

Tree of Life, Forest of Life, Web of Life, Mess of Life – What is it after all?

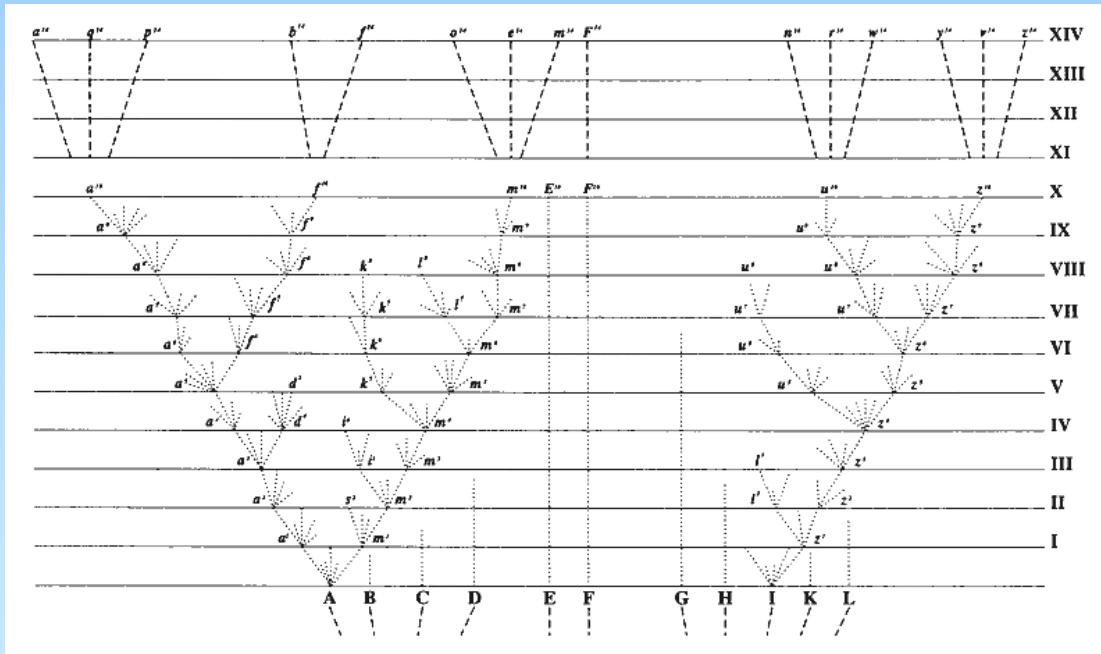
Eugene V. Koonin

National Center for Biotechnology Information, NLM, NIH

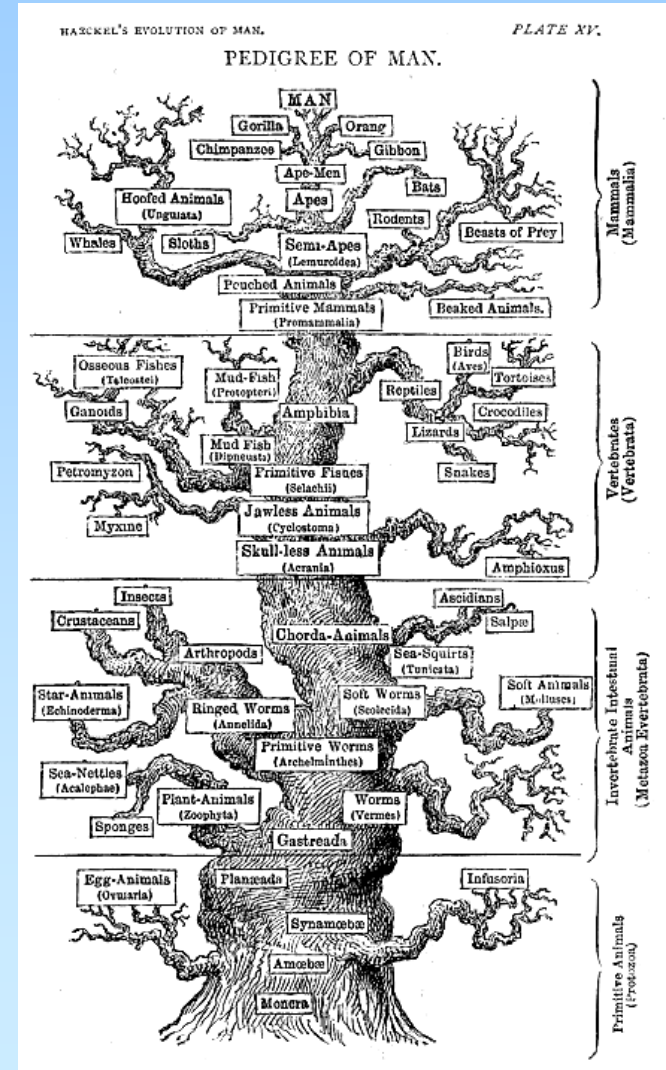
02/07/2011, UCSB

A brief history of TOL

Thinking of the history of life in terms of phylogenetic trees is as old as scientific biology (if not older)



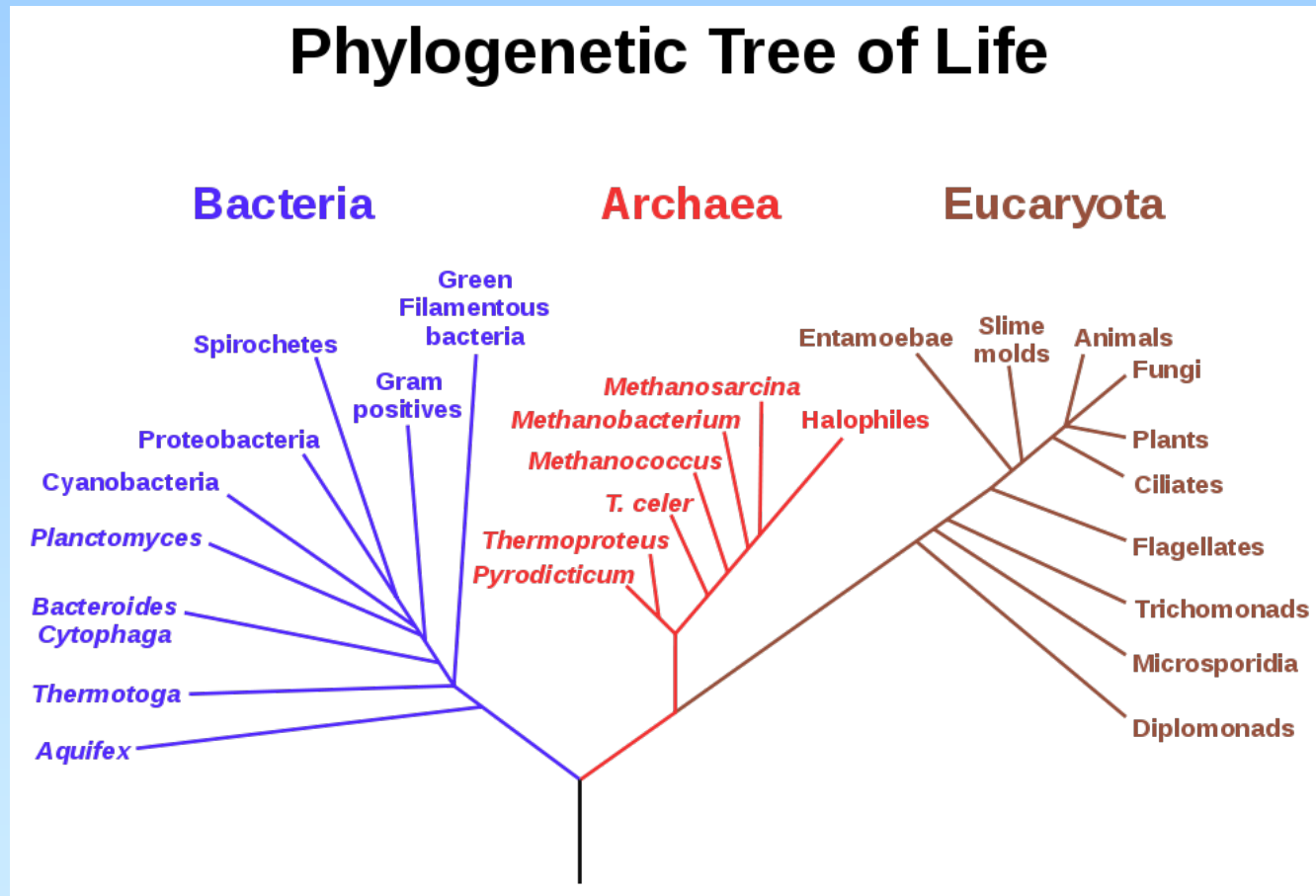
Charles Darwin (1859) *Origin of Species* [one and only illustration]: "descent with modification"



Ernst Haeckel (1879)
The Evolution of Man

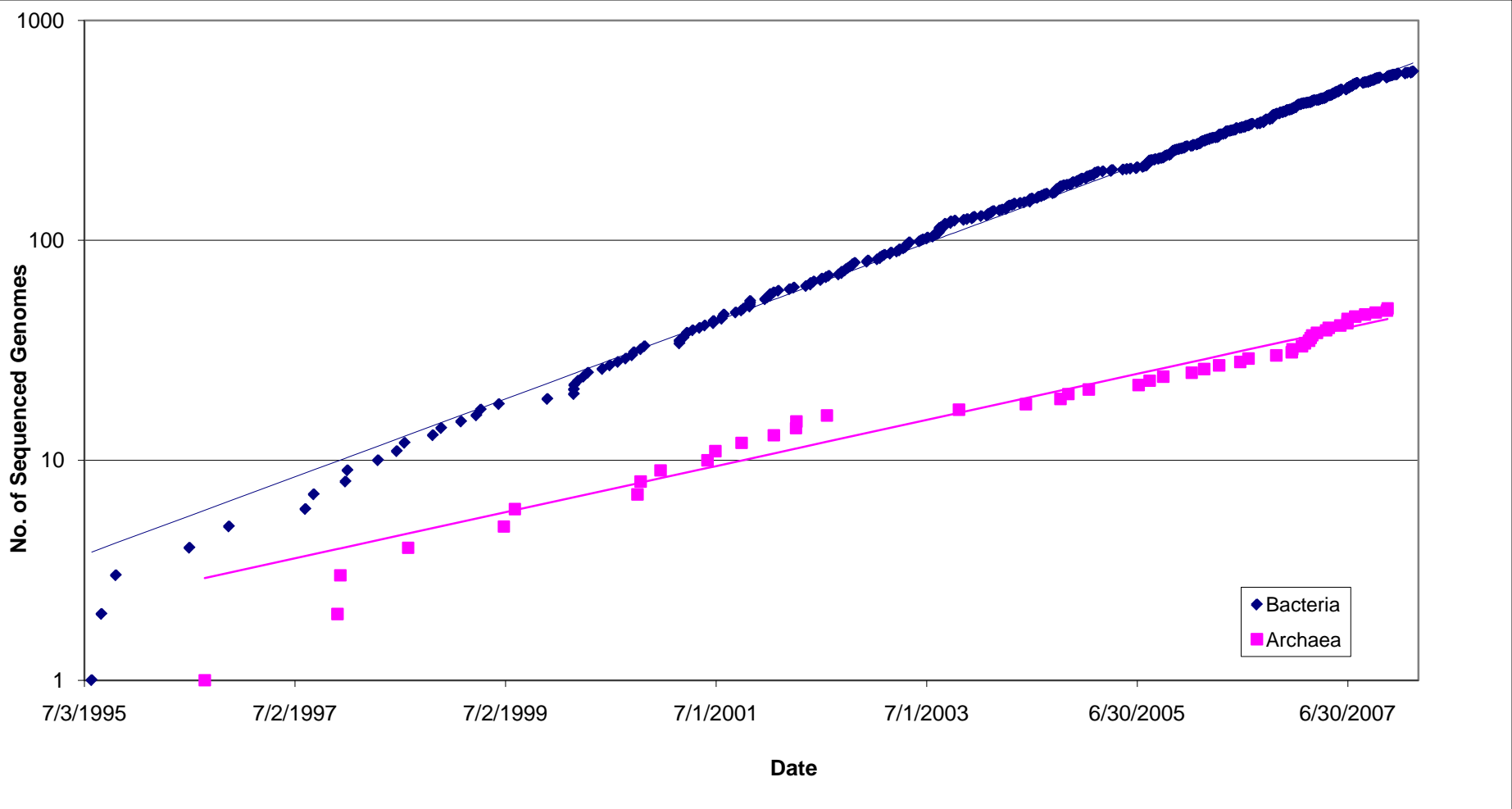
A brief history of TOL

Advent of molecular phylogenetics – expectations of objectively reconstructed complete Tree of Life.



Woese et al. (1990) *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.* PNAS 87, 4576-4579 [Figure 1, modified]

Exponential accumulation of prokaryotic genome sequences

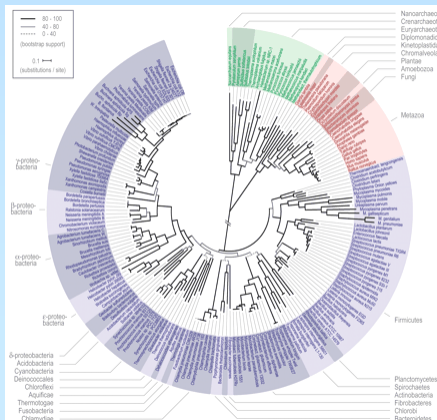


A brief history of TOL

Genomic era – growing frustration with discrepancies between the trees reconstructed for individual genes and heroic efforts to overcome the noise. Role of horizontal gene transfer in the evolution of prokaryotic genomes is established.

Major approaches:

- gene repertoire and gene order
- distribution of distances between orthologs
- concatenated alignments of "non-transferable" gene cores
- consensus trees and supertrees



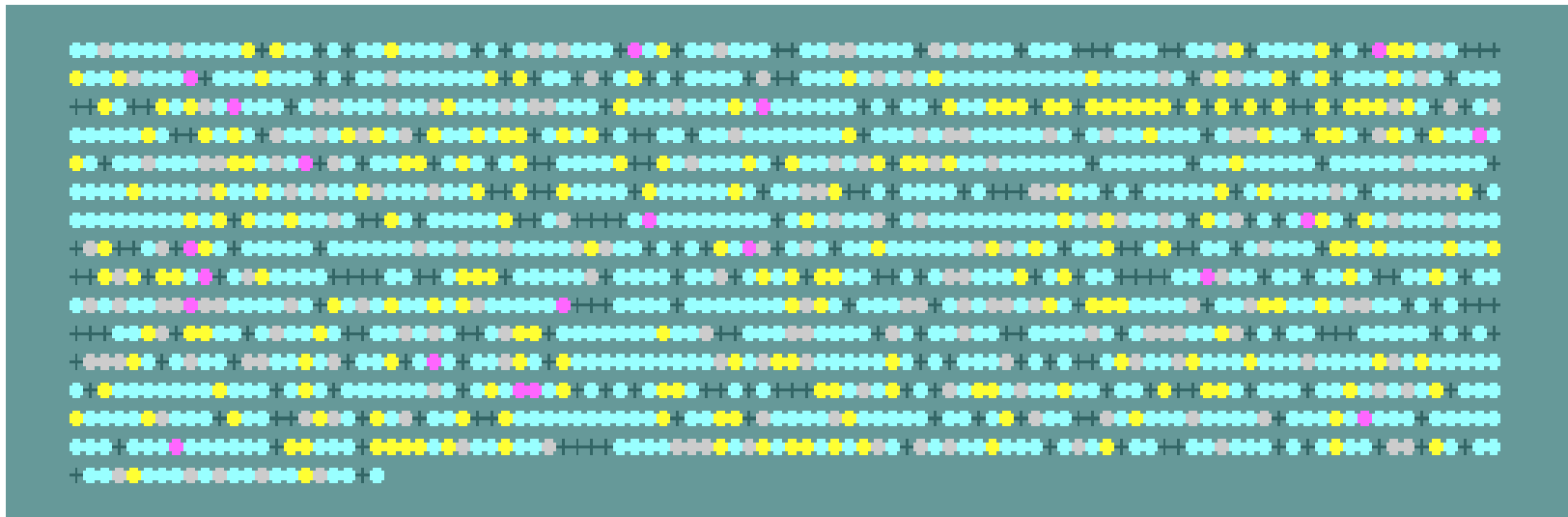
Ciccarelli et al. (2006) *Towards automatic reconstruction of a highly resolved tree of life.* Science 311, 1283-1287 [Figure 2]

Apparent massive horizontal gene transfer (HGT) from archaea to a hyperthermophilic bacterium

Aquifex aeolicus TaxTable - Netscape

File Edit View Go Communicator Help

1522 *Aquifex aeolicus* proteins: taxonomic distribution of the homologs



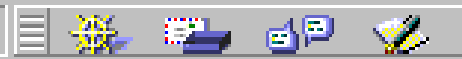
Cut-Off:

Cut-Off+:

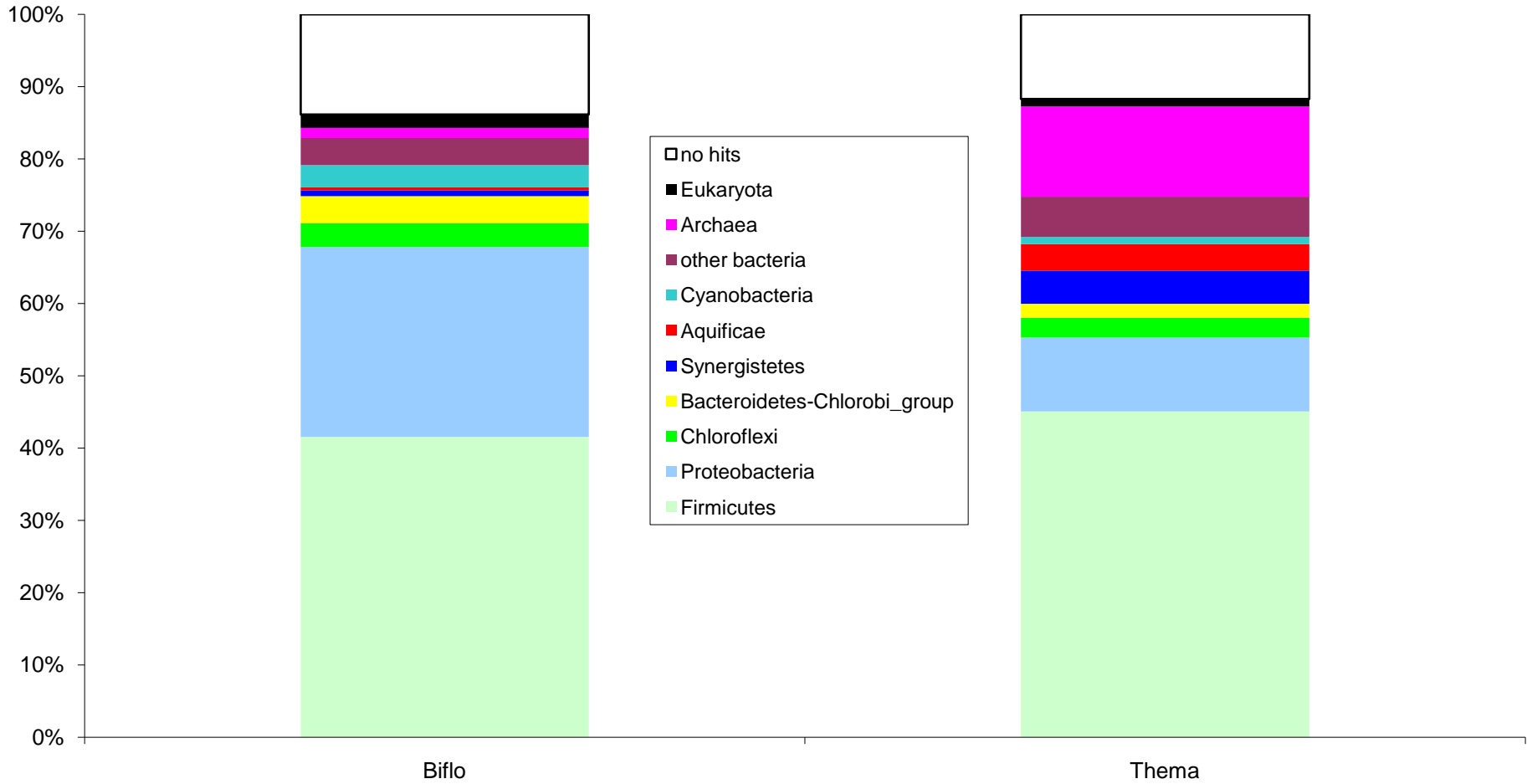
Redraw

HELP

Document: Done

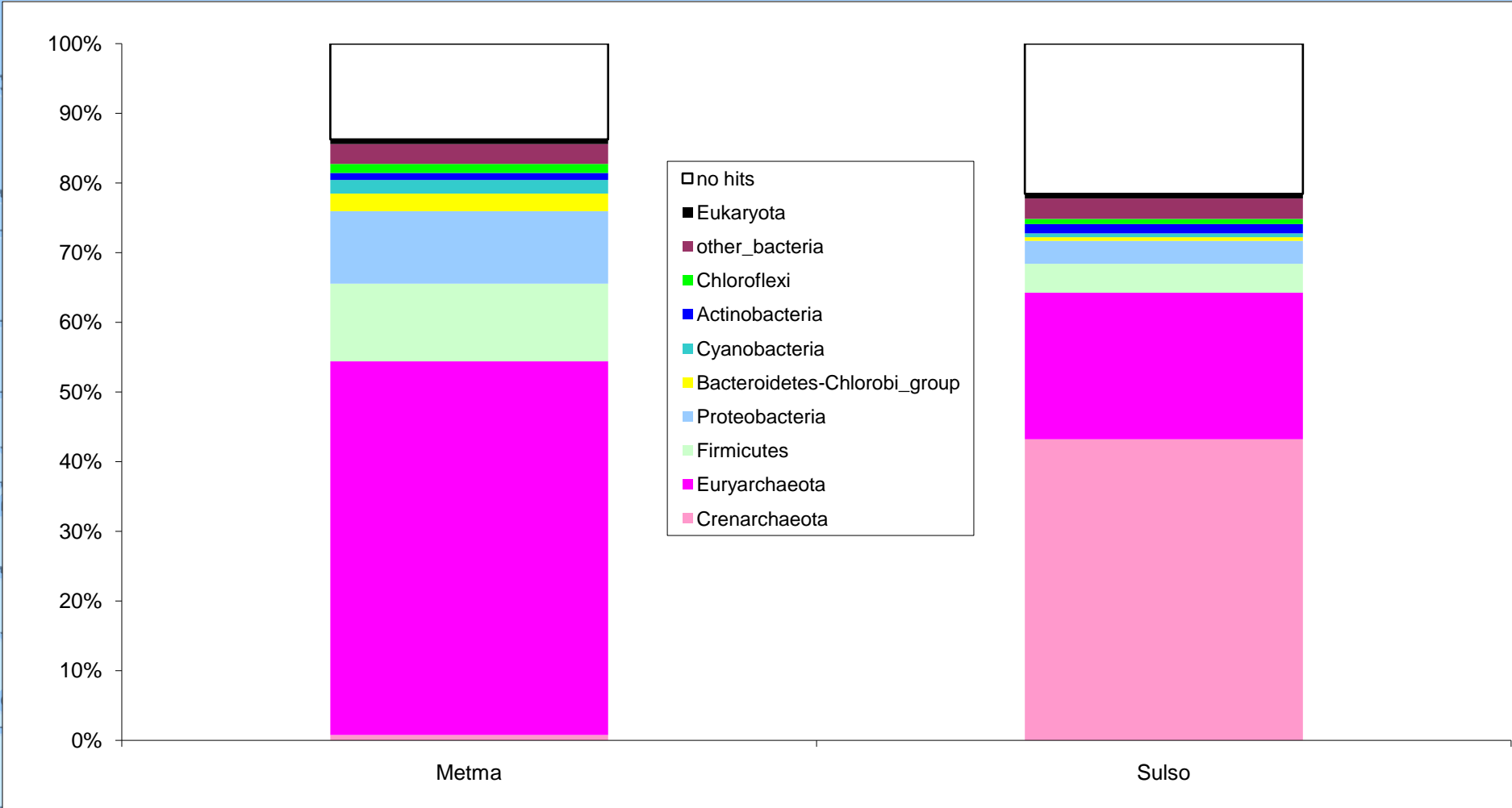


Differences in gene provenance between mesophilic and thermophilic bacteria: Telltale sign of HGT

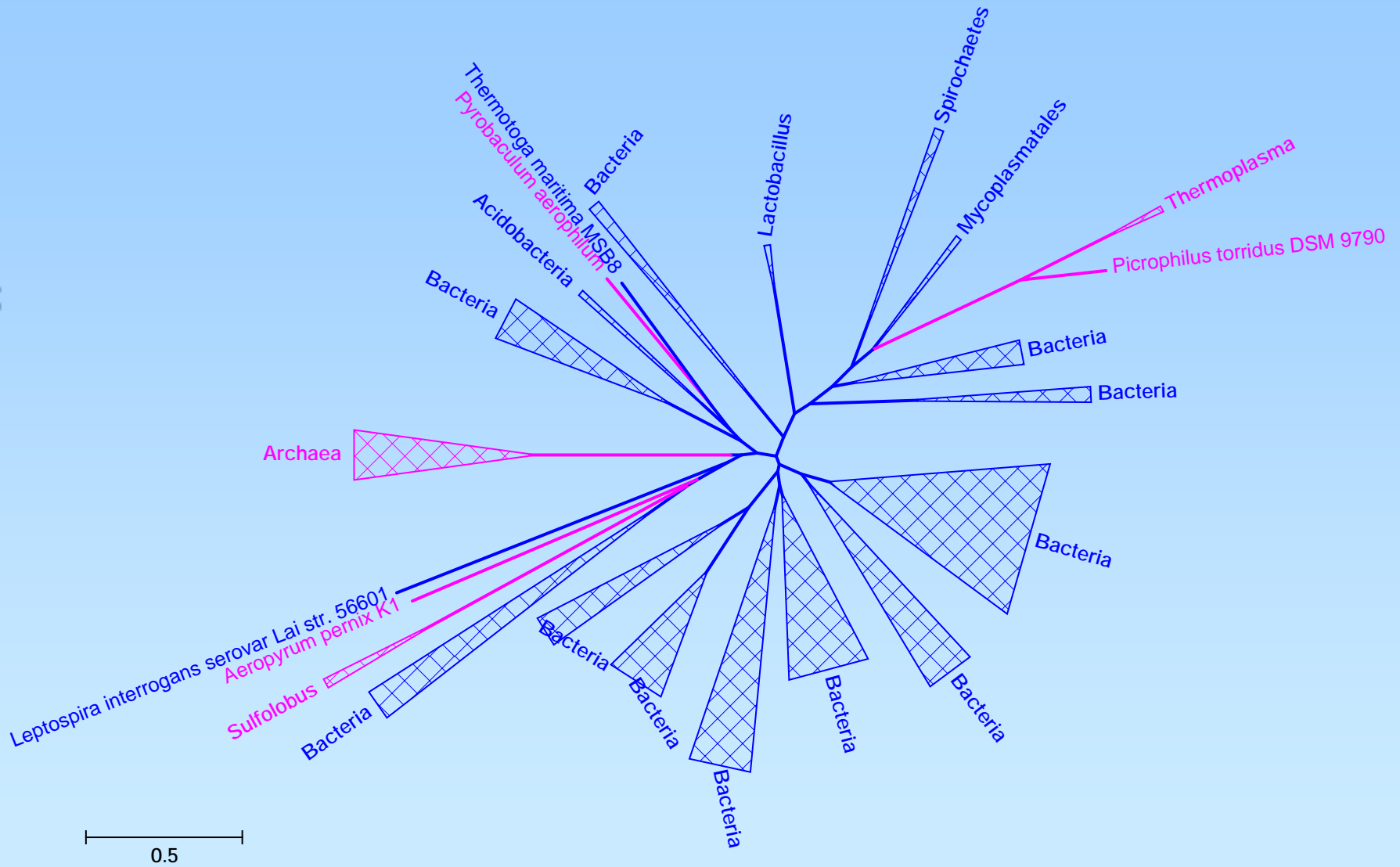


Nation

Differences in gene provenance between mesophilic and thermophilic archaea: Telltale sign of HGT

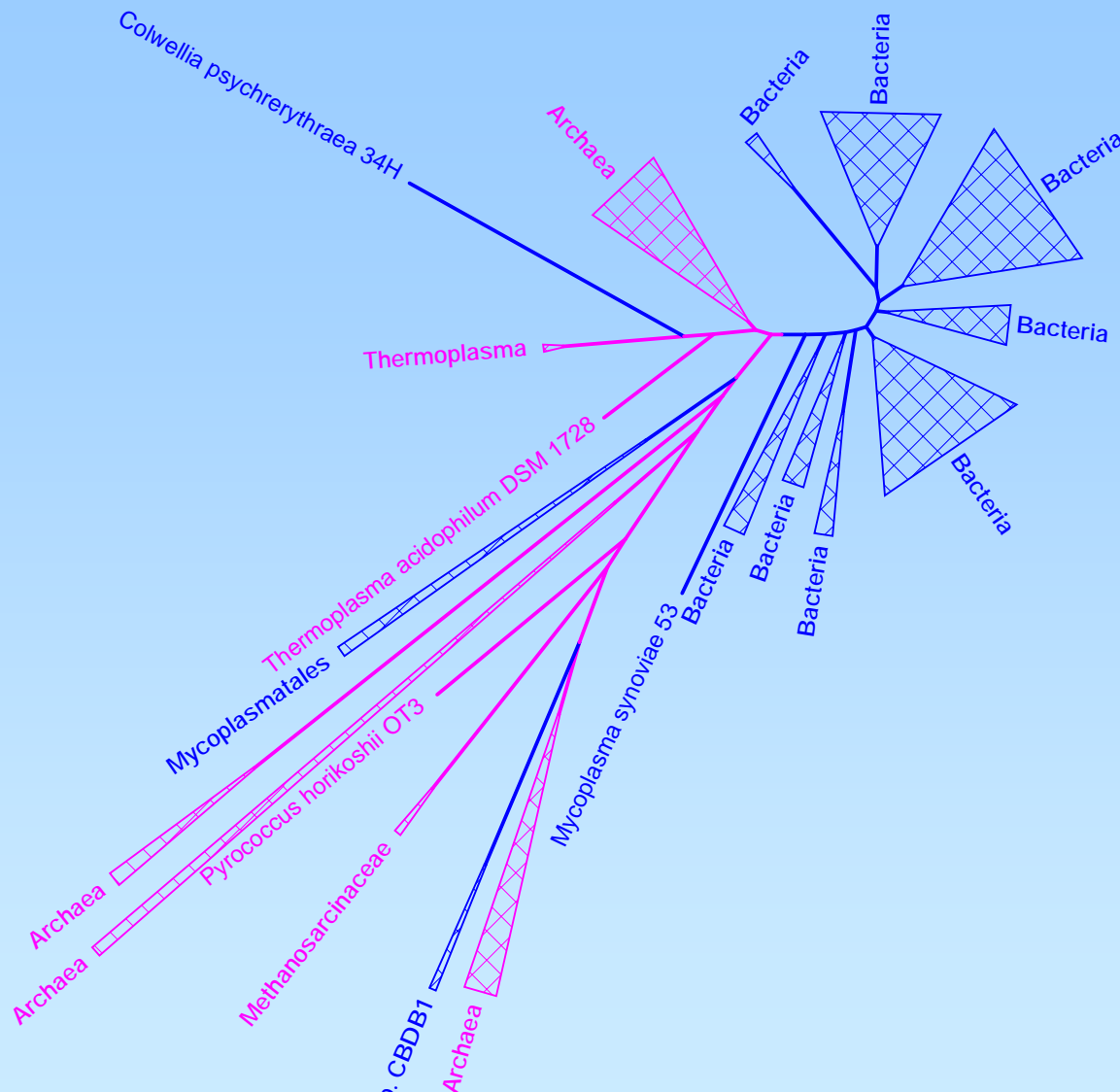


COG0030, dimethyladenosine transferase, an enzyme involved in rRNA methylation



Evolutionary history is gene-specific: different genes generally yield trees with different topologies

FtsZ, a GTPase involved in cell division

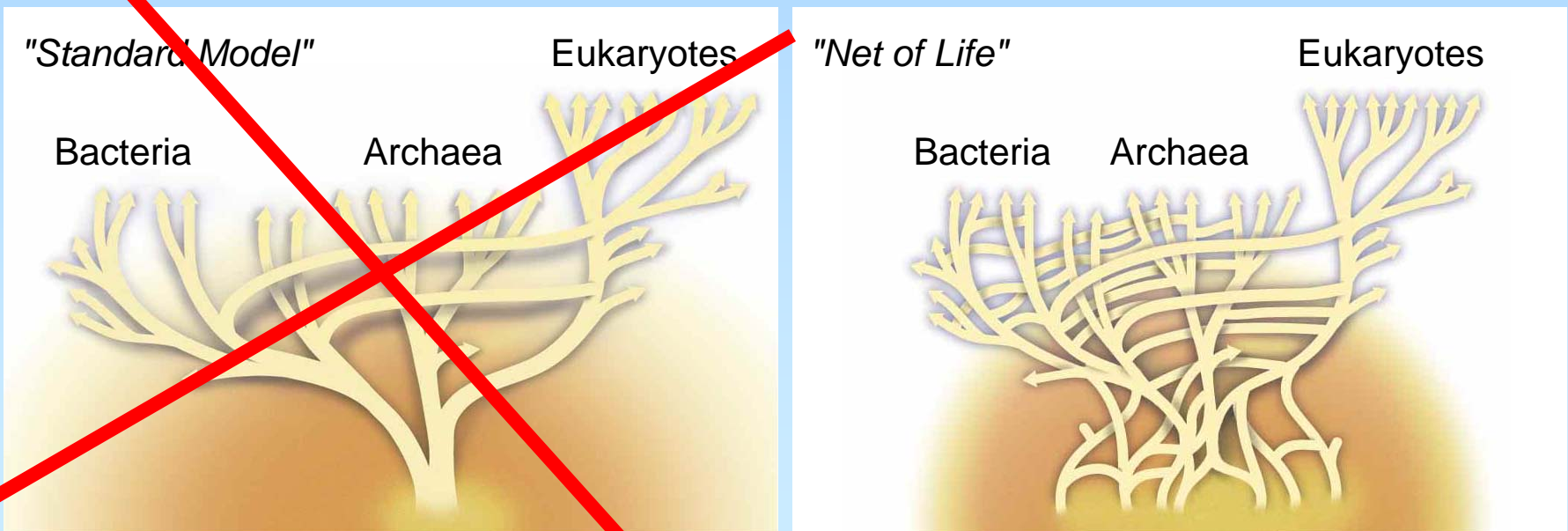


Evolutionary history is gene-specific: different genes generally yield trees with different topologies

A brief history of TOL

Troubled times – "uprooting" the TOL for prokaryotes (which comprise the great majority of cells on earth)

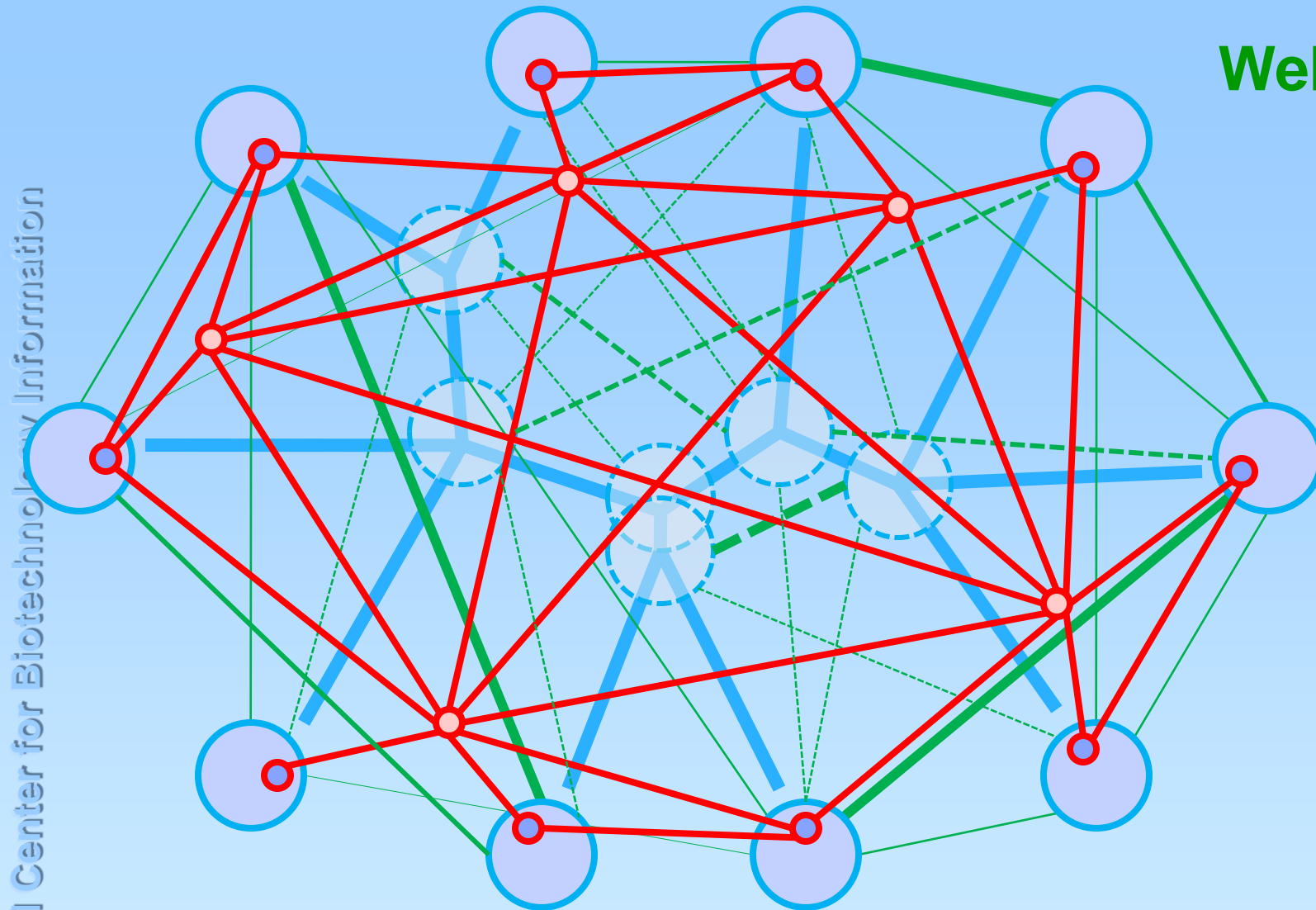
- horizontal gene transfer is rampant; no gene is exempt
- histories of individual genes are in general different
- tree-like signal is completely lost (or never existed at all)
- there are no species (or other taxa) in prokaryotes
- whatever consistent tree signal is observed, **is created by biases in HGT**

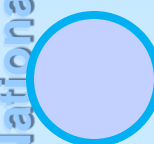




Doolittle WF. (2000) *Uprooting the tree of life*. Sci. Am. 282, 90-95 [modified]




Web of Life

National Center for Biotechnology Information

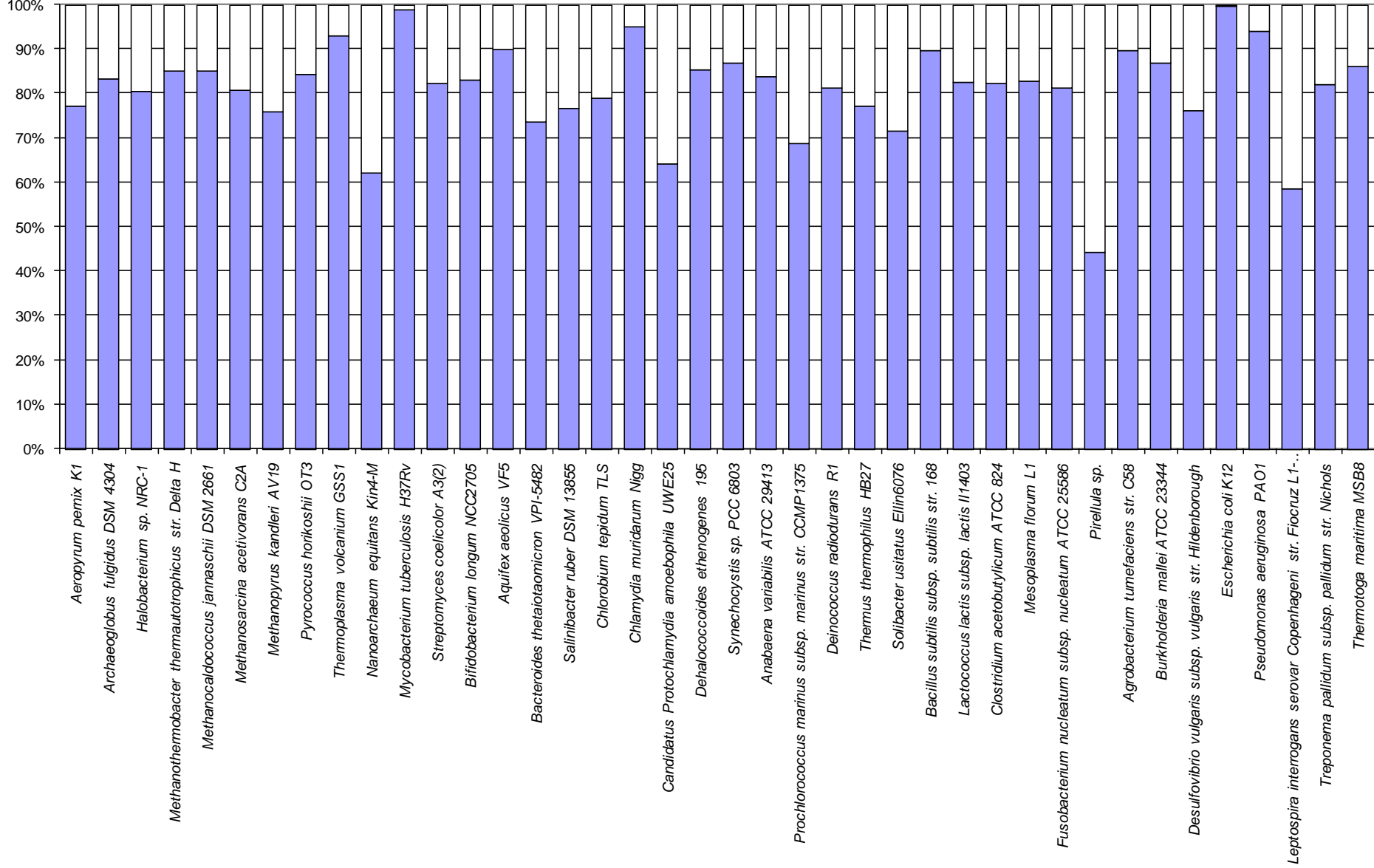


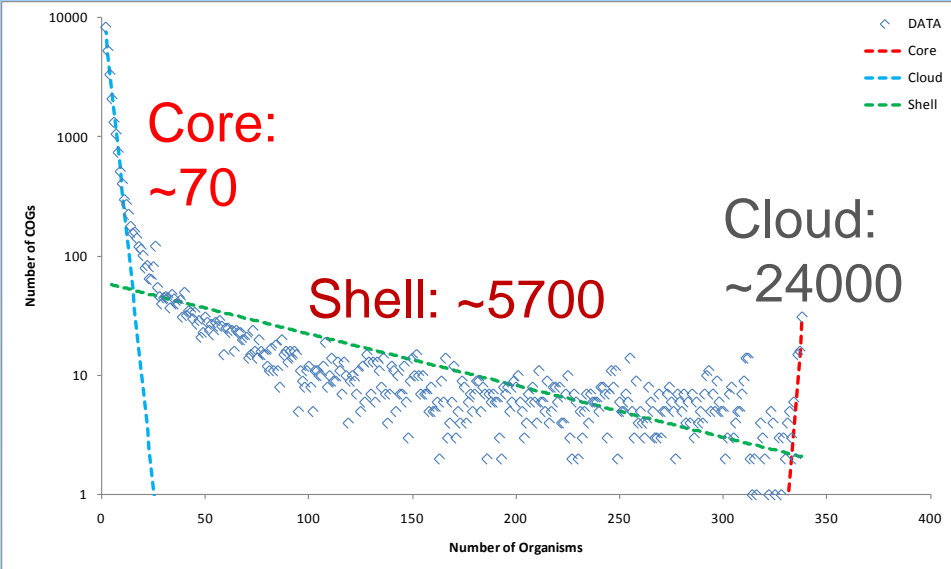
 extant genomes
 ancestral genomes

 extra- and intra-cellular mobilome elements

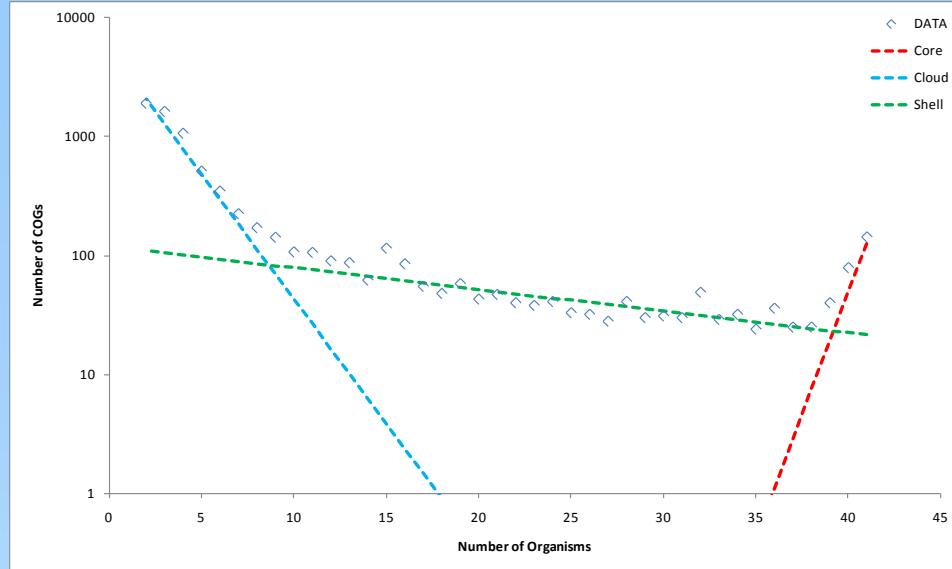
 vertical inheritance
 horizontal exchange
 mobilome exchange

Most (~70-80%) of genes in prokaryotic genomes are evolutionarily conserved –belong to COGs – orthologous lineages - distinct units of evolution

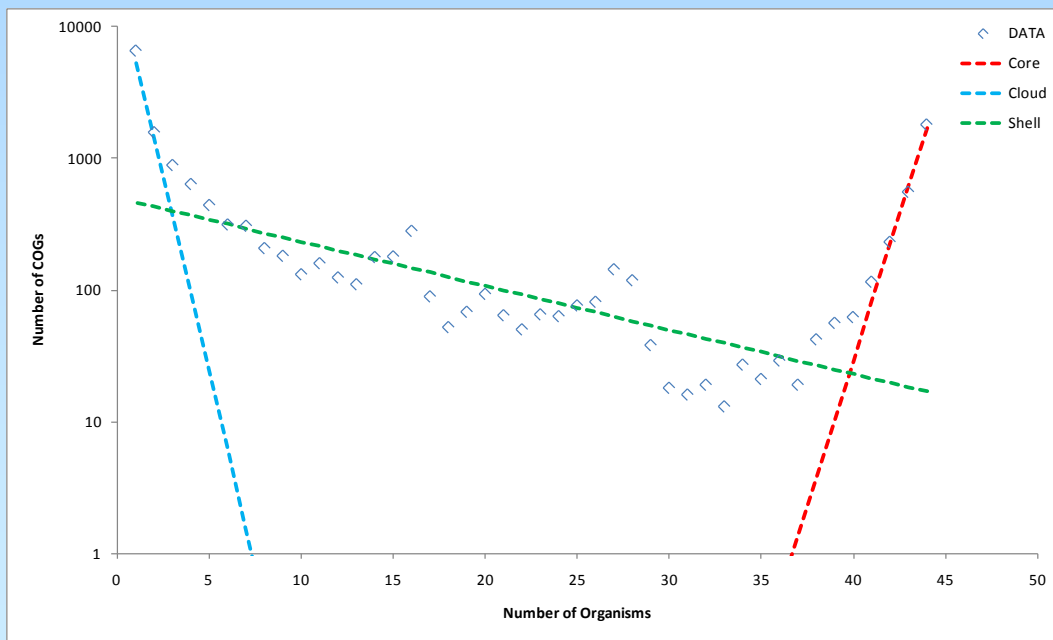




338 Archaea and Bacteria



41 Archaea



44 Escherichia and Salmonella

Fractal structure of the prokaryotic gene space:
Tripartite organization of pangenomes at all levels

The phylogenetic Forest of Life: Tree and net signals in the evolution of prokaryotes



Some basic facts on TOL ...and a few firmly held beliefs

- replication of genetic material creates a tree-like graph of relationships at the most basic ("atomic") level
- recombination (of all kinds – from typical sex to cross-kingdom HGT) turn a tree into a more general **directed acyclic graph (DAG)**
- "tree thinking" is inevitable and fundamentally relevant in biology but cannot be literally and strictly true – **TOL is an abstraction**
- gene histories in a population coalesce fast compared to characteristic evolutionary times
- concepts of "clades" and "ancestral genomes" (and, therefore, of orthology) are relevant
- reconstructed phylogenetic trees reflect the true evolutionary histories of genes (with all disclaimers about possible errors and artifacts)
- sequence-based phylogenetic reconstructions are useful in evolutionary research

Spectrum of Positions

Tree of Life is the dominant trend in evolution; HGT is rare and overhyped; most observed "transfers" are artifacts

Tree of Life is the common history of (nearly) non-transferable core of genes, surrounded by "vines" of HGT

All genes have their own evolutionary history blending HGT and vertical inheritance; there might exist a statistical trend in the pattern of gene histories (possibly even tree-like)

Ubiquity of HGT makes TOL concept totally obsolete; prokaryotic species and higher taxa do not exist; microbial "taxonomy" is created by pattern of shared HGTs

accept TOL

reject TOL

Kurland, Logsdon, Faguy

Daubin, Moran, Woese, Fitz-Gibbon, Fraser, Eisen, Salzberg, Kunin, Ouzounis, Bork, Galtier, Kim

Olsen, Koonin, Martin, Boucher

Doolittle, Gogarten, Baptiste

after **O'Malley & Boucher.** (2005) *Paradigm change in evolutionary microbiology.* Stud. Hist. Philos. Biol. Biomed. Sci. 36, 183-208

Charting the Forest of Life: Data and Methods

Source data and basic analysis methods:

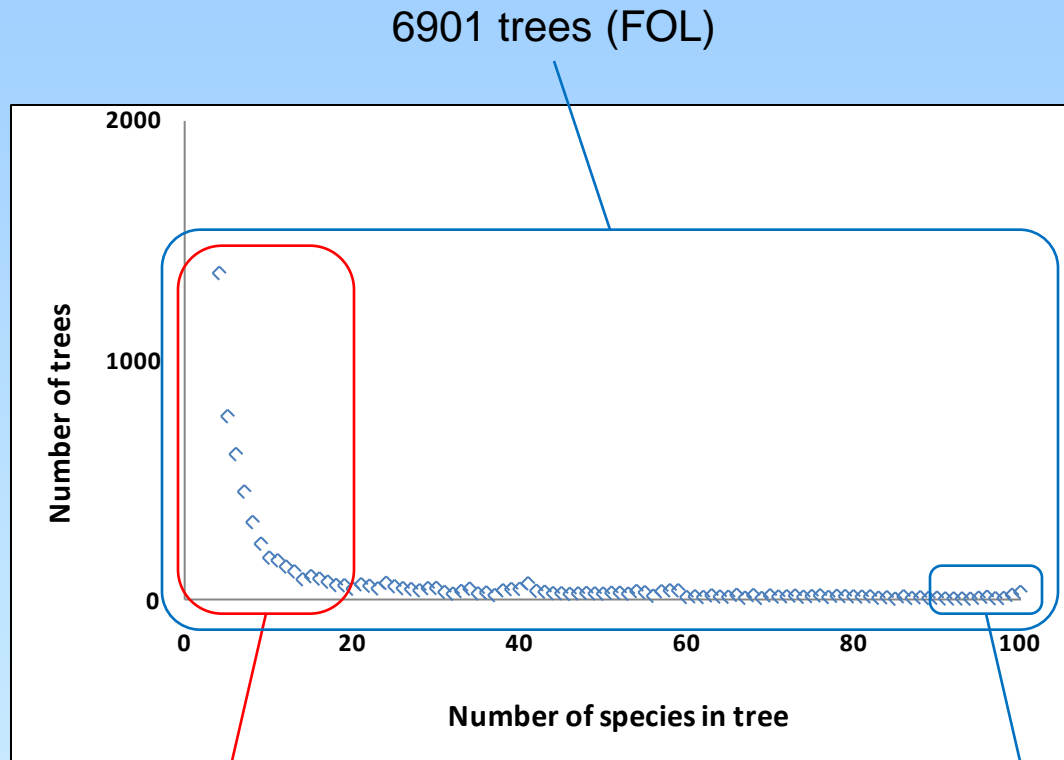
- 100 hand-picked microbial genomes (41 archaea and 59 bacteria) representing a "fair" sample of prokaryote diversity (as known in 2008)
- 6901 clusters of orthologous genes (NCBI COGs and EMBL EggNOGs)
- multiple protein sequence alignments → index orthologs → ML phylogenetic trees

Comparative analysis of trees:

- "Split Distance" (Puigbo et al. 2007) computes the distance between trees based on a fraction of shared bipartitions
- "Boot-Split Distance" (Puigbo et al. 2009) does the same but weighs bipartitions by bootstrap support of the corresponding internal branch
- both procedures produce a distance between a pair of (unrooted) trees that share at least 4 leaves in the range of [0..1]

FOL and NUTs

Forest of Life (**FOL**) and Nearly Universal Trees (**NUTs**)

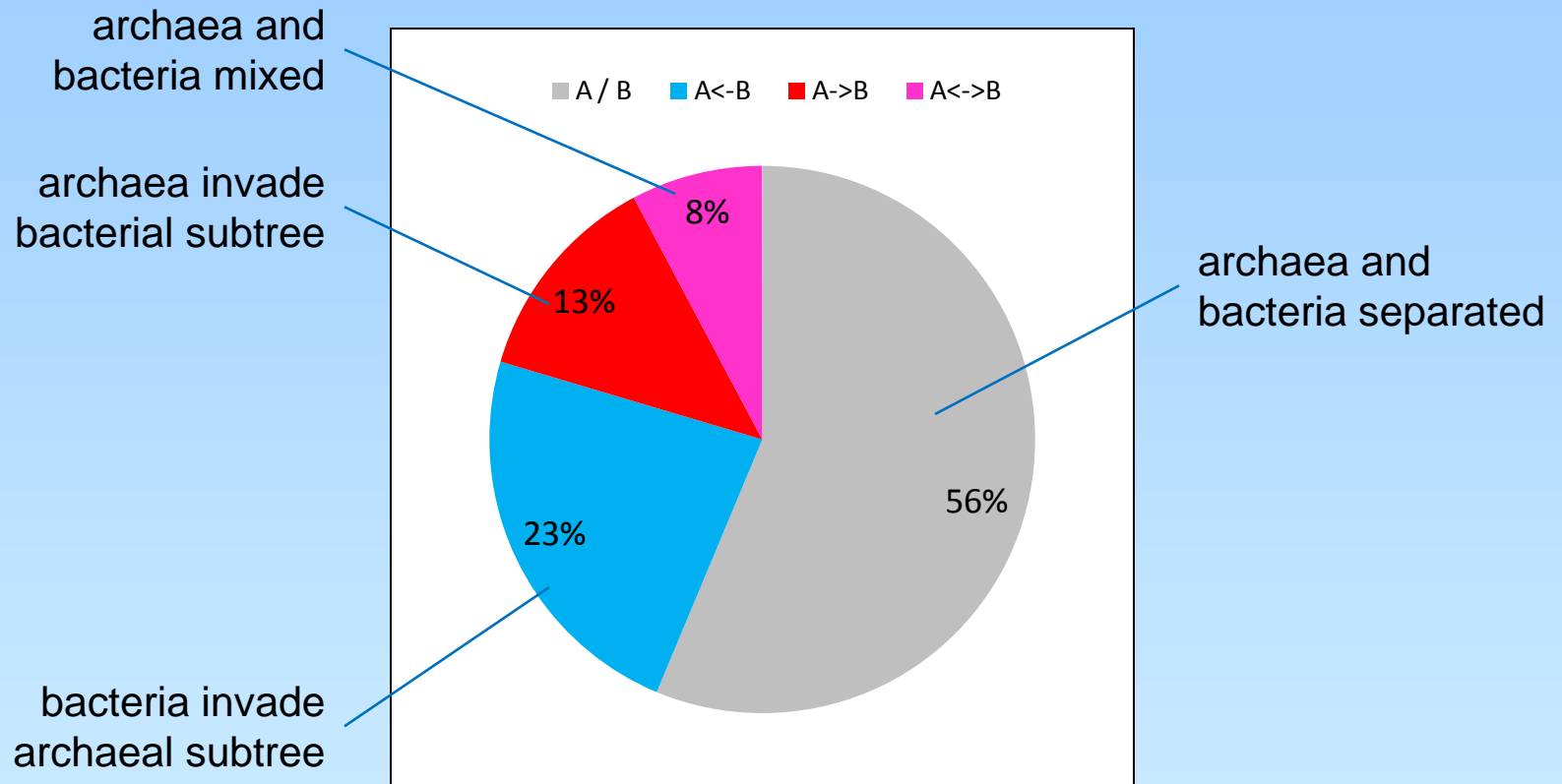


most of the trees
contain relatively
few species

102 "nearly universal"
trees (90+ species):
NUTs

Archaea and Bacteria going NUTs

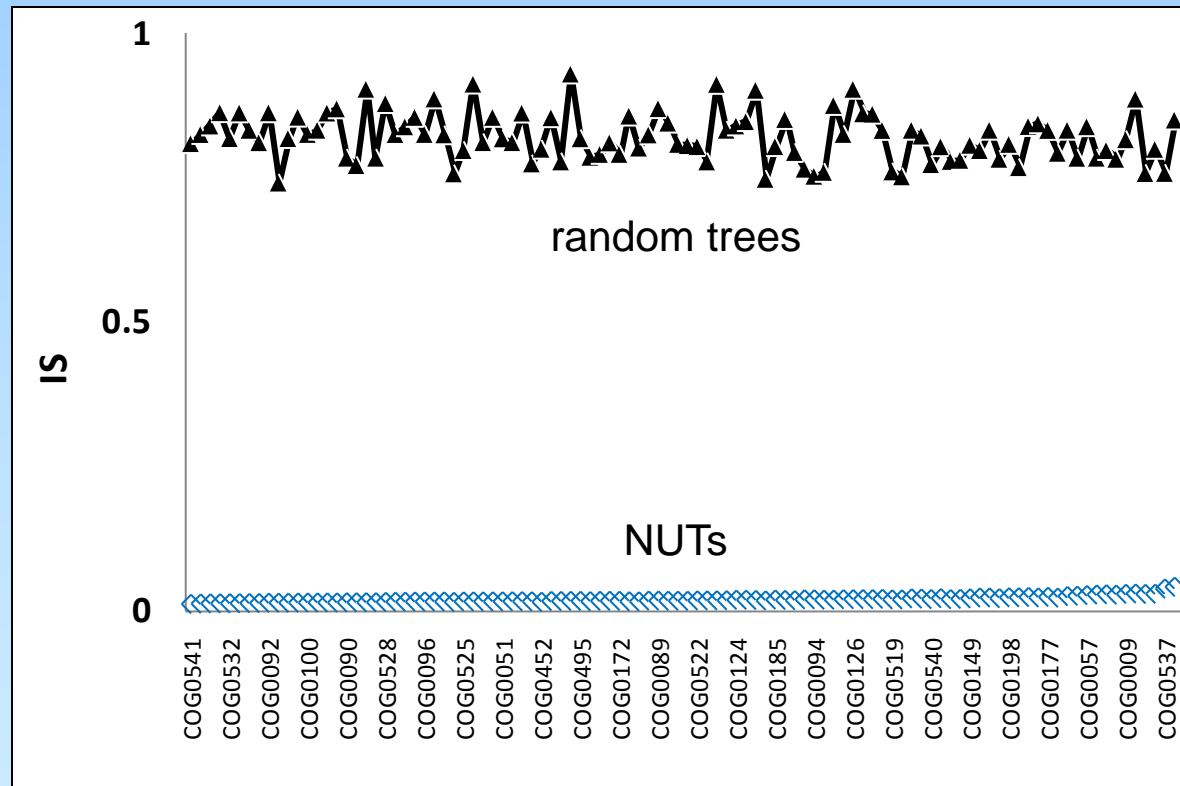
How well separated are archaea and bacteria?



56% of NUTs show perfect separation between archaea and bacteria;
92% of NUTs show partial, non-random separation

NUTs vs Random Trees

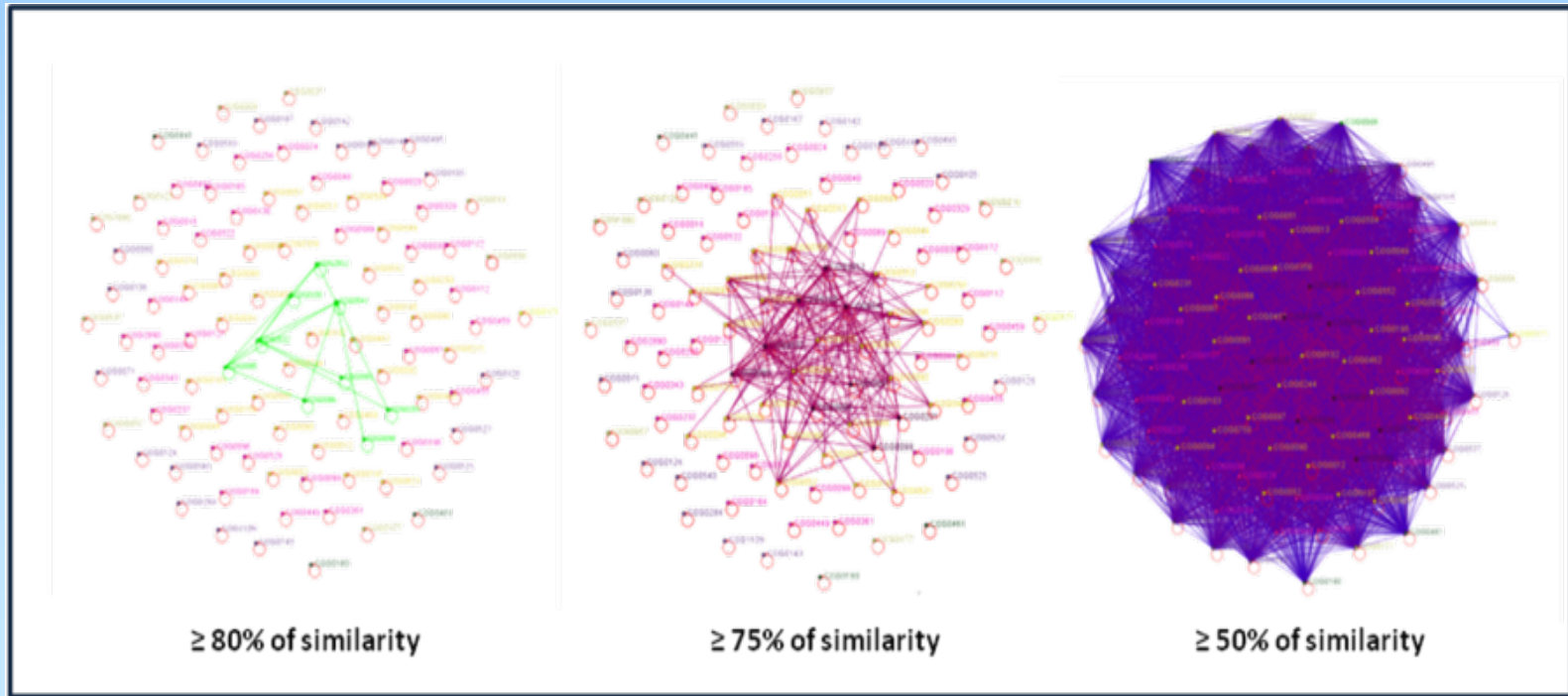
Inconsistency Score (IS) compares a tree to a set of trees. The score is based on the frequency of bipartitions derived from the given tree among all trees in the FOL. Range $\sim[0..1]$.



NUTs are incomparably more topologically consistent than random trees

NUTs Pattern of Similarity

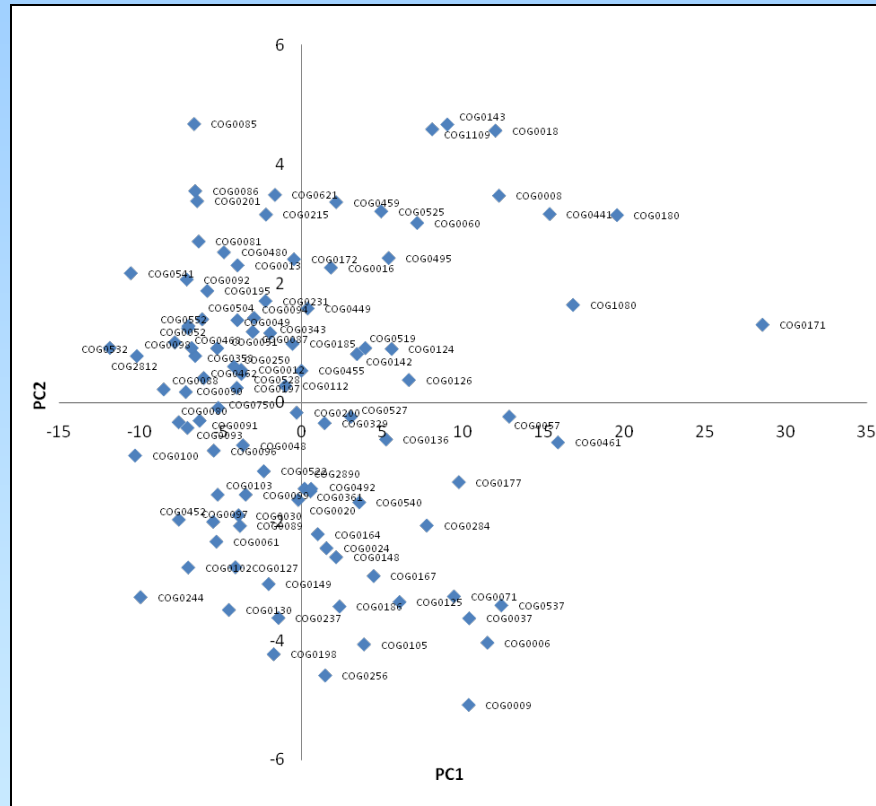
An edge connects two vertices (trees) if the distance between them is below the threshold.



A single connected cluster appears and gradually grows to encompass all NUTs as the threshold is lowered.

Are NUTs Clustered?

The 102x102 matrix of distances between NUTs is projected into a lower-dimensional space using CMDS.



Analysis using gap function approach (Tibshirani et al. 2001) shows lack of distinct separate clusters in the tree topology space

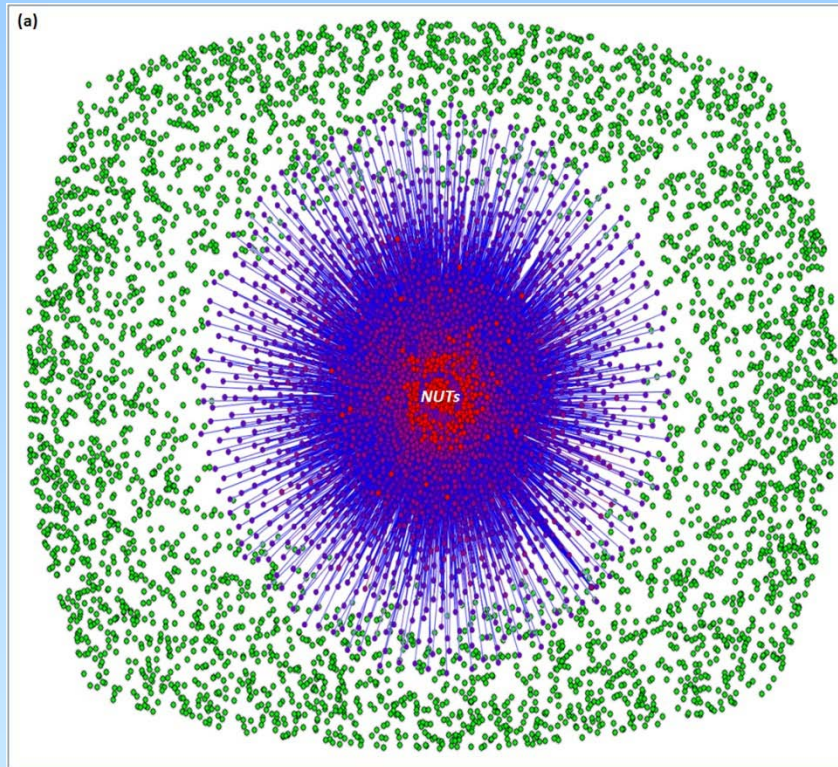
Conclusions – 1

Analysis of 102 Nearly Universal Trees (90+ species) shows that:

- most of the NUTs maintain at least some degree of separation between the bacterial and archaeal domains
 - mutual inconsistency among NUTs is dramatically lower than expected by chance
 - similarity pattern shows single central cluster of highly similar trees which gradually grows as the threshold is lowered
 - NUTs do not form distinct clusters in the tree topology space
- there exists a single common evolutionary pattern among the nearly universally conserved genes; individual NUTs seem to show random deviations from this pattern – **tree-like signal**

NUTs vs FOL

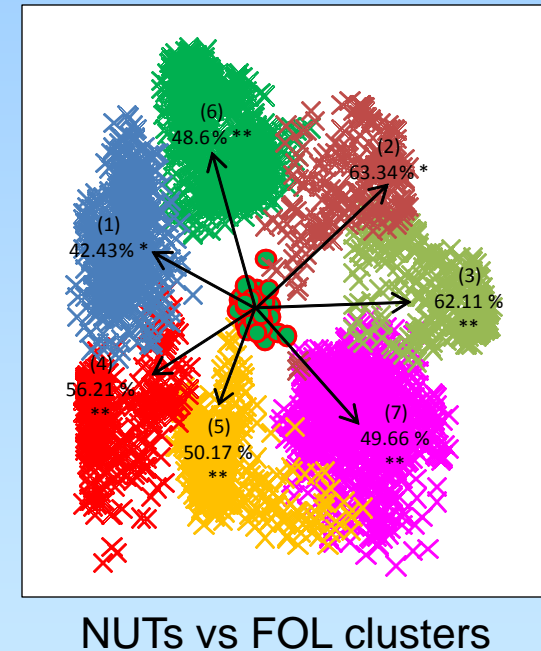
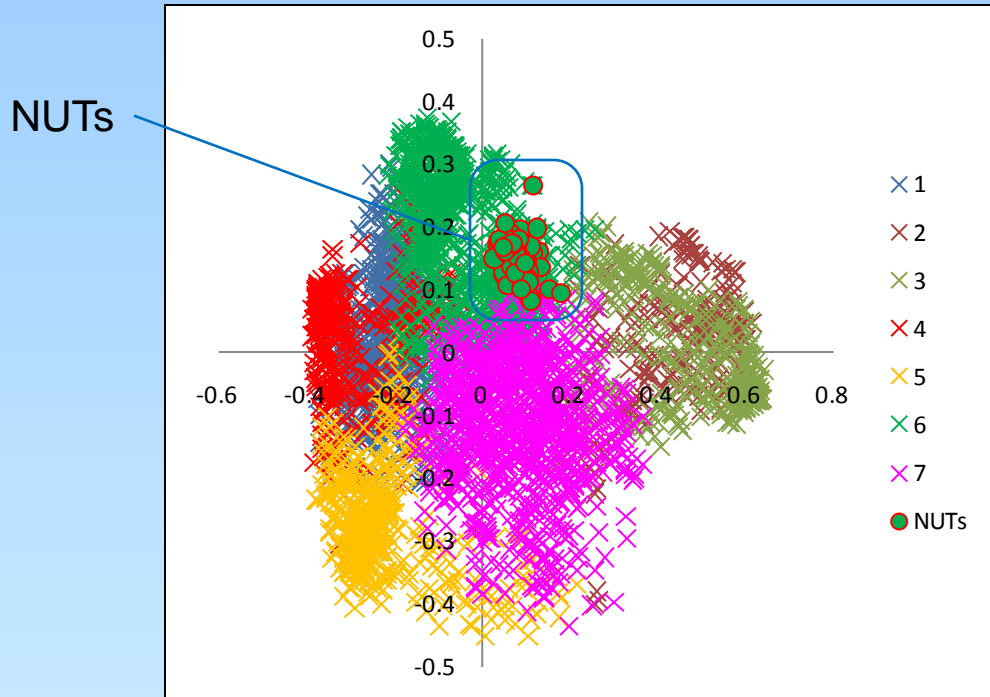
Similarity between NUTs and the rest of the FOL.



The NUTs are connected to 2505 trees (36%) from the FOL at a 0.8 similarity cut-off. The mean similarity between the NUTs and the FOL is ~ 0.5 (only ~ 0.1 for random trees).

NUTs vs FOL

The distance matrix for the entire is projected into a lower-dimensional space using CMDS.



FOL trees **form 7 distinct clusters** in tree topology space. Clusters differ largely by phyletic patterns. NUTs form a tight group within one of the clusters and are approximately equidistant to all clusters.

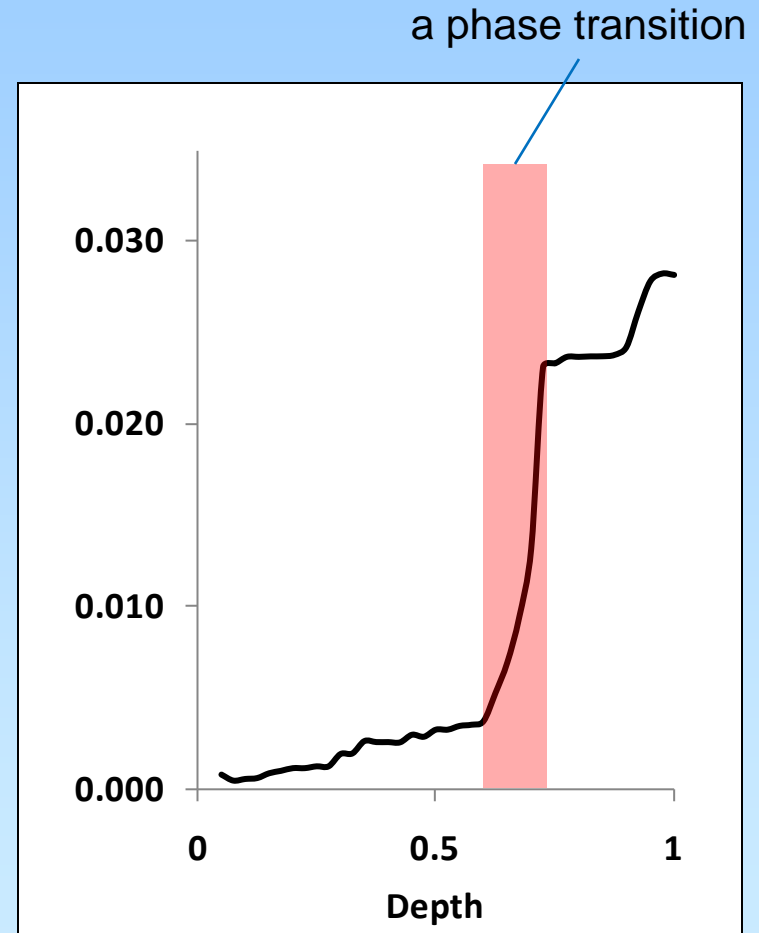
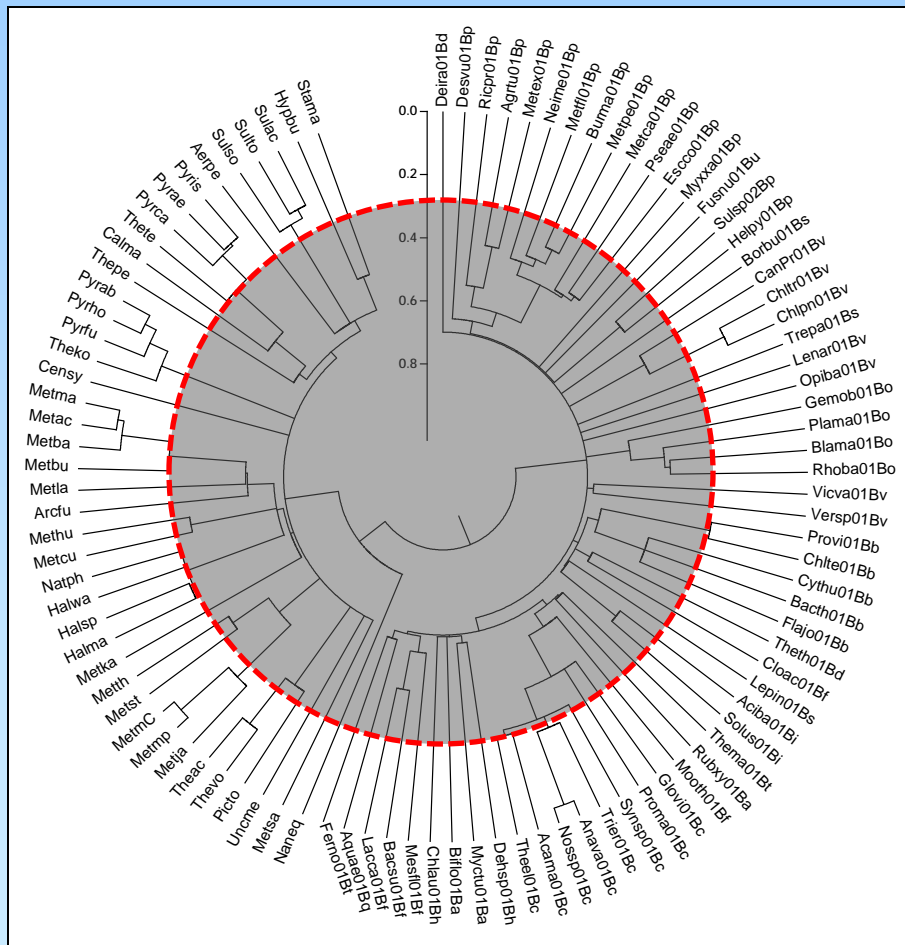
Conclusions – 2

Analysis of the full Forest of Life in comparison to NUTs shows that:

- a considerable fraction of FOL trees are very similar to NUTs: average FOL-NUTs similarity is dramatically above the random level
 - unlike NUTs, topologies of the FOL trees show distinct clustering largely determined by the phyletic patterns (i.e. set of species present)
 - in the tree topology space NUTs form a comparatively tight centrally located group
-
- compared to NUTs, FOL trees show much wider diversity of their topologies; however, the "central" tree-like signal still exists for a large part of the FOL
 - a "consensus" tree make sense at least as a crude representation of the common trend in the FOL (especially so for the NUTs).

NUTs: The Finer Structure

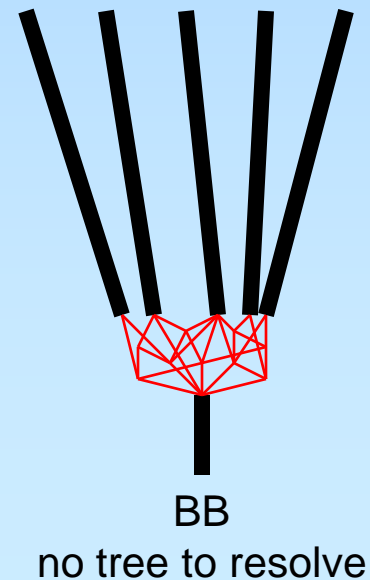
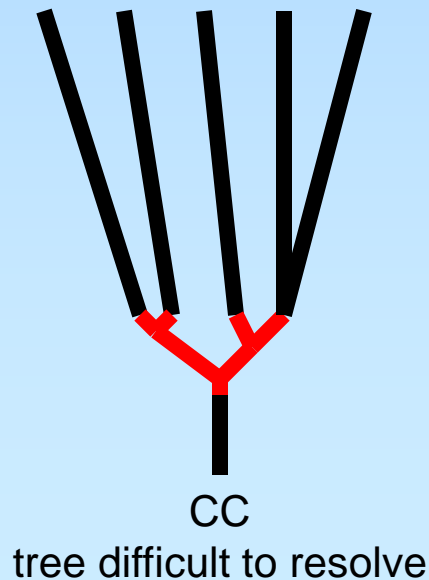
Tree inconsistency as a function of depth in a supertree (ignore all bipartitions below the threshold).



Coherence between NUTs takes a dramatic drop at depths of 0.65-0.75.

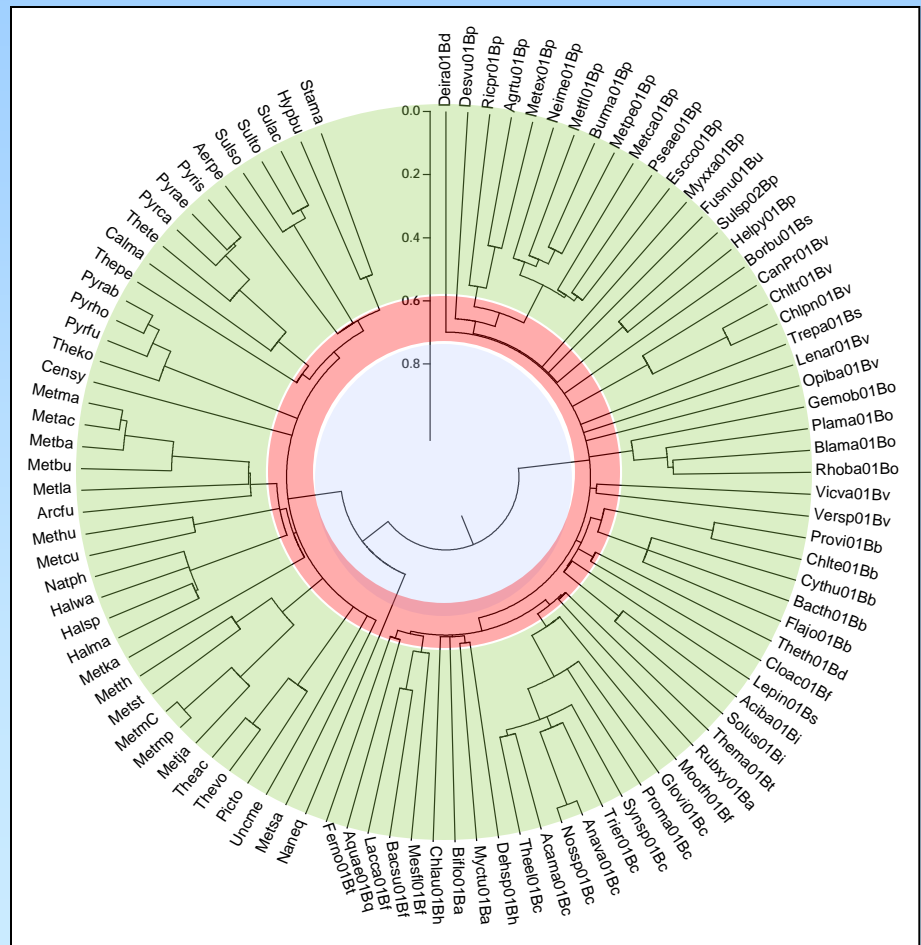
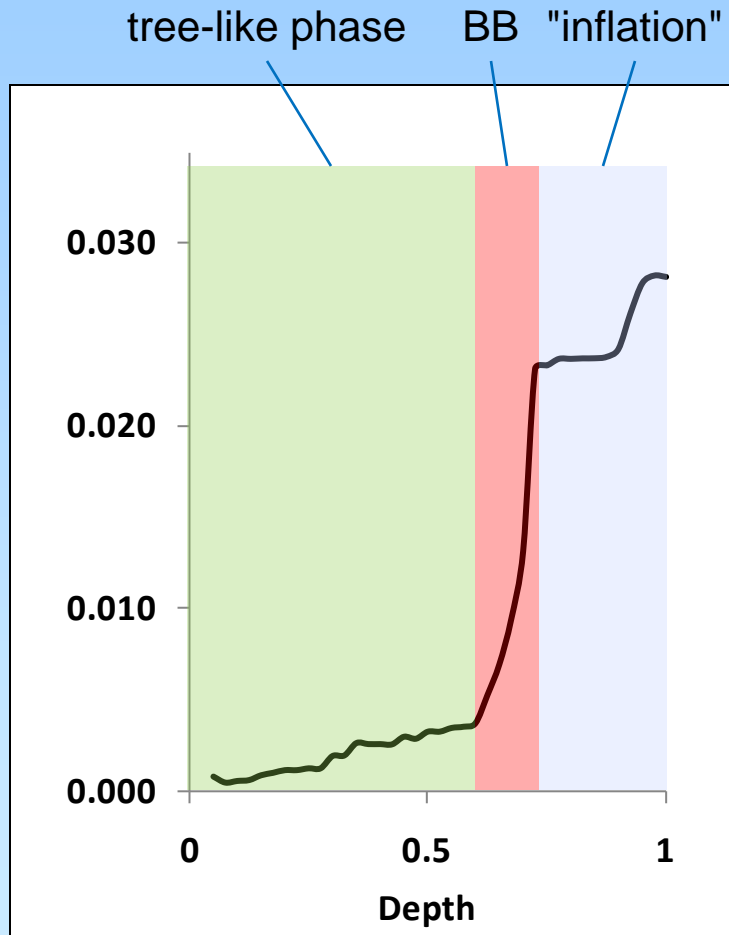
NUTs: CC or BB?

- rapid loss of tree-like signal at depths of 0.65-0.75 (roughly corresponding to divergence of bacterial and archaeal phyla). What could have happened?
 - "*Compressed Cladogenesis*" (CC) type of event – rapid diversification with short, difficult to resolve branches
 - "*Big Bang*" (BB) type of event – cladogenesis accompanied with a burst of HGT making the tree representation irresolvable and inapplicable



NUTs: CC or BB?

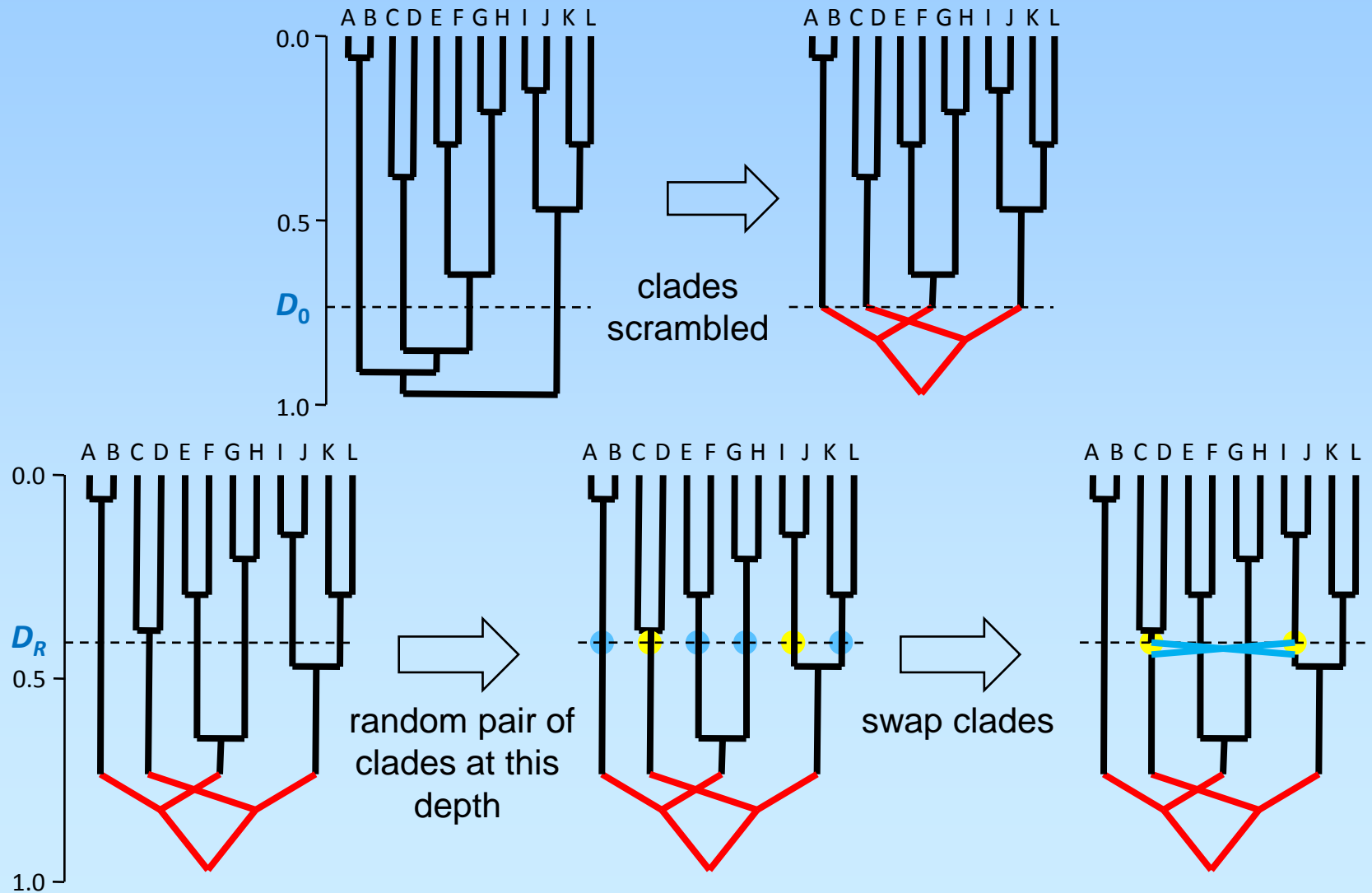
Possible interpretation for the "phase transition":



Was the phylogenetic signal destroyed by a "Big Bang" type of event?

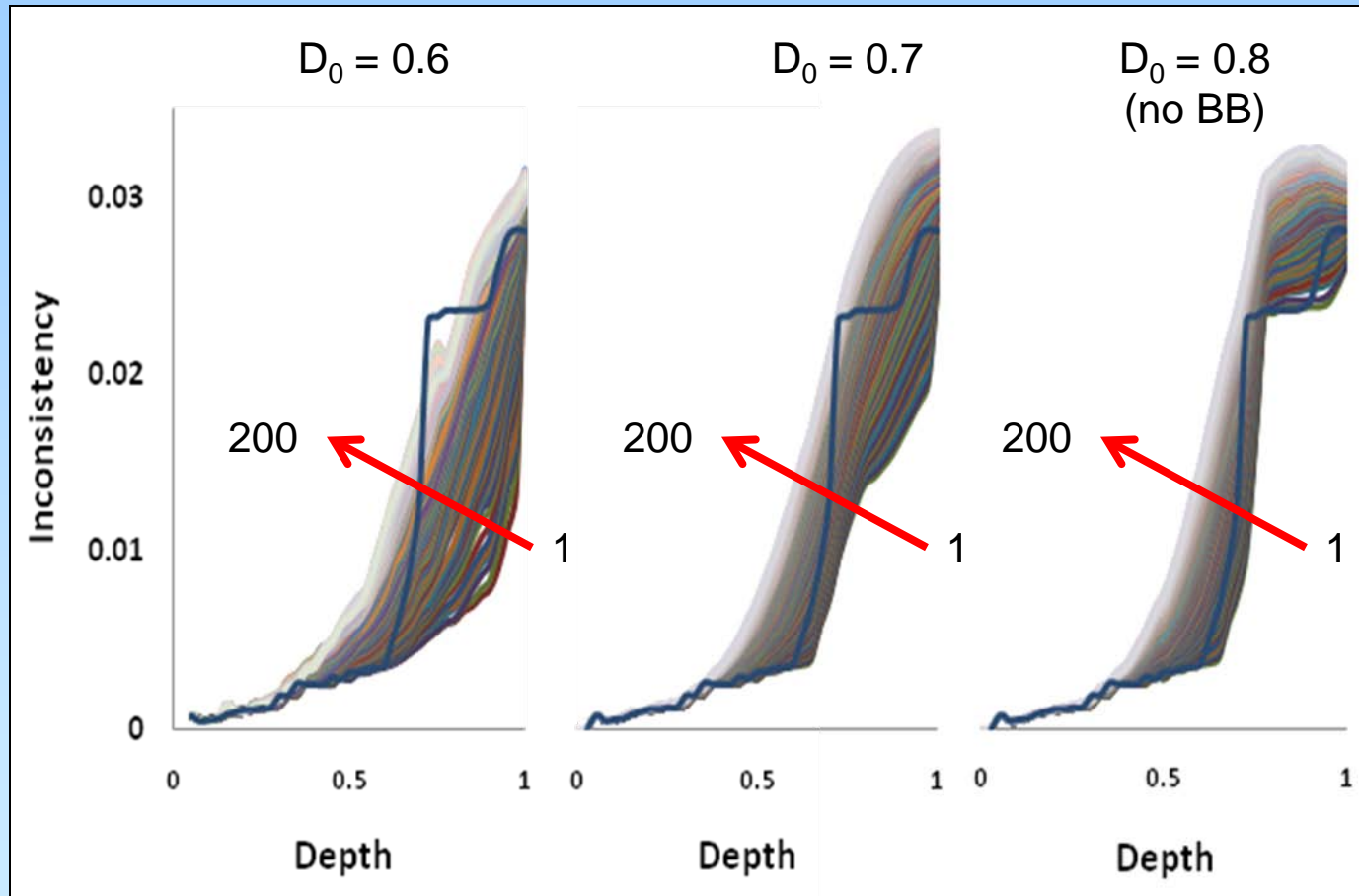
NUTs: CC or BB?

Computer simulation of BB followed by HGT



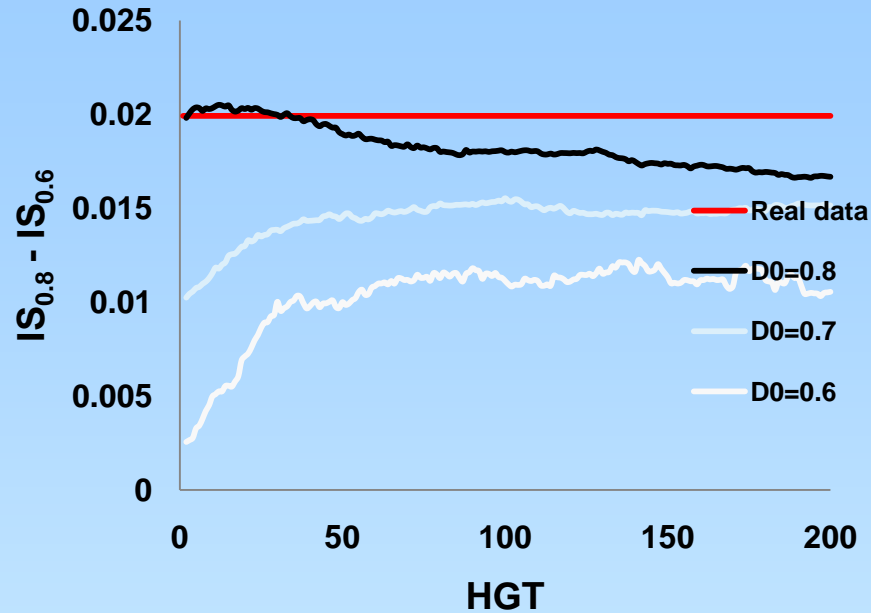
NUTs: CC or BB?

Computer simulation of BB followed by HGT (1 to 200 HGTs).



The best fit to the observed inconsistency curve comes from ~50 HGTs with **no BB**

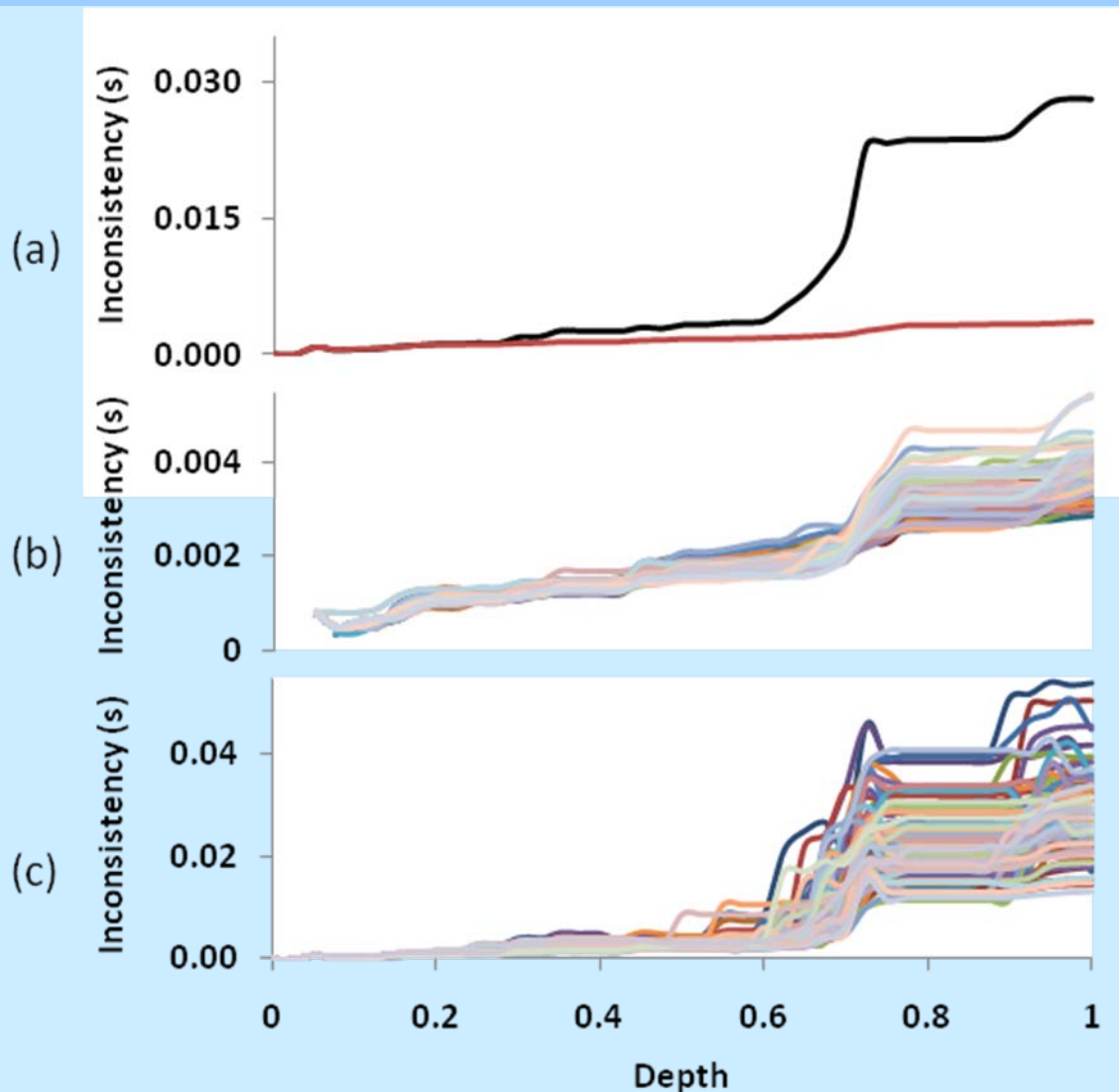
Testing the BBB model



Inconsistency drop between depths 0.6 and 0.8 depending on the number of simulated HGT events:



Testing the BBB model: NUTs vs FOL



a) Inconsistency vs depth plot: FOL in black, mean of 102 NUTs in red

b) Inconsistency vs depth plot: 102 NUTs

c) Inconsistency vs depth plot: 102 random trees from the FOL

Conclusions – 3

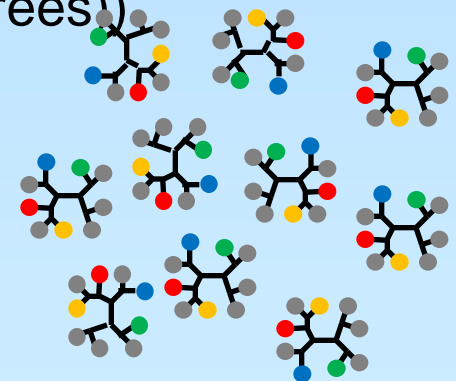
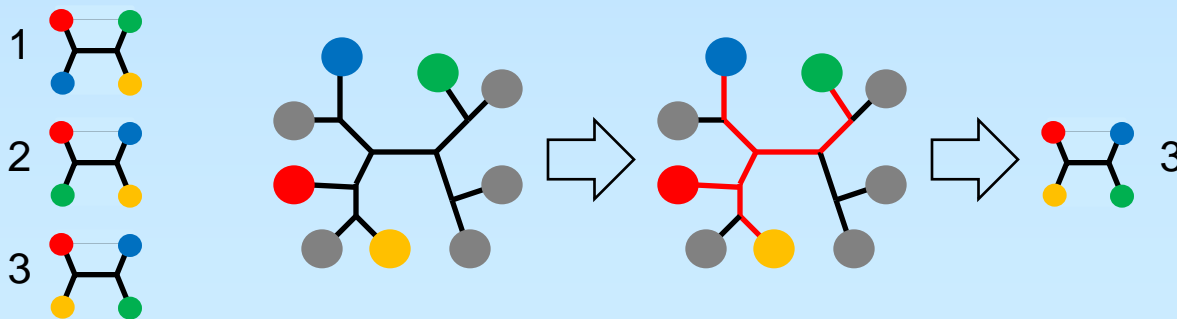
Computer simulation of BB+HGT model of NUTs evolution shows that:

- a "Big Bang" type of event is poorly compatible with the observed pattern of decline of tree consistency with phylogenetic depth
- the best fit is produced by ~50 random HGTs
- the observed phase-transition-like drop of tree consistency at the level of divergence of the major bacterial and archaeal clades (phyla) is probably a consequence of **compressed cladogenesis**
- a tree-like evolution of prokaryote lineages probably occurred but we might never resolve the deep TOL topology because the signal is weak and largely obliterated by subsequent HGTs

Distinguishing Tree and Net signals in the FOL using Quartets of species

The FOL shows a great diversity of phyletic and phylogenetic patterns. We employ the quartet analysis to measure the vertical and horizontal evolutionary signal in different areas of the Forest.

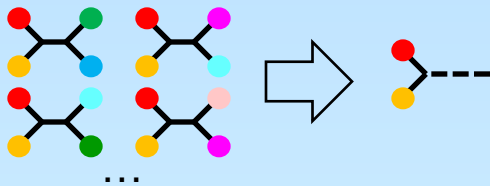
- there are $C^4_{100} \approx 4 \times 10^6$ species quartets from the set of 100 species
- each quartet of species can resolve into 3 different tree topologies ($\sim 12 \times 10^6$ combinations total)
- any tree containing all 4 species from a given quartet resolves them into one of these 3 topologies
- for any quartet one can compute the support for all 3 topologies within a set of trees (i.e. relative frequencies of the topologies); $\sim 8 \times 10^{10}$ comparisons for the whole FOL (or any subset of trees))



$$f_1 + f_2 + f_3 = 1$$

Charting the FOL with Quartets

- for any pair of species, the support across all quartets that put these species together can be averaged to calculate distance d_{ij}
- construct a 100x100 distance matrix showing how often any pair of species comes out as neighbors in this set of trees



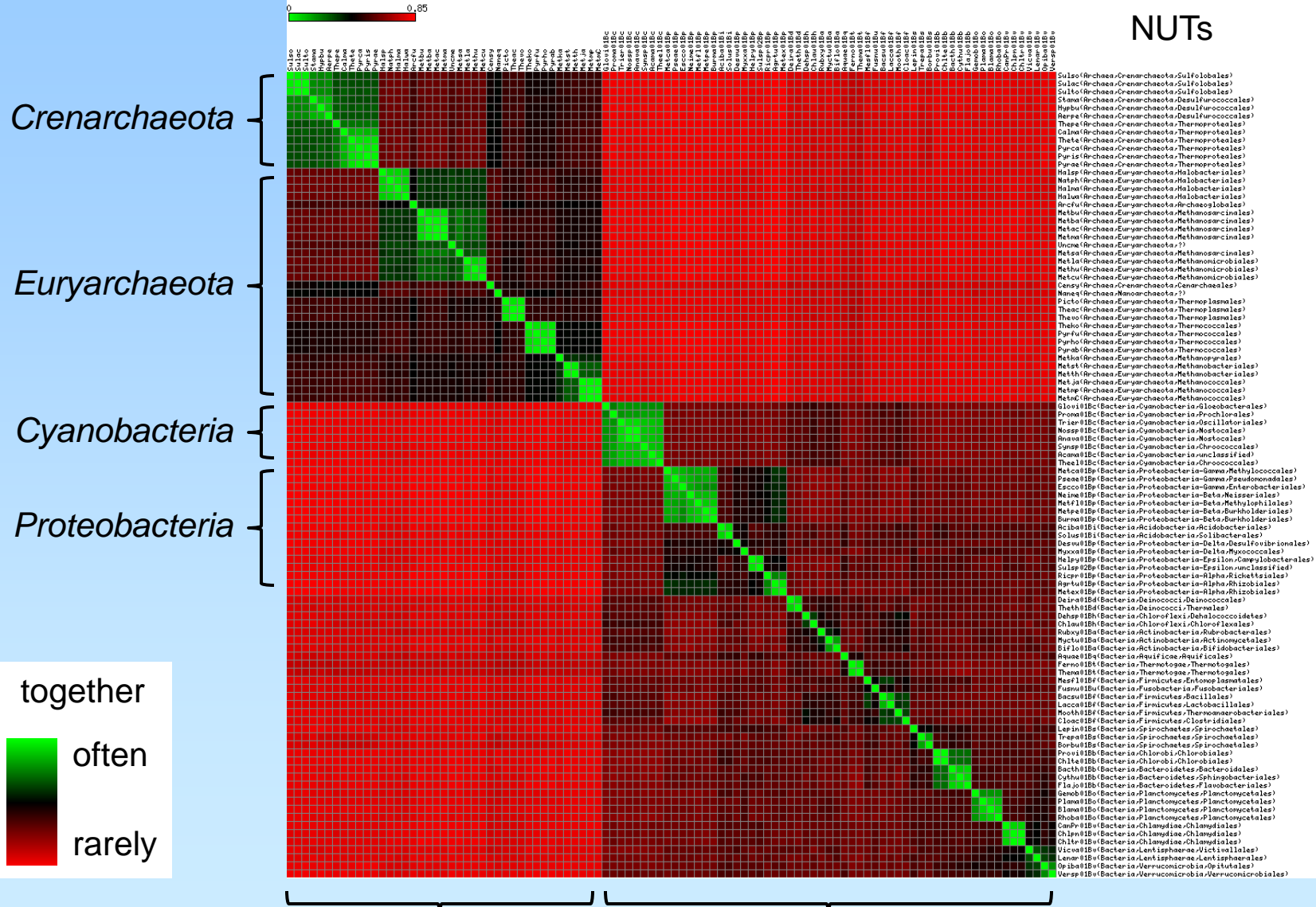
<i>species</i>	1	2	3	...
1	0.00
2	...	0.00
3	0.00	...
...

	support	distance
Always	1.00	0.00
Random	0.33	0.67
Never	0.00	1.00

Quartet Species Matrix – NUTs species order according to supernetwork of NUTs

National Center for Biotechnology Information

NUTs

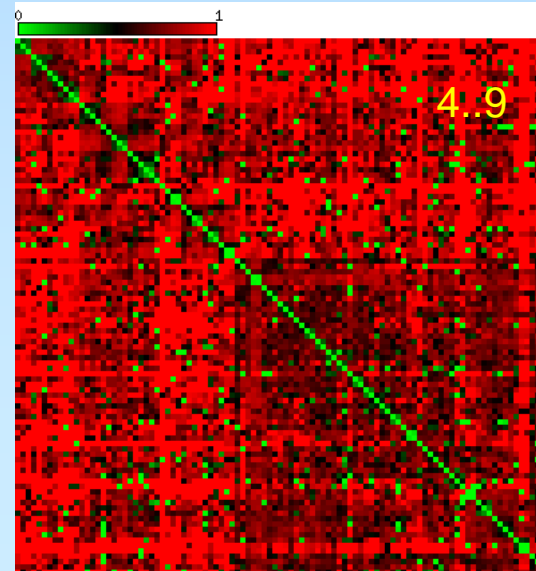
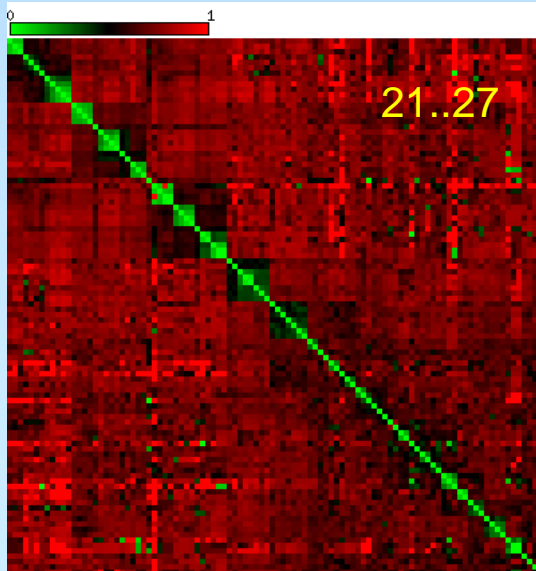
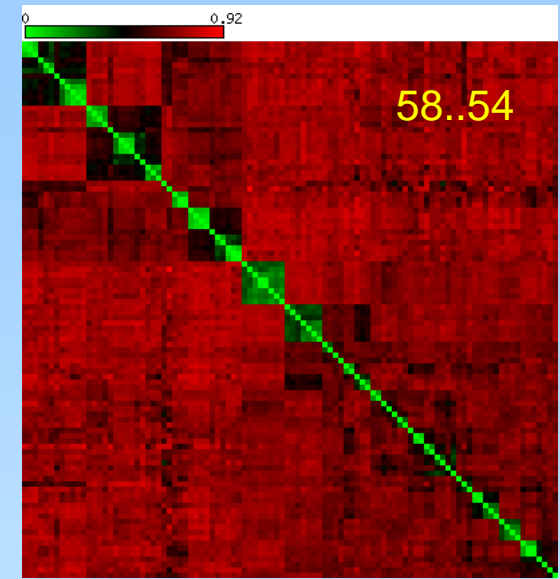
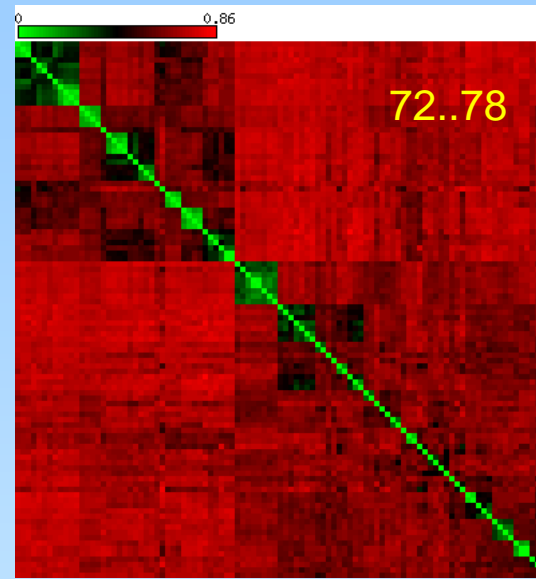
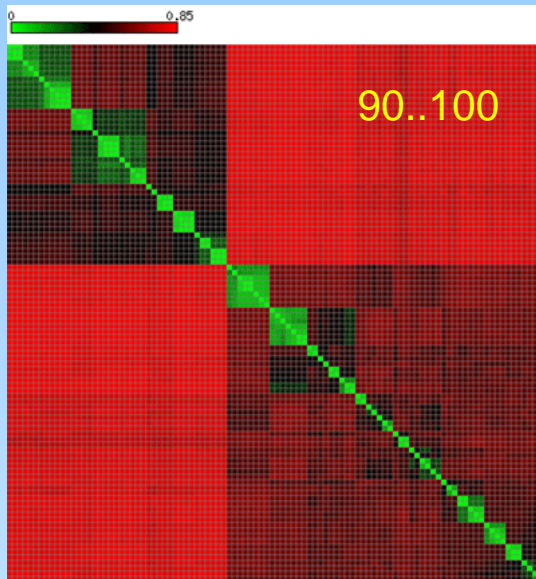


Archaea

Bacteria

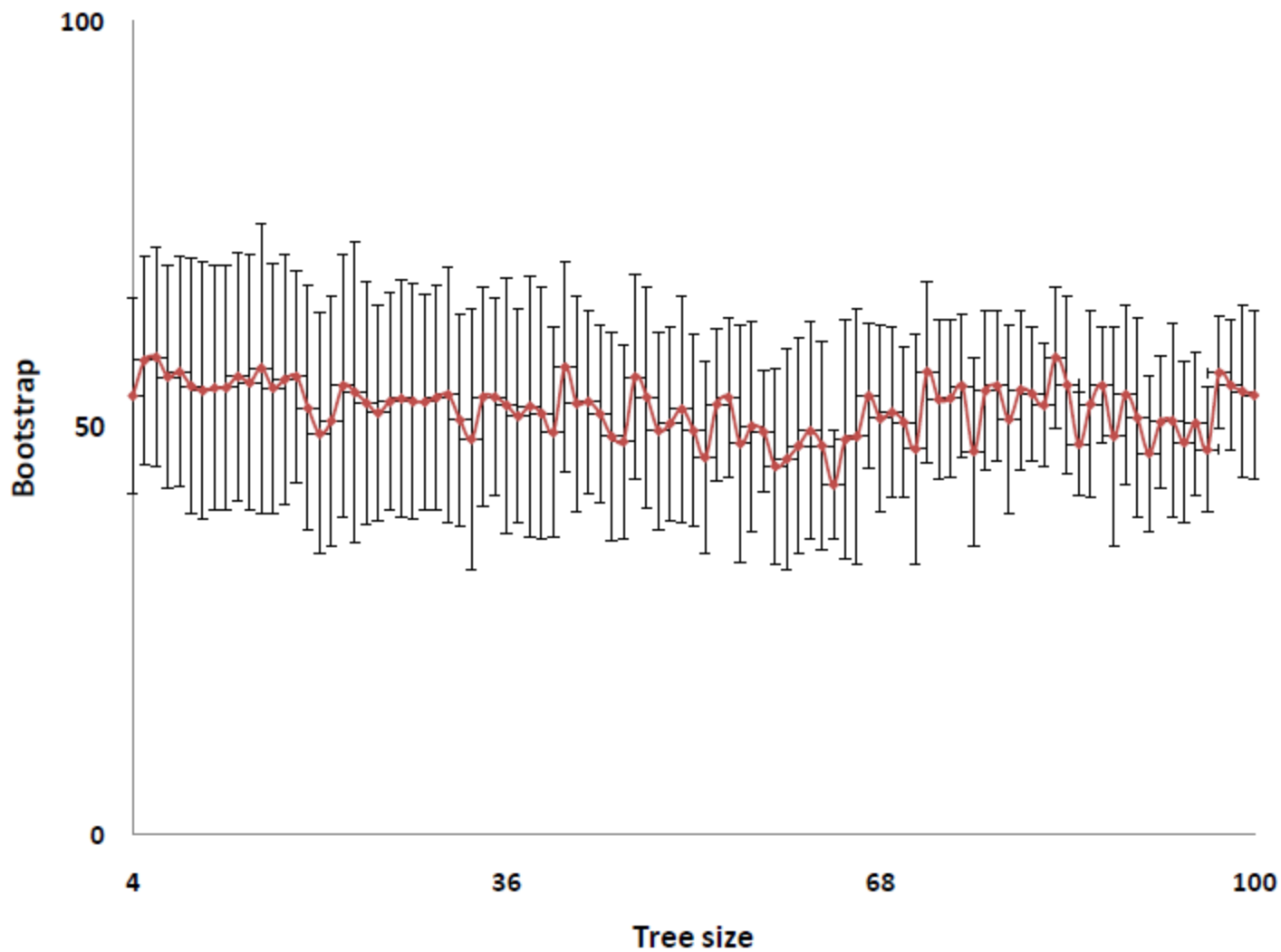
together
often
rarely

Quartet Species Matrix vs Tree Size



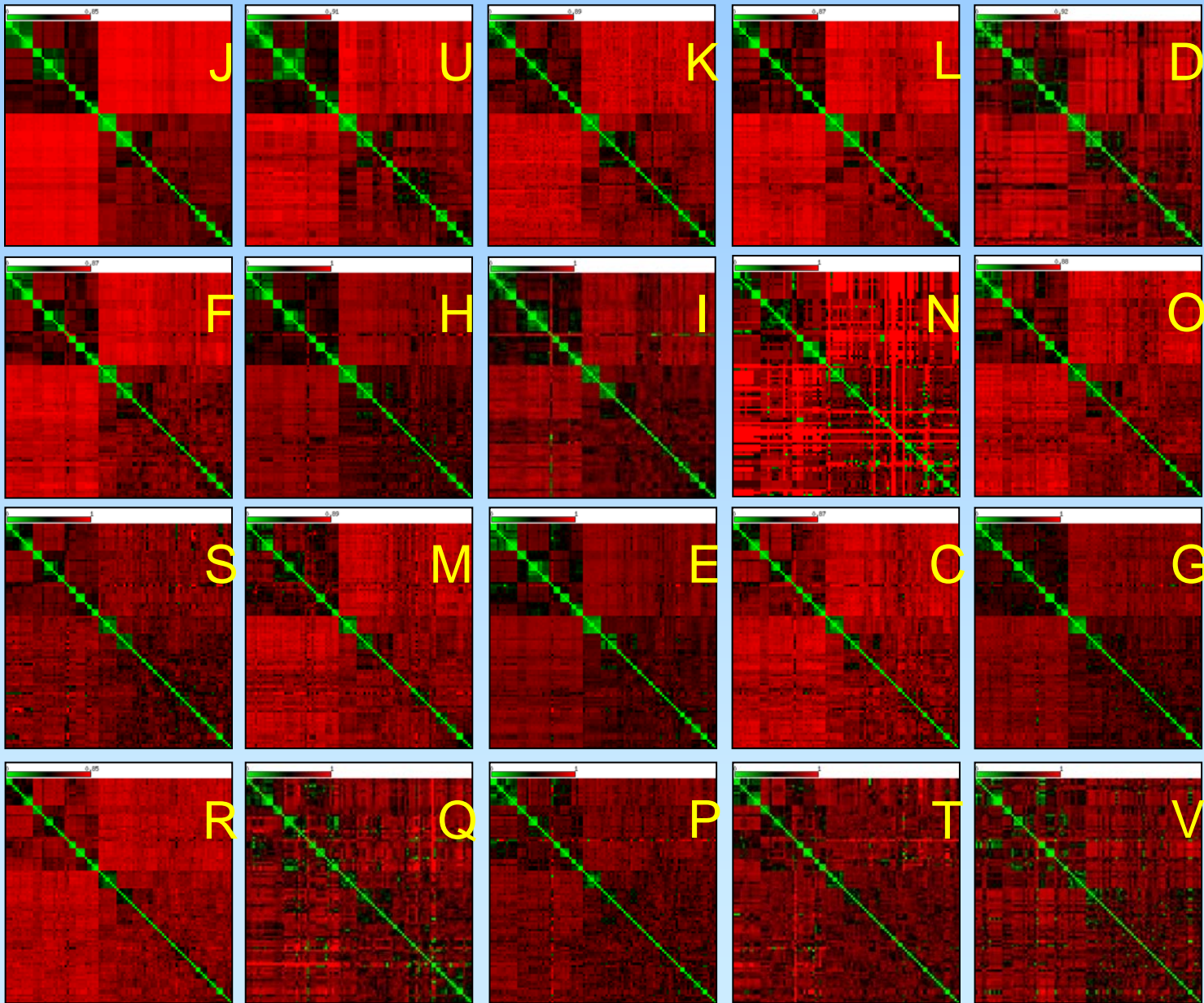
In smaller trees the consistent signal congruent with taxonomy degrades into a quasi-random pattern

Figure S3

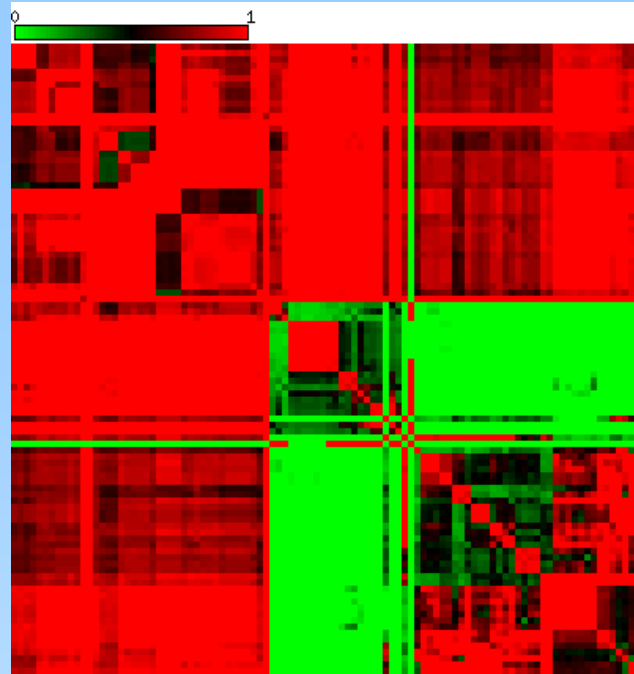
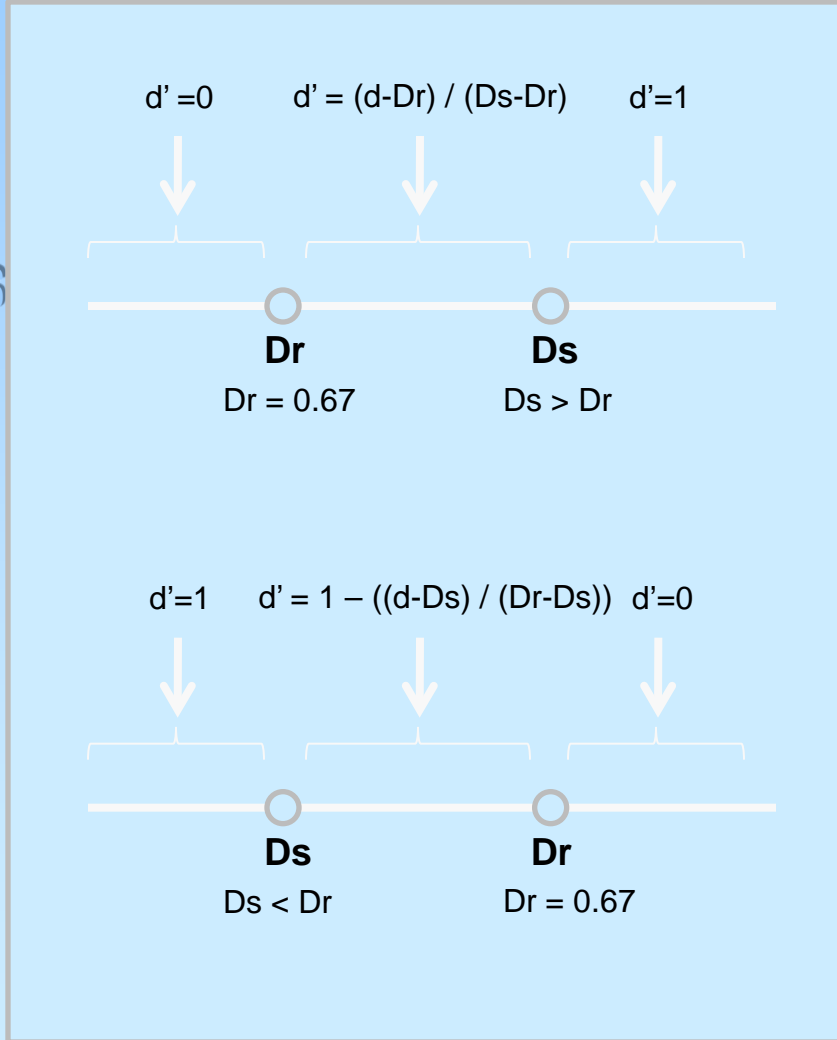


Quartet Species Matrix vs Function

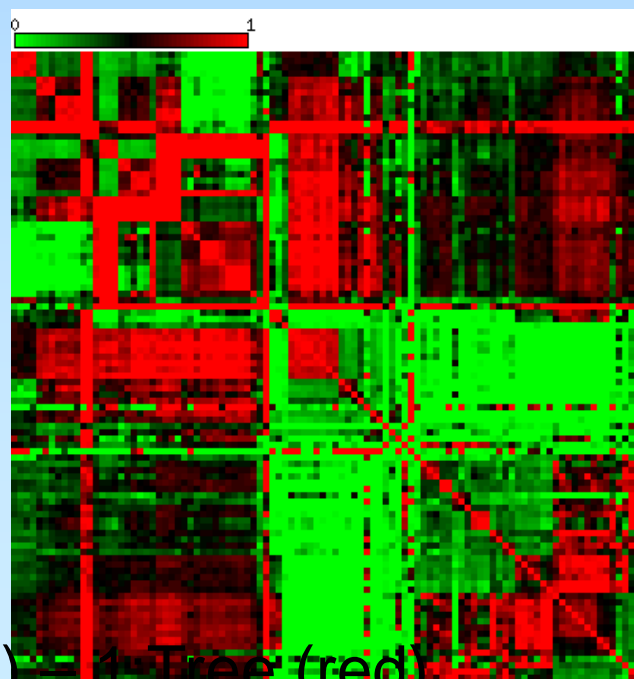
National Center for Biotechnology Information



TNT (Tree/Net Trend): scoring tree-like and net-like evolution quantitatively



NUTs
0.63 +/- 0.35



FOL
0.39 +/- 0.31

0: Network(green) – Neutral (black) 1: Tree (red)

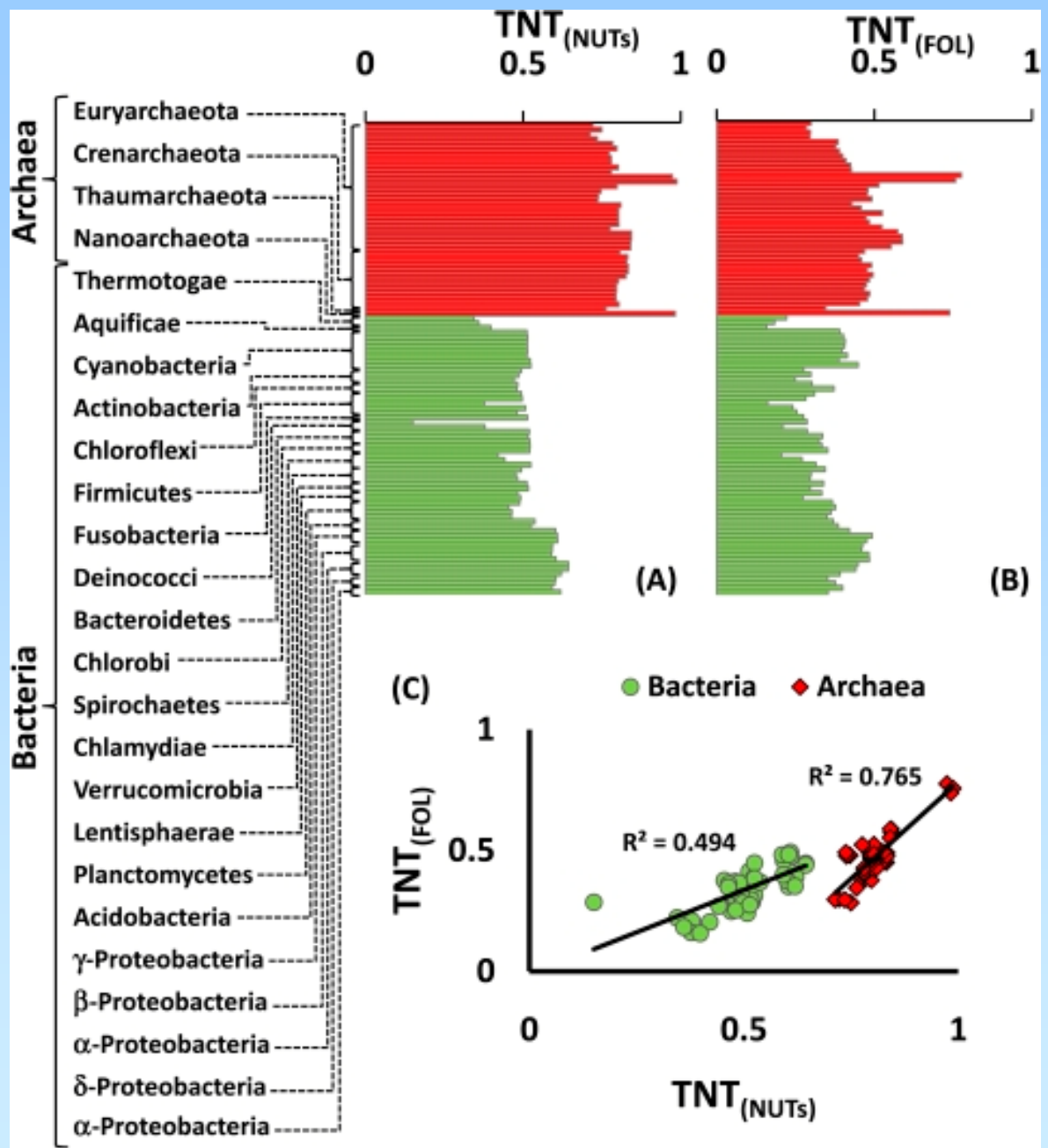
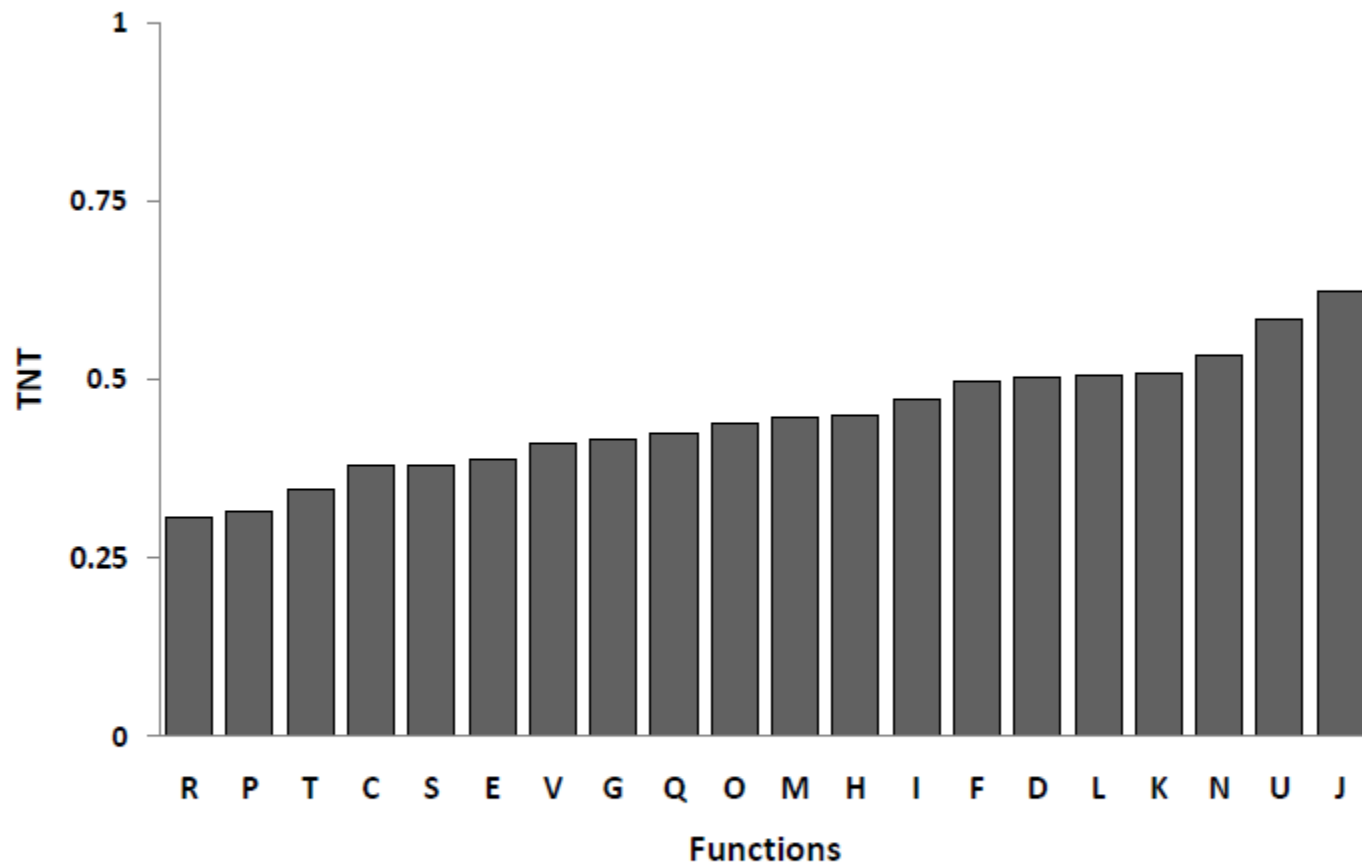


Figure S17



Conclusions – 4

Quartet-based analysis of the Forest of Life shows that:

- relationships between species in NUTs roughly follow the conventional microbial taxonomy (presumably a consequence of significant contribution of tree-like evolution)
 - the “TOL” signal decays with decreasing number of species in the tree
 - different functional classes of genes show substantially different balance between tree-like and net-like modes of gene transfer and possibly in preferred routes of HGT
 - **Evolution among the NUTs is “2/3 tree” but the evolution in the entire FOL is “almost 2/3 net”**
- these observations are compatible with the "core-shell-cloud" concept with propensity for HGT tending to decrease for universally conserved genes involved in the key information storage and processing machinery.

“GDL conjecture”:

-appearance of trees in prokaryotes might be explained by a gradient of HGT from “closely related” to “distantly related” organisms

Gogarten JP, Doolittle WF, Lawrence JG.

Prokaryotic evolution in light of gene transfer.

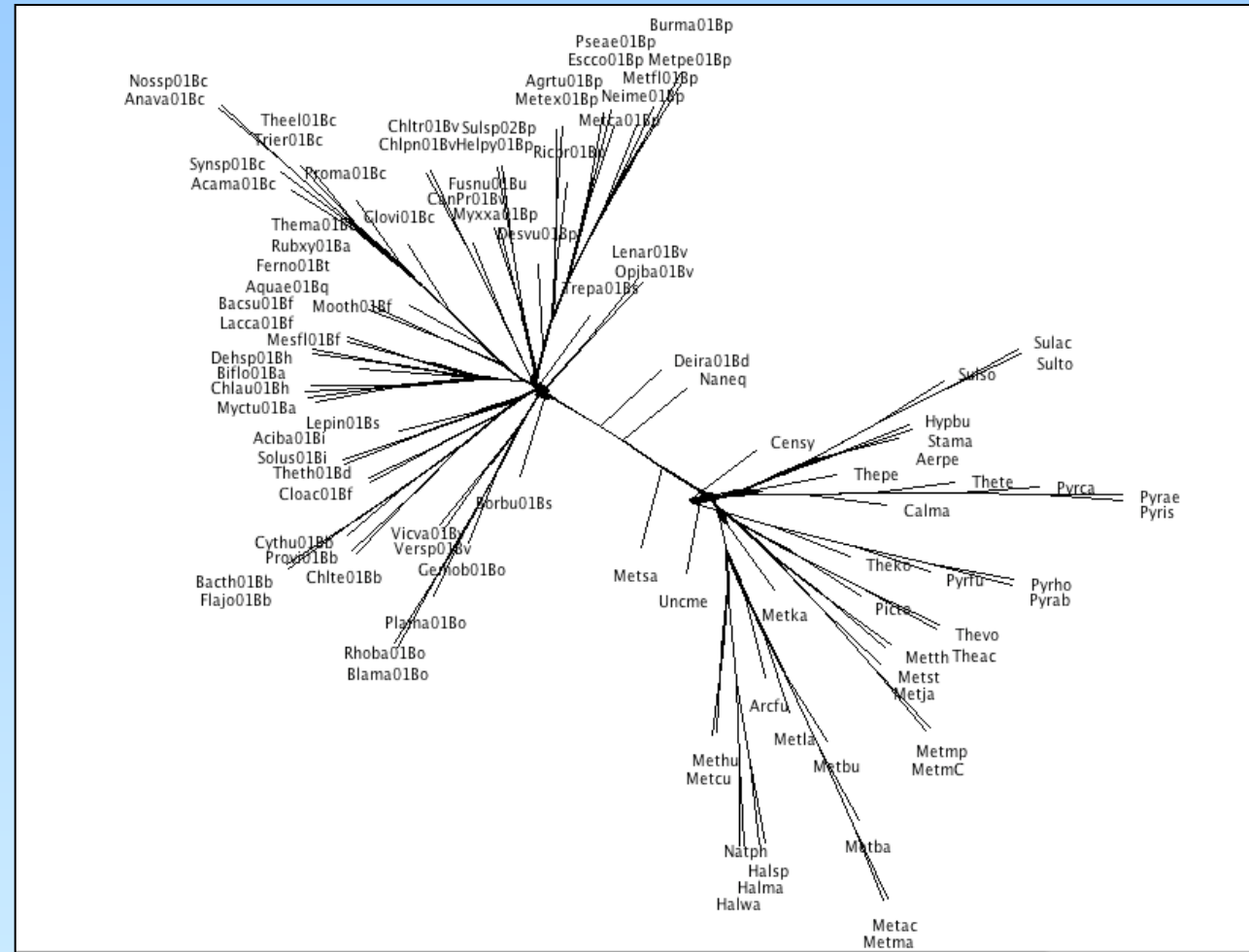
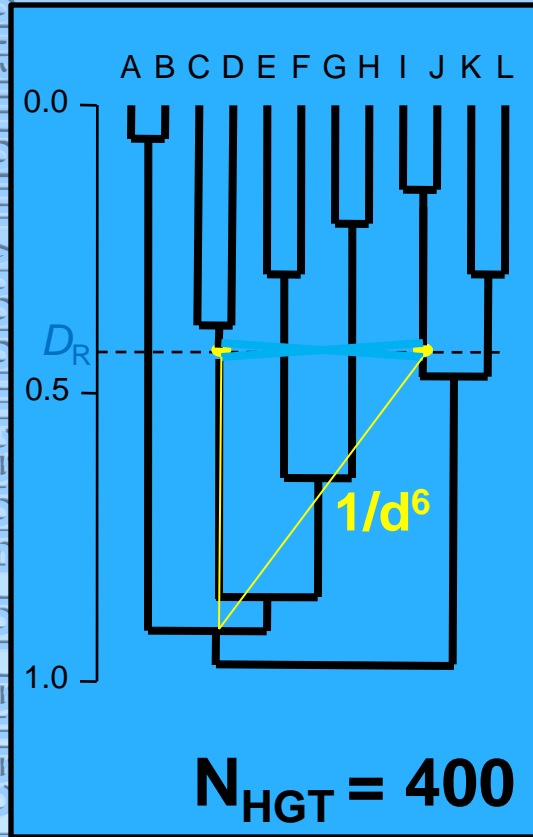
Mol Biol Evol. 2002 Dec;19(12):2226-38.

The traditional view, that prokaryotic evolution can be understood primarily in terms of clonal divergence and periodic selection, must be augmented to embrace gene exchange as a creative force, itself responsible for much of the pattern of similarities and differences we see between prokaryotic microbes. Rather than replacing periodic selection on genetic diversity, gene loss, and other chromosomal alterations as important players in adaptive evolution, gene exchange acts in concert with these processes to provide a rich explanatory paradigm-some of whose implications we explore here. In particular, we discuss the role of recombination and HGT in giving phenotypic "coherence" to prokaryotic taxa at all levels of inclusiveness,

- (1) the implications of these processes for the reconstruction and meaning of "phylogeny,"
- (3) new views of prokaryotic adaptation and diversification based on gene acquisition and exchange.

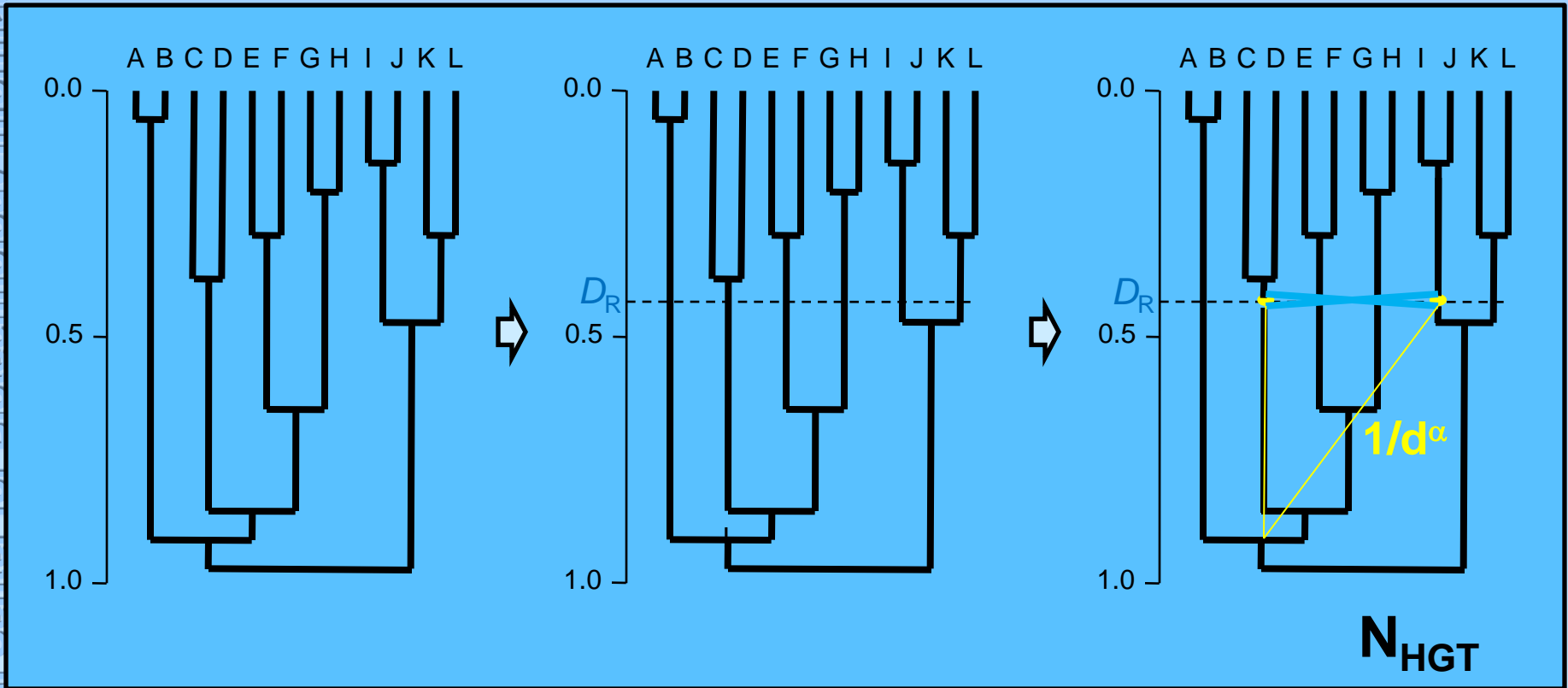
SIMULATIONS

Weak GDL conjecture: are the data compatible with HGT gradient?



-no iterations: keeping the tree matrix derived from the NUTs supertree throughout

SIMULATIONS

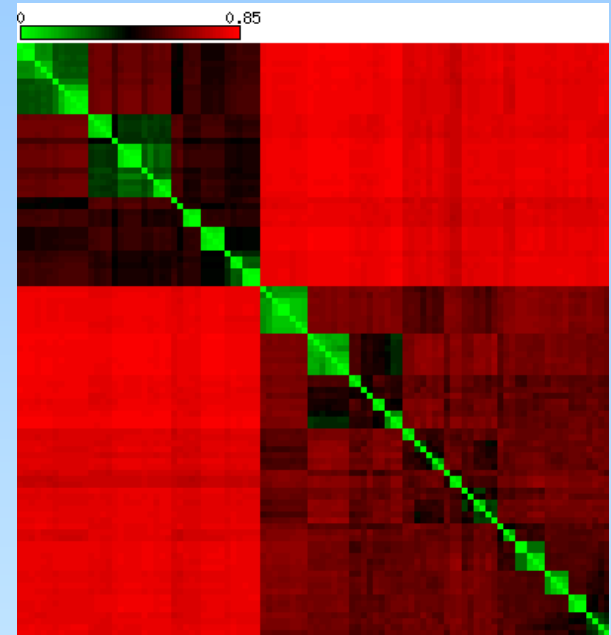
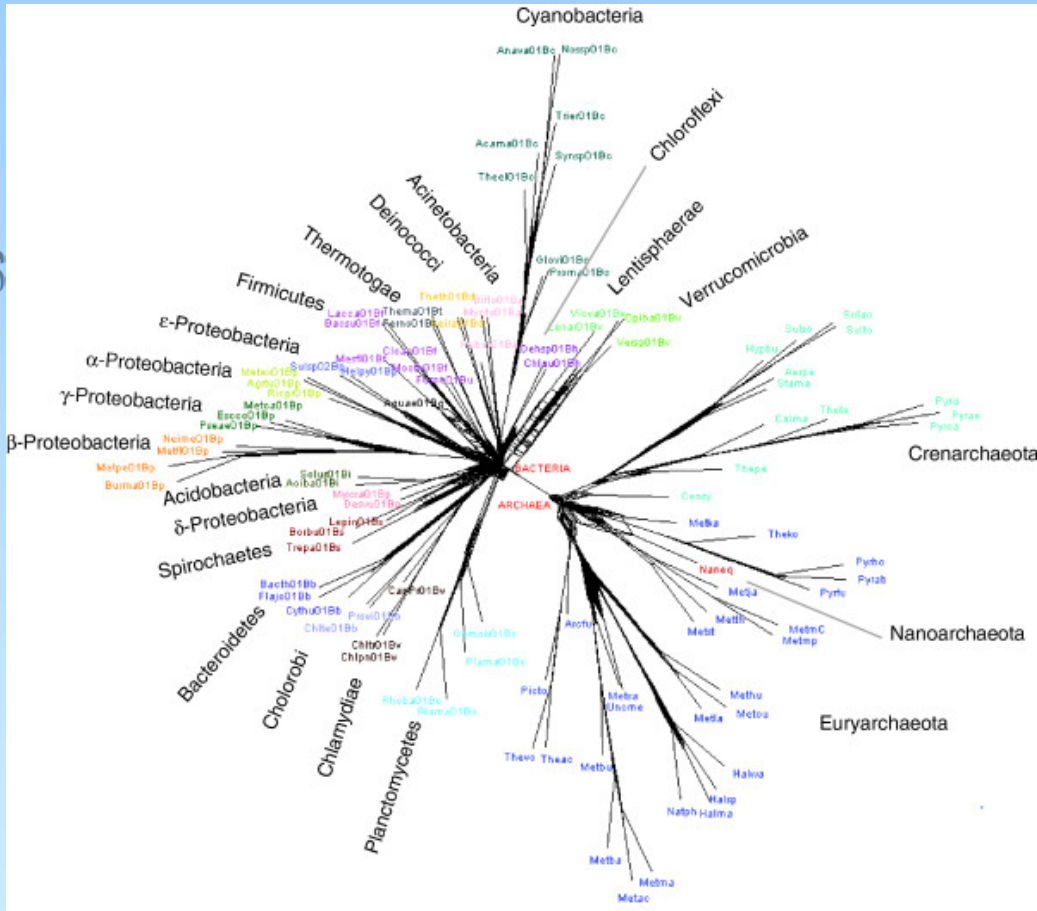


- variables: N_{HGT} ; gradient of HGT rate from tips to root (α)
- target values: $SS_{\text{archaea/bacteria}}$; mean bootsplit distance (D_R)



Supernetwork of NUTs

Heatmap

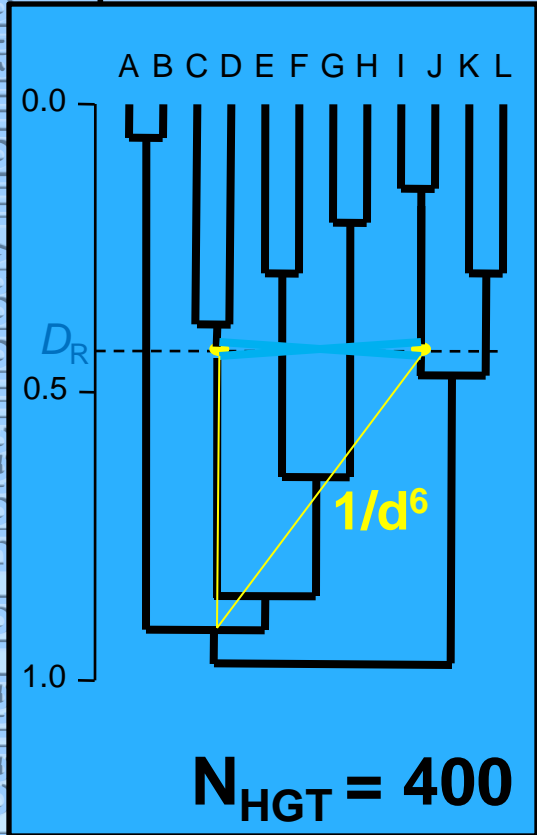


- $SS_{B/A} = 1$ in $\sim 65\%$ of NUTs.
- Mean Split Distance NUTs ~ 0.65 .

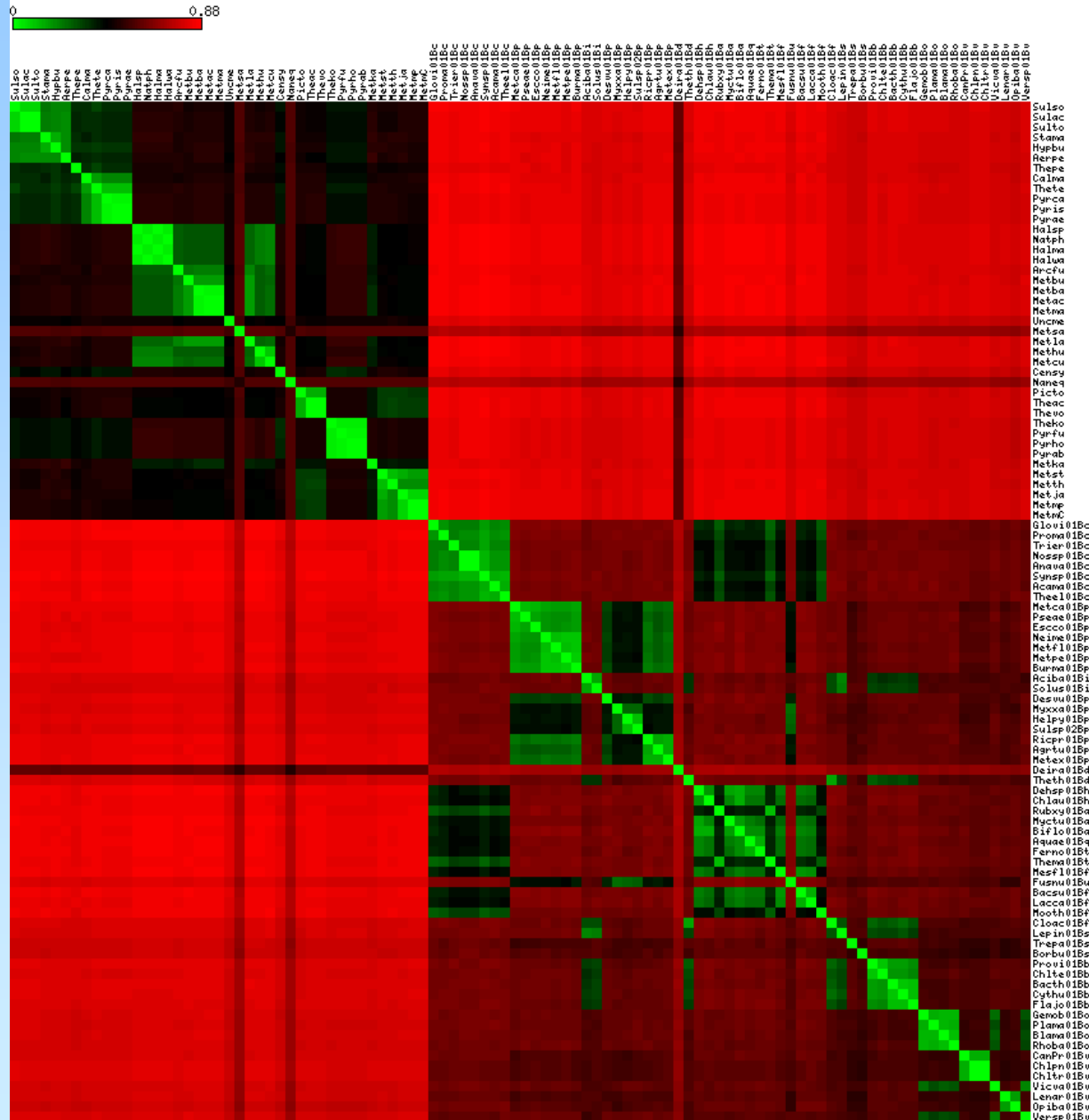
SIMULATIONS



Stationary simulation:
Keep matrix constant

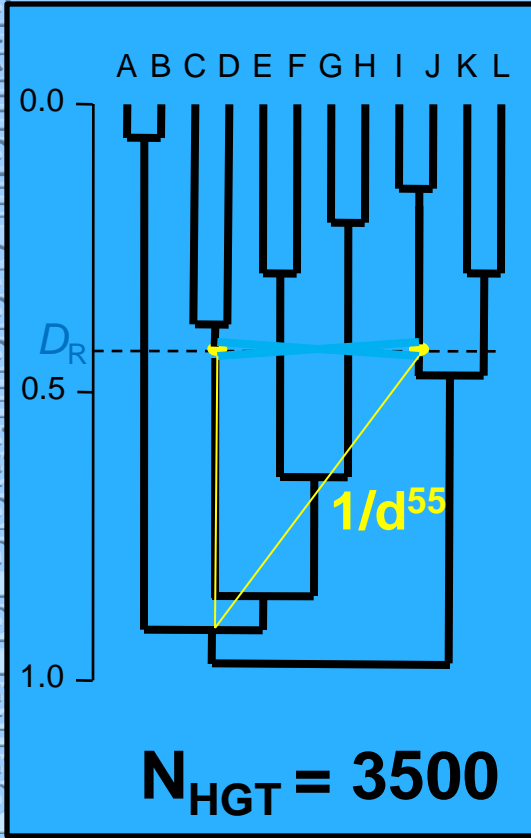


-moderate HGT,
moderate gradient
from tips to root,
Results similar to the
FOL matrix

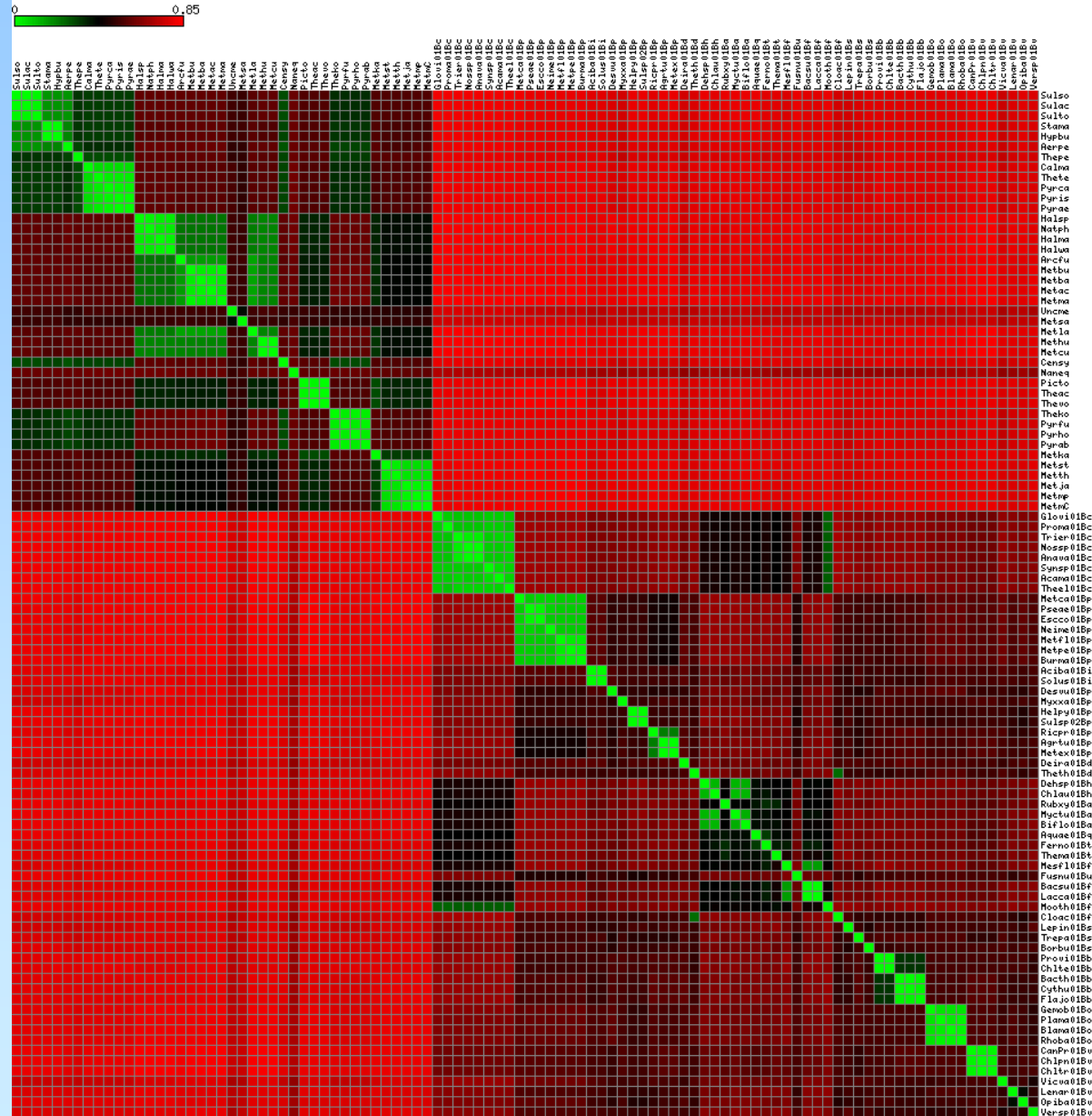


SIMULATIONS

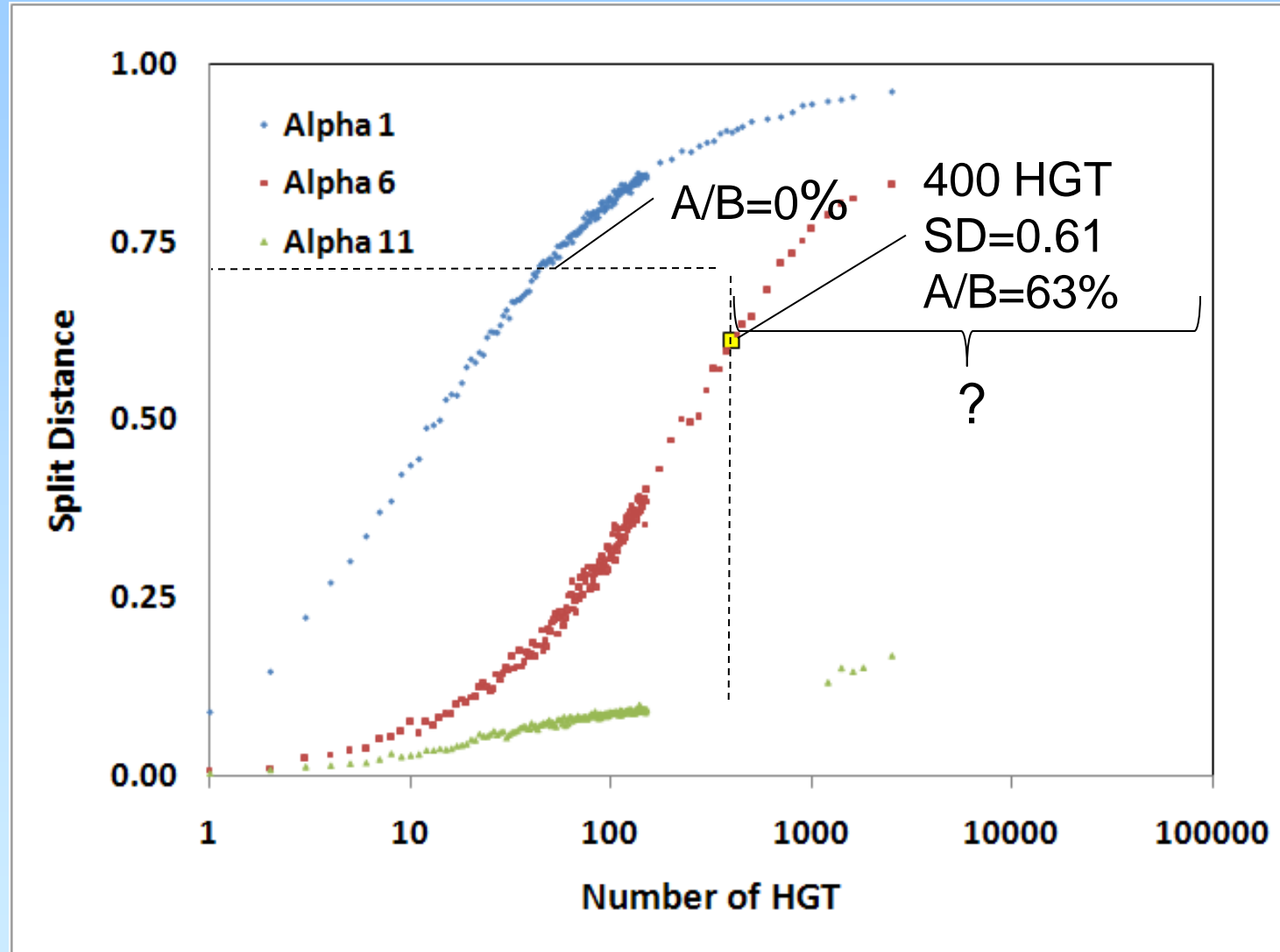
Dynamic simulation:
Change matrix each time



-very high rate of HGT between closely related species, hence many HGT events altogether
-almost no HGT between distant species

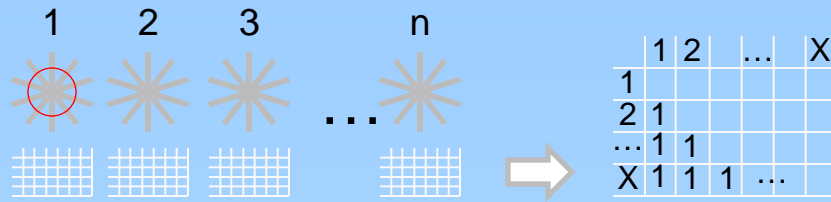


Exploration of the parameter space of HGT simulation: weak GDL conjecture holds

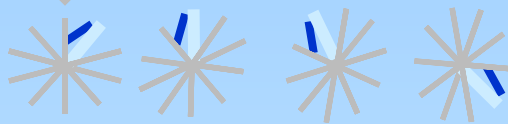


Strong GDL conjecture: no tree, just HGT gradient

$\alpha=1$



HGT 1

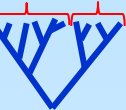


HGT N

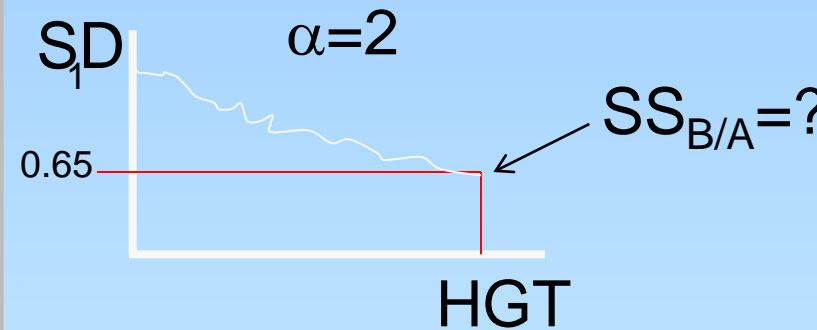
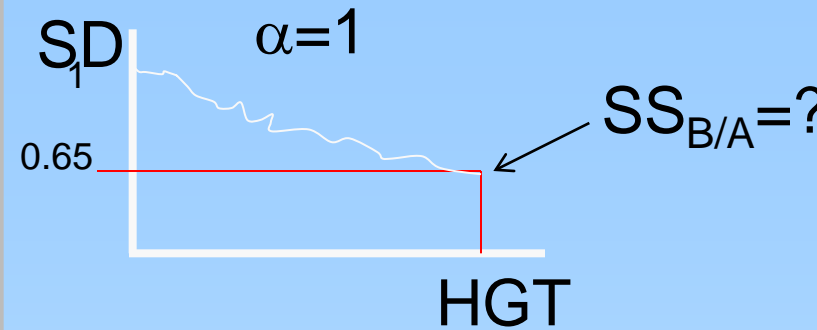


Supertree

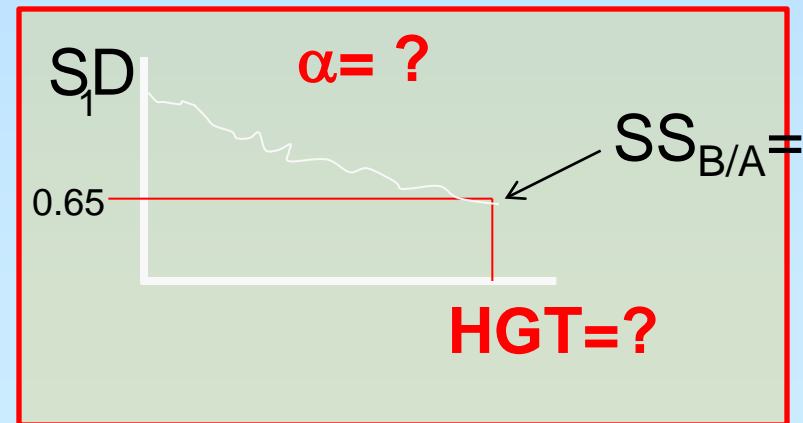
$SS_{B/A}$



Define 'Archaea' and 'Bacteria'

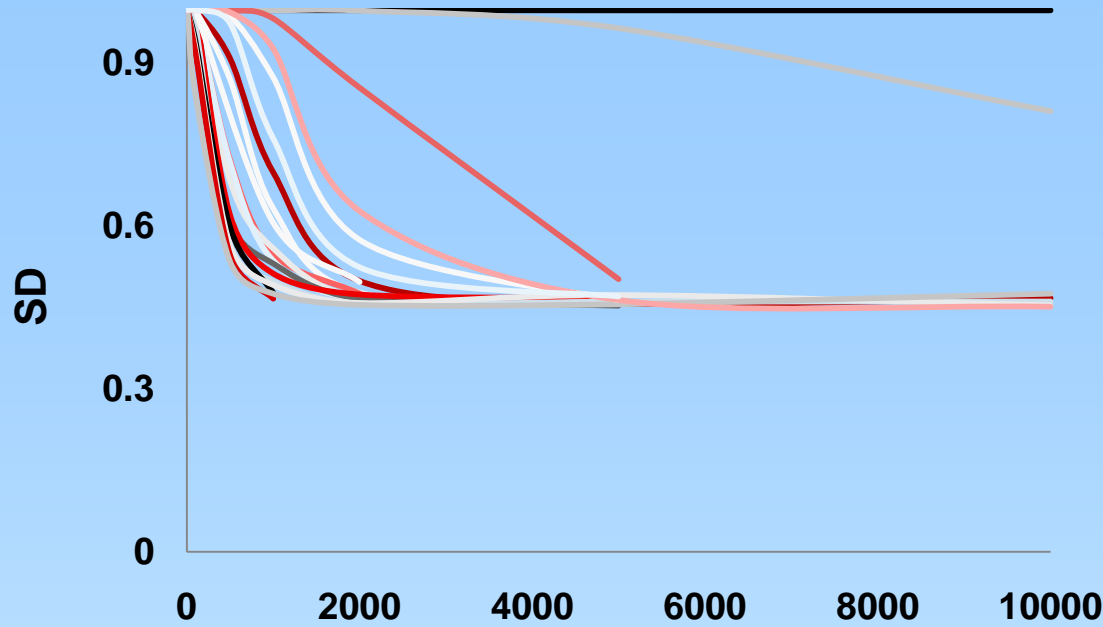


...



Strong GDL conjecture fails

otechnology Information

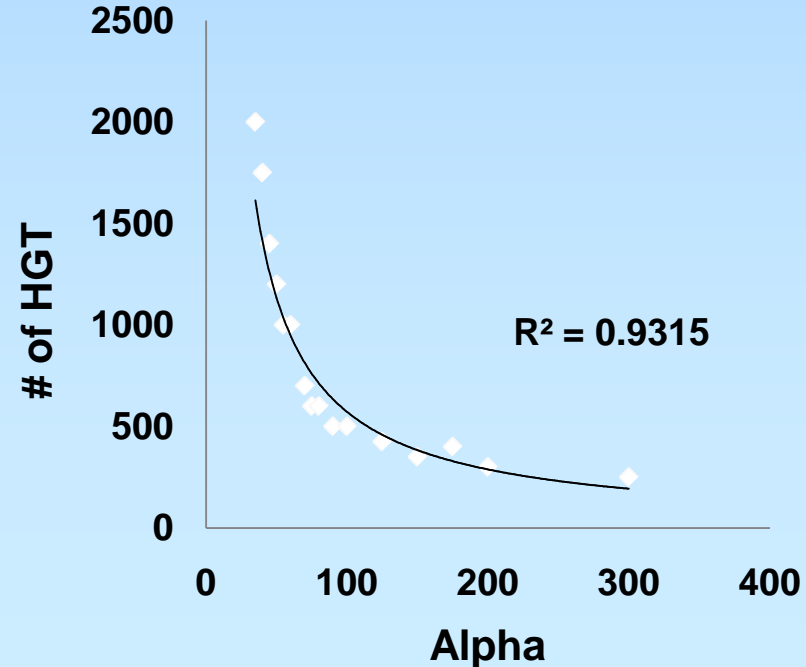


Alpha

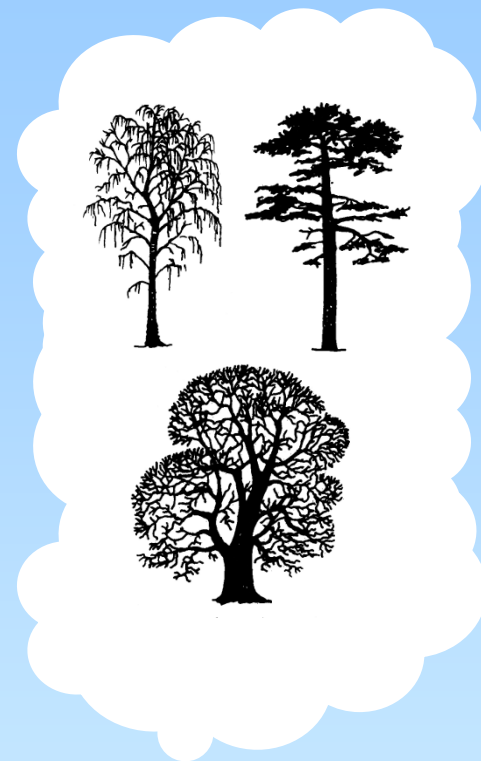
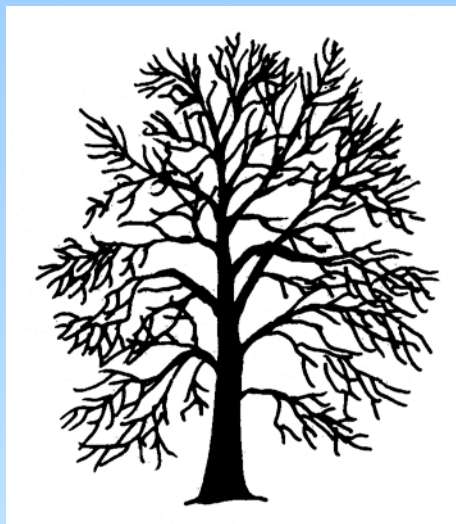
- 1
- 6
- 10
- 20
- 50
- 60
- 70
- 75
- 80
- 90
- 100
- 125
- 150
- 175
- 200
- 300
- 55
- 30
- 40
- 35
- 45

HGT

Alpha	# of HGT	Mean SD	% of trees with SSb/a=1
35	2000	0.62776424	0
40	1750	0.66168697	0
45	1400	0.61601791	0
50	1200	0.62177028	0
55	1000	0.64779965	0
60	1000	0.63291263	0
70	700	0.61033011	0
75	600	0.65819848	0
80	600	0.63711134	0
90	500	0.61186712	0
100	500	0.62150161	0
125	425	0.62970947	0
150	350	0.64022076	0
175	400	0.60737061	0
200	300	0.62646673	0
300	250	0.6414787	0



To see the forest for the trees...



...but also trees for the forest

The Take-Home Message on TOL

- There is no single "Tree of Life" describing the evolutionary history of all or even the majority of the prokaryote genomes
- Yet, there is a central tree-like trend of evolution compatible with a common history of descent of prokaryotic groups
- This trend is more evident at shallow phylogenetic depths, in more ubiquitous genes and among some functional categories of genes (eg, translation)
- Observations are compatible with the ancient divergence between the Bacterial and Archaea followed by explosive radiation of major phyla followed by HGT that distorted but did not destroy the tree-like signal
- **Altogether, HGT might dominate evolution but the tree-like signal is stronger than the signal from any particular route of HGT**

Puigbo P, Wolf YI, Koonin EV. (2009) *Search for a 'Tree of Life' in the thicket of the phylogenetic forest.* J. Biol. 8, 59

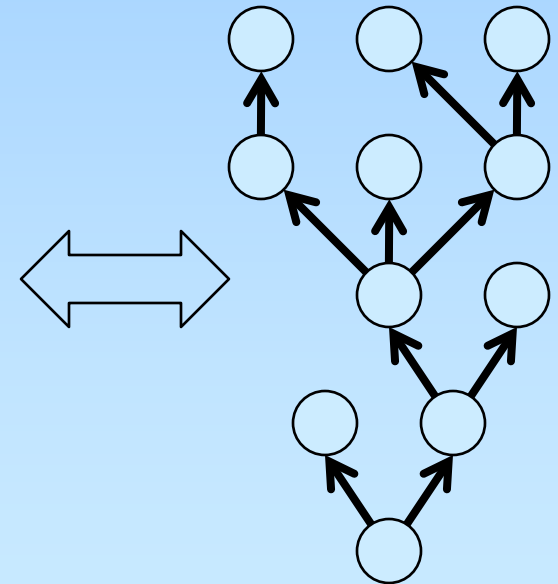
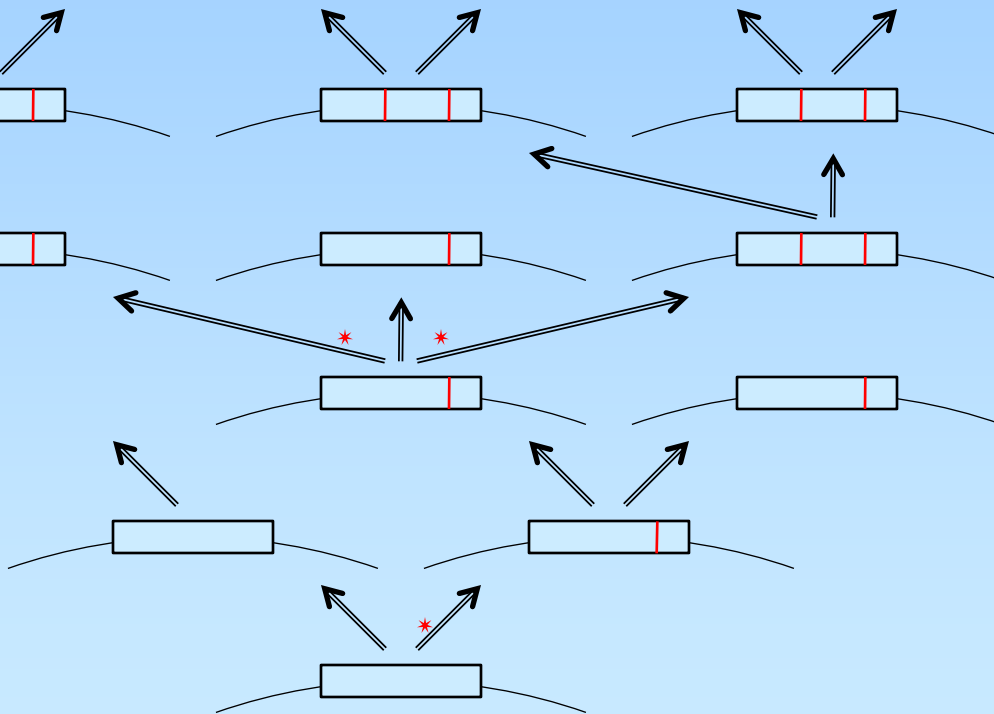
Koonin EV, Wolf YI, Puigbo P. (2009) *The Phylogenetic Forest and the Quest for the Elusive Tree of Life.* Cold Spring Harb Symp Quant Biol.

Koonin EV, Wolf YI. (2009) *The fundamental units, processes and patterns of evolution, and the tree of life conundrum.* Biol Direct. 4, 33

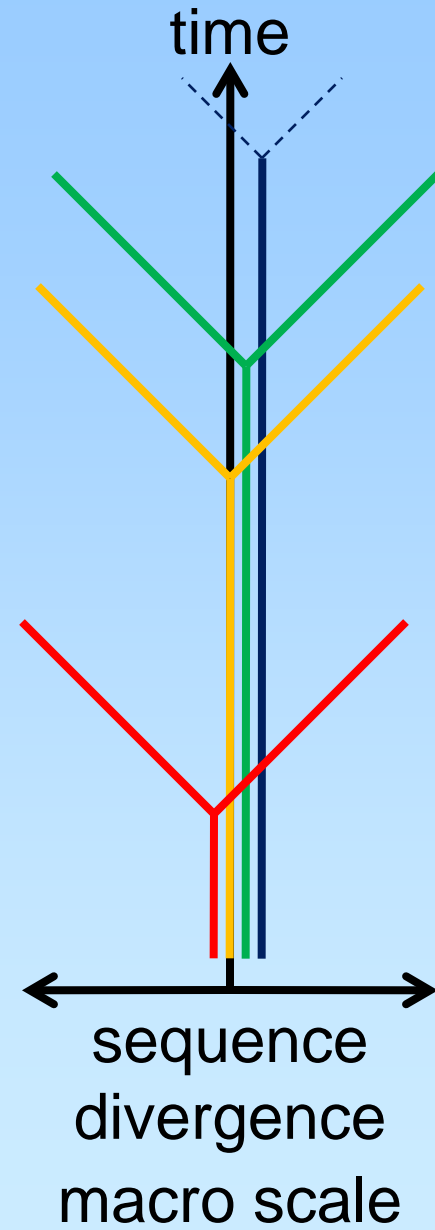
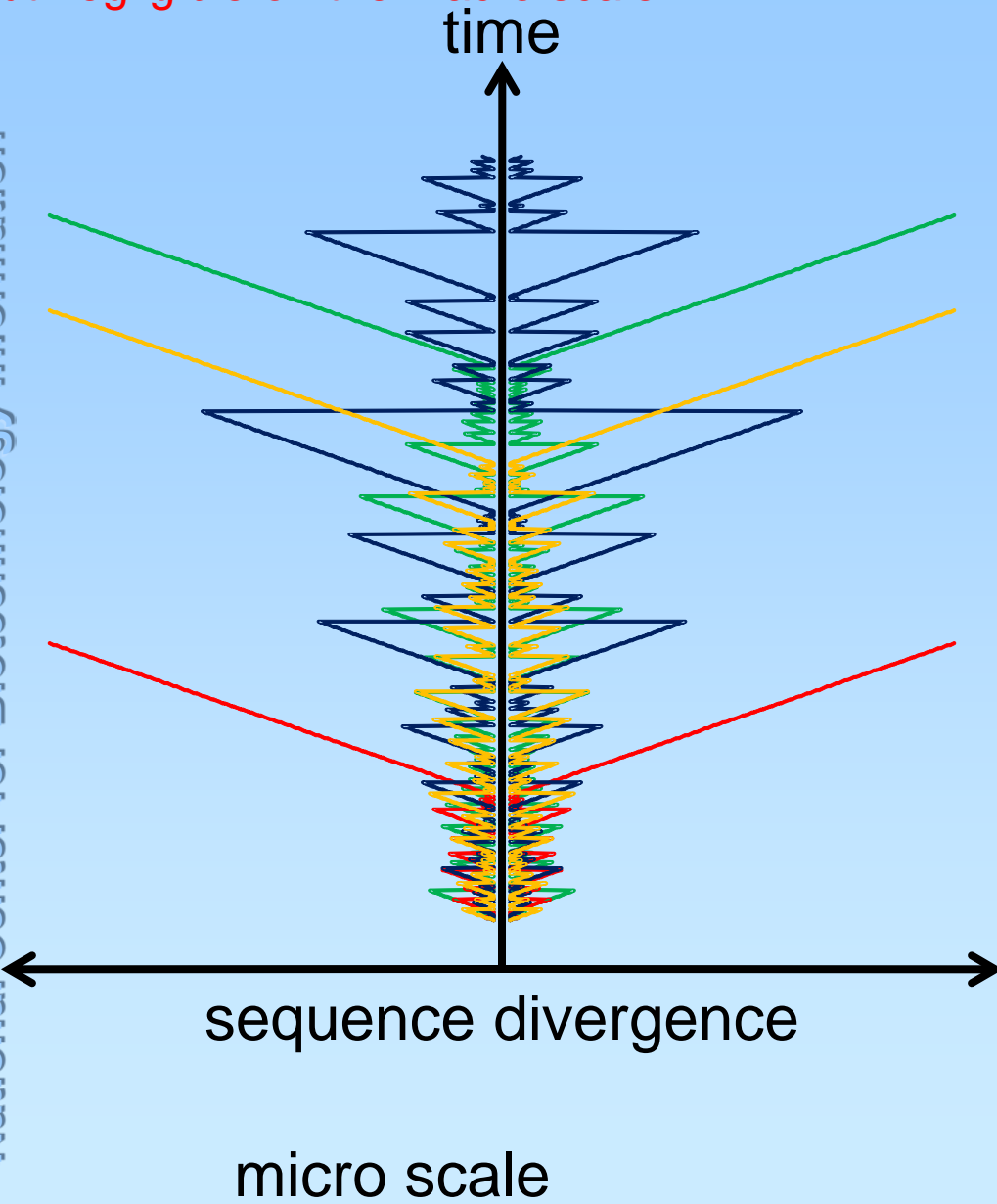
Puigbo P, Wolf YI, Koonin EV. (2010) *The tree and net signals in the evolution of prokaryotes.* GBE

Is tree thinking valid/necessary/relevant at all?

Yes: in the absence of intragenic recombination, the history of a replicating genetic element (gene) is isomorphously represented by a generalized tree graph



Intergenic recombination is important on the micro scale (homologous recombination) but negligible on the macro scale



So was Darwin wrong about the “tree simile”?

In principle, NO! The fundamental pattern of evolution IS tree-like.

Only, Darwin could not (for obvious reasons) correctly define the fundamental unit of evolution



Pere Puigbo



Yuri Wolf

Bill Martin
W. Ford Doolittle
J. Peter Gogarten
Maureen O'Malley