am-N

* Monophyly, Paraphyly, and Polyphyly

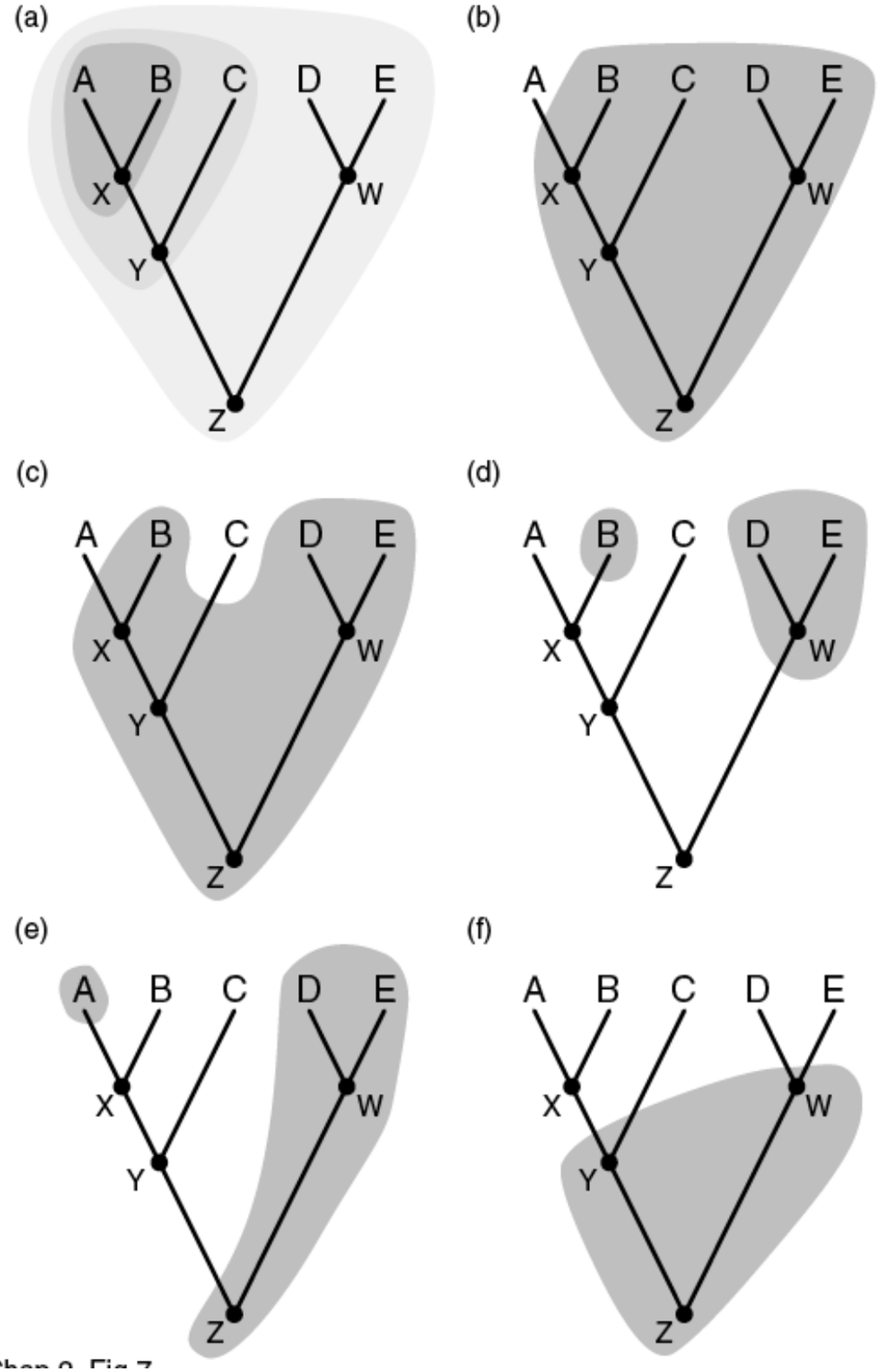
"natural" groups = monophyletic

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

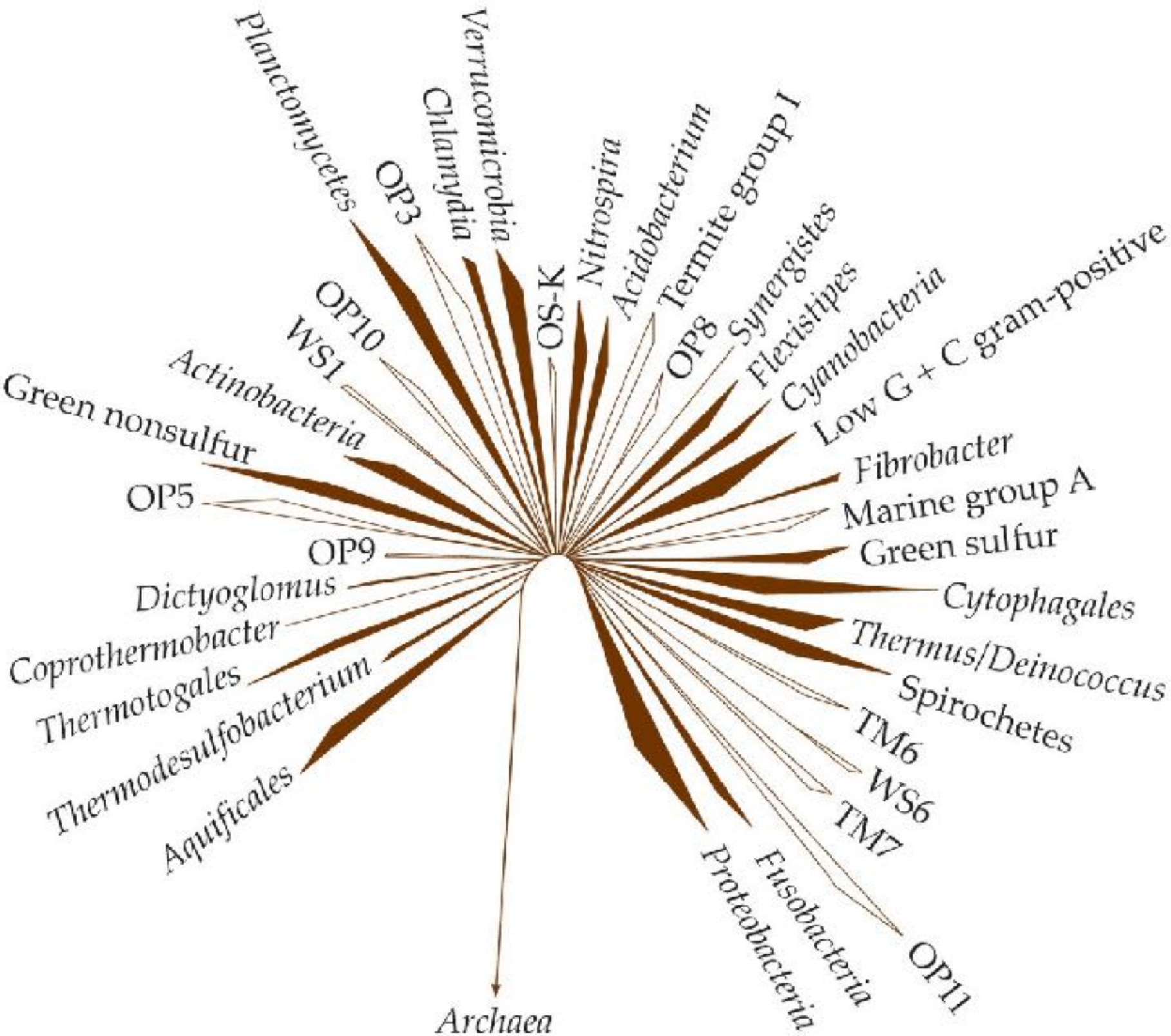
2. polyphyletic: multiple ancestries

3. paraphyletic: ancestor but not all descendants

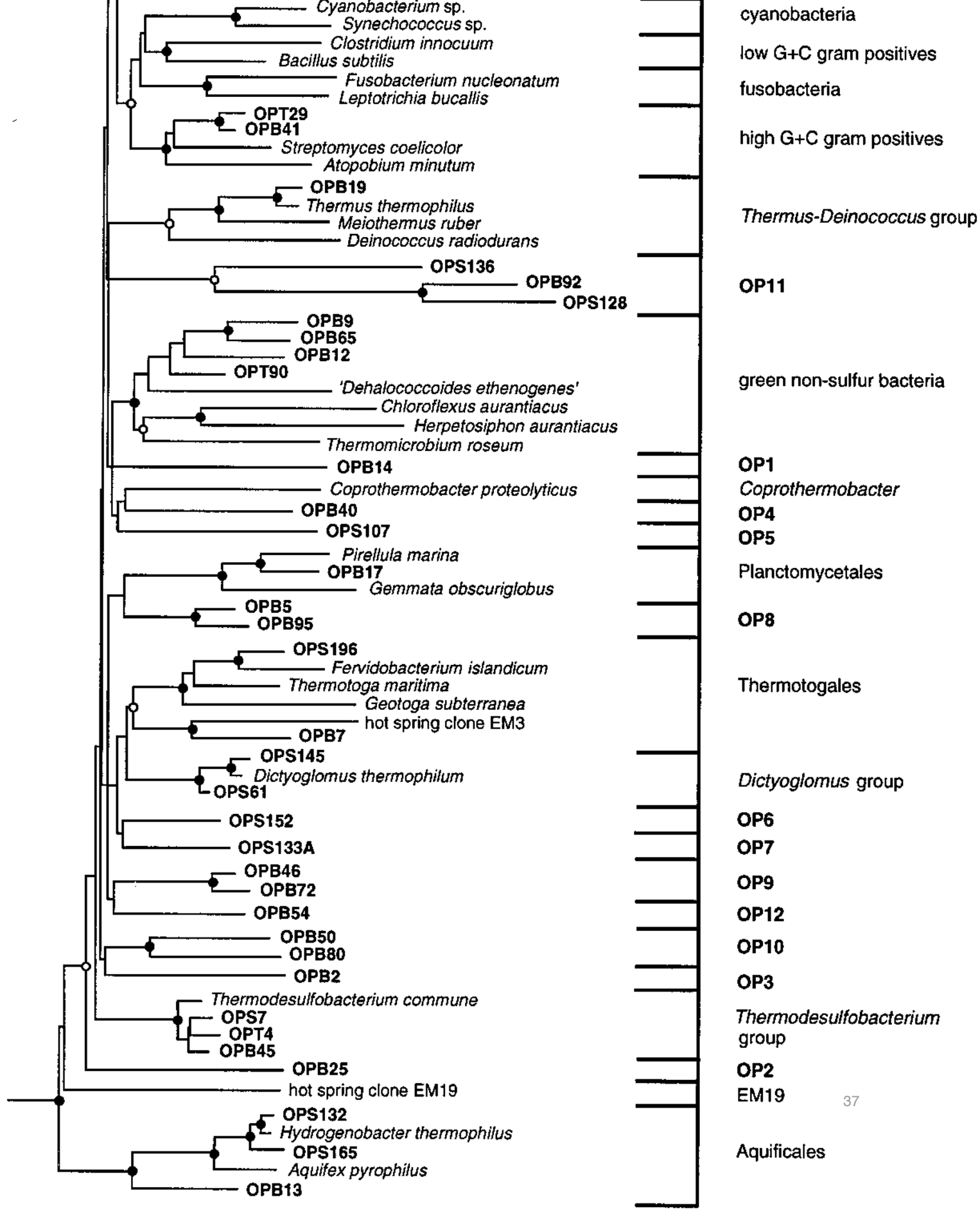
What about dark groups to the right?



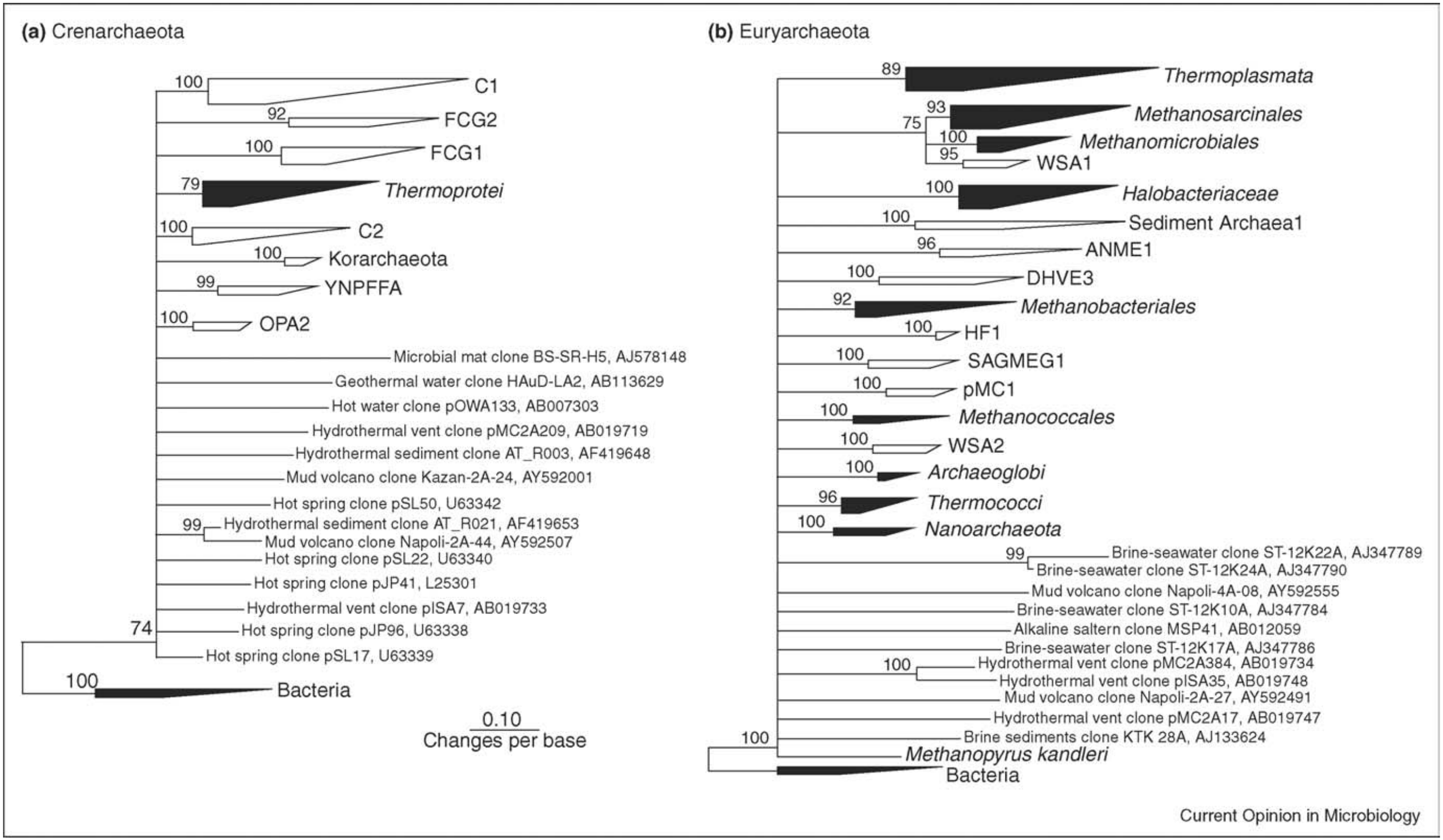
Bacterial divisions



MICROBIAL LIFE, **Box 17.5** © 2002 Sinauer Associates, Inc

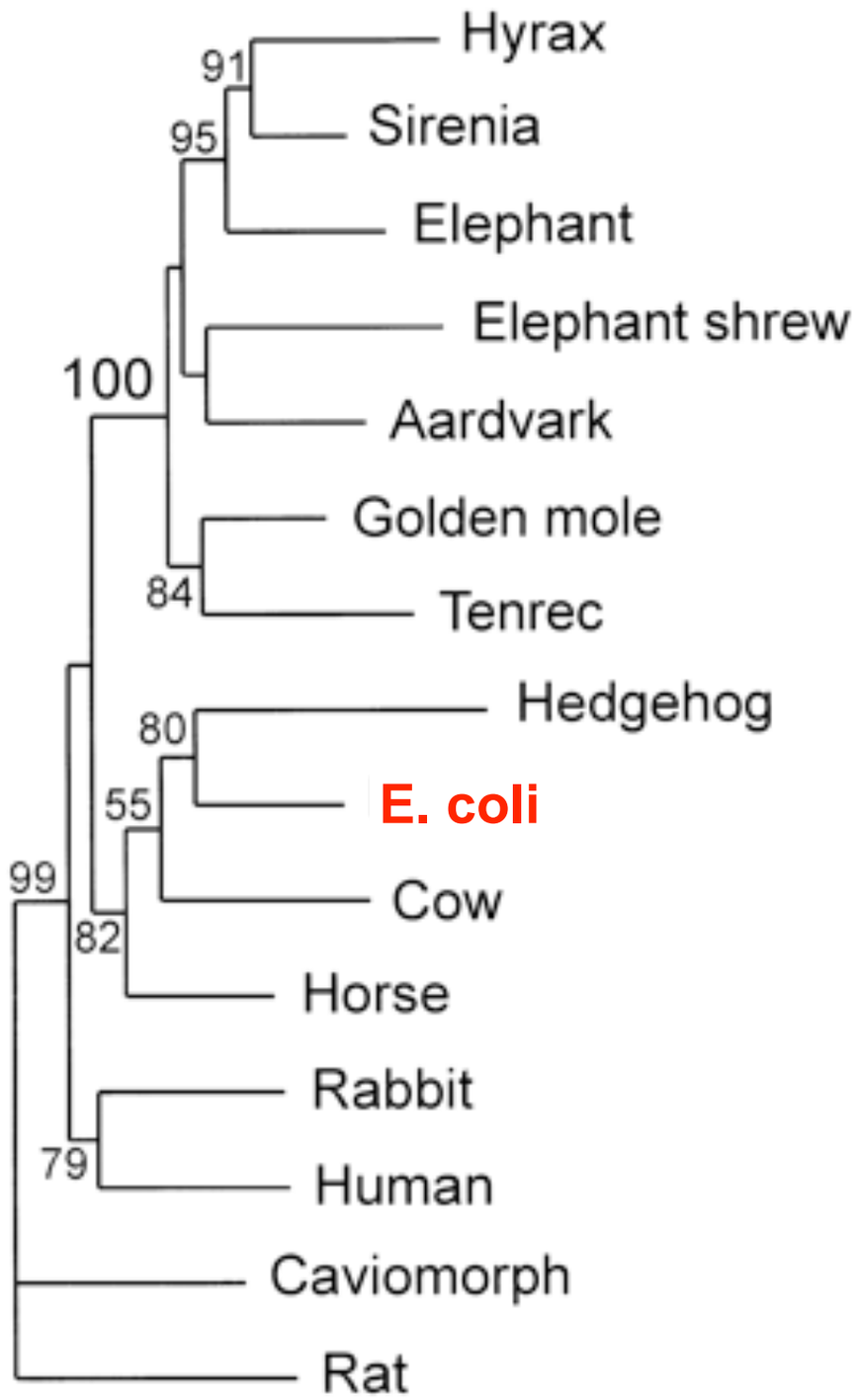

Cyanobacterium sp.
Synechococcus sp.
Clostridium innocuum
Bacillus subtilis
Fusobacterium nucleonatum
Leptotrichia bucallis
OPT29
OPB41
Streptomyces coelicolor
Atopobium minutum
OPB19
Thermus thermophilus
Meiothermus ruber
Deinococcus radiodurans
OPS136
OPB92
OPS128
OPB9
OPB65
OPB12
OPT90
'Dehalococcoides ethenogenes'
Chloroflexus aurantiacus
Herpetosiphon aurantiacus
Thermomicrobium roseum
OPB14
Coprothermobacter proteolyticus
OPB40
OPS107
Pirellula marina
OPB17
Gemmata obscuriglobus
OPB5
OPB95
OPS196
Fervidobacterium islandicum
Thermotoga maritima
Geotoga subterranea
hot spring clone EM3
OPB7
OPS145
Dictyoglomus thermophilum
OPS61
OPS152
OPS133A
OPB46
OPB72
OPB54
OPB50
OPB80
OPB2
Thermodesulfobacterium commune
OPS7
OPT4
OPB45
OPB25
hot spring clone EM19
OPS132
Hydrogenobacter thermophilus
OPS165
Aquifex pyrophilus
OPB13
0.10
cyanobacteria
low G+C gram positives
fusobacteria
high G+C gram positives
Thermus-Deinococcus group
OP11
green non-sulfur bacteria
OP1
Coprothermobacter
OP4
OP5
Planctomycetales
OP8
Thermotogales
Dictyoglomus group
OP6
OP7
OP9
OP12
OP10
OP3
Thermodesulfobacterium group
OP2
EM19
Aquificales

Archaeal Diversity (ssu rRNA)



Horizontal Gene Transfer?

Xenologs: Horizontal gene transfer



Horizontal gene transfer: genes acquired not through common "vertical" ancestry

*

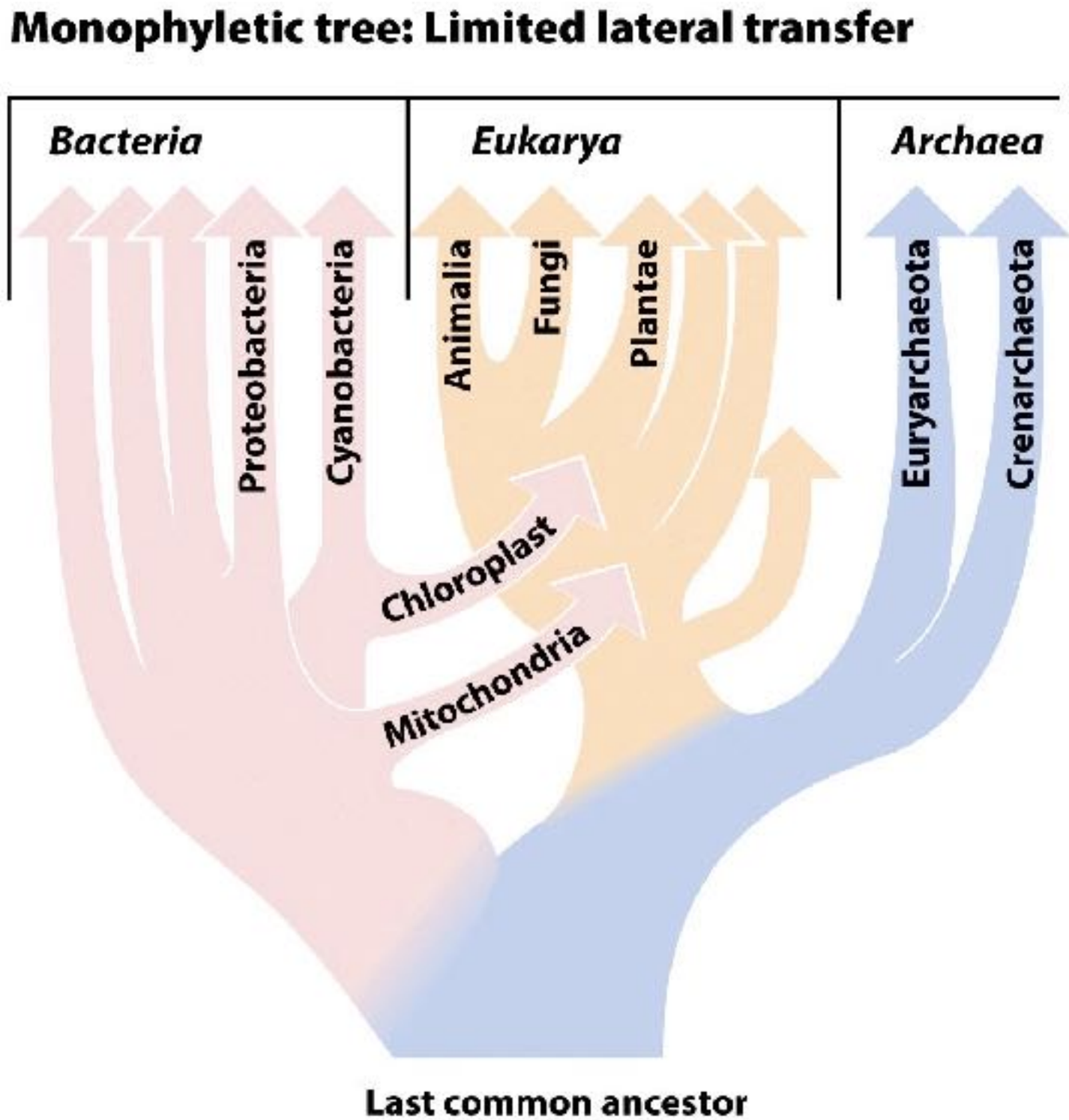
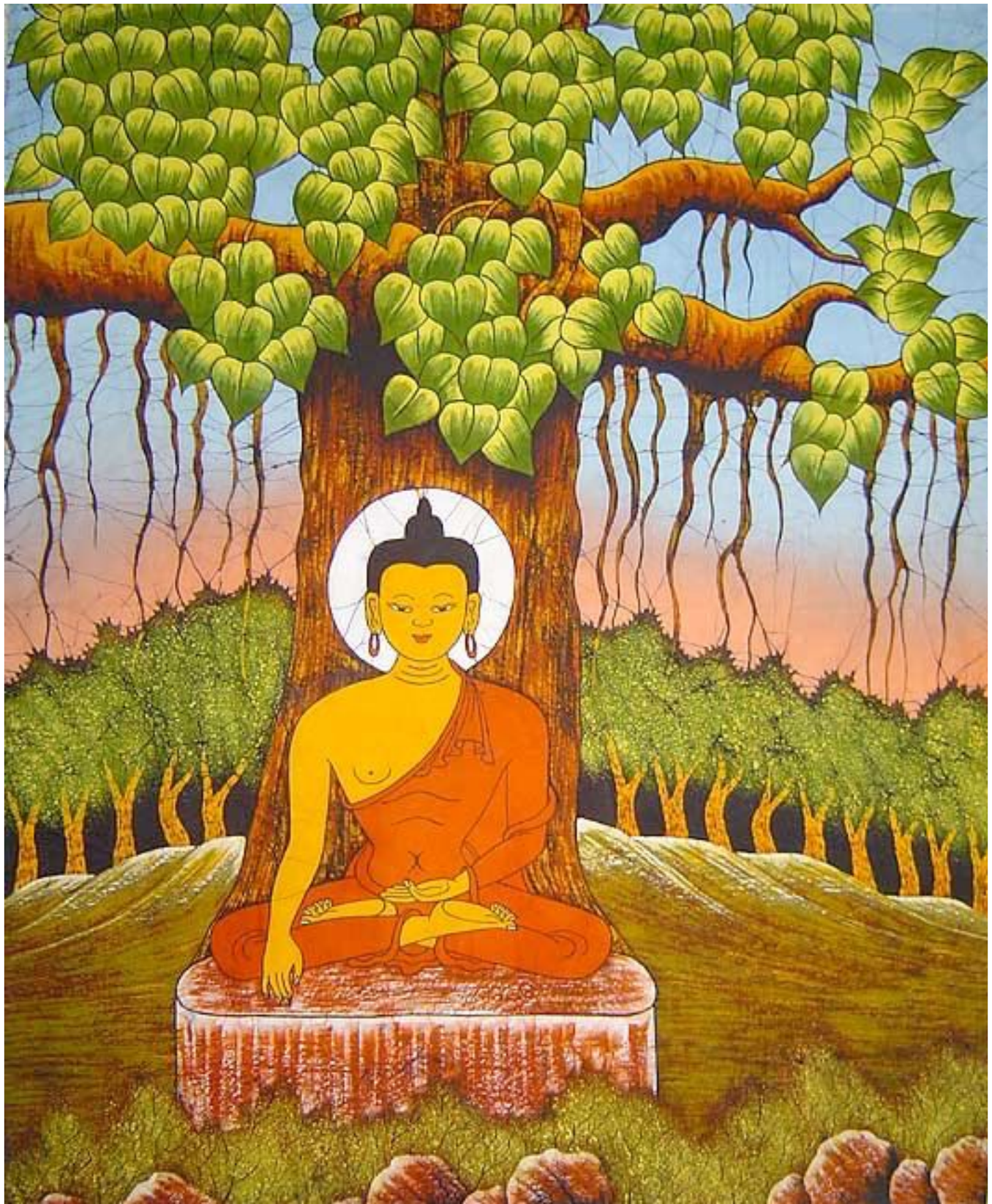


Figure 17.27a Microbiology: An Evolving Science
Source: W. Ford Doolittle. 1999. *Science* 284:2126.

Alpha proteobacterial genes in eukaryotes?
Cyanobacterial genes in eukaryotes?

What's the sound of ten species diverging?

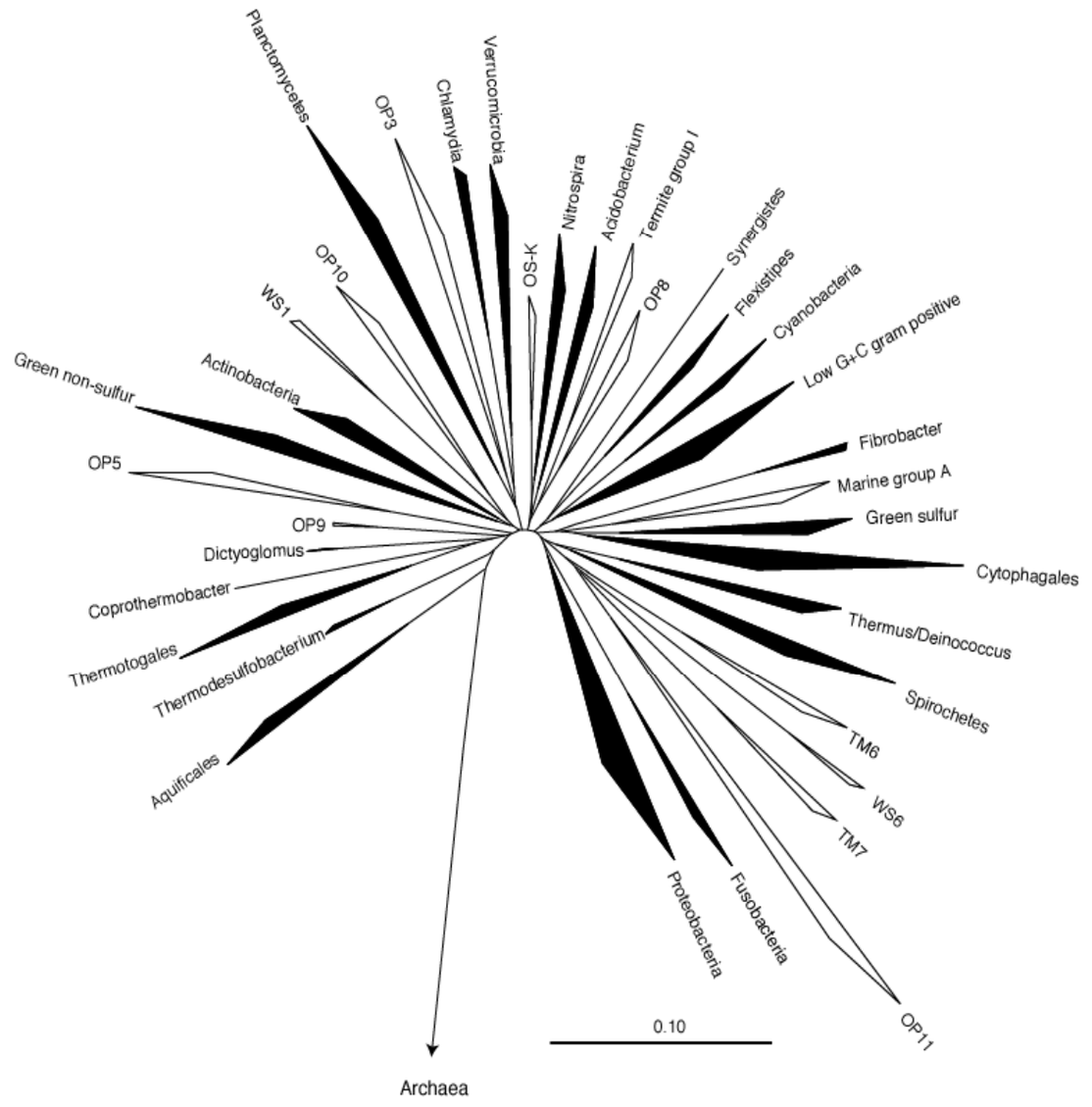


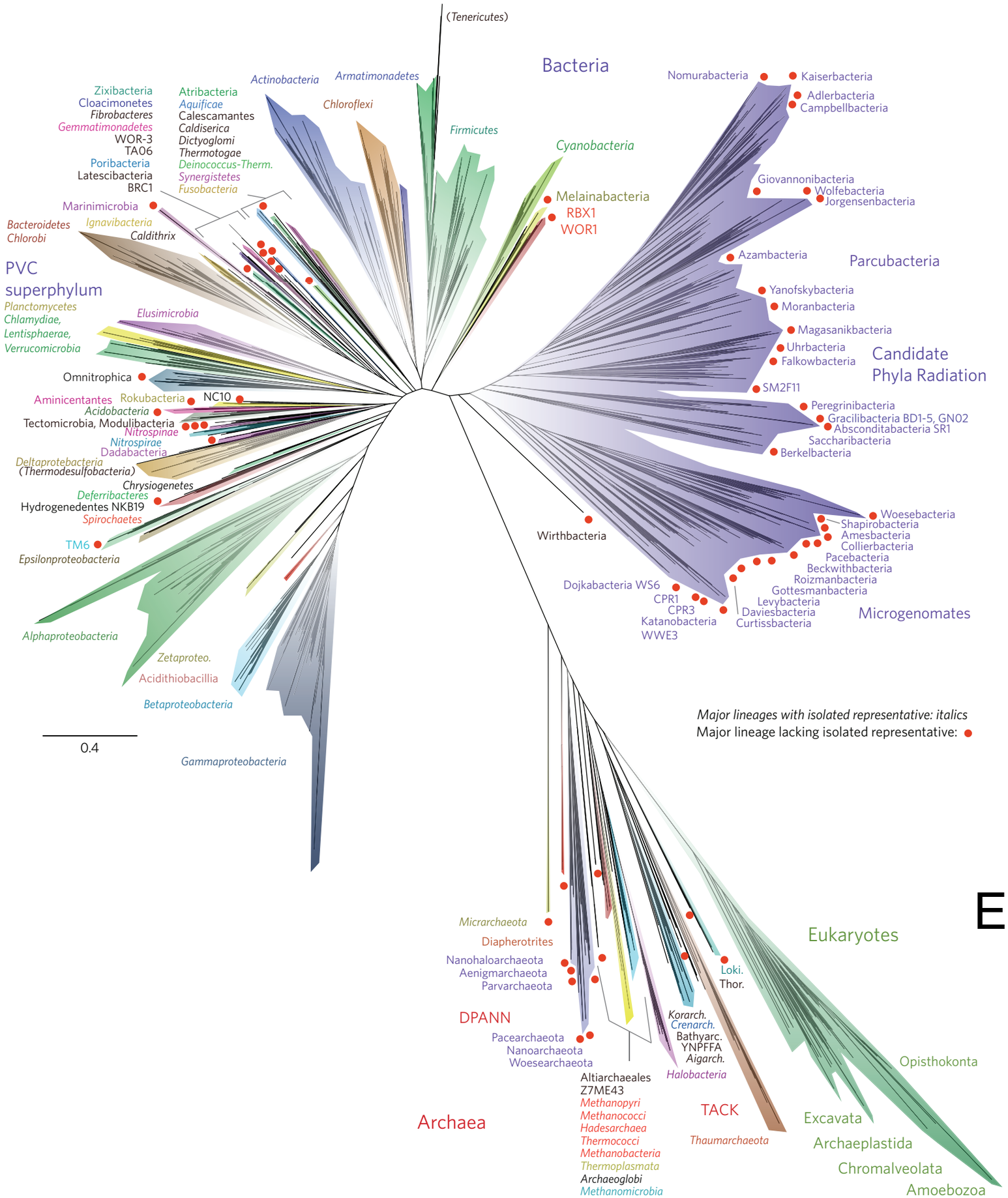
Polytomy:

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.

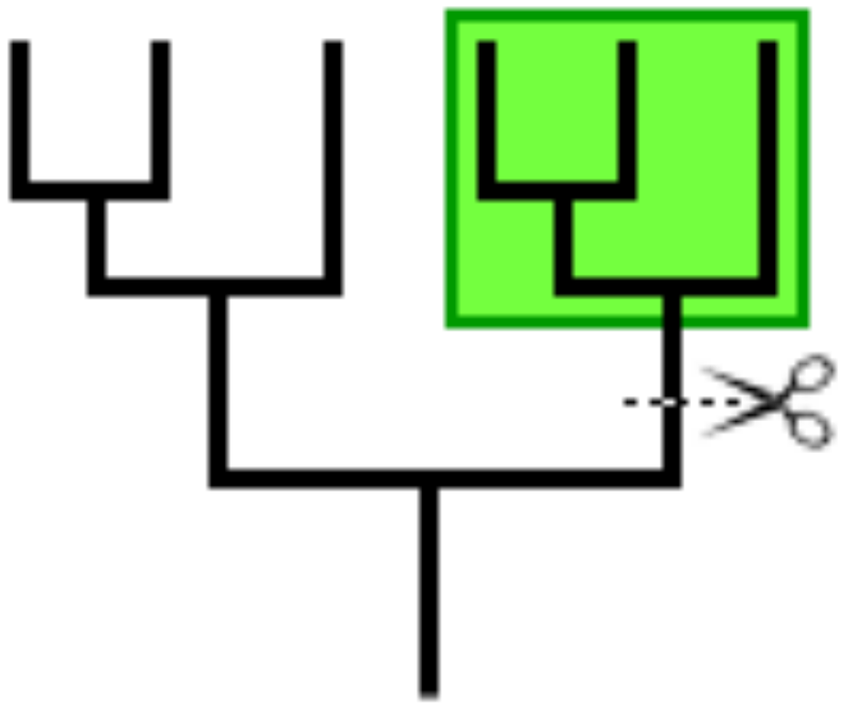


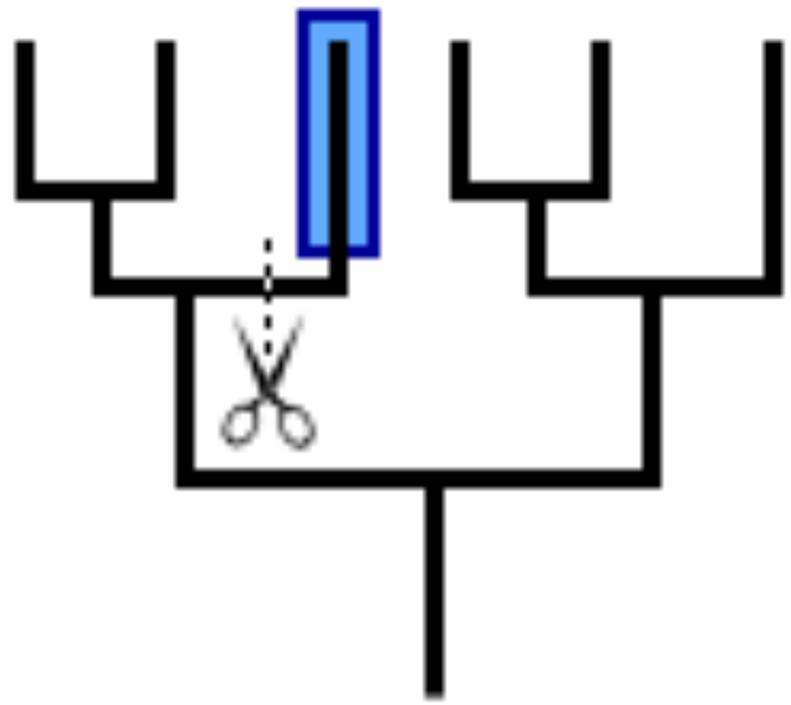


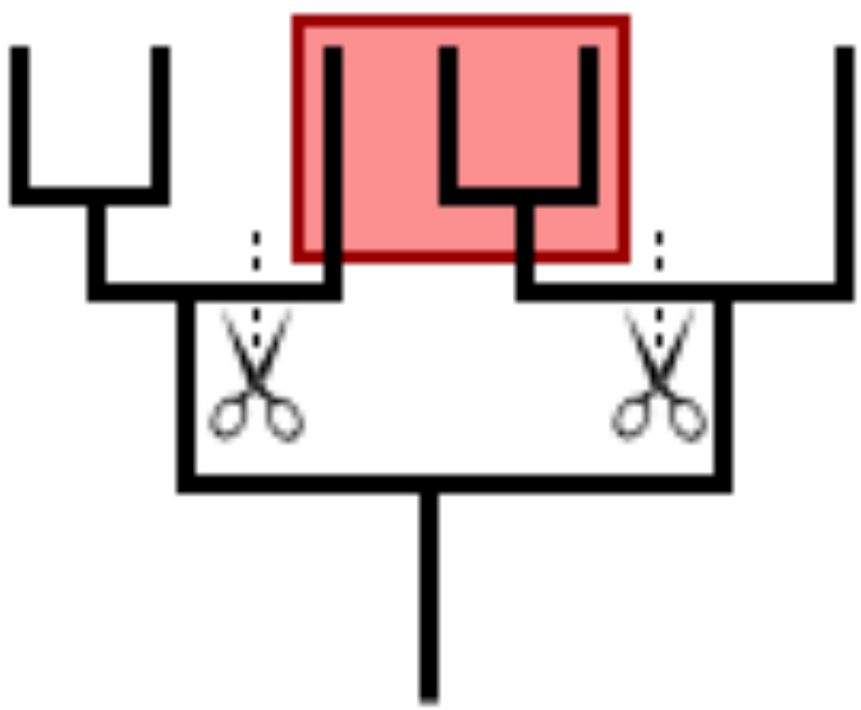
What's the scale?

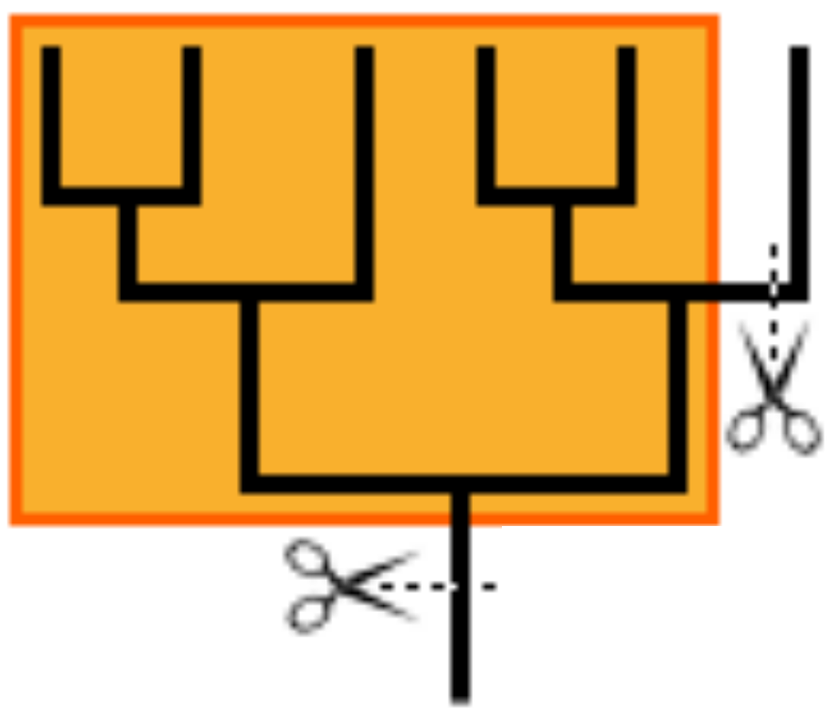
Euks are archaea?

"scissor test" for monophyly

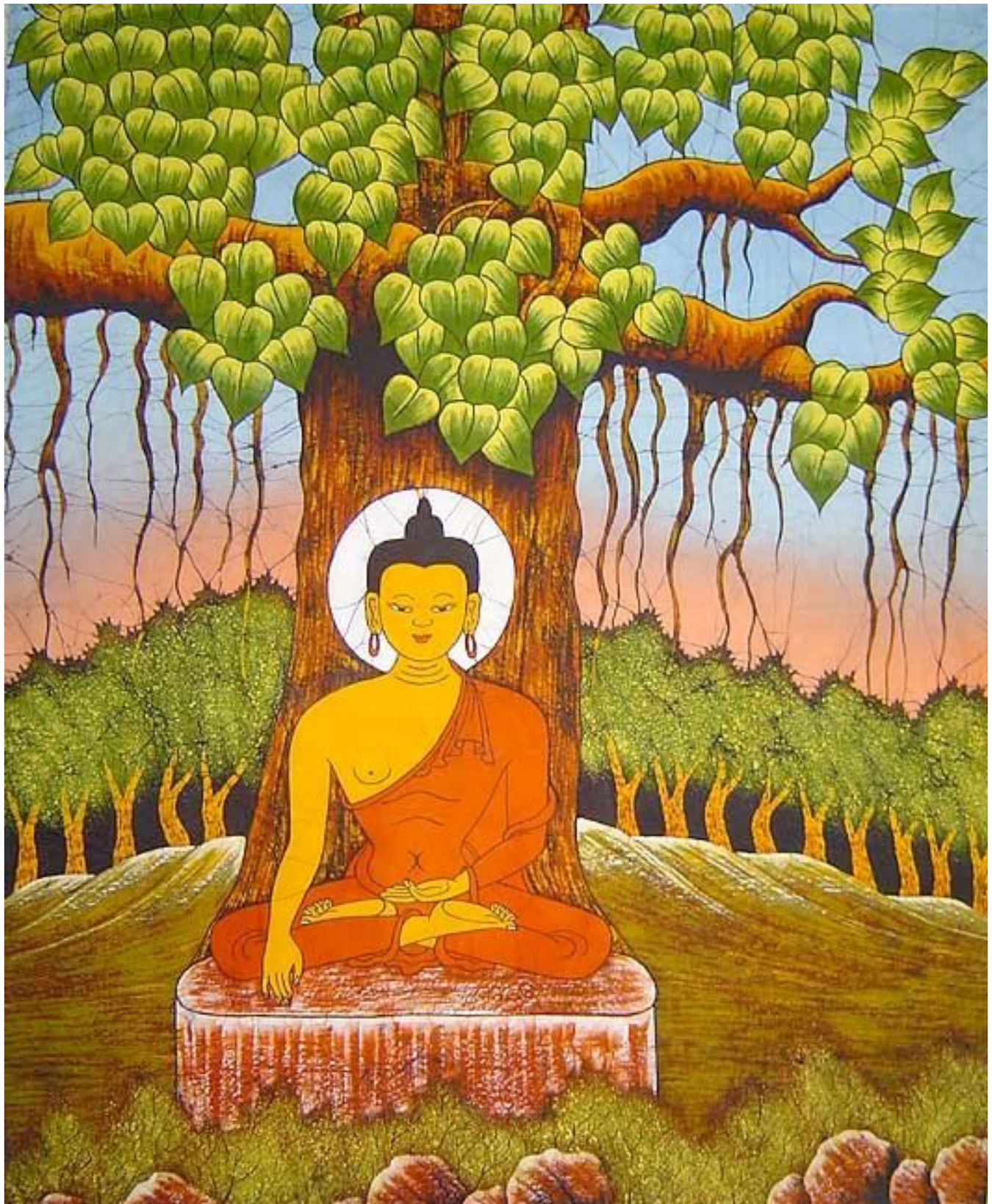




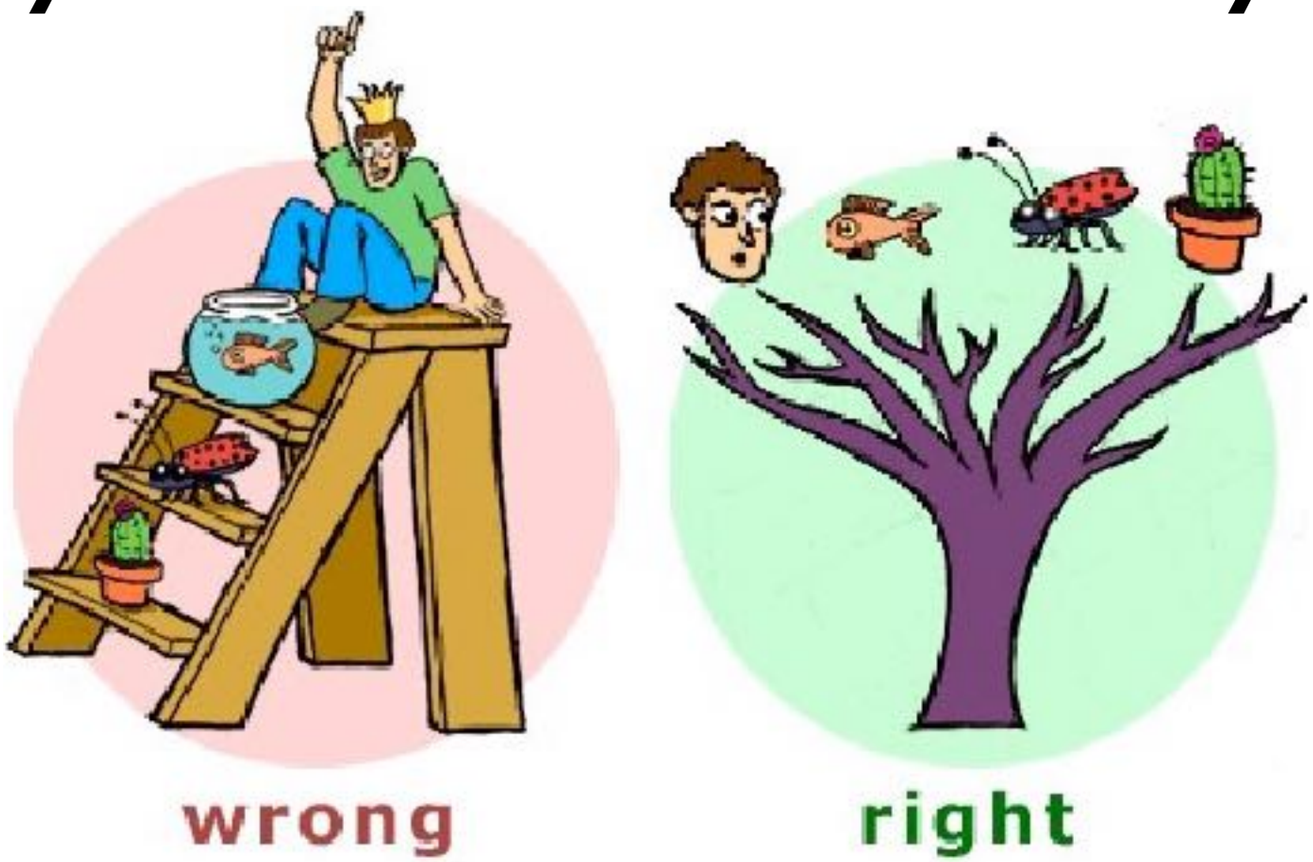




Does life really get better and better?

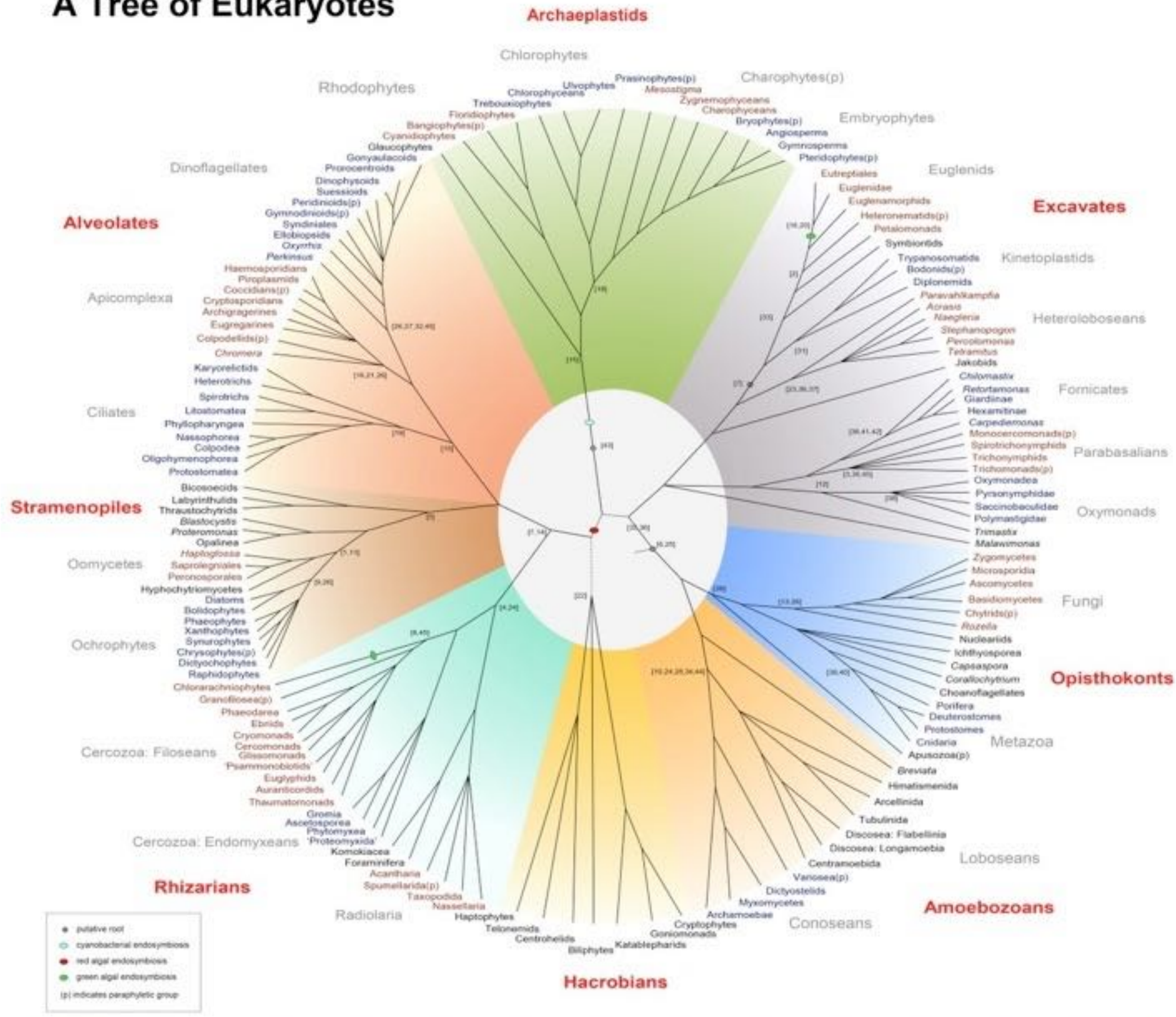


The myth of the evolutionary ladder



- 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



ARTICLE

doi:10.1038/nature14447

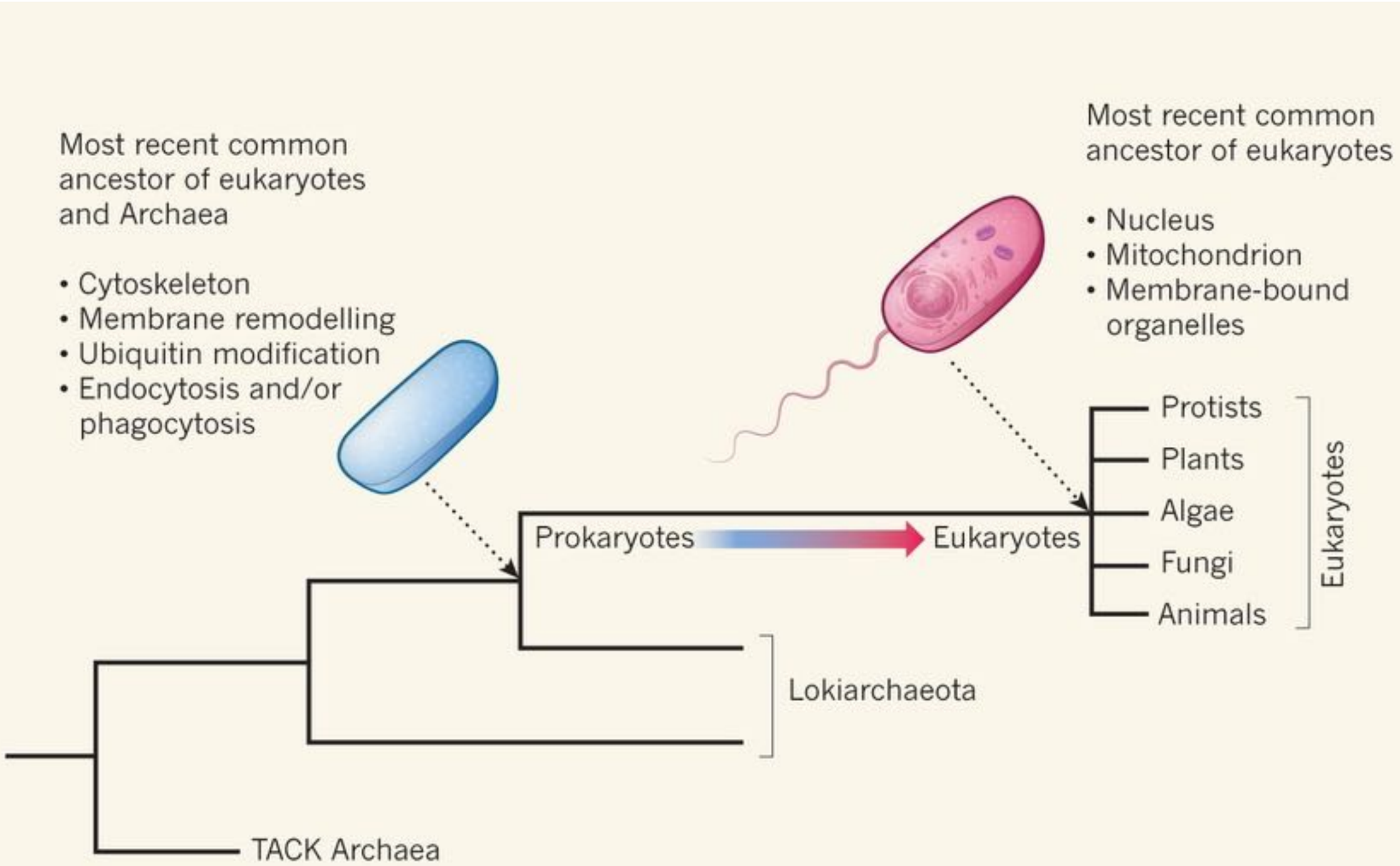
Complex archaea that bridge the gap between prokaryotes and eukaryotes

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

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.

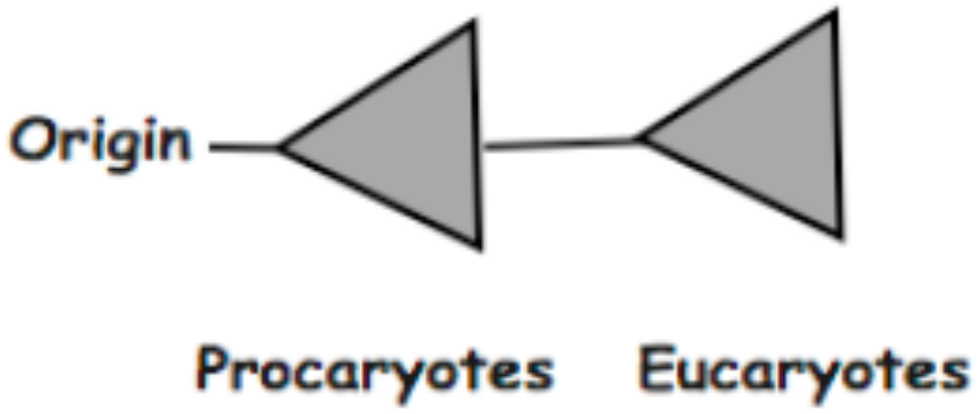
"eukaryotic host evolved from a bona fide archaeon…"

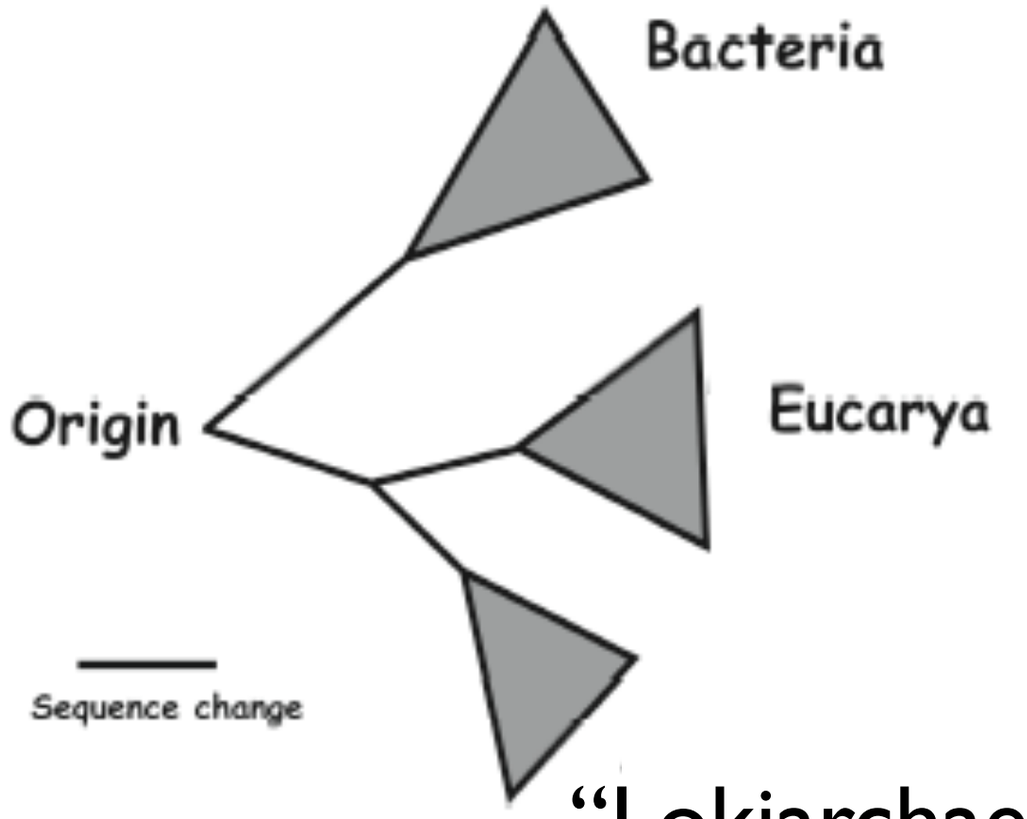
what's wrong with this picture?



iMessage with Charles Darwin

Today, 9:49 PM



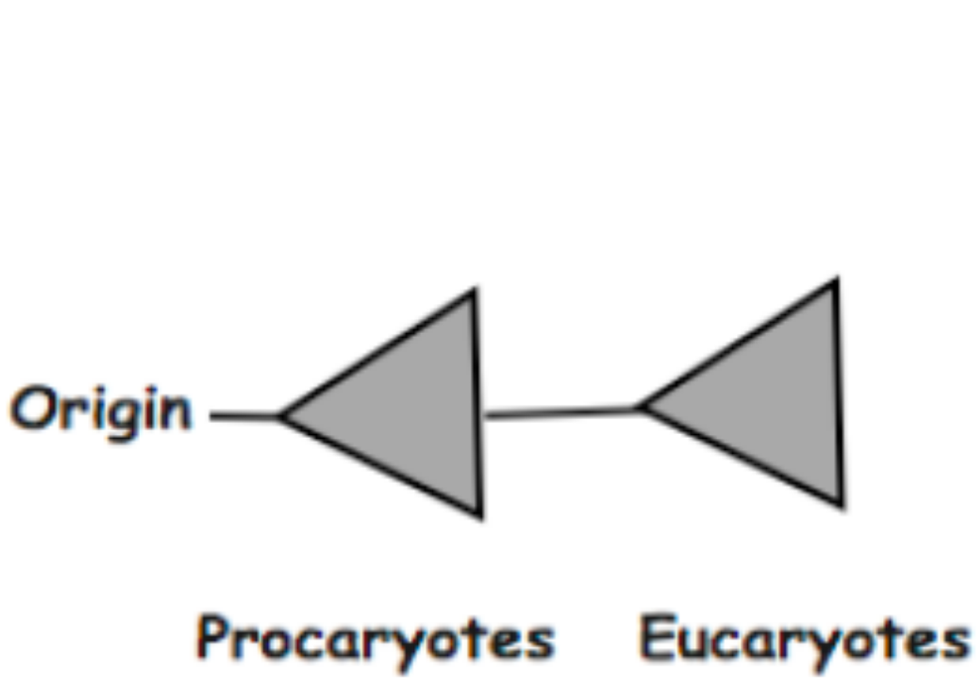


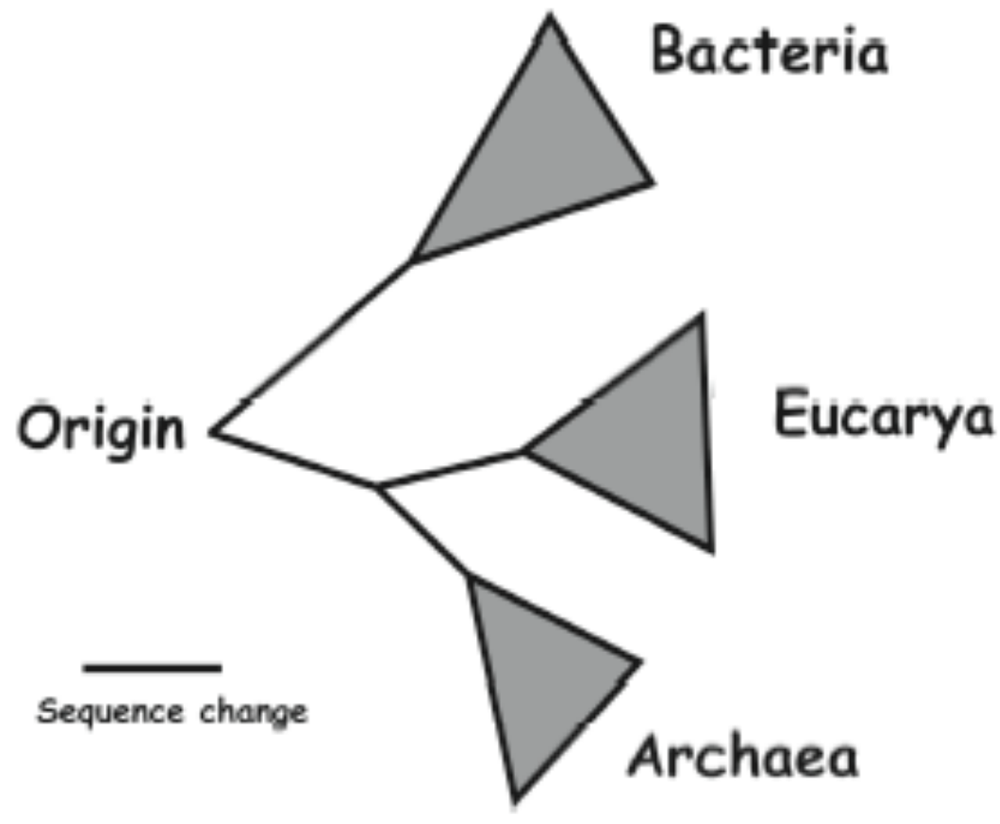
"Lokiarchaea"?

"Prokaryote" is still not a monophyletic group

"Prokaryote"= "pre-eukaryote"

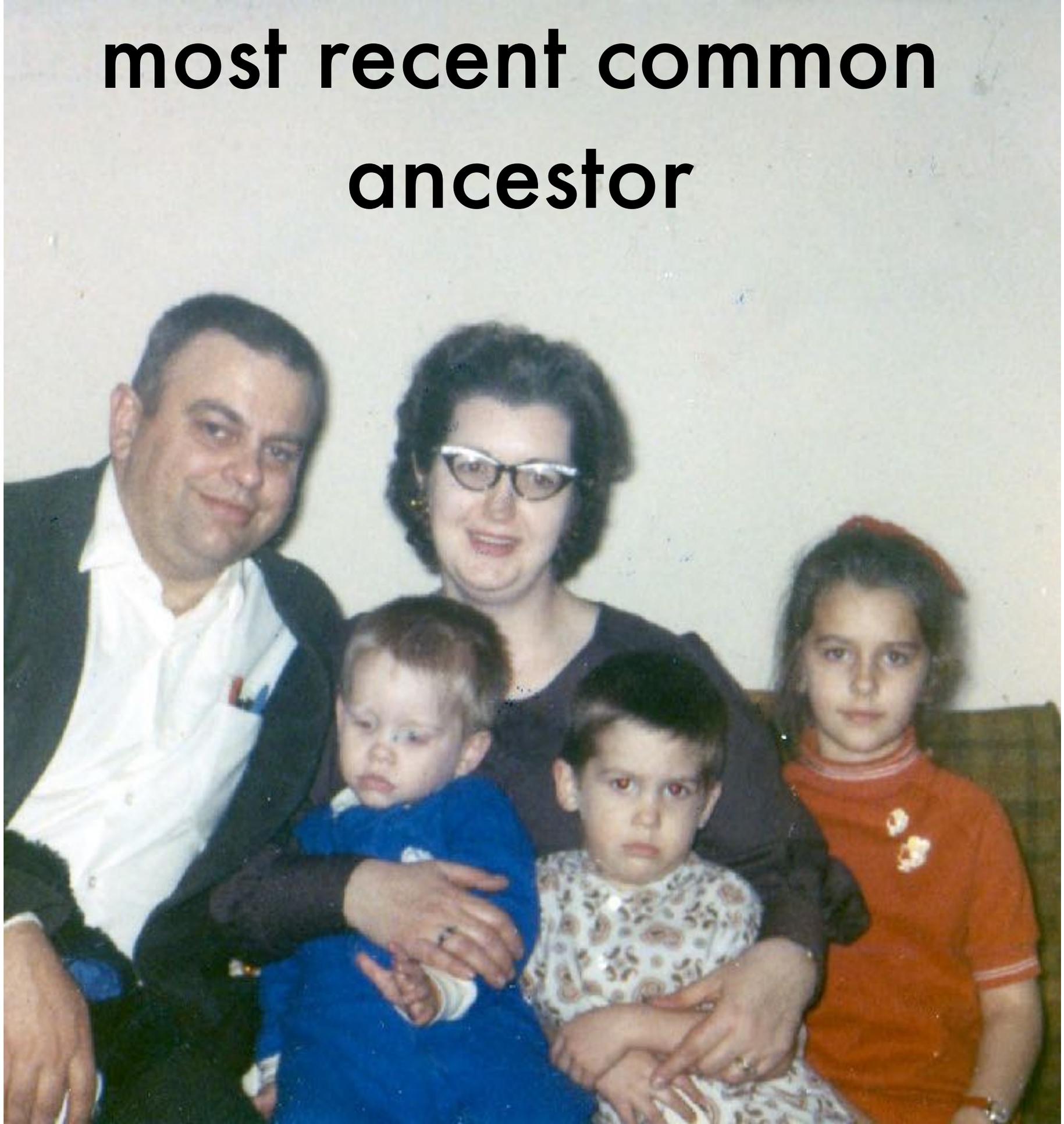
This is a hypothesis, not supported by evidence



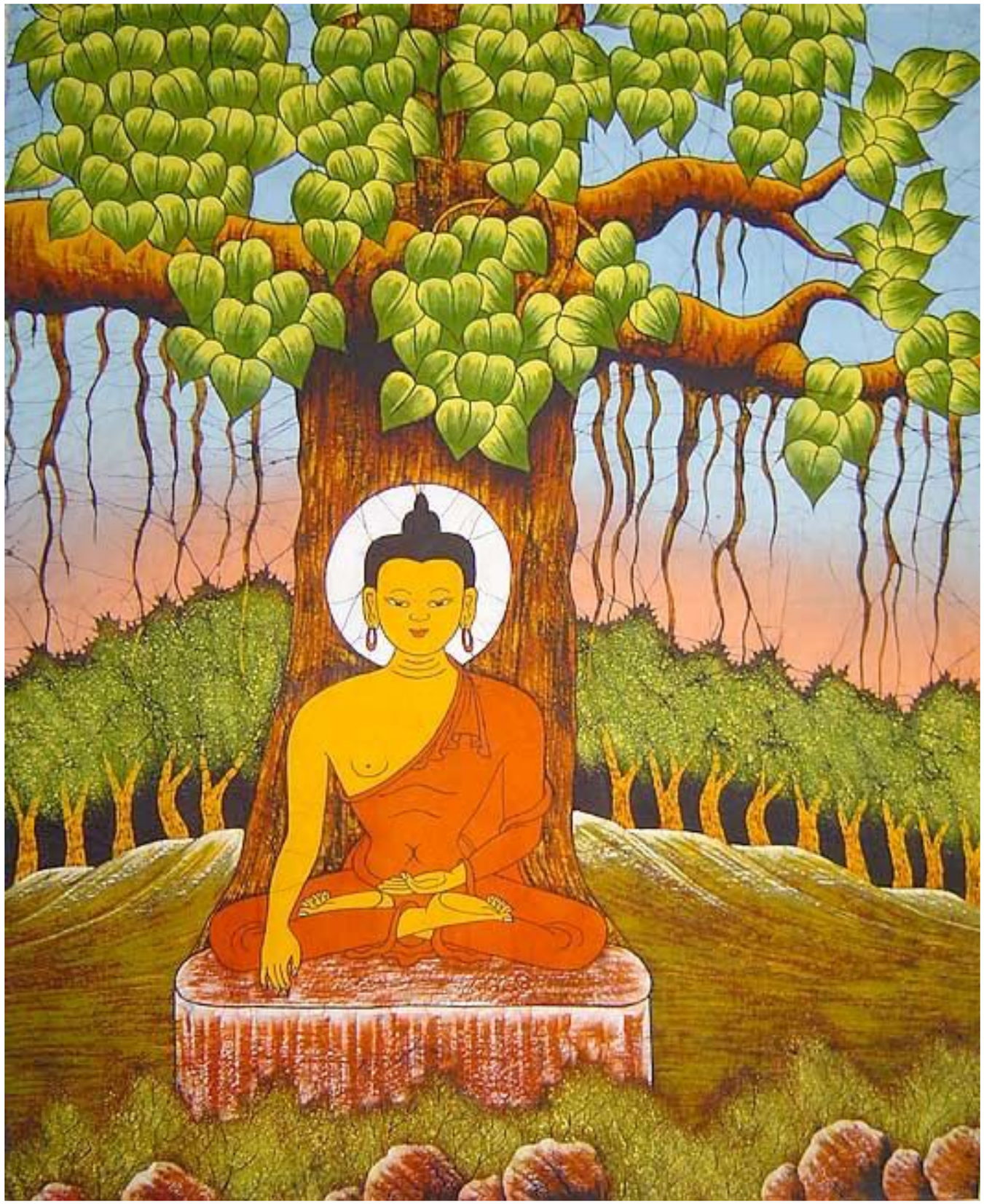


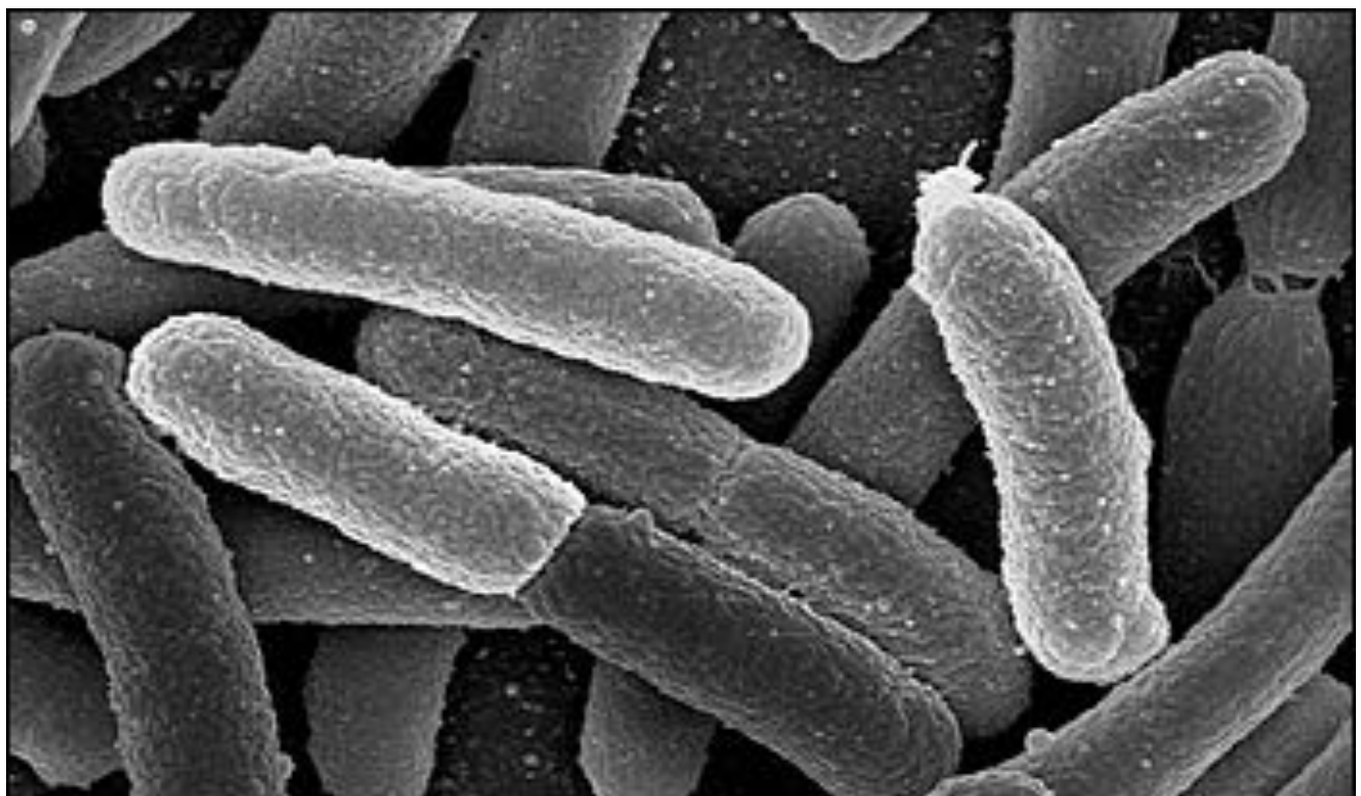
"Prokaryote" is not a clade

most recent common ancestor



What's common to all life?





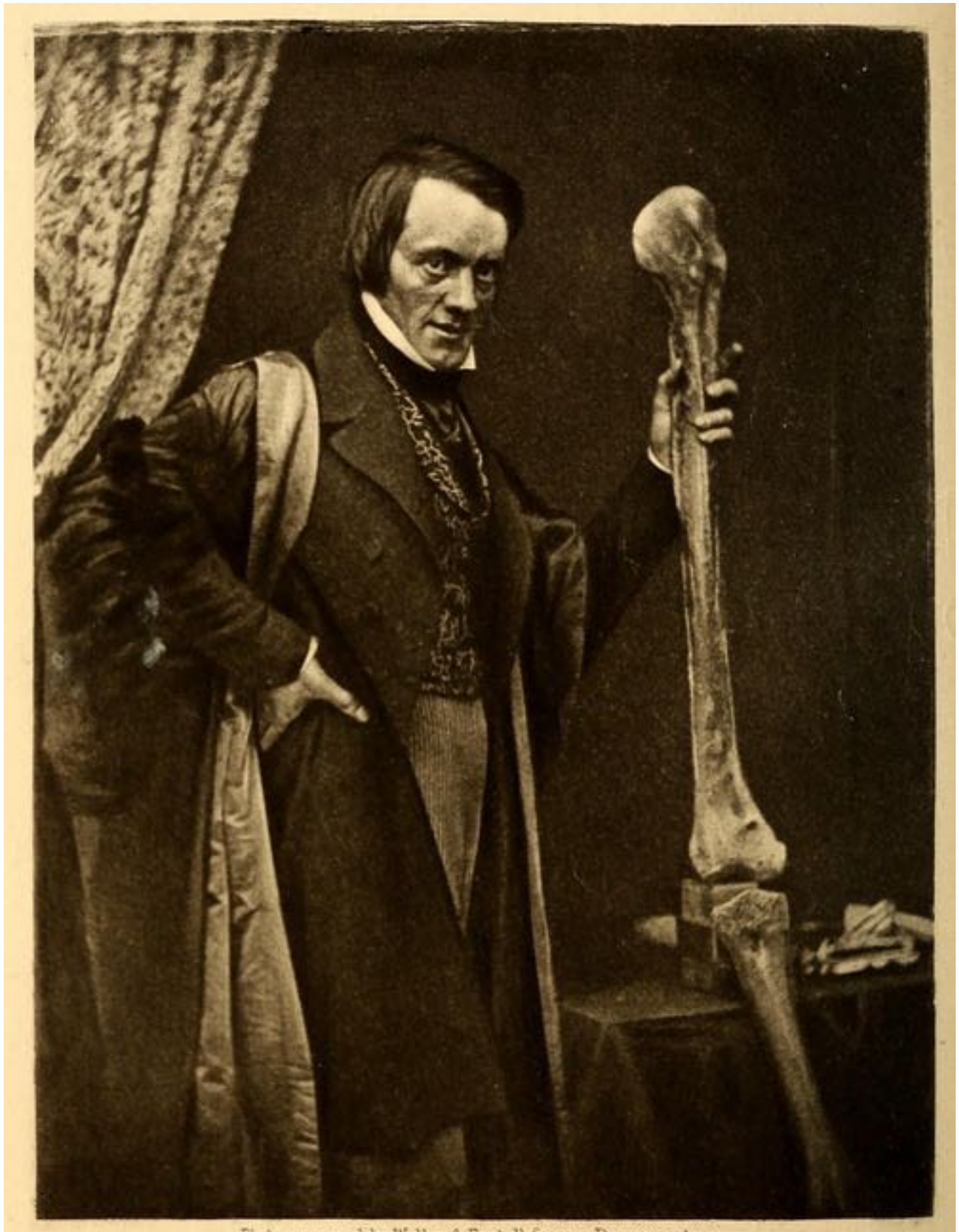


Photo-engraved by Walker & Boutall from a Daguerreotype.

Richard Owen

Homology

"The <u>same</u> organ in different animals under a variety of form and function"

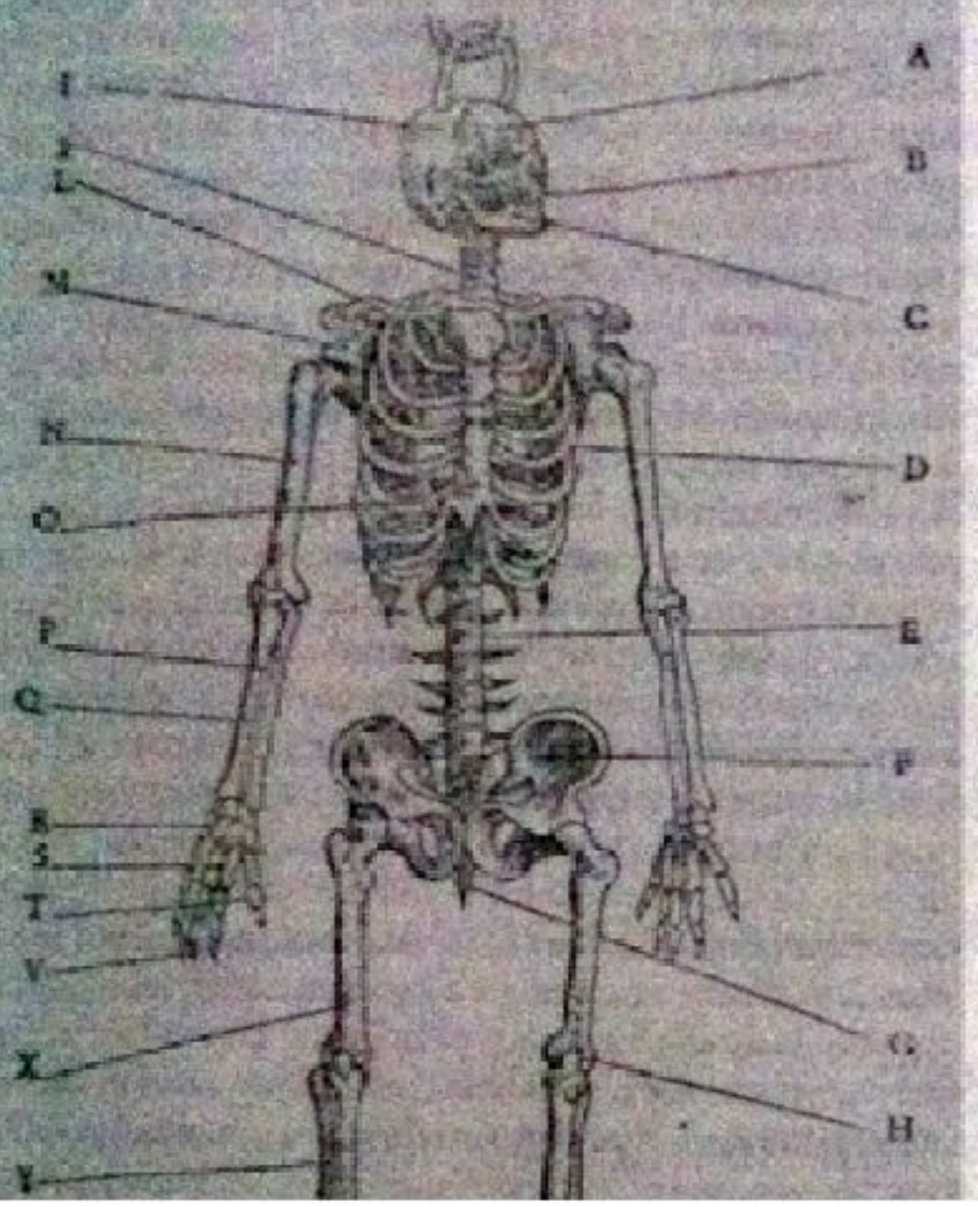
Sir Richard Owen (1843)

Homology

- Because all life shares a common ancestry - so should the "parts" of life

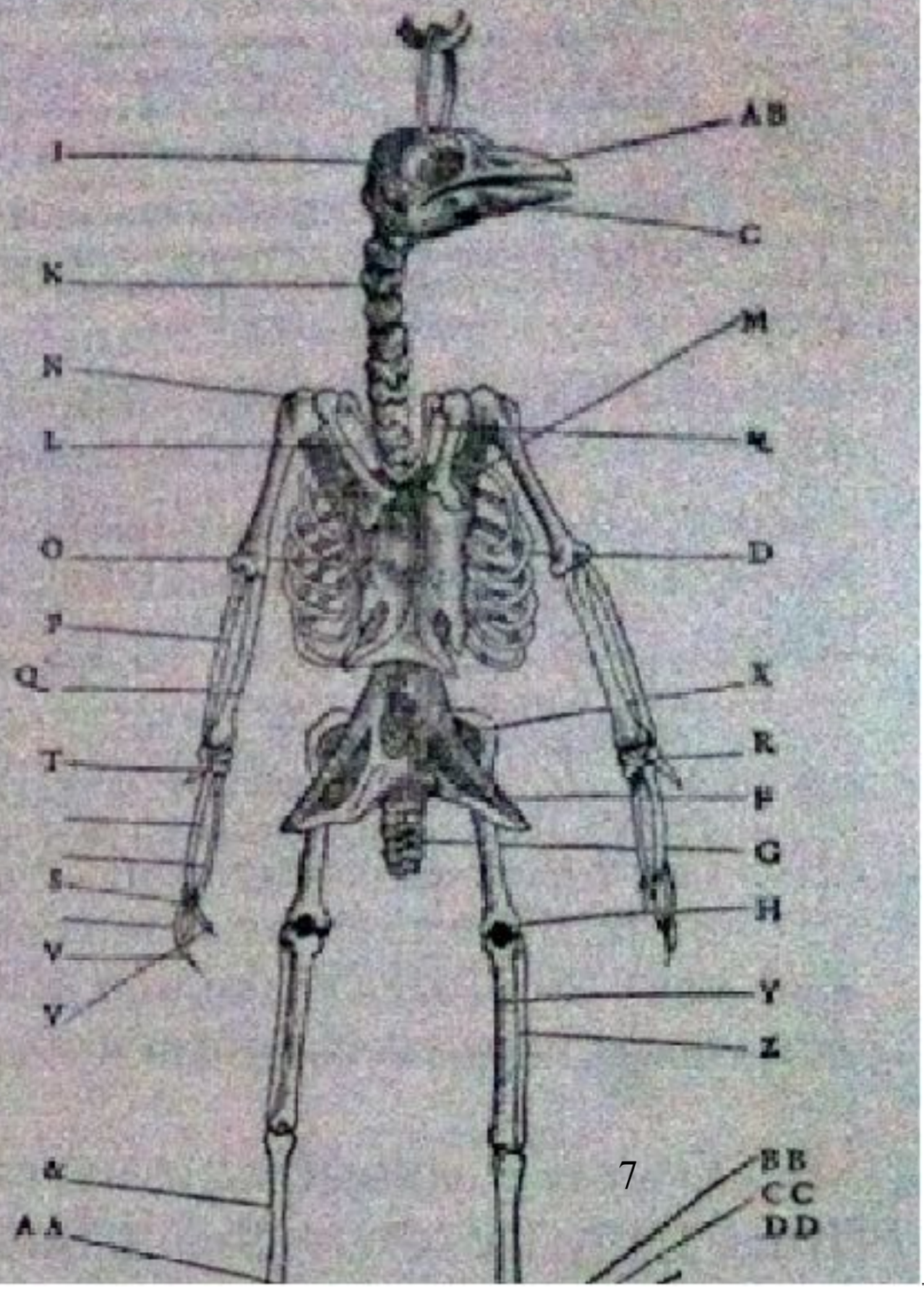
- Homology is similarity due to common descent
- Analogy is due to adaptation to the same environmental condition (convergence)
- "Levels" of homology

Portraict de l'arrest des os humains, mis en comparaison de l'anatomie de ceux des oyseaux, faisant que les lettres d'icelle se rapportent à celle cy, pour faire apparoistre combien l'affinité est grande des vns aux autres.



La comparaison du totale portraict des os humains monstre combien celluy cy qui est d'vn oyseau en est proche.

Portraict des os de l'oyseau.



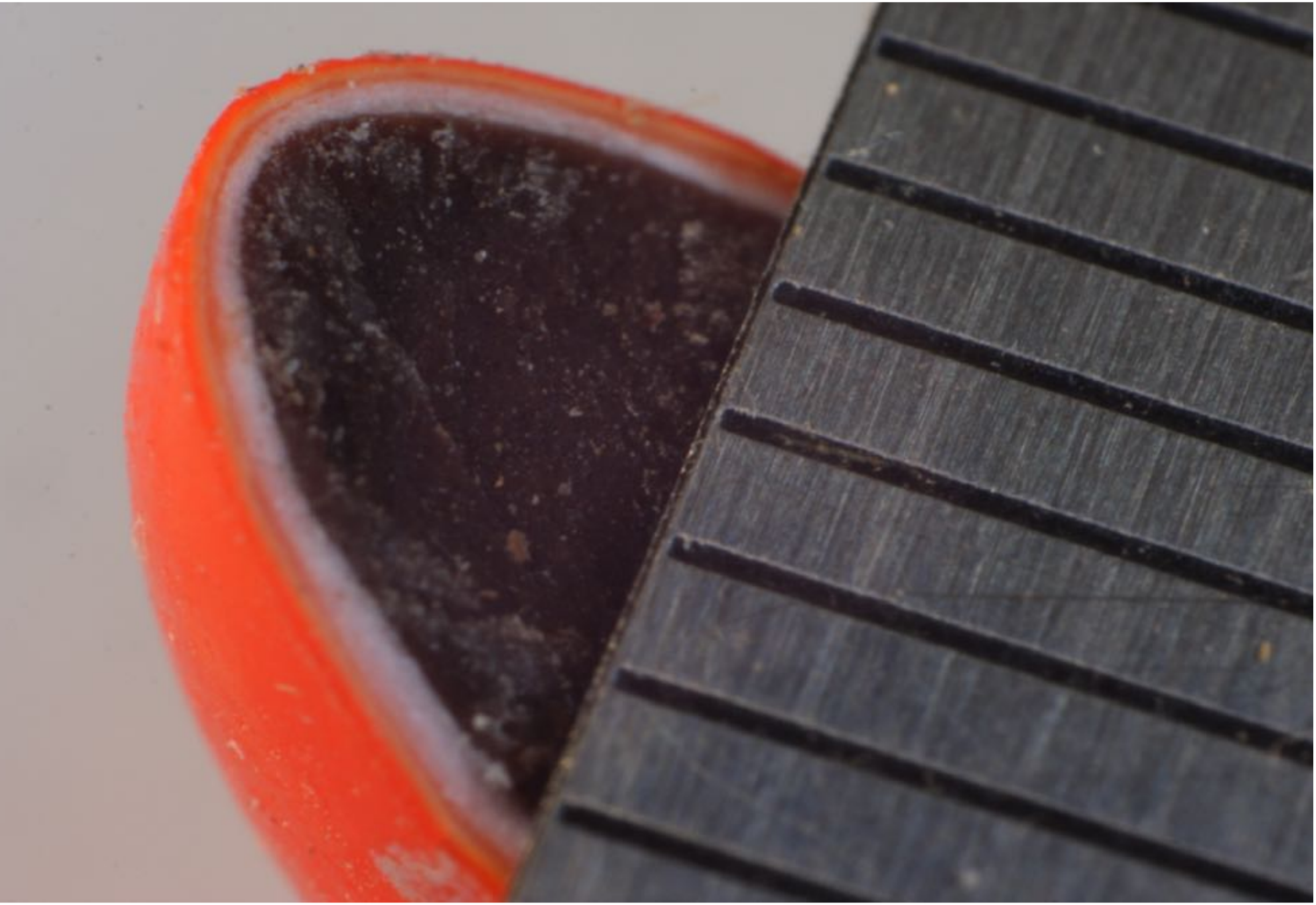
steps in assessing homology

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


PEANUT
Made with Real Milk Chocolate
m&m's
CHOCOLATE CANDIES
250


PEANUT BUTTER
Made with Real Peanut Butter
m&m's
CHOCOLATE CANDIES
240


Reese's
pieces
Peanut Butter



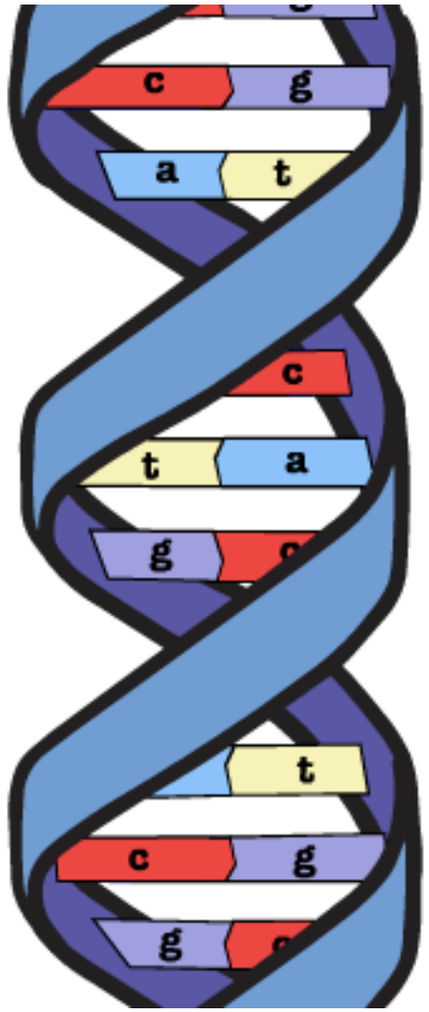
M & M diversity

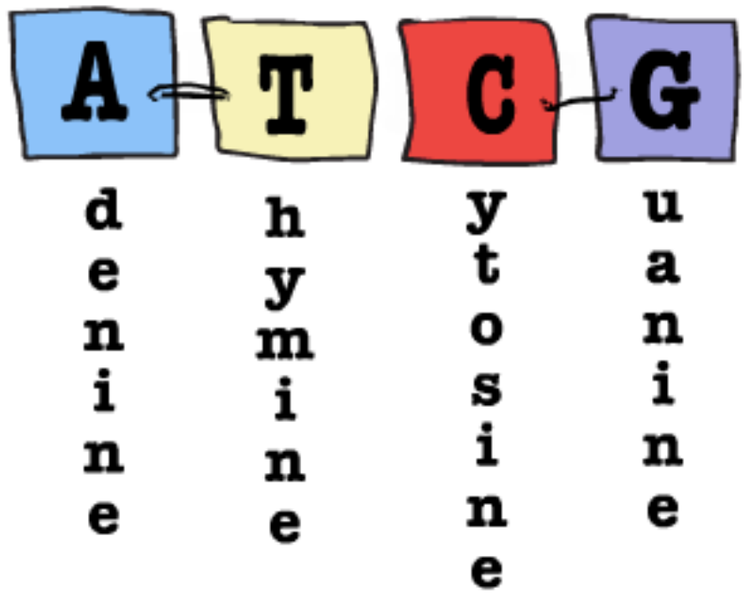
What are the criteria for defining homology at the molecular level?

Why do we care?

"Molecules as documents of evolutionary history"

Zuckerkandl and Pauling (1965)

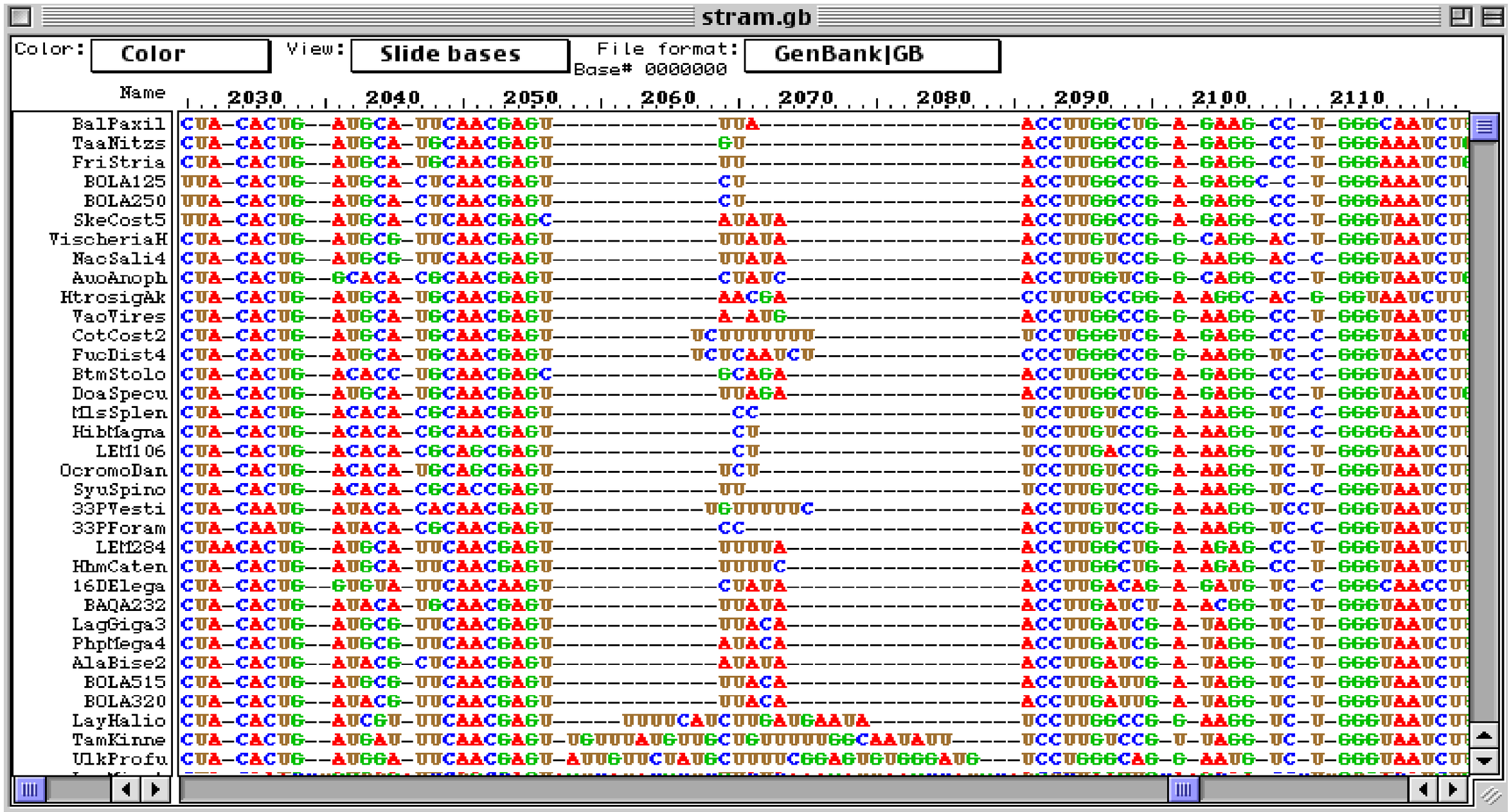




Homology

- 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?

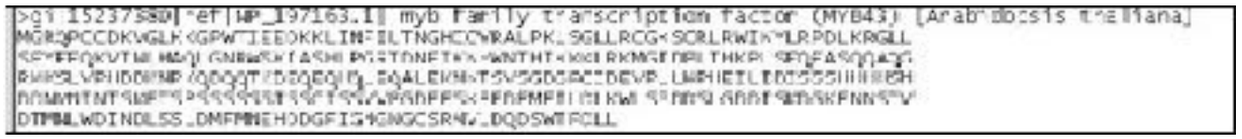
similarity vs. homology

- ***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

BLAST (Basic Local Alignment Search Tool)

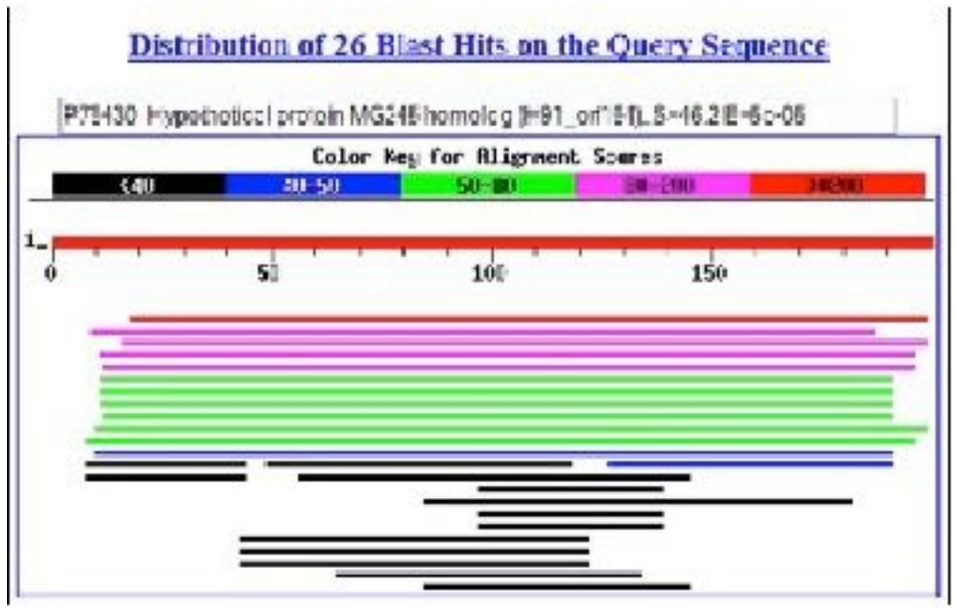
QUERY sequence(s)



BLAST program

BLAST database

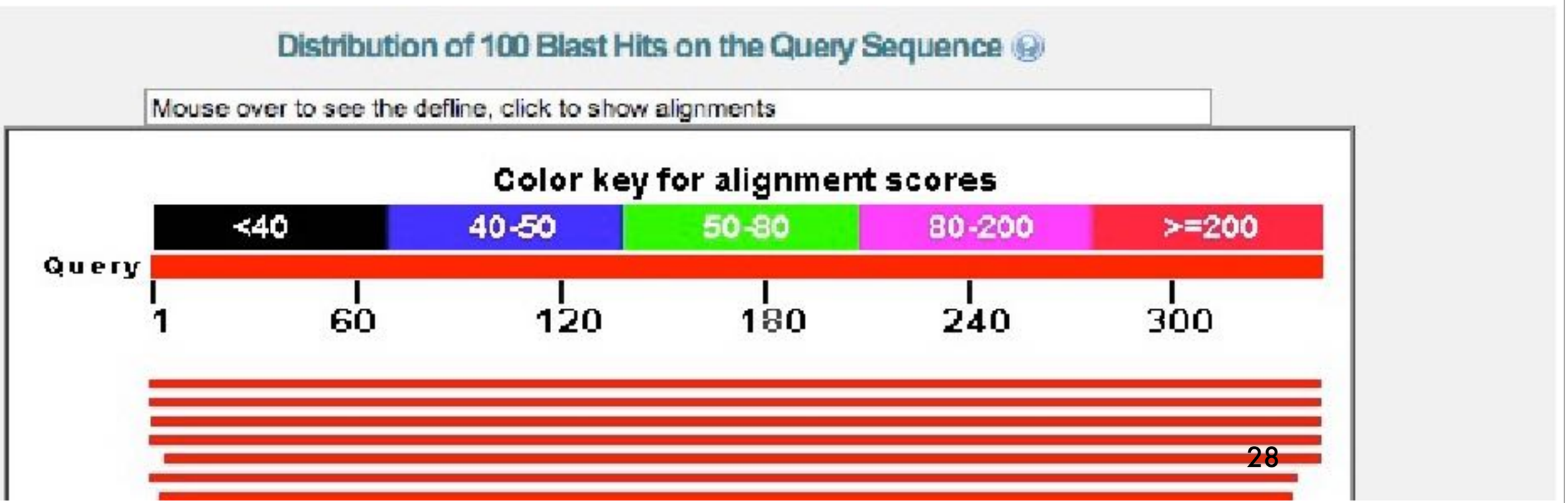
BLAST results



- Search for similarity to infer "homology"
- "mutual best hits" or reciprocal BLAST

What are the homologs?

Formatting options (for output)
Accession or gi: unique record of GB record
Max Score
Total Score
Query Coverage
E value: significance threshold, but really an EXPECT score of searching a given database (depends on database size)



Key concepts in BLAST

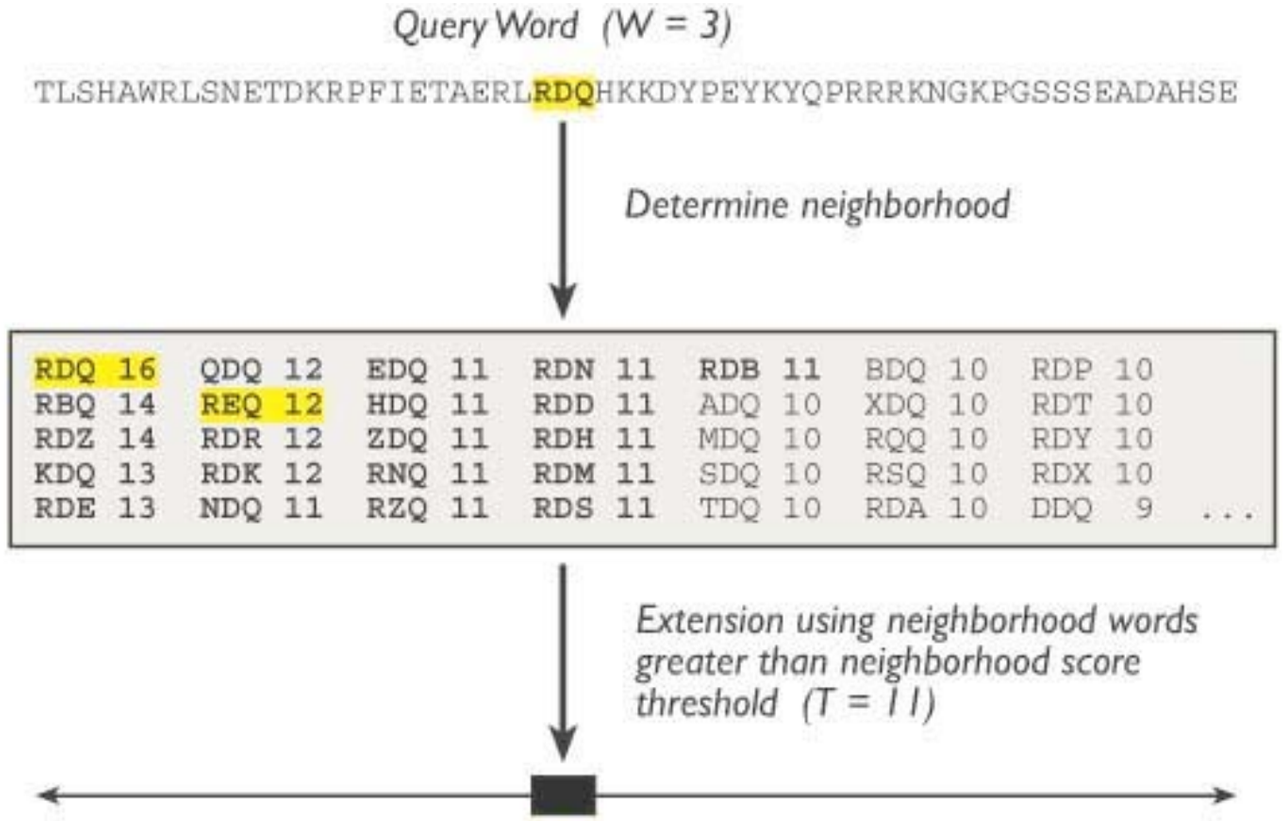
BLAST essentially computes "similarity", not alignment

- Given 2 proteins…

BLAST breaks search into "chunks" by finding all subsequences (stretches of similarity, or "words") of length k (e.g., k=4) that occur in both seqs

- build score on matches (scoring matrix, gap cost)
- extend from subsequences to see if you can increase score
- compute total score (when no more extensions are possible)

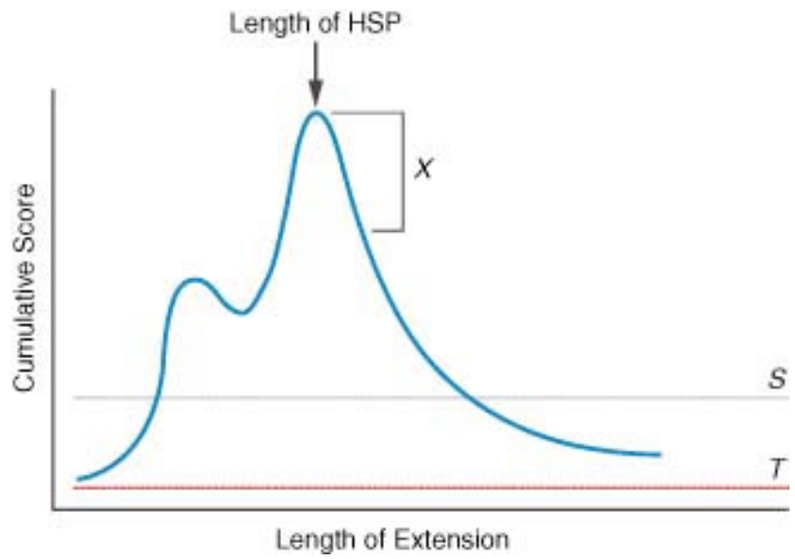
- Then computes a BLAST score against precomputed scores for all proteins in database
- Then ranks the score



HSP = high scoring segment pair

<u>Score (S)</u> = alignment quality (gaps + scoring matrix)

<u>E value (E)</u> = number of different alignments with scores => S that are expected to occur by chance



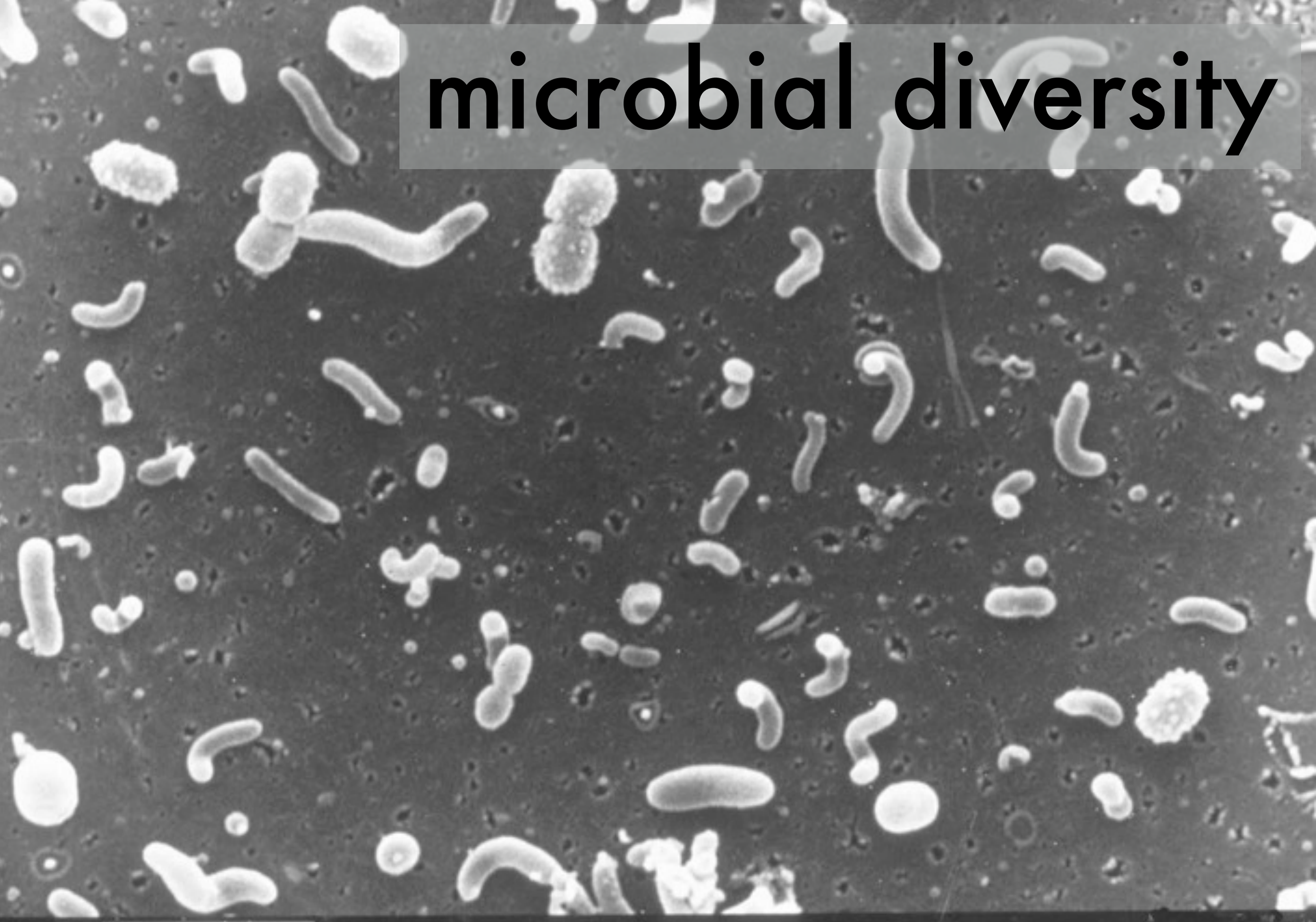
It's all about that (data)'base.

-Scott Dawson (2016)

BLAST analyses

- how conserved is your rpoB homolog as compared to the next homolog from another species?
- how many rpoB homologs are there ?
- how evolutionary diverse are the homologs (e.g, archaea, bacteria, eukaryotes?)

microbial diversity

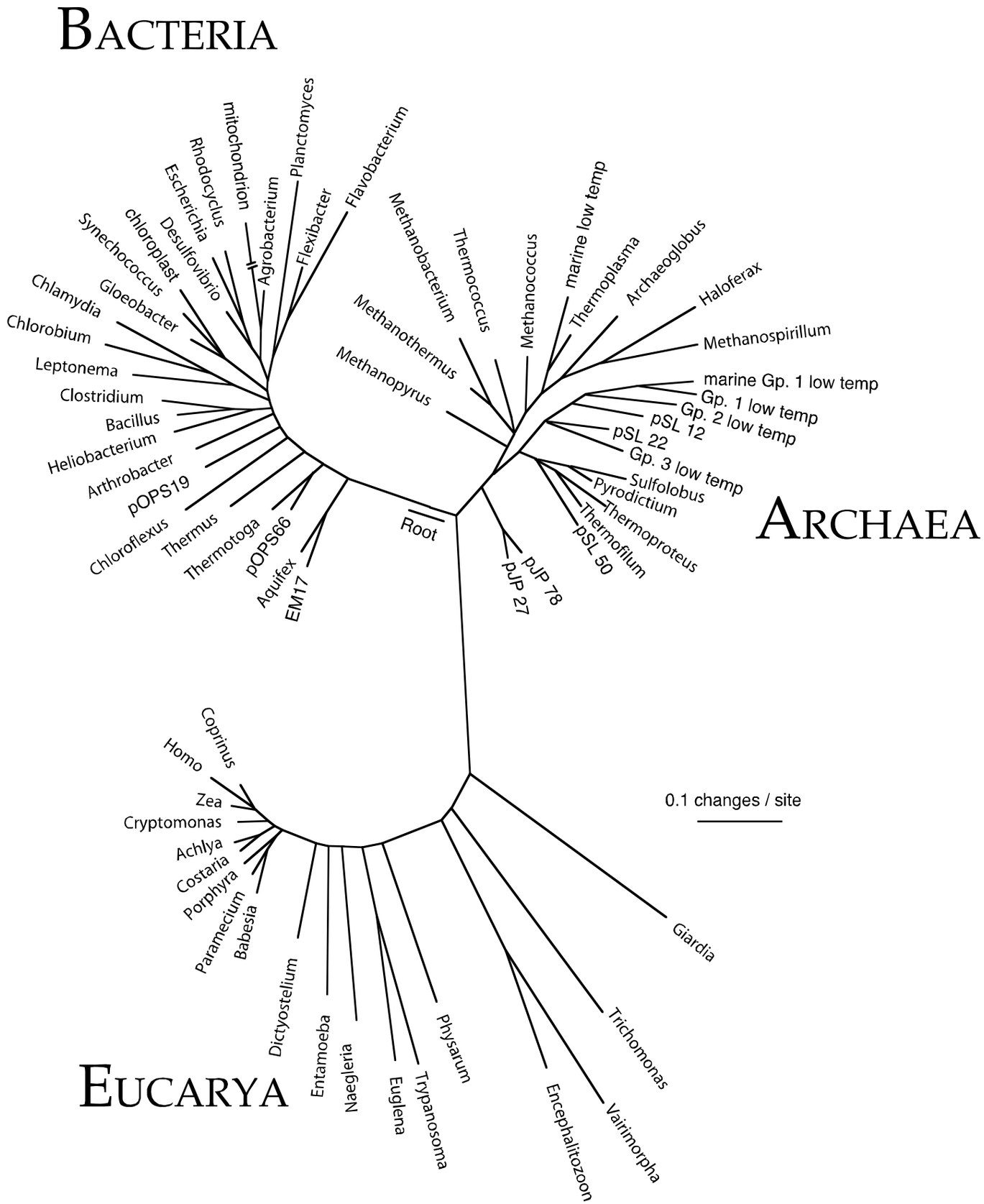


*

Carl Woese used rRNA genes to determine "natural" or evolutionary relationships among microbes

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

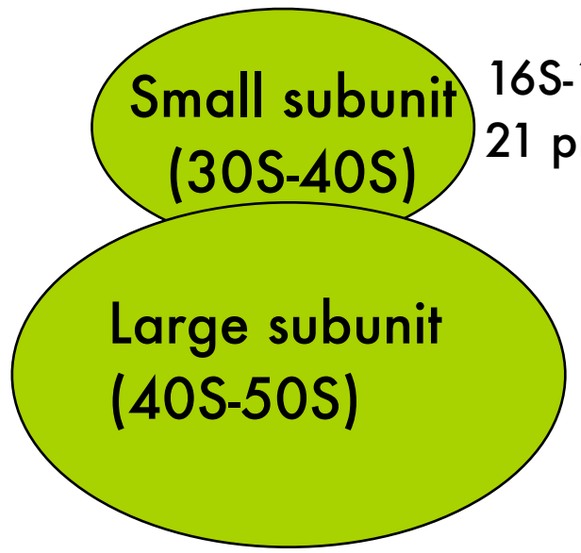
Gene Trees as Proxies for Organism Trees



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)

The Ribosome
(70S-80S)



16S-18S; 1500-1900 nts
21 proteins

23S-28S; 2,900-4700 nts
5S; 120 nts
5.8S euks; 160 nts
34 proteins

Lessons from the "Big Tree"

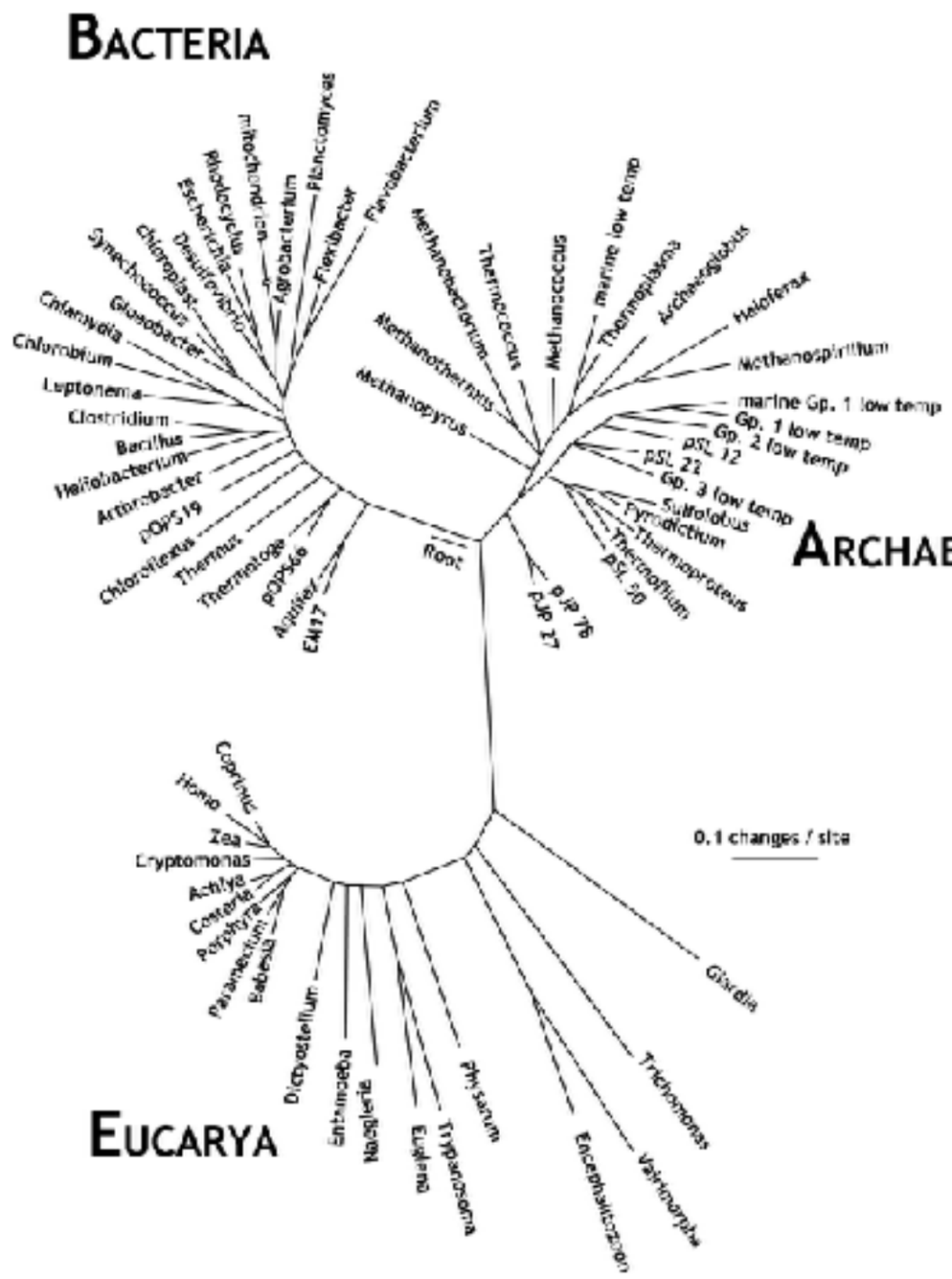
- All cellular life is related (one origin)

- Three domains of life, not two: Bacteria, Archaea, Eukarya

- Most life is microbial (still)

- <u>Natural taxonomy:</u> 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)



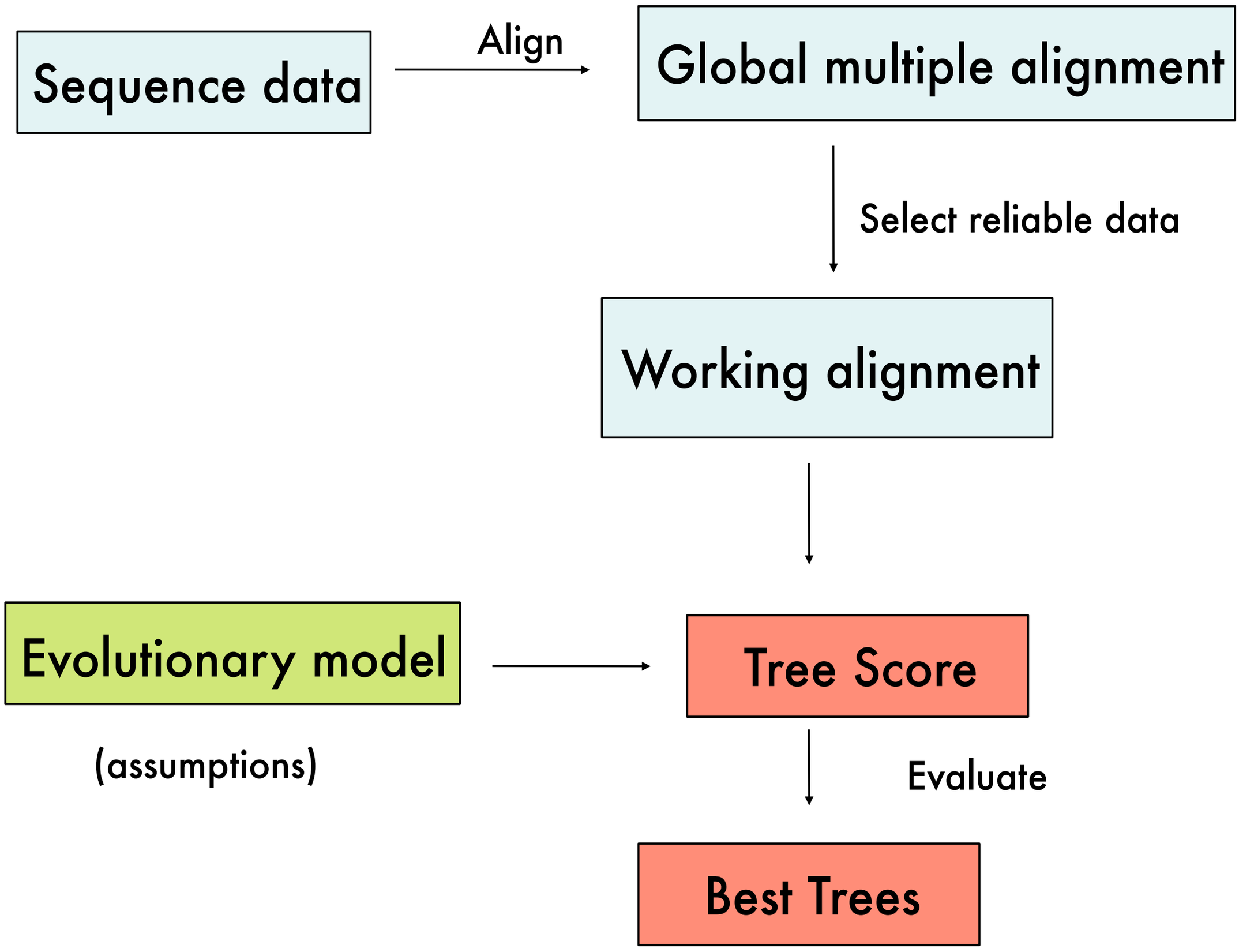
Zen and the Art of Molecular Phylogeny

Part 2: Building Trees

The road to Phylogenetic Nirvana by overcoming the Five Hindrances:

Desire
Anger
Sloth
Worry
Doubt

*


Sequence data
Align
Global multiple alignment
Select reliable data
Working alignment
Evolutionary model
(assumptions)
Tree Score
Evaluate
Best Trees

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----------UUA-------------------ACCUUGGCUG-A-GAAG-CC-U-GGGCAAUCUU
TaaNitzs CUA-CACUG--AUGCA-UGCAACGAGU----------GU--------------------ACCUUGGCCG-A-GAGG-CC-U-GGGAAAUCUG
FriStria CUA-CACUG--AUGCA-UGCAACGAGU----------UU--------------------ACCUUGGCCG-A-GAGG-CC-U-GGGAAAUCUG
BOLA125  UUA-CACUG--AUGCA-CUCAACGAGU----------CU--------------------ACCUUGGCCG-A-GAGGC-C-U-GGGAAAUCUU
BOLA250  UUA-CACUG--AUGCA-CUCAACGAGU----------CU--------------------ACCUUGGCCG-A-GAGG-CC-U-GGGAAAUCUU
SkeCost5 UUA-CACUG--AUGCA-CUCAACGAGC----------AUAUA-----------------ACCUUGGCCG-A-GAGG-CC-U-GGGUAAUCUU
VischeriaH CUA-CACUG--AUGCG-UUCAACGAGU----------UUAUA-----------------ACCUUGUCCG-G-CAGG-AC-U-GGGUAAUCUU
NacSali4 CUA-CACUG--AUGCG-UUCAACGAGU----------UUAUA-----------------ACCUUGUCCG-G-AAGG-AC-C-GGGUAAUCUU
AuoAnoph CUA-CACUG--GCACA-CGCAACGAGU----------CUAUC-----------------ACCUUGGUCG-G-CAGG-CC-U-GGGUAAUCUG
HtrosigAk CUA-CACUG--AUGCA-UGCAACGAGU----------AACGA-----------------CCUUUGCCGG-A-AGGC-AC-G-GGUAAUCUUU
YaoYires CUA-CACUG--AUGCA-UGCAACGAGU----------A-AUG-----------------ACCUUGGCCG-G-AAGG-CC-U-GGGUAAUCUU
CotCost2 CUA-CACUG--AUGCA-UGCAACGAGU----------UCUUUUUUU-------------UCCUGGGUCG-A-GAGG-CC-C-GGGUAAUCUG
FucDist4 CUA-CACUG--AUGCA-UGCAACGAGU----------UCUCAAUCU-------------CCCUGGGCCG-G-AAGG-UC-C-GGGUAACCUU
BtmStolo CUA-CACUG--ACACC-UGCAACGAGC----------GCAGA-----------------ACCUUGGCCG-A-GAGG-CC-C-GGGUAAUCUU
DoaSpecu CUA-CACUG--AUGCA-UGCAACGAGU----------UUAGA-----------------ACCUUGGCUG-A-GAGG-CC-U-GGGUAAUCUG
MlsSplen CUA-CACUG--ACACA-CGCAACGAGU----------CC--------------------UCCUUGUCCG-A-AAGG-UC-C-GGGUAAUCUU
HibMagna CUA-CACUG--ACACA-CGCAACGAGU----------CU--------------------UCCUUGUCCG-A-AAGG-UC-C-GGGGAAUCUU
LEM106   CUA-CACUG--ACACA-CGCAGCGAGU----------CU--------------------UCCUUGACCG-A-AAGG-UC-U-GGGUAAUCUU
OcromoDan CUA-CACUG--ACACA-UGCAGCGAGU----------UCU-------------------UCCUUGUCCG-A-AAGG-UC-U-GGGUAAUCUU
SyuSpino CUA-CACUG--ACACA-CGCACCGAGU----------UU--------------------UCCUUGUCCG-A-AAGG-UC-C-GGGUAAUCUU
33PYesti CUA-CAAUG--AUACA-CACAACGAGU----------UGUUUUUC--------------ACCUUGUCCG-A-AAGG-UCCU-GGGUAAUCUU
33PForam CUA-CAAUG--AUACA-CGCAACGAGU----------CC--------------------ACCUUGUCCG-A-AAGG-UC-C-GGGUAAUCUU
LEM284   CUAACACUG--AUGCA-UUCAACGAGU----------UUUUA-----------------ACCUUGGCUG-A-AGAG-CC-U-GGGUAAUCUU
HhmCaten CUA-CACUG--AUGCA-UUCAACGAGU----------UUUUC-----------------ACCUUGGCUG-A-AGAG-CC-U-GGGUAAUCUU
16DElega CUA-CACUG--GUGUA-UUCAACAAGU----------CUAUA-----------------ACCUUGACAG-A-GAUG-UC-C-GGGCAACCUU
BAQA232  CUA-CACUG--AUACA-UGCAACGAGU----------UUAUA-----------------ACCUUGAUCU-A-ACGG-UC-U-GGGUAAUCUU
LagGiga3 CUA-CACUG--AUGCG-UUCAACGAGU----------UUACA-----------------ACCUUGAUCG-A-UAGG-UC-U-GGGUAAUCUU
PhpMega4 CUA-CACUG--AUGCG-UUCAACGAGU----------AUACA-----------------ACCUUGAUCG-A-UAGG-UC-U-GGGUAAUCUU
AlaBise2 CUA-CACUG--AUACG-CUCAACGAGU----------AUAUA-----------------ACCUUGAUCG-A-UAGG-UC-U-GGGUAAUCUU
BOLA515  CUA-CACUG--AUGCG-UUCAACGAGU----------UUACA-----------------ACCUUGAUUG-A-UAGG-UC-U-GGGUAAUCUU
BOLA320  CUA-CACUG--AUACG-UUCAACGAGU----------UUACA-----------------ACCUUGAUUG-A-UAGG-UC-U-GGGUAAUCUU
LayHalio CUA-CACUG--AUCGU-UUCAACGAGU-----UUUUCAUCUUGAUGAAUA----------UCCUUGGCCG-G-AAGG-UC-U-GGGUAAUCUU
TamKinne CUA-CACUG--AUGAU-UUCAACGAGU-UGUUUAUGUUGCUGUUUUUGGCAAUAUU-----UCCUUGUCCG-U-UAGG-UC-U-GGGUAAUCUU
UlkProfu CUA-CACUG--AUGGA-UUCAACGAGU-AUUGUUCUAUGCUUUUCGGAGUGUGGGAUG---UCCUGGGCAG-G-AAUG-UC-U-GGGUAAUCUU
```

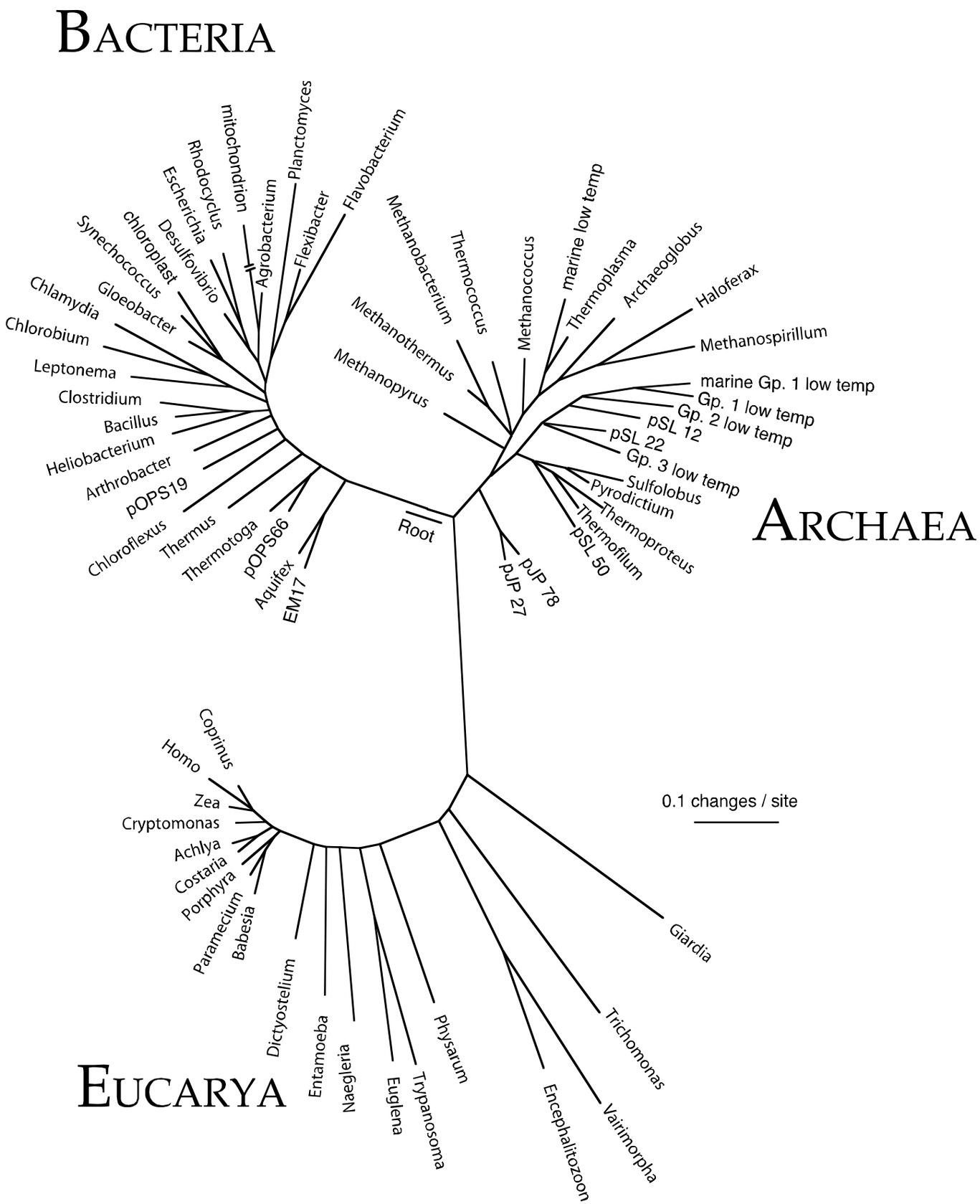
Figure 17.20 Microbiology: An Evolving Science

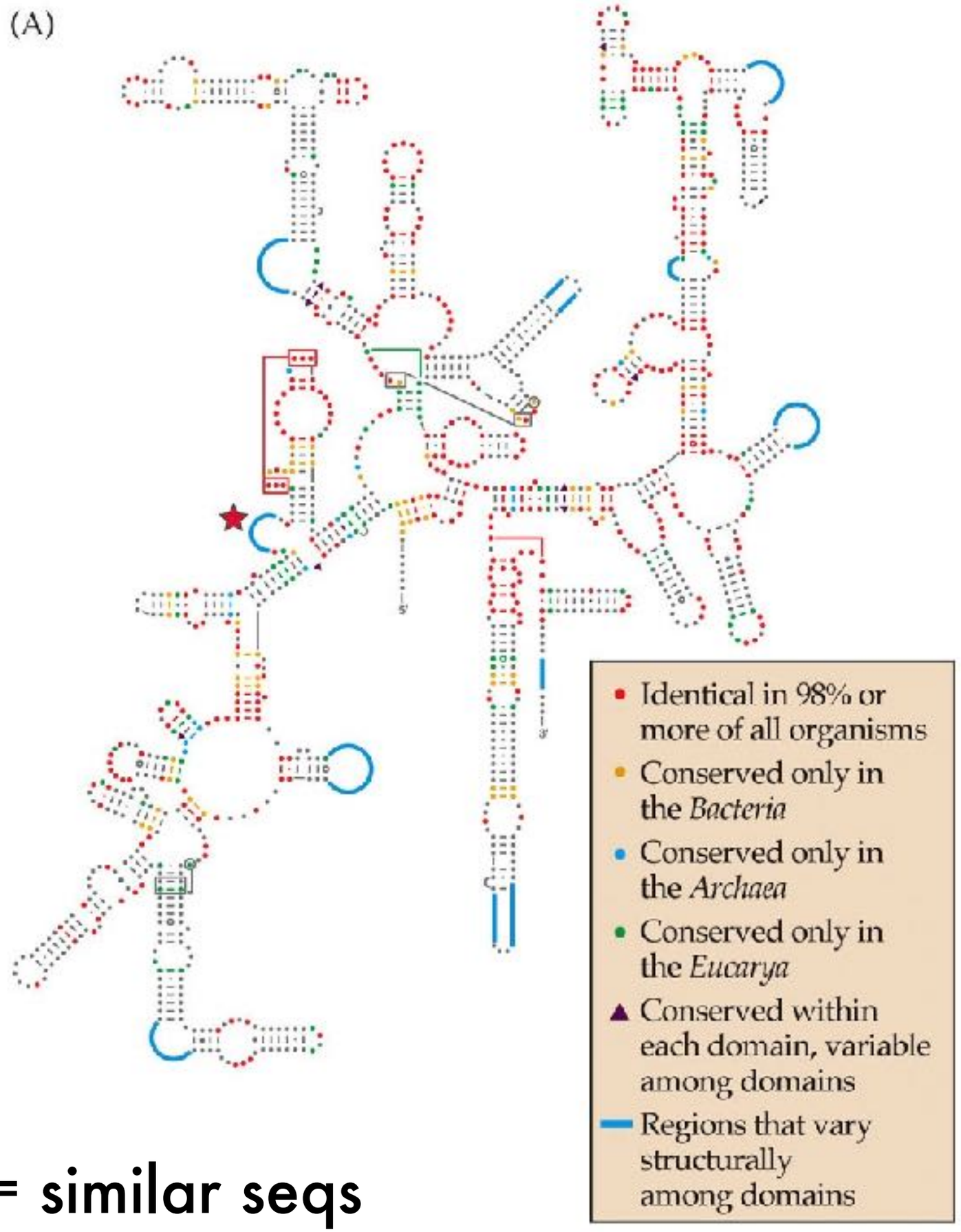
% identity?

<u>Multiple sequence alignment:</u>

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

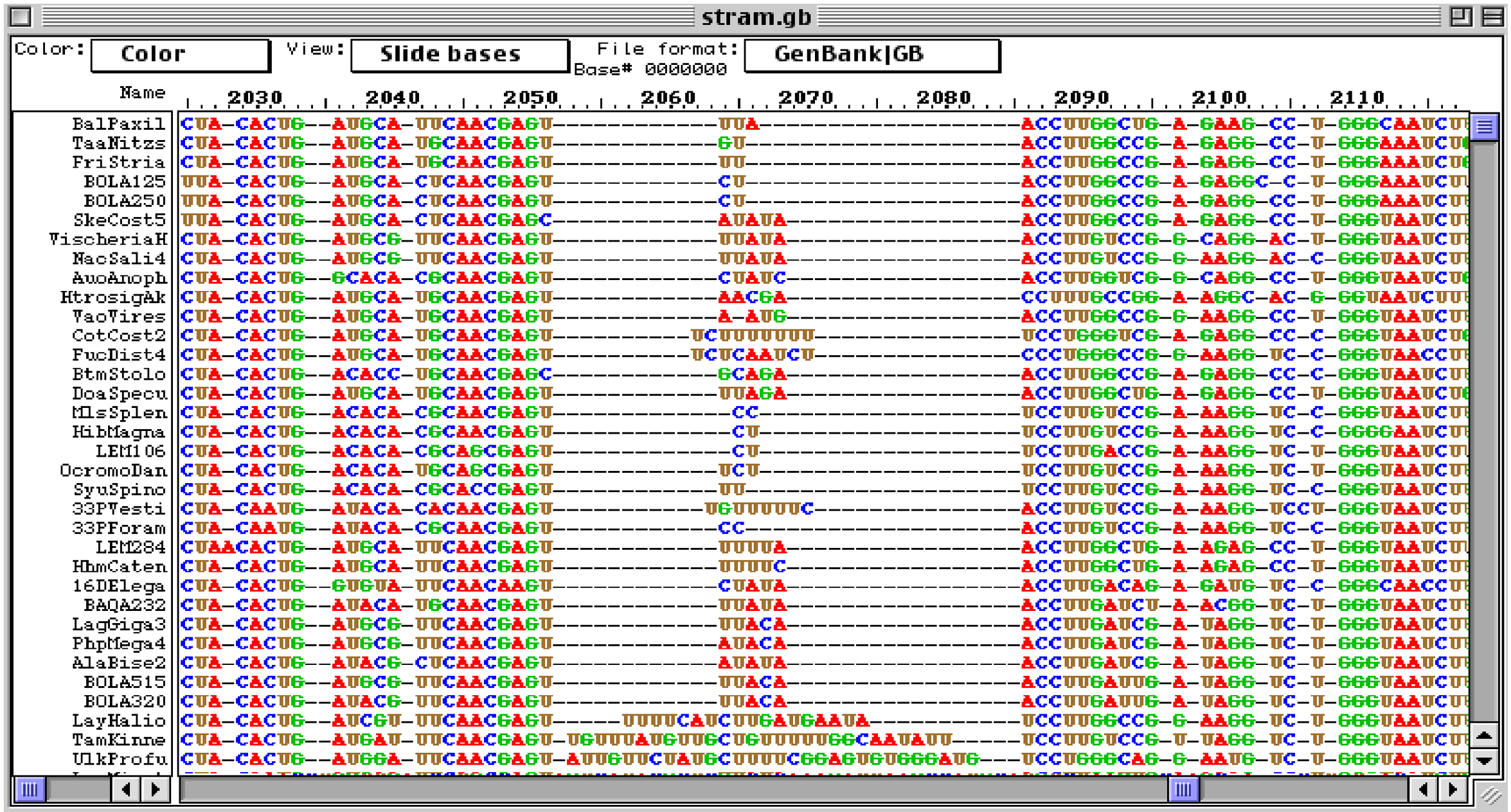
Gene Trees as Proxies for Organism Trees





similar “phylotypes”= similar seqs

Structure-based sequence alignment



Are different sized stems and loops homologous?

Sequence alignment = hypothesis

- Common origin (homology)
- Descended from fragment of an ancestral genome
- Compared residues trace back to ancestral sequence

* Positional homology (columns)

–> Omit non-homologous regions (uncertain ancestry)

Aligning molecular sequences

- why?
  - functional information
  - structural information
  - molecular evolution
  - genome assembly
- if similar enough may have similar function (homologous??)
- more diversity, more information?

What would be some problems in comparing all the genes in a genome to those in another genome?

What about aligning a genome to another genome? or many genomes?

Global vs. local alignment

- global: aligned start to finish (whole sequence, include as many matching pairs)

- local: region of high similarity; limited to highly similar conserved regions; length differences

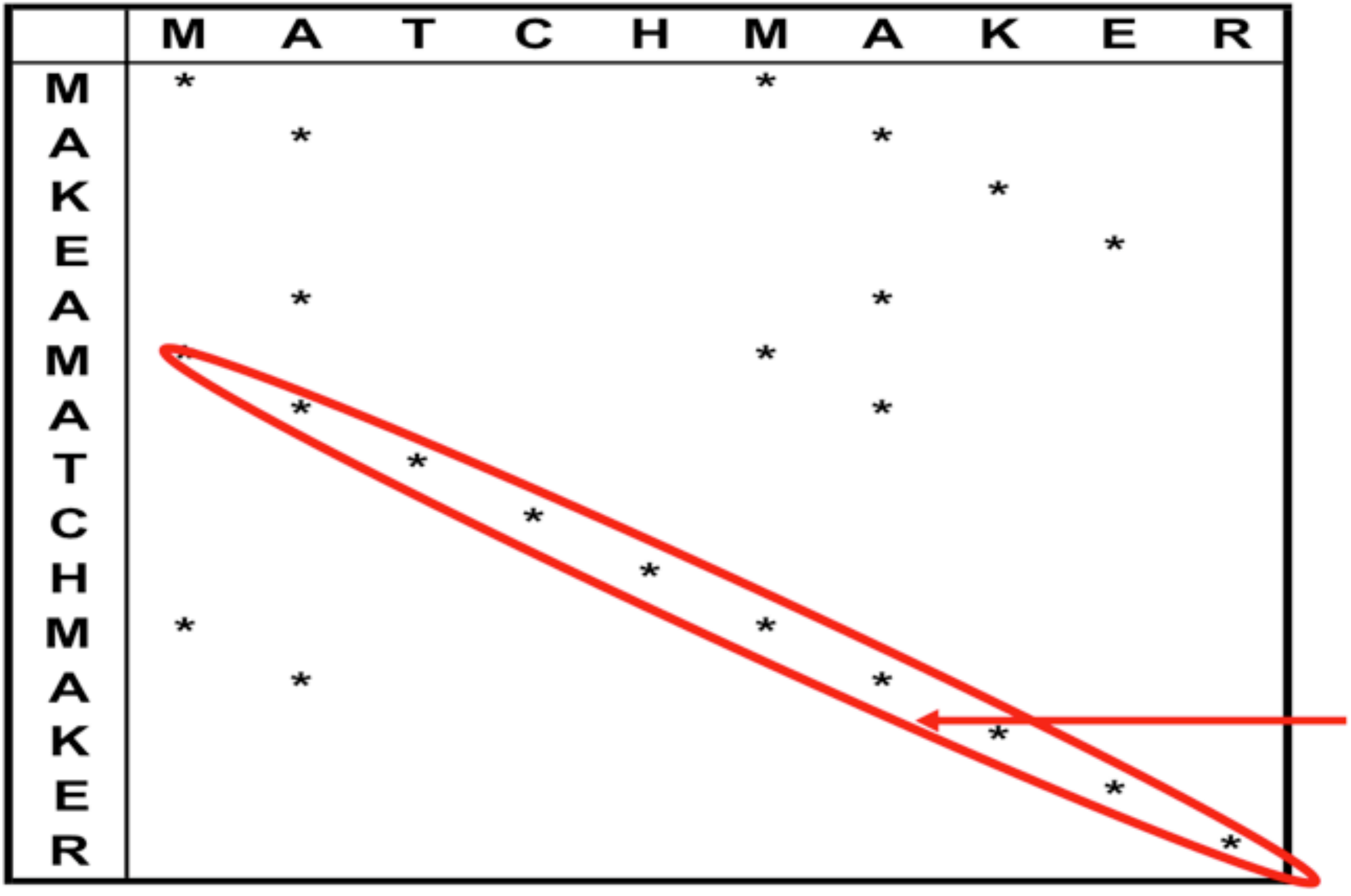
```
 LGPSSKQTGKGS-SRIWDN
|     |   |||    |   |
 LN-ITKSAGKGAIMRLGDA
```

GLOBAL

```
 --------TGKG---------
         |||
 --------AGKG---------
```

LOCAL

Dot Matrix Analysis



Dot plots (or dot matrices)

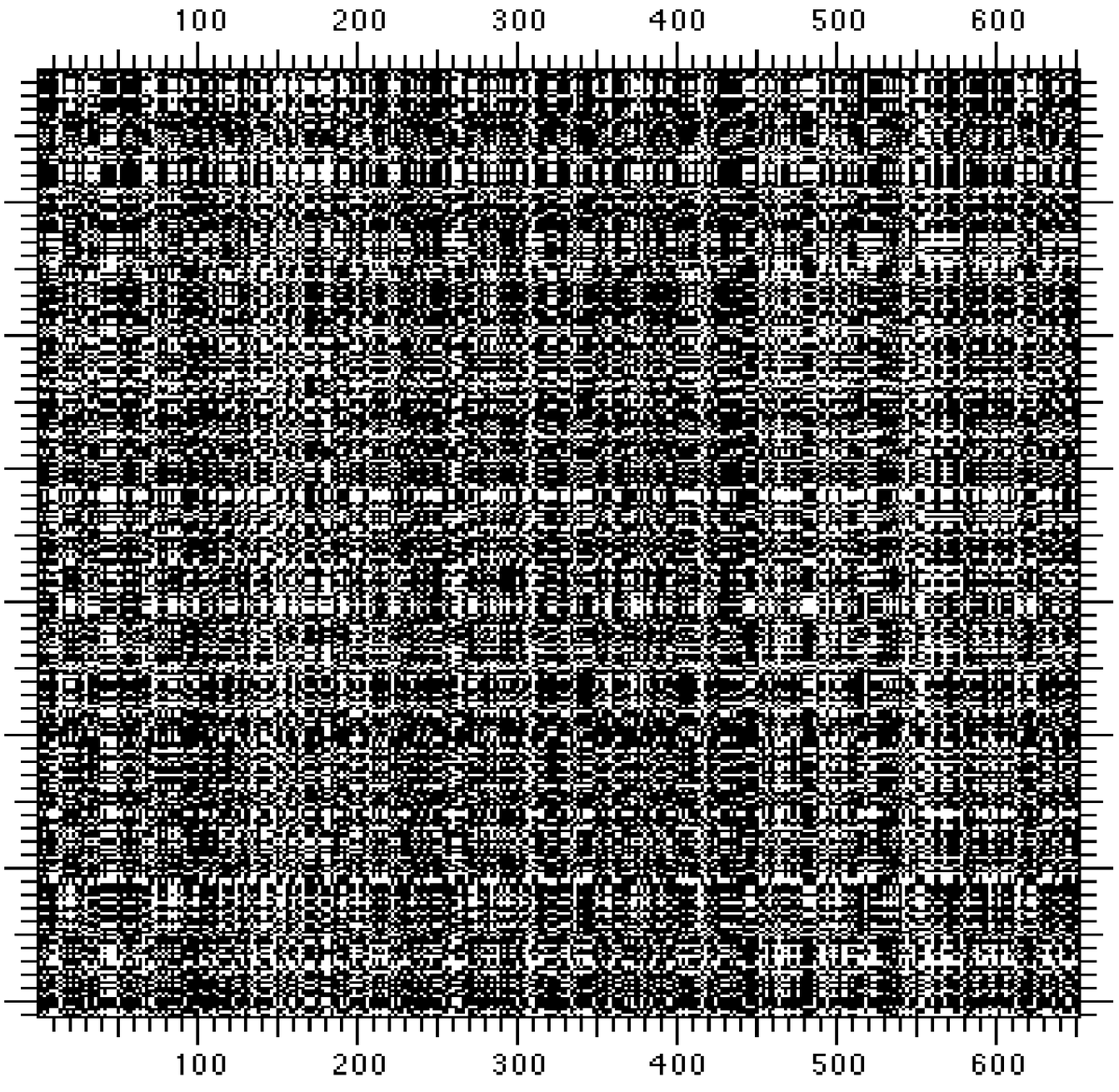
- improve visualization by filtering random matches
- compare a block of positions rather than comparing a single sequence position
- find direct or inverted repeats

Sliding windows improve dot plot visualization

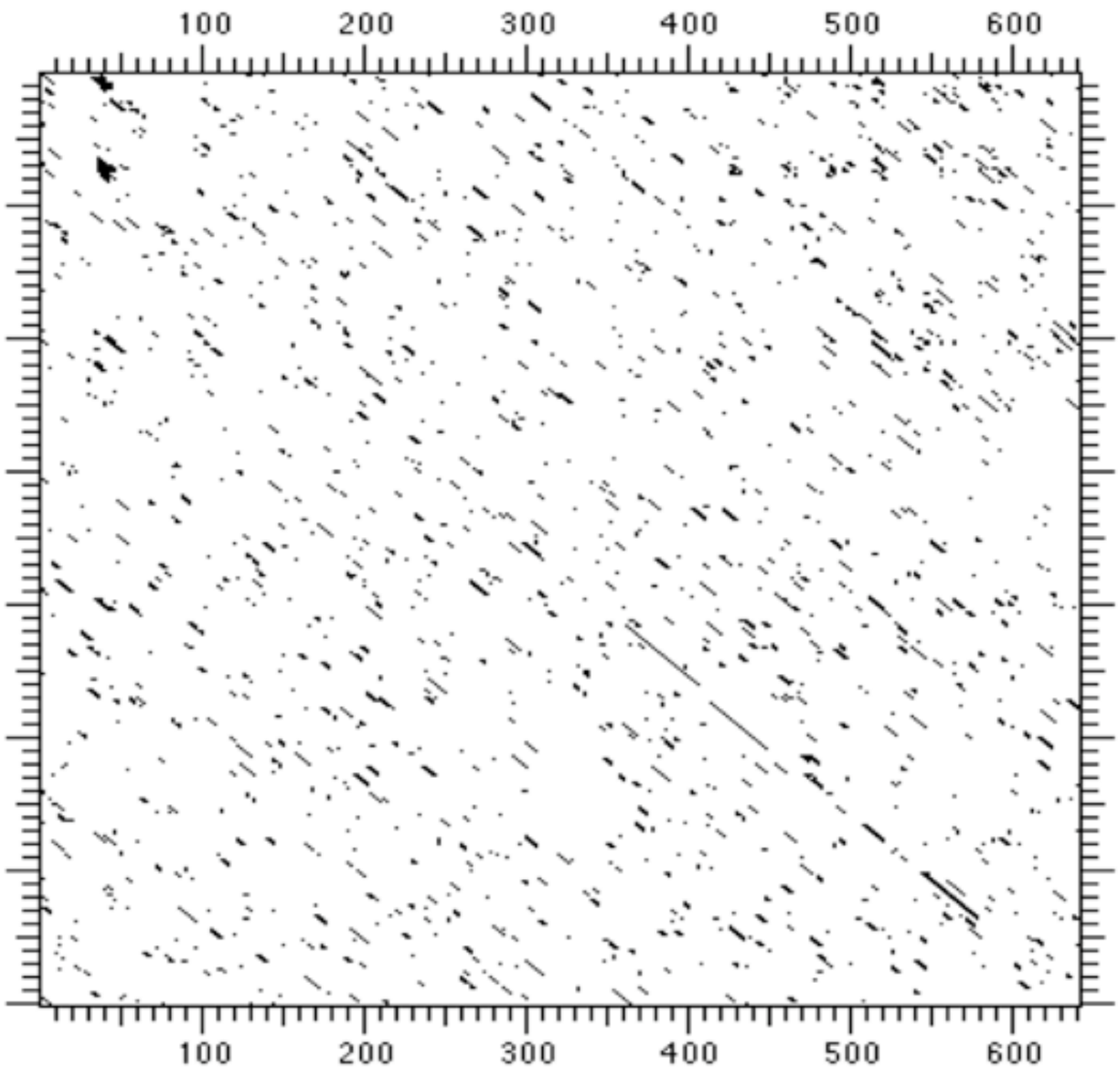
- instead of dots for each character compare a number of positions (window size)
- mark a dot where there is minimum number (stringency) of identical characters
- stringency = some arbitrary threshold
- different values for different problems

  - for DNA, set sliding window size = 15; stringency=10
  - for proteins set sliding window size = 2 or 3; stringency=2

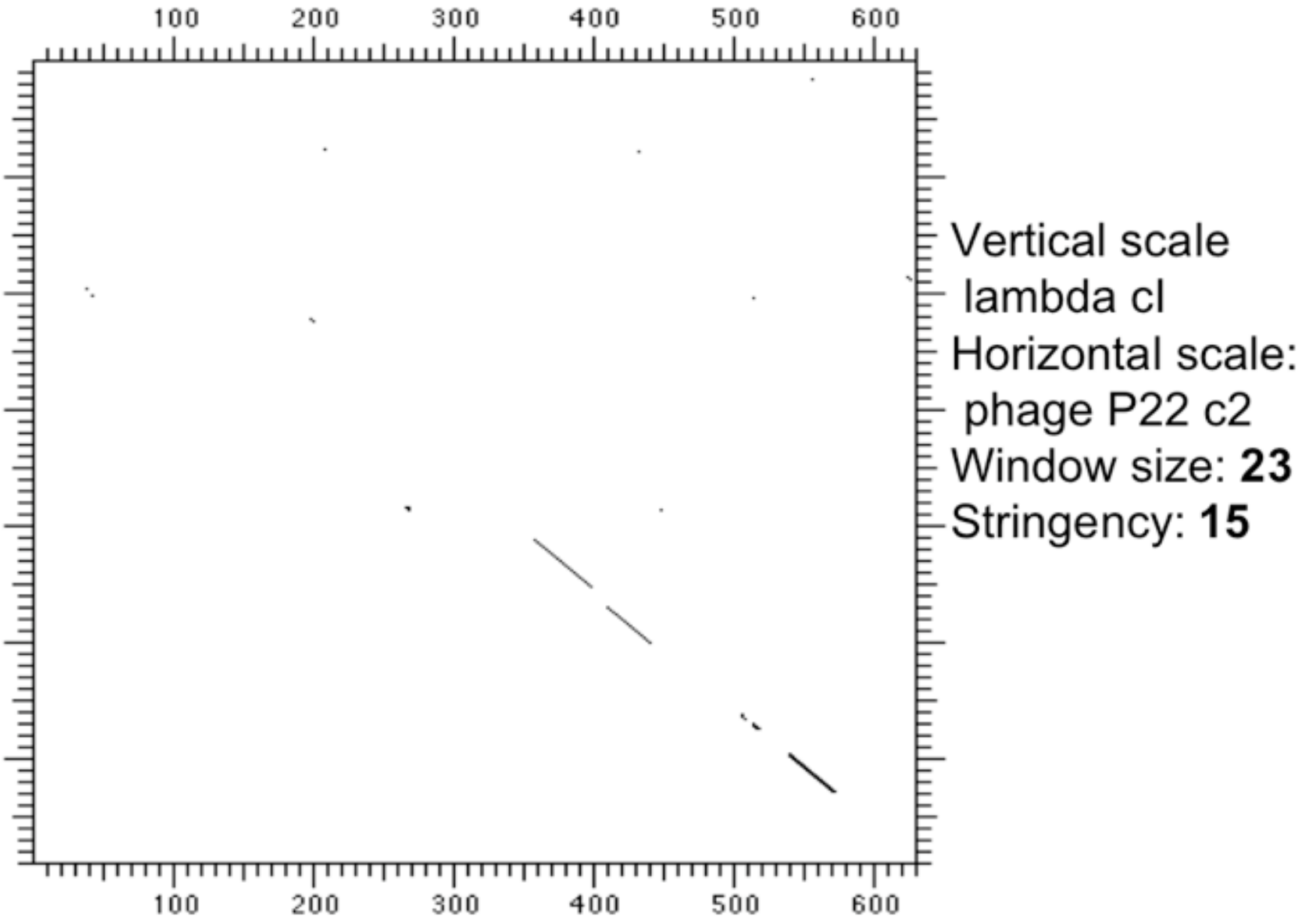
Sliding windows (too small?)



Vertical scale:
lambda cl
Horizontal scale:
P hage P22 c2
Window size: **1**
Stringency: **1**



Vertical scale
lambda cI
Horizontal scale:
phage P22 c2
Window size: **11**
Stringency: **7**


100
200
300
400
500
600
Vertical scale
lambda cl
Horizontal scale:
phage P22 c2
Window size: 23
Stringency: 15

*

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-AGC-TCC-TGG-TTA-CTC-CTG-GGA-TCC
GAG-GGT-TCC-AGC-TAT-CTA-TTA-ATT-GGT-AGC
GAC-GGC-AGT-GCA-TGG-TTG-CTT-TTG-GGC-AGT
GAT-GGG-TCA-GCT-TAC-CTC-CTG-GCC-GGG-TCA
```

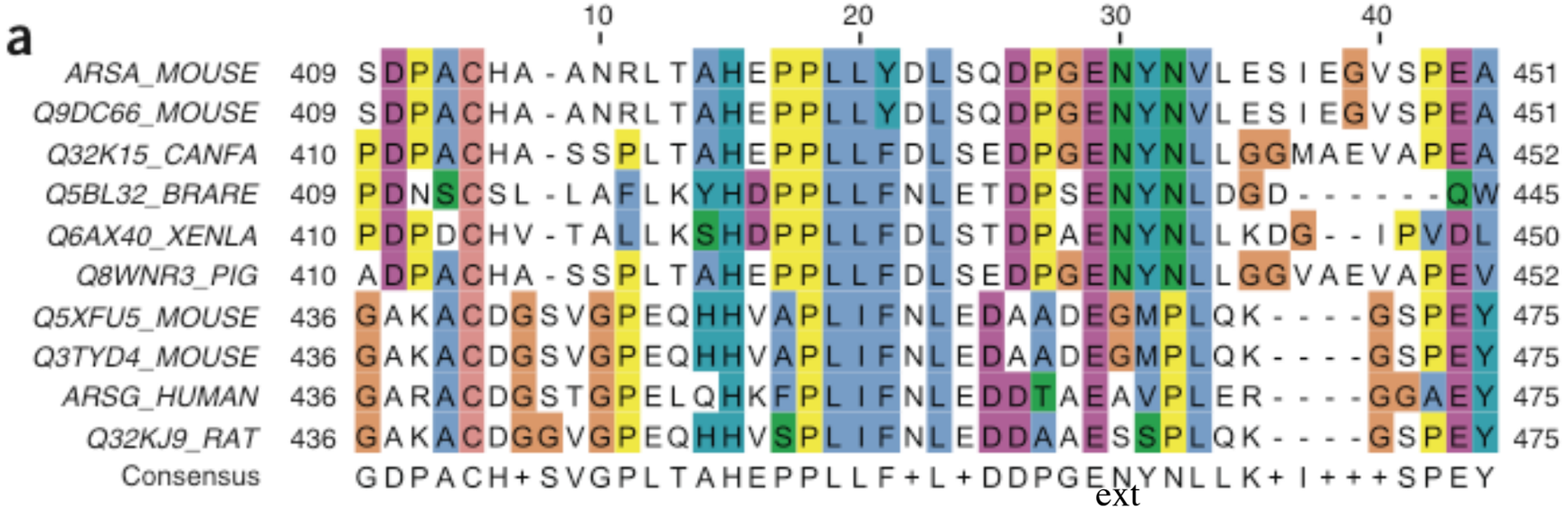
Calculate % identity both ways

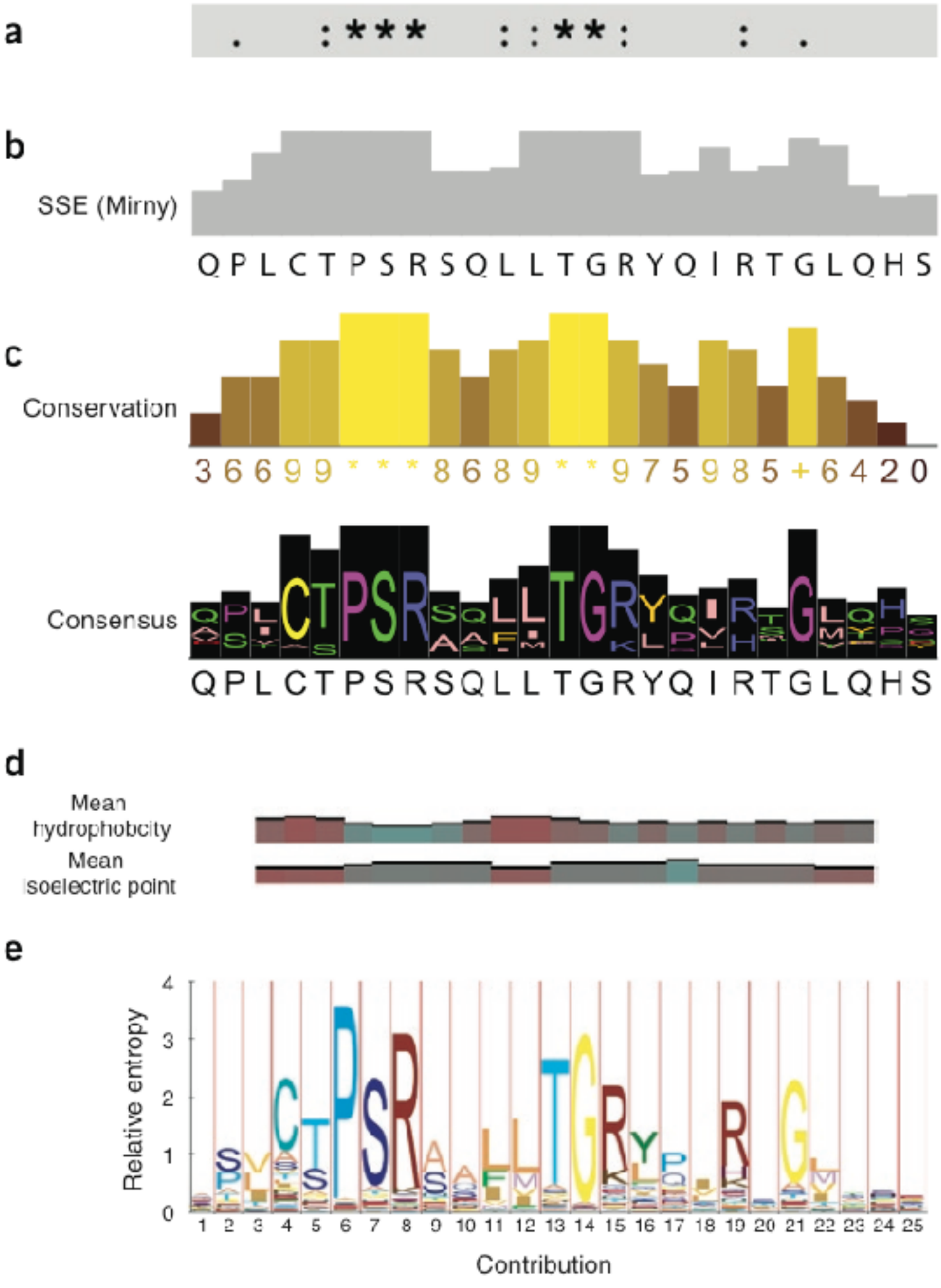
9/10 (aa) : 90%
14/30 (nt) : 47%

PhyloServer

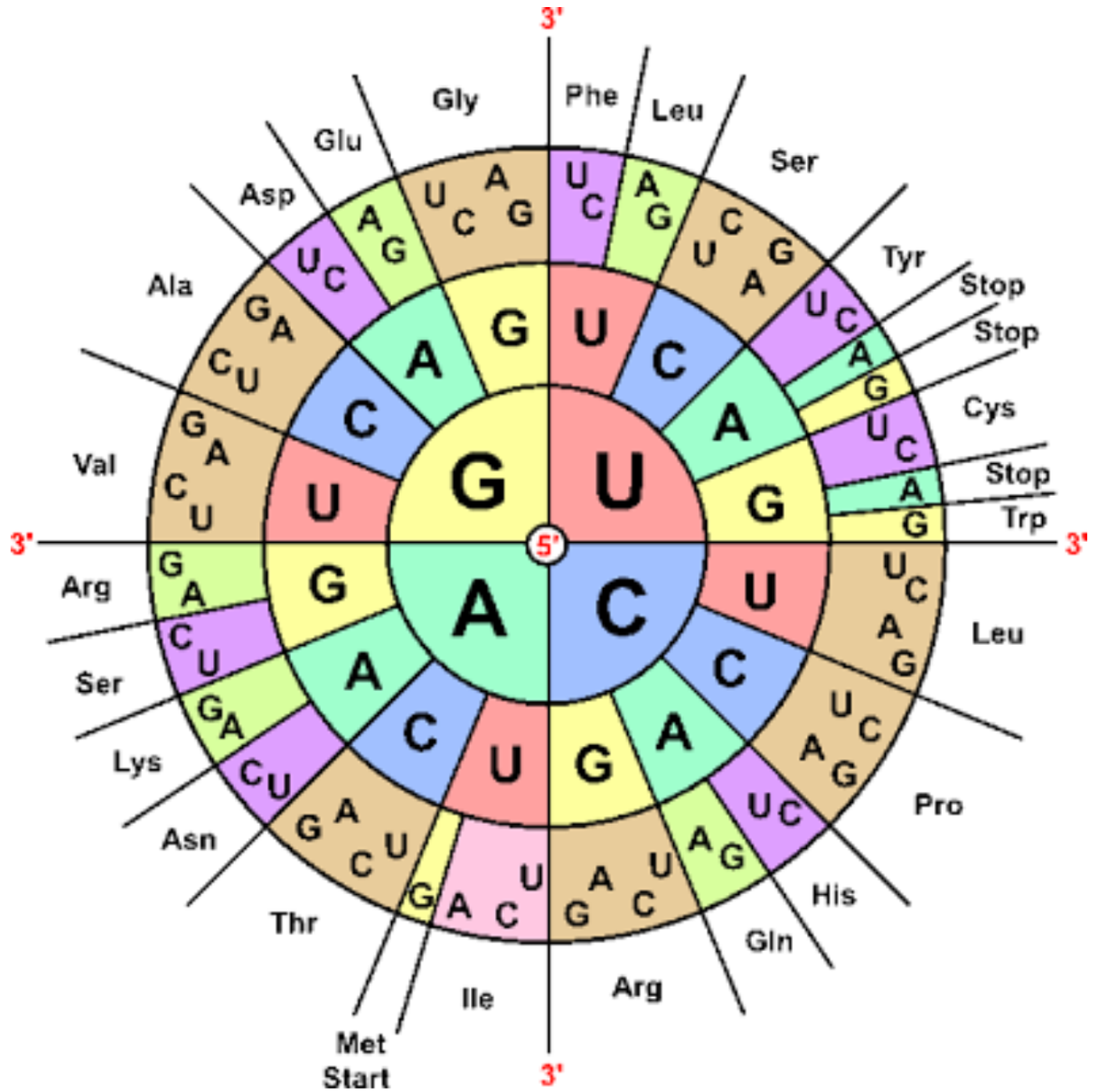
http://phylo.cs.mcgill.ca

Jalview


a
10
20
30
40
ARSA_MOUSE 409 SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA 451
Q9DC66_MOUSE 409 SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA 451
Q32K15_CANFA 410 PDPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGMAEVAPEA 452
Q5BL32_BRARE 409 PDNSCSL-LAFLKYHDPPLLFNLETDPSENYNLDGD------QW 445
Q6AX40_XENLA 410 PDPDCHV-TALLKSHDPPLLFDLSTDPAENYNLLKDG--IPVDL 450
Q8WNR3_PIG 410 ADPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGVAEVAPEV 452
Q5XFU5_MOUSE 436 GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY 475
Q3TYD4_MOUSE 436 GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY 475
ARSG_HUMAN 436 GARACDGSTGPELQHKFPLIFNLEDDTAEAVPLER----GGAEY 475
Q32KJ9_RAT 436 GAKACDGGVGPEQHHVSPLIFNLEDDAAESSPLQK----GSPEY 475
Consensus GDPACH+SVGPLTAHEPPLLF+L+DDPGENYNLLK+I+++SPEY
ext
b
Amino acid
Buried index
Turn propensity
Strand propensity
Helix propensity
Hydrophobicity
Taylor
Zappo
ClustalX
A
R
N
D
C
Q
E
G
H
I
L
K
M
F
P
S
T
W
Y
V
B
X
Z
c
Nucleotide
Nucleic Acid Database A C G T U
Jalview/TOPALi A C G T U
ClustalX A C G T U
MacClade A C G T U
GC versus AT (Geneious) A C G T U
Purine versus pyrimidine A C G T U
d
Hydrophobic
Small
Tiny
Aliphatic
Aromatic
Polar
Neg. charge
Pos. charge
P
A
G
S
N
T
D
E
K
R
Q
H
Y
W
F
M
I
L
V
C S–S
C S–H


a
. :*** ::**: : .
b
SSE (Mirny)
Q P L C T P S R S Q L L T G R Y Q I R T G L Q H S
c
Conservation
3 6 6 9 9 * * * 8 6 8 9 * * 9 7 5 9 8 5 + 6 4 2 0
Consensus
Q P L C T P S R S Q L L T G R Y Q I R T G L Q H S
d
Mean hydrophobcity
Mean isoelectric point
e
Relative entropy
0
1
2
3
4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Contribution

How to design degenerate PCR primers?



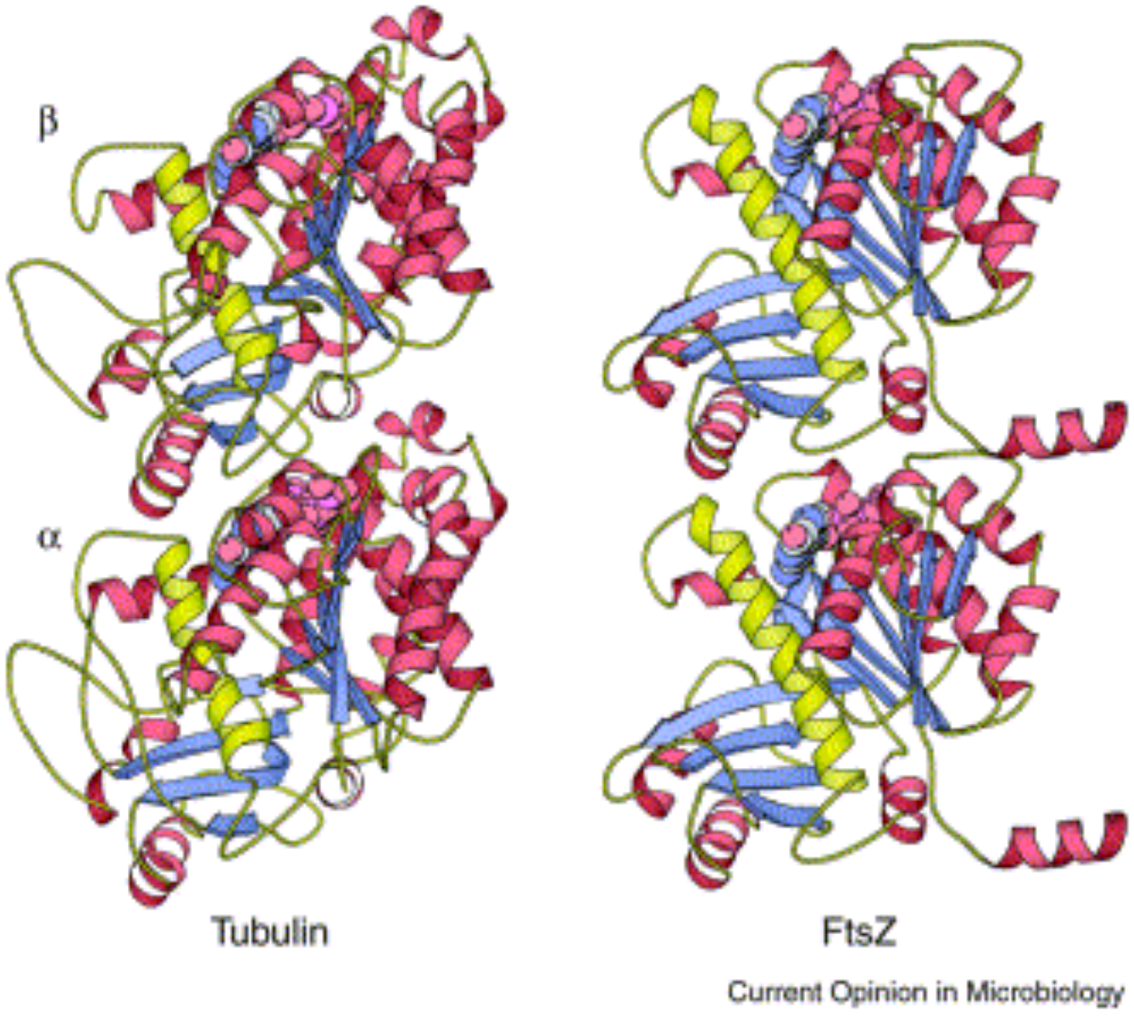
https://github.com/andand/DEGEPRIME

How to recognize deep homology?

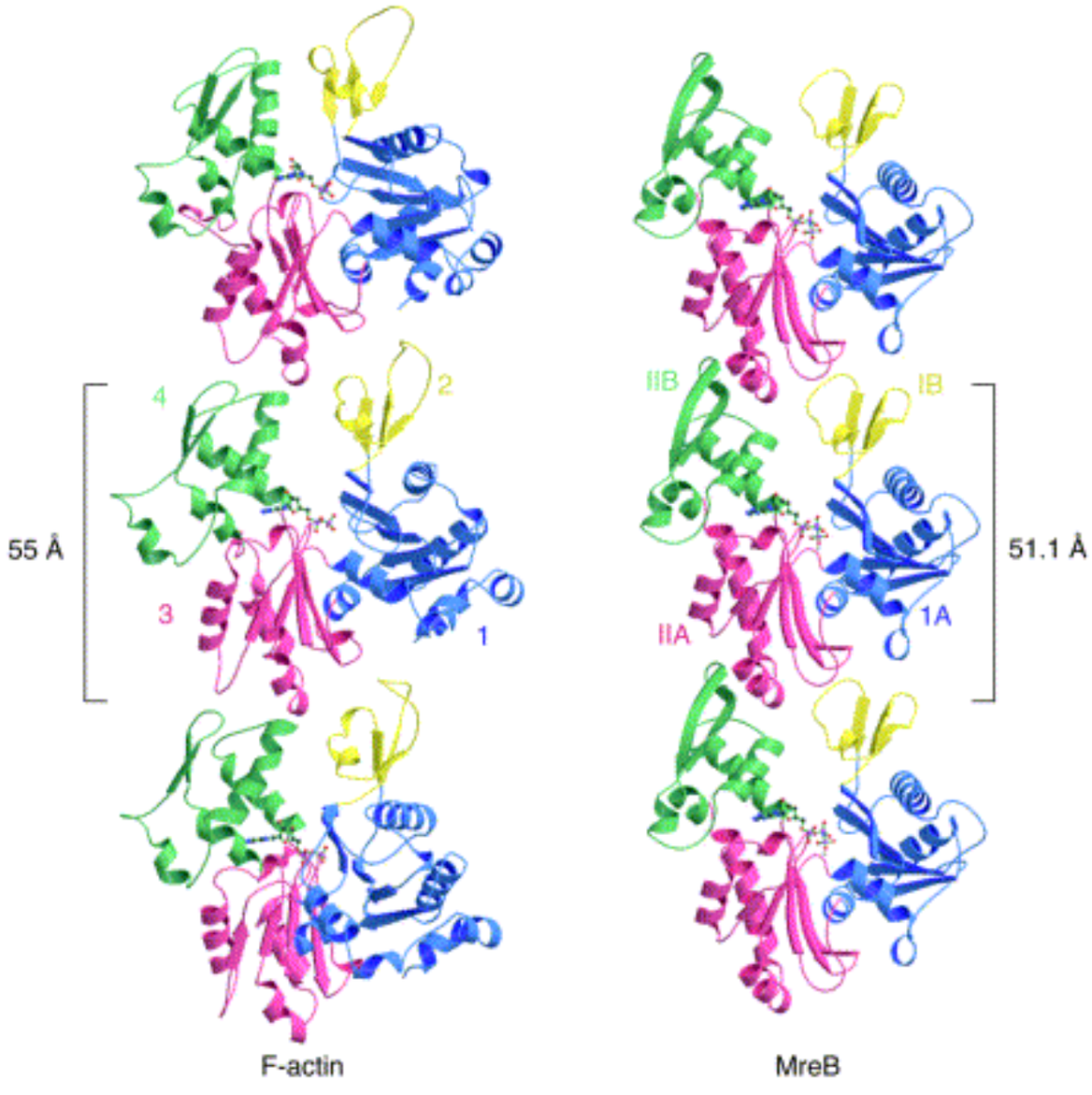
what are potential problems?

* Conserved tertiary structure, but not primary structure (sequence)

Use tertiary structure as alignment guide!

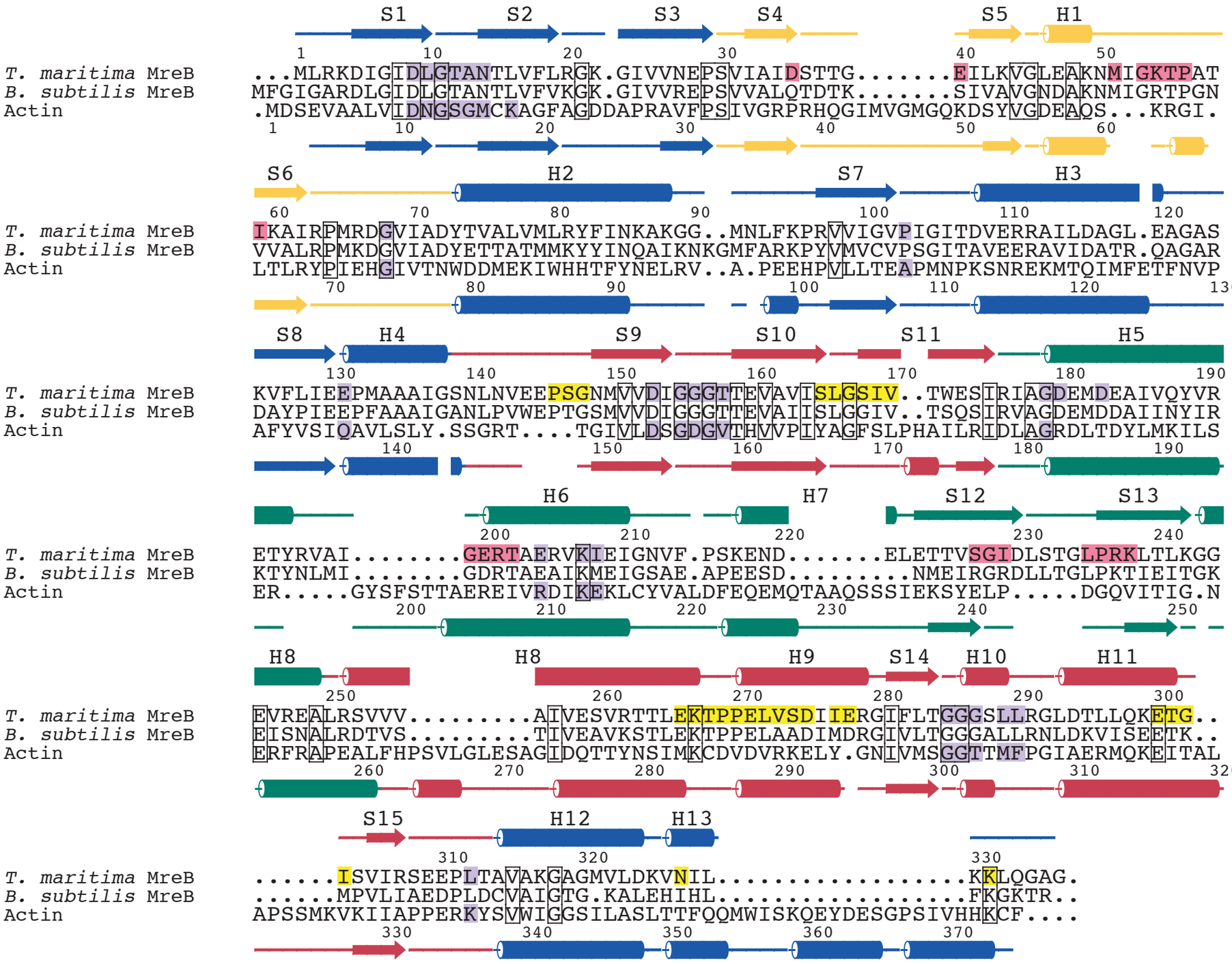


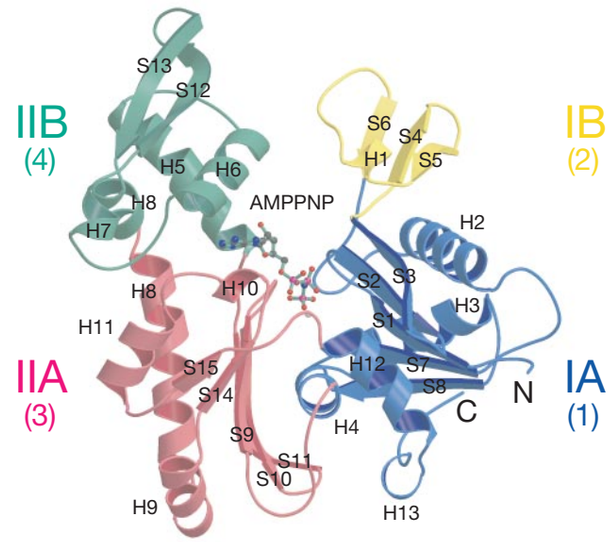
10% identity



10% identity

Mre B vs. Actin Structure-based Alignment





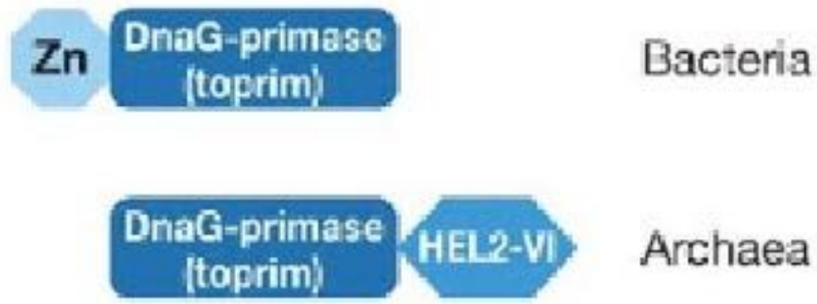
boxed = all 3

active sites = purple

MreB identities = 56%

MreB v. actin = 15%

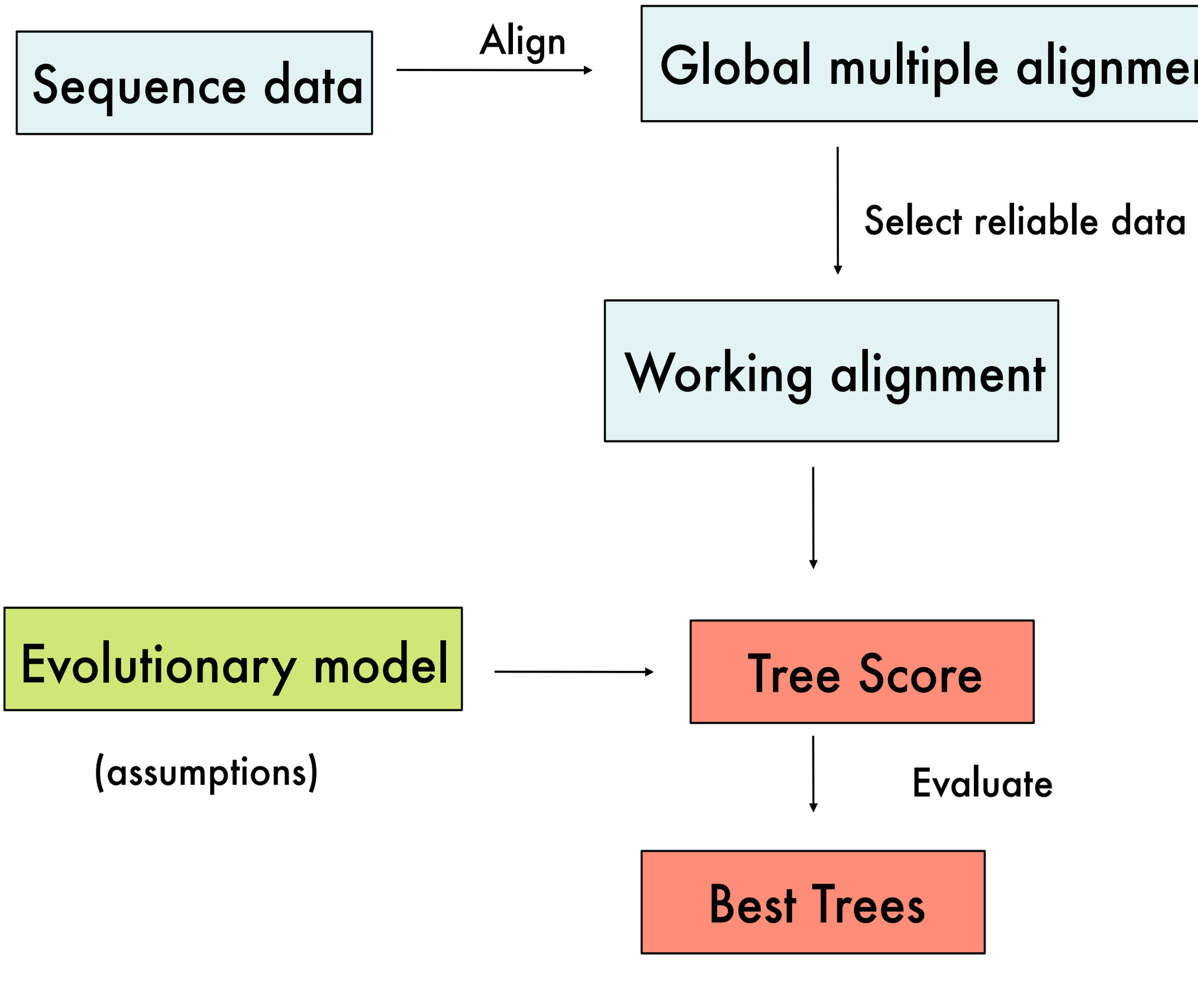
a


Zn
DnaG-primase (toprim)
Bacteria
DnaG-primase (toprim)
HEL2-VI
Archaea

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

*

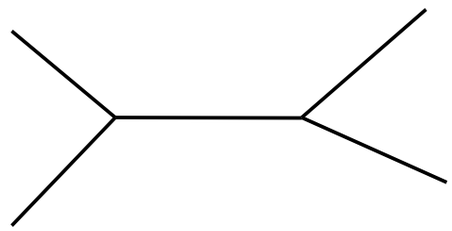

Sequence data
Align
Global multiple alignment
Select reliable data
Working alignment
Evolutionary model
(assumptions)
Tree Score
Evaluate
Best Trees

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=

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

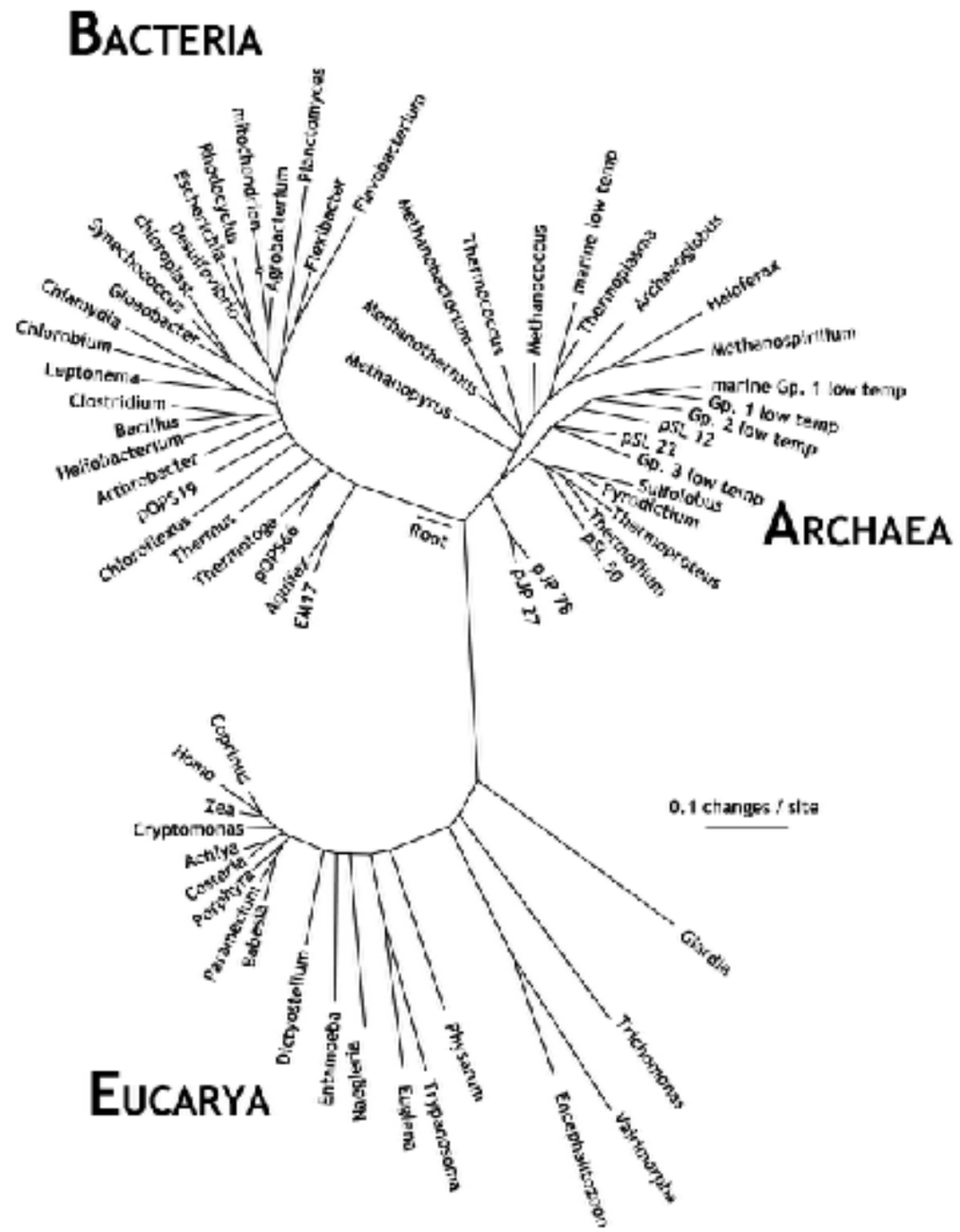
1, 3, 15, 105, 945, 10395....

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

red = GEBA (new genomes)

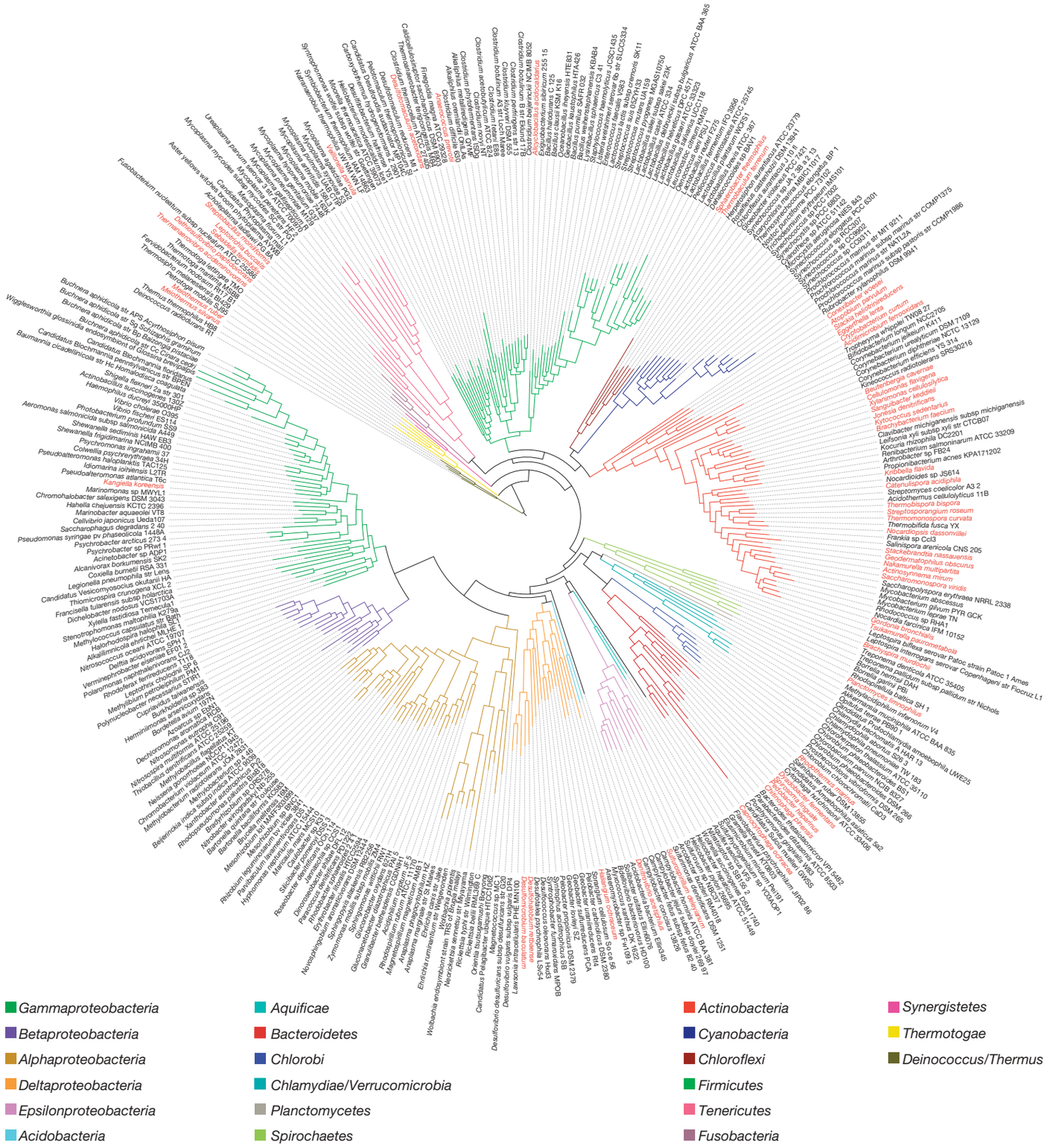


Figure 1 | Maximum-likelihood phylogenetic tree of the bacterial domain based on a concatenated alignment of 31 broadly conserved protein-coding genes[16]. Phyla are distinguished by colour of the branch and GEBA genomes are indicated in red in the outer circle of species names.

Genome-enabled strategies for determining evolutionary relationships

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

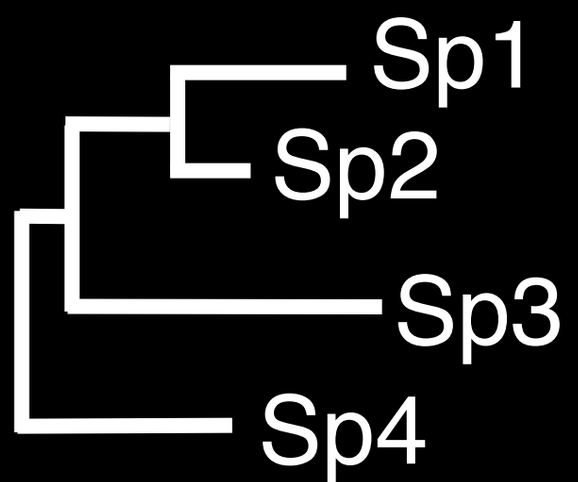
do all genes have similar evolutionary histories? (i.e., are all genes good proxies?)

Multiple genes separate analysis

Gene 1

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

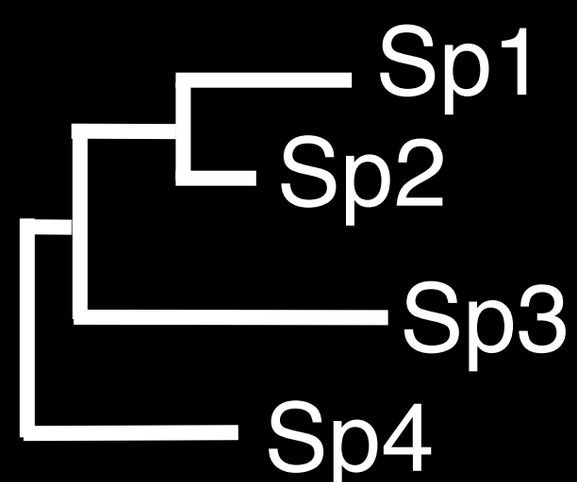
Evolutionary model



Gene 2

```
Sp1: TCTTT…GAA
Sp2: TCGCT…GGA
Sp3: CTATT…GGA
Sp4: CCATT…GGA
```

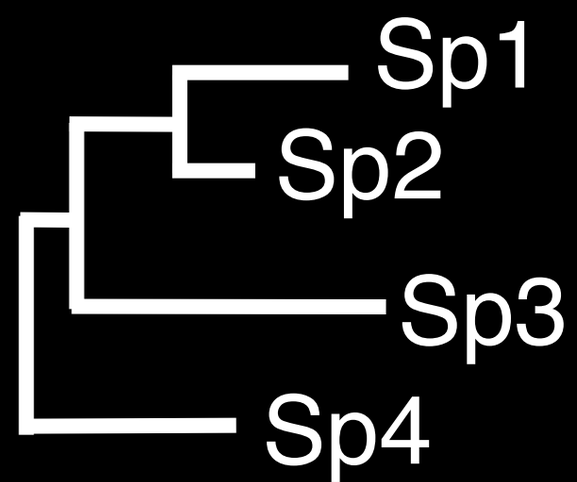
Evolutionary model



Gene 3

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

Evolutionary model



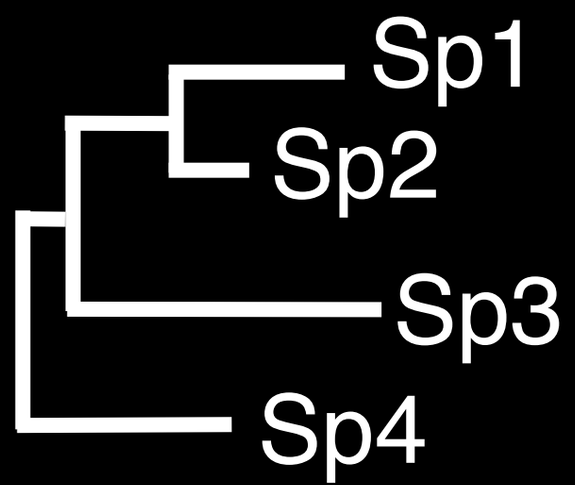
e.g., Murphy *et al.* (2001)

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)

Article

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

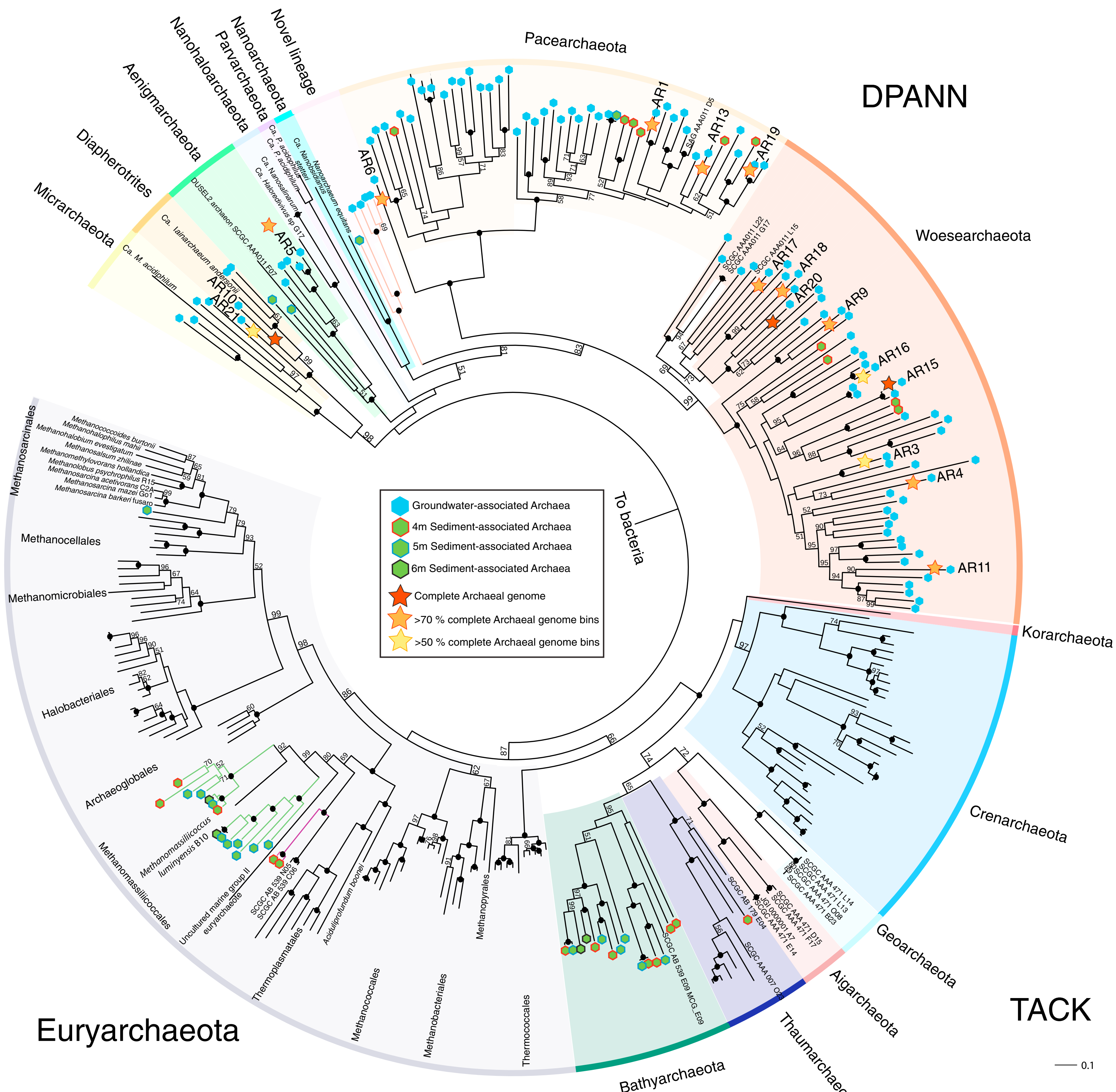
Correspondence

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.

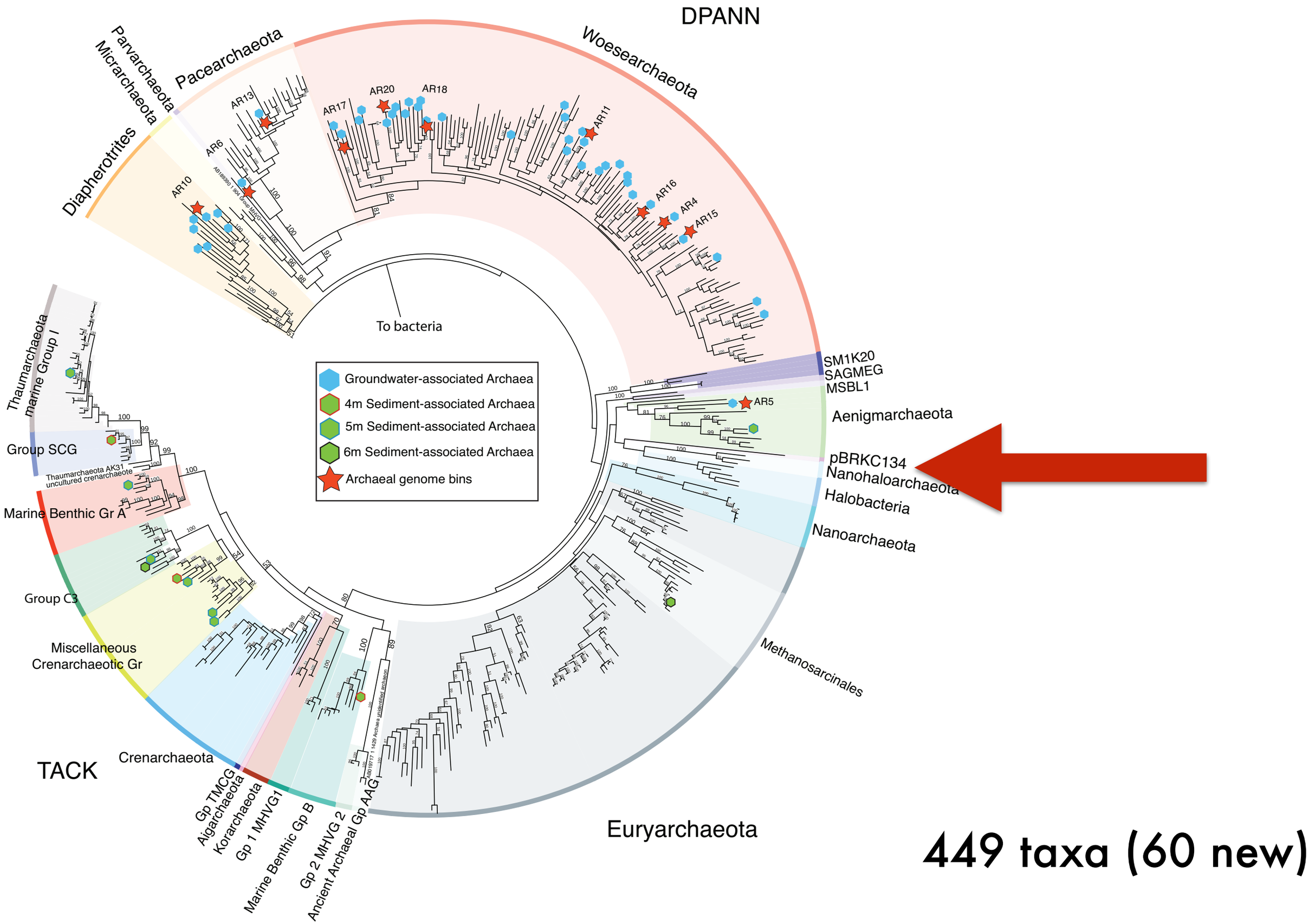
concatenated ribosomal protein phylogeny



153 archaeal taxa

15 proteins

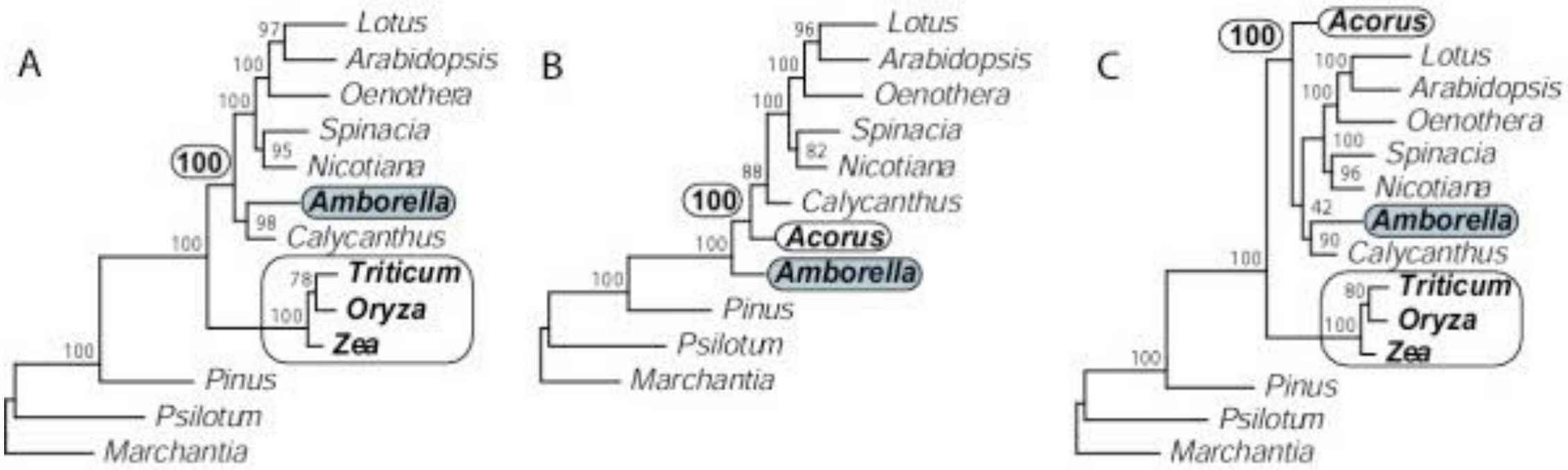
ssu rRNA phylogeny



449 taxa (60 new)

*

Taxon sampling is a big deal

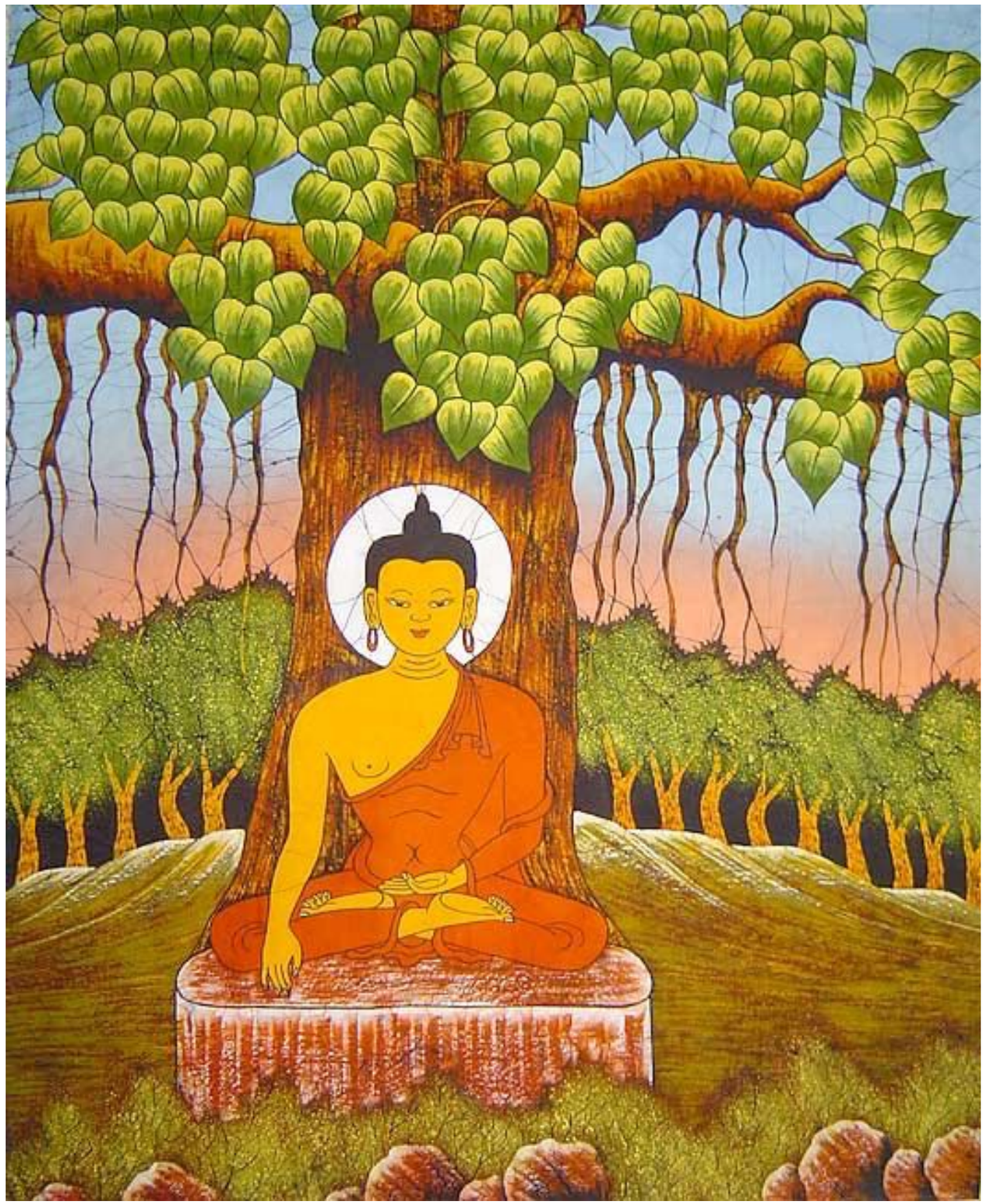


What is the closest relative of Amborella in each tree?
(each is Maximum parsimony, different taxa)

Taxonomic sampling= more variation

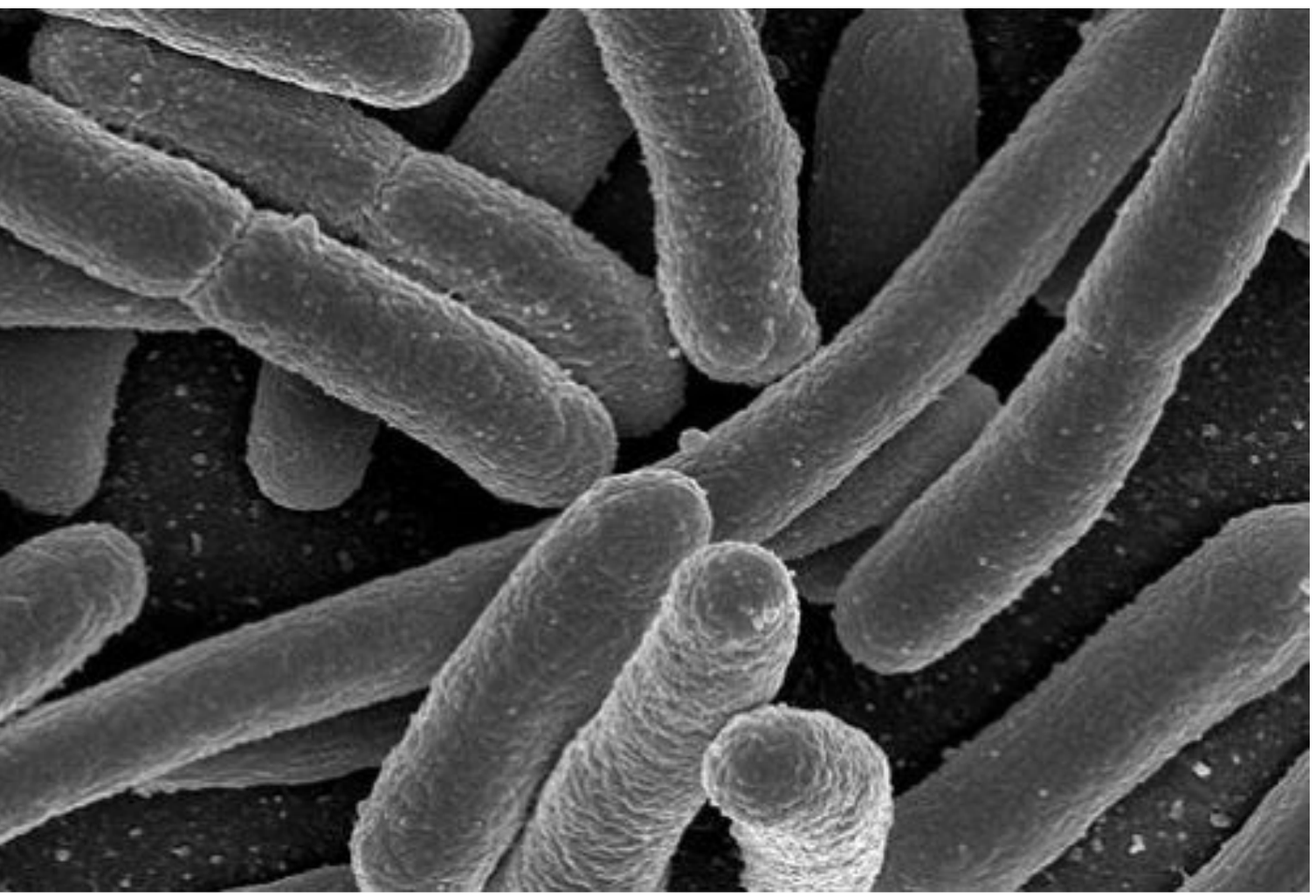
- In general, more taxa (diverse) better trees
- less taxa has negative impact on the robustness of phylogenies
- affects the number of groups/relationships among groups
- causes trouble in sorting out gene families

What's a bacterial "species"?

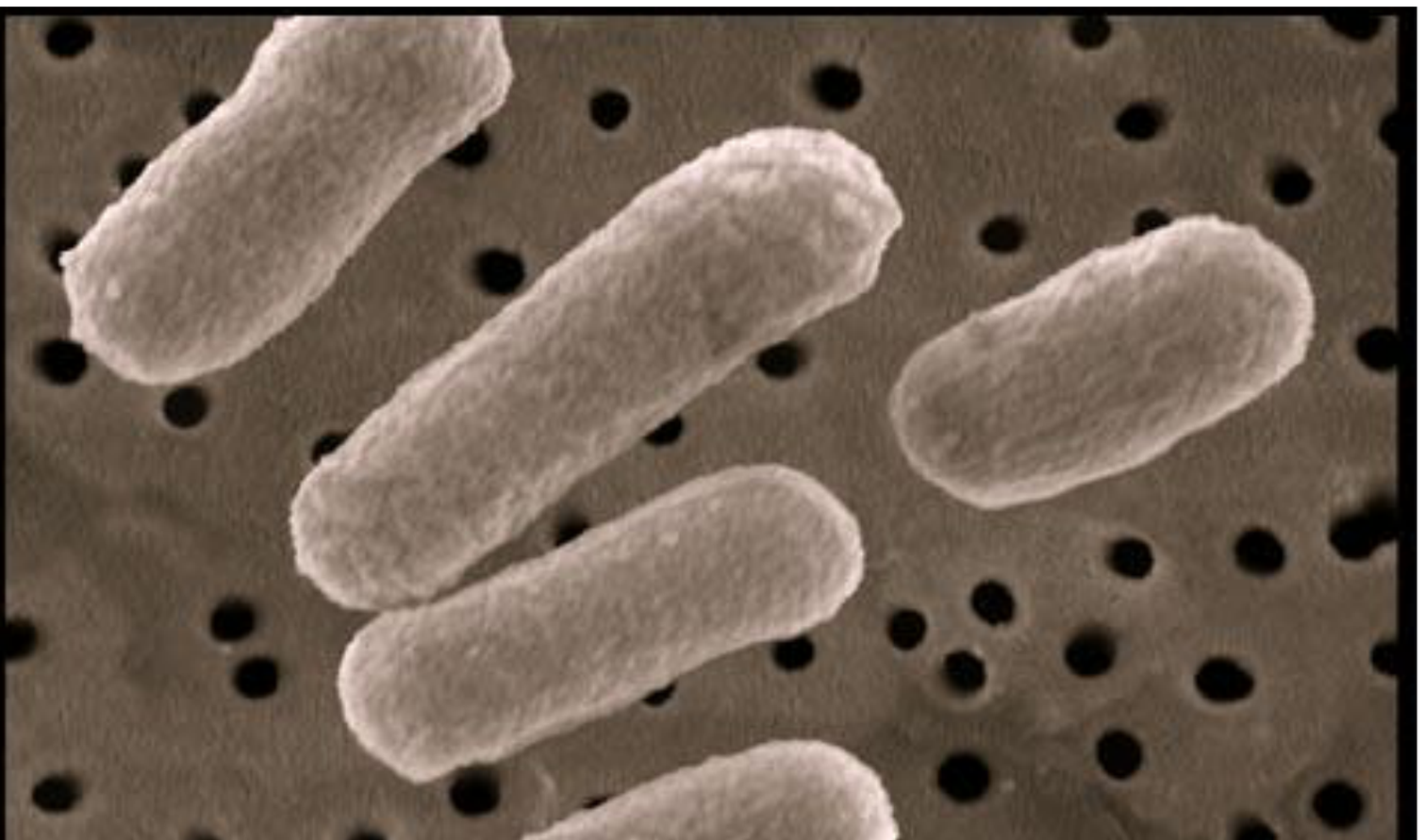


genotype or phenotype?

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[a])	Phylogenetic group[b]	Extraintestinal mouse model phenotype[c] (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
IAI1	**Human**	**Faeces**	**Commensal**	**B1**	**NK (0)**	**This work**
55989	**Human**	**Faeces**	**Diarrhoea (EAEC)**	**B1**	**K (10)**	**This work**
S. boydii 4 227 (Sb 227)	Human	Faeces	Shigellosis	S1	ND[d]	[116]
S. sonnei 046 (Ss 046)	Human	Faeces	Shigellosis	SS	ND	[116]
S. flexneri 2a 301 (Sf 301)	Human	Faeces	Shigellosis	S3	ND	[117]
S. flexneri 2a 2457T (Sf 2457T)	Human	Faeces	Shigellosis	S3	NK (0)	[118]
S. flexneri 5b 8401 (Sf 8401)	Human	Faeces	Shigellosis	S3	ND	[119]
S. dysenteriae 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]
UMN026	**Human**	**Urine**	**Cystitis (ExPEC)**	**D**	**K (10)**	**This work**
IAI39	**Human**	**Urine**	**Pyeloneprhitis (ExPEC)**	**D**	**K (8)**	**This work**
UTI89	Human	Urine	Cystitis (ExPEC)	B2	K (10)	[122]
APEC O1	Chicken	Lung	Colisepticemia (ExPEC)	B2	K (10)	[123]
S88	**Human**	**Cerebro-spinal fluid**	**New born meningitis (ExPEC)**	**B2**	**K (10)**	**This work**
CFT073	Human	Blood	Pyeloneprhitis (ExPEC)	B2	K (10)	[30]
ED1A	**Human**	**Faeces**	**Healthy subject**	**B2**	**NK (0)**	**This work**
536	Human	Urine	Pyeloneprhitis (ExPEC)	B2	K (10)	[124]
E. fergusonii	**Human**	**Faeces**	**Unknown**	**Outgroup**	**NK (1)**	**This work**

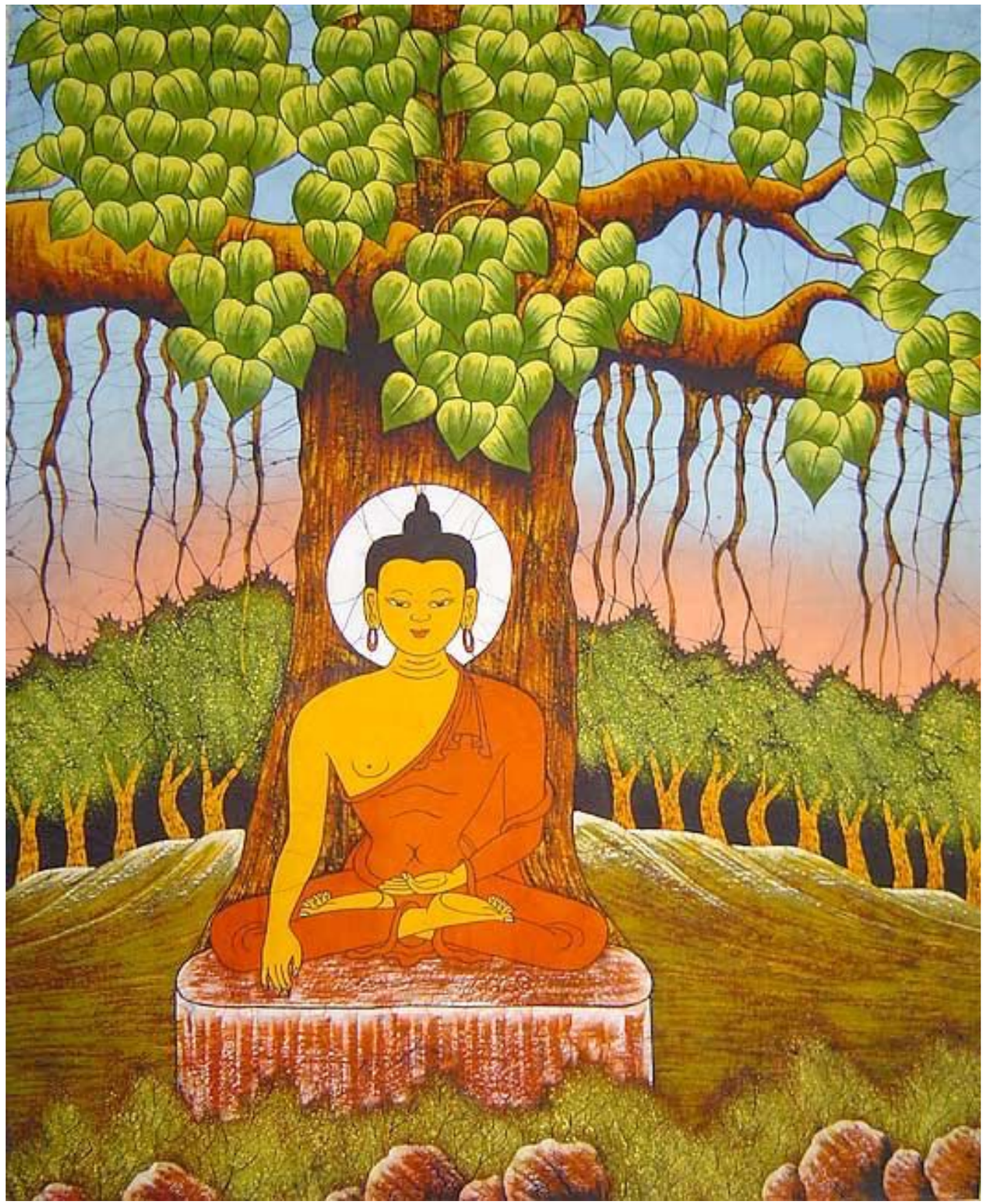
E coli strain genome characteristics

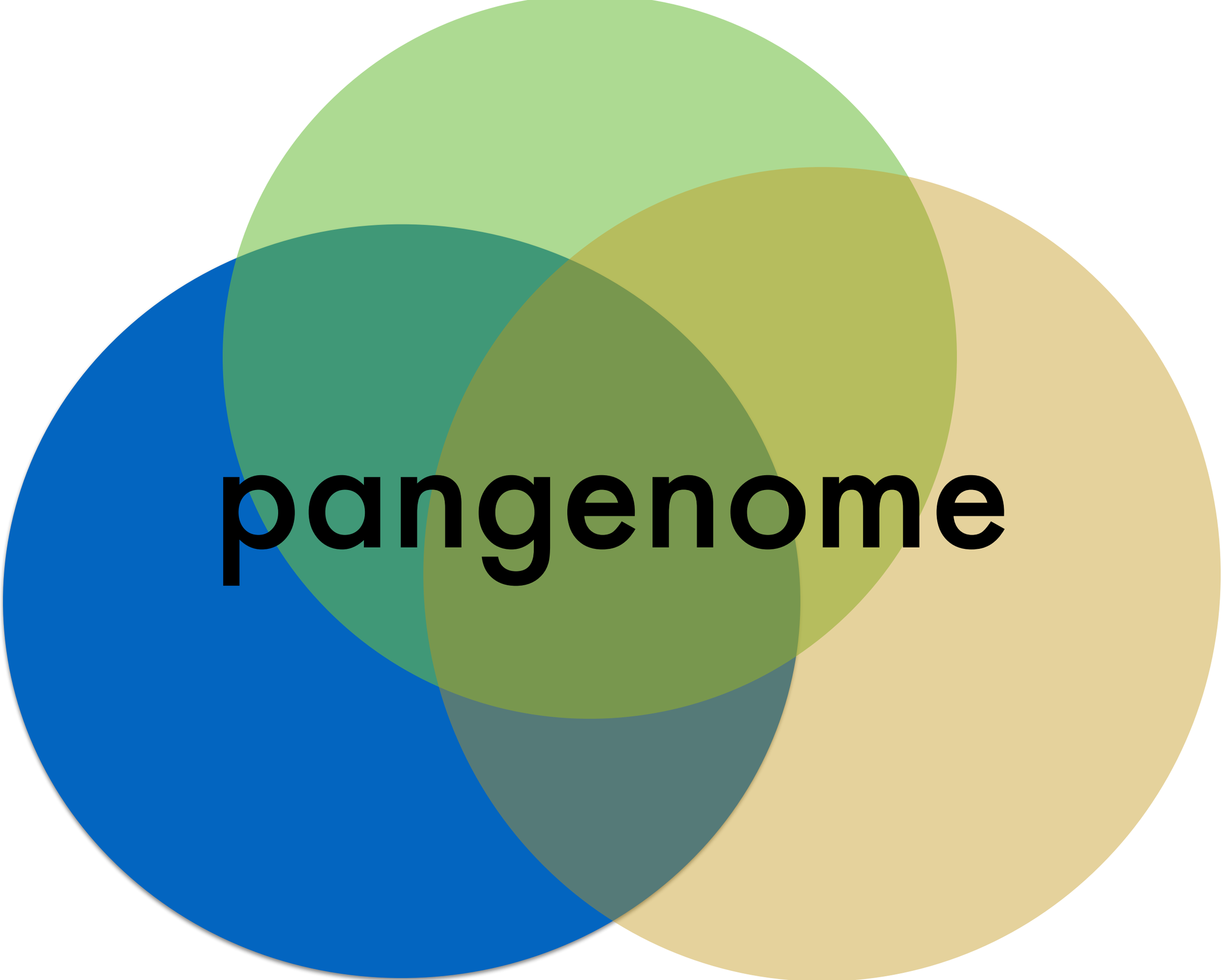
Table 2. General features of the *Escherichia coli* and *E. fergusonii* genomes sequenced in this work with *E. coli* K-12 MG1655 as reference (chromosome features).

Chromosome features	***E. coli* K-12 MG1655**	***E. coli* strains**						***E. fergusonii* 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	4 588 711
G+C content (%)	50.8	50.7	50.8	50.7	50.7	50.6	50.7	49.9
rRNA operons	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)	7 (+5S)
tRNA genes	86	94	86	91	91	88	88	87
Total Protein-coding genes[a]	4306	4969	4491	5129	4859	4906	4918	4336
Pseudogenes[b] (nb)	81	79	51	95	90	80	45	22
Protein coding density[c]	85.7	87.4	87.6	86.2	87	86.1	87.8	84.7
Assigned function[d] (%)	80	74	77	74	77	78	76.5	77
Conserved hypothetical (%)	12.5	23	21.5	23	22	20	22	20
Orphans (%)	7.5	3	1.5	3	1	2	1.5	3
IS-like genes (nb)	66	150	42	118	47	224	92	29
Phage-associated genes (nb)	231	406	201	657	507	393	429	235

[a]The number of protein-coding genes is given without the number of coding sequences annotated as artifactual genes (Supplementary Table 2A).
[b]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.
[c]The computed protein coding density takes into account the total length of protein genes excluding overlaps between genes, artifacts, and RNA genes.
[d]Protein genes with assigned function include the total number of definitive and putative functional assignments.
doi:10.1371/journal.pgen.1000344.t002

Please discuss.




pangenome

core

Escherichia coli

Core genome

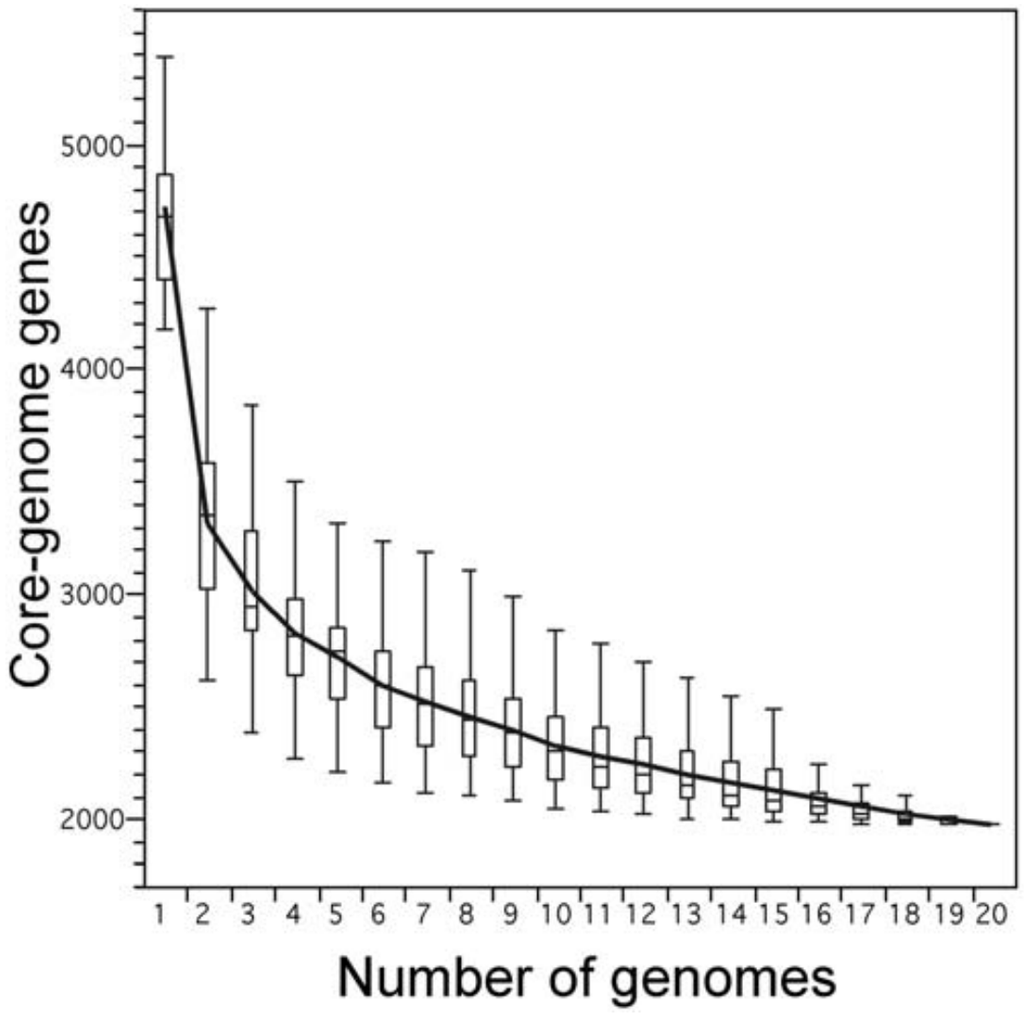
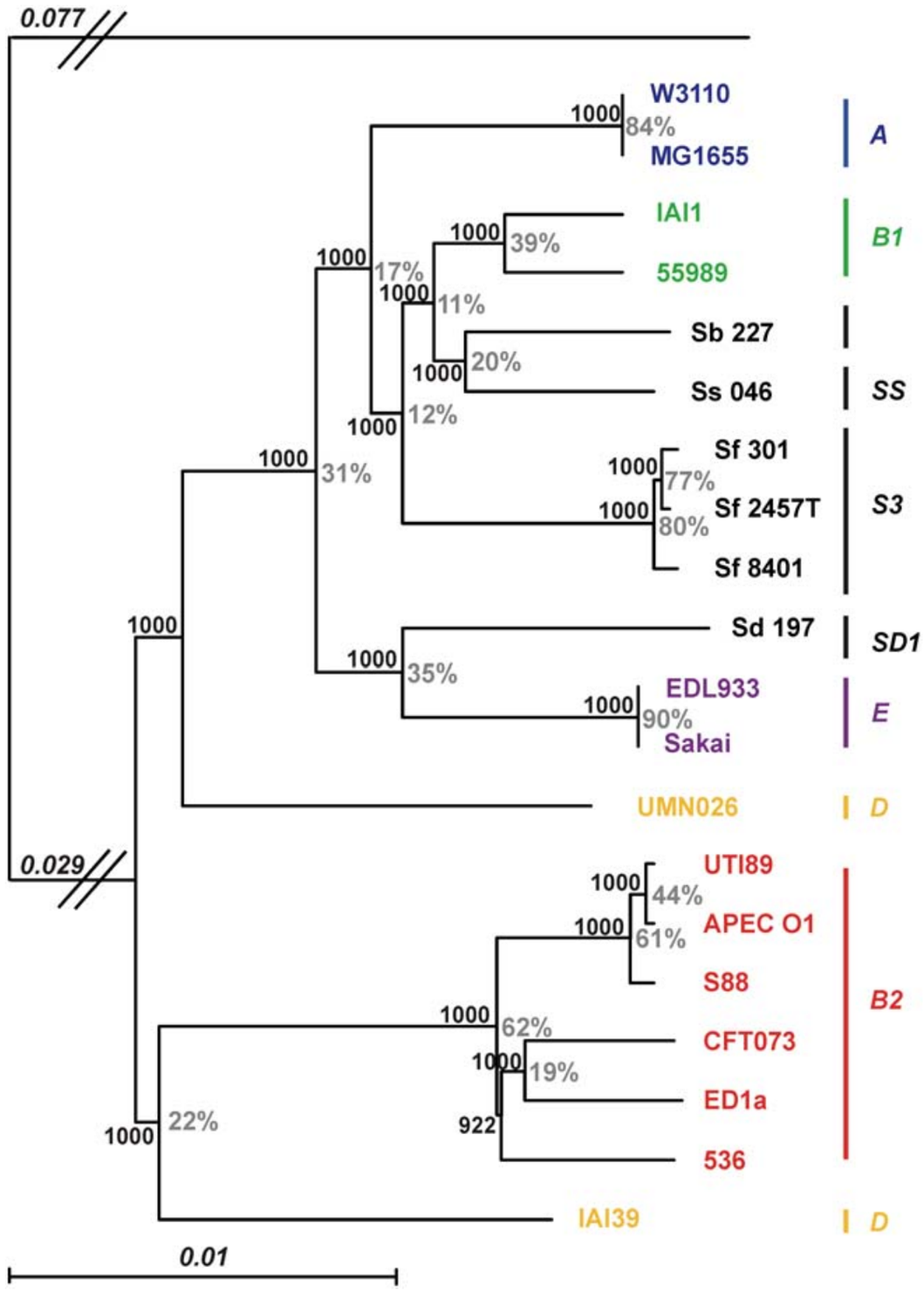


Figure 1. ***Escherichia coli*** **core and pan-genome evolution according to the number of sequenced genomes.** Number of genes in common (left) and total number of non-orthologous genes (right) for a given number of genomes analysed for the different strains of *E. coli*. The upper and lower edges 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

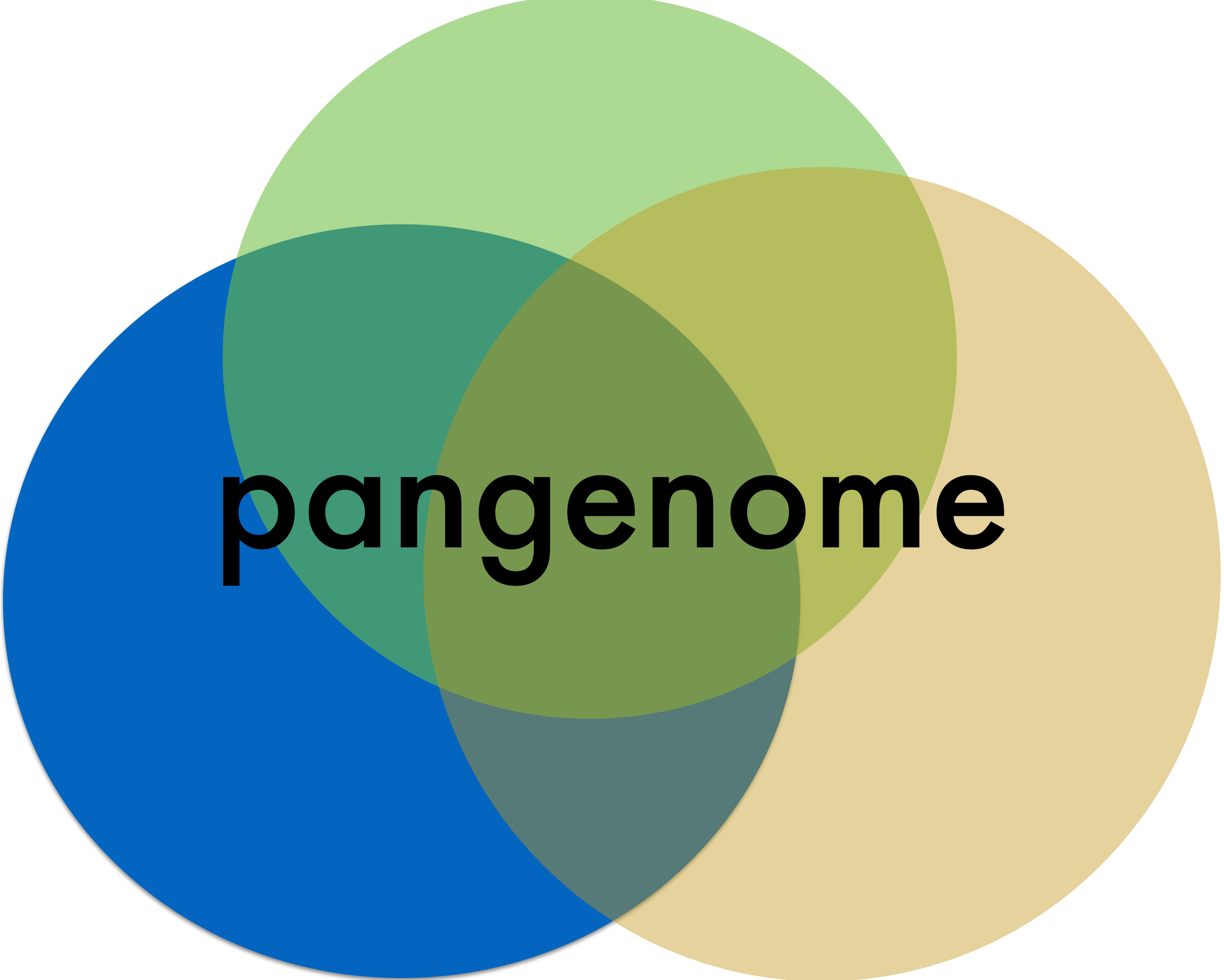


1878 gene
Core genome phylogeny

bootstrap values (1000)

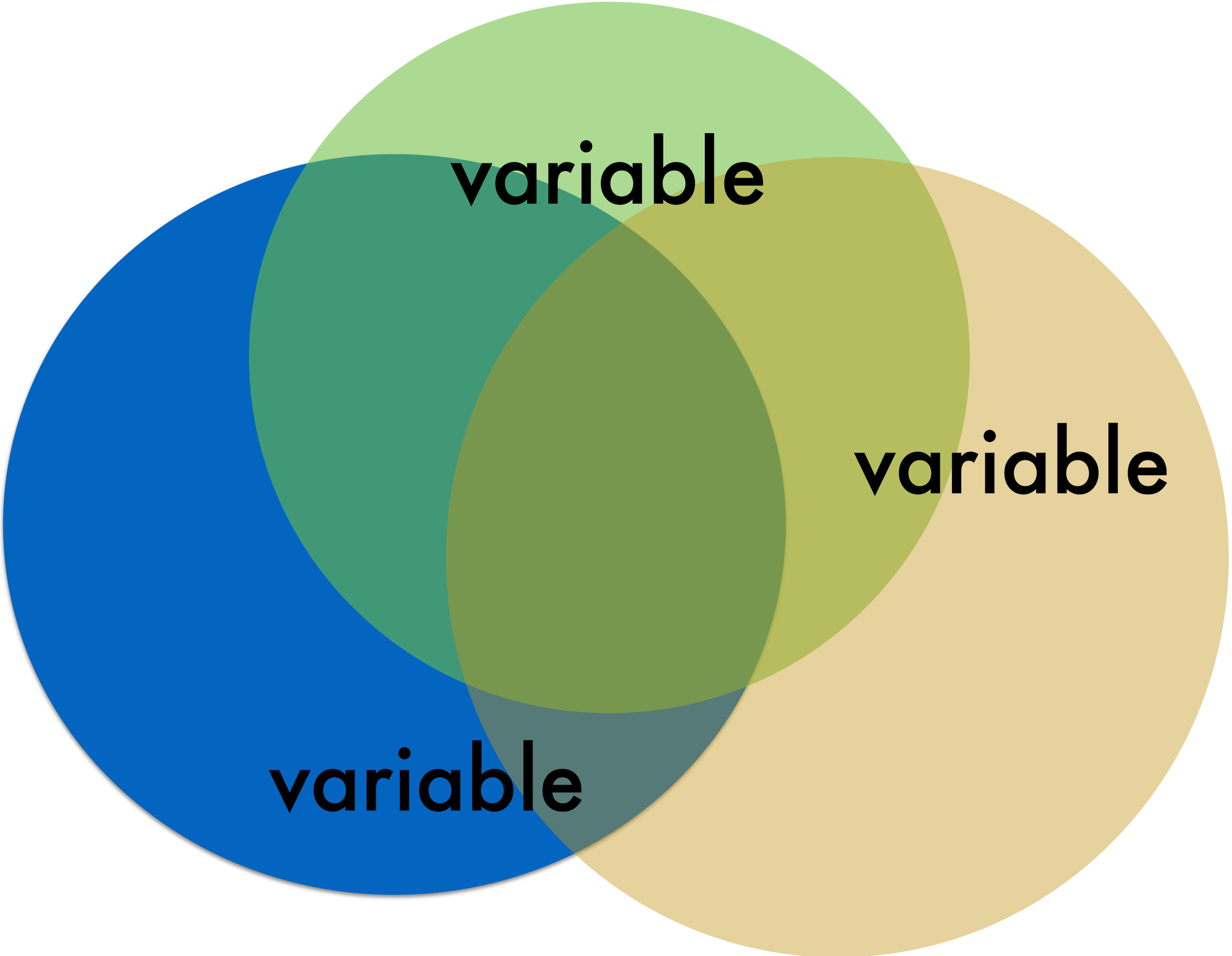
% = consensus strength
(number of genes confirming that node)

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'

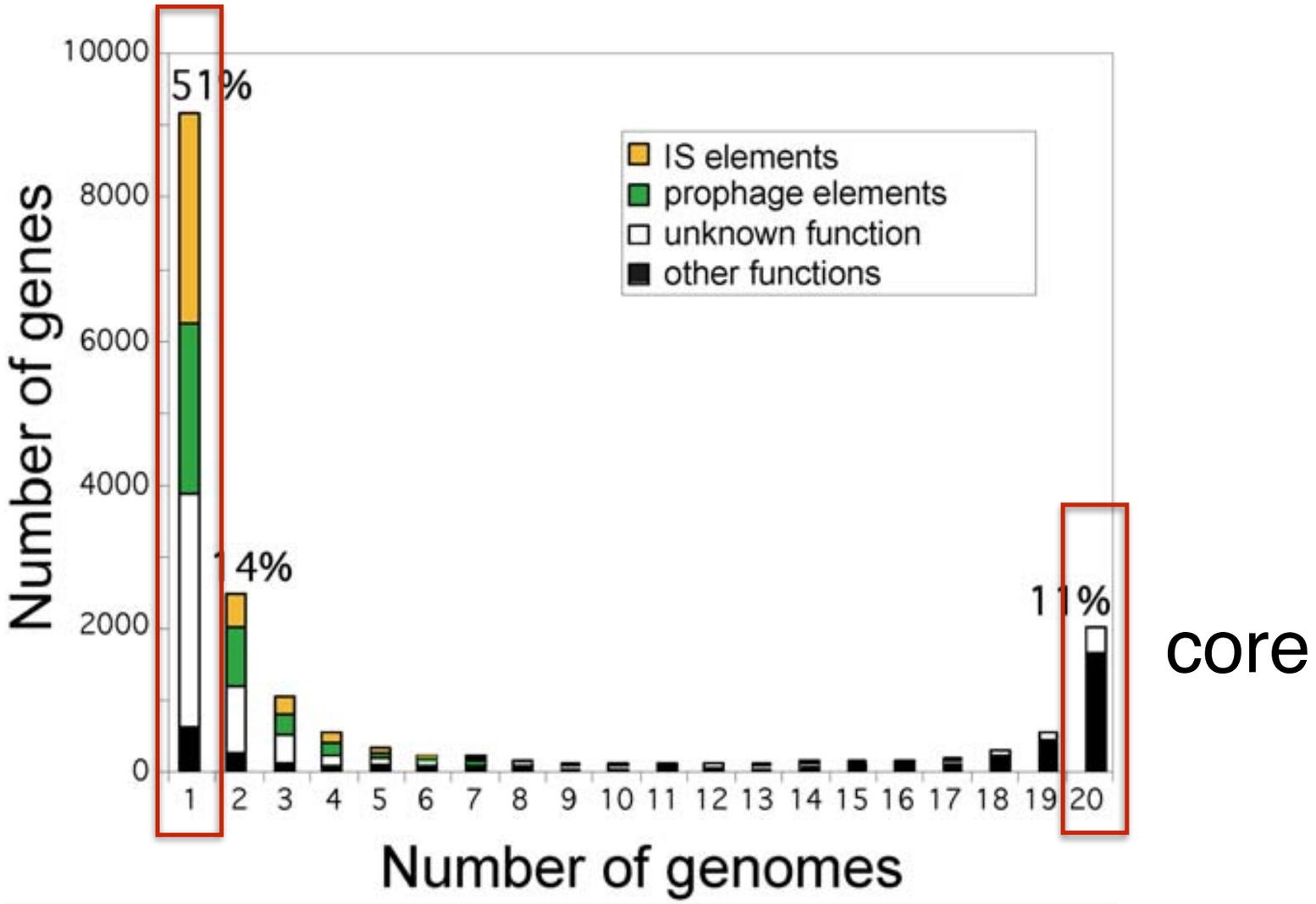
The determination of the pan-genome for a species depends on choice of strains and on the detection of homologs (orthologs and paralogs)


variable
variable
variable

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

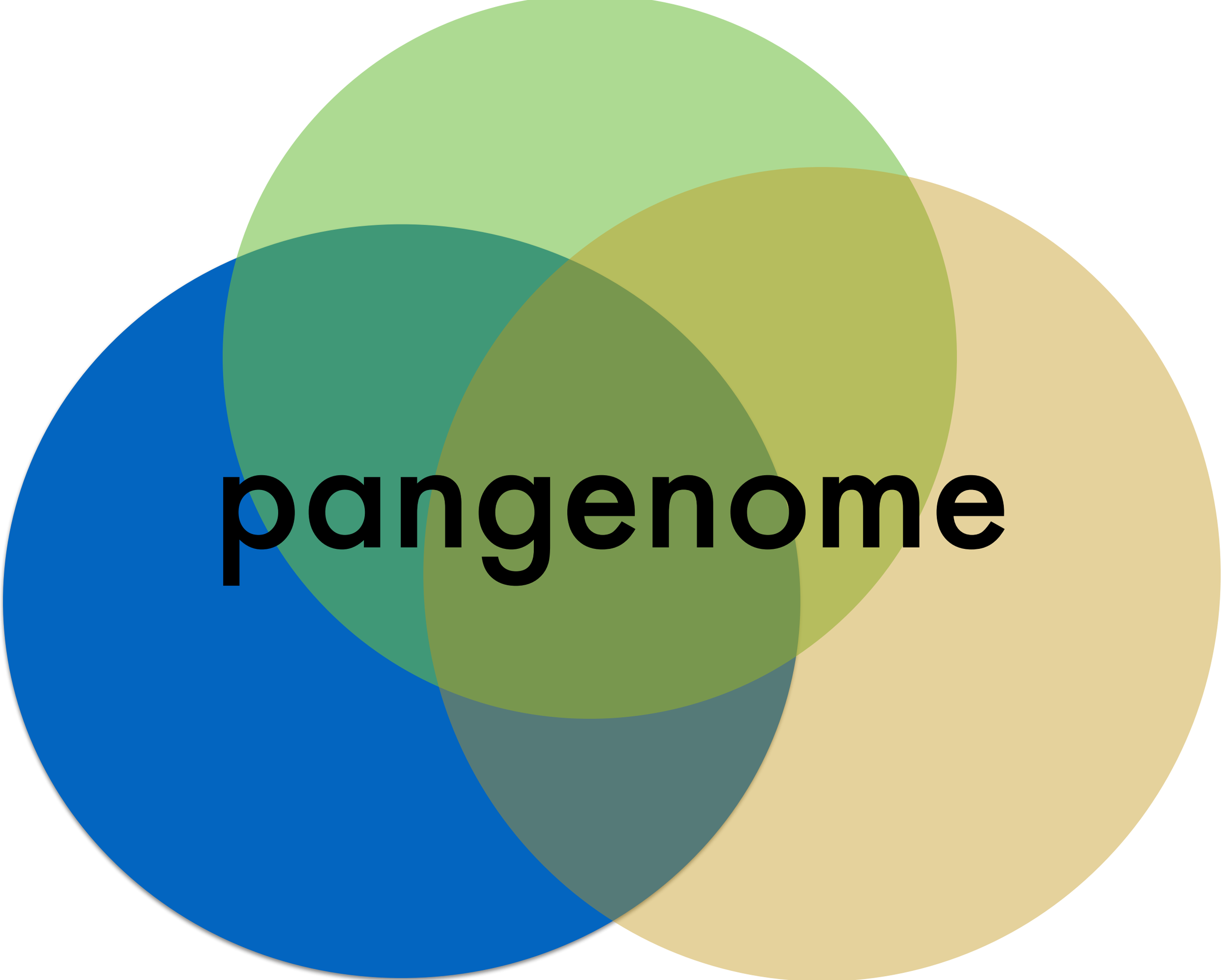
(tend to be hypothetical or unknown function)

strain-specific

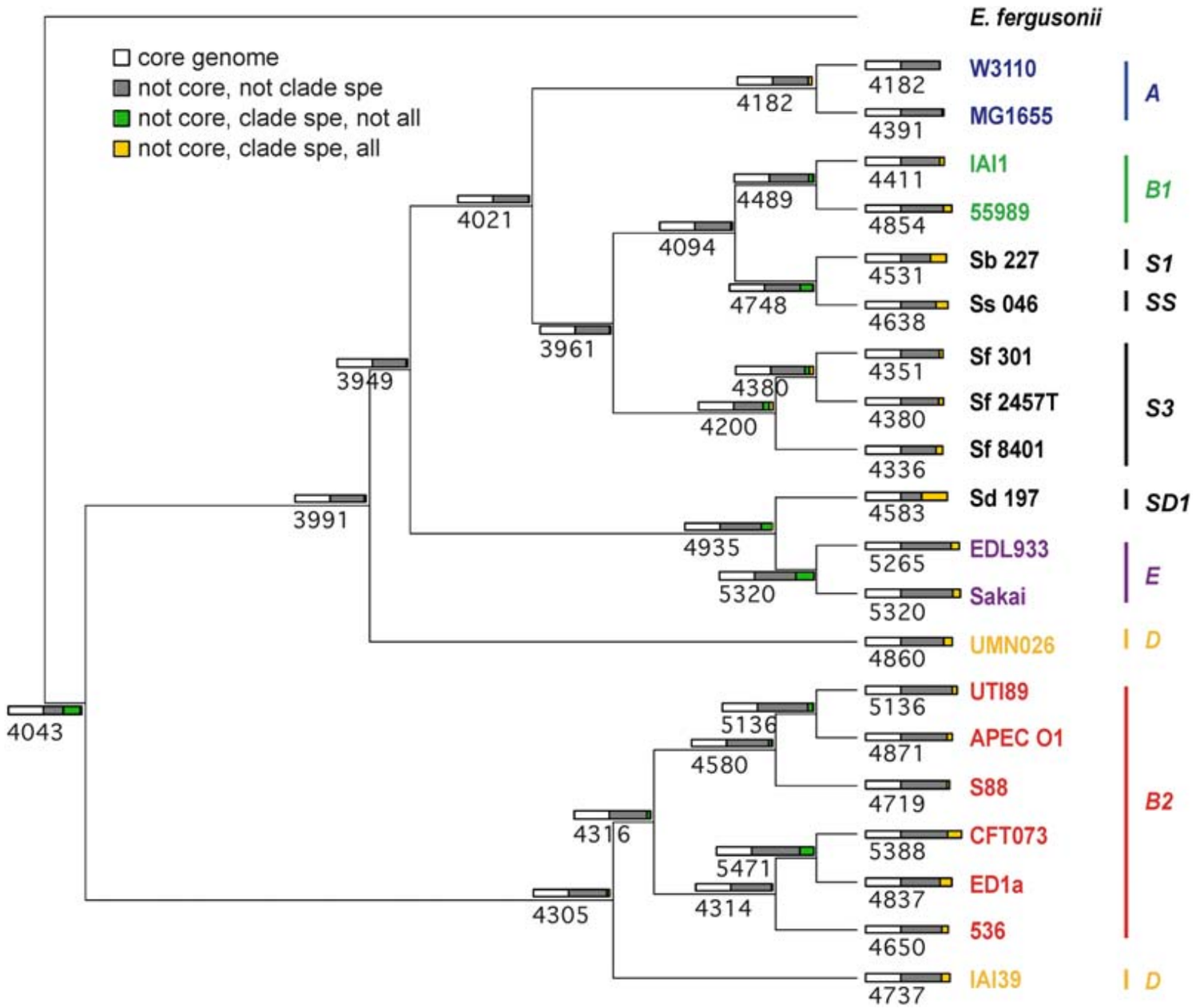


core

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



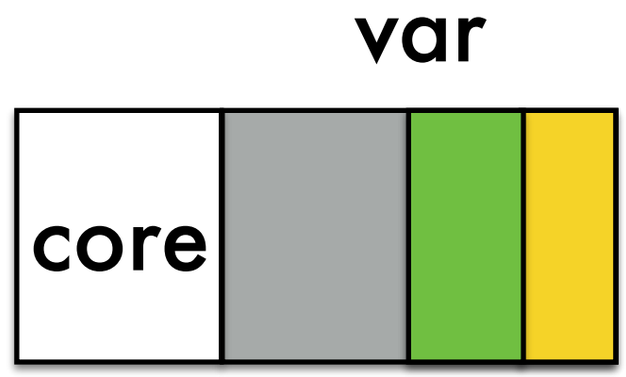
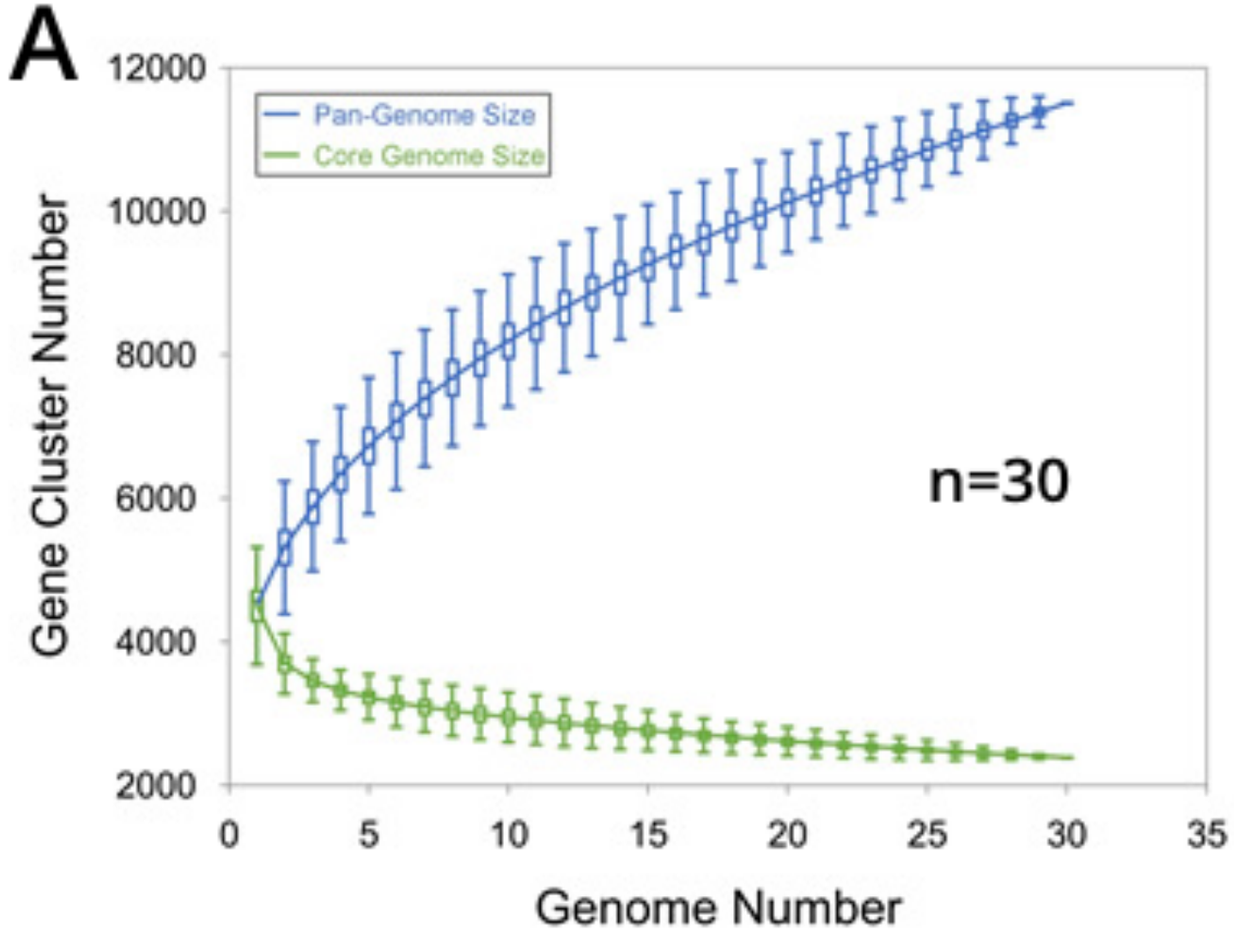
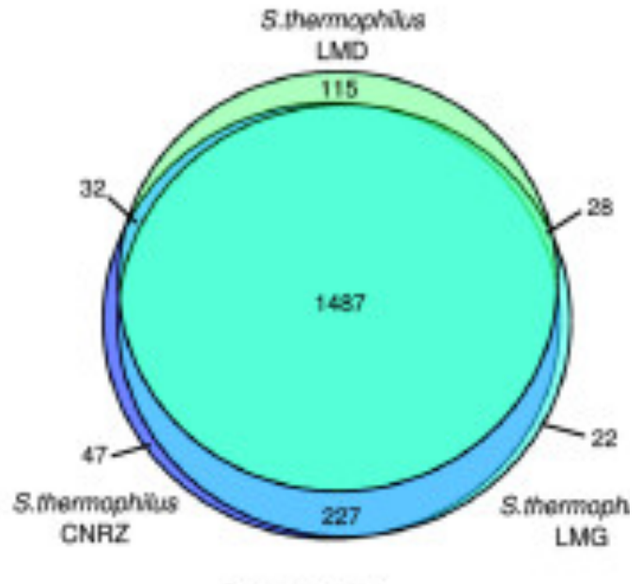


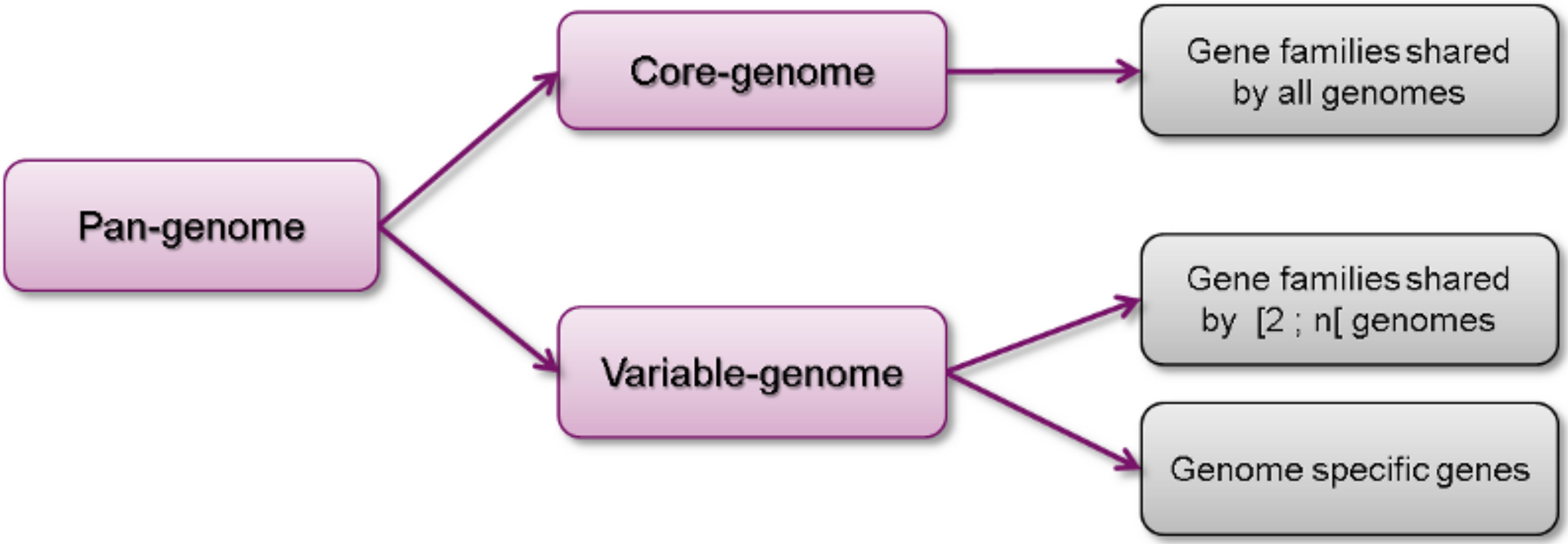
Figure 6. Inferred gene content evolution in the lineage of *Escherichia coli*. The cladogram shows the phylogenetic relationships among the 20 *E. coli/Shigella* genomes rooted on the *E. fergusonii* genome, as in Figure 4, but ignoring branch lengths. The major phylogenetic groups are indicated by 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

A

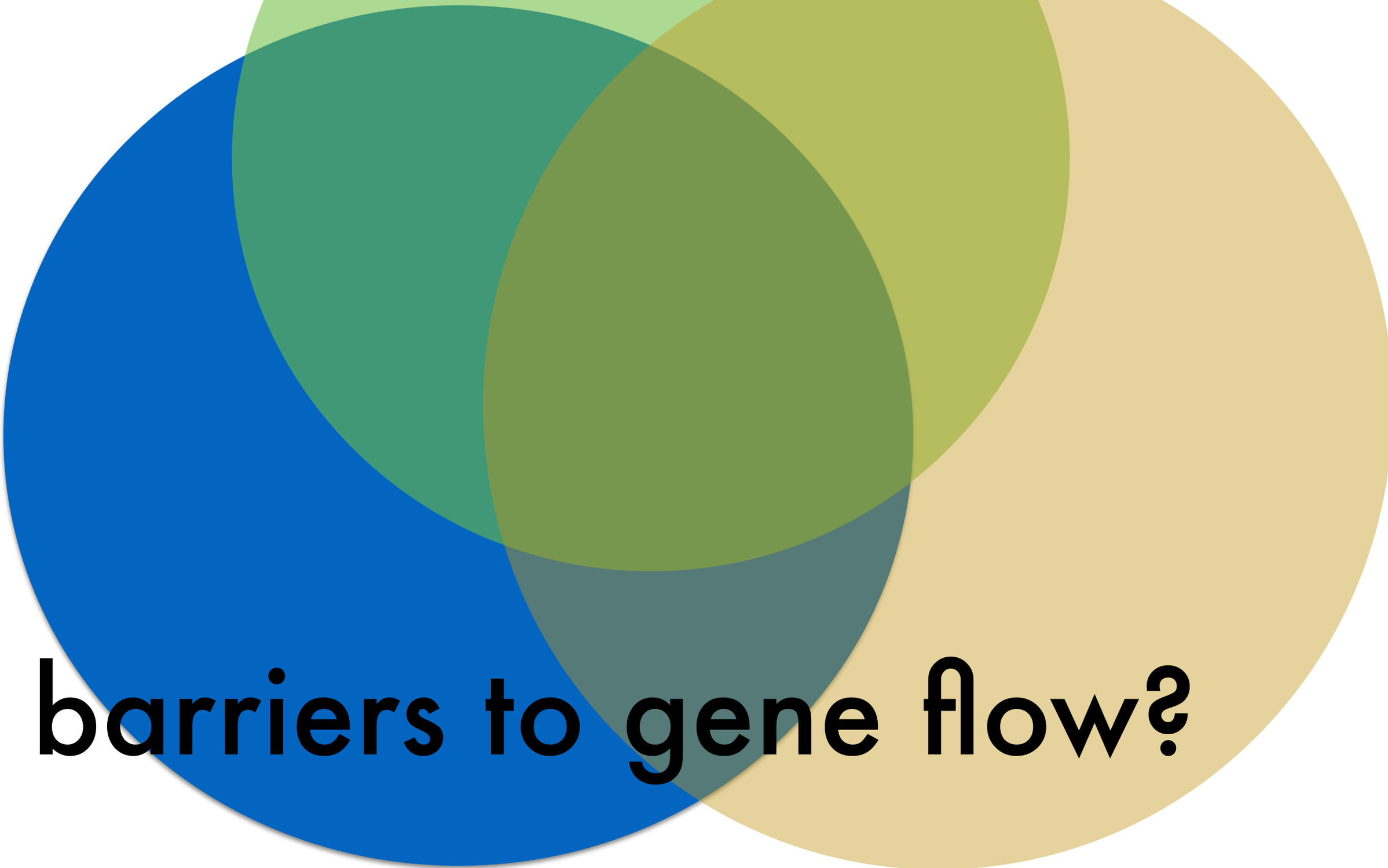

Pan-Genome Size
Core Genome Size
n=30
Gene Cluster Number
Genome Number
12000
10000
8000
6000
4000
2000
0
5
10
15
20
25
30
35

same species




Pan-genome
Core-genome
Variable-genome
Gene families shared by all genomes
Gene families shared by [2 ; n[genomes
Genome specific genes

Genomes are dynamic



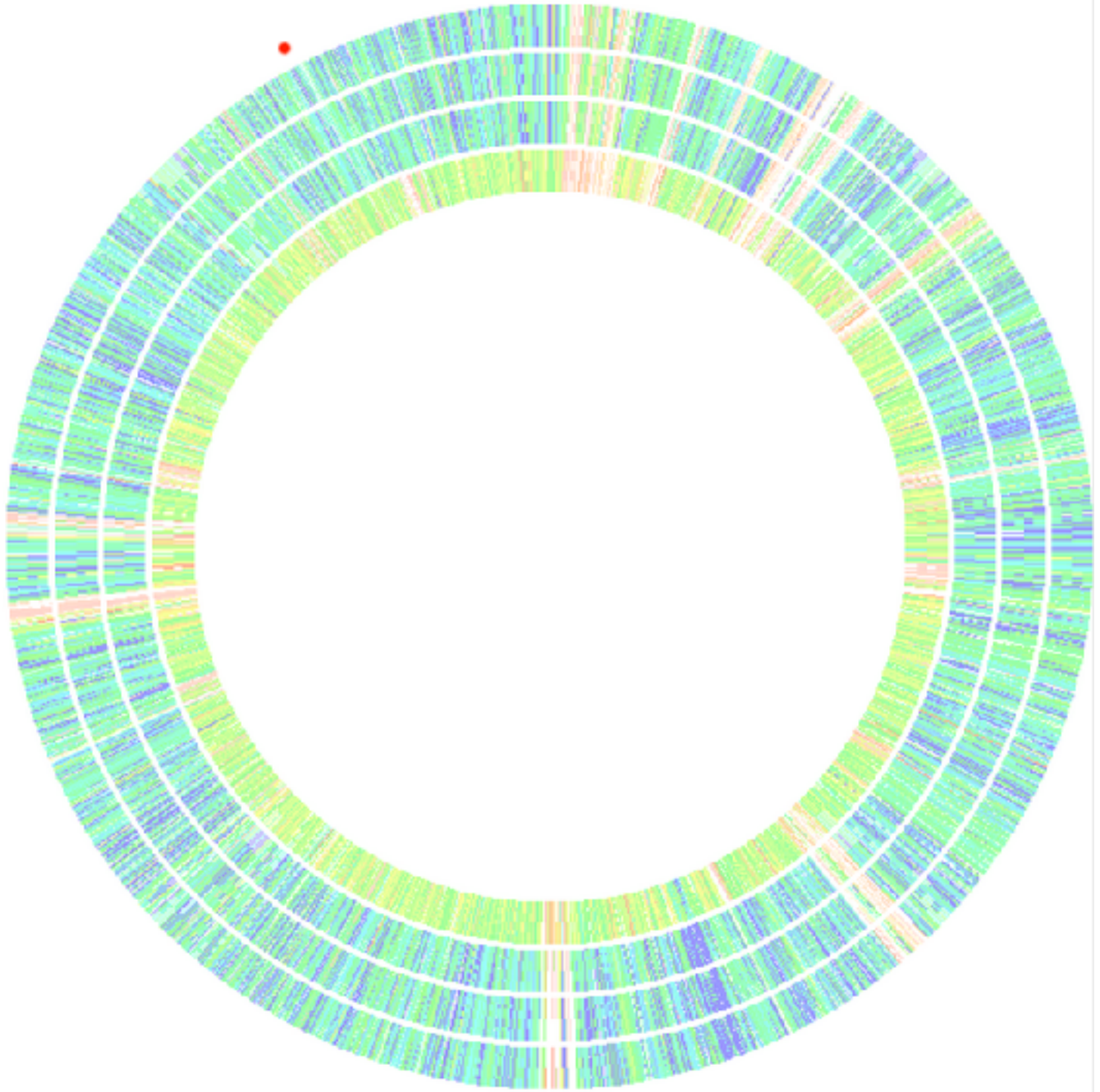
barriers to gene flow?

display 30 items per page

«first «prev displaying 4089 - 4118 of 4446 next» last»

percent identity 408.16	<	percent identity 408.18	<
percent identity 408.19	<	percent identity 408.17	<

419610.8			408.16			408.18			408.19			408.17		
Contig 1	Gene	Length	Hit	Contig all	Gene	Hit	Contig all	Gene	Hit	Contig all	Gene	Hit	Contig all	Gene
.	4110	535	bi	23	3273	bi	45	4256	bi	23	3313	bi	34	3176
.	4111	602	bi	23	3274	bi	45	4257	bi	23	3314	bi	34	3177
.	4112	414	bi	23	3275	bi	45	4258	bi	23	3315	bi	34	3178
.	4113	265	bi	23	3276	bi	45	4259	bi	23	3316	bi	34	3179
.	4114	287	bi	23	3277	bi	45	4260	bi	23	3317	bi	34	3180
.	4115	488	bi	23	3278	bi	45	4261	bi	23	3318	bi	34	3181
.	4116	339	bi	23	3279	bi	45	4262	bi	23	3319	bi	34	3182
.	4117	622	bi	23	3280	bi	45	4263	bi	23	3320	bi	34	3183
.	4118	620	bi	23	3282	bi	45	4265	bi	23	3322	bi	34	3185
.	4119	692	bi	23	3283	bi	45	4266	bi	23	3323	bi	34	3186
.	4120	942	bi	23	3284	bi	45	4267	bi	23	3324	bi	34	3187
.	4121	215	bi	23	3285	bi	45	4268	bi	23	3325	bi	34	3188
.	4122	500	bi	23	3287	bi	45	4270	bi	23	3327	bi	34	3190
.	4123	293	bi	23	3289	bi	45	4272	bi	23	3329	bi	34	3192
.	4124	412	bi	23	3290	bi	45	4273	bi	23	3330	uni	41	4687
.	4125	1070	bi	23	3291	bi	45	4274	bi	23	3331	bi	34	3193
.	4126	382	bi	23	3292	bi	45	4275	bi	23	3332	bi	34	3194
.	4127	114	bi	23	3293	bi	45	4276	bi	23	3333	bi	34	3196
.	4128	457	bi	23	3294	bi	45	4277	bi	23	3334	bi	34	3197
.	4129	209	bi	23	3295	bi	45	4278	bi	23	3335	bi	34	3198
.	4130	212	bi	23	3296	bi	45	4279	bi	23	3336	bi	34	3199
.	4131	244	bi	23	3298	bi	45	4281	bi	23	3338	bi	34	3200
.	4132	497	-			-			-			bi	12	1819
.	4133	86	-			-			-			-		
.	4134	320	bi	23	3300	bi	45	4283	bi	23	3340	bi	34	3274
.	4135	123	bi	23	3301	bi	45	4284	bi	23	3341	bi	34	3275
.	4136	1067	bi	23	3302	bi	45	4285	bi	23	3342	bi	34	3276
.	4137	525	-			-			-			-		
.	4138	174	bi	23	3305	bi	45	4288	bi	23	3345	bi	34	3280
.	4139	639	bi	23	3306	bi	45	4289	bi	23	3346	bi	34	3281

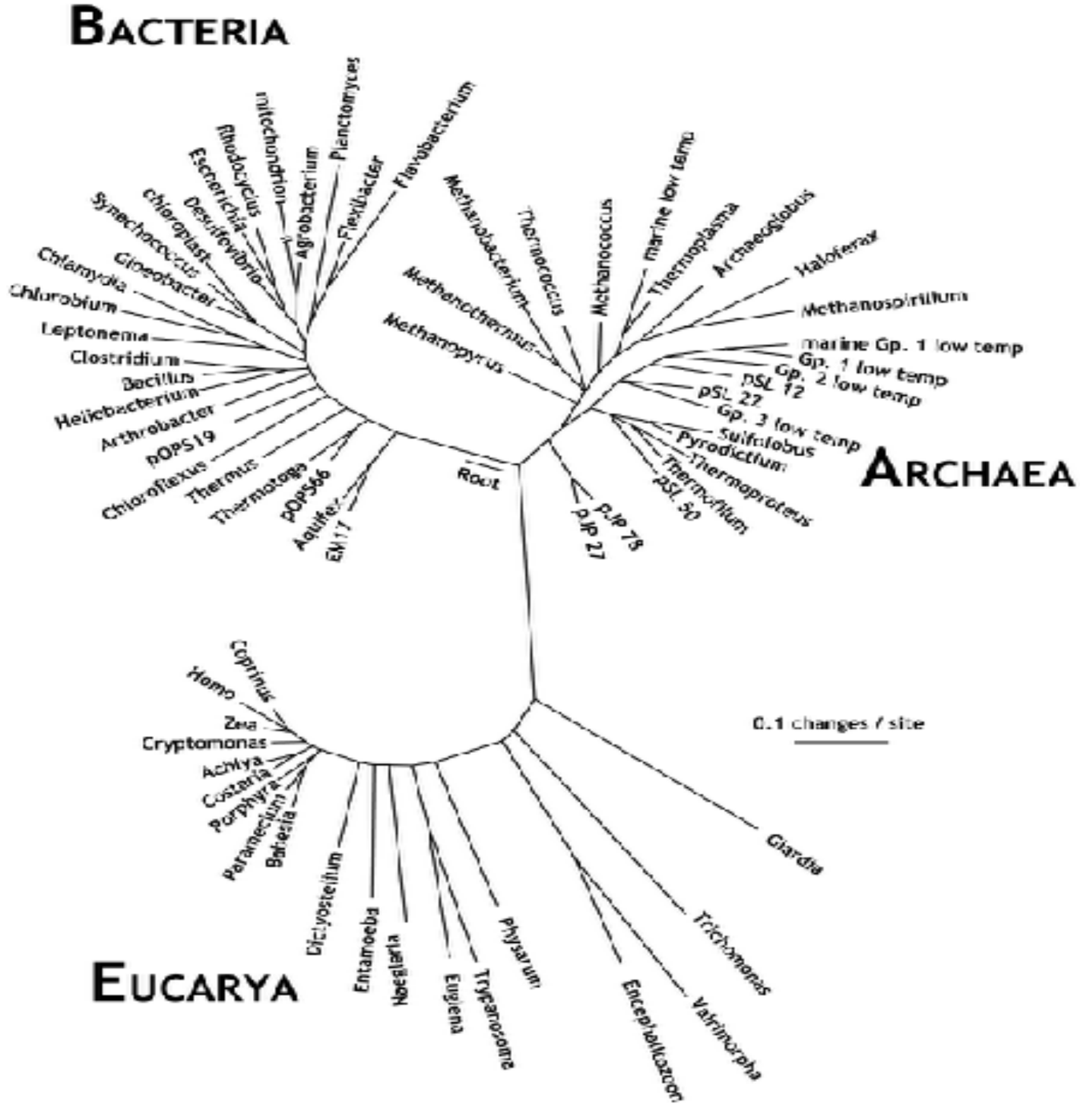


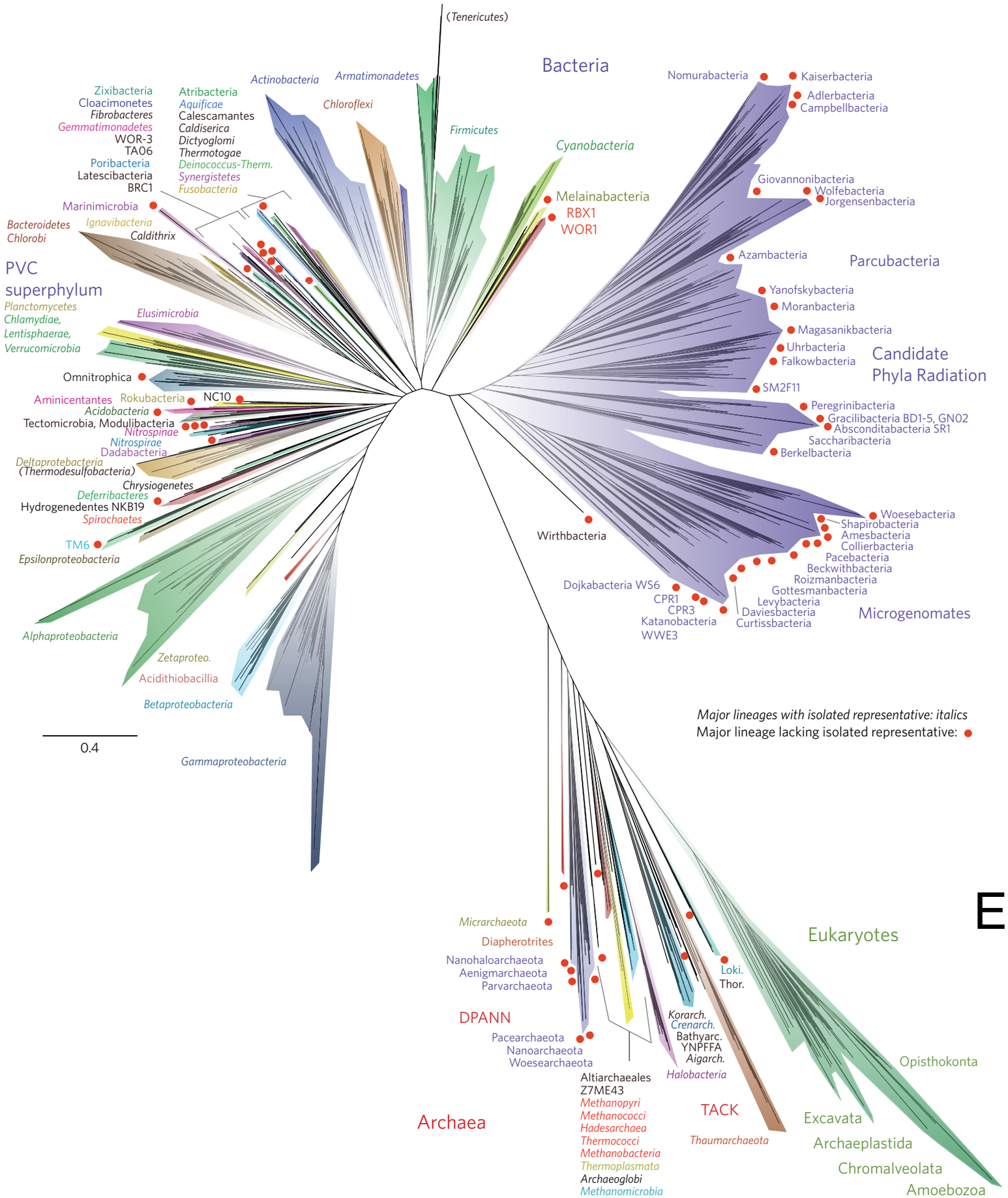
darker colors = higher % similarity

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

-Scott Dawson (2015)

Is this the best map of life?

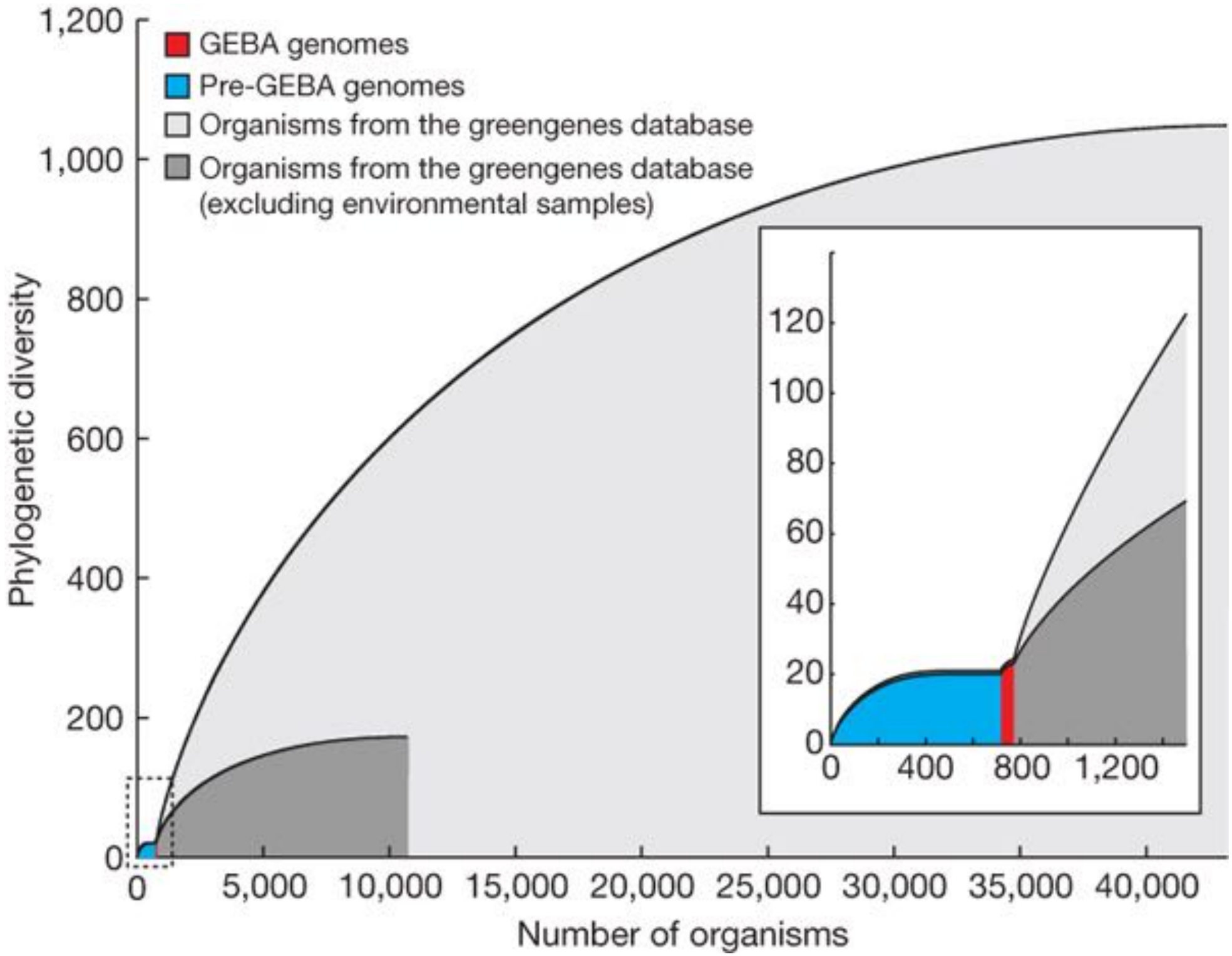




What's the scale?

Euks are archaea?

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



DY Wu *et al. Nature* **462**, 1056-1060 (2009) doi:10.1038/nature08656

https://www.arb-silva.de

silva
high quality ribosomal RNA databases

SILVA SSU / LSU 123 - full release

	SSU Parc	SSU Ref	SSU Ref NR	L
Minimal length	300	1200/900	1200/900	3
Quality filtering	basic	strong	strong	b
Guide Tree	no	no	yes	r
Release date	23.07.15	23.07.15	23.07.15	2
Aligned rRNA sequences	**4,985,791**	1,756,783	597,607	5

http://rdp.cme.msu.edu

https://rdp.cme.msu.edu

ABOUT RDP | ASSIGNMENT GENERATOR | CITATION | CONTACTS | RELATED SITES | RESOURCES | TUTORIALS | USER WIKI

my rdp login

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.

Hierarchy Browser | Classifier | Probe Match | FunGene
MbcS GoogleSheets | Library Compare | Sequence Match | RDPipeline
Aligner | Tree Builder | RDP Open Source | Tutorials

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

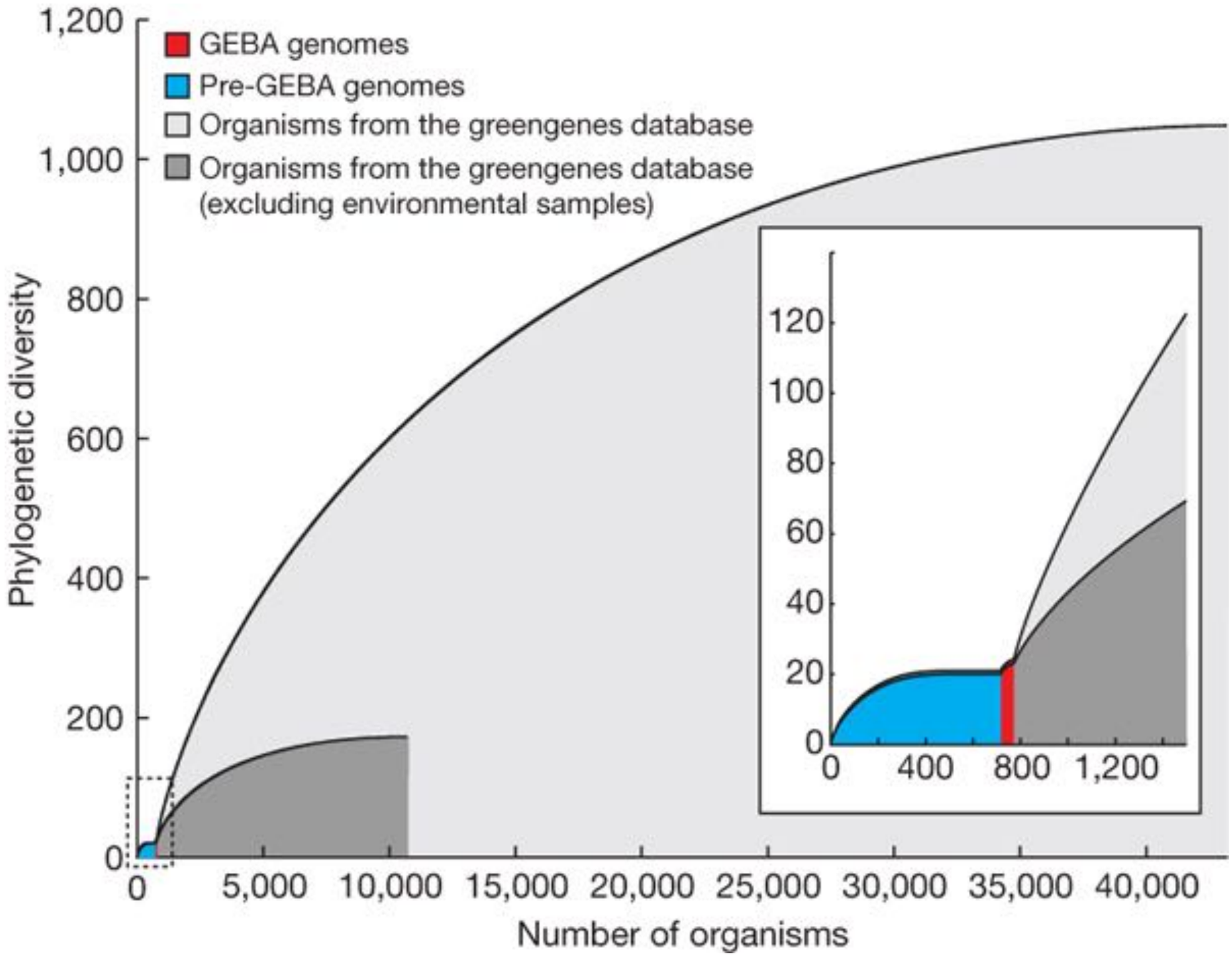
05/28/2015 RDP Staff attending ASM Meeting in New Orleans
RDP staff will be attending the ASM General Meeting in New Orleans in the coming week. Two RDP posters will be presented: first on Tuesday morning:...

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.

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



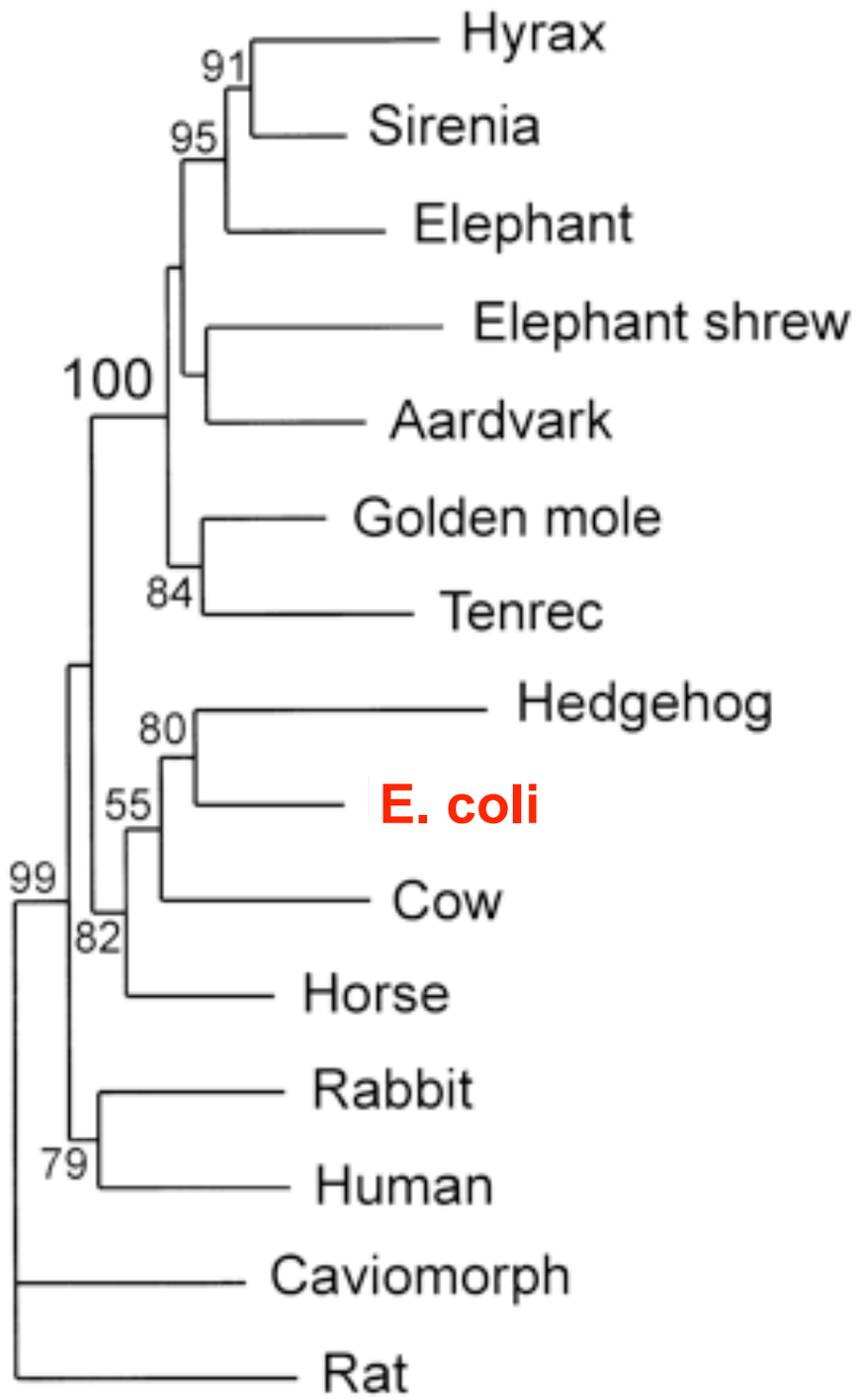
DY Wu *et al. Nature* **462**, 1056-1060 (2009) doi:10.1038/nature08656

It's all about the bass.

-Megan Trainor (2014)

Horizontal Gene Transfer?

Xenologs: Horizontal gene transfer



Horizontal gene transfer: genes acquired not through common "vertical" ancestry

*

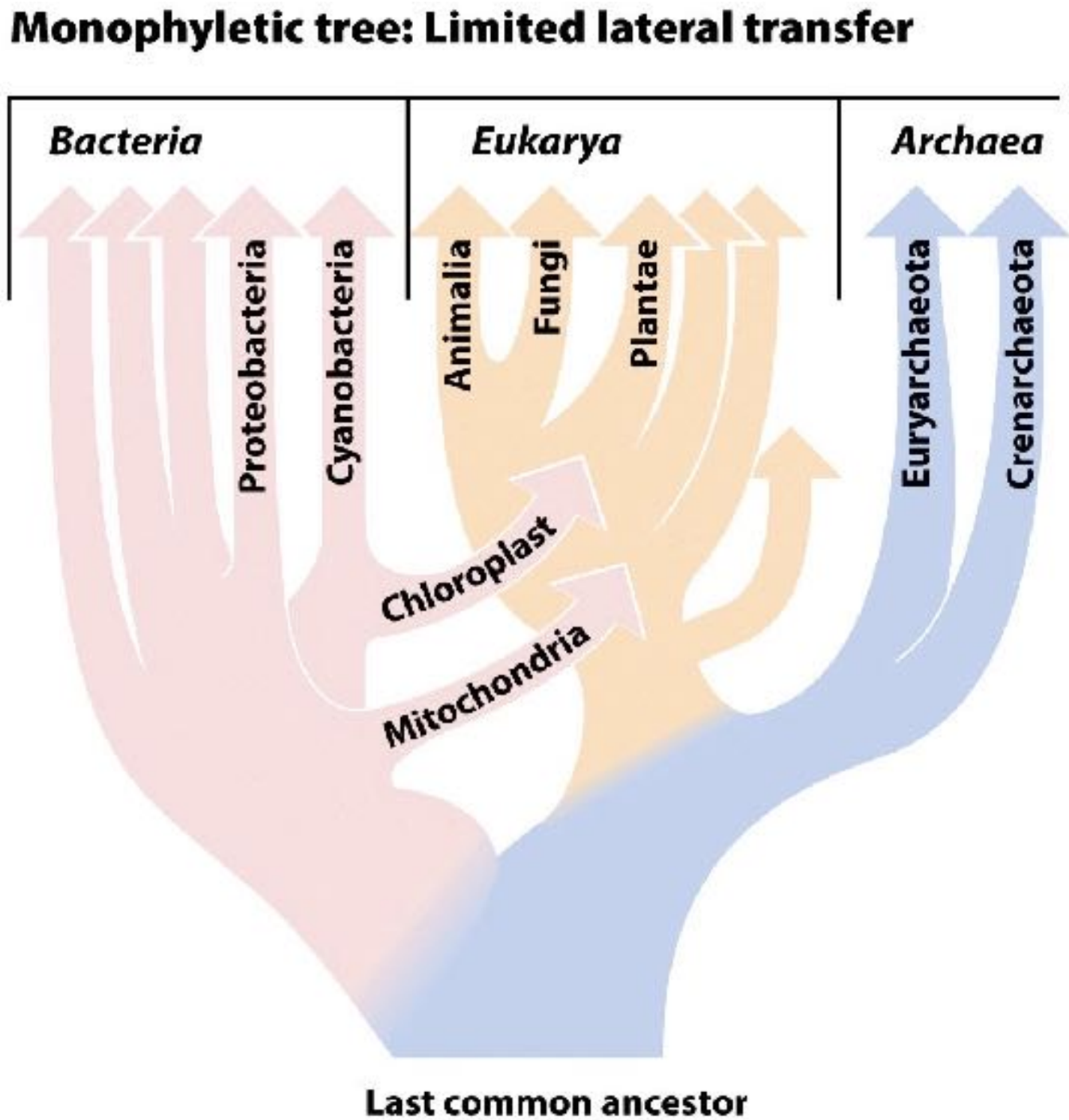


Figure 17.27a Microbiology: An Evolving Science
Source: W. Ford Doolittle. 1999. *Science* 284:2126.

Alpha proteobacterial genes in eukaryotes?
Cyanobacterial genes in eukaryotes?