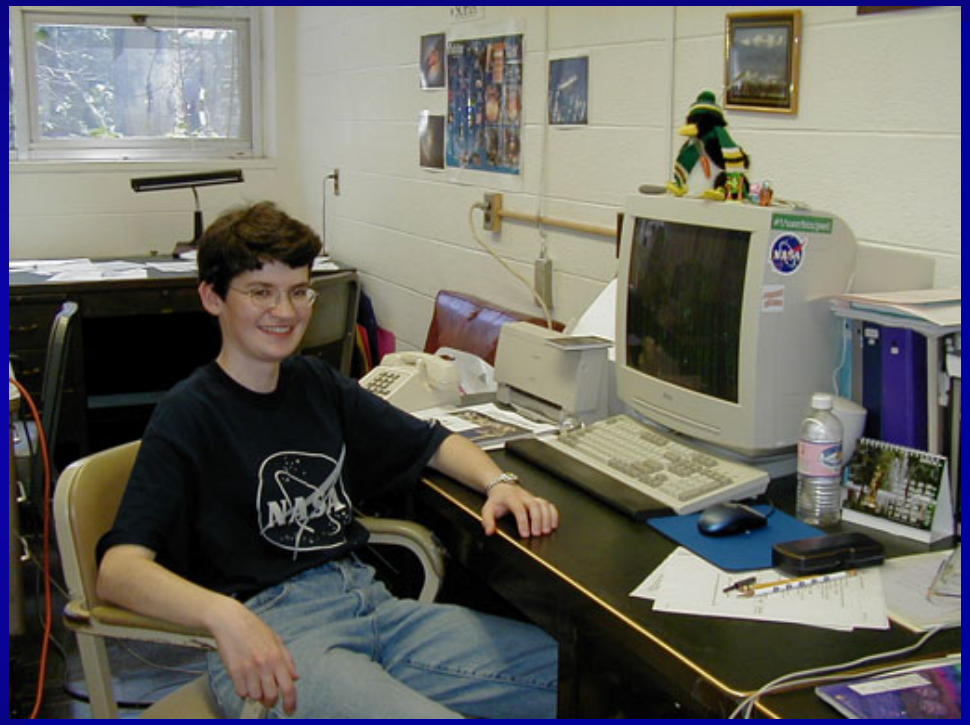
Probability Mapping and Bipartition Analysis to Study Genome Histories

J. Peter Gogarten
and
Olga Zhaxybayeva
Dept. of Molecular and Cell Biology, Univ. of Connecticut

DIMACS Workshop on Reticulated Evolution, DIMACS Center, Rutgers University, September 20 - 22, 2004

Acknowledgements

Olga Zhaxybayeva:



HGT:

Lutz Hamel (URI)
Paul Lewis (UConn)
Robert Blankenship (ASU)
Jason Raymond (ASU)
Ford Doolittle (Dalhousie)
Jeffery Lawrence (Pittsburgh)
Gary Olsen (Urbana)

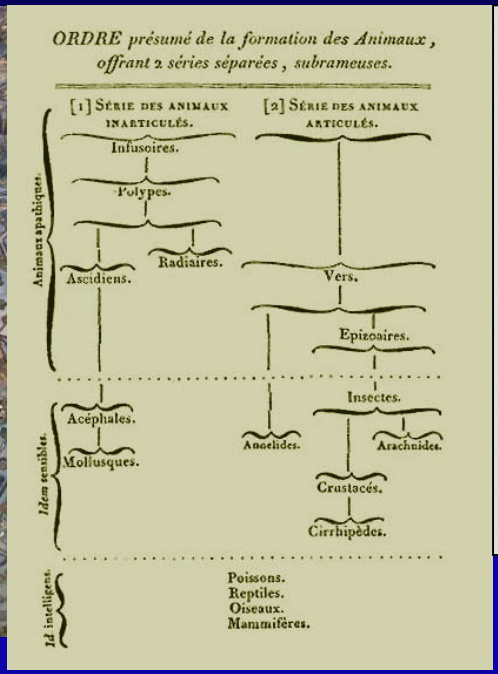
Coalescence:

Andrew Martin (U of C Boulder)
Joe Felsenstein (U of Wash)
Hyman Hartman (MIT)
Yuri Wolf (NCBI)

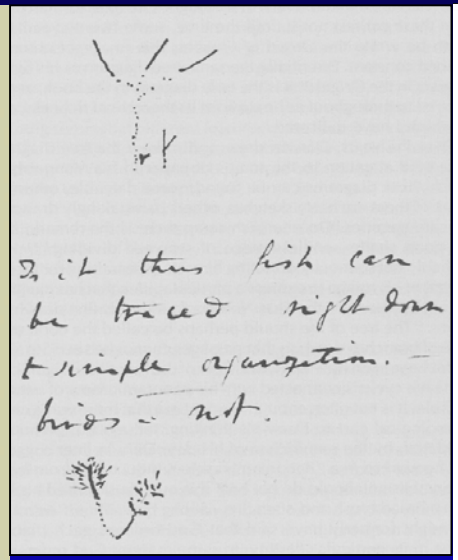
NASA Exobiology Program
NSF Microbial Genetics

Trees as a Visualization of Evolution

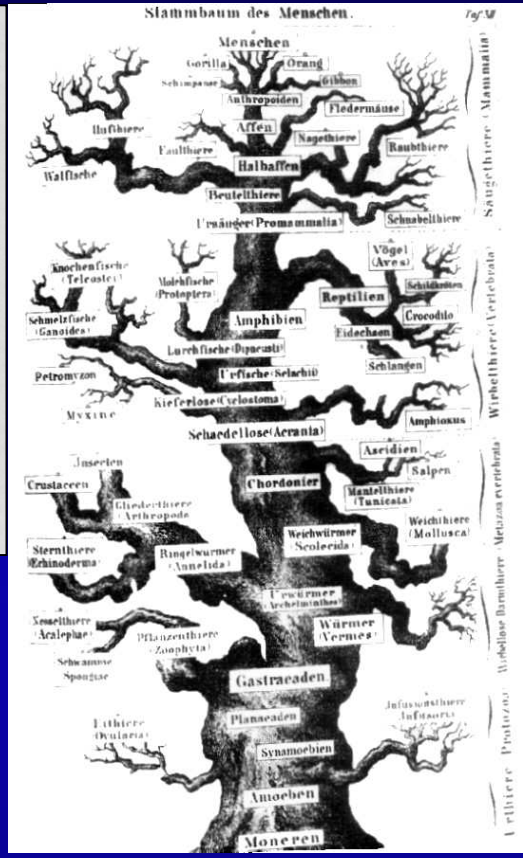
Genealogy (Church Ceiling, Santo Domingo, Oaxaca)



Lamarck's Tree of Life (1815)



Page B26 from Charles Darwin's (1809-1882) notebook (1837):



Lebensbaum (German for "Tree of Life") from Ernst Haeckel, 1874

SSU-rRNA Tree of Life

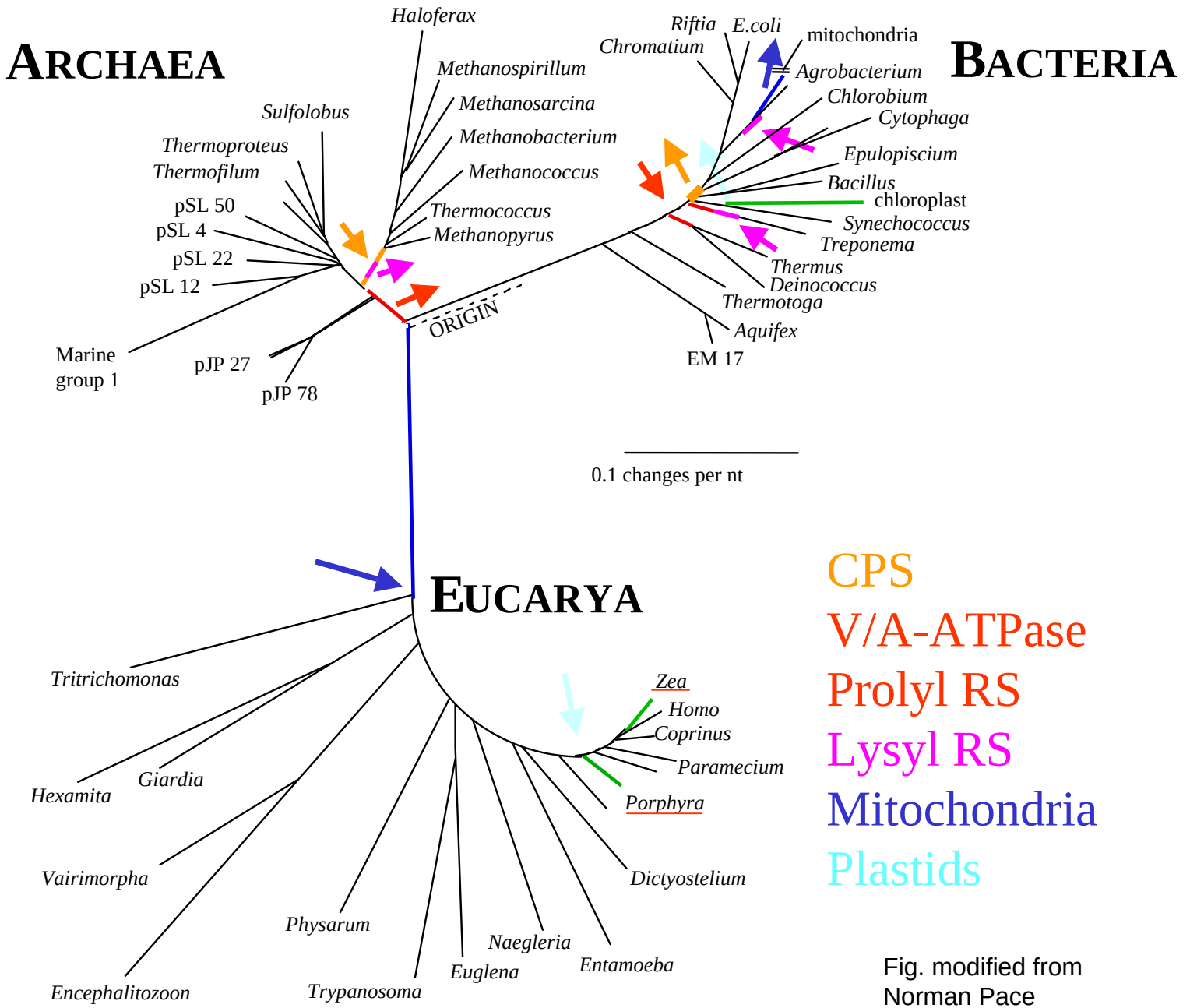
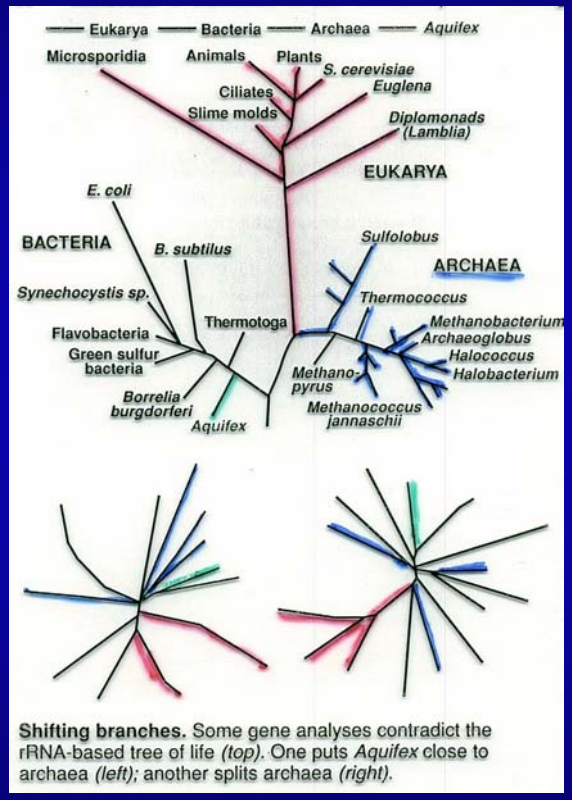


Fig. modified from Norman Pace

RESEARCH NEWS

Genome Data Shake Tree of Life

New genome sequences are mystifying evolutionary biologists by revealing unexpected connections between microbes thought to have diverged hundreds of millions of years ago



Shifting branches. Some gene analyses contradict the rRNA-based tree of life *(top)*. One puts *Aquifex* close to archaea *(left)*; another splits archaea *(right)*.

Science, **280** p.672ff (1998)

Horizontal Gene Transfer leads to Mosaic Genomes, where different parts of the genome have different histories.

Publicly Available Prokaryotic Genomes:

181 - completed

236 - in progress

(as of September 8, 2004)

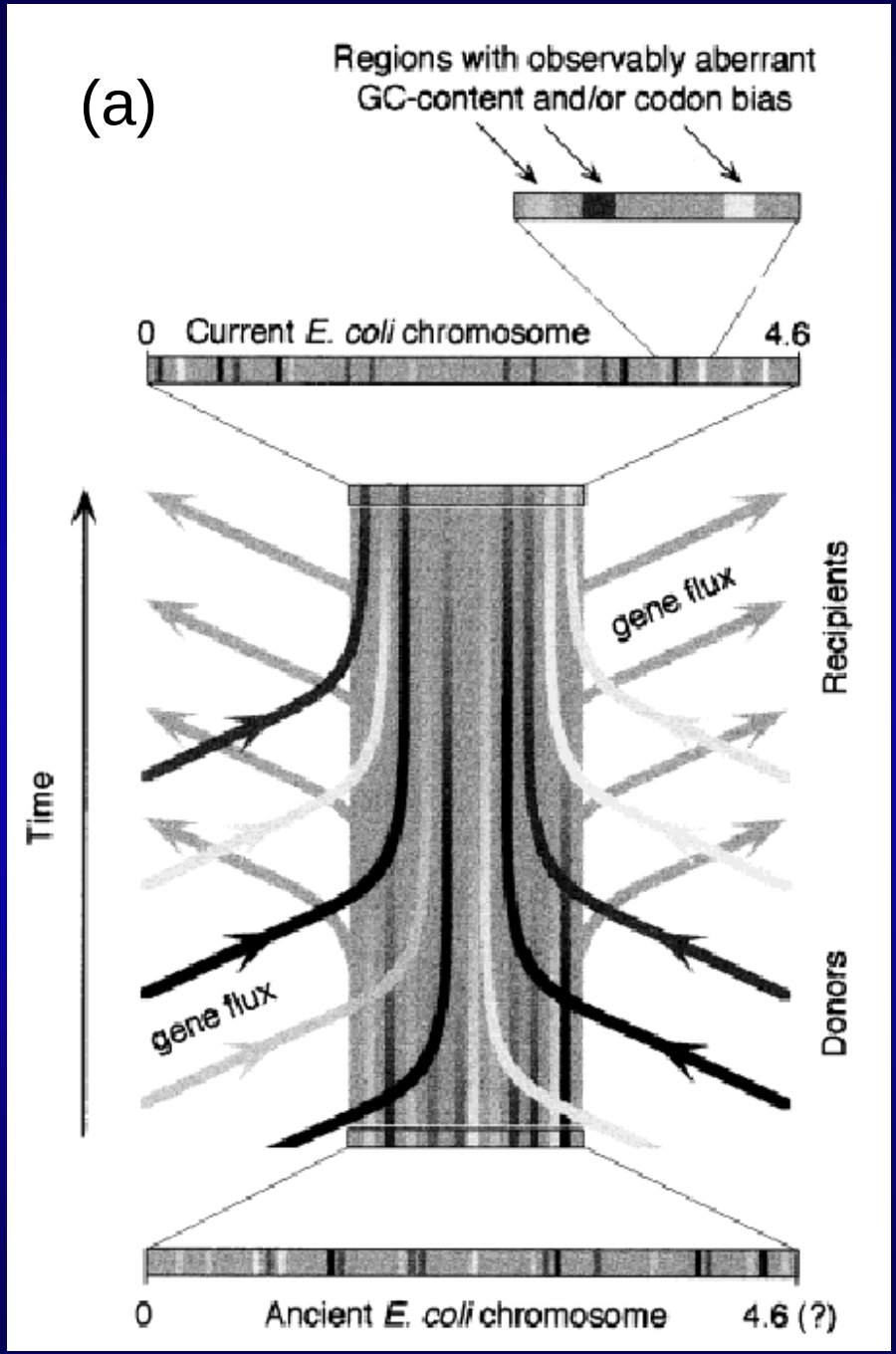
Transferred genes can be detected using:

(a) unusual composition,

(b) the comparison between closely related species, or

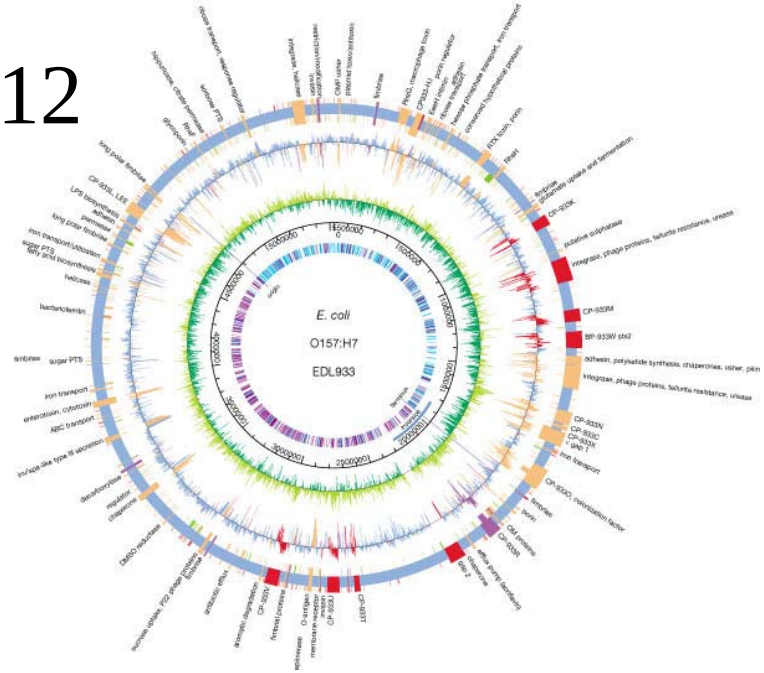
(c) conflicting molecular phylogenies.

From Bill Martin
BioEssays 21 (2), 99-104.



E. coli O157:H7 *versus* *E. coli* K12

- divergence about 4.5 million years ago



"We find that lateral gene transfer is far more extensive than previously anticipated. In fact, **1,387 new genes encoded in strain-specific clusters** of diverse sizes were found in O157:H7."

Common: **4,100,000** bp; **3,574** protein-coding genes
(about 95% identical each on the nucleotide level)

Only in O157:H7: **1,340,000** bp; **1,387** protein-coding genes

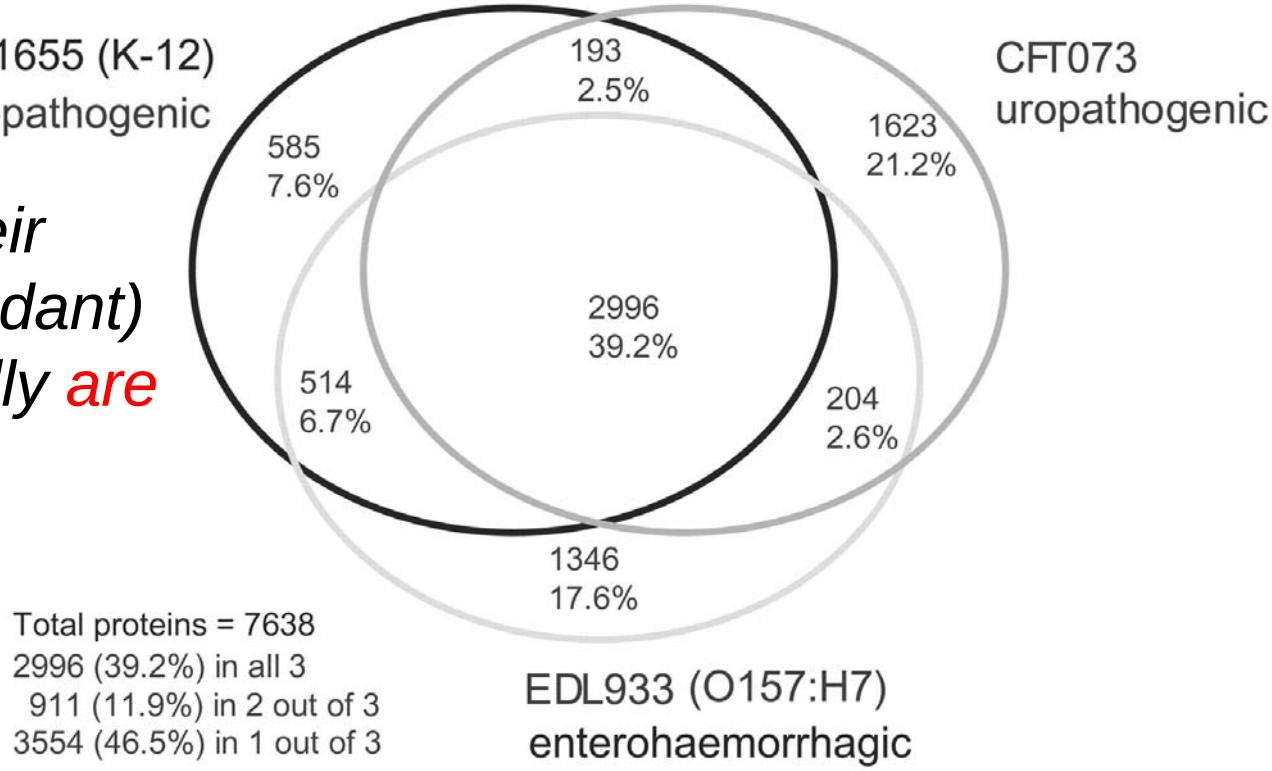
Only in K12: **530,000** bp, **528** protein-coding genes

From: **Perna *et al.*** (2001) Nature 409: 529-33
see also **Hayashi *et al.*** (2001) DNA Res. 8:11-22

Welch RA, *et al*.
Proc Natl Acad Sci U S A. 2002; 99:17020-4

Escherichia coli, strain CFT073, uropathogenic
Escherichia coli, strain EDL933, enterohemorrhagic
Escherichia coli K12, strain MG1655, laboratory strain,

"… only 39.2% of their combined (nonredundant) set of proteins actually are common to all three strains."

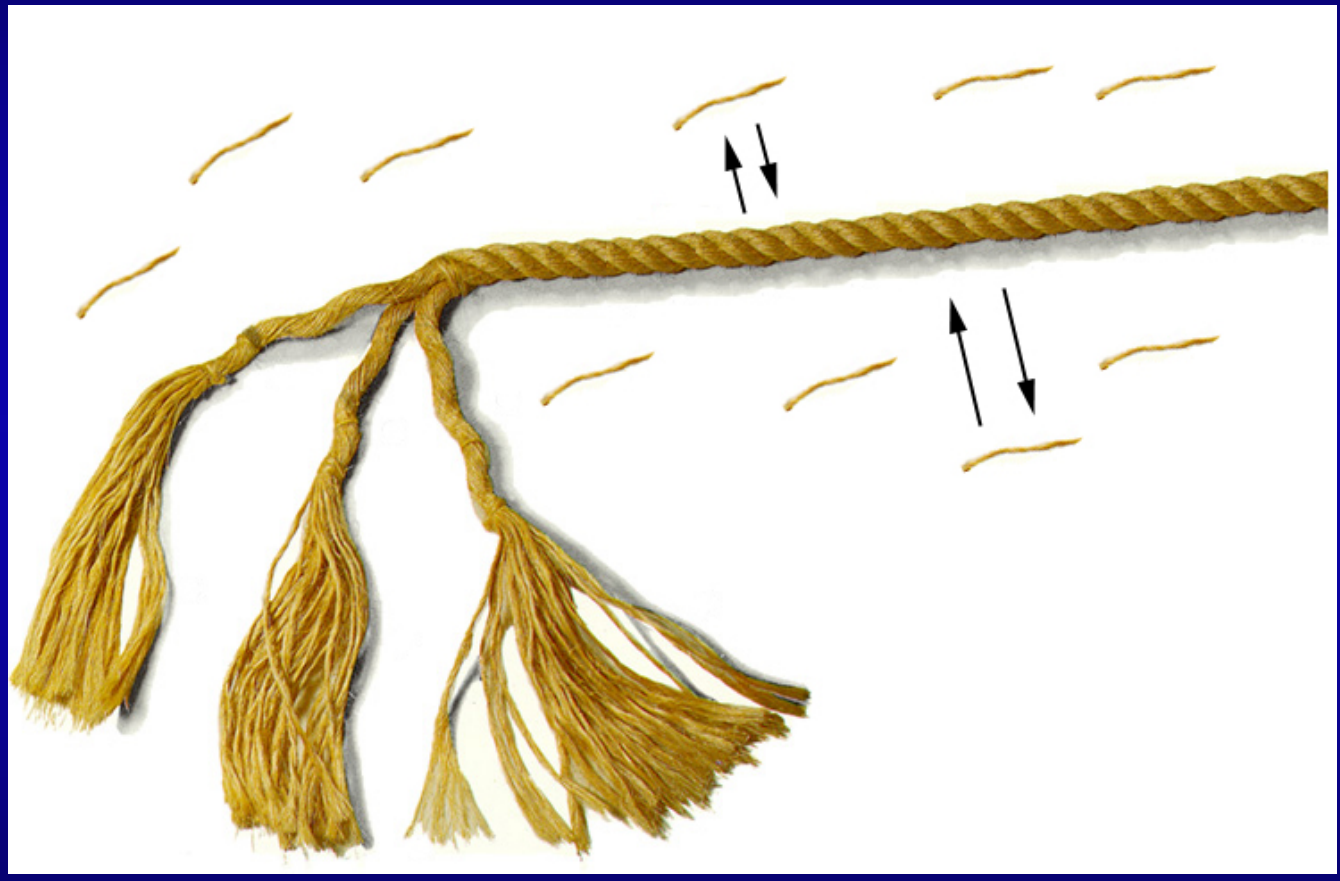


What is an “organismal lineage” in light of horizontal gene transfer?

Over very **short** time intervals an organismal lineage can be defined as the majority consensus of genes.
This definition only “fails”, if two organisms make co-equal contributions (e.g. endosymbiosis).

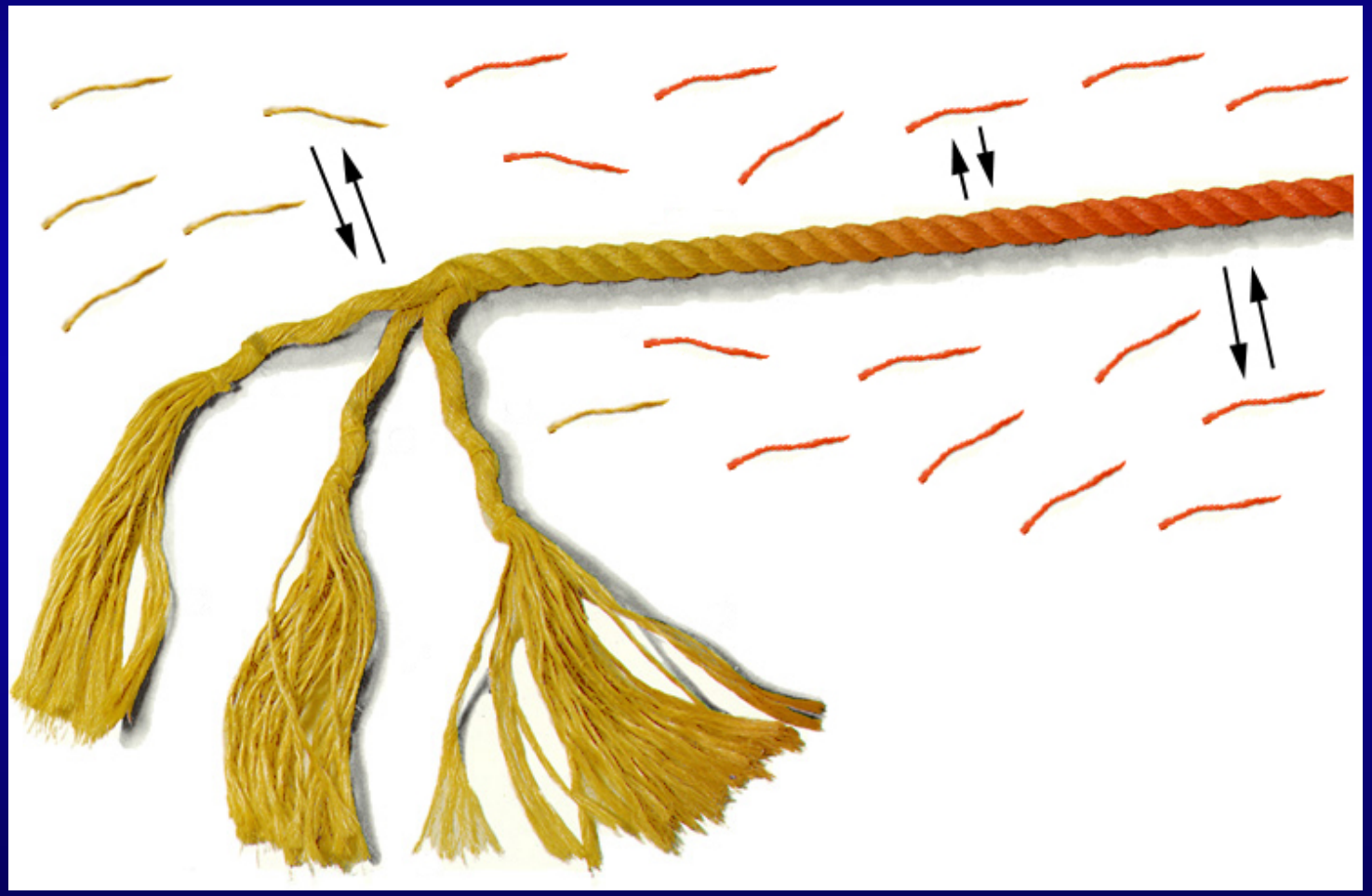
Rope as a metaphor to describe an organismal lineage *(Gary Olsen)*

Individual fibers = genes that travel for some time in a lineage.

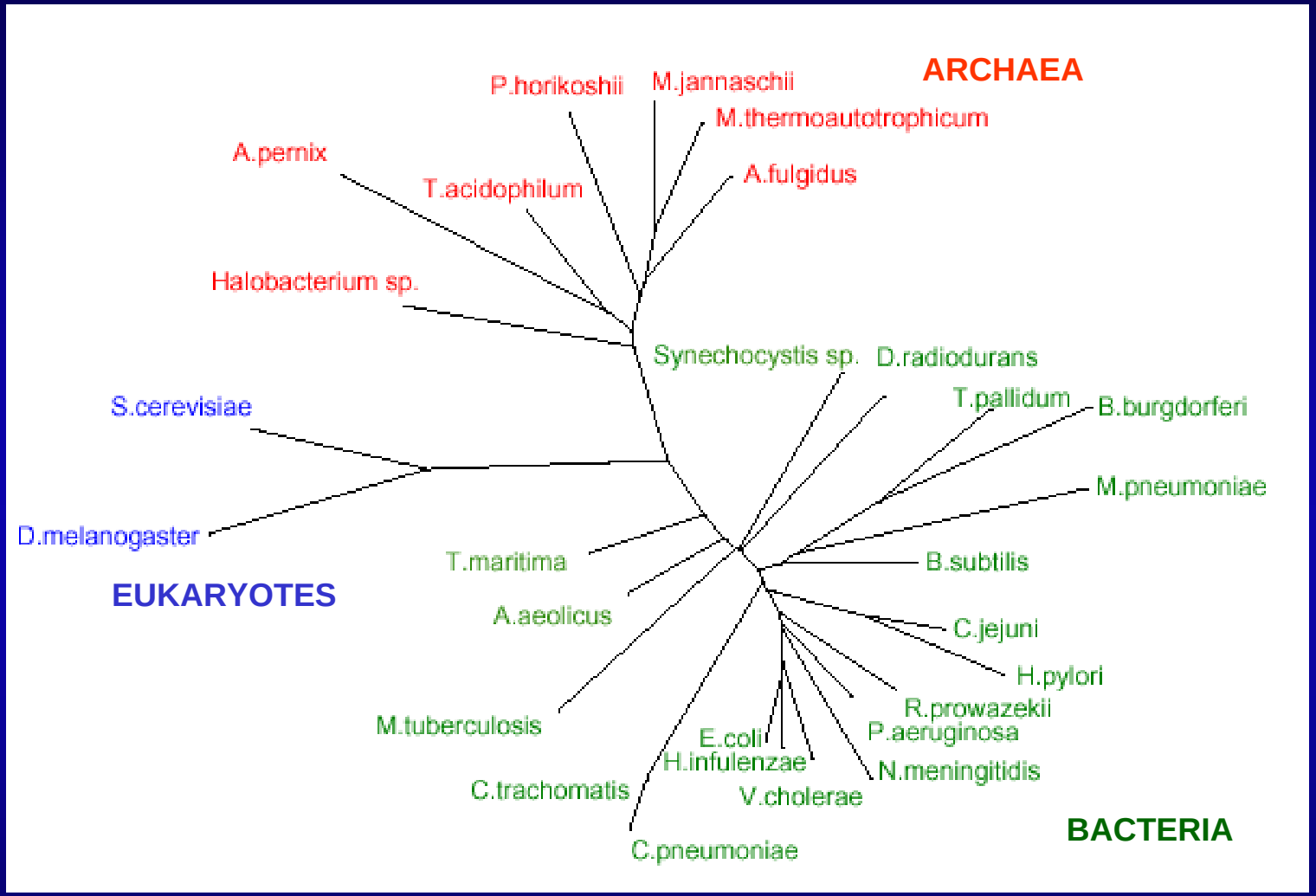


While no individual fiber present at the beginning might be present at the end, the rope (or the organismal lineage) nevertheless has continuity.

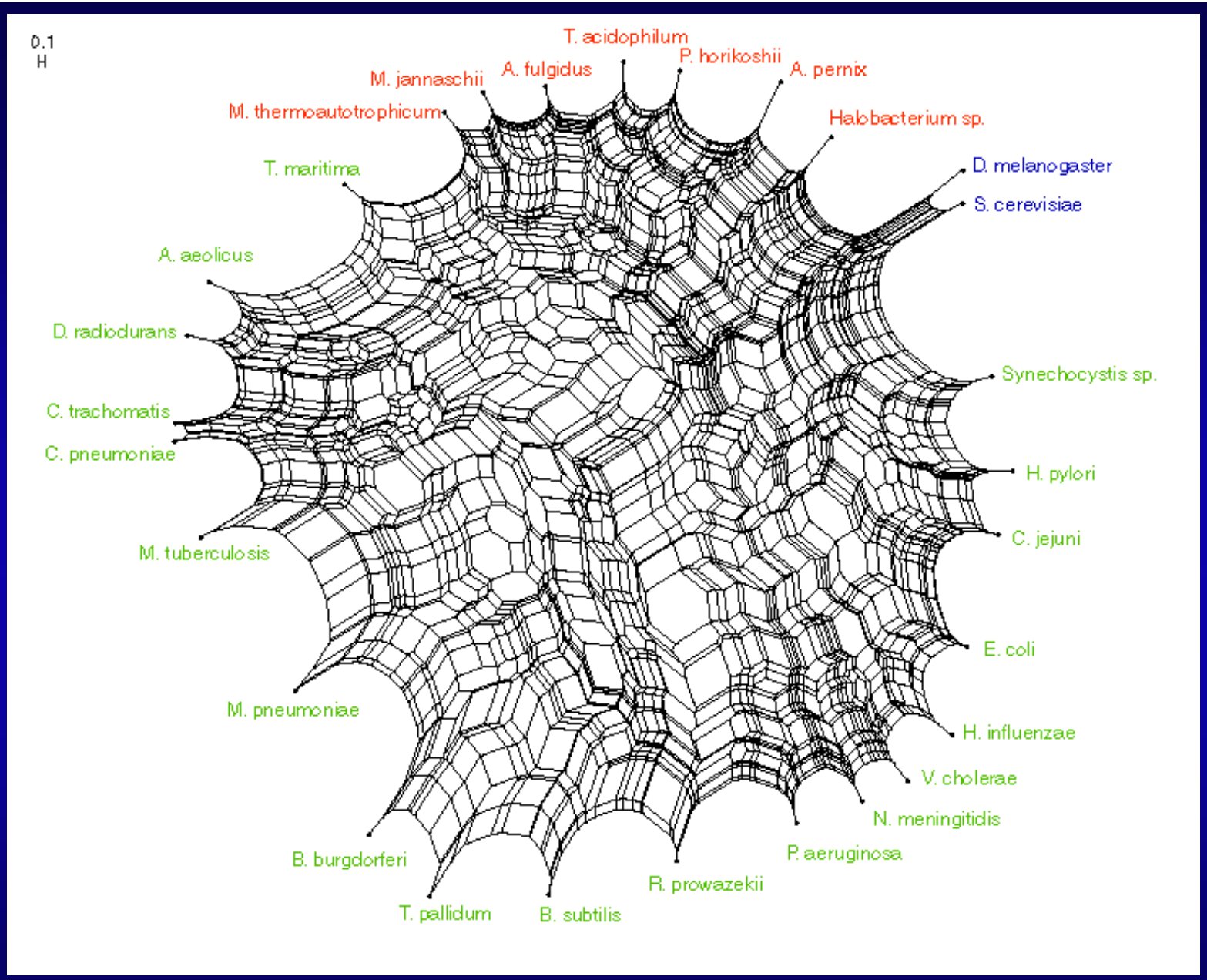
However, the genome as a whole will acquire the character of the incoming genes (the rope turns solidly red over time).



Genome Content Tree



Other genome content trees: Tekaia et al. (1999) *Genome Res* **9:**550- 557; Snel et al. (1999) *Nat Genet* **21:**108-110; Lin & Gerstein (2000) *Genome Res* **10:**808-818; Fitz-Gibbon & House (1999) *Nucleic Acids Res* **27:**4218-4222 and (2002) *J Mol Evol* **54**:539-47; Charlebois et al. (2003) *Nature* **421**:217; Wolf et al. (2001), *BMC Evol. Biol* **1**:8



Same data as before, but network calculated using NeighborNet (David Bryant 2002, http://www.mcb.mcgill.ca/~bryant/NeighborNet/)

Visualization of Mosaic Genome Content

Bayes' Theorem

Reverend Thomas Bayes
(1702-1761)

$$P(model|data, I) = P(model, I) \frac{P(data|model, I)}{P(data,I)}$$

Likelihood

describes how well the model predicts the data

Posterior Probability

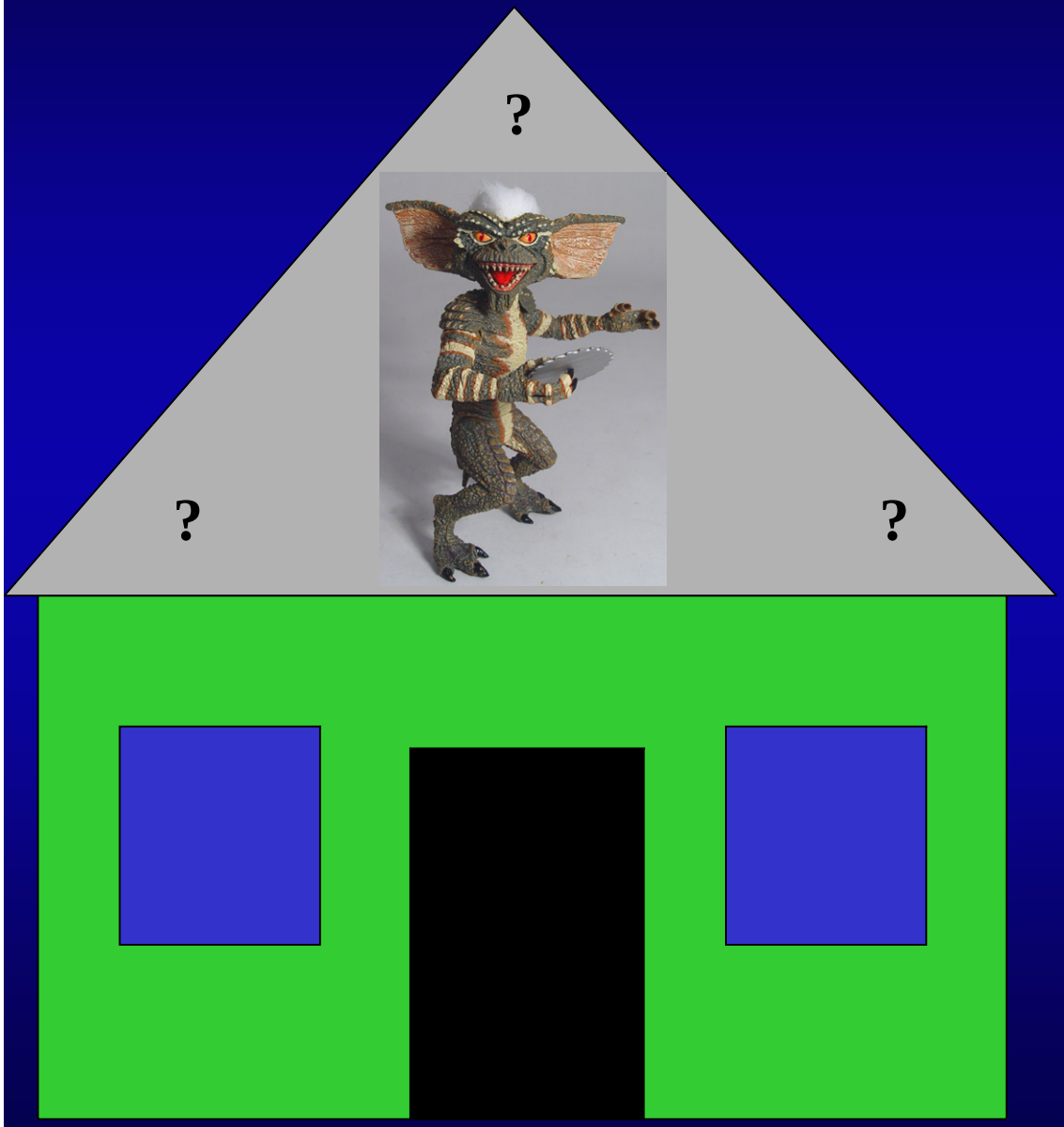
represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

Elliot Sober's Gremlins



Observation: Loud noise in the attic

Hypothesis: *gremlins in the attic playing bowling*

Likelihood =
P(noise|gremlins in the attic)
very high

Posterior Probability =
P(gremlins in the attic|noise)
very low

ML Mapping

(Strimmer and von Haeseler, 1997)

Data: Alignment of four sequences

Hypotheses: All possible unrooted tree topologies

T_1, T_2, T_3

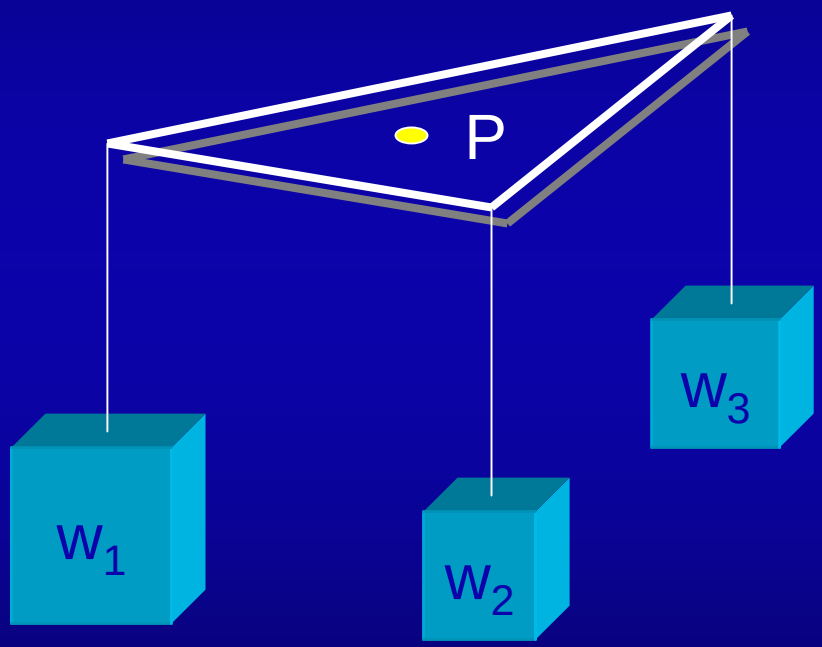
Prior: Equal Probabilities

For each set of 4 sequences:

- Calculate maximum-likelihood L_i for each tree T_i
- Calculate posterior probabilities p_i for each tree T_i
- Plot the point (p_1, p_2, p_3) into equilateral triangle

Barycentric Coordinates

(August Ferdinand Möbius, 1827)



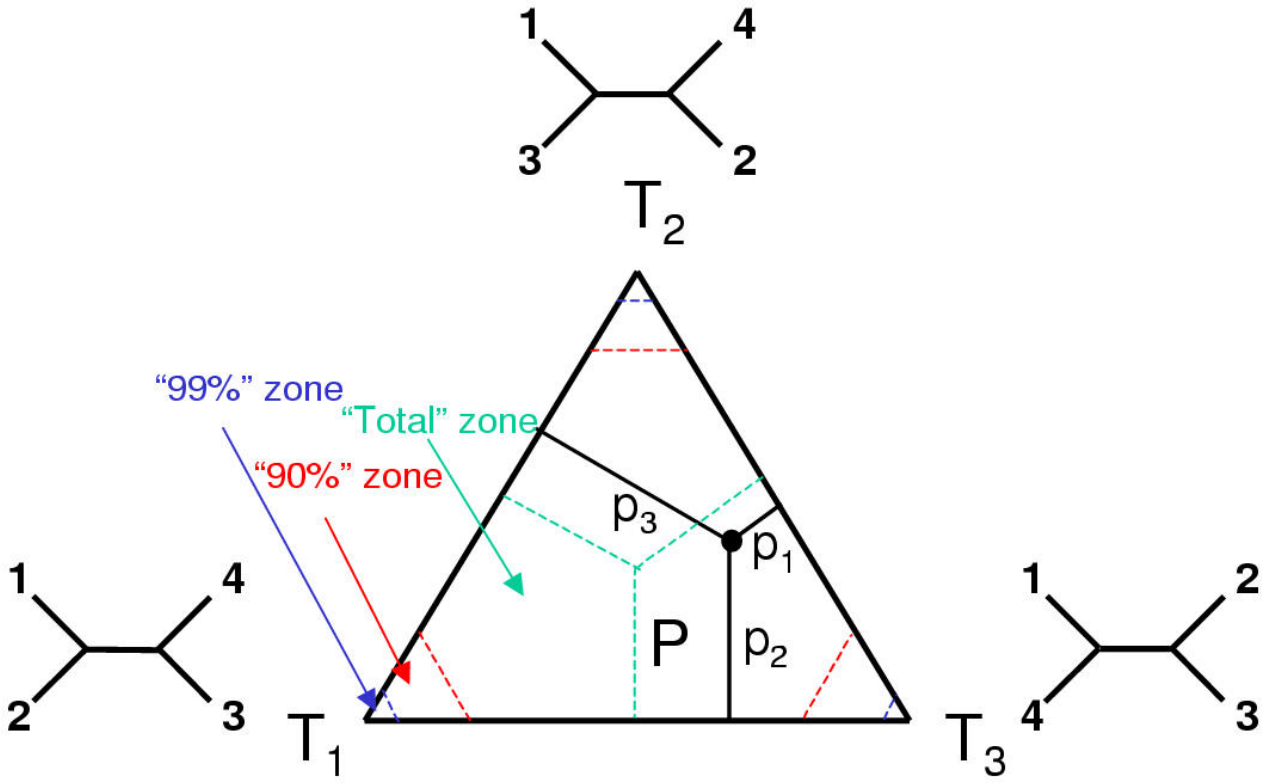
P : barycenter=center of gravity

For any point P inside the triangle, there exist masses w_1, w_2, w_3 such that if placed at the corresponding vertices of the triangle, their center of gravity will coincide with point P.

Barycentric coordinates are defined uniquely for every point inside the triangle (given that $w_1+w_2+w_3=1$) .

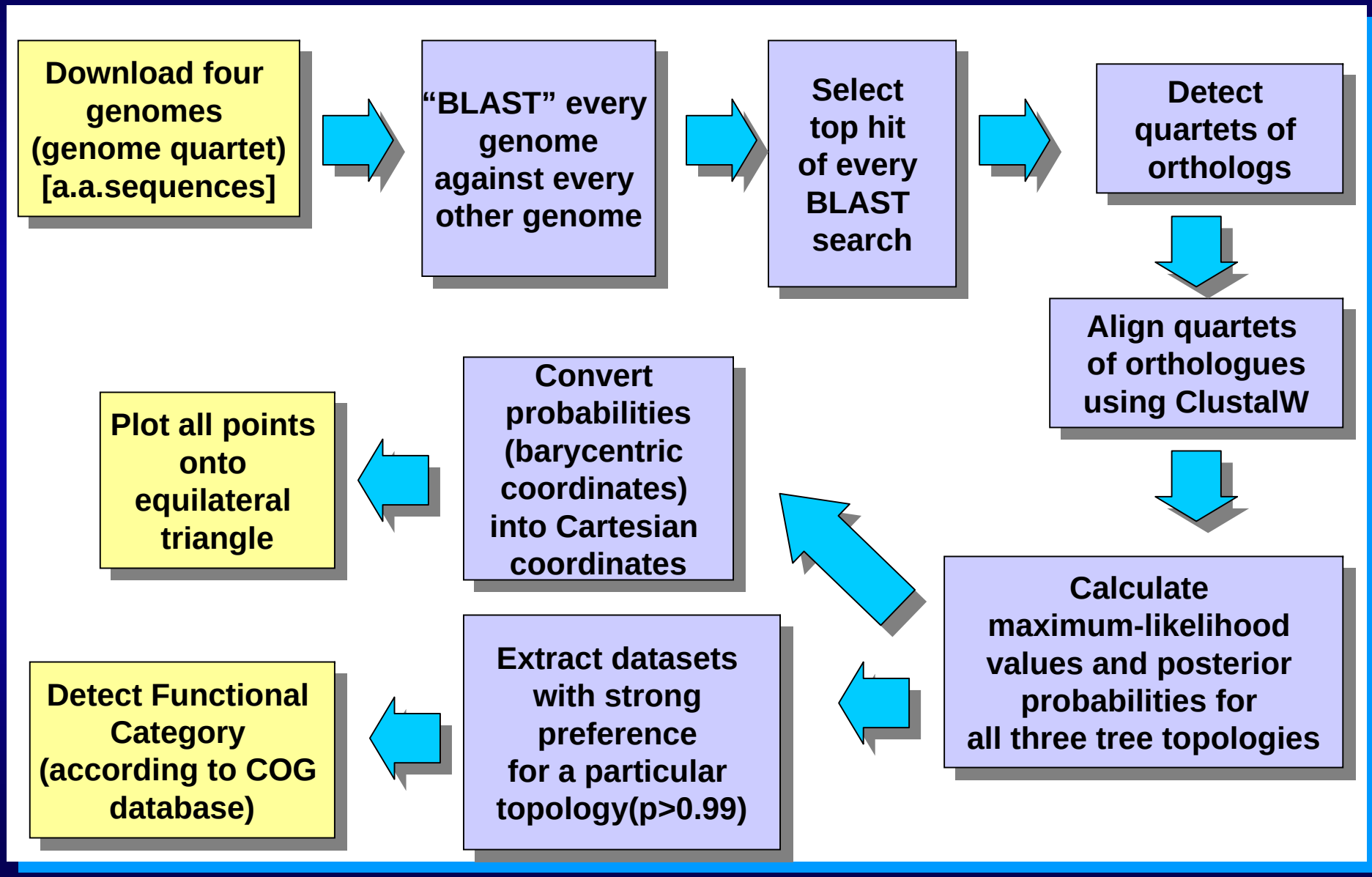
ML Mapping

(Fig. modified from Strimmer)



p_1, p_2 and p_3 are barycentric coordinates of point P

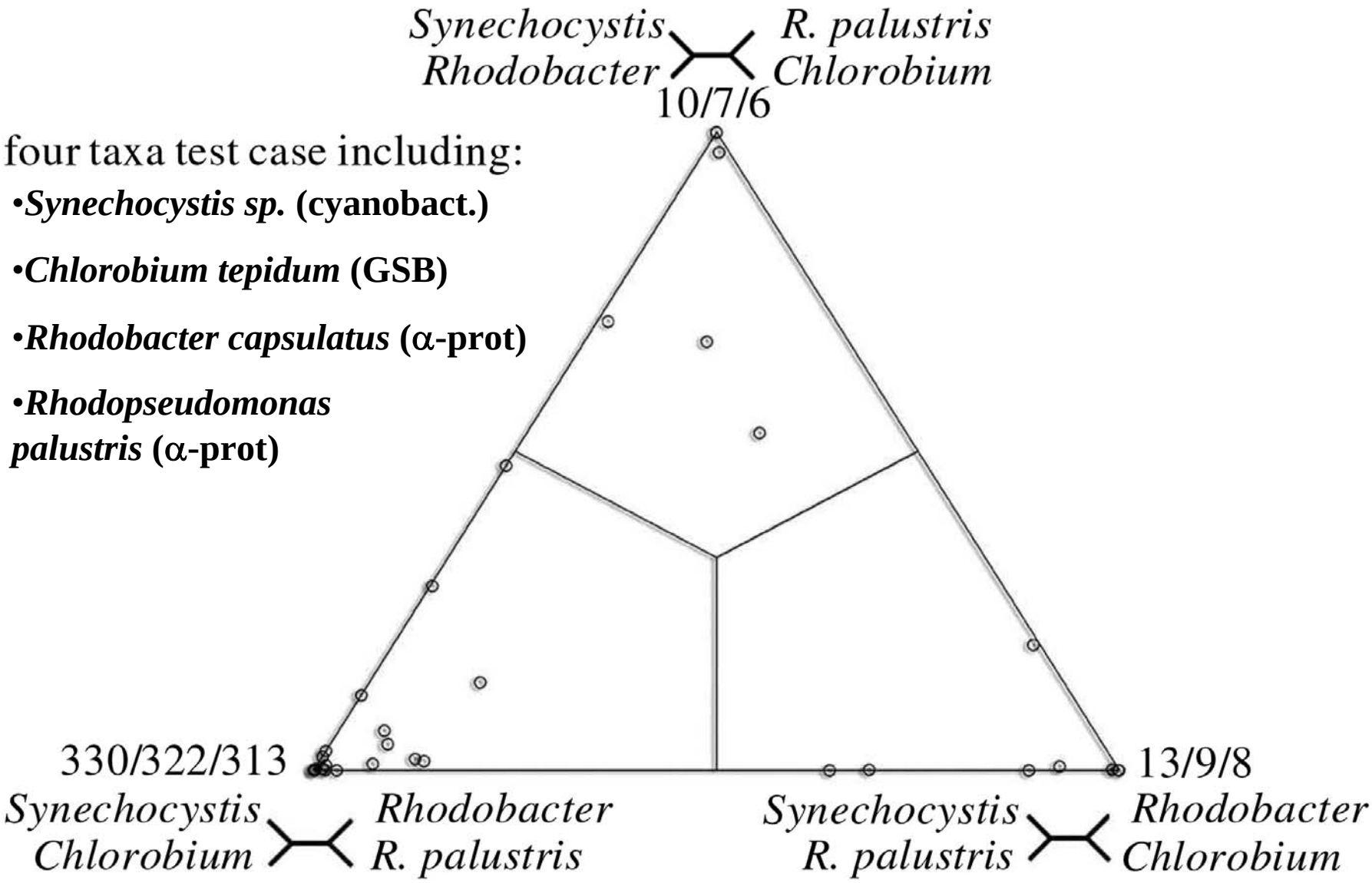
Data Flow


Download four genomes (genome quartet) [a.a.sequences]
"BLAST" every genome against every other genome
Select top hit of every BLAST search
Detect quartets of orthologs
Align quartets of orthologues using ClustalW
Calculate maximum-likelihood values and posterior probabilities for all three tree topologies
Extract datasets with strong preference for a particular topology(p>0.99)
Detect Functional Category (according to COG database)
Convert probabilities (barycentric coordinates) into Cartesian coordinates
Plot all points onto equilateral triangle

TEST CASE

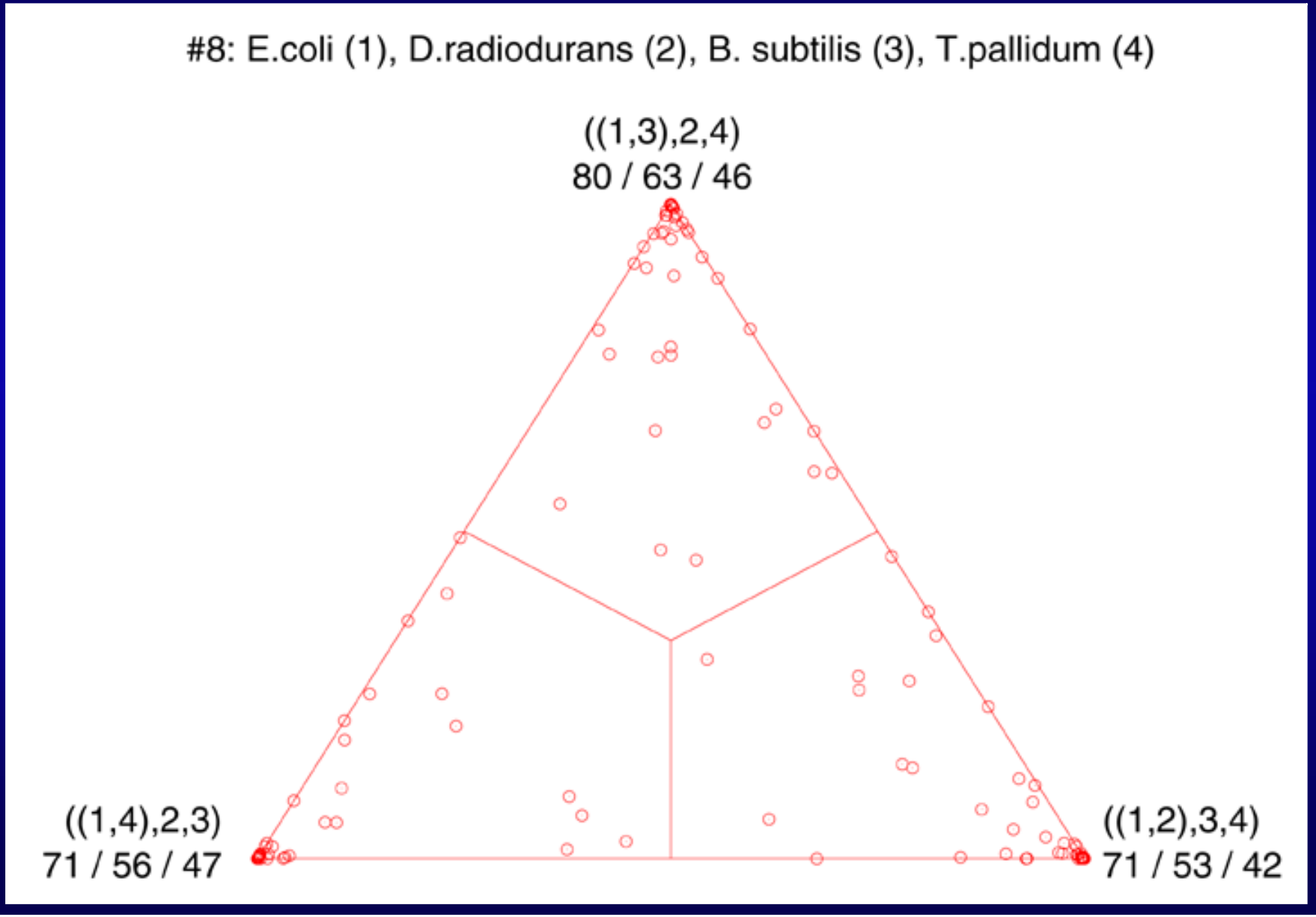
four taxa test case including:

- ***Synechocystis sp.*** **(cyanobact.)**
- ***Chlorobium tepidum*** **(GSB)**
- ***Rhodobacter capsulatus*** **(α-prot)**
- ***Rhodopseudomonas palustris*** **(α-prot)**



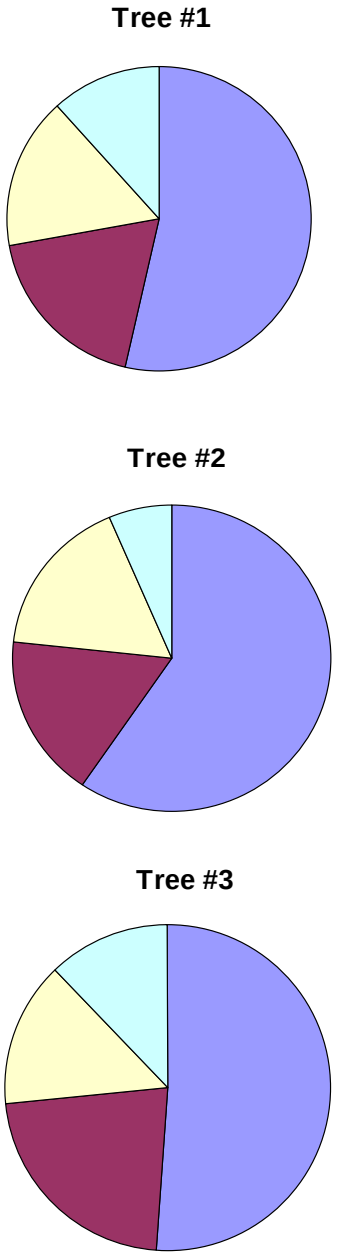
Raymond, Zhaxybayeva, Gogarten, Blankenship, *Phil. Trans. R. Soc. Lond. B* 2003, **358**: 223-230.

Inter-phylum relationships (bacteria) - there is no obvious core



Zhaxybayeva and Gogarten, *BMC Genomics* 2002, **3**:4

			#8	
Functional Categories of COGs:		**1**	**2**	**3**
Information storage and processing		23	28	25
J	Translation, ribosomal structure and biogenesis	15	22	15
K	Transcription	0	0	4
L	DNA replication, recombination and repair	8	6	6
Cellular processes		8	8	11
D	Cell division and chromosome partitioning	0	2	0
O	Posttranslational modification, protein turnover, chaperones	4	2	4
M	Cell envelope biogenesis, outer membrane	3	3	1
N	Cell motility and secretion	1	1	5
P	Inorganic ion transport and metabolism	0	0	1
T	Signal transduction mechanisms	0	0	0
Metabolism		7	8	7
C	Energy production and conversion	1	1	0
G	Carbohydrate transport and metabolism	2	2	3
E	Amino acid transport and metabolism	2	1	1
F	Nucleotide transport and metabolism	0	2	1
H	Coenzyme metabolism	2	1	2
I	Lipid metabolism	0	1	0
Poorly characterized		5	3	6
R	General function prediction only	5	3	3
S	Function unknown	0	0	3



Zhaxybayeva and Gogarten, *BMC Genomics* 2002, 3:4

Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes

(Huelsenbeck and Ronquist, 2001)

Problem:

Strimmer's formula $p_i = \frac{L_i}{L_1+L_2+L_3}$

only considers 3 trees
(those that maximize the likelihood for the three topologies)

Solution:

Exploration of the tree space by sampling trees using a biased random walk
(Implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

$$p_i \approx \frac{N_i}{N_{total}}$$

,where N_i - number of sampled trees of topology *i*, *i*=1,2,3

N_{total} – total number of sampled trees (has to be large)

Illustration of a biased random walk

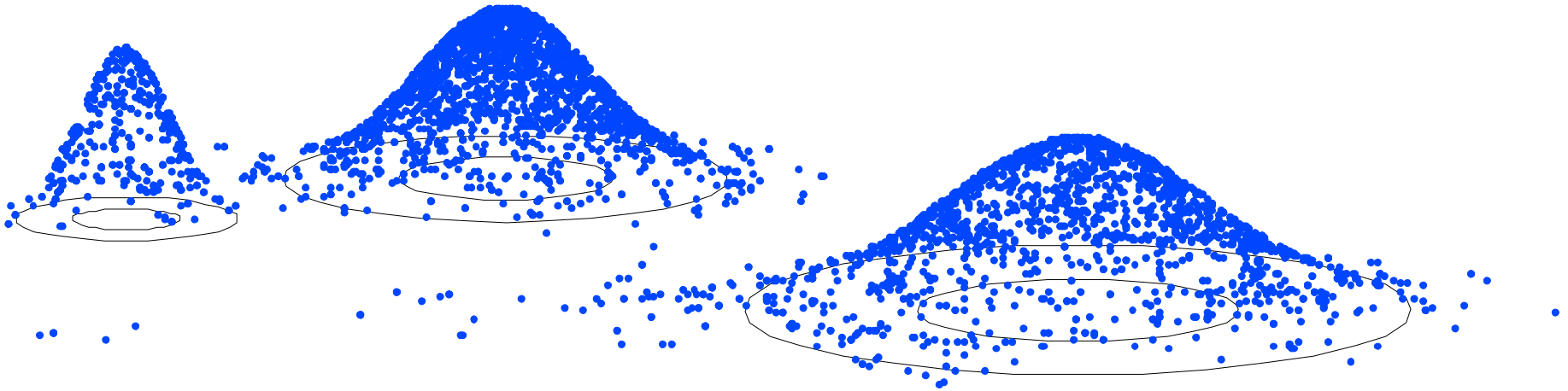
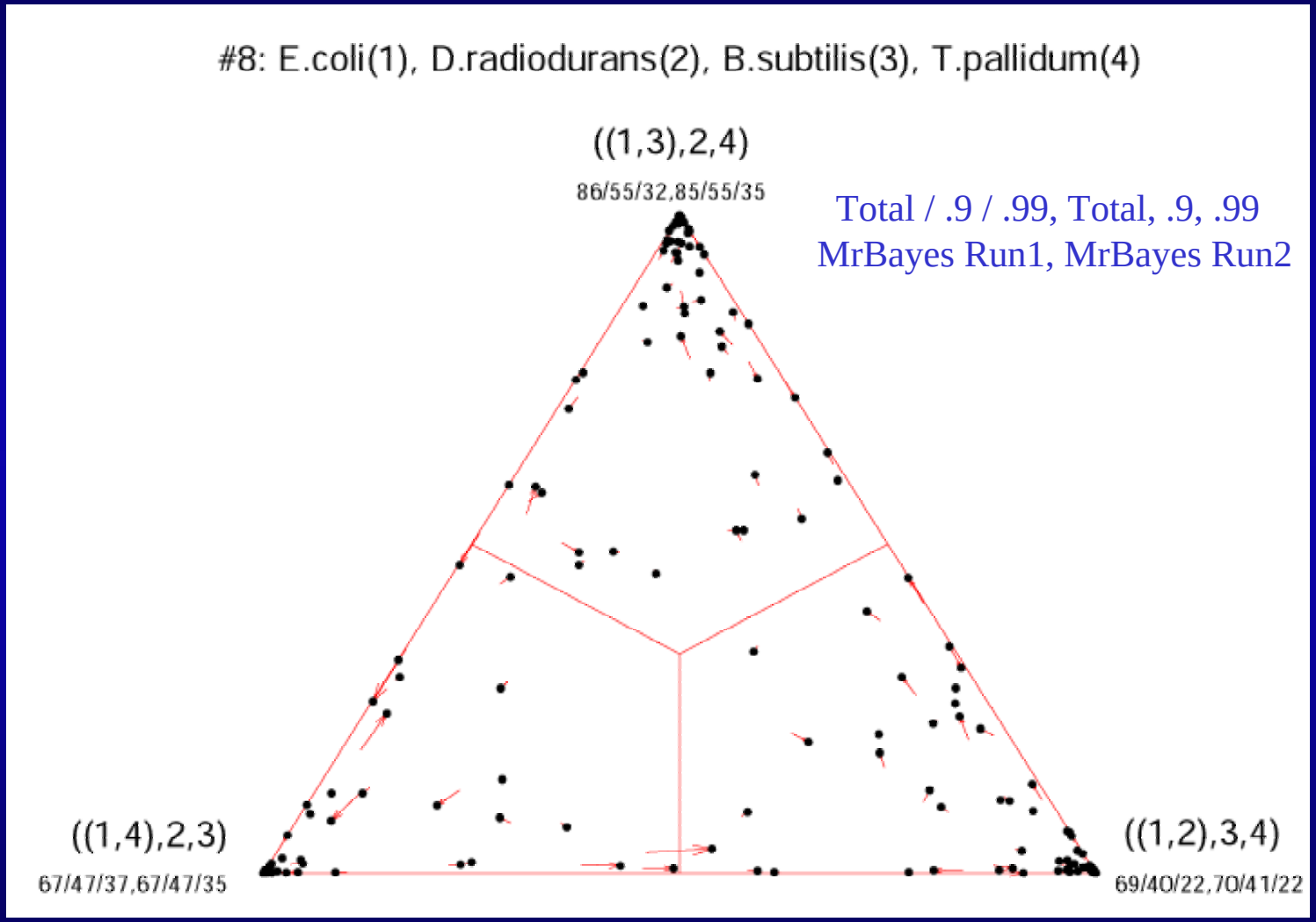


Figure generated using MCRobot program (Paul Lewis, 2001)

Inter-phylum relationships (bacteria) - there is no obvious core

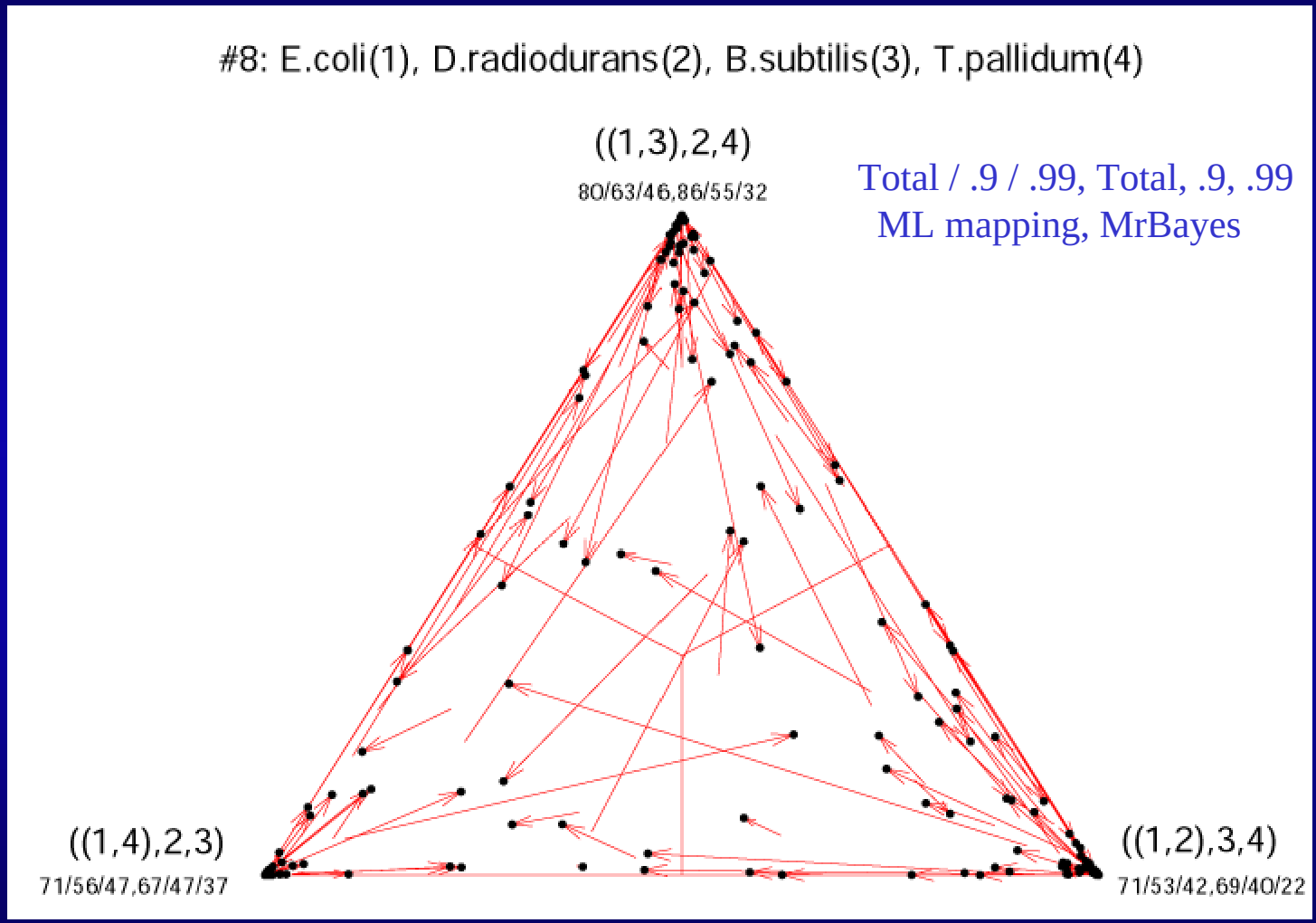


P-vector with MrBayes Run#1: Start of arrow

P-vector with MrBayes Run#2: Black dot at tip of arrow

Zhaxybayeva and Gogarten, *BMC Genomics* 2002, **3**:4

Comparing ML-mapping to Bayesian posterior probabilities



P-vector with ML-mapping: Start of arrow

P-vector with MrBayes: Black dot at tip of arrow

Zhaxybayeva and Gogarten, *BMC Genomics* 2002, **3**:4

Alternative Approaches to Estimate Posterior Probabilities (2)

Bootstrap Support Values Mapping:

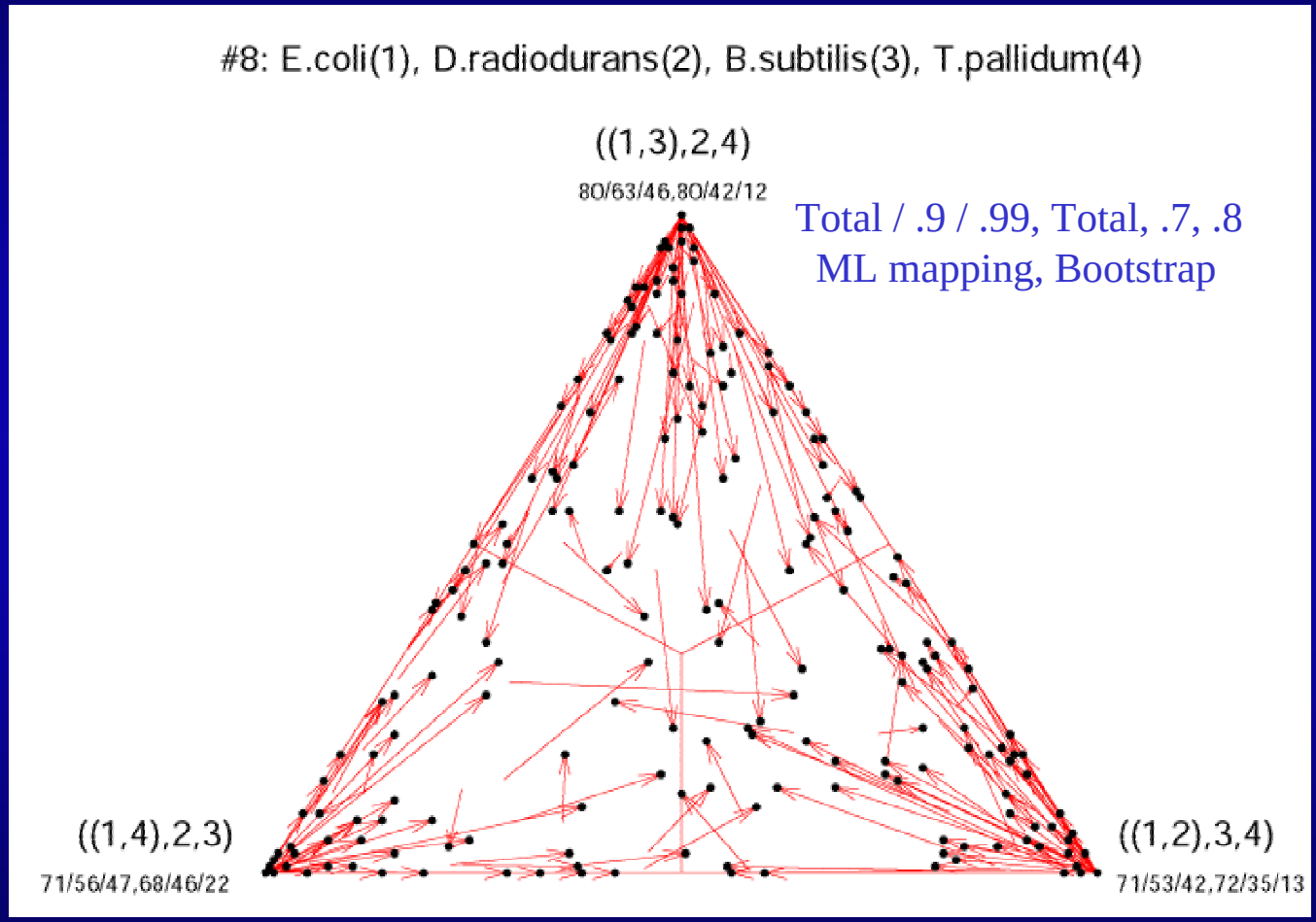
For each Quartet of Orthologous Proteins:

1) Create 100 bootstrapped samples

2) Evaluate three tree topologies for each of 100 samples

3) Construct bootstrap support values vector,
i.e., percent of bootstrapped samples
that have the highest likelihood value
for each tree topology.

Comparing ML-Mapping to Bootstrap Support Values



P-vector with ML-mapping: Start of arrow

P-vector with Bootstrap: Black dot at tip of arrow

Zhaxybayeva and Gogarten, *BMC Genomics* 2002, **3**:4

Comparing Support Measures:

99%≈90%≈70%

posterior probability calculated according to **ML mapping**

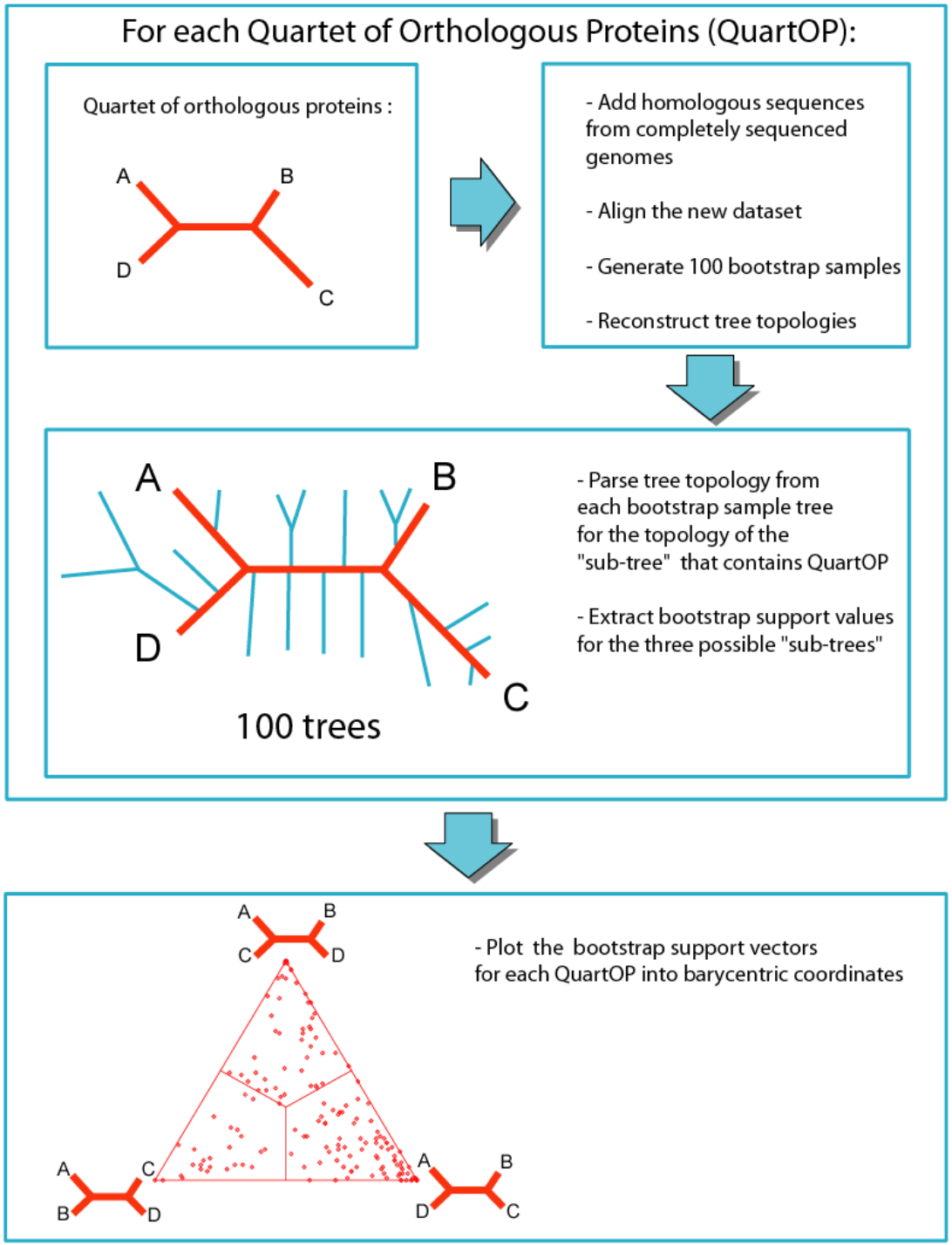
posterior probability estimated using **MCMC** (MrBayes)

bootstrap support

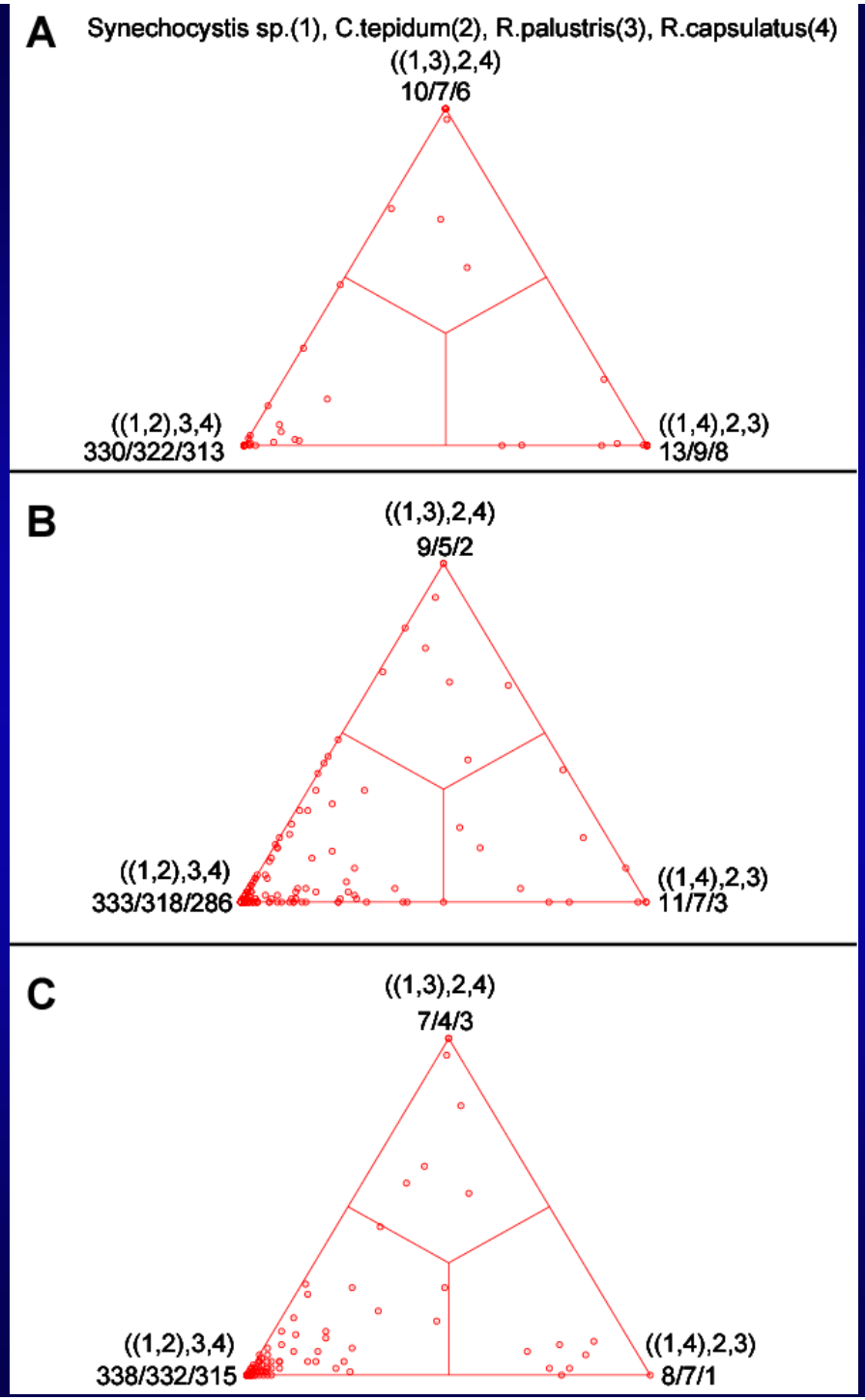
Increasing Reliability

Phylogenetic reconstruction becomes more reliable when more sequences are included.

DATA FLOW
analyses of extended datasets



Zhaxybayeva and Gogarten, *BMC Genomics* 2003 **4**: 37



Zhaxybayeva and Gogarten, *BMC Genomics* 2003 **4**: 37

COMPARISON OF DIFFERENT SUPPORT MEASURES

A: mapping of posterior probabilities according to Strimmer and von Haeseler

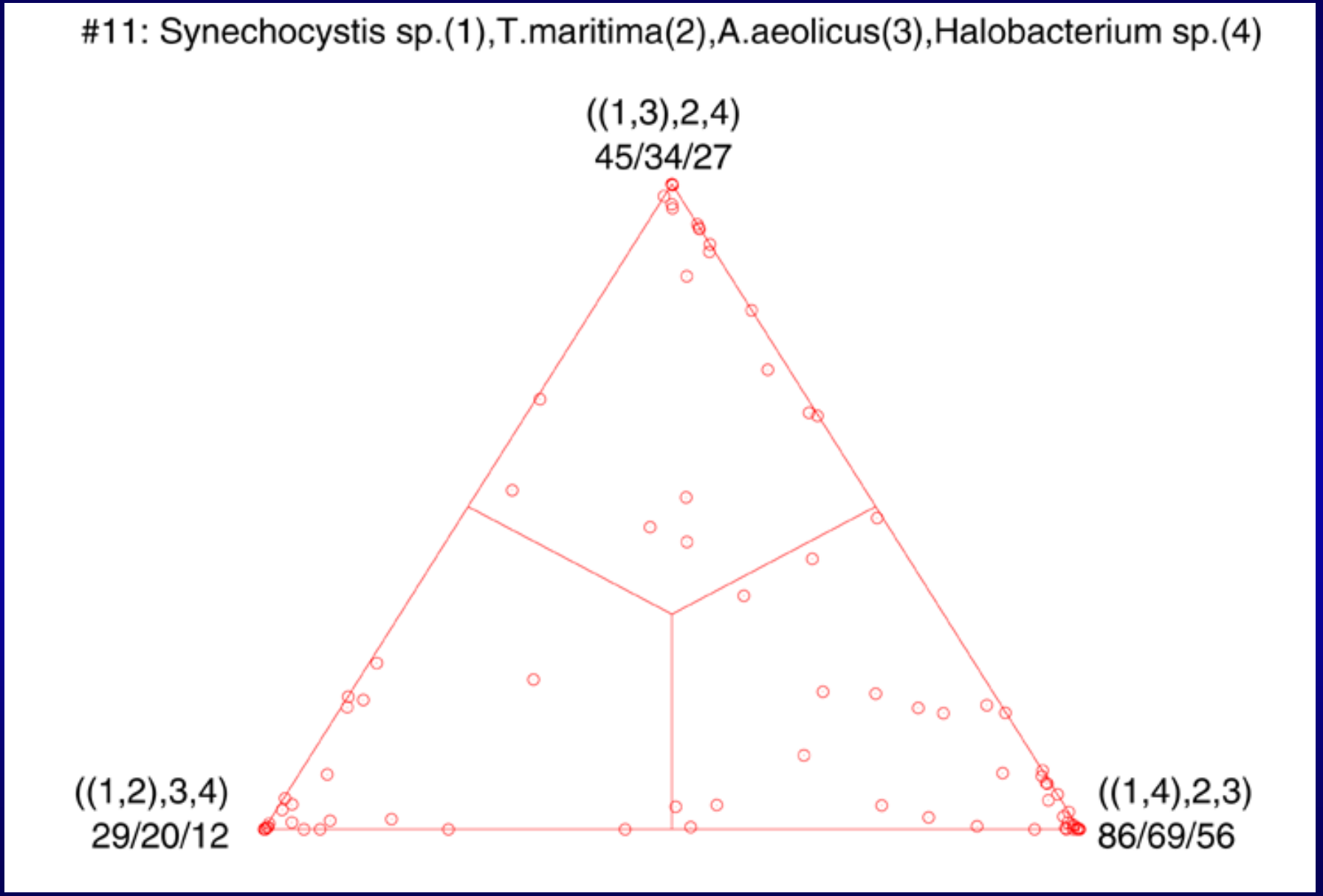
B: mapping of bootstrap support values

C: mapping of bootstrap support values from extended datasets

Inter-Domain Genome Comparisons

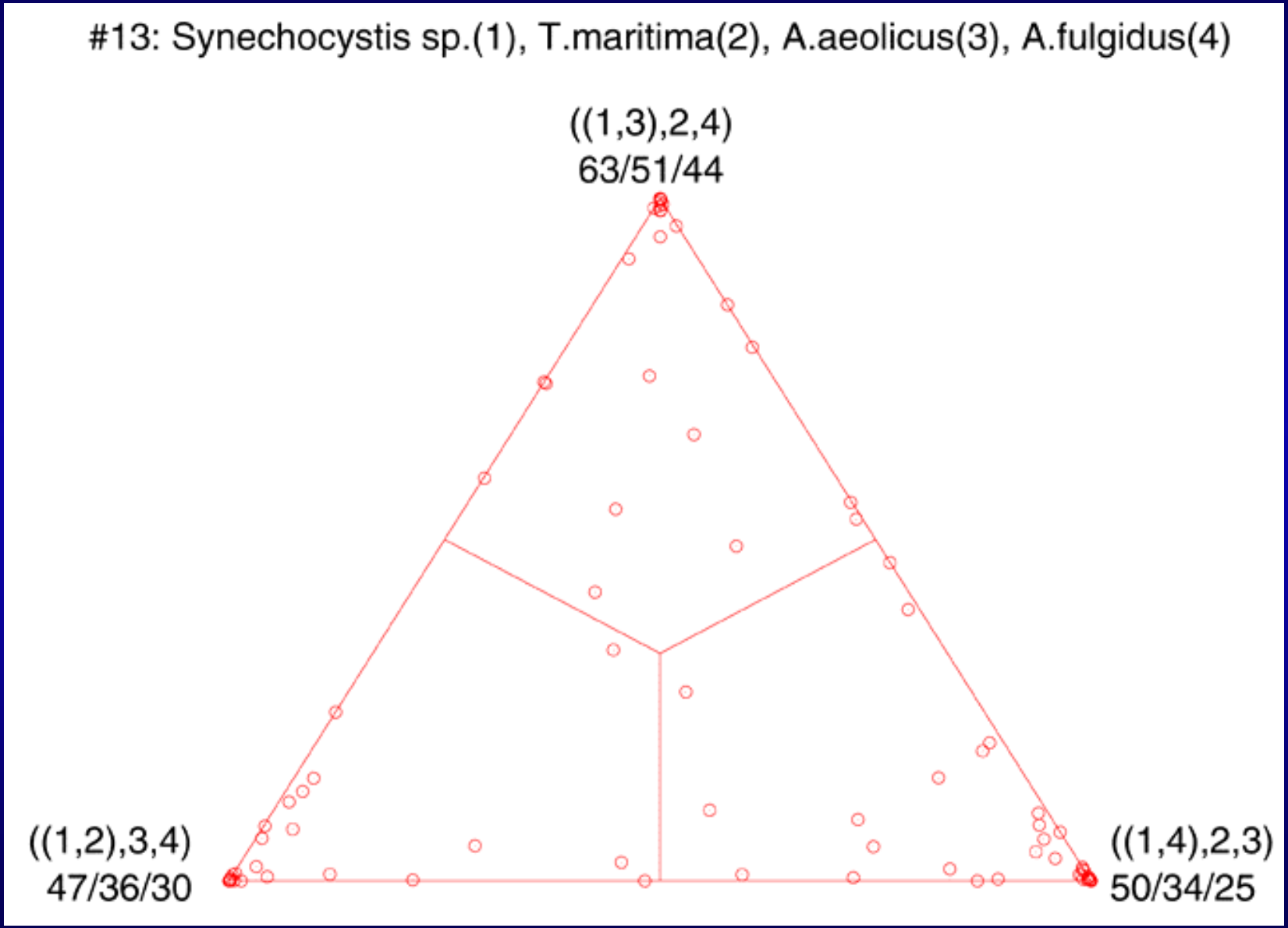
- ✓*Synechocystis* sp. – cyanobacterium
- ✓*Thermotoga maritima* – thermophilic bacterium
- ✓*Aquifex aeolicus* – thermophilic bacterium
- ✓*Halobacterium* sp. – salt-loving euryarchaeon

ML Map



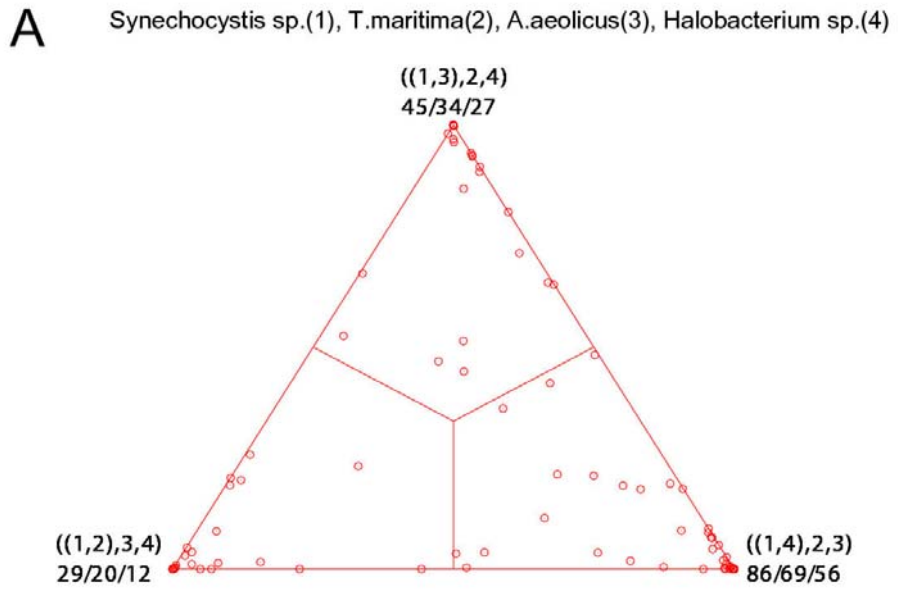
Zhaxybayeva and Gogarten, *BMC Genomics* 2002, 3:4

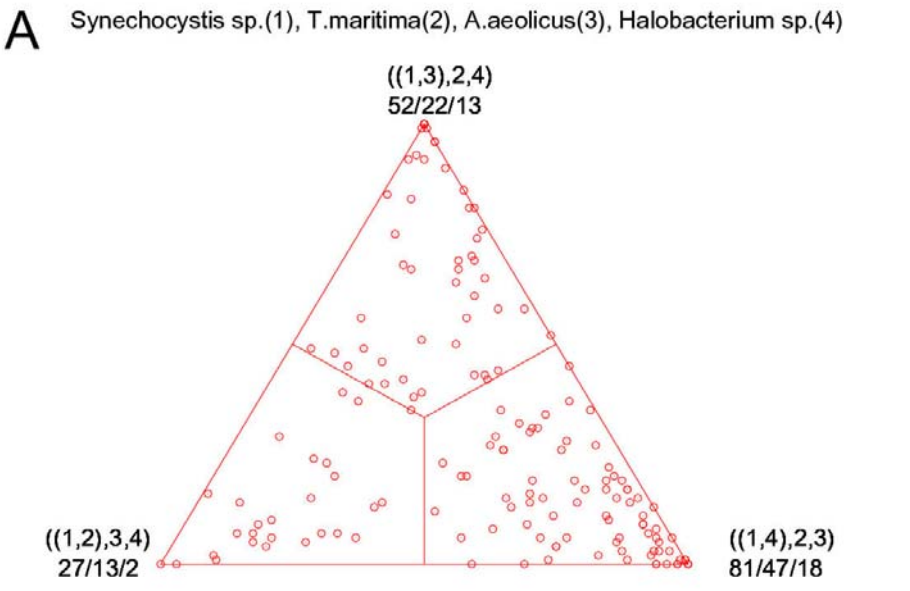
ML Map

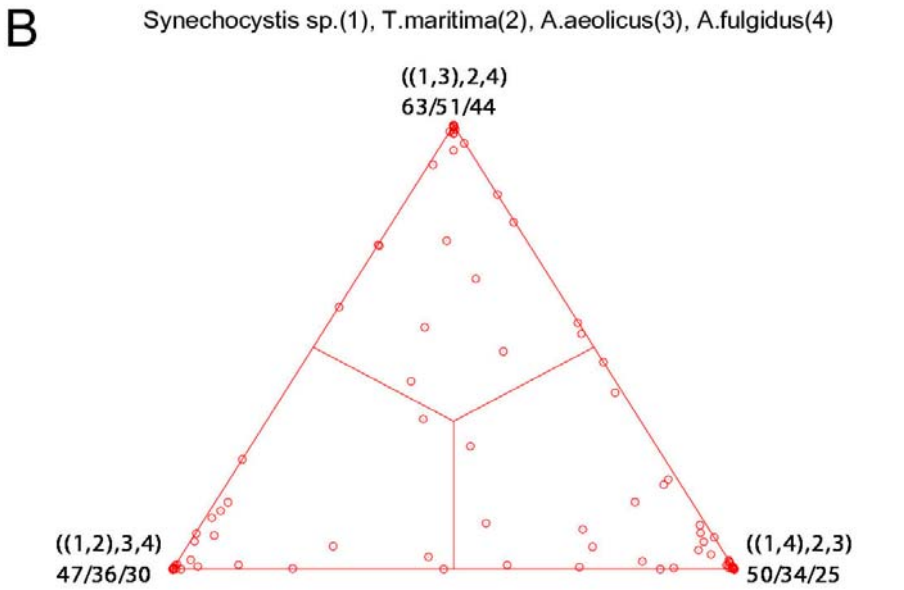


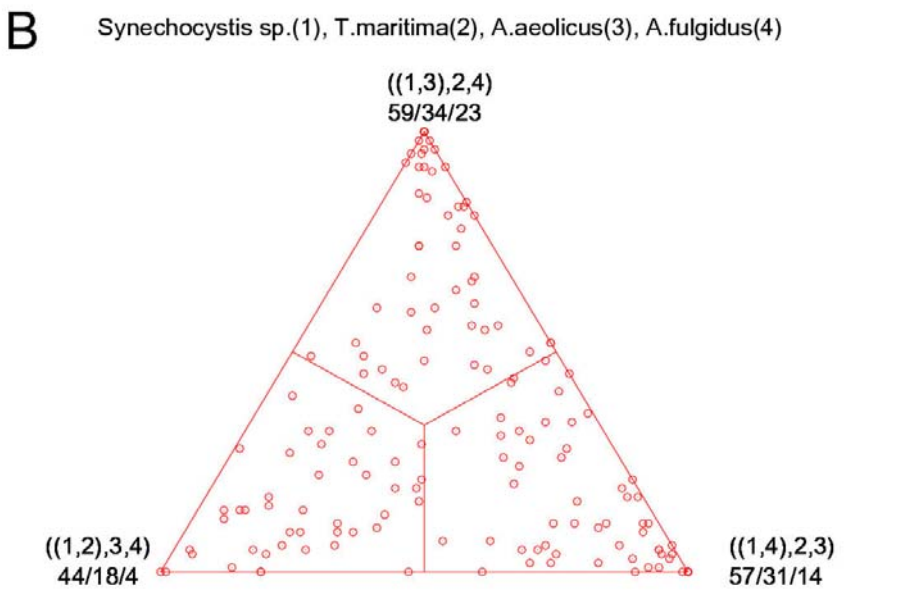
Zhaxybayeva and Gogarten, *BMC Genomics* 2002, 3:4

ml-mapping *versus* bootstrap values from extended datasets









Proteins in the ***information storage and processing category*** that group orthologs from ***Halobacterium*** with ***Synechocystis*** *and* ***Thermotoga*** *with* ***Aquifex*** (Topology #3 – putative identification)

Nucleotide modifying Enzymes

- tRNA-pseudouridine synthase
- dimethyladenosine transferase

Enzymes involved in DNA repair and recombination

- DNA mismatch repair protein
- excision nuclease A,B,C chains (involved in DNA repair)
- Endonuclease V (involved in DNA repair)

Enzymes involved in translation

- putative translation factor SUA5
- initiation factor IF2
- translation initiation factor eIF-2B subunit alpha
- Glu-tRNA amidotransferase subunits A,B
- ribosomal proteins L1,L11,L3,S4
- amino acyl tRNA synthetases for serine, valine, methionine, cysteine, proline, phenylalanine (α SU)

Other

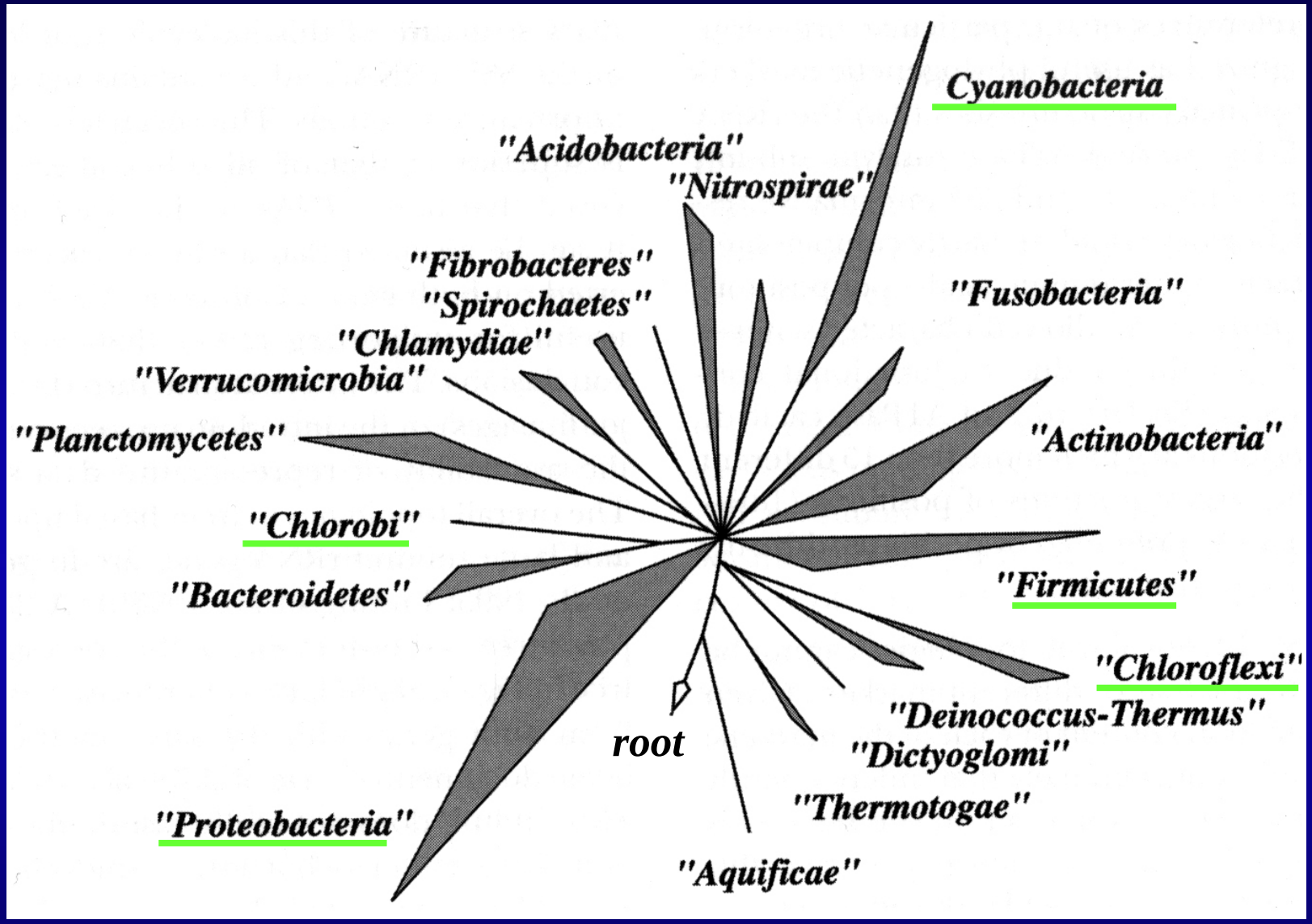
- DNA gyrase subunits [A,B]
- DNA helicase

NUMBER OF GENES PER CONFIDENCE LEVEL FOR DIFFERENT TYPES OF MAPPINGS

Genome Quartet	99% posterior probability	90% bootstrap support from non-extended datasets	90% bootstrap support from extended datasets
Interdomain quartet consisting of *Synechocystis* sp., *Halobacterium* sp., *Aquifex aeolicus* and *Thermotoga maritima.*	95	42	33
Interdomain quartet consisting of *Synechocystis* sp., *Archaeoglobus fulgidus*, *Aquifex aeolicus* and *Thermotoga maritima.*	99	42	41
Interphylum quartet of *Synechocystis* sp., *Chlorobium tepidum*, *Rhodobacter capsulatus* and *Rhodopseudomonas palustris.*	327	291	319

Zhaxybayeva and Gogarten, *BMC Genomics* 2003 **4**: 37

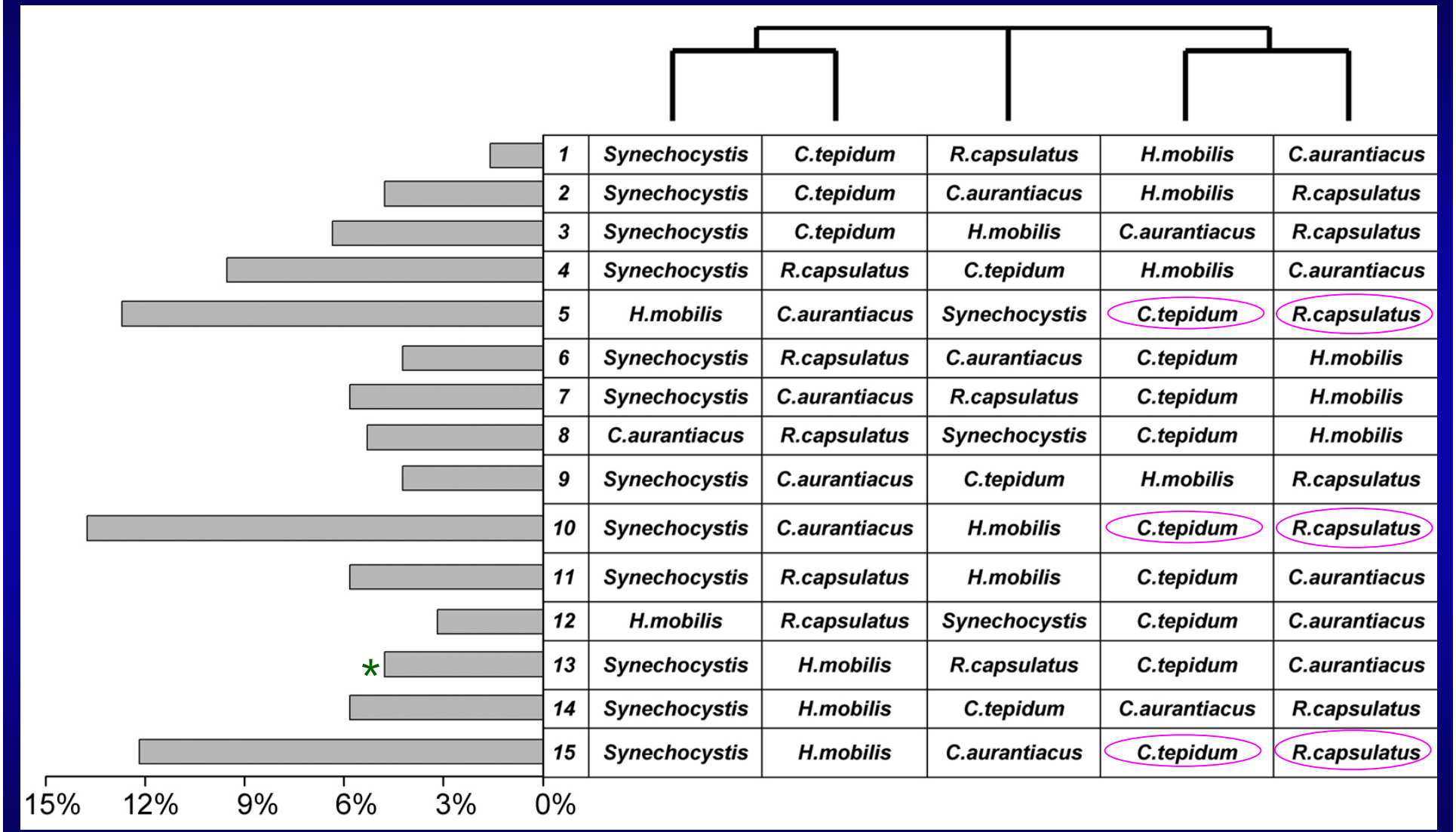
Extension of Mapping to Five Genomes



23S rRNA tree depicting the major bacterial phyla

(from Bergey's Manual of Systematic Bacteriology, 2nd Ed.)

Distribution of orthologs among 15 possible trees



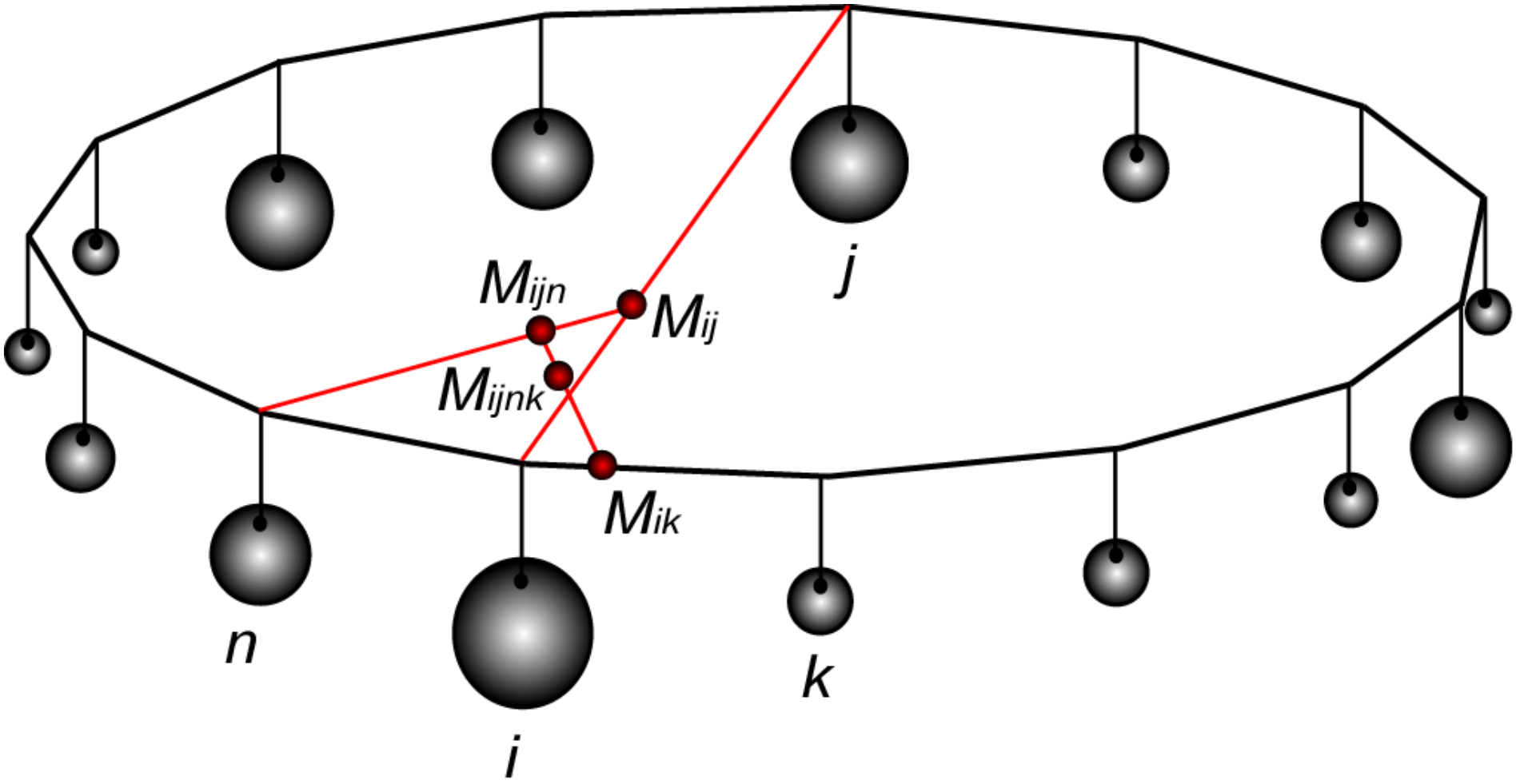
1	Synechocystis	C.tepidum	R.capsulatus	H.mobilis	C.aurantiacus
2	Synechocystis	C.tepidum	C.aurantiacus	H.mobilis	R.capsulatus
3	Synechocystis	C.tepidum	H.mobilis	C.aurantiacus	R.capsulatus
4	Synechocystis	R.capsulatus	C.tepidum	H.mobilis	C.aurantiacus
5	H.mobilis	C.aurantiacus	Synechocystis	C.tepidum	R.capsulatus
6	Synechocystis	R.capsulatus	C.aurantiacus	C.tepidum	H.mobilis
7	Synechocystis	C.aurantiacus	R.capsulatus	C.tepidum	H.mobilis
8	C.aurantiacus	R.capsulatus	Synechocystis	C.tepidum	H.mobilis
9	Synechocystis	C.aurantiacus	C.tepidum	H.mobilis	R.capsulatus
10	Synechocystis	C.aurantiacus	H.mobilis	C.tepidum	R.capsulatus
11	Synechocystis	R.capsulatus	H.mobilis	C.tepidum	C.aurantiacus
12	H.mobilis	R.capsulatus	Synechocystis	C.tepidum	C.aurantiacus
13	Synechocystis	H.mobilis	R.capsulatus	C.tepidum	C.aurantiacus
14	Synechocystis	H.mobilis	C.tepidum	C.aurantiacus	R.capsulatus
15	Synechocystis	H.mobilis	C.aurantiacus	C.tepidum	R.capsulatus

188 datasets of orthologous genes

Raymond, J., Zhaxybayeva, O., Gogarten, J.P., Gerdes, S., Blankenship, R.E.: Whole Genome Analysis of Photosynthetic Prokaryotes. *Science* 2002, **298**: 1616-1620.

CALCULATION OF THE CENTER OF GRAVITY OF THE DEKAPENTAGON

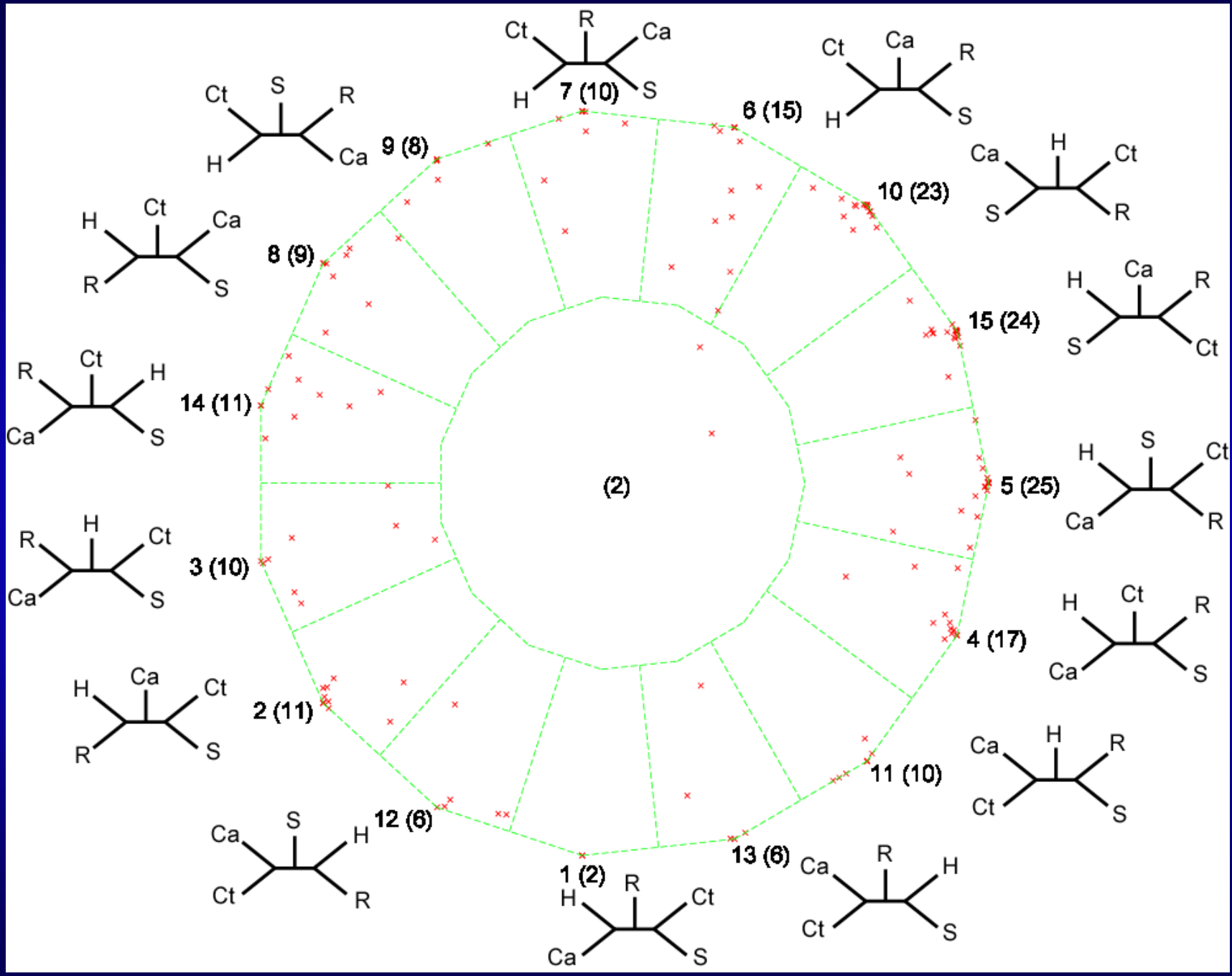
Illustration of the principle



Olga Zhaxybayeva, Lutz Hamel, Jason Raymond, and J. Peter Gogarten, *Genome Biology* 2004, 5: R20

PHYLOGENETIC DEKAPENTAGON

Posterior probabilities

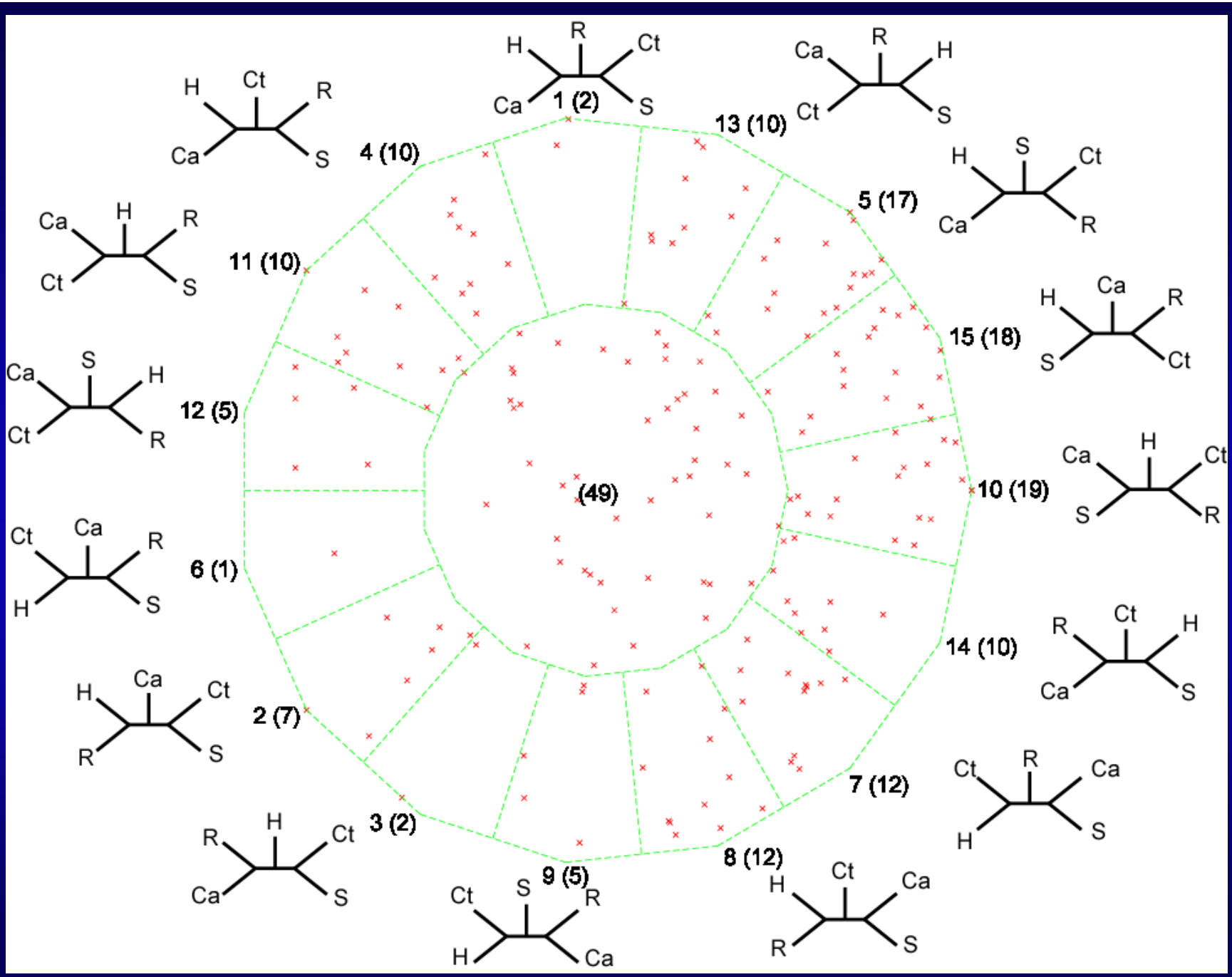


R: *Rhodobacter capsulatus*, **H**: *Heliobacillus mobilis*, **S**: *Synechocystis* sp., **Ct**: *Chlorobium tepidum*, **Ca**: *Chloroflexus aurantiacus*

Olga Zhaxybayeva, Lutz Hamel, Jason Raymond, and J. Peter Gogarten, *Genome Biology* 2004, **5**: R20

PHYLOGENETIC DEKAPENTAGON

Bootstrap support values



R: *Rhodobacter capsulatus*, **H**: *Heliobacillus mobilis*, **S**: *Synechocystis* sp., **Ct**: *Chlorobium tepidum*, **Ca**: *Chloroflexus aurantiacus*

Olga Zhaxybayeva, Lutz Hamel, Jason Raymond, and J. Peter Gogarten, *Genome Biology* 2004, **5**: R20

Extension of the analyses to more than five genomes

PROBLEM:

Number of possible unrooted tree topologies is equal to $(2n-5)!/[2^{n-3}(n-3)!]$

$\Rightarrow$ Polygon becomes a circle

$\Rightarrow$ Many topologies are not supported by data

SOLUTION:

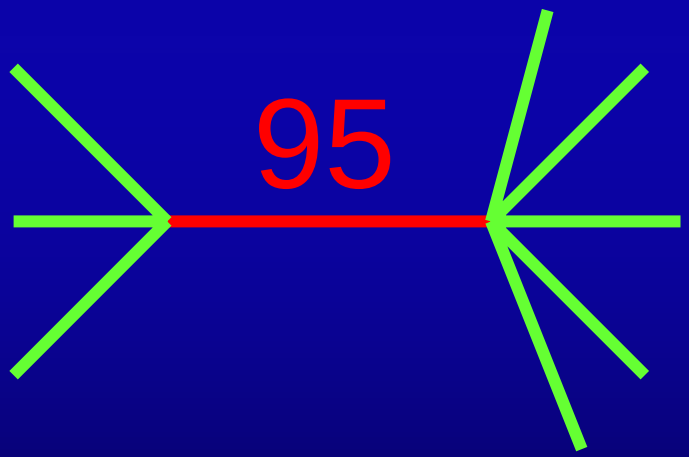
Switching from topologies to bipartitions of data

BIPARTITION PLOTS

(Modified Lento Plots)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition – a division of a phylogenetic tree into two parts that are connected by a single branch.
It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.



Number of bipartitions for *N* genomes is equal to $2^{(N-1)}-N-1$.

WHY BIPARTITIONS?

1. The number of possible bipartitions is much smaller than number of possible tree topologies, which makes it possible to evaluate all possible partitions.

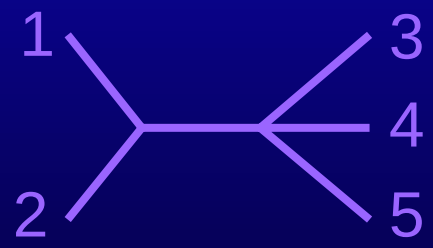
2. Analyses of bipartitions allows to consider datasets that otherwise would be considered as non-informative due to lack of resolution in one or the other part of the tree.

3. Putatively horizontally transferred genes can be detected because they give rise to partitions significantly conflicting with plurality partitions.

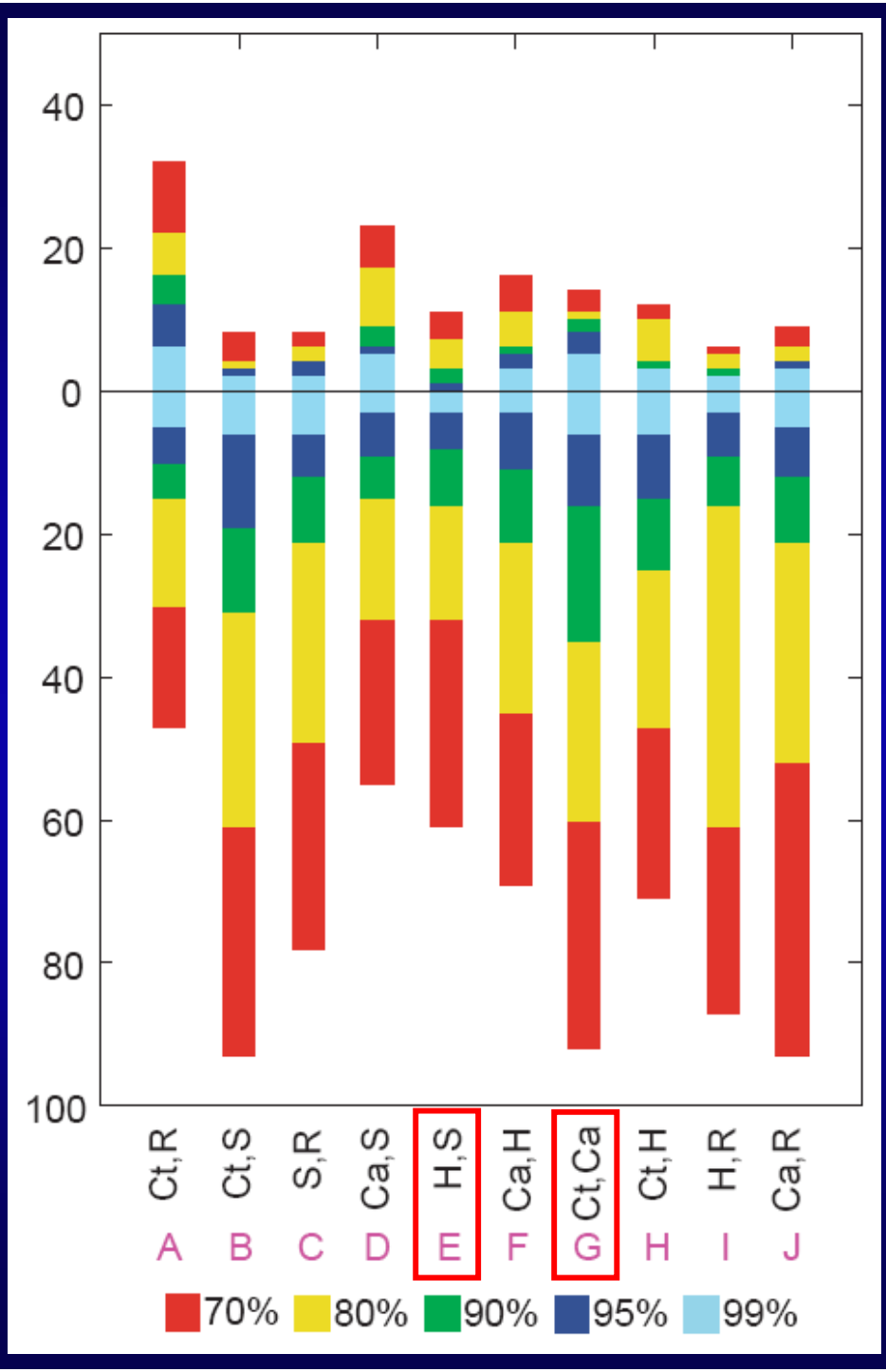
Example of bipartition analysis for five genomes of photosynthetic bacteria

R: *Rhodobacter capsulatus*,
H: *Heliobacillus mobilis*,
S: *Synechocystis* sp.,
Ct: *Chlorobium tepidum*,
Ca: *Chloroflexus aurantiacus*

Bipartitions supported by genes from chlorophyll biosynthesis pathway



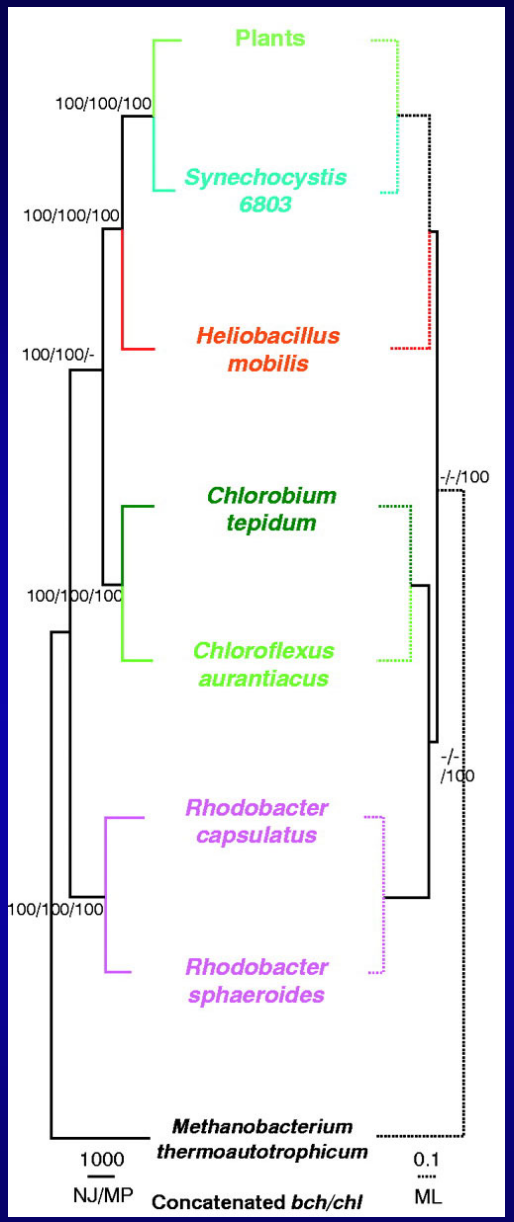
10 bipartitions



Zhaxybayeva, Hamel, Raymond, and Gogarten, *Genome Biology* 2004, **5**: R20

Phylogenetic Analyses of Genes from chlorophyll biosynthesis pathway

(extended datasets)



R: *Rhodobacter capsulatus*, **H**: *Heliobacillus mobilis*, **S**: *Synechocystis* sp., **Ct**: *Chlorobium tepidum*, **Ca**: *Chloroflexus aurantiacus*

Xiong *et al. Science*, 2000 **289**:1724-30

Gene	Plurality Partition (Ct, H, S, R, Ca)	(H, Ca, R, S, Ct)	(Ca, R, H, Ct, S)
BchB	0 0 0 0/0/0	100 100 100 100/100/100	16 15 34 42/41/41
BchD	4 2 19 0/0/0	74 76 24 98/98/98	100 99 100 100/100/100
BchH	0 0 0 0/0/0	100 100 100 100/100/100	100 100 98 100/100/100
BchI	0 2 5 2/1/2	11 9 20 0/0/0	98 93 70 100/100/100
BchL	0 0 0 0/0/0	18 33 89 73/75/74	99 100 100 100/100/100
BchN	0 0 0 0/0/0	100 99 100 100/100/100	3 6 100 6/5/6

TREE-PUZZLE Distances/NJ tree
TREE-PUZZLE Distances/FITCH tree
Parsimony with bootstrap
MrBayes (3 independent runs)

Zhaxybayeva, Hamel, Raymond, and Gogarten, *Genome Biology* 2004, 5: R20

PLOS BIOLOGY

From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the γ-Proteobacteria

Emmanuelle Lerat[1], Vincent Daubin[2], Nancy A. Moran[1*]

1 Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, United States of America, 2 Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, Arizona, United States of America

13 gamma-proteobacterial genomes:

- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

Detected 205 strictly selected orthologous datasets

Concatenated into one dataset

One consensus tree

Constructed 13 possible hypotheses for tree topologies and evaluated them with each dataset

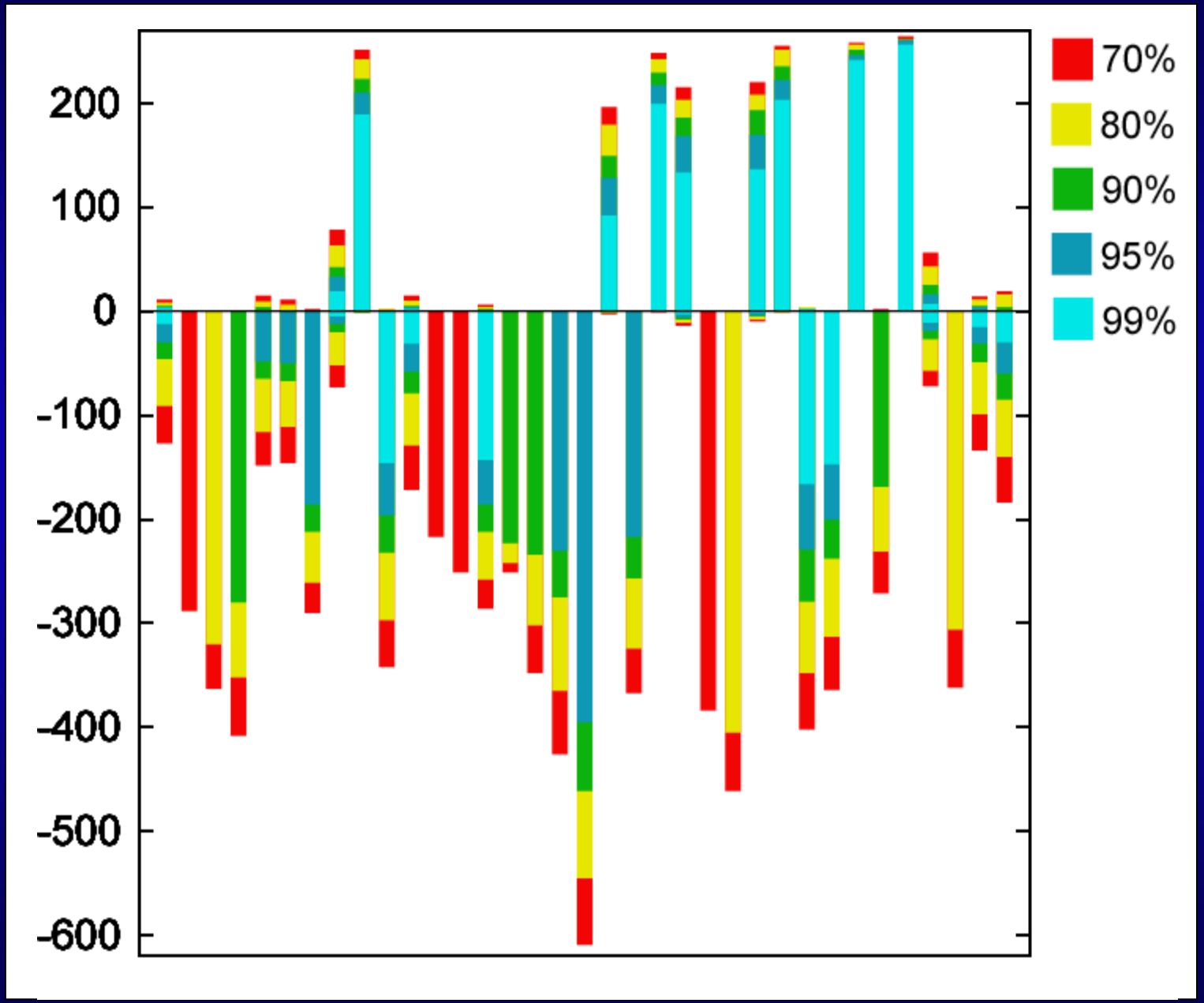
Majority support for one tree topology = species tree (?) for gamma proteobacteria

“Lento”-plot of 35 supported bipartitions (out of 4082 possible)

13 gamma-proteobacterial genomes (258 putative orthologs):

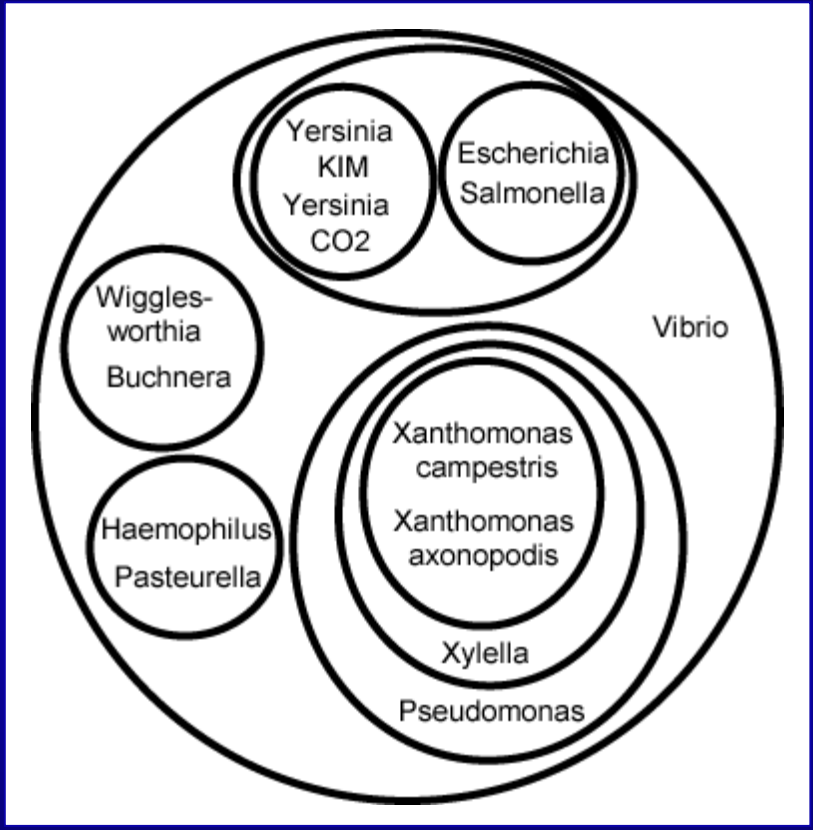
- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

There are 13,749,310,575 possible unrooted tree topologies for 13 genomes

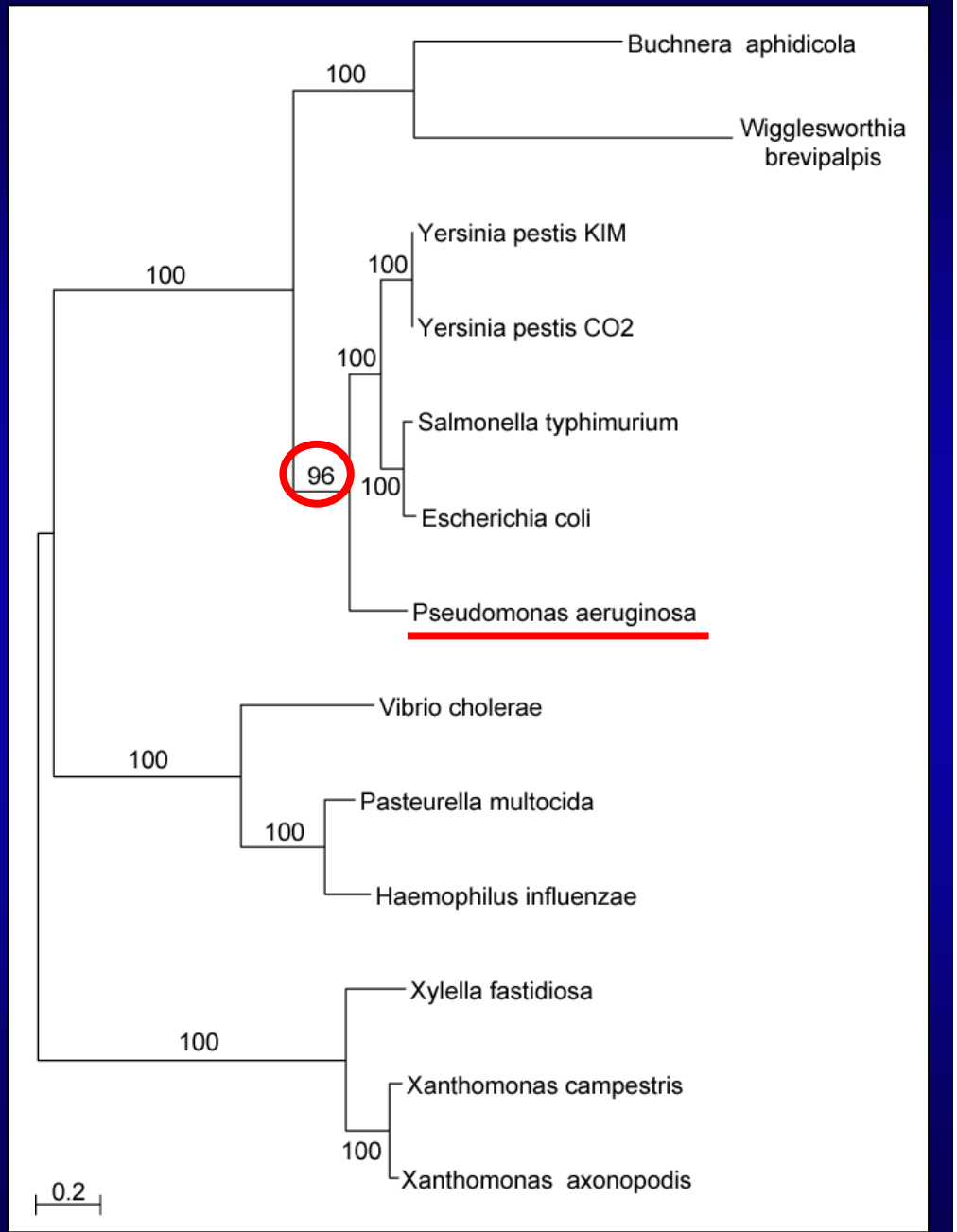


Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

Consensus cluster of significantly supported bipartitions



Phylogeny of virulence factor homologs (mviN)

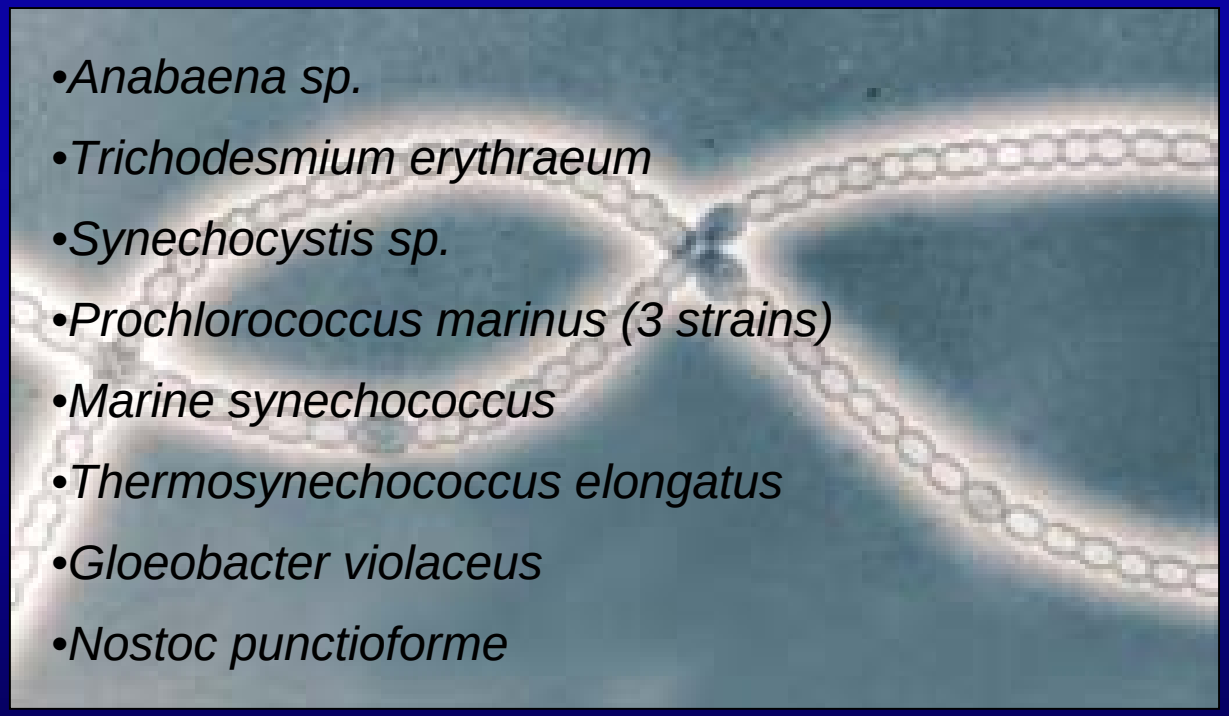


Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

Case of Cyanobacteria

Based on 16S rRNA:

- 13 gamma proteobacteria have up to 19.8% sequence divergence,
- 10 cyanobacteria are at most 14% divergent.

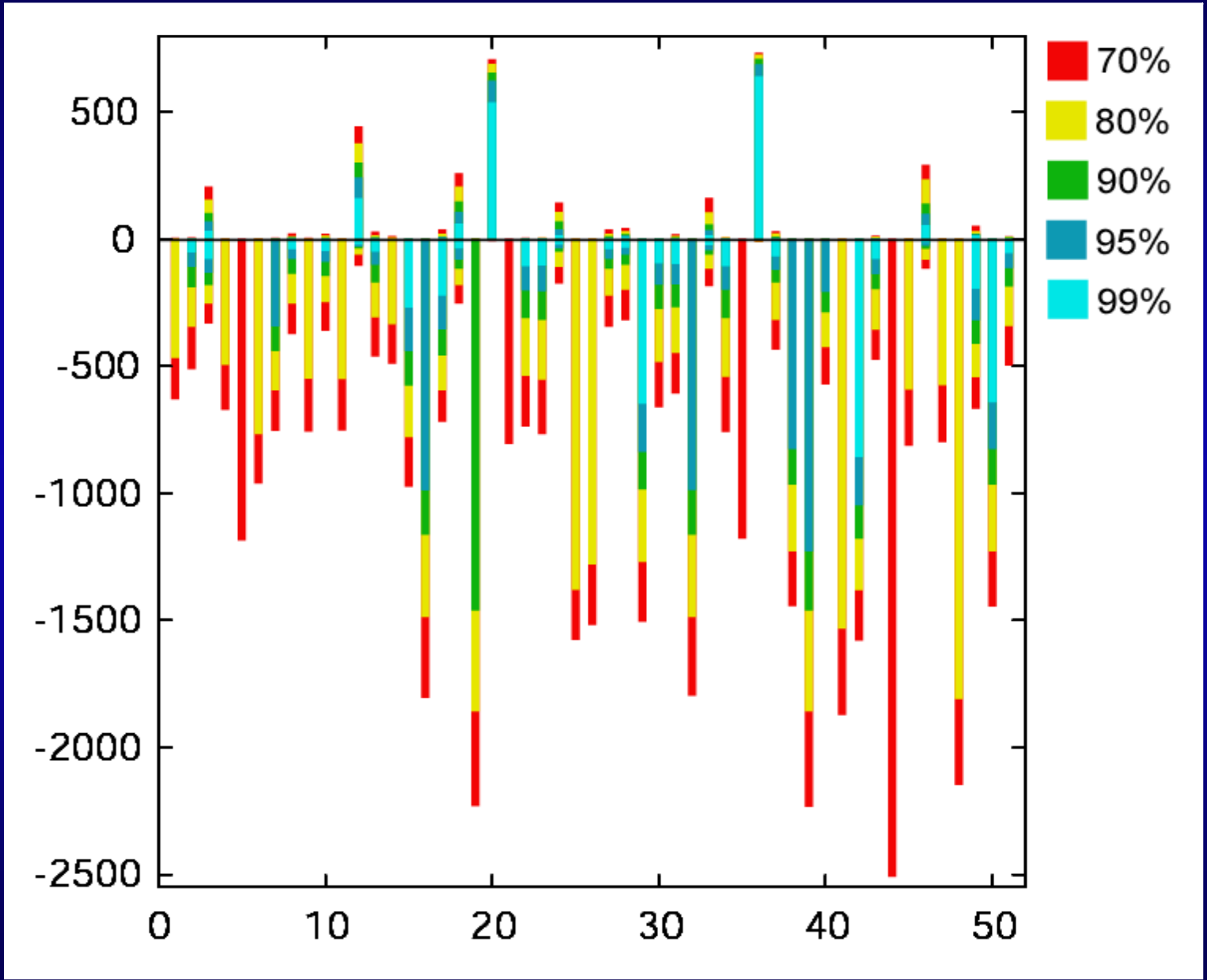


There are 678 orthologous genes detected by the reciprocal hit scheme.

“Lento”-plot of 51 supported bipartitions (out of 501 possible)

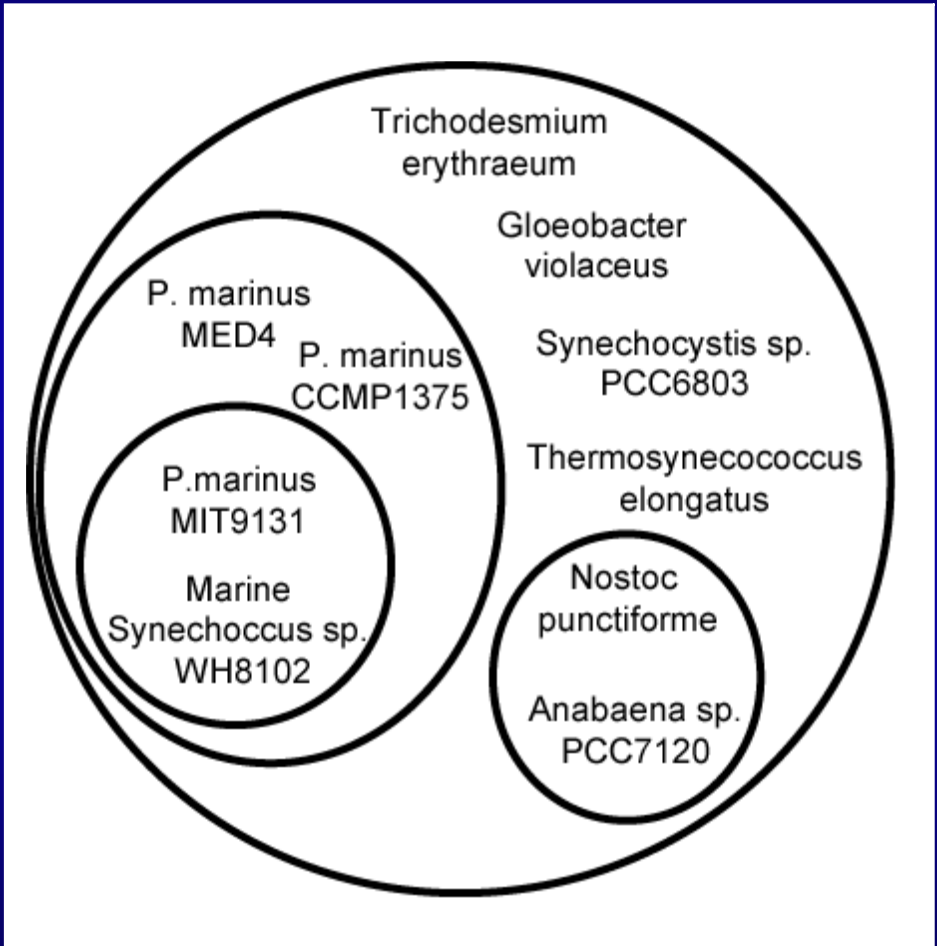
10 cyanobacteria:

- *Anabaena*
- *Trichodesmium*
- *Synechocystis* sp.
- *Prochlorococcus marinus* (3 strains)
- Marine *Synechococcus*
- *Thermo-synechococcus elongatus*
- *Gloeobacter*
- *Nostoc punctioforme*

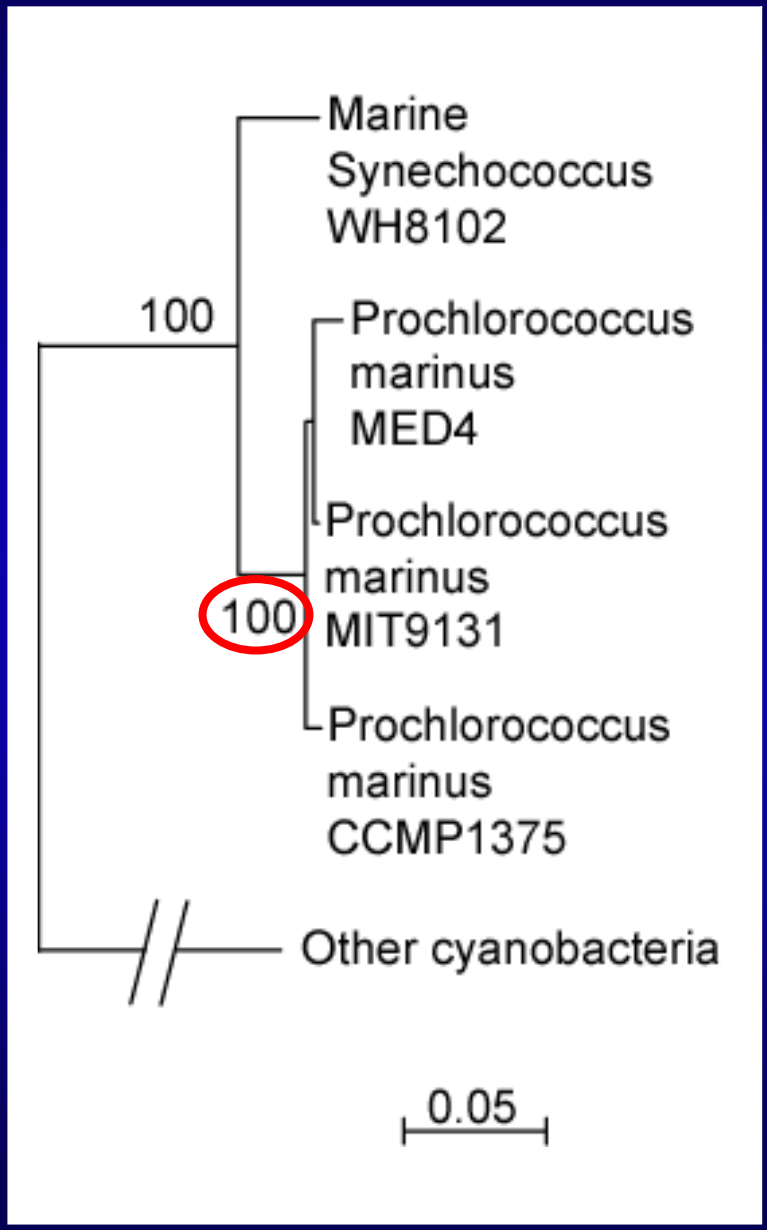


Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

Consensus cluster of significantly supported bipartitions



The phylogeny of ribulose bisphosphate carboxylase large subunit



Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

Other genes in conflict with the consensus at >=99% bootstrap support:

- cell division protein FtsH,
- translation initiation factor IF-2,
- ferredoxin, *petF*
- geranylgeranyl hydrogenase, *chlP*
- amidophosphoribosyltransferase,
- photosystem II reaction center core protein D2, *psbD*
- photosystem II CP43 core antenna protein, *psbC*
- photosystem II CP47 core antenna protein, *psbB*
- photosystem I reaction center core protein A2, *psaB*
- photosystem I reaction center core protein A1, *psaA*
- photosystem II manganese-stabilizing protein, *psbO*
- 5'-methylthioadenosine phosphorylase.

Zhaxybayeva, Lapierre and Gogarten, *Trends in Genetics*, 2004, **20(5)**: 254-260.

Transfer of photosynthesis genes to and from *Prochlorococcus* viruses

Debbie Lindell*†, Matthew B. Sullivan†‡, Zackary I. Johnson*, Andrew C. Tolonen‡, Forest Rohwer§, and Sallie W. Chisholm*¶‖

www.pnas.org/cgi/doi/10.1073/pnas.0401526101 PNAS | July 27, 2004 | vol. 101 | no. 30 | 11013–11018

Photosynthetic genes found in *Prochlorococcus* phages:

- PSII core reaction center protein D1 (*psbA*)
- PSII core reaction center protein D2 (*psbD*)
- ferredoxin (*petF*)
- plastocyanin (*petE*)
- HLIP cluster 14-type protein (*hli14* – high light inducible protein)

CONCLUSIONS I

- Genomes are mosaic
- Support value mapping is a useful tool to dissect mosaic genomes
- While ML mapping can provide a quick assessment of genome mosaicism, it grossly overestimates reliability
- Analyzing extended datasets using embedded subtrees solves the problems associated with taxon sampling without sacrificing the visually appealing graphical representation

CONCLUSIONS II

- Bipartition plots are a useful tool for comparative genome analyses. They allow to identify the plurality consensus cluster of genes contained in genomes as well as genes that conflict with the plurality consensus.

- In many instances majority or at least plurality signals are obtained from the analysis of individual genes.

- Sometimes clade-defining characteristics are among the genes that are transferred. E.g., for photosynthetic bacteria: plurality consensus phylogeny of genes $\neq$ phylogeny of the chlorophyll biosynthetic enzymes.

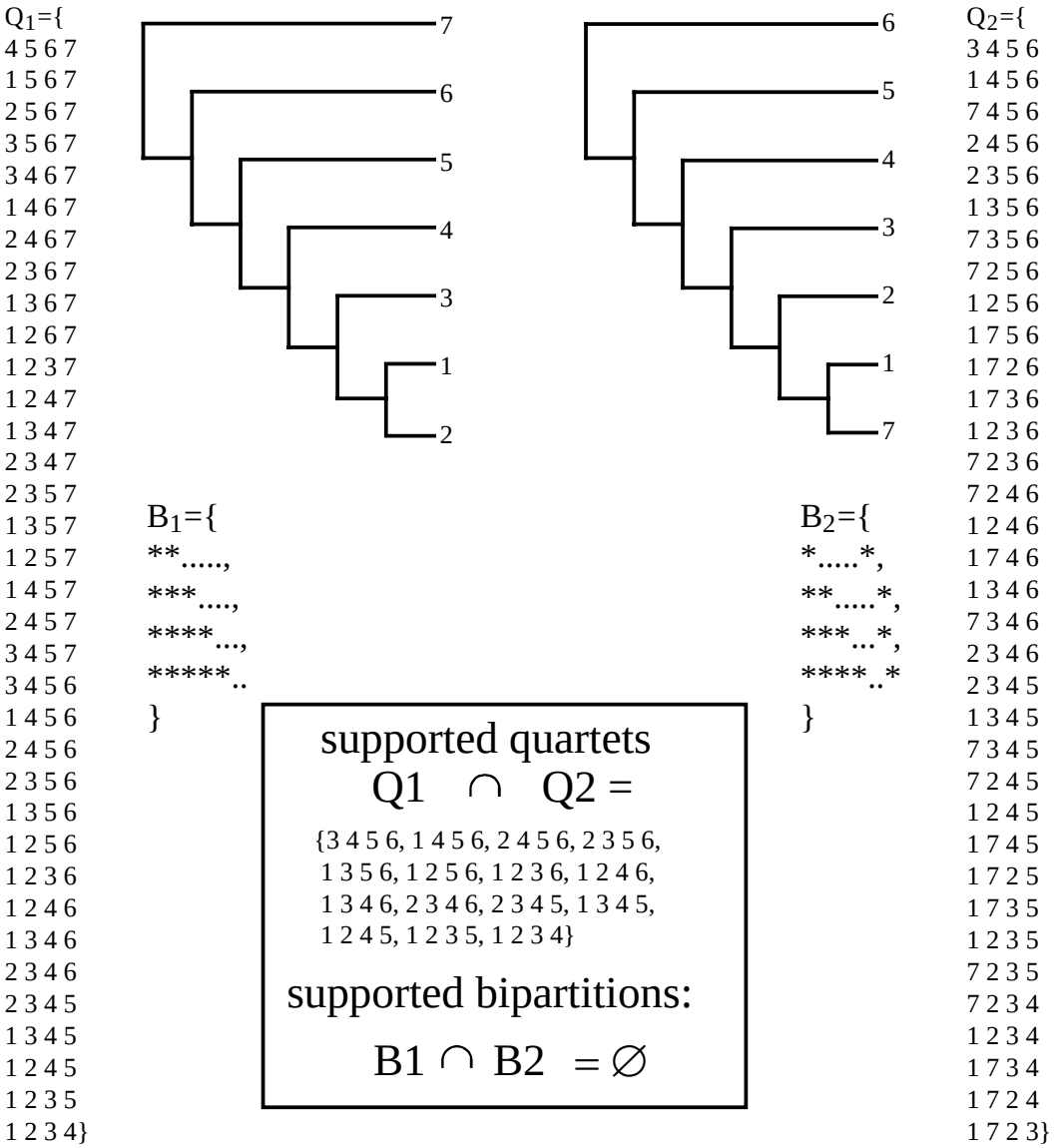


Illustration of a topology where quartet analyses are more useful than bipartition analyses

FUTURE RESEARCH

"Replace" bipartitions with Embedded Quartets in spectral analyses

+ Gene families that are not represented in all genomes can be included

+ adding more sequences does not deteriorate support values

+ a single "rogue" sequence does not erase all of the captured phylogenetic information