DNA Sequencing and the Hidden World of Microbes

Jonathan A. Eisen
University of California, Davis

Talk for ISB Symposium
April 15, 2012

The New York Times
Sunday, April 1, 2007

Health

Scientist Reveals Secret of the Ocean: It's Him

By NICHOLAS WADE
Published: April 1, 2007

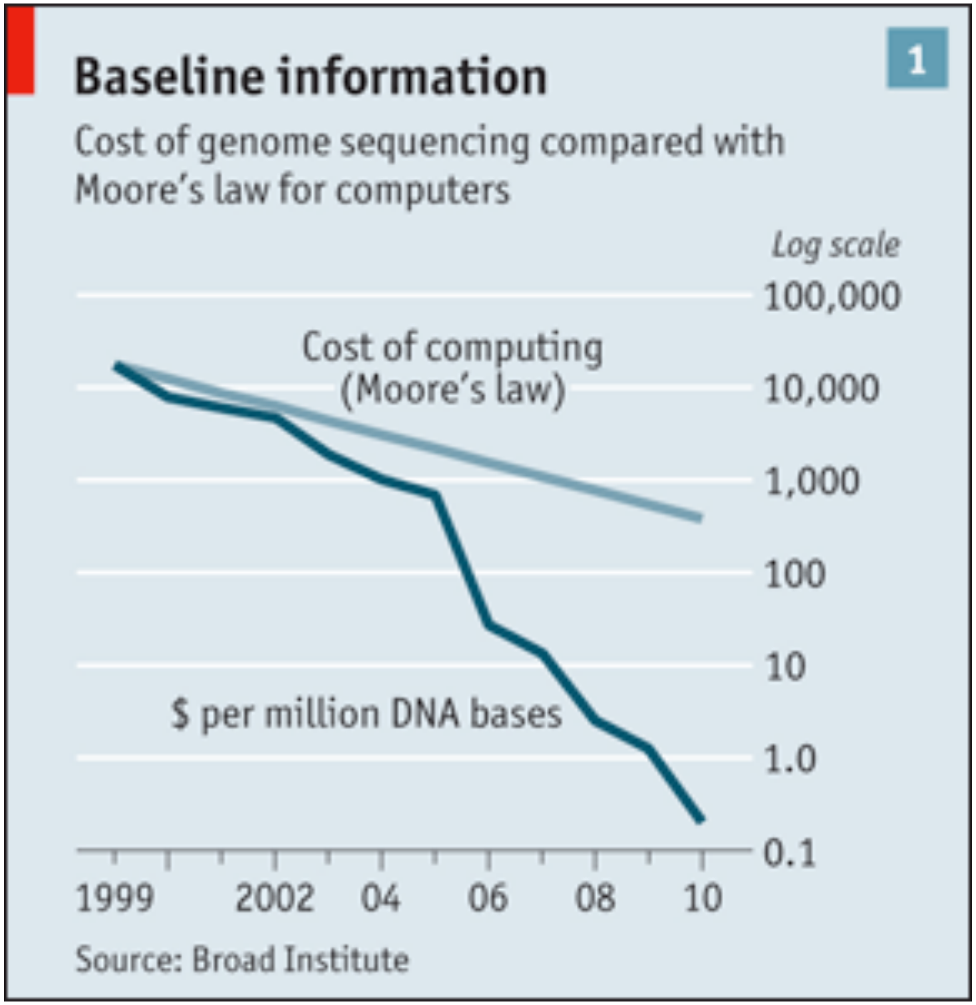
PRINT
SINGLE-PAGE
SHARE
SHARE

Maverick scientist J. Craig Venter has done it again. It was just a few years ago that Dr. Venter announced that the human genome sequenced by Celera Genomics was in fact, mostly his own. And now, Venter has revealed a second twist in his genomic self-examination. Venter was discussing his Global Ocean Voyage, in which he used his personal yacht to collect ocean water samples from around the world. He then used large filtration units to collect microbes from the water samples which were then brought back to his high tech lab in Rockville, MD where he used the same methods that were used to sequence the human genome to study the genomes of the 1000s of ocean dwelling microbes found in each sample. In discussing the sampling methods, Venter let slip his latest attack on the standards of science – some of the samples were in fact not from the ocean, but were from microbial habitats in and on his body.

"The human microbiome is the next frontier," Dr. Venter said. "The ocean voyage was just a cover. My main goal has always been to work on the microbes that live in and on people. And now that my genome is nearly complete, why not use myself as the model for human microbiome studies as well. "

It is certainly true that in the last few years, the microbes that live in and on people have become a hot research topic. So hot that the same people who were involved in the race to sequence the human genome have been involved in this race too. Francis Collins, Venter main competitor and still the director of the National Human Genome Research Institute

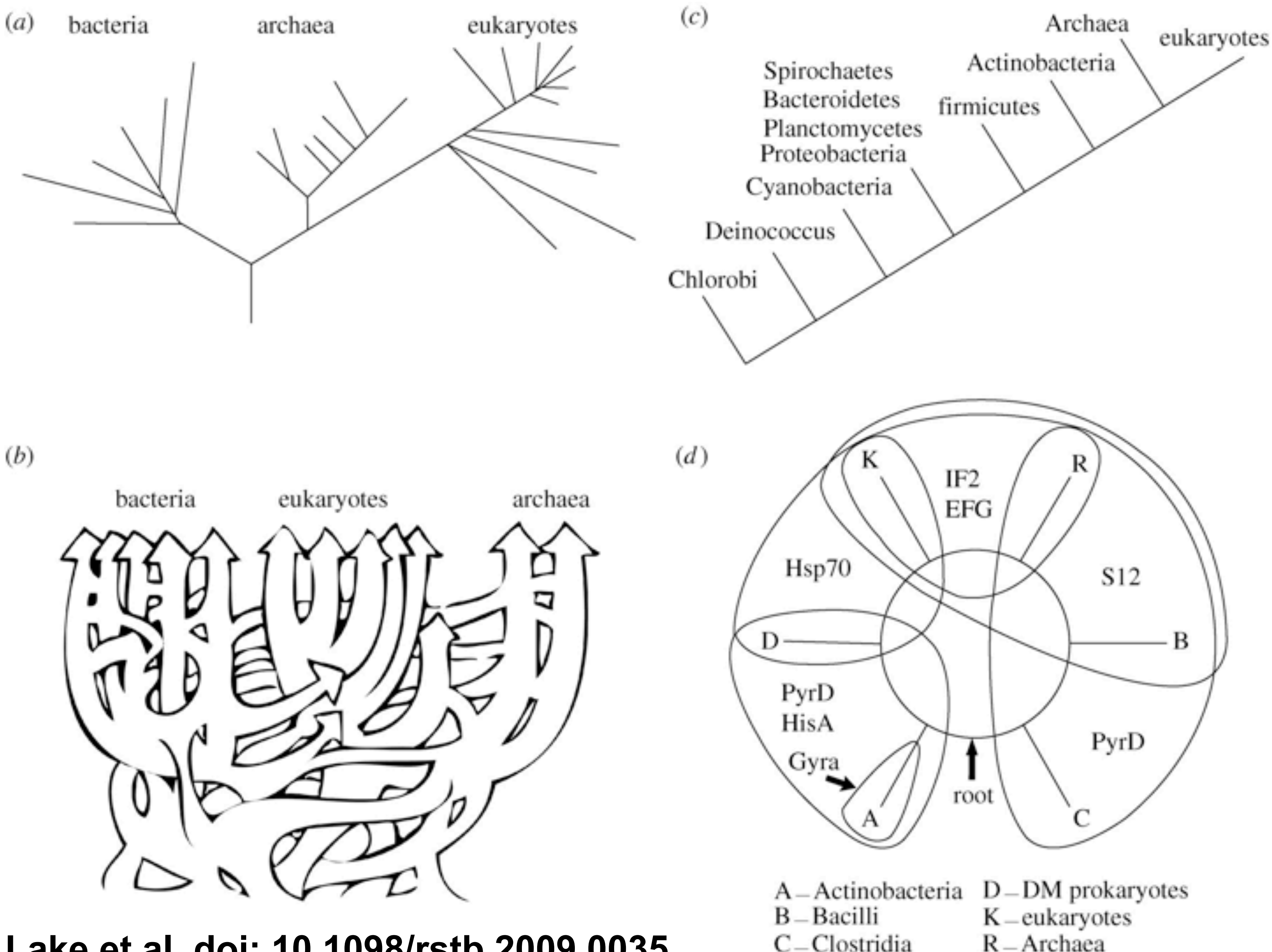
The Sequencing Revolution



Microbial Diversty w/ New Data

- Sequence data continues to revolutionize studies of microbial diversity
- To best take advantage of it:
  - Better genomic sampling
  - Better methods to use phylogeny

Whatever the History: Trying to Incorporate it is Critical



from Lake et al. doi: 10.1098/rstb.2009.0035

Part I: Better Sampling

rRNA Tree of Life

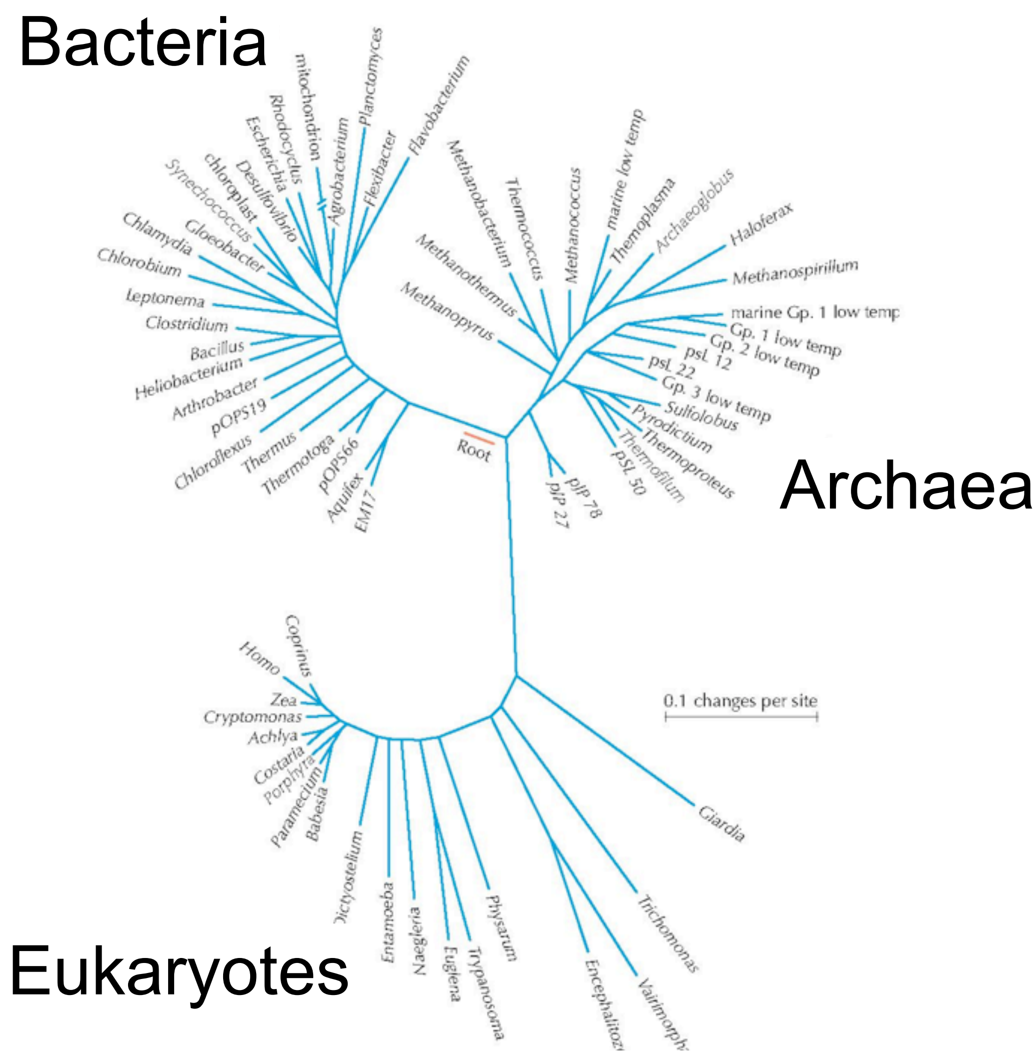
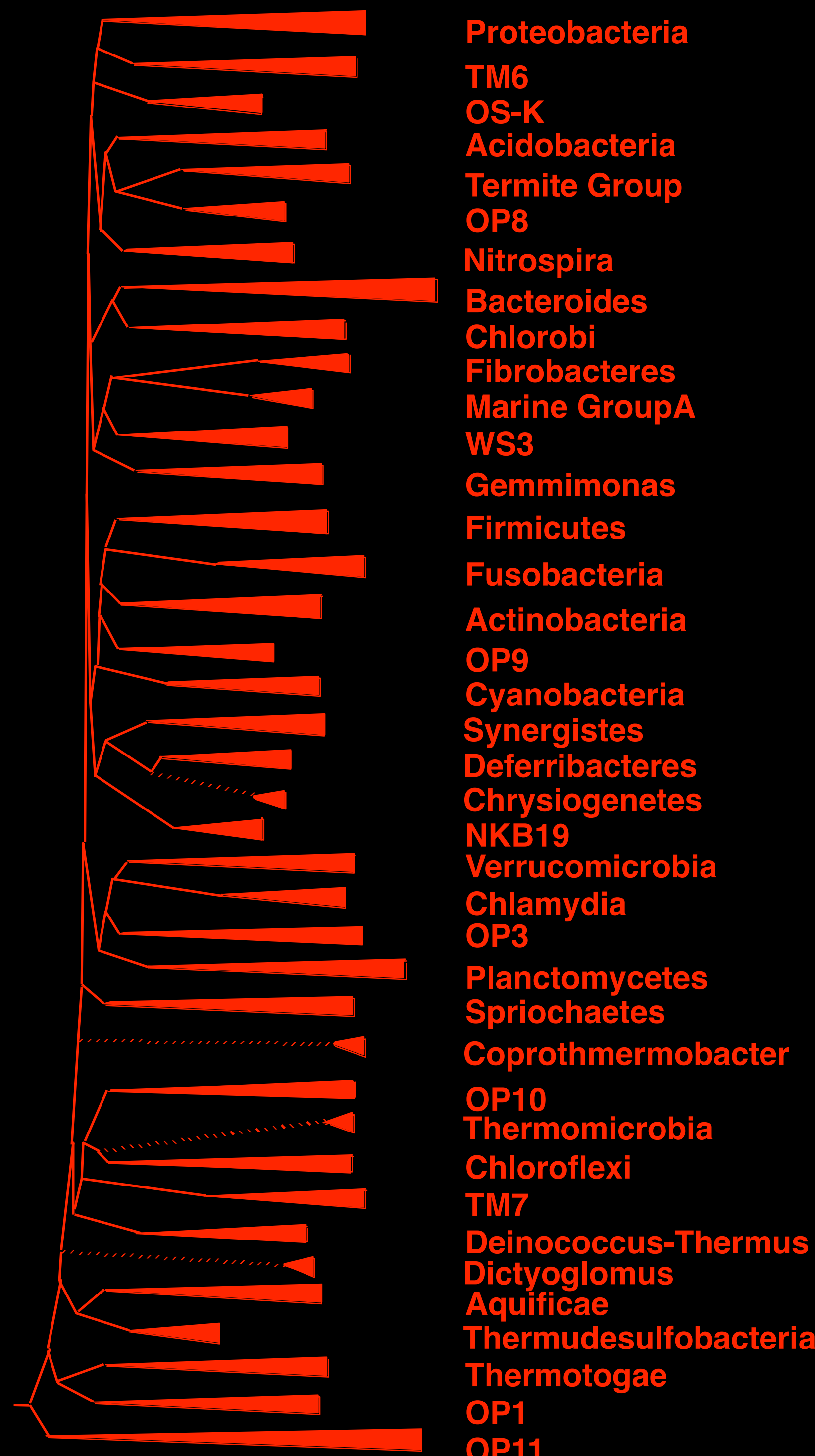
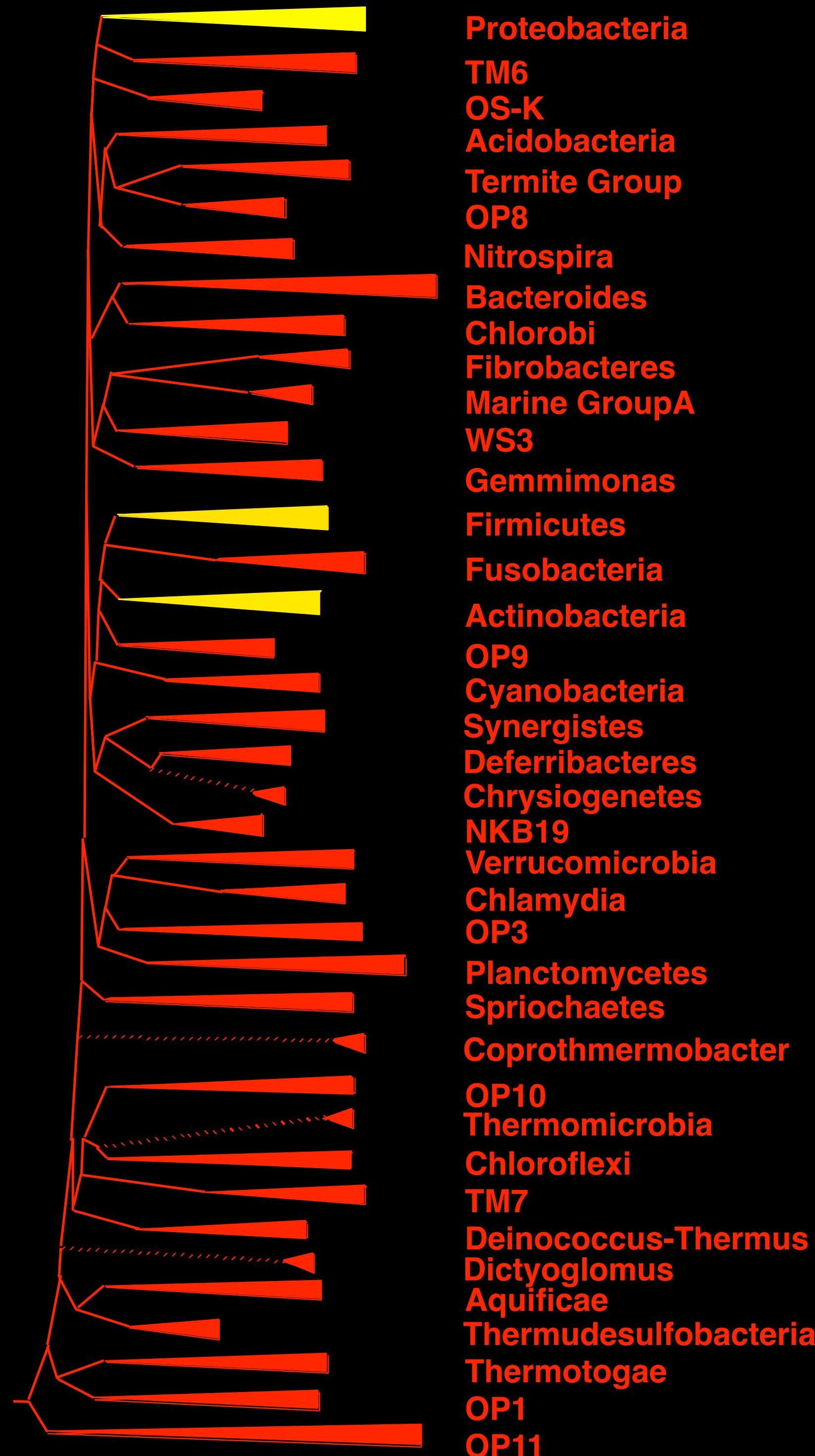


Figure from Barton, Eisen et al. "Evolution", CSHL Press. 2007.

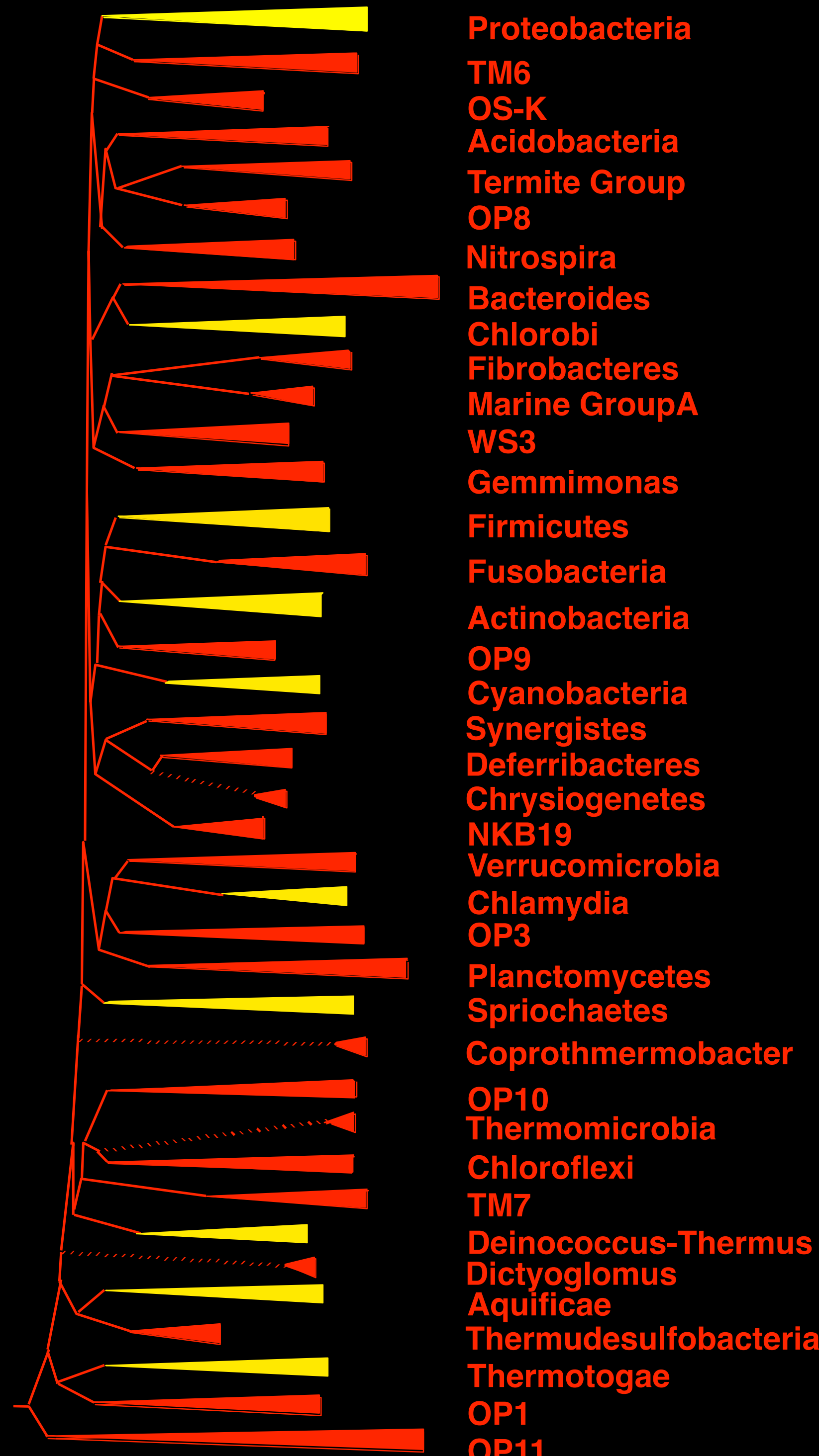
Based on tree from Pace 1997 Science 276:734-740



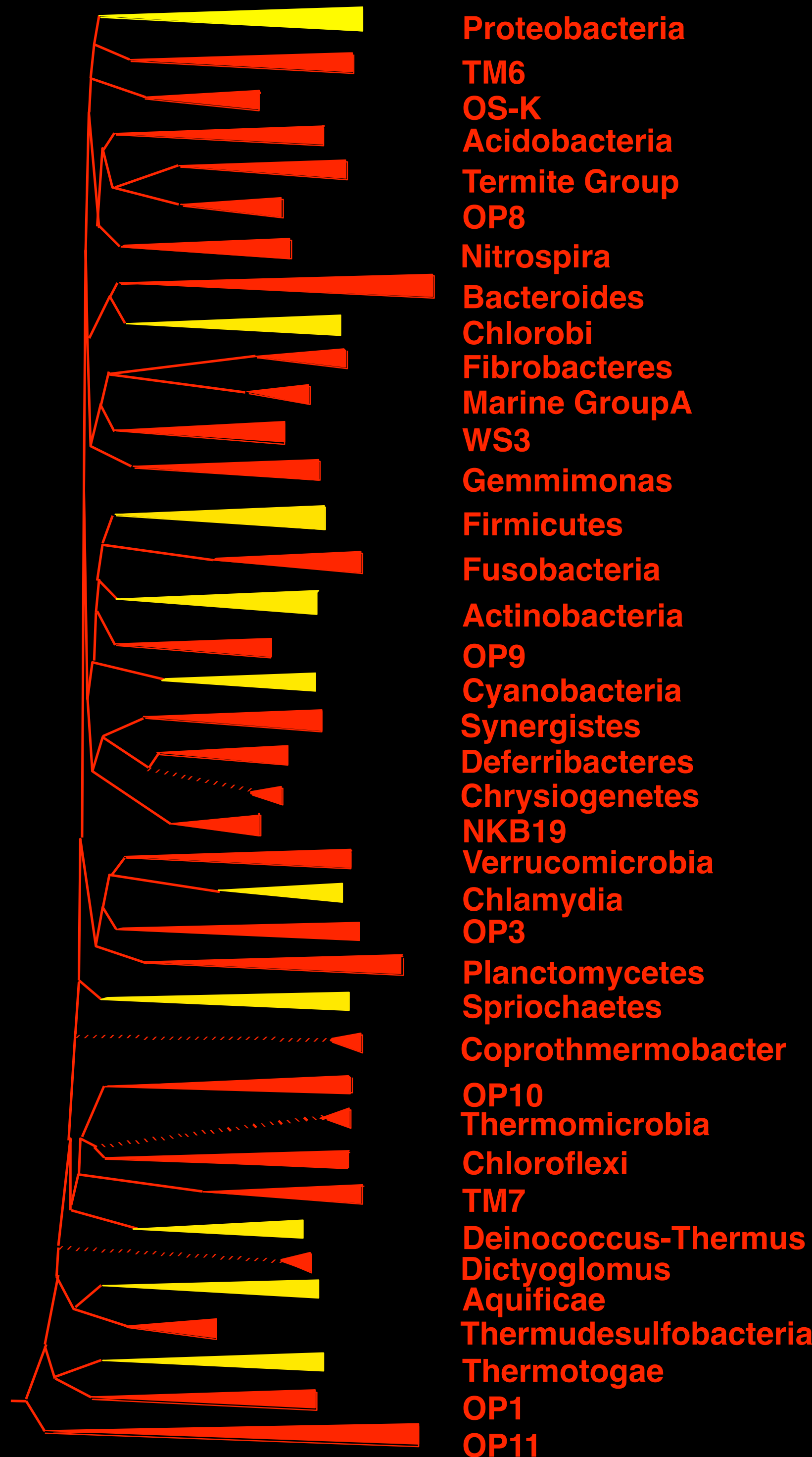
- At least 40 phyla of bacteria



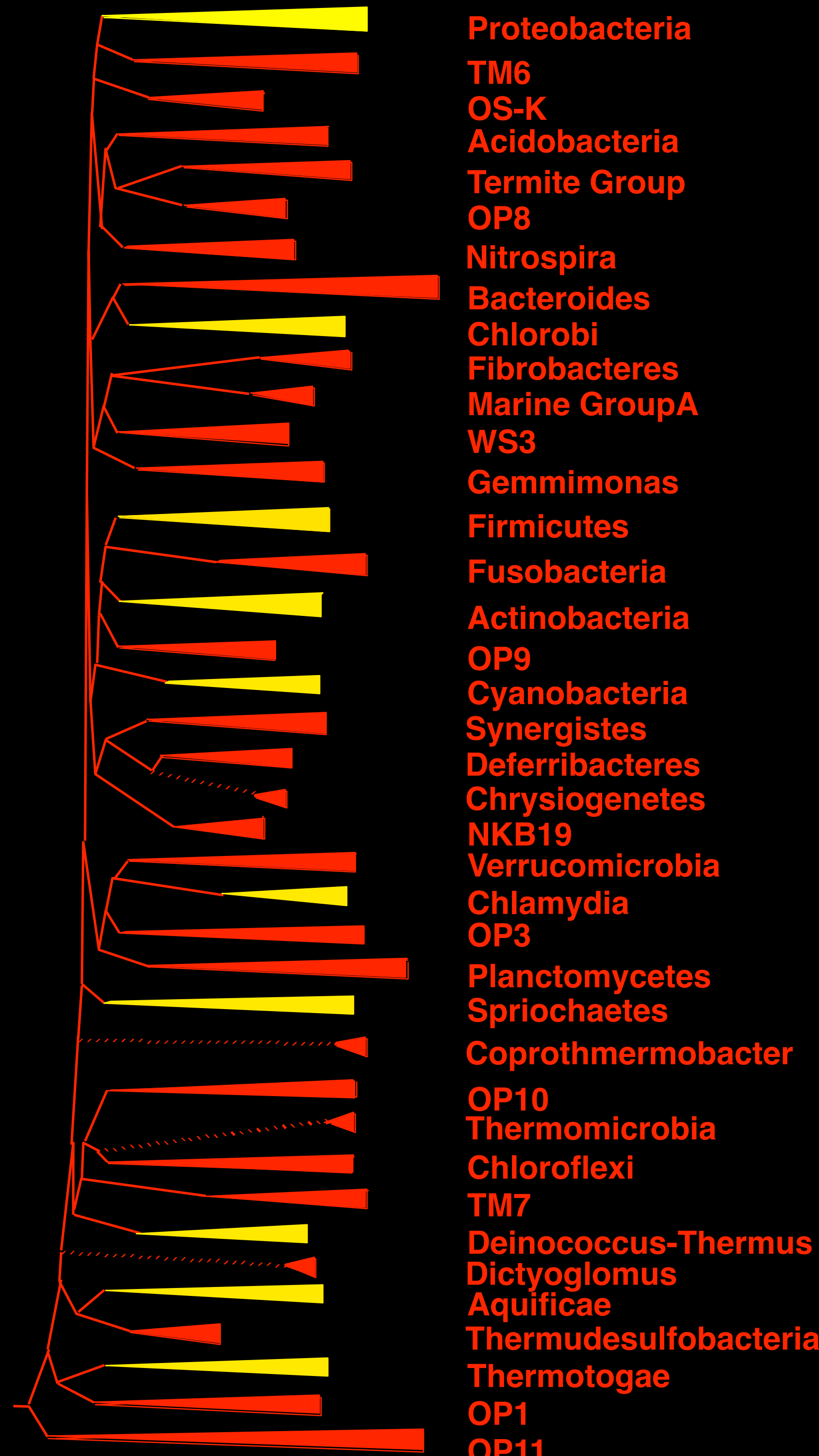
- At least 40 phyla of bacteria
- Most genomes from three phyla



- At least 40 phyla of bacteria
- Most genomes from three phyla
- Some studies in other phyla



- At least 40 phyla of bacteria
- Most genomes from three phyla
- Some other phyla are only sparsely sampled
- Same trend in Eukaryotes



- At least 40 phyla of bacteria
- Most genomes from three phyla
- Some other phyla are only sparsely sampled
- Same trend in Viruses

Tree of Life -- Home

http://www.tigr.org/tol/

Berkeley Ph... Home Page Pathema Mutation Rates Msg Cal ISC login BITS Tet PKS/NRPS JTC Giant

TREE of LIFE

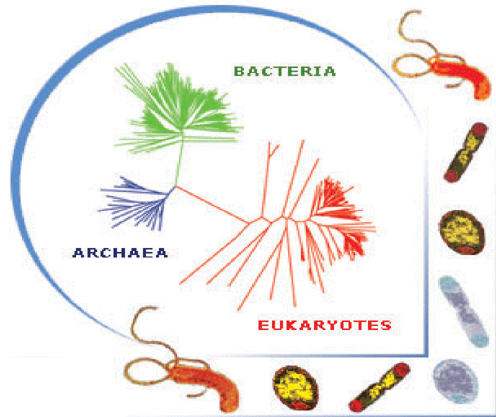
Phylogenomics: A Genome Level Approach to Assembling the Bacterial Branches of the Tree of Life

- Home
- What's New
- Project Summary
- People
- Results
- News/Press
- Broader Impact
- Links
- SAB
- TIGR

Home

In this project, the power of genome sequencing will be harnessed to better understand the bacterial branches of the Tree of Life. This "phylogenomics" project, which will integrate genomic and phylogenetic studies, is one segment of the National Science Foundation's "Assembling the Tree of Life" program, announced this week.

The first complete genome sequence of a free-living organism was determined at The Institute for Genomic Research (TIGR) in 1995. In the years since then, complete genome sequences have been determined for a wide diversity of species including bacterial pathogens, extremophiles, protistan parasites, plants, animals, and fungi. While these include representatives from each of the three major branches in the tree of life (the Bacteria, Archaea and Eukarya) only a limited subset of each of these groups has been studied. In particular, despite the vast diversity of bacterial species and their importance as pathogens and the foundation of many ecosystems, many major groups have been neglected.



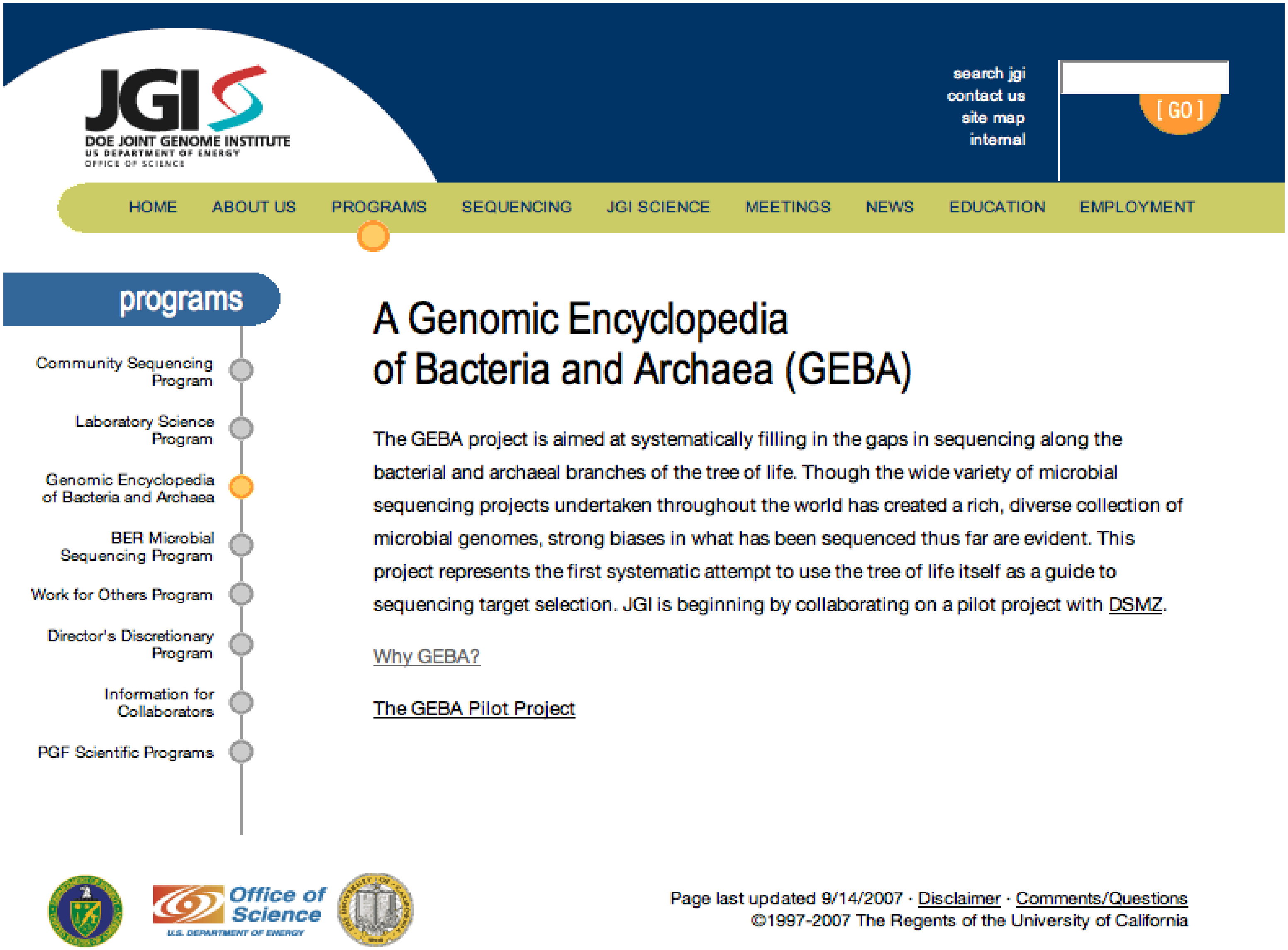
The goal of this project, a collaboration between TIGR and the Center of Marine Biotechnology (COMB), is to fill some of the gaps in our knowledge in the bacterial domain. Specifically, this project involves sequencing the complete genomes of representative strains of eight bacterial phyla. Each phylum represents a major branch in the "Tree of Life" and the phyla that were selected (Chrysiogenetes, Deferribacteres, Dictyoglomus, Nitrospira, Coprothermobacter, Synergistes, Thermodesulfobacteria, and Thermomicrobium) have not yet been explored using whole genome sequencing.

The data from these genome sequences will then be used

1. to improve the understanding of the complex phylogenetic relationships among major bacterial phyla
2. to provide information and resources that will allow scientists to examine the evolutionary relationships within these poorly understood phyla in more depth than is now possible and
3. to launch experimental studies on the biology and physiology of organisms in these phyla.

Funded by NSF Tree of Life Program

 Last Update:07/02/04


search jgi
contact us
site map
internal
[GO]

HOME | ABOUT US | PROGRAMS | SEQUENCING | JGI SCIENCE | MEETINGS | NEWS | EDUCATION | EMPLOYMENT

programs

- Community Sequencing Program
- Laboratory Science Program
- Genomic Encyclopedia of Bacteria and Archaea
- BER Microbial Sequencing Program
- Work for Others Program
- Director's Discretionary Program
- Information for Collaborators
- PGF Scientific Programs

A Genomic Encyclopedia of Bacteria and Archaea (GEBA)

The GEBA project is aimed at systematically filling in the gaps in sequencing along the bacterial and archaeal branches of the tree of life. Though the wide variety of microbial sequencing projects undertaken throughout the world has created a rich, diverse collection of microbial genomes, strong biases in what has been sequenced thus far are evident. This project represents the first systematic attempt to use the tree of life itself as a guide to sequencing target selection. JGI is beginning by collaborating on a pilot project with DSMZ.

Why GEBA?

The GEBA Pilot Project

Page last updated 9/14/2007 · Disclaimer · Comments/Questions

http://www.jgi.doe.gov/programs/GEBA/pilot.html

GEBA Lesson 1:
The rRNA Tree of Life is a Useful Tool

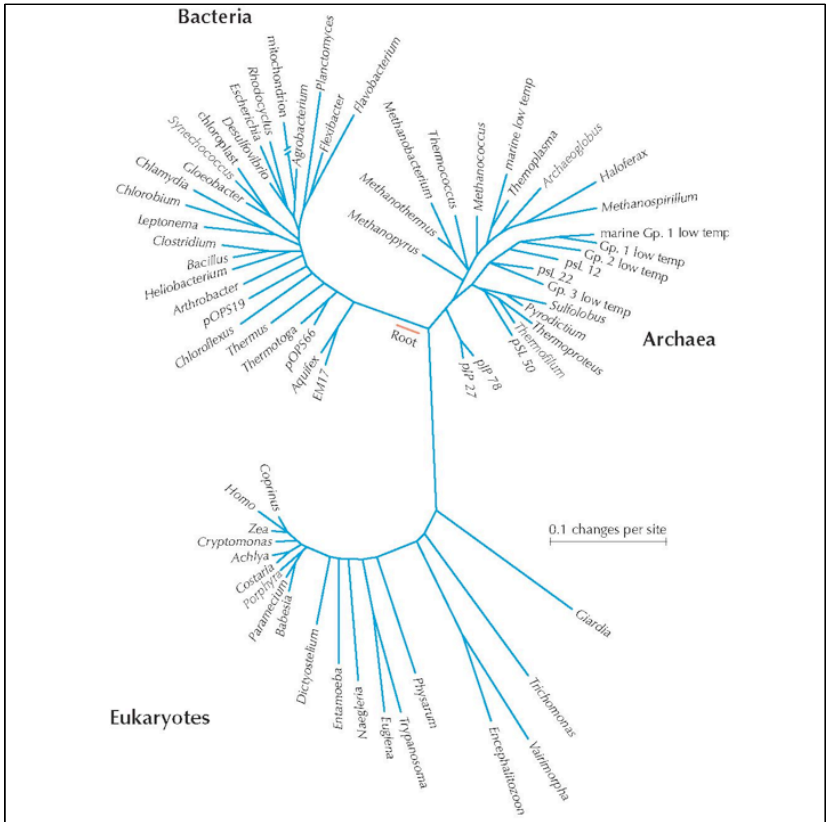
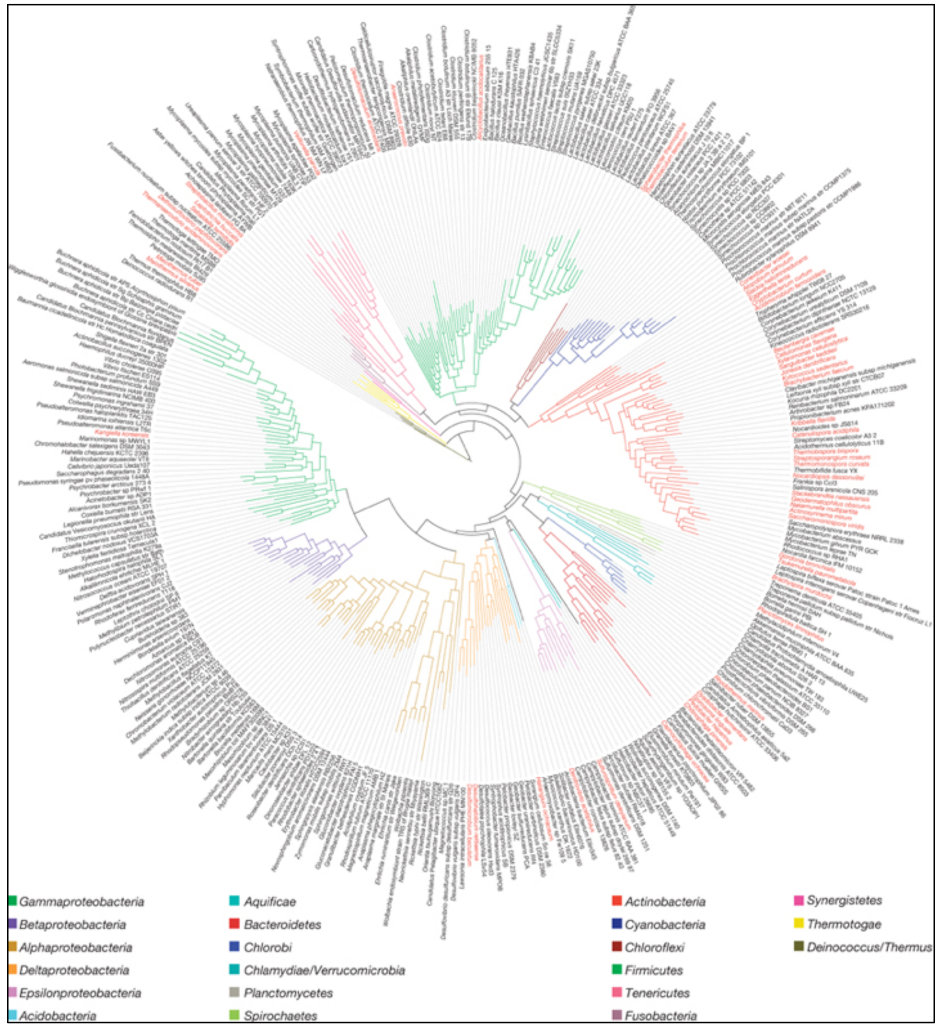


FIGURE 5.17. A phylogenetic tree with representatives of each of the three domains of life. The tree is based primarily on analysis of sequences of small-subunit rRNA molecules. It is believed that the positions of some organisms in this tree do not accurately reflect their true phylogenetic position (e.g., see Box 8.1 for a more thorough discussion of the evolution of eukaryotes and the back endpaper for a more up-to-date tree). Nevertheless, the division of life into three domains is supported by analyses of many other characteristics.

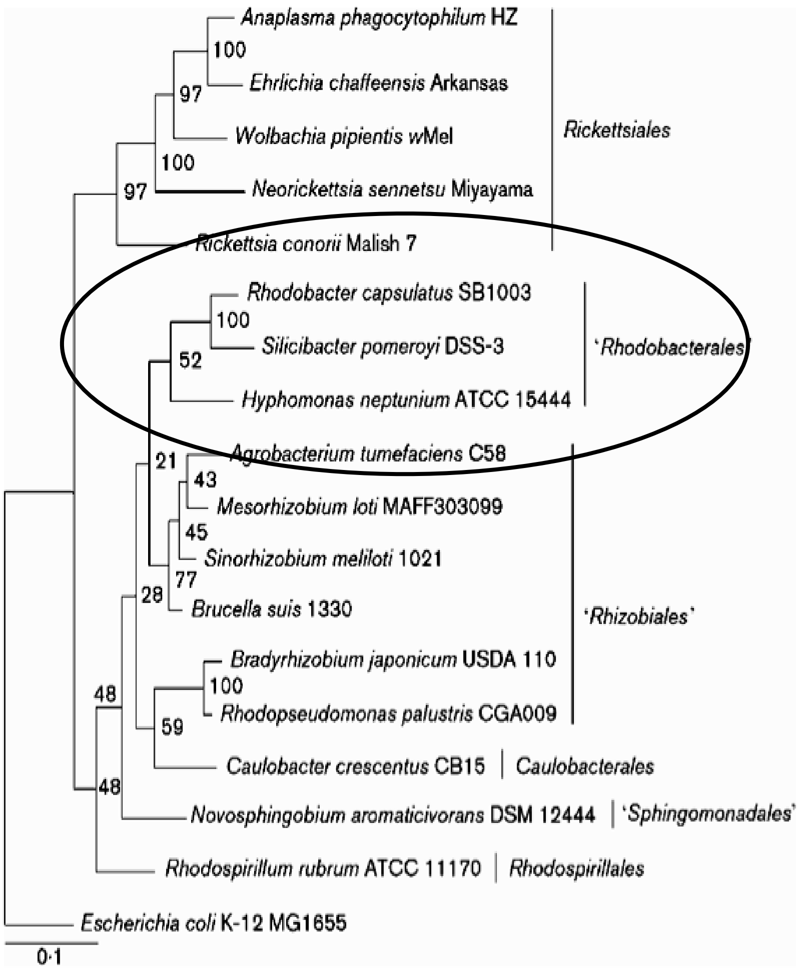
5.17, redrawn from Pace N.R., *Science* 276: 734–740, © 1977 American Association for the Advancement of Science



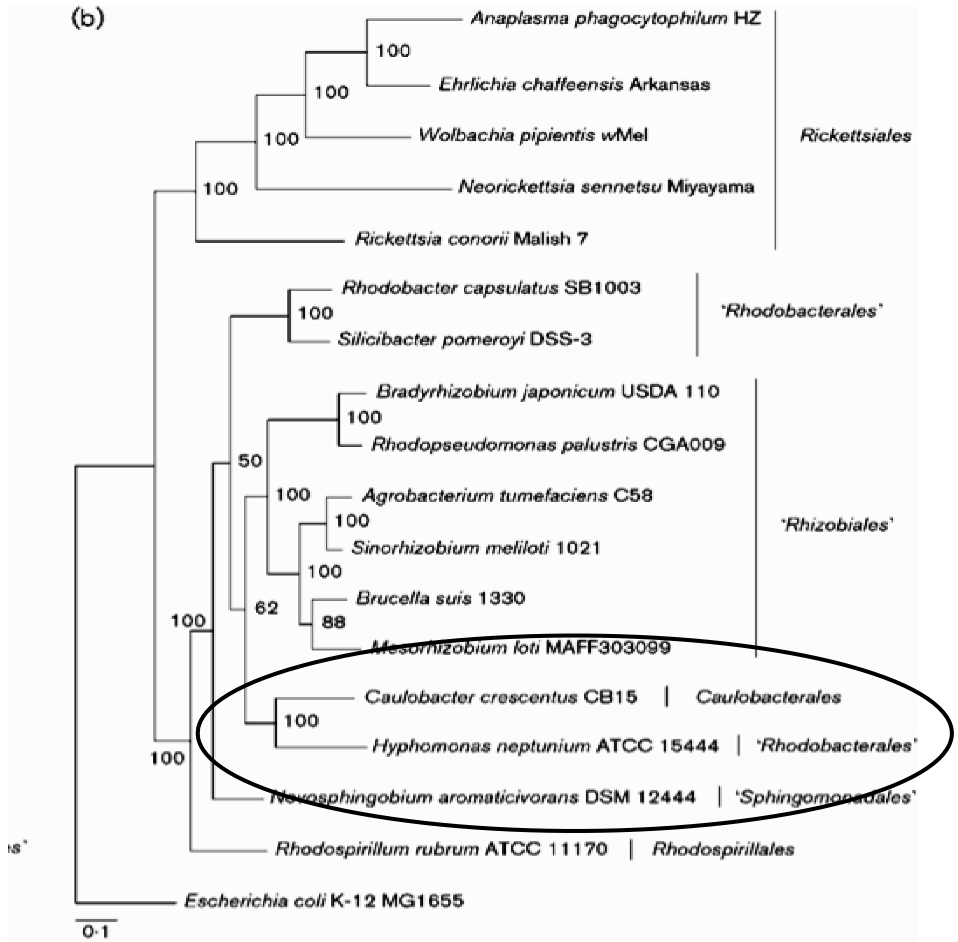
From Wu et al. 2009 Nature 462, 1056-1060

GEBA Lesson 2:
The rRNA Tree of Life is not perfect ...

16s



WGT, 23S



Badger et al. 2005 Int J System Evol Microbiol 55: 1021-1026.

GEBA Lesson 3: Phylogeny improves genome annotation

- Took 56 GEBA genomes and compared results vs. 56 randomly sampled new genomes
- Better definition of protein family sequence “patterns”
- Greatly improves “comparative” and “evolutionary” based predictions
- Conversion of hypothetical into conserved hypotheticals
- Linking distantly related members of protein families
- Improved non-homology prediction





GEBA Lesson 4 : Metadata Important

SIGS

Standards in Genomic Sciences

An Open Access Journal of the Genomic Standards Consortium

FOUNDING MEMBERS

Sam Angiuoli
Baltimore, MD, USA

Patrick Chain
Los Alamos, NM, USA

Dawn Field
Oxford, UK

HOME ABOUT LOG IN REGISTER
SEARCH CURRENT ARCHIVES
ANNOUNCEMENTS FOR AUTHORS

Home > **Standards in Genomic Sciences**

Standards in Genomic Sciences

Standards in Genomic Sciences (2009) 1:21-28 DOI:4056/sigs.1162

Complete genome sequence of *Beutenbergia cavernae* type strain (HKI 0122^{T})

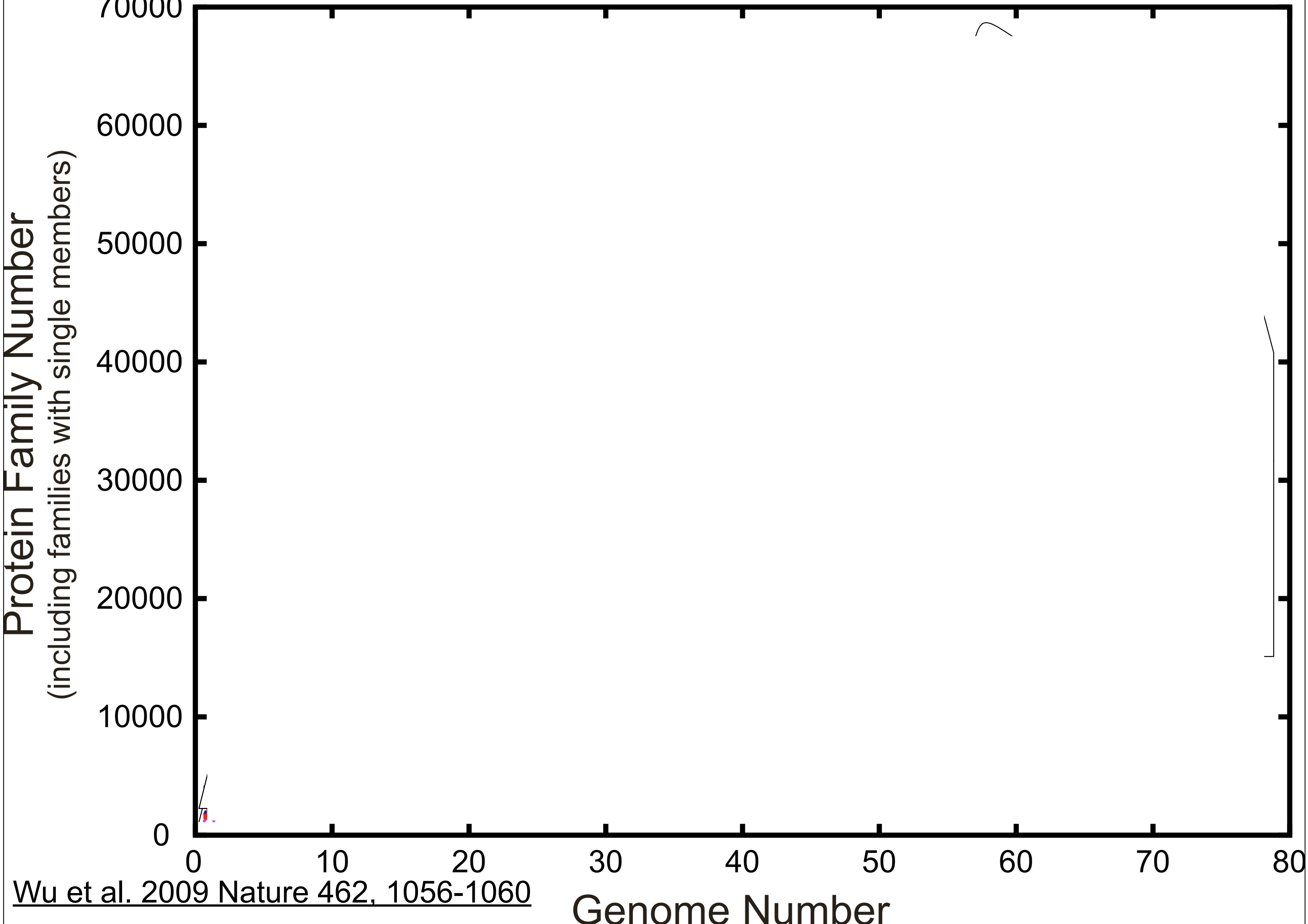
Miriam Land[1,2], Rüdiger Pukall[3], Birte Abt[3], Markus Göker[3], Manfred Rohde[4], Tijana Glavina Del Rio[1], Hope Tice[1], Alex Copeland[1], Jan-Fang Cheng[1], Susan Lucas[1], Feng Chen[1], Matt Nolan[1], David Bruce[1,5], Lynne Goodwin[1,5], Sam Pitluck[1], Natalia Ivanova[1], Konstantinos Mavromatis[1], Galina Ovchinnikova [1], Amrita Pati[1], Amy Chen[6], Krishna Palaniappan[6], Loren Hauser[1,2], Yun-Juan Chang[1,2], Cynthia C. Jefferies[1,2], Elizabeth Saunders[5], Thomas Brettin[1,5], John C. Detter[1,5], Cliff Han[1,5], Patrick Chain[1,7], James Bristow[1], Jonathan A. Eisen[1,8], Victor Markowitz[6], Philip Hugenholtz[1], Nikos C. Kyrpides[1], Hans-Peter Klenk[3], and Alla Lapidus[1*]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[3] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[4] HZI - Helmholtz Centre for Infection Research, Braunschweig, Germany
[5] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico USA
[6] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[7] Lawrence Livermore National Laboratory, Livermore, California, USA
[8] University of California Davis Genome Center, Davis, California, USA

GEBA Lesson 5: Improves discovering new genetic diversity

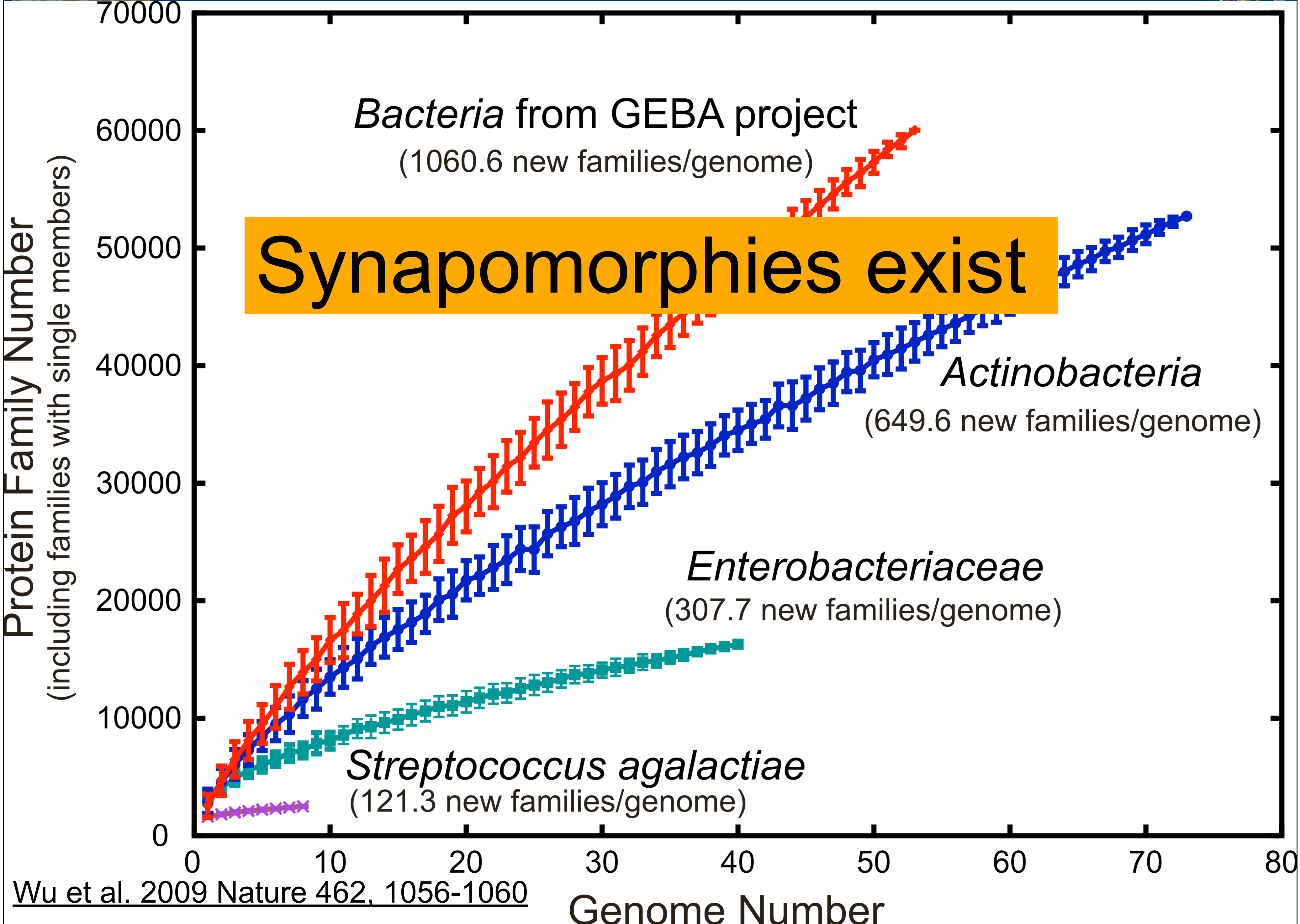
Protein Family Rarefaction

- Take data set of multiple complete genomes
- Identify all protein families using MCL
- Plot # of genomes vs. # of protein families



Wu et al. 2009 Nature 462, 1056-1060

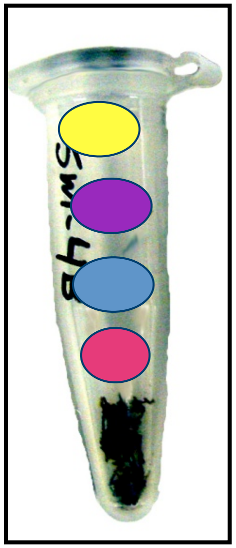
Synapomorphies exist



Wu et al. 2009 Nature 462, 1056-1060

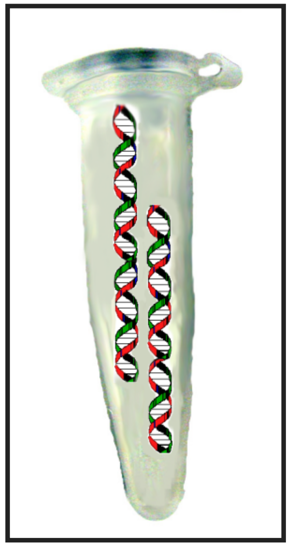
GEBA Lesson 6:
Improves Analysis of Uncultured

rRNA Phylotyping



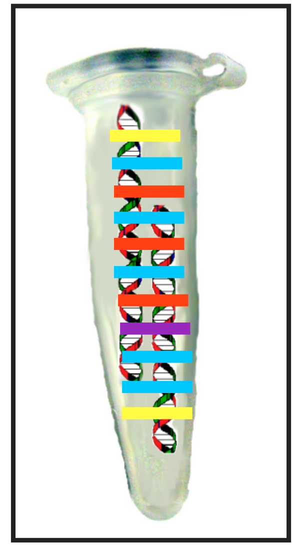
DNA extraction

PCR



PCR

Makes lots of copies of the rRNA genes in sample



Sequence rRNA genes

rRNA1
5'...ACACACATAGGTGGAGCTA
GCGATCGATCGA... 3'

rRNA2
5'..TACAGTATAGGTGGAGCTAG
CGACGATCGA... 3'

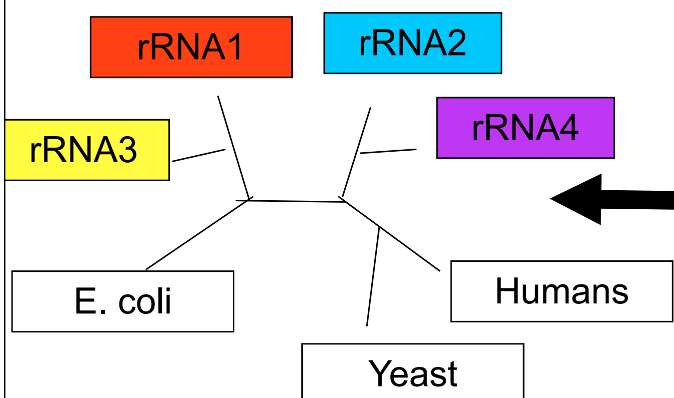
rRNA3
5'...ACGGCAAAATAGGTGGATT
CTAGCGATATAGA... 3'

rRNA4
5'...ACGGCCCGATAGGTGGATT
CTAGCGCCATAGA... 3'

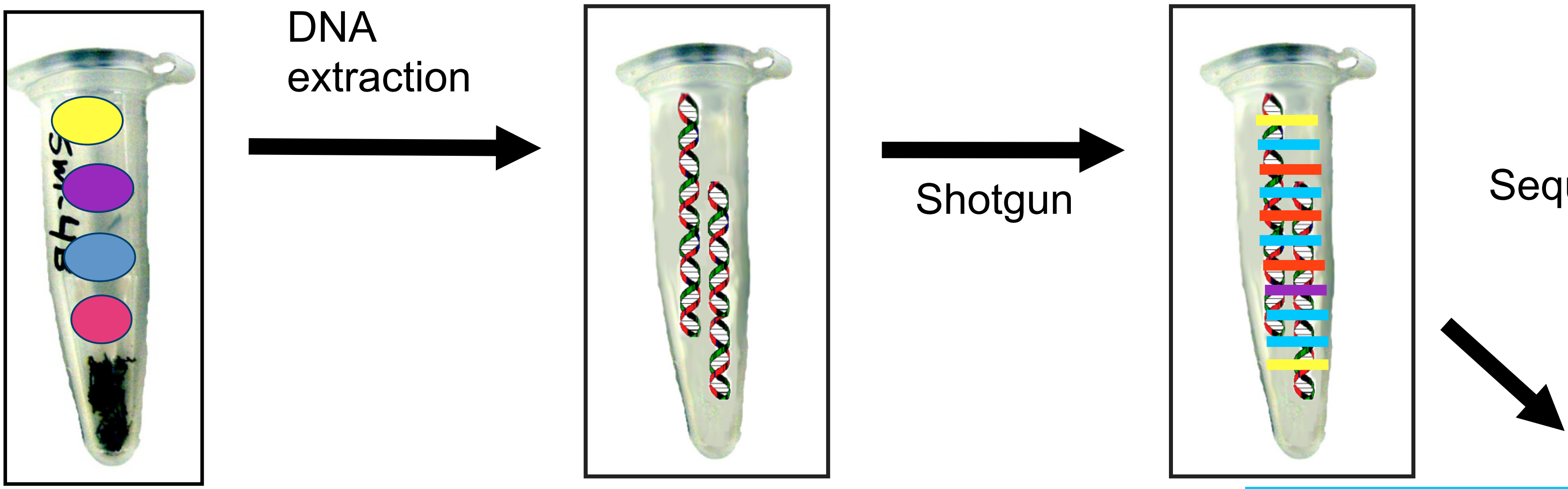
Sequence alignment = Data matrix

rRNA1	A	C	A	C	A	C
rRNA2	T	A	C	A	G	T
rRNA3	C	A	C	T	G	T
rRNA4	C	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T
Yeast	T	A	C	A	G	T

Phylogenetic tree



Metagenomic Phylotyping



Gene 1
5'...ACACACATAGGTGGAGCTA
GCGATCGATCGA... 3'

Gene 2
5'..TACAGTATAGGTGGAGCTAG
CGACGATCGA... 3'

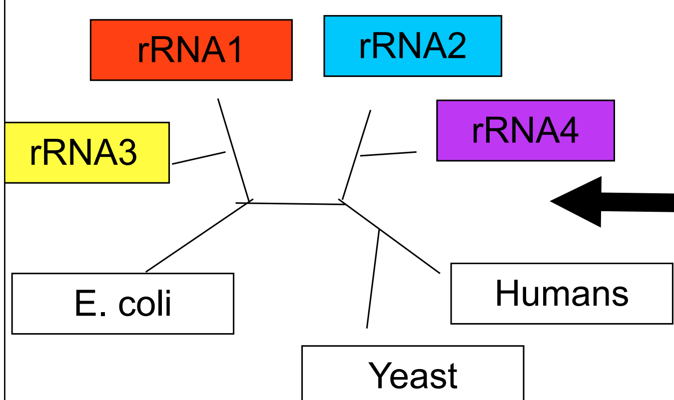
Gene 3
5'...ACGGCAAAATAGGTGGATT
CTAGCGATATAGA... 3'

Gene 4
5'...ACGGCCCGATAGGTGGATT
CTAGCGCCATAGA... 3'

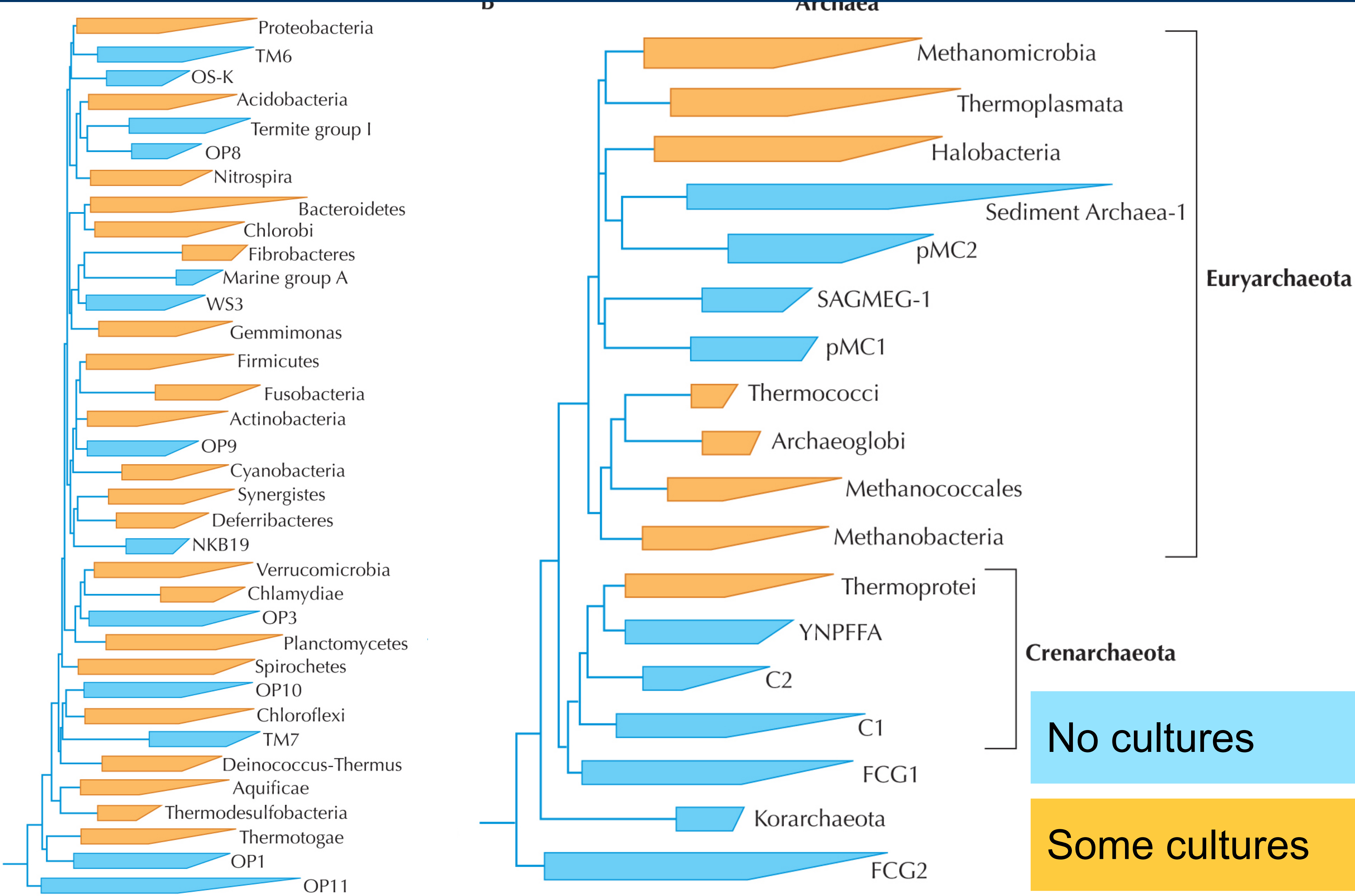
Sequence alignment = Data matrix

Gene 1	A	C	A	C	A	C
Gene 2	T	A	C	A	G	T
Gene 3	C	A	C	T	G	T
Gene 4	C	A	C	A	G	T
E. coli	A	G	A	C	A	G
Humans	T	A	T	A	G	T
Yeast	T	A	C	A	G	T

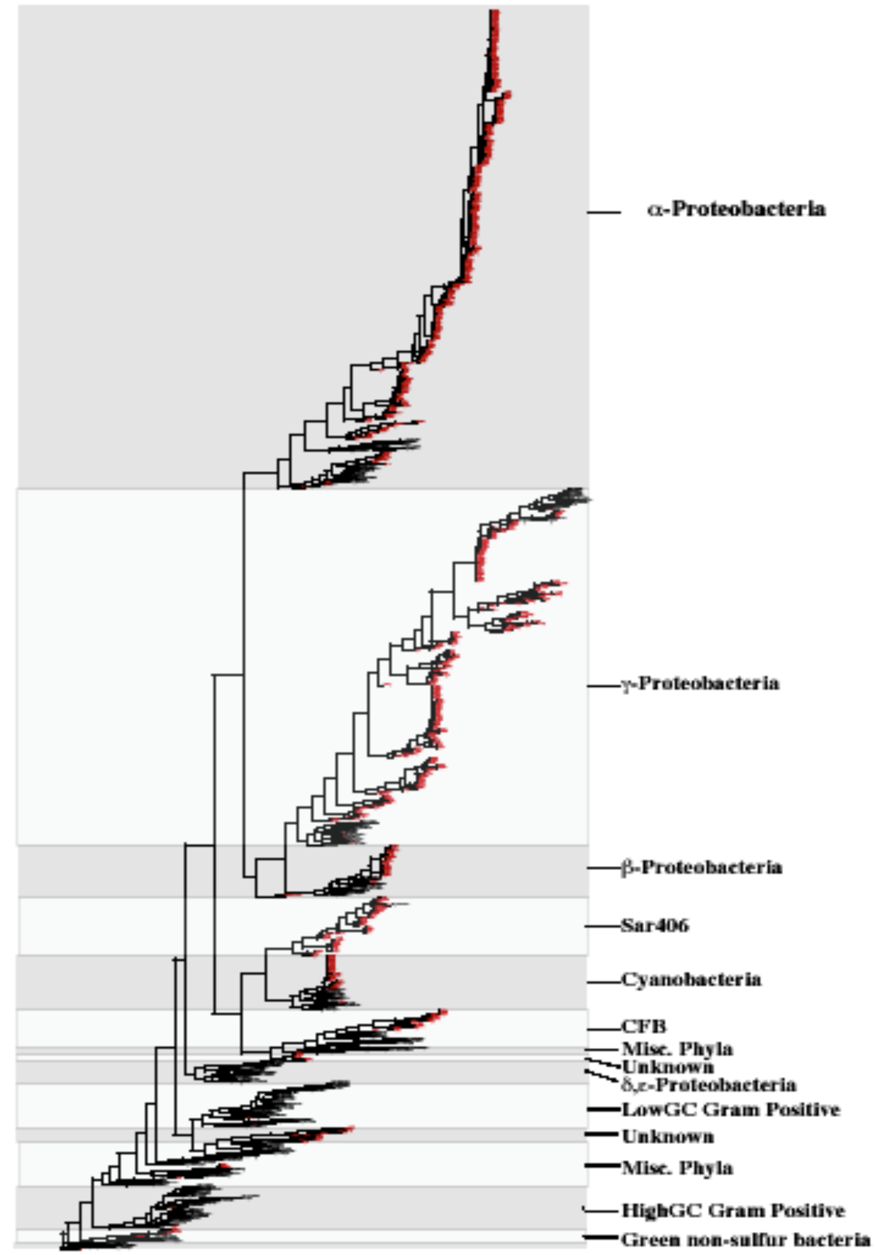
Phylogenetic tree



Major phyla of bacteria & archaea



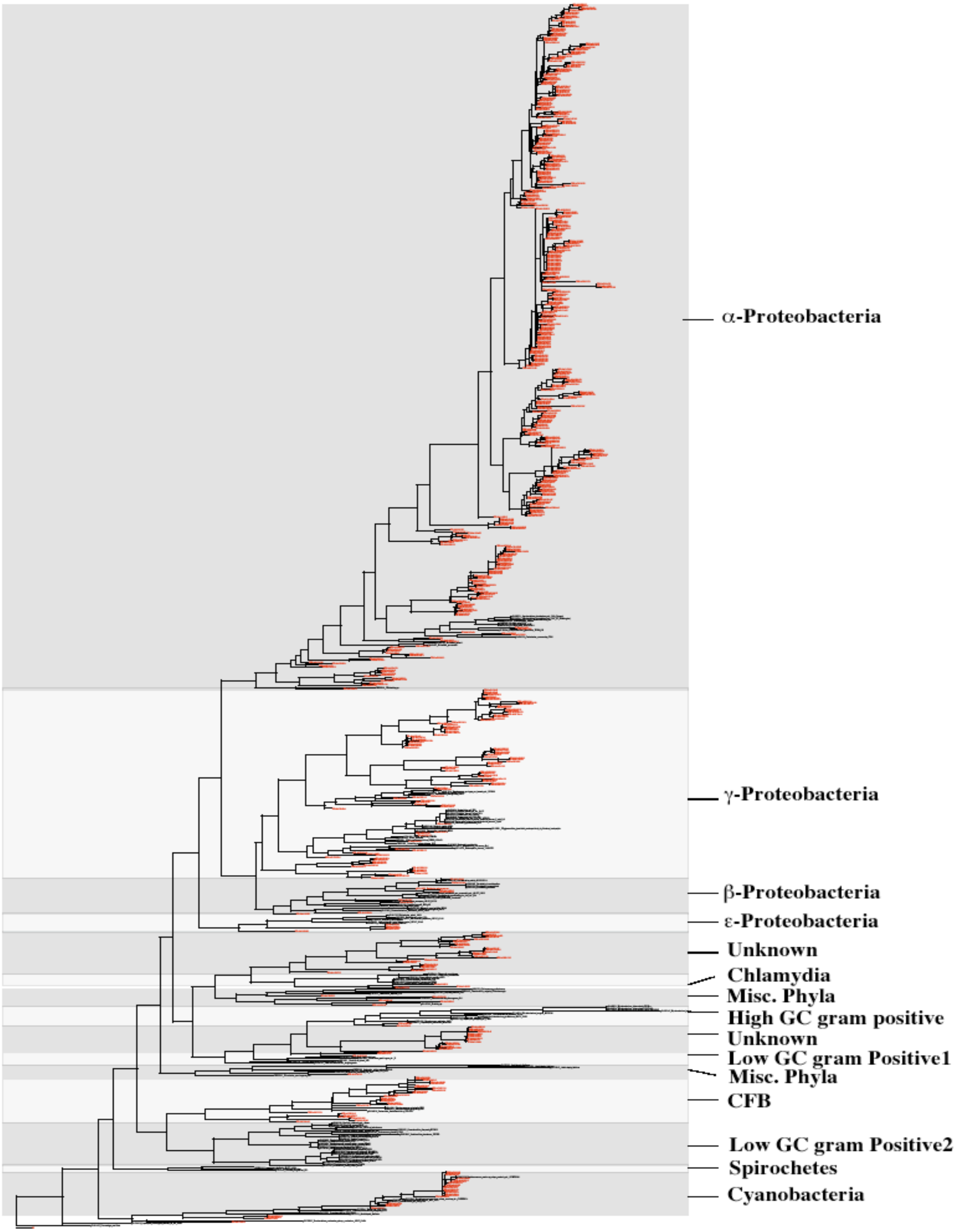
rRNA Phylotyping Sargasso Data



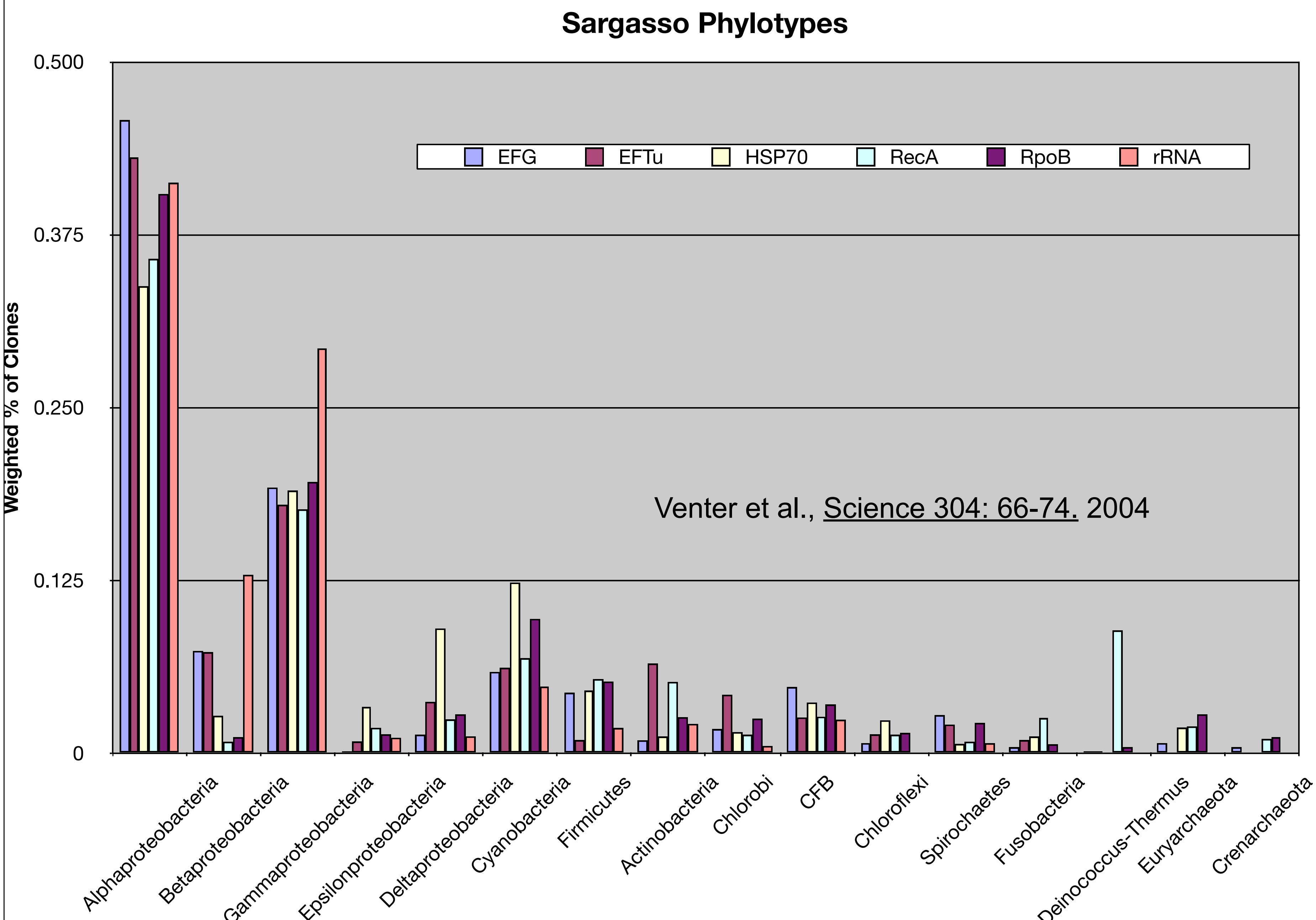
Venter et al., Science 304: 66. 2004

RecA Phylotyping Sargasso Data

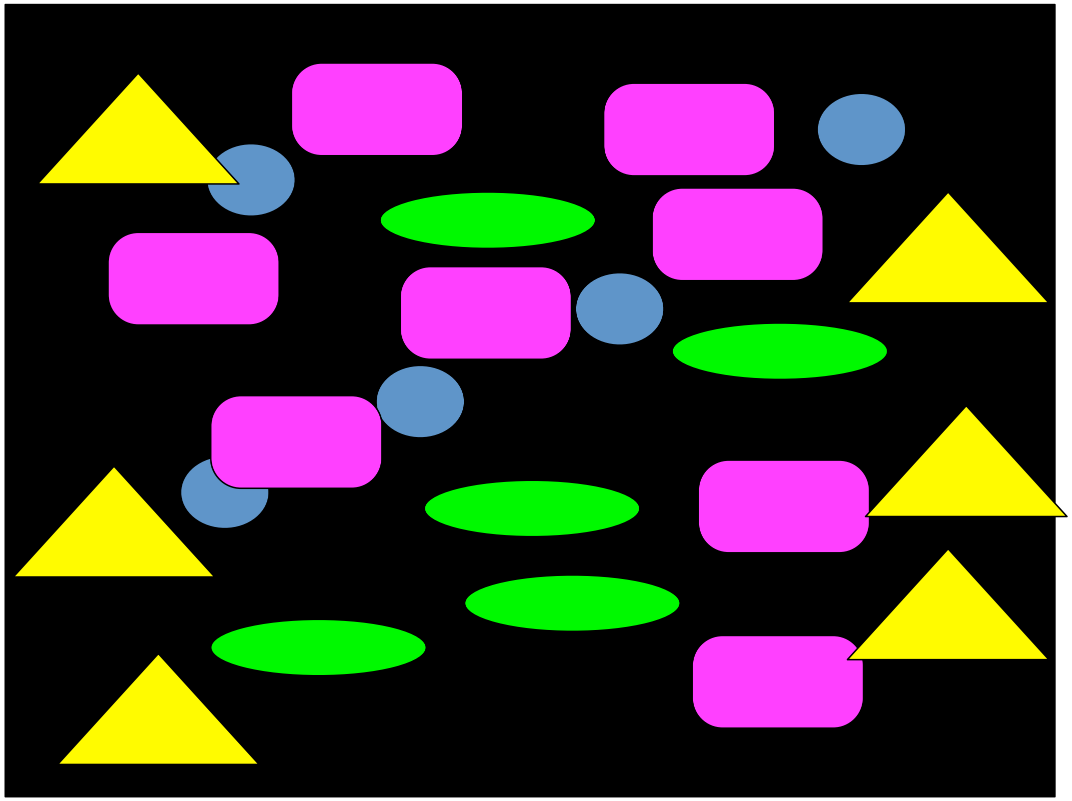
Venter et al., Science 304: 66. 2004



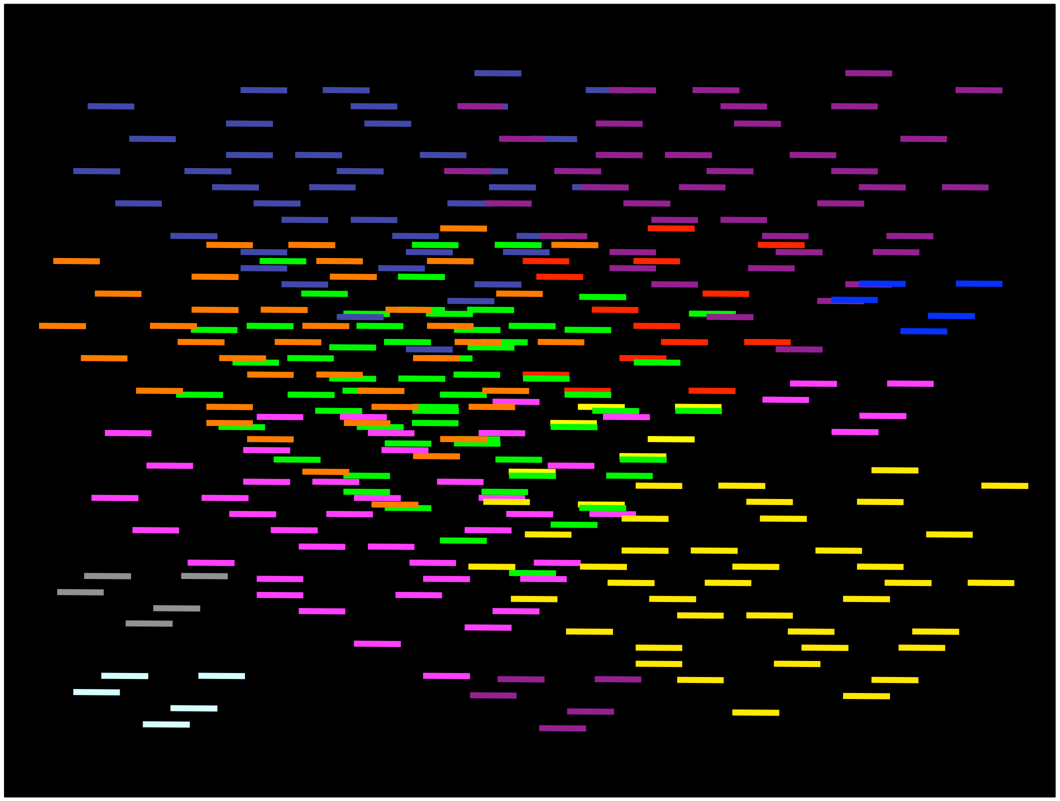
Protein vs. rRNA Sargasso Data



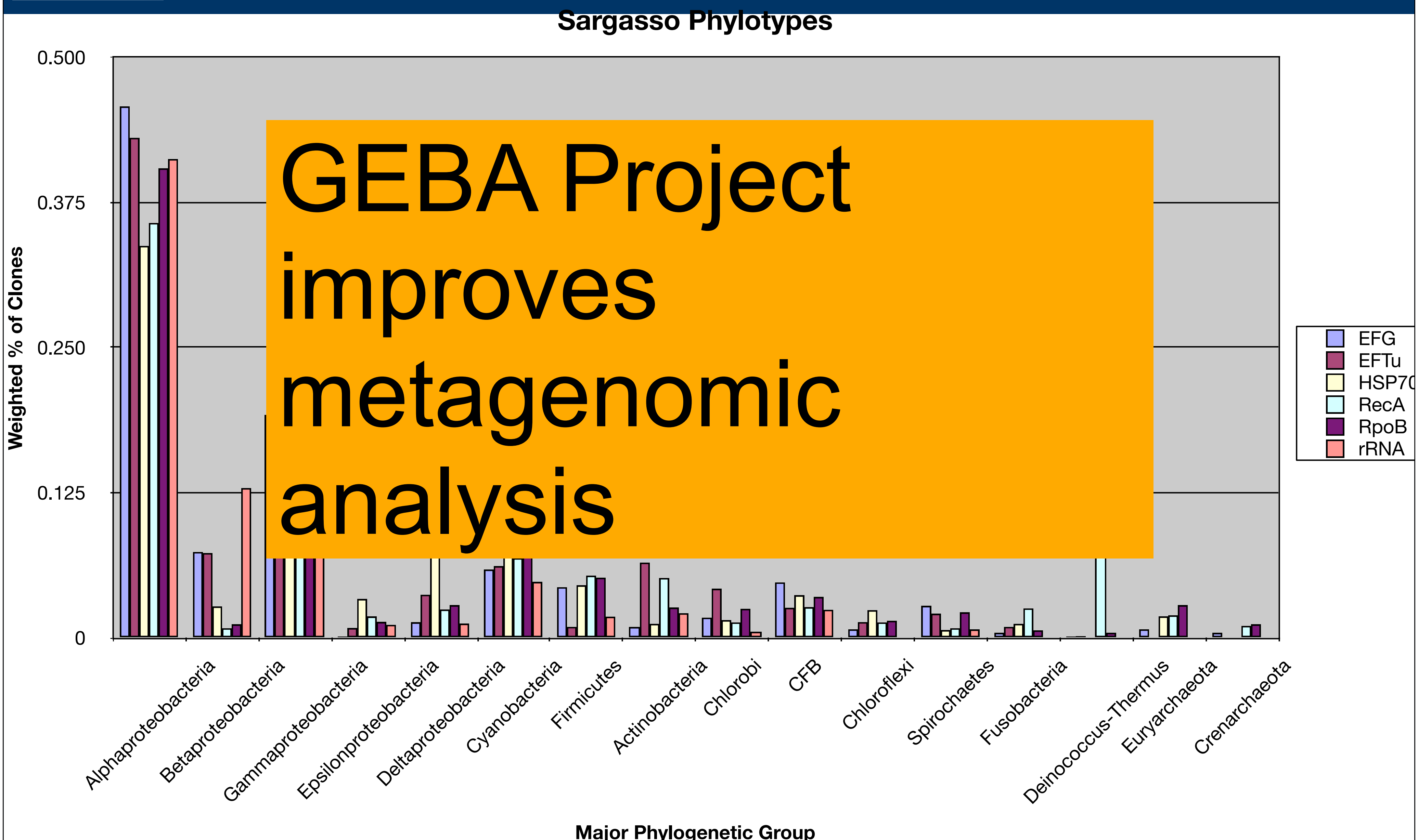
Metagenomics



Binning challenge



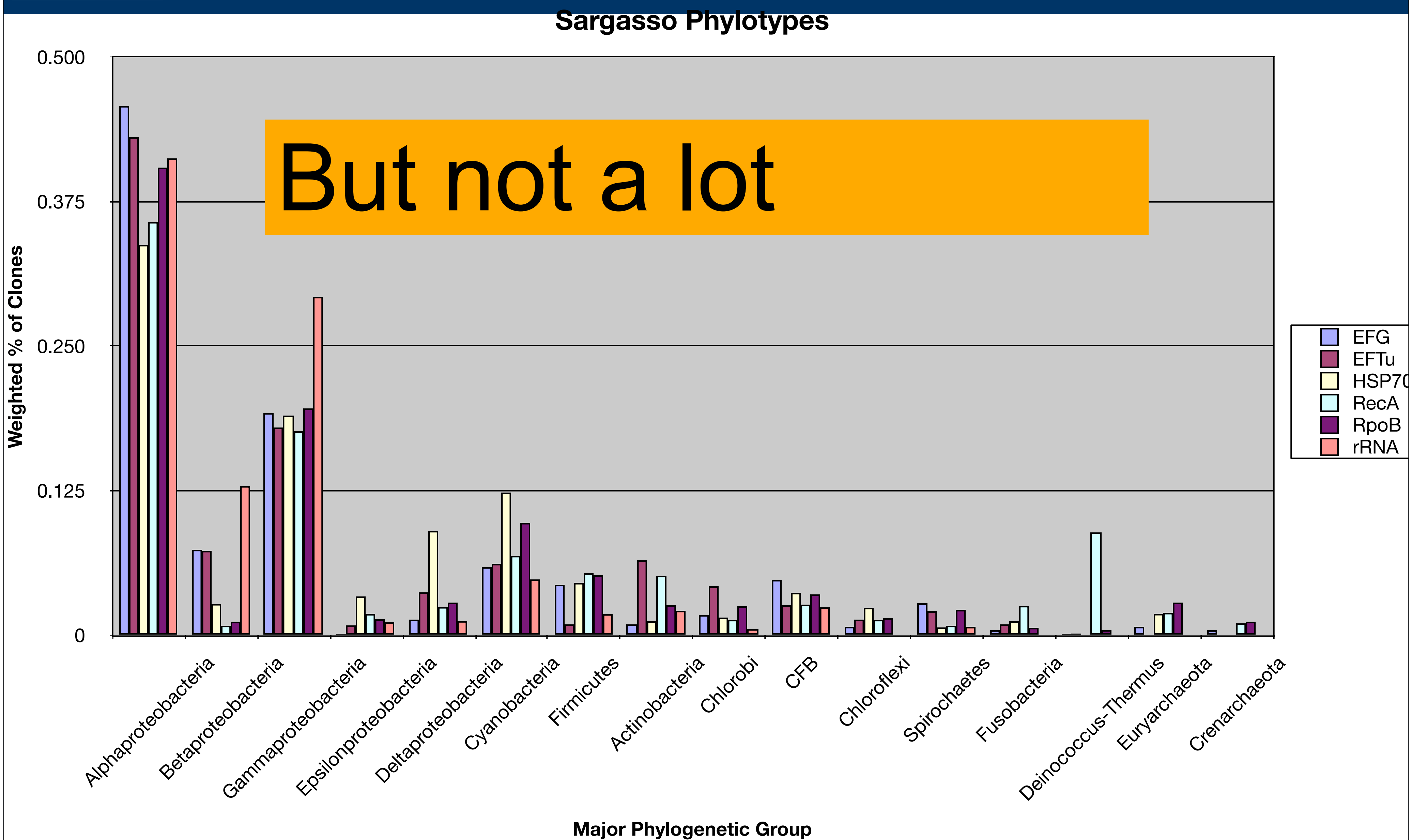
Other Markers



GEBA Project improves metagenomic analysis

Venter et al., Science 304: 66-74, 20

Other Markers



Venter et al., Science 304: 66-74, 20

GEBA Now

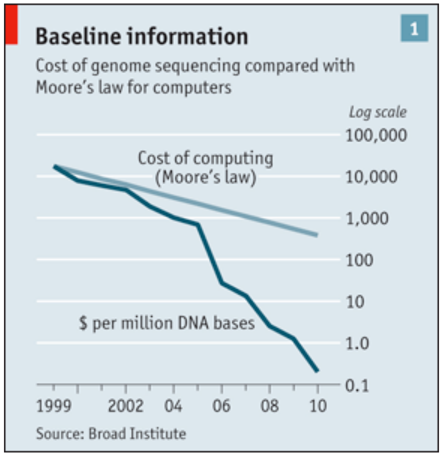
- 300+ genomes
- Rich sampling of major groups of cultured organisms
- Zoomed in sampling of haloarchaea, cyanobacteria and more

Part 2: Better Methods

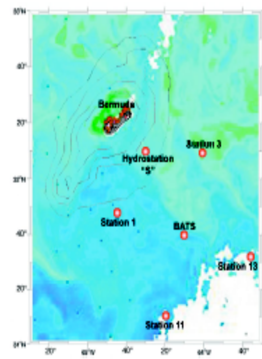
I: Better Phylogenetic Methods

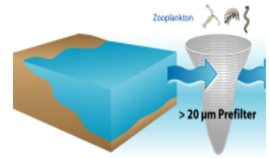
Major Issues in Phylotpying

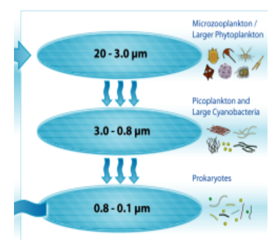
Beyond Moore's Law

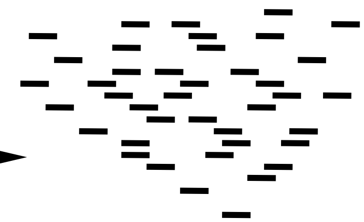


Metagenomics

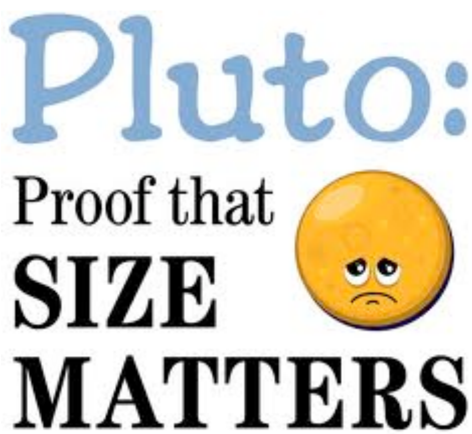






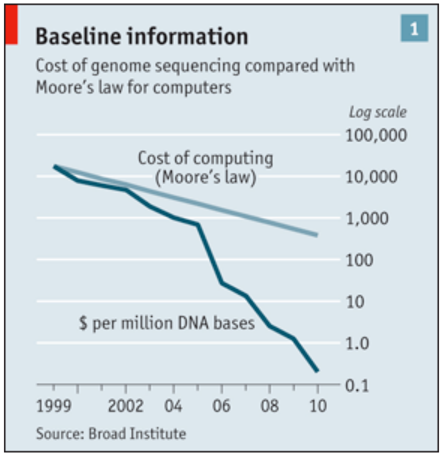


Short reads

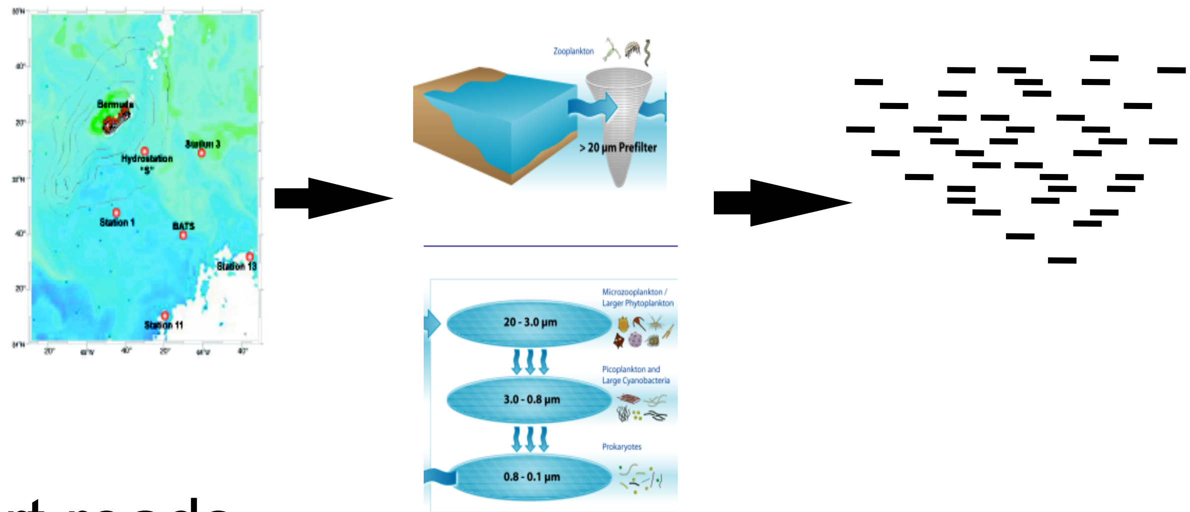


Major Issues in Phylotpying

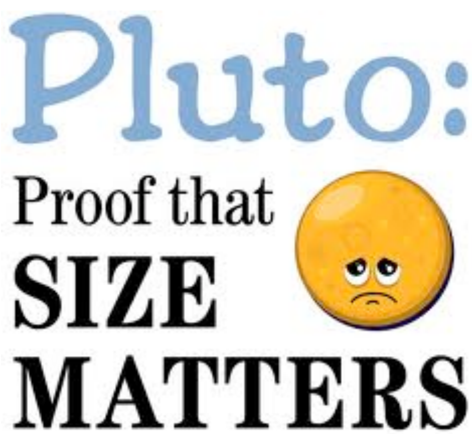
Beyond Moore's Law



Metagenomics



Short reads



WE NEED NEW METHODS

Method 1: Each is an island

- Each new sequences is an island

- Take reference data
- Build alignment, models, trees
- Add new sequence to reference alignment and build tree

STAP

ss-rRNA query sequence

↓

Step 1: make domain assignment

↓

Step2: use BLASTN to select related data set

↓

build and trim alignments

↓

build a maximum likelihood tree

↓

make initial taxonomic assignment using mid-point rooting

↓

Step 3: create mini database of sequences one level up plus 2 outgroup sequences

↓

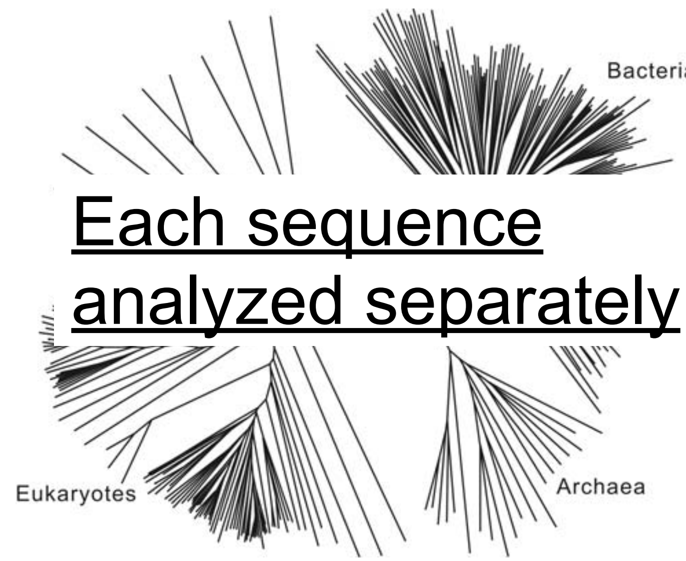
build and trim alignments

↓

build a maximum likelihood or neighbor-joining tree

↓

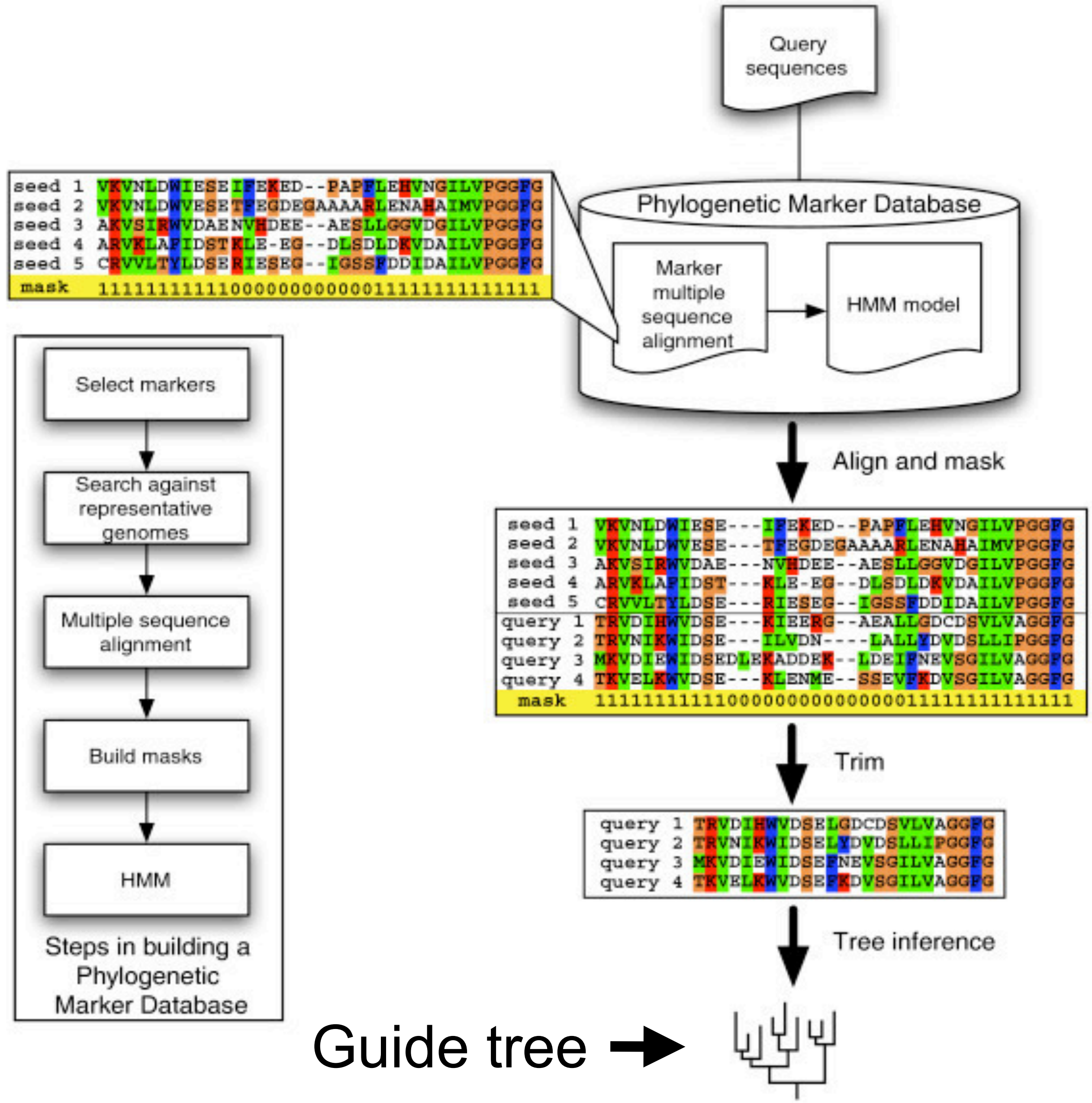
make taxonomic assignment using outgroup rooting



Each sequence analyzed separately

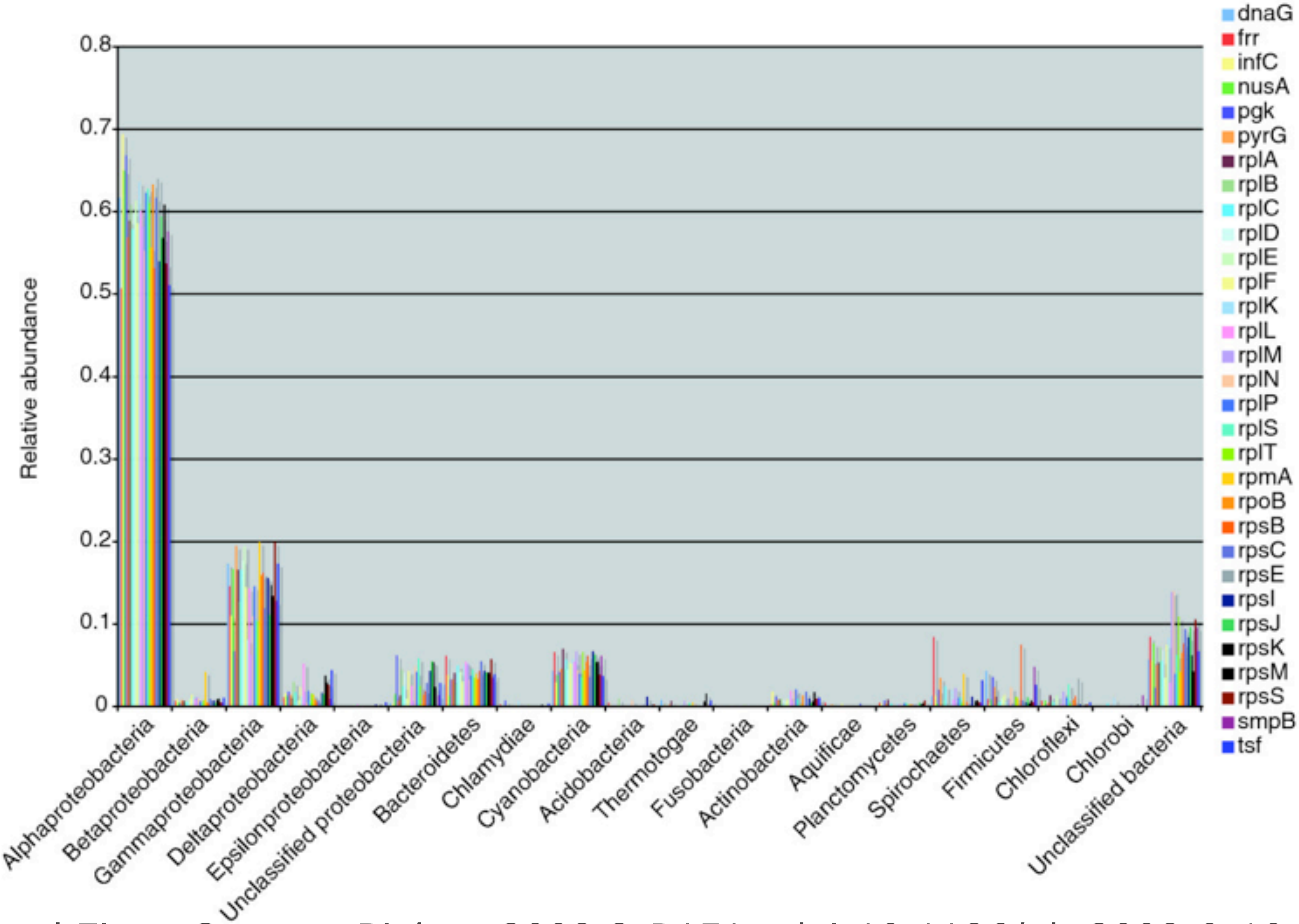
Wu et al. 2008 PLoS One

AMPHORA



Wu and Eisen Genome Biology 2008 9:R151 doi:10.1186/ gb-2008-9-10-r151

Phylotyping w/ Proteins



Wu and Eisen *Genome Biology* 2008 **9**:R151 doi:10.1186/gb-2008-9-10-r151

Method 2: All in the family

- Combine new sequences into one tree

- Take reference data
- Build alignment, models, trees
- Add all sequences to reference alignment and build tree

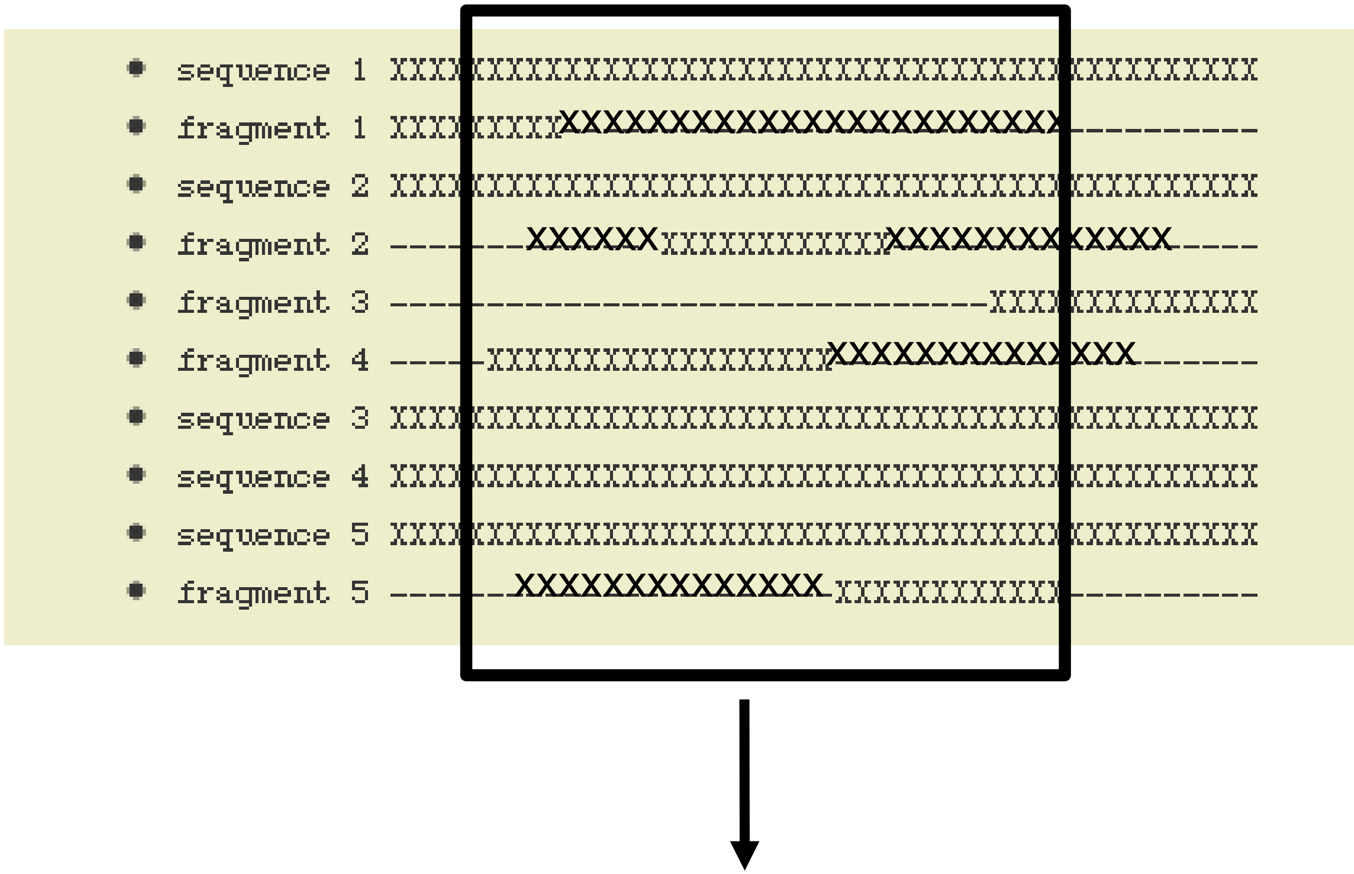
Phylogenetic Challenge

- sequence 1 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 1 XXXXXXXXX**XXXXXXXXXXXXXXXXXXXXXXXX**----------
- sequence 2 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 2 -------**XXXXXX**XXXXXXXXXXXX**XXXXXXXXXXXXX**----
- fragment 3 -------------------------------XXXXXXXXXXXXXX
- fragment 4 -----XXXXXXXXXXXXXXXXXX**XXXXXXXXXXXXXX**------
- sequence 3 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- sequence 4 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- sequence 5 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 5 ------**XXXXXXXXXXXXXX**-XXXXXXXXXXXX----------

↓

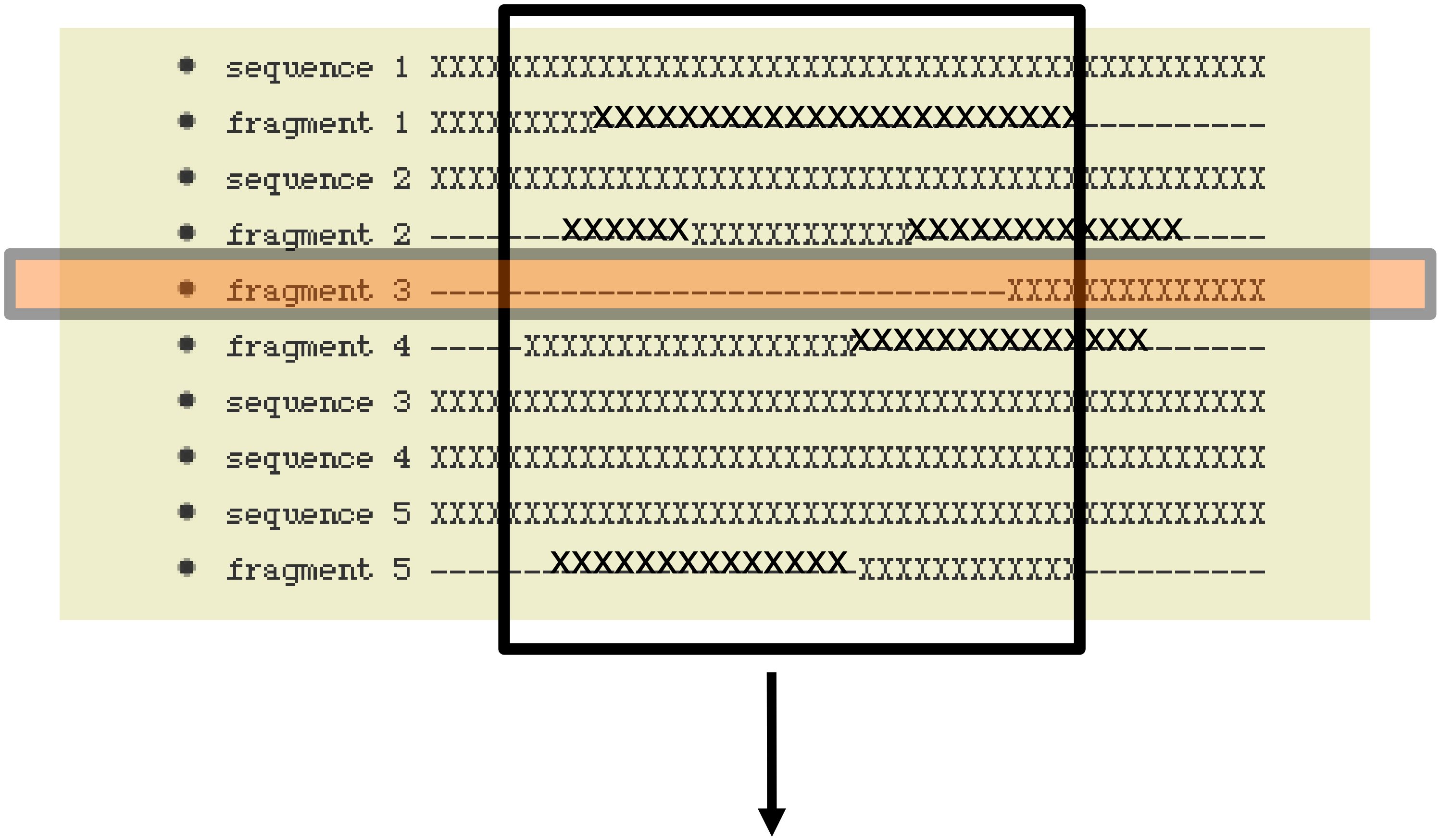
A single tree with everything?

Phylogenetic Challenge



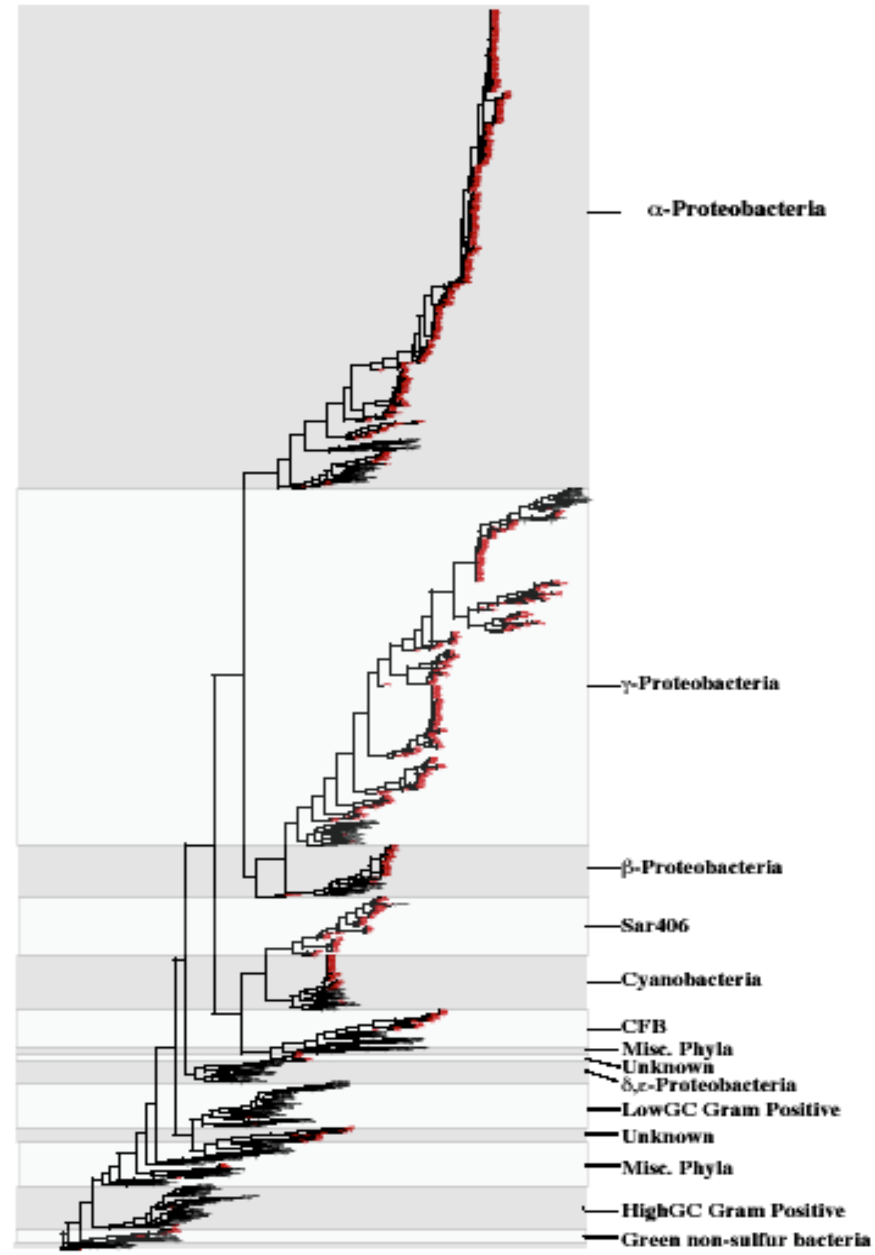
A single tree with everything
(as long as there is a lot of overlap)

Phylogenetic Challenge



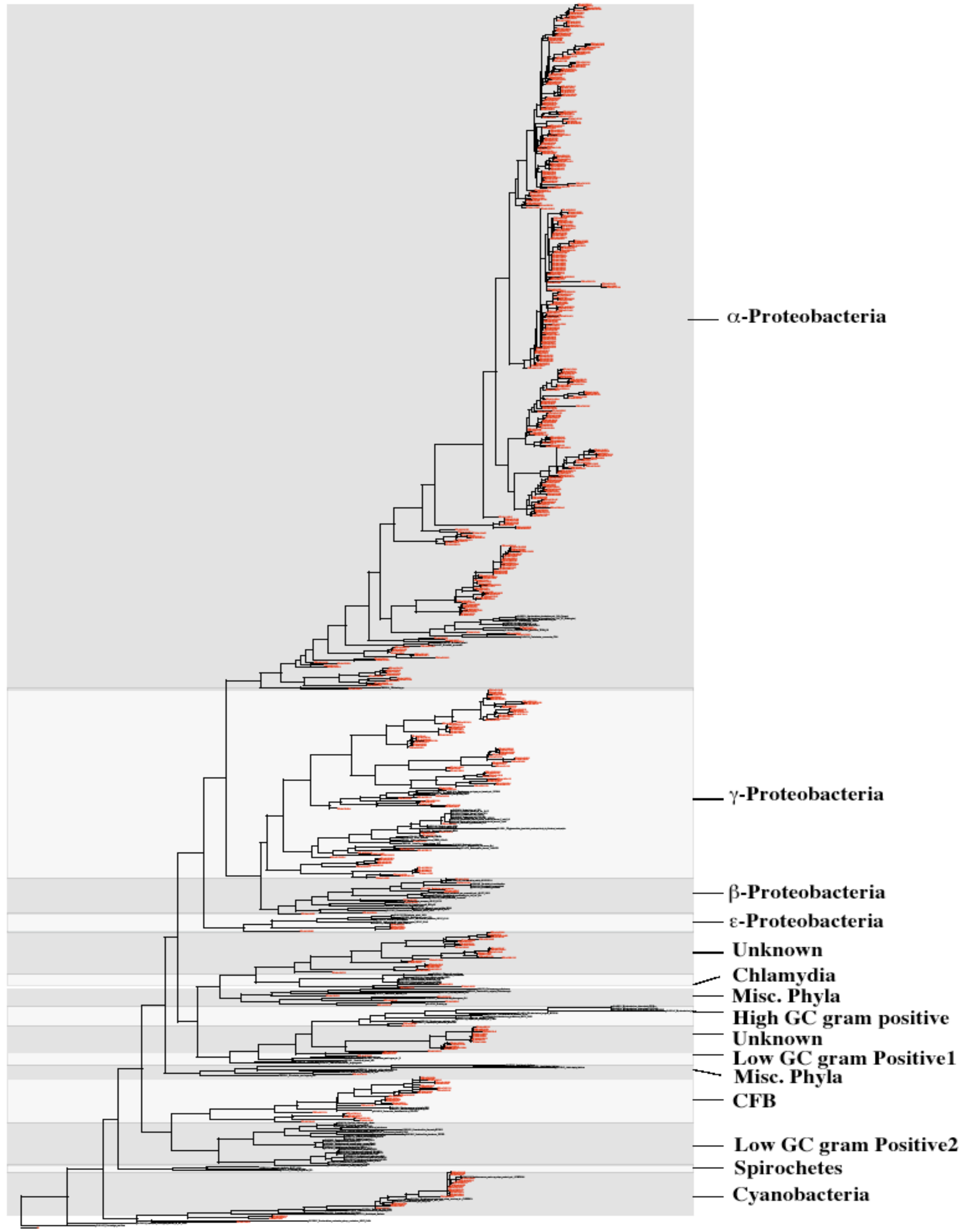
A single tree with everything
(as long as there is a lot of overlap)

rRNA Phylotyping Sargasso Data



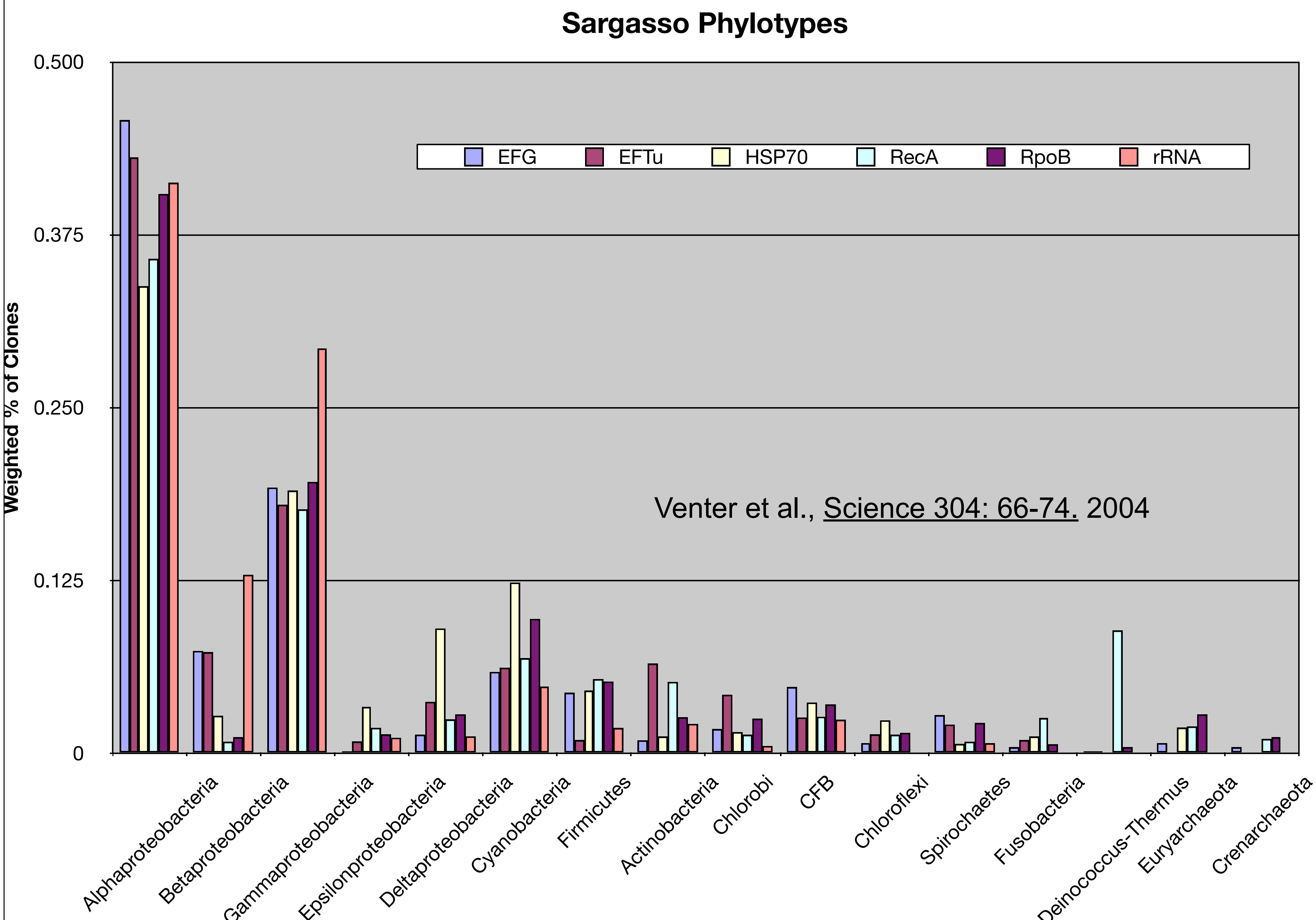
Venter et al., Science 304: 66. 2004

RecA Phylotyping Sargasso Data



Venter et al., Science 304: 66. 2004

Protein vs. rRNA Sargasso Data



Phylogenetic Challenge

- sequence 1 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 1 XXXXXXXXX------------------------------------
- sequence 2 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 2 ------------- XXXXXXXXXXXX-------------------
- fragment 3 -------------------------------XXXXXXXXXXXXXX
- fragment 4 -----XXXXXXXXXXXXXXXXXX----------------------
- sequence 3 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- sequence 4 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- sequence 5 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
- fragment 5 -----------------------XXXXXXXXXXXX----------

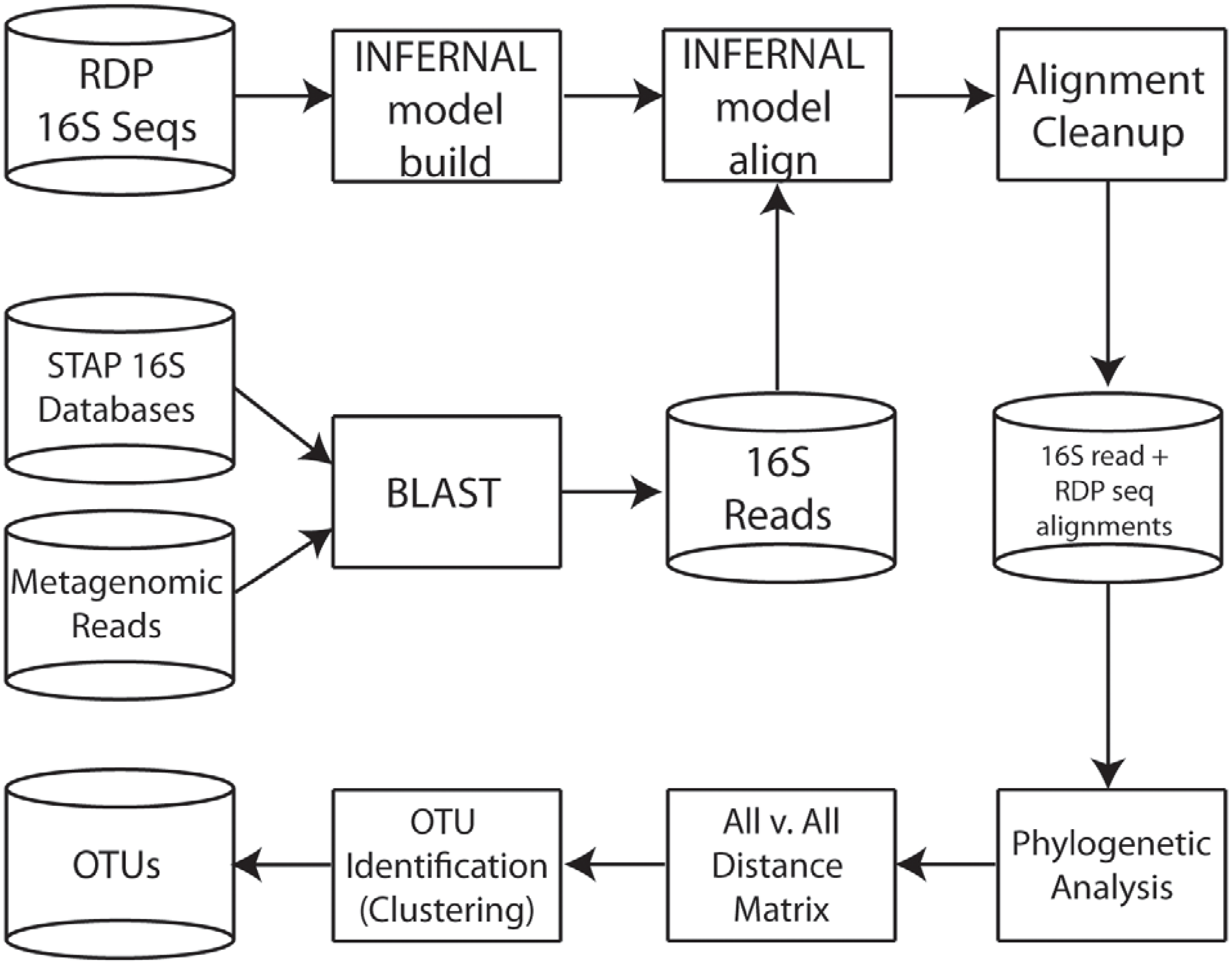
A single tree with everything?

Phylogenetic Challenge

```
• sequence 1 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
• fragment 1 XXXXXXXXX------------------------------------
• sequence 2 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
• fragment 2 ------------- XXXXXXXXXXXX-------------------
• fragment 3 -------------------------------XXXXXXXXXXXXXX
• fragment 4 -----XXXXXXXXXXXXXXXXXX----------------------
• sequence 3 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
• sequence 4 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
• sequence 5 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
• fragment 5 -----------------------XXXXXXXXXXXX----------
```

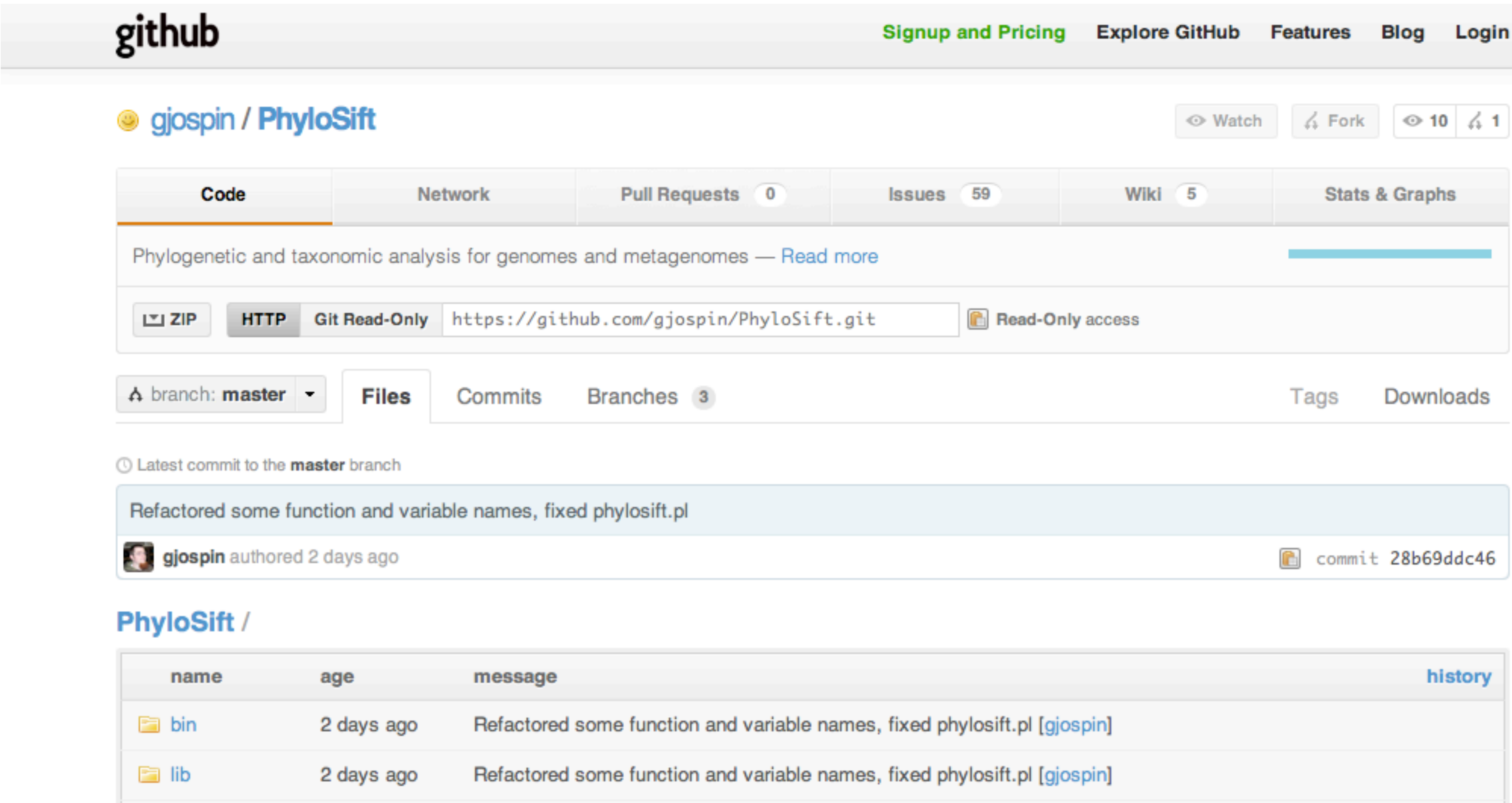
A single tree with everything?

PhylOTU



PhylOTU - Sharpton et al. PLoS Comp. Bio 2011

Phylosift/ pplacer

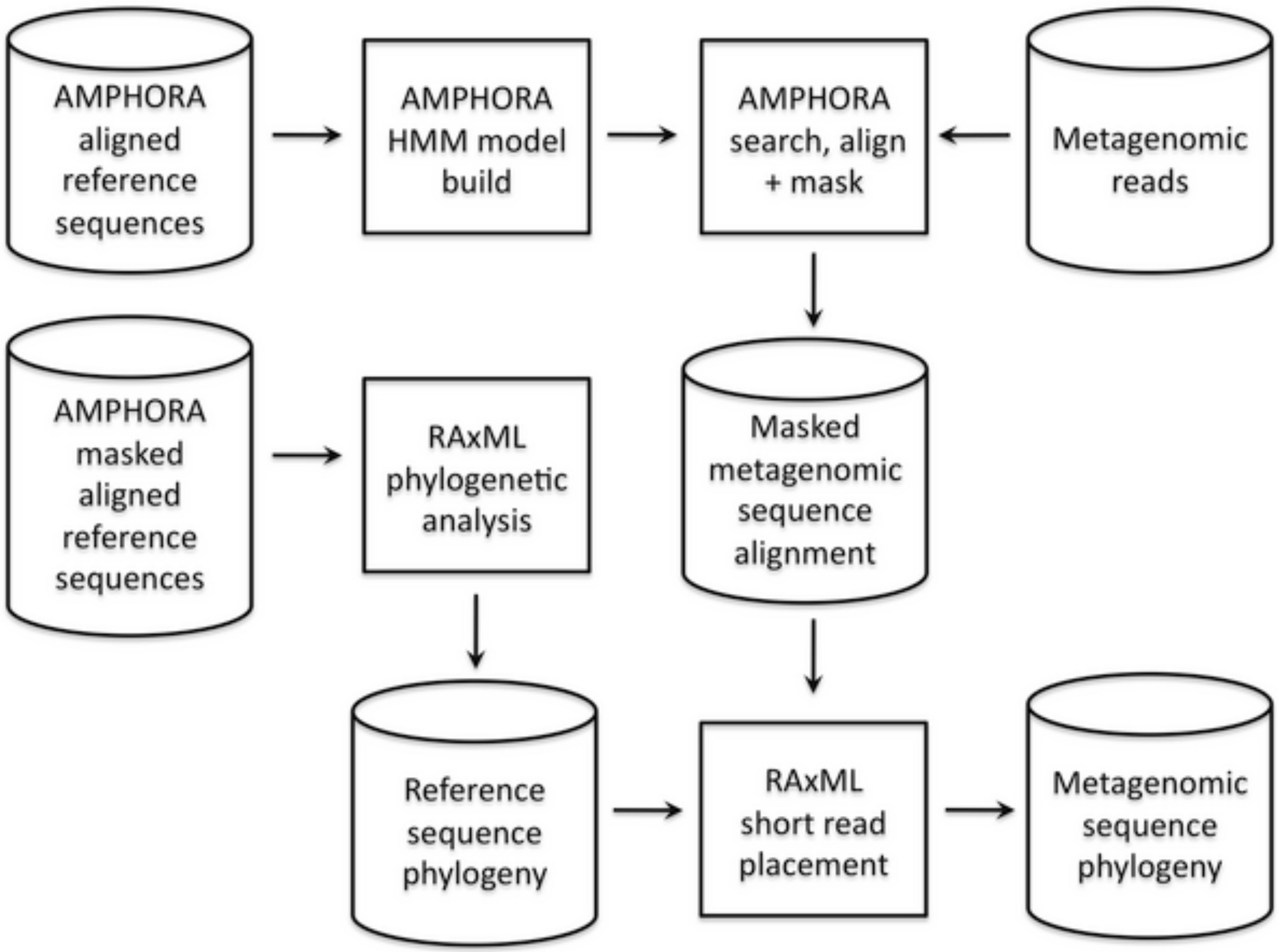


Method 3: All in the genome

- Combine new sequences from different gene families into one tree

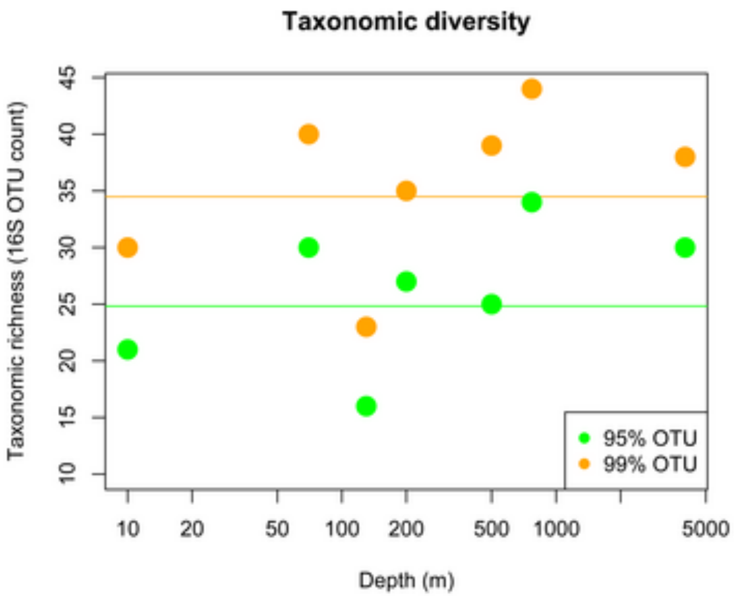
- Take reference data
- Build alignment, models
- Concatenate
- Add all sequences to reference alignment and build tree

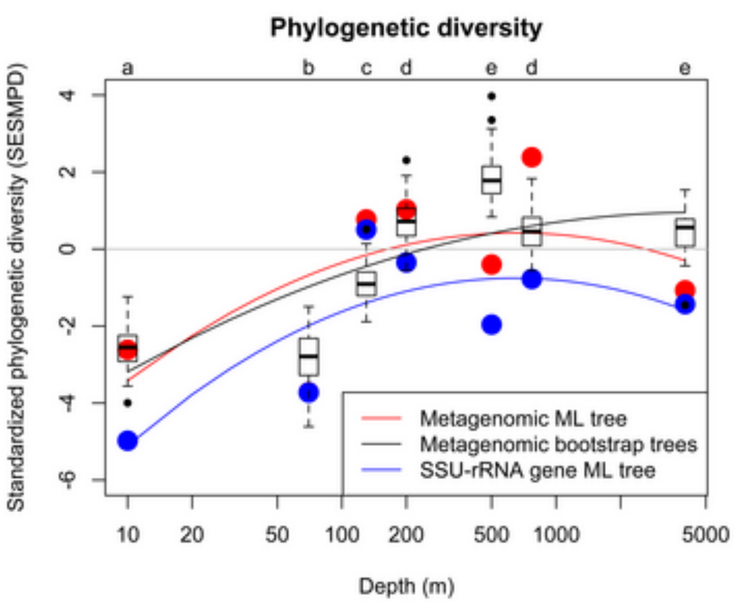
Kembel Combiner

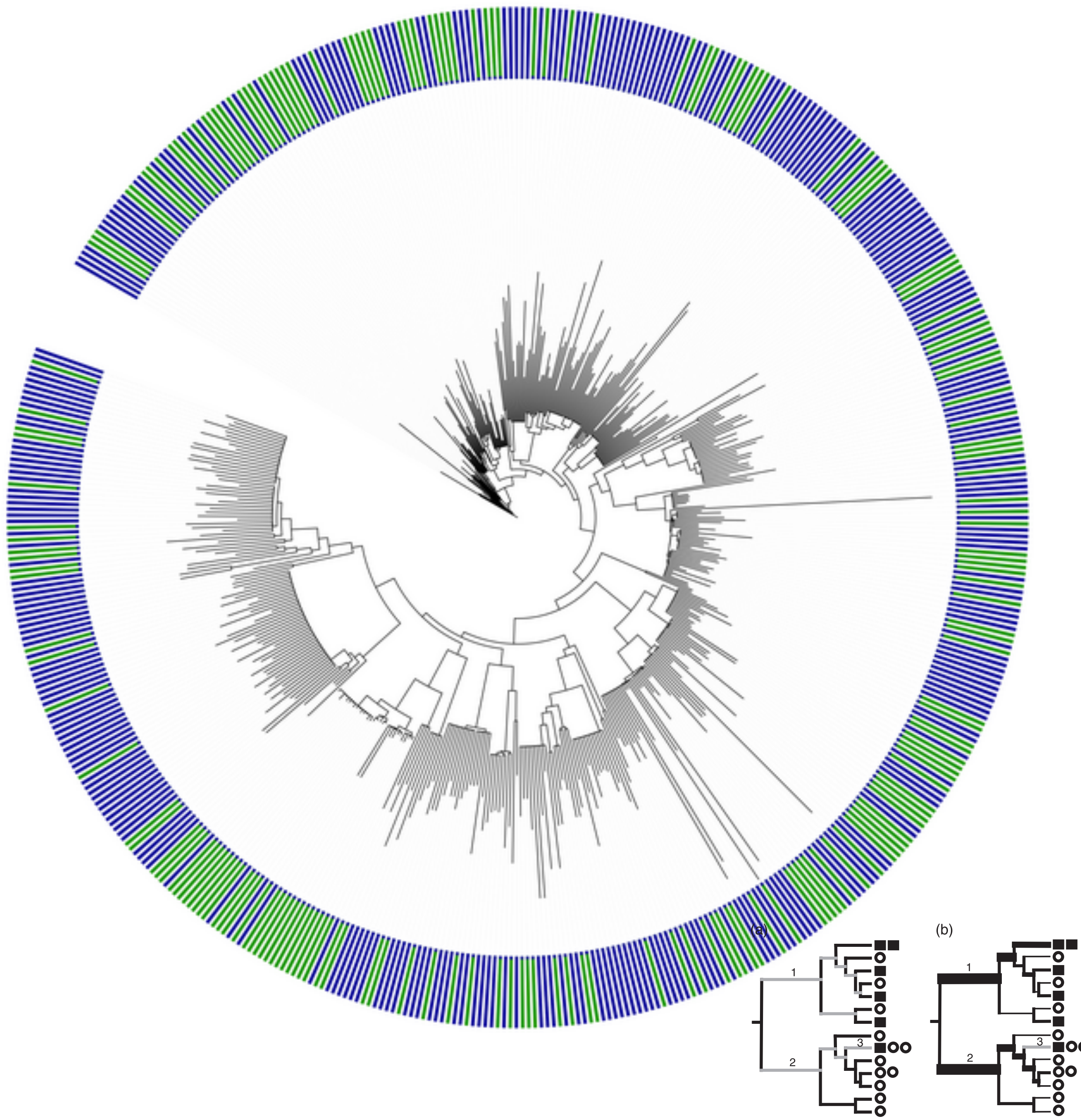


Kembel et al. The phylogenetic diversity of metagenomes. PLoS One 2011

Kembel Combiner

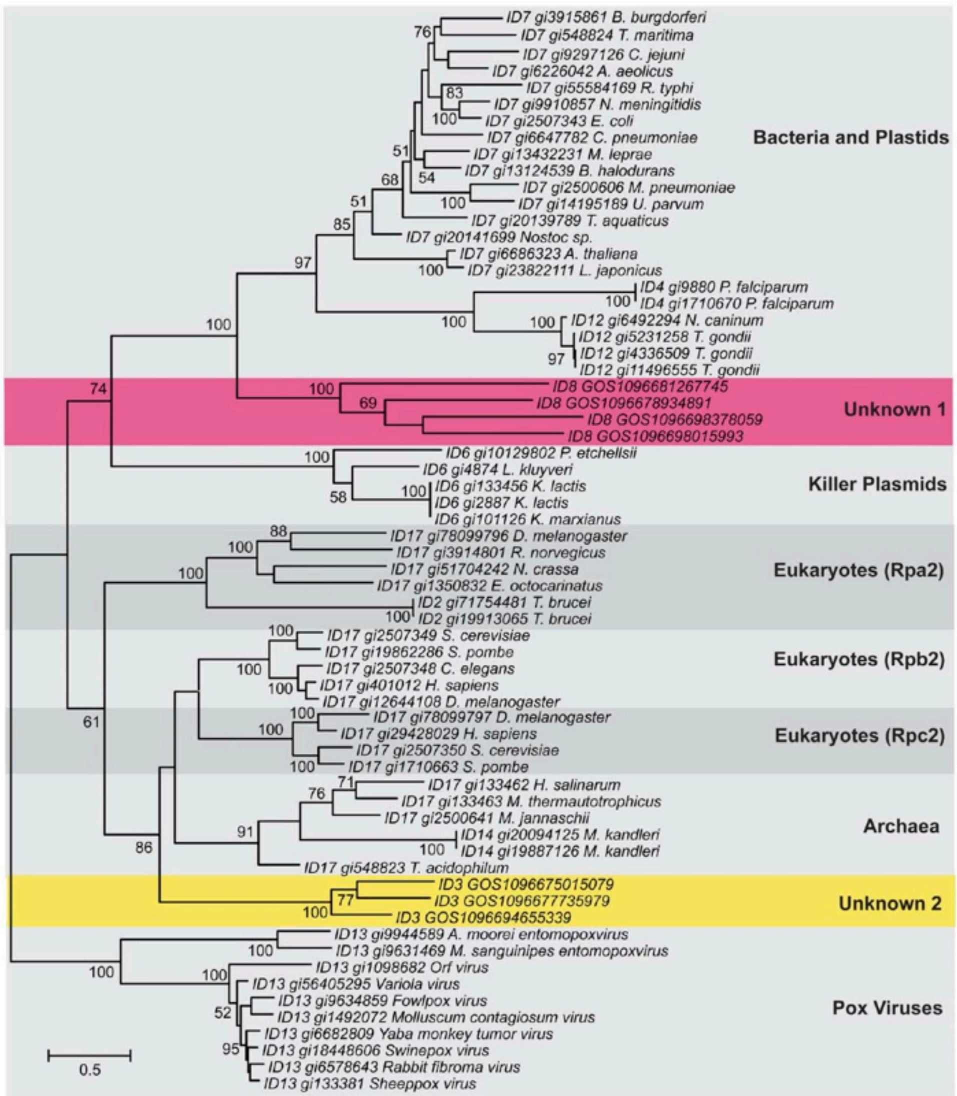




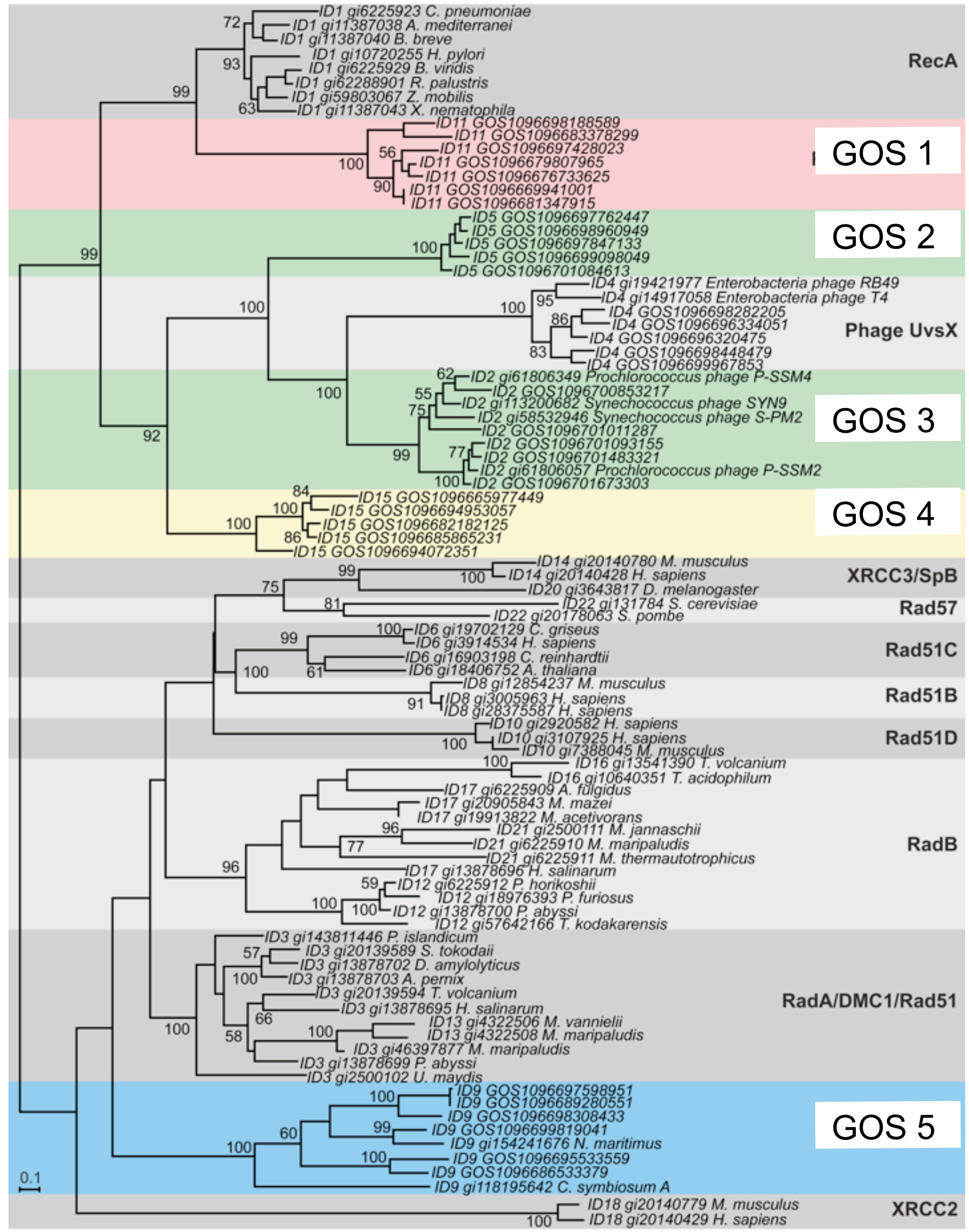


Weird Stuff is Out There

RecA, RpoB in GOS

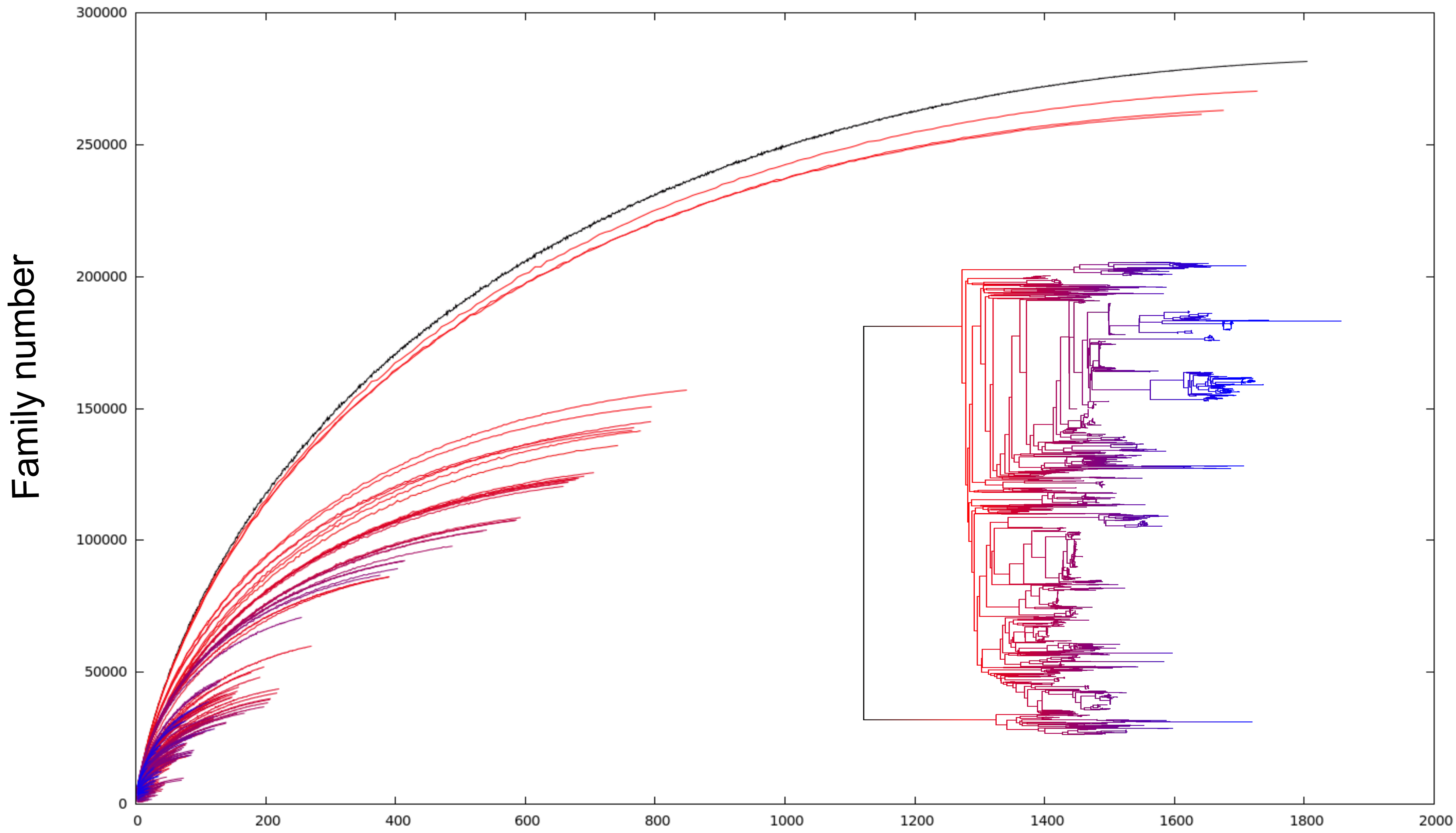


Wu et al PLoS One 2011

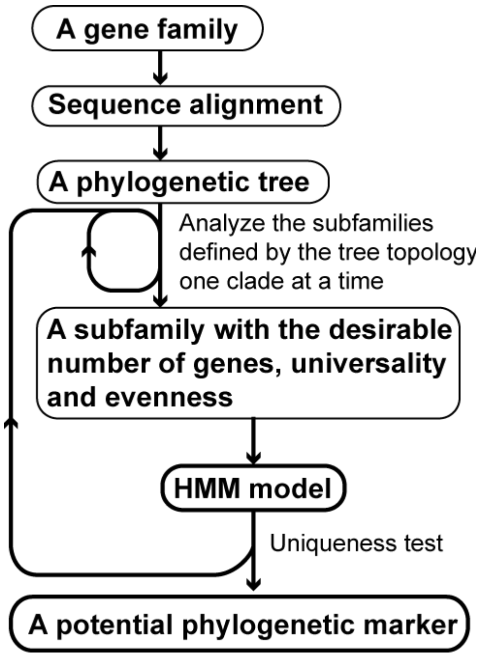


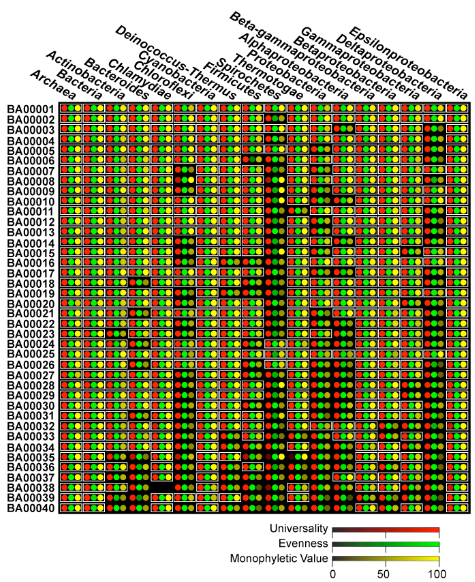
II: More Gene Families

Families/PD not uniform



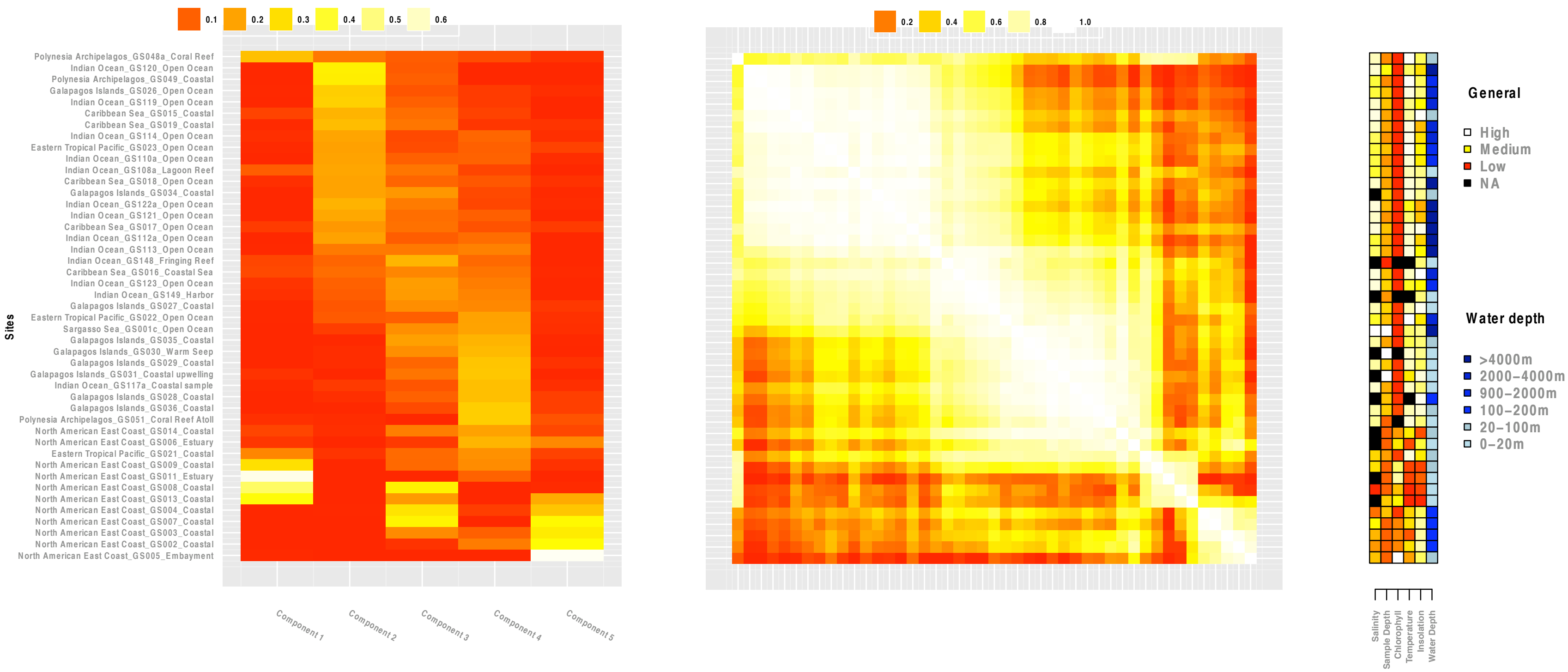
More Markers





Phylogenetic group	Genome Number	Gene Number	Maker Candidates
Archaea	62	145415	106
Actinobacteria	63	267783	136
Alphaproteobacteria	94	347287	121
Betaproteobacteria	56	266362	311
Gammaproteobacter	126	483632	118
Deltaproteobacteria	25	102115	206
Epislonproteobacter	18	33416	455
Bacteriodes	25	71531	286
Chlamydae	13	13823	560
Chloroflexi	10	33577	323
Cyanobacteria	36	124080	590
Firmicutes	106	312309	87
Spirochaetes	18	38832	176
Thermi	5	14160	974
Thermotogae	9	17037	684

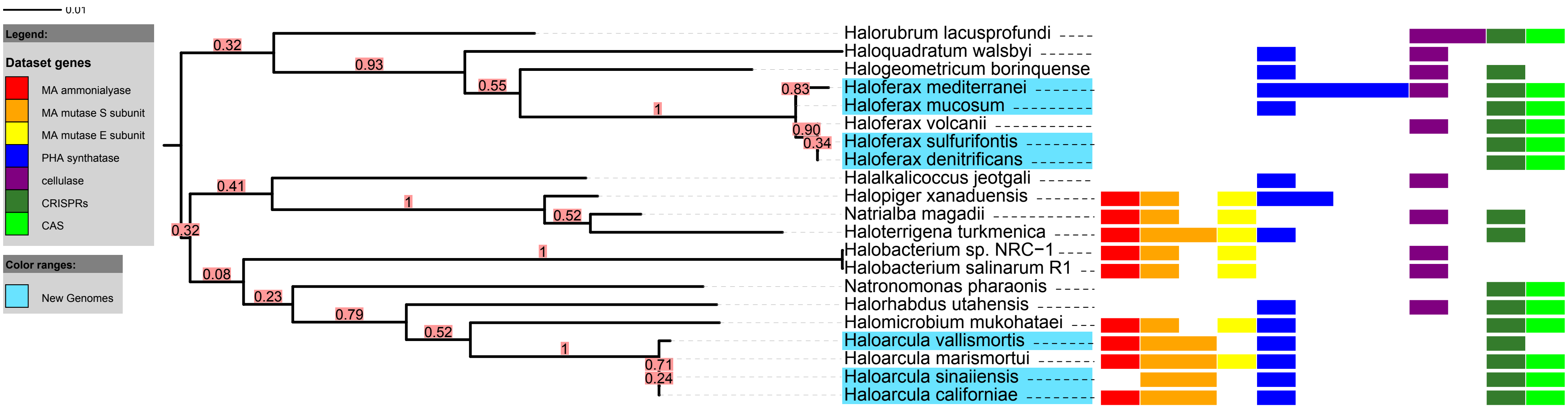
Gene Fams in Metagenomes



(a) (b) (c)

III: Zoom in on Lineages

Haloarchaea



Haloarchaea TBPs

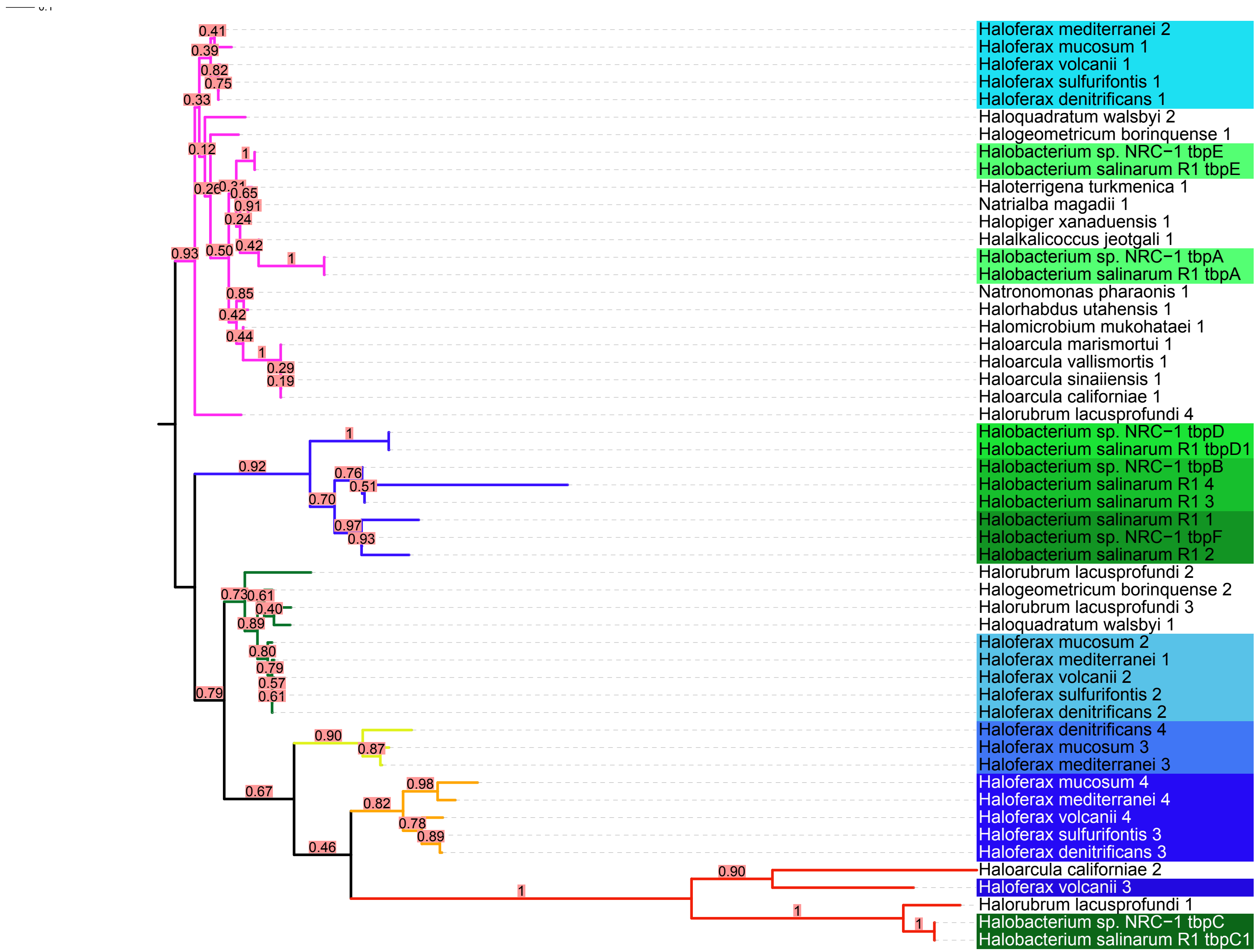
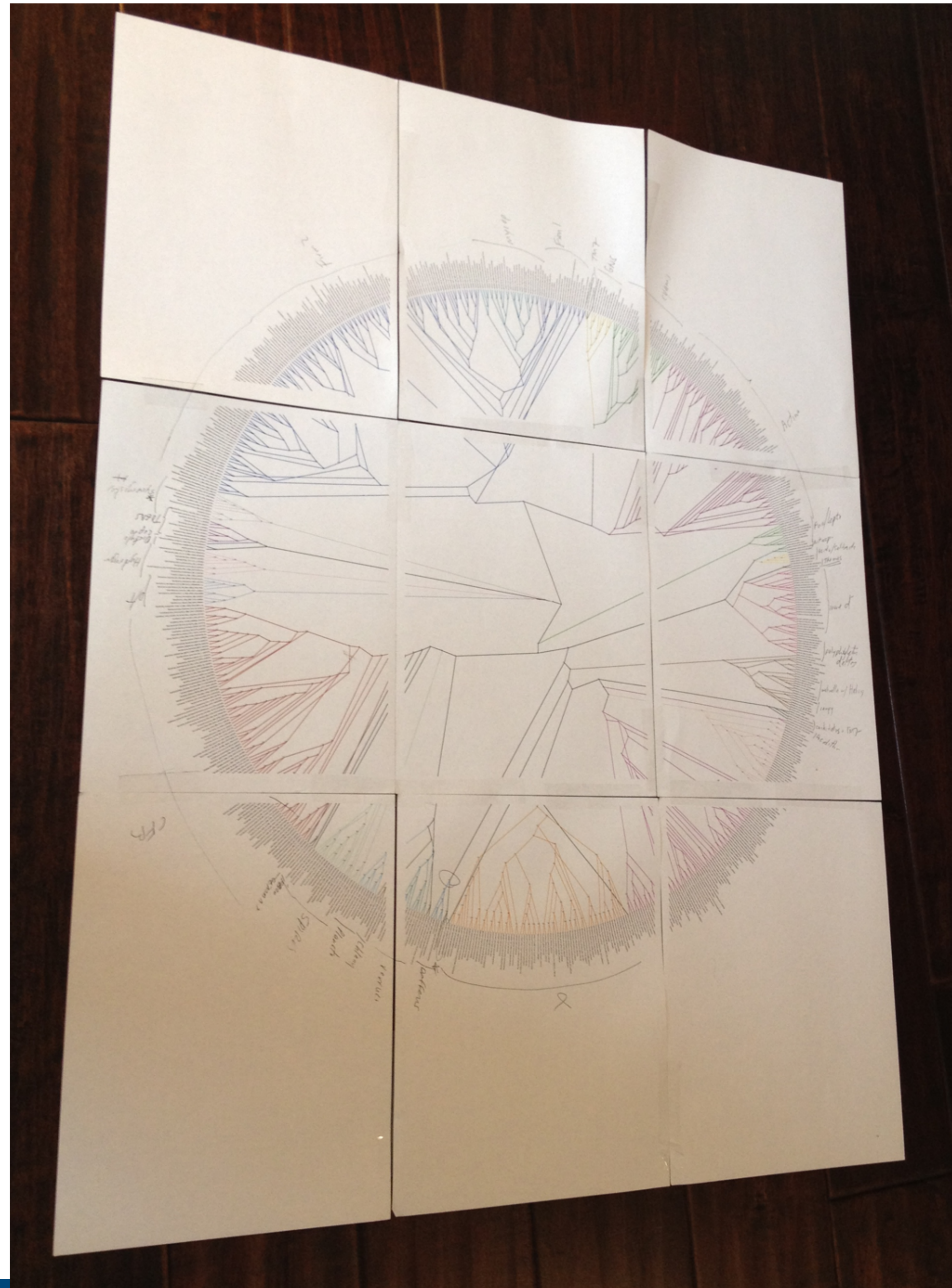


Figure 8. Independent expansion of the TATA-binding protein family in two haloarchaeal genera. Phylogeny of TATA-binding protein (TBP) homologs identified by RAST with Bootstrap values shown. Colored branches represent duplication events (with the dark blue branch representing four duplications). Ancestral TBP (found in all genomes) is shown on the purple branch. Successive duplications are shown in darkening shades of green (*Halobacterium*) or blue (*Haloferax*).

Lynch et al. in preparation

IV: Better Reference Tree



V: Uncultured Lineages

rRNA Tree of Life

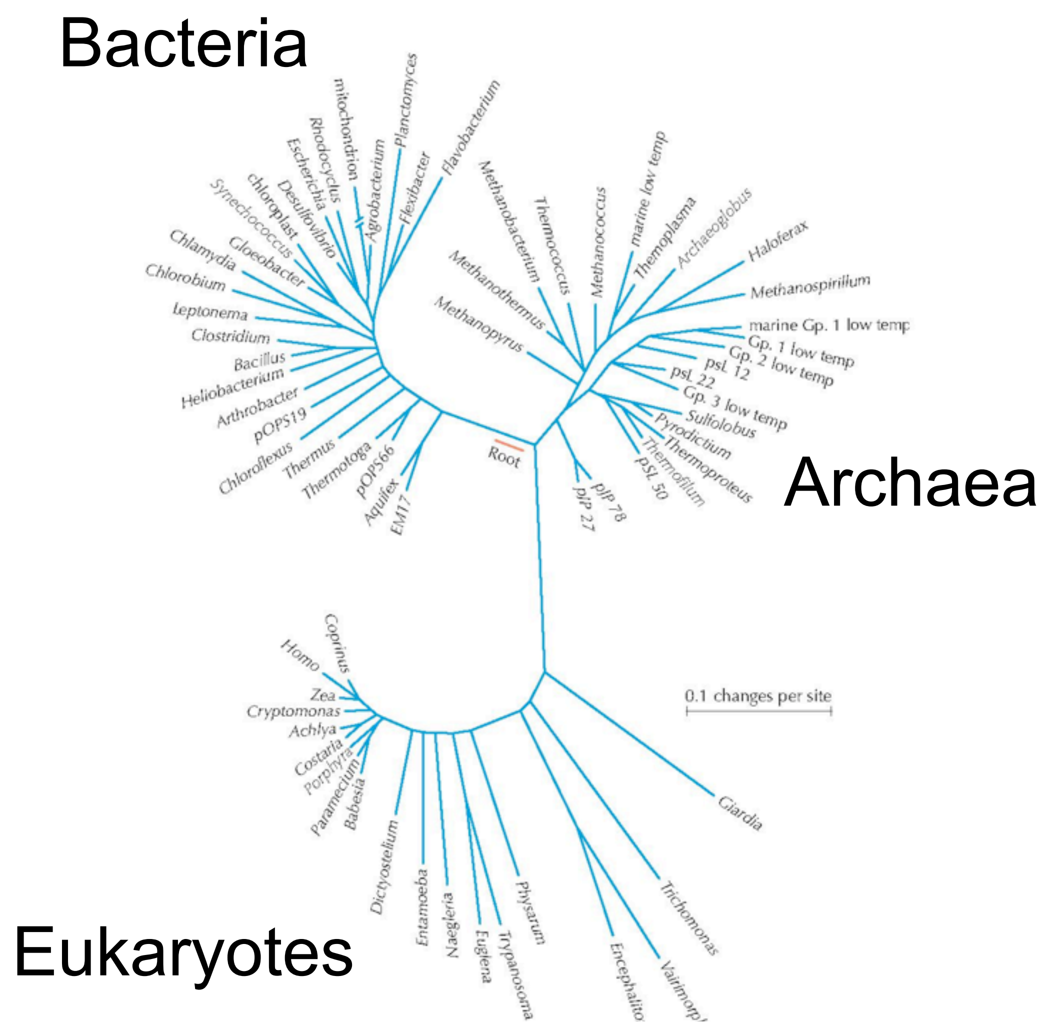
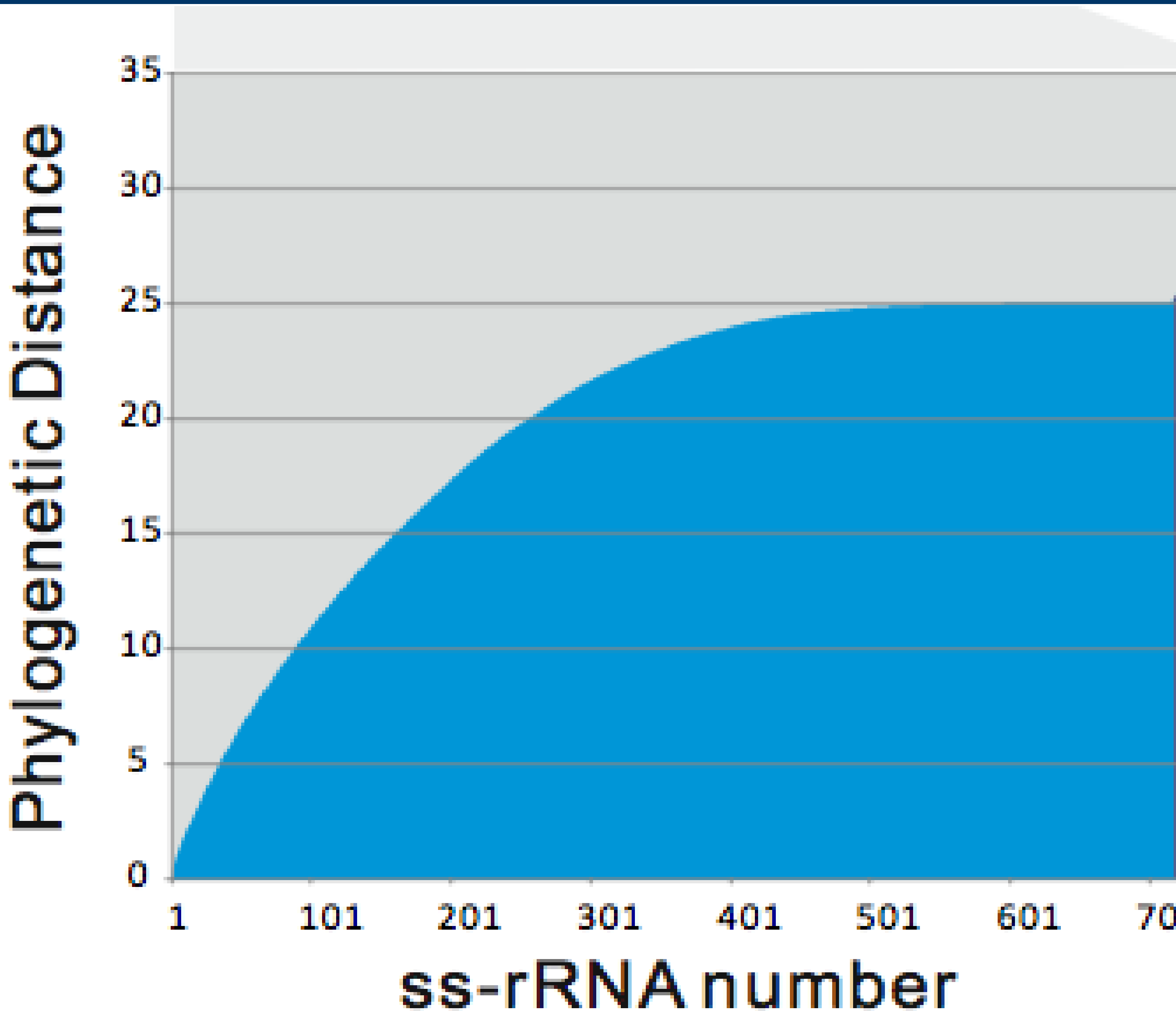


Figure from Barton, Eisen et al. "Evolution", CSHL Press. 2007.

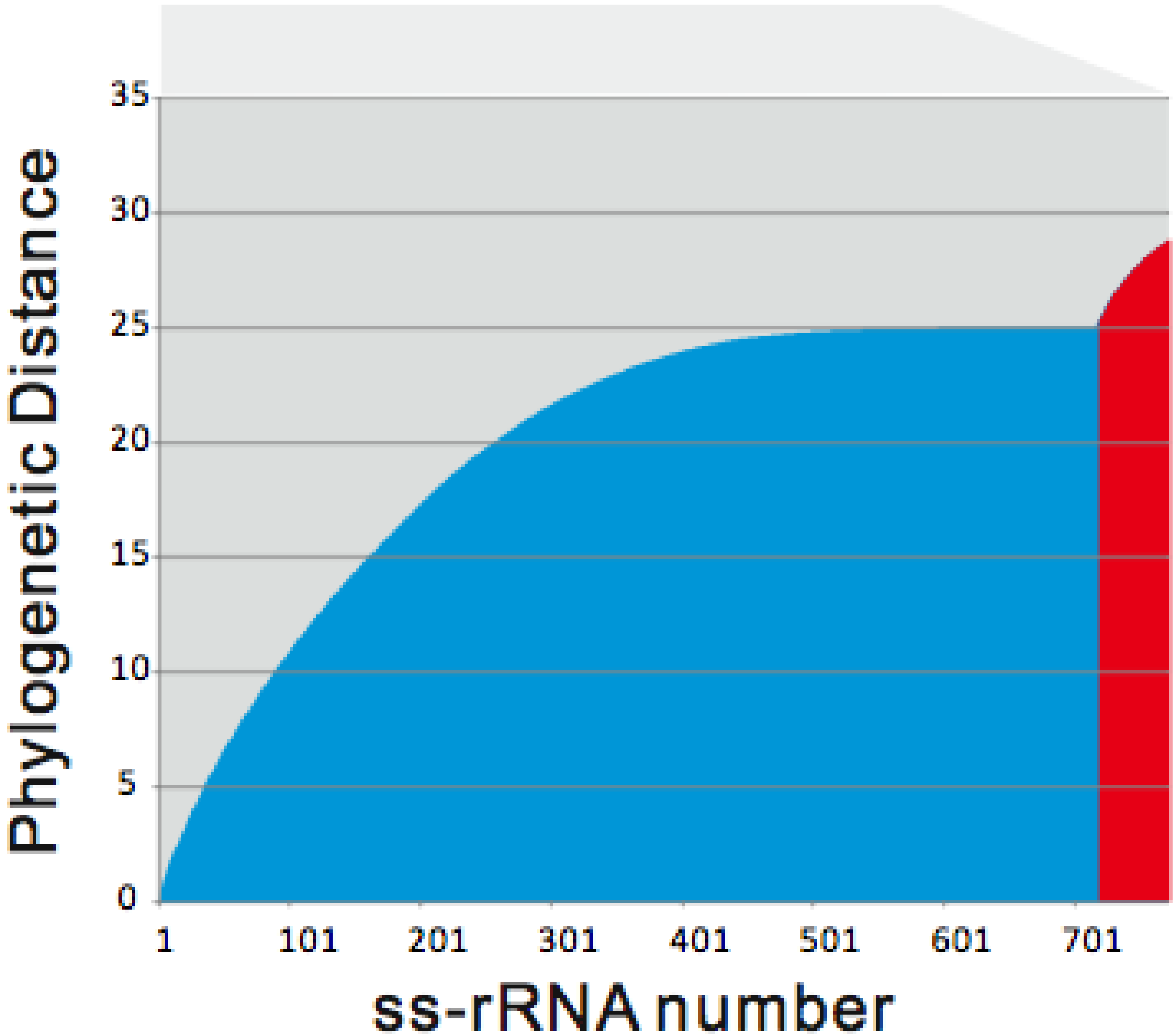
Based on tree from Pace 1997 Science 276:734-740

PD: Genomes



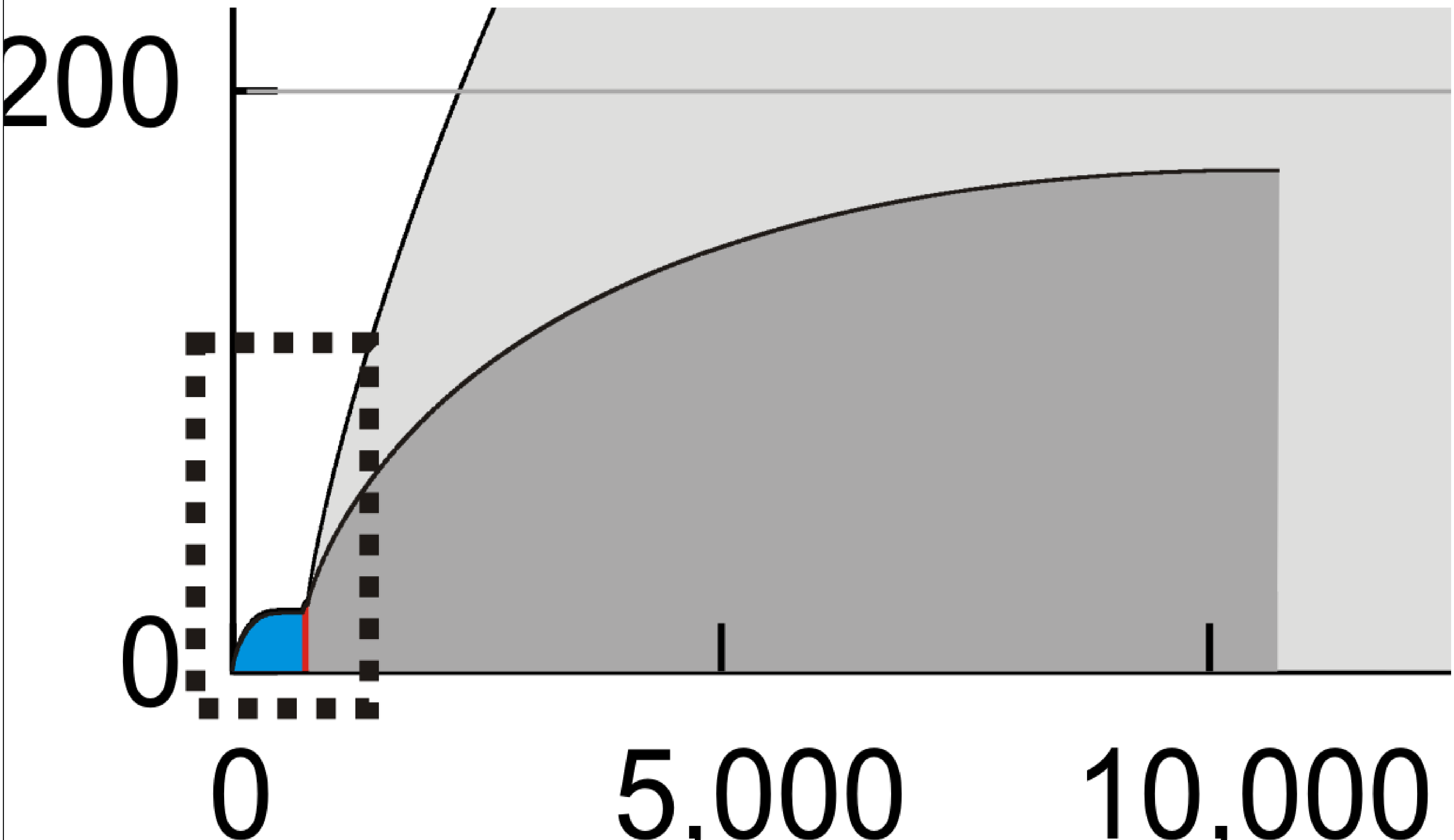
From Wu et al. 2009 Nature 462, 1056-1060

PD: Genomes + GEBA



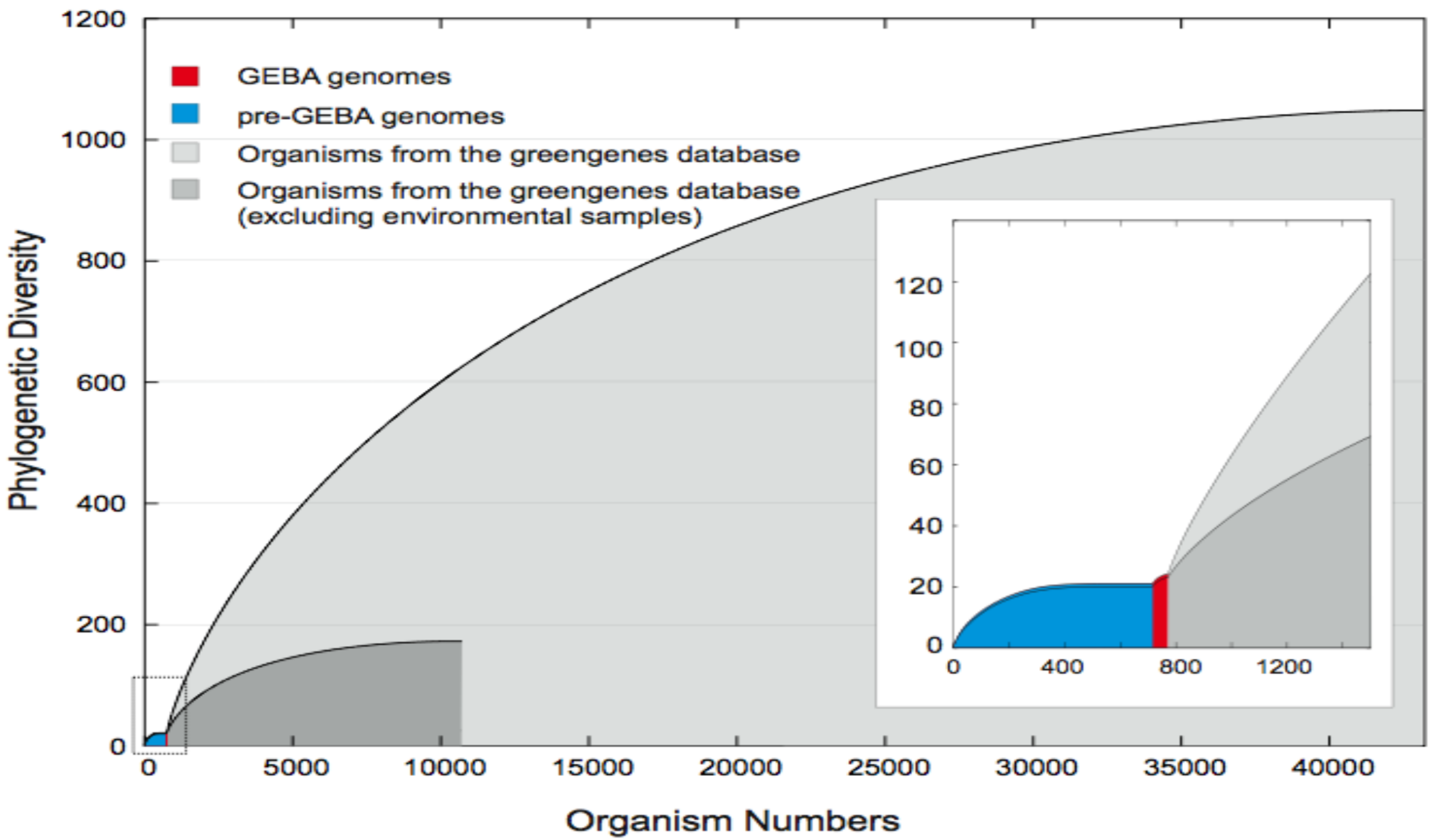
From Wu et al. 2009 Nature 462, 1056-1060

PD: Isolates



From Wu et al. 2009 Nature 462, 1056-1060

PD: All



From Wu et al. 2009 Nature 462, 1056-1060

Uncultured Lineages:

- Get into culture
- Enrichment cultures
- If abundant in low diversity ecosystems
- Flow sorting
- Microbeads
- Microfluidic sorting
- Single cell amplification

UCDAVIS UNIVERSITY OF CALIFORNIA

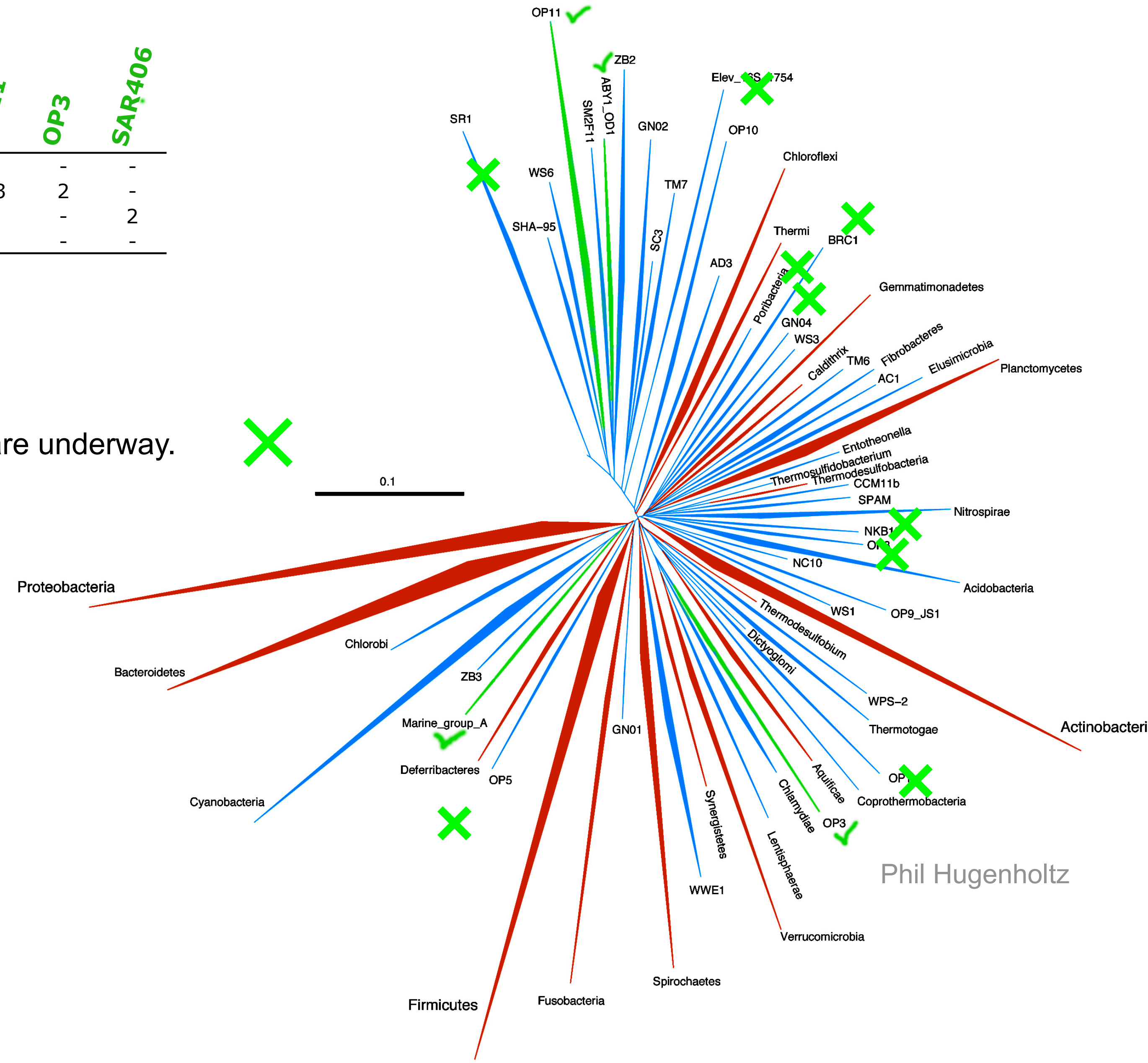
UC DAVIS genome center

GEBA uncultured

Number of SAGs from Candidate Phyla

	OD1	OP11	OP3	SAR406
Site A: Hydrothermal vent	4	1	-	-
Site B: Gold Mine	6	13	2	-
Site C: Tropical gyres (Mesopelagic)	-	-	-	2
Site D: Tropical gyres (Photic zone)	1	-	-	-

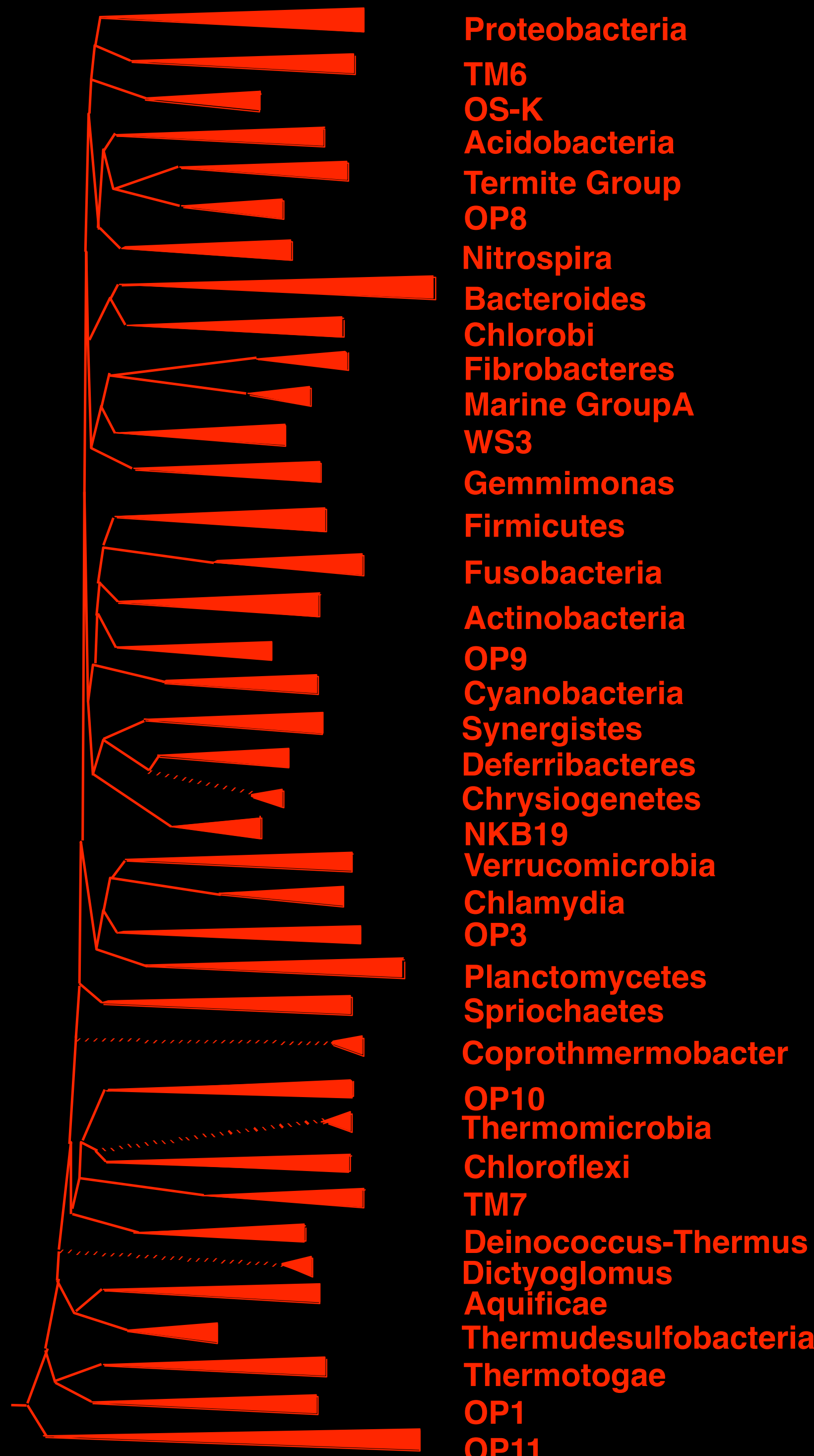
Sample collections at 4 additional sites are underway.



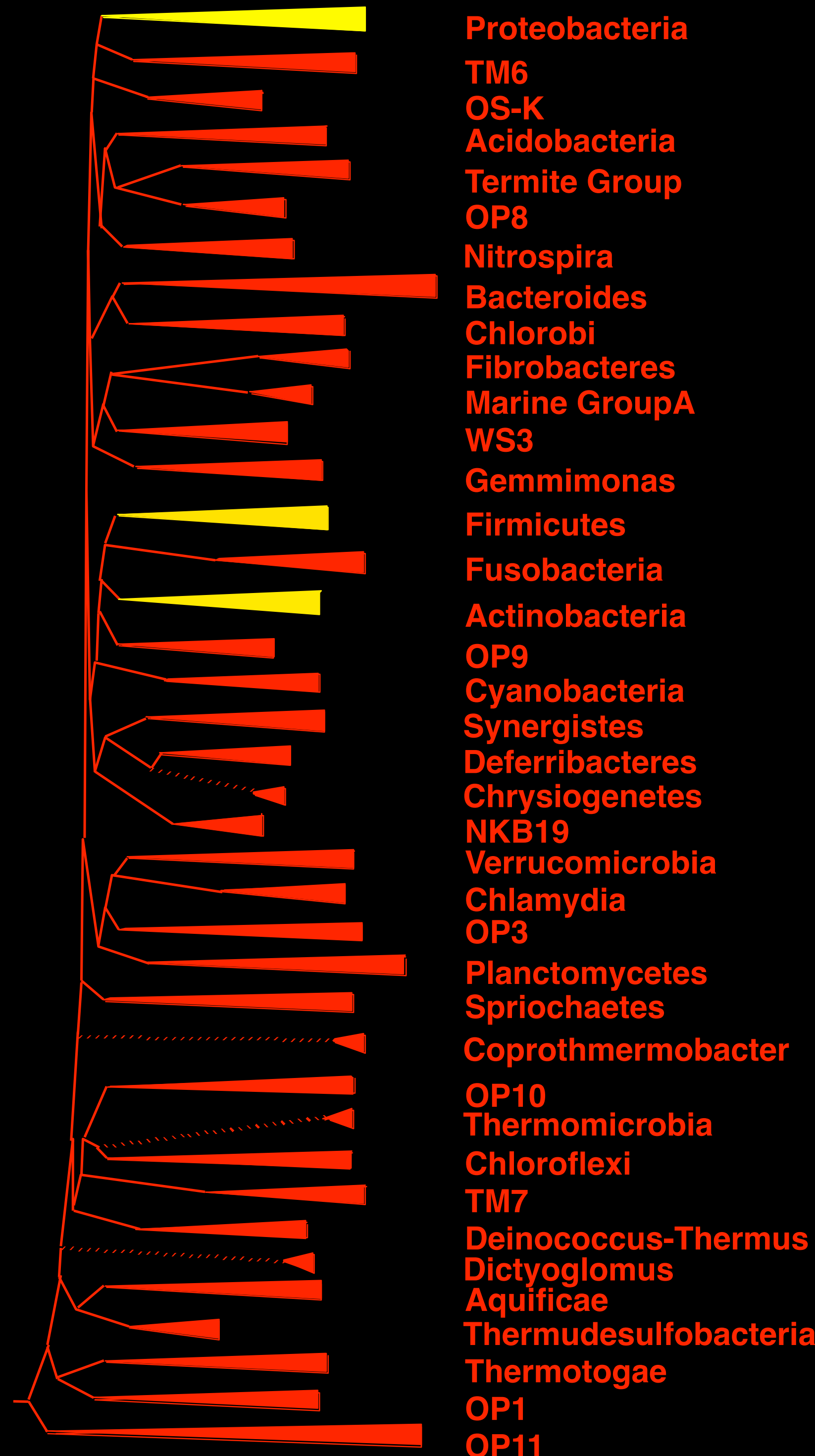
Phil Hugenholtz

Experiments?

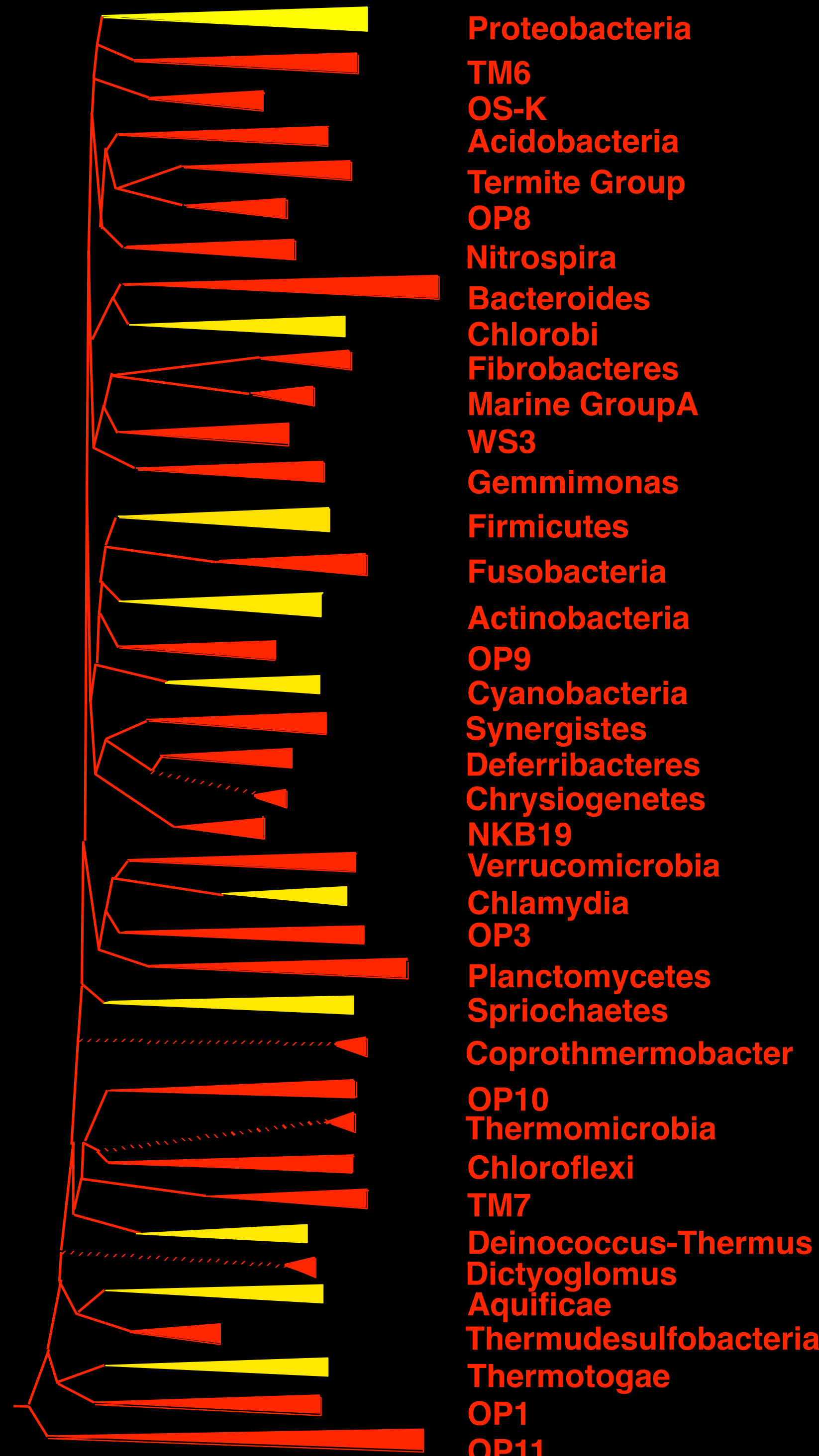
Need Experiments from Across the Tree of Life too



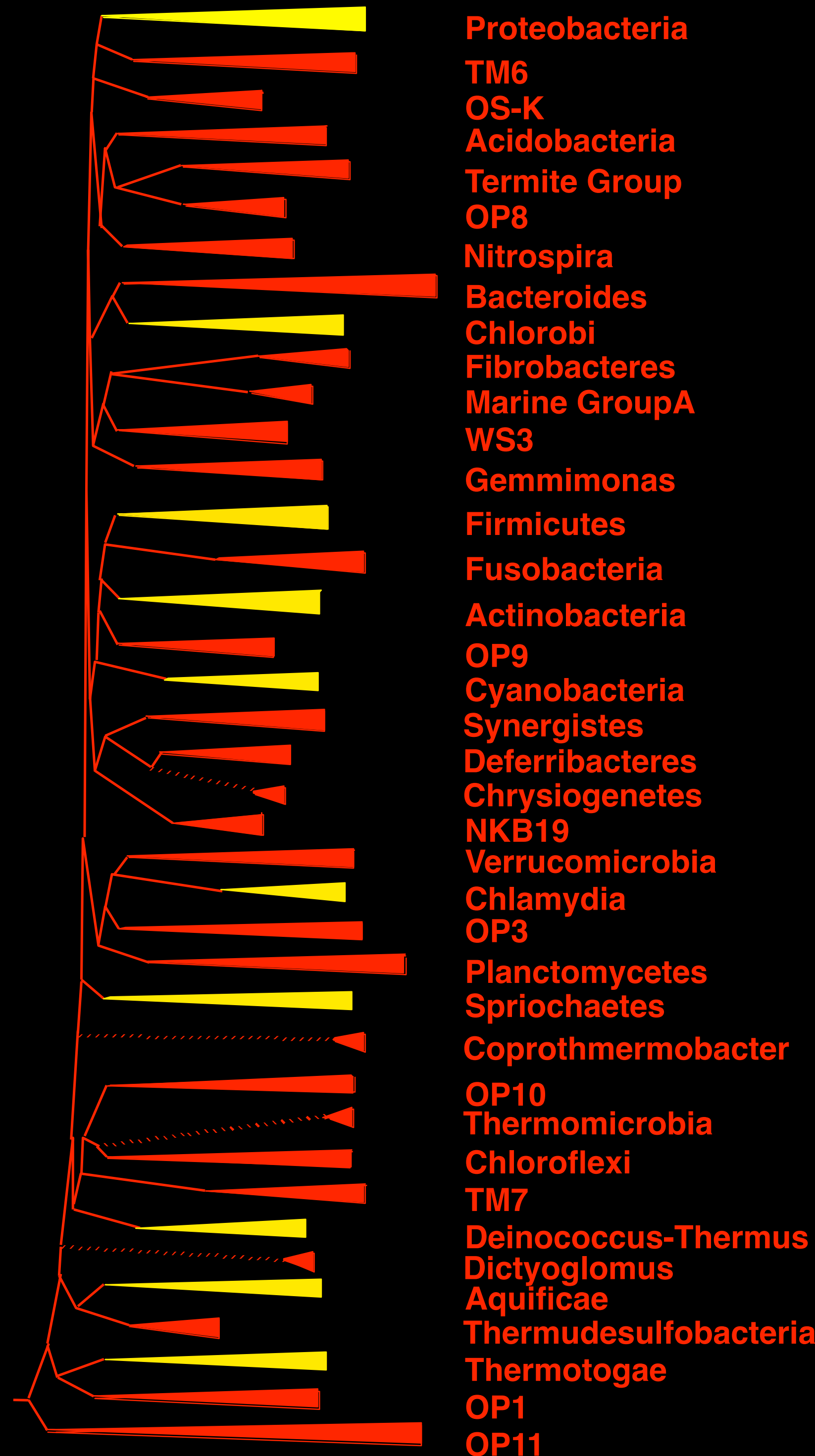
- **At least 40 phyla of bacteria**



- At least 40 phyla of bacteria
- Experimental studies are mostly from three phyla

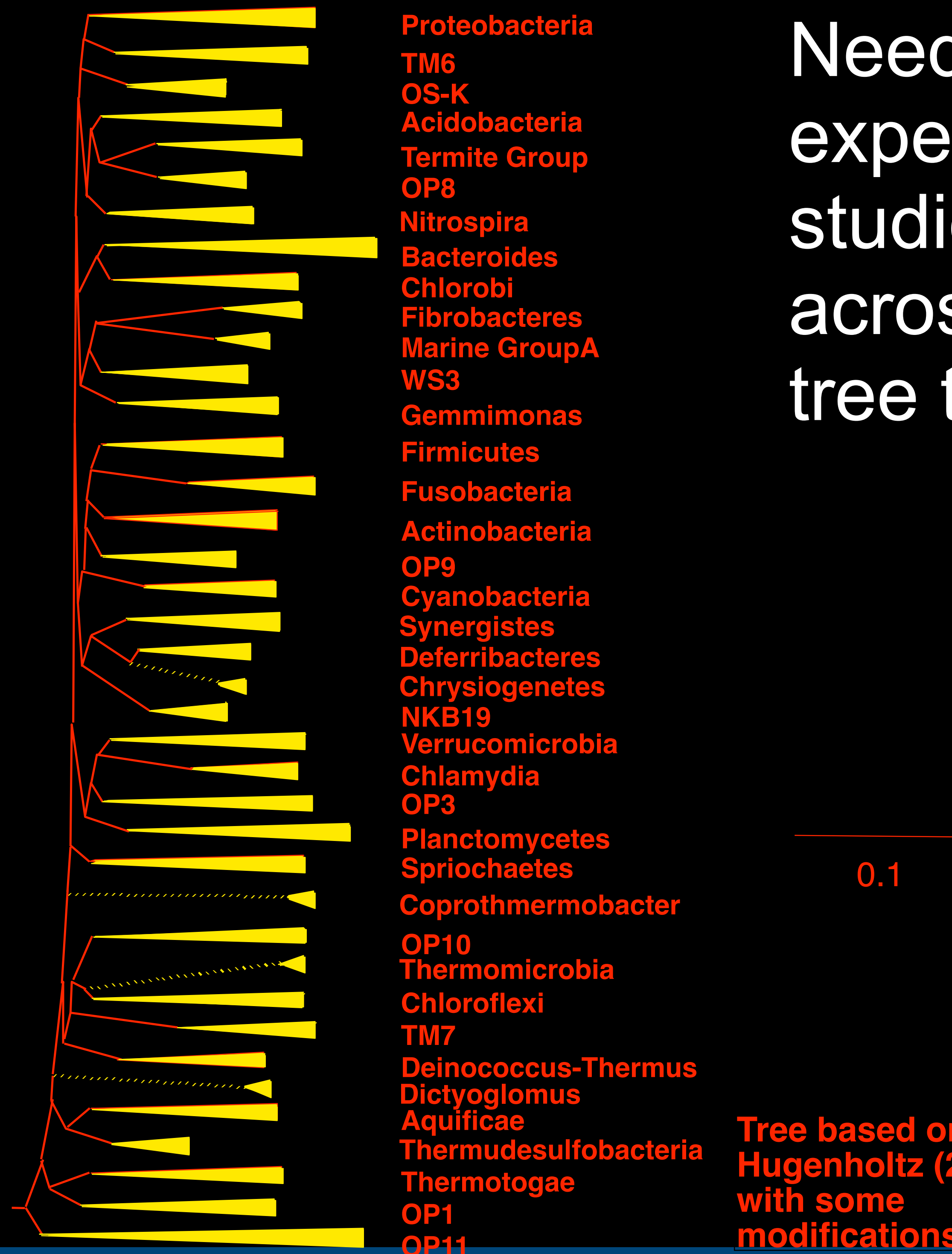


- At least 40 phyla of bacteria
- Experimental studies are mostly from three phyla
- Some studies in other phyla



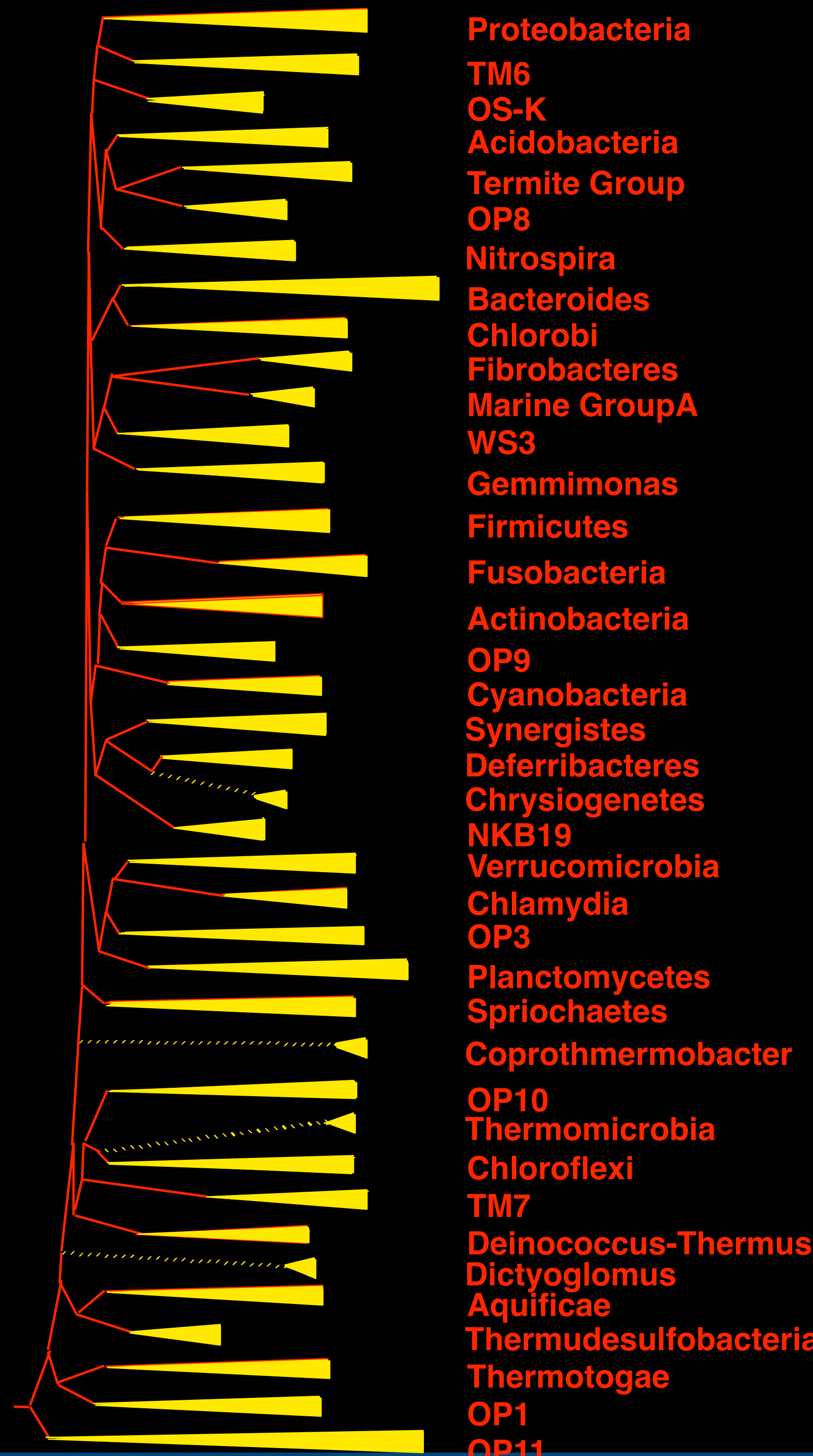
- At least 40 phyla of bacteria
- Experimental studies are mostly from three phyla
- Some studies in other phyla
- Same trend in Eukaryotes

Need experimental studies from the tree too



Tree based on Hugenholtz (2002) with some modifications.

Adopt a Microbe



Tree based on Hugenholtz (2002) with some modifications.

Acknowledgements

- $$$
  - DOE
  - NSF
  - GBMF
  - Sloan
  - DARPA
  - DSMZ
  - DHS
- People, places
  - DOE JGI: Eddy Rubin, Phil Hugenholtz, Nikos Kyrpides
  - UC Davis: Aaron Darling, Dongying Wu, Holly Bik, Russell Neches, Jenna Morgan-Lang
  - Other: Jessica Green, Katie Pollard, Martin Wu, Tom Slezak, Jack Gilbert, Steven Kembel, J. Craig Venter, Naomi Ward, Hans-Peter Klenk