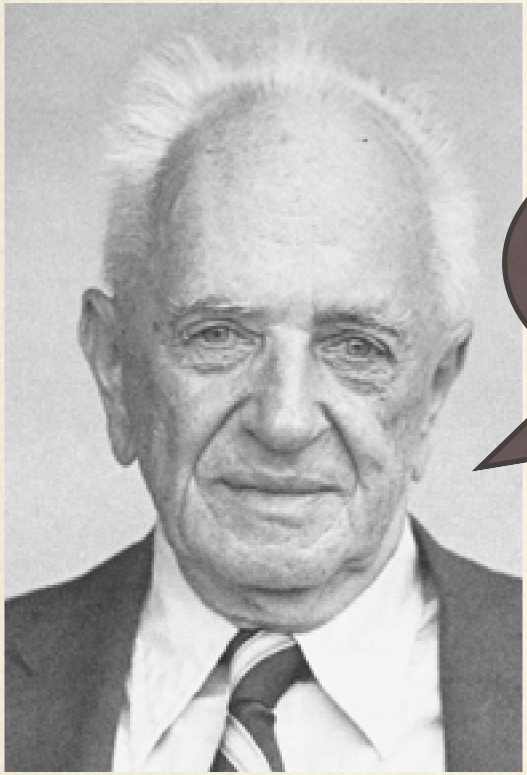
Nothing in (Micro)biology Makes Sense Except in the Light of Evolution

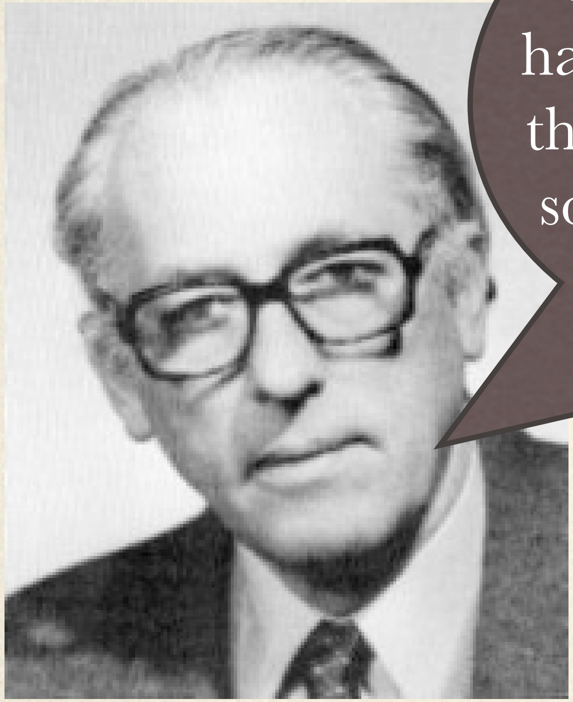
Jonathan H. Badger
March 29, 2007

Theodosius Dobzhansky, 1973



Nothing in biology makes sense except in the light of evolution!

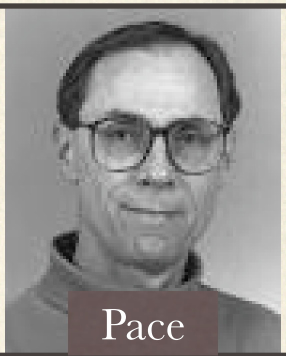
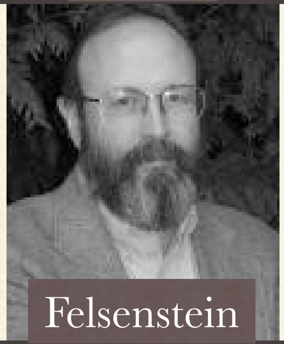
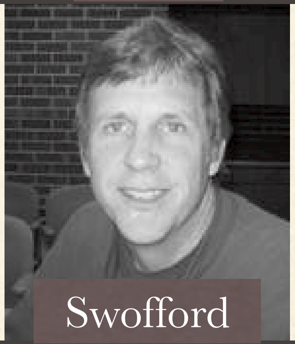
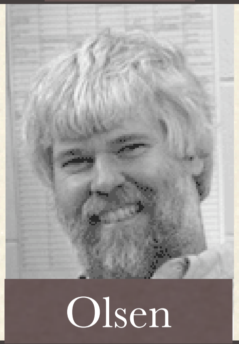
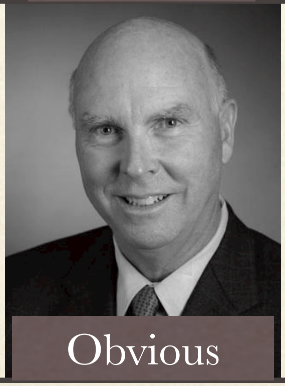
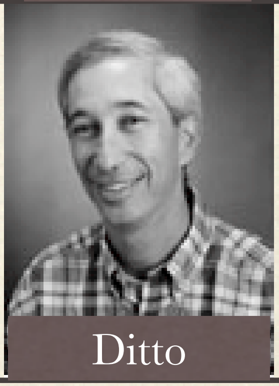
Roger Stanier, 1970



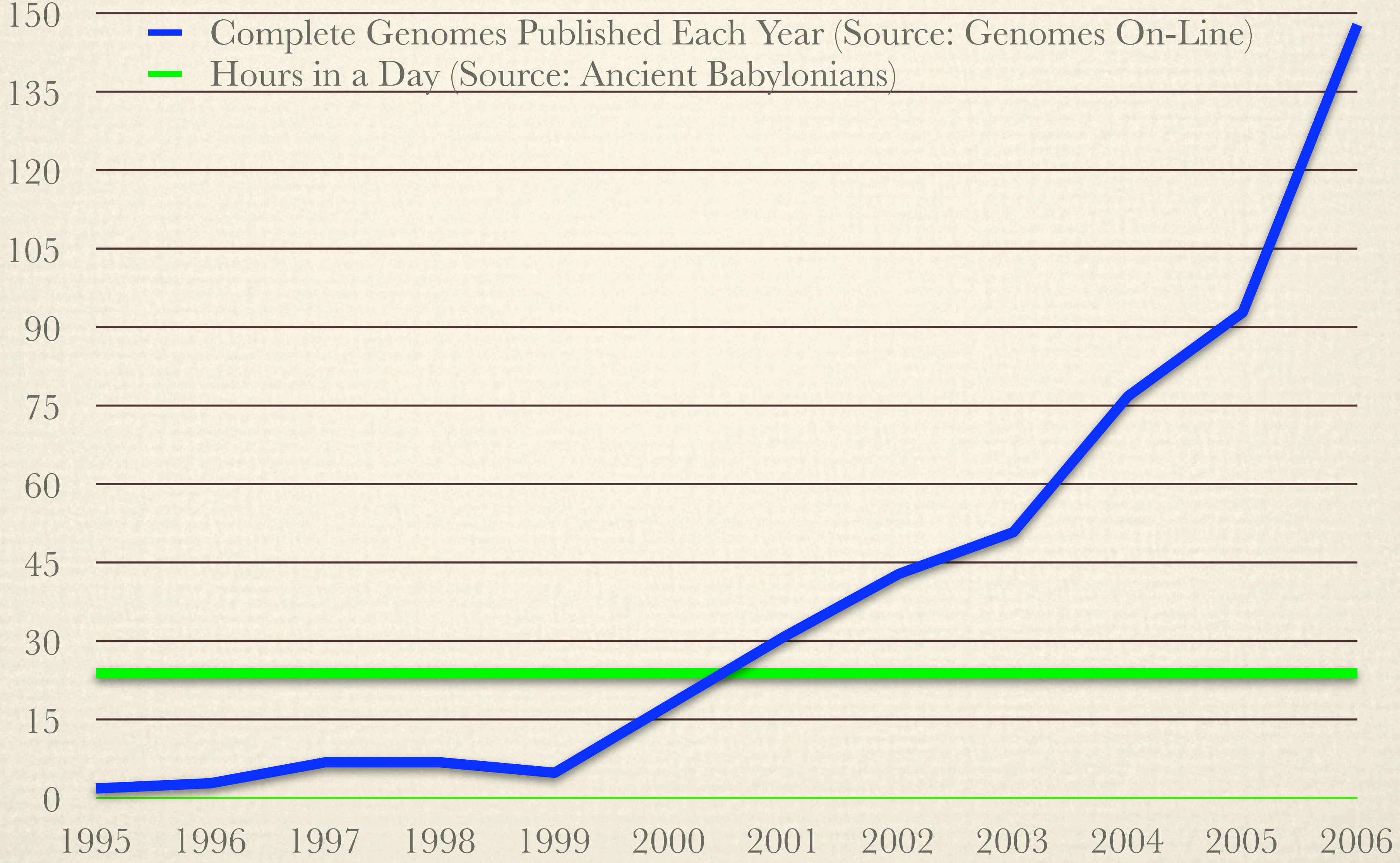
> Evolutionary speculation constitutes a kind of metascience, which has the same fascination for some biologists that metaphysical speculation possessed for some medieval scholastics. [...] a harmless habit, unless it becomes an obsession.

Why the difference in opinion? What was lacking?

The Molecular Revolution in Microbiology

rRNA - 1977	Woese	Fox	 Pace
Phylogenetic Tools - 1980s	 Felsenstein	 Swofford	 Olsen
Genomes - 1995	 Obvious	You know	 Ditto

Genomes Sequenced Each Year



Outline of Talk

- Part I: Evolutionary tools I’ve developed that can help you
  - phylipFasta/paupFasta
  - APIS (Automated Phylogenetic Inference System)
  - ECFinder
  - IDEA (contributed to)
- Part II: Examples of how my tools have helped my own research
  - Positive Selection in *Geobacter*
  - Reverse Gyrase Phylogeny
  - Influenza Phylogeny

phylipFasta/paupFasta

Phylogenetic Methods & Packages

- Methods
  - Neighbor Joining (NJ) - Quick, based on clustering of distances (which can be generated from sequences)
  - Maximum Parsimony (MP) - Slow, based on finding tree requiring fewest changes in data (sequences)
  - Maximum Likelihood (ML) - Even slower, based on finding tree maximizing the probability of generating the sequences given a model of sequence evolution
- Packages
  - PHYLIP (free) - focused on ML, but can do all
  - PAUP (commercial) - focused on MP, but can do all but protein ML

Inferring Phylogenies

- In theory, although large phylogenies may take up large amounts of computer time, they shouldn’t take much human time
  - Make Alignment
  - Run phylogeny program
  - Go get coffee (or go home & sleep)
  - Have nice tree figure to include in papers & presentations
- But in practice....

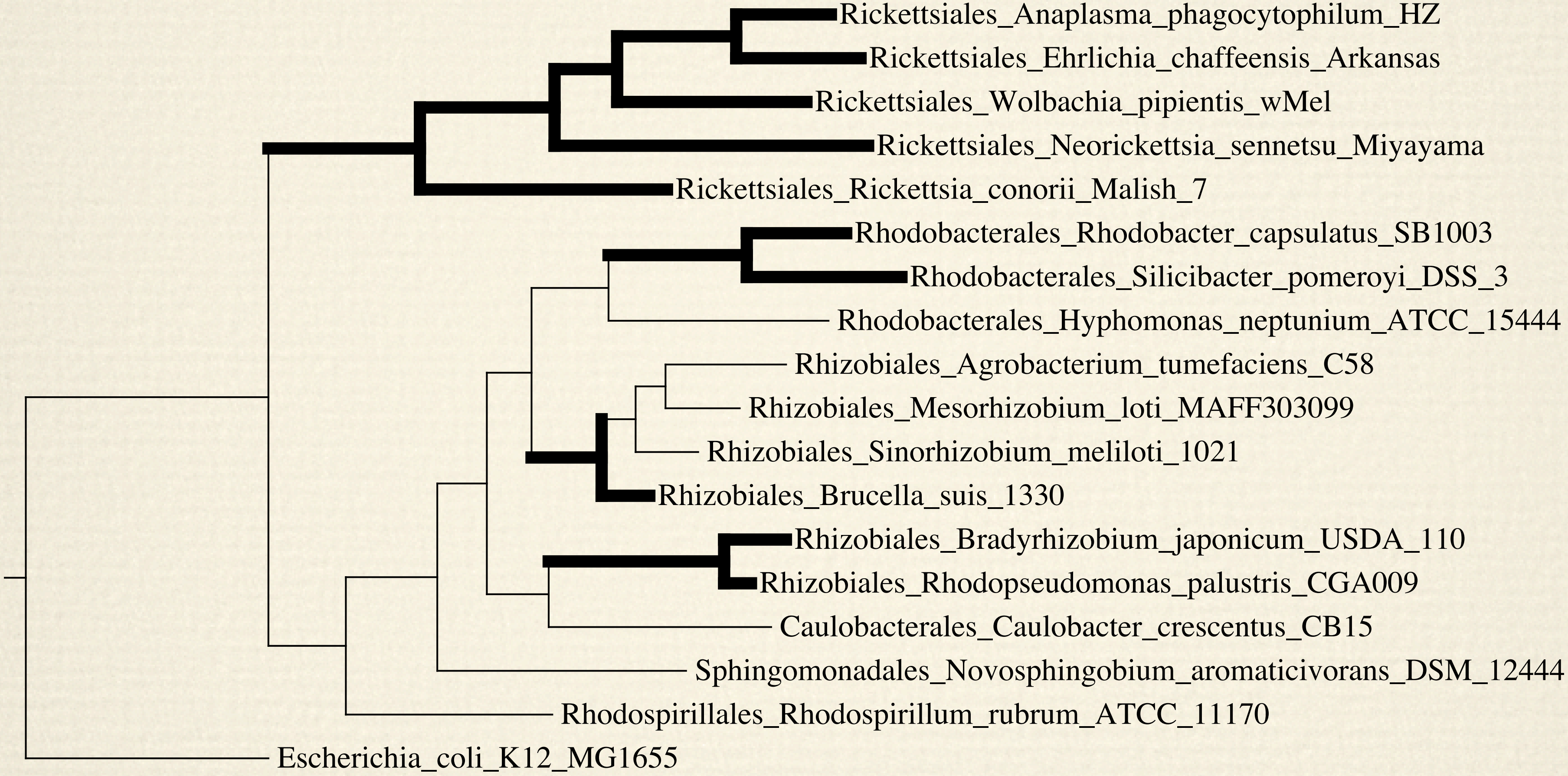
The Pains of PHYLIP

- Just to make a bootstrapped tree
  - Format Sequence (10 char header)
  - Make alignment & convert it to PHYLIP format
  - Run Seqboot
  - Run DNA(Prot)dist (if NJ)
  - Run Neighbor/DNA(Prot)Pars/DNA(Prot)ML
  - Run Consense
  - Reedit tree to get back meaningful name
- And programs not easily scriptable (use interactive input)

Enter phylipFasta

- From about 20 mins of human time to a couple of seconds
- Creates PDF of trees (as seen in this presentation)
- To make ML tree -- just two lines!
    - muscle -in seqs.fasta -out seqs.afa
    - phylipFasta -m ml seqs.afa
- So easy, even faculty members can do it!

Example of Tree


Rickettsiales_Anaplasma_phagocytophilum_HZ
Rickettsiales_Ehrlichia_chaffeensis_Arkansas
Rickettsiales_Wolbachia_pipientis_wMel
Rickettsiales_Neorickettsia_sennetsu_Miyayama
Rickettsiales_Rickettsia_conorii_Malish_7
Rhodobacterales_Rhodobacter_capsulatus_SB1003
Rhodobacterales_Silicibacter_pomeroyi_DSS_3
Rhodobacterales_Hyphomonas_neptunium_ATCC_15444
Rhizobiales_Agrobacterium_tumefaciens_C58
Rhizobiales_Mesorhizobium_loti_MAFF303099
Rhizobiales_Sinorhizobium_meliloti_1021
Rhizobiales_Brucella_suis_1330
Rhizobiales_Bradyrhizobium_japonicum_USDA_110
Rhizobiales_Rhodopseudomonas_palustris_CGA009
Caulobacterales_Caulobacter_crescentus_CB15
Sphingomonadales_Novosphingobium_aromaticivorans_DSM_12444
Rhodospirillales_Rhodospirillum_rubrum_ATCC_11170
Escherichia_coli_K12_MG1655

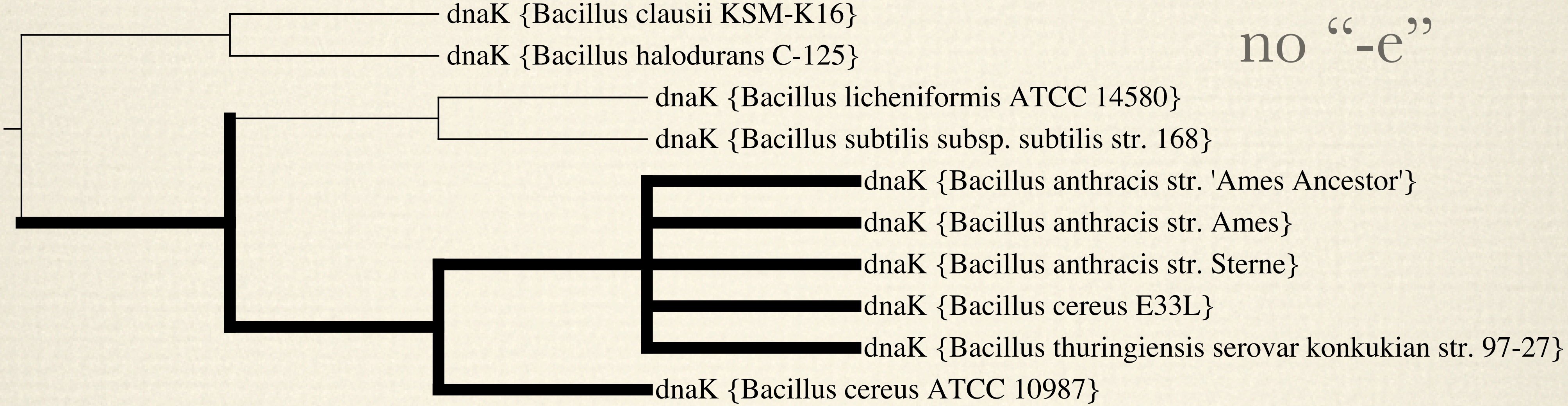
In trees created by my script, branch lengths are proportional to amount of substitution/replacement, dark branches are clades >75% bootstrap support.

phylipFasta options

```
Usage: phylipFasta [options] fasta [fasta..]
    -b, --noboot             don't bootstrap (default false)
    -n, --numboot            number of bootstrap reps  (default 100)
    -e, --estimate           estimate branch lengths (default false)
    -g, --gamma              use gamma distrib. rates (default false)
    -m, --method             method (default nj -  nj, pars, ml valid)
    -o, --outgroup           outgroup
    -p, --pam                use PAM model (default false)
```

The joys of “-e”

no “-e”



dnaK {Bacillus clausii KSM-K16}

dnaK {Bacillus halodurans C-125}

dnaK {Bacillus licheniformis ATCC 14580}

dnaK {Bacillus subtilis subsp. subtilis str. 168}

dnaK {Bacillus anthracis str. 'Ames Ancestor'}

dnaK {Bacillus anthracis str. Ames}

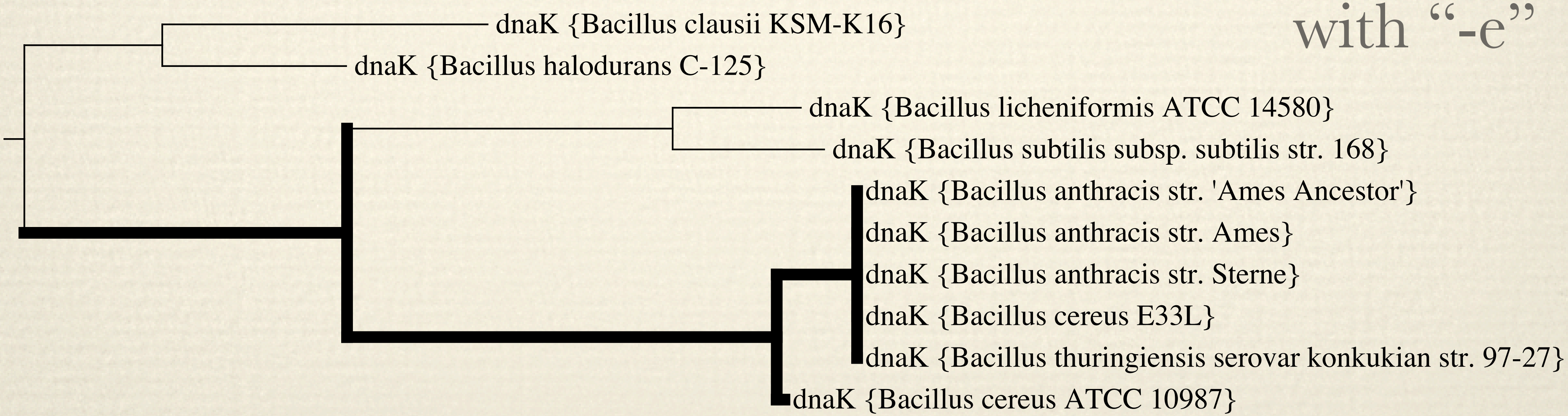
dnaK {Bacillus anthracis str. Sterne}

dnaK {Bacillus cereus E33L}

dnaK {Bacillus thuringiensis serovar konkukian str. 97-27}

dnaK {Bacillus cereus ATCC 10987}

with “-e”



dnaK {Bacillus clausii KSM-K16}

dnaK {Bacillus halodurans C-125}

dnaK {Bacillus licheniformis ATCC 14580}

dnaK {Bacillus subtilis subsp. subtilis str. 168}

dnaK {Bacillus anthracis str. 'Ames Ancestor'}

dnaK {Bacillus anthracis str. Ames}

dnaK {Bacillus anthracis str. Sterne}

dnaK {Bacillus cereus E33L}

dnaK {Bacillus thuringiensis serovar konkukian str. 97-27}

dnaK {Bacillus cereus ATCC 10987}

paupFasta

```
Usage: paupFasta [options] fasta [fasta...]
    -b, --noboot         don't bootstrap (default false)
    -n, --numboot        number of bootstrap reps  (default 100)
    -e, --estimate       estimate branch lengths (default false)
    -m, --method         method (default pars)
    -o, --outgroup       outgroup
```

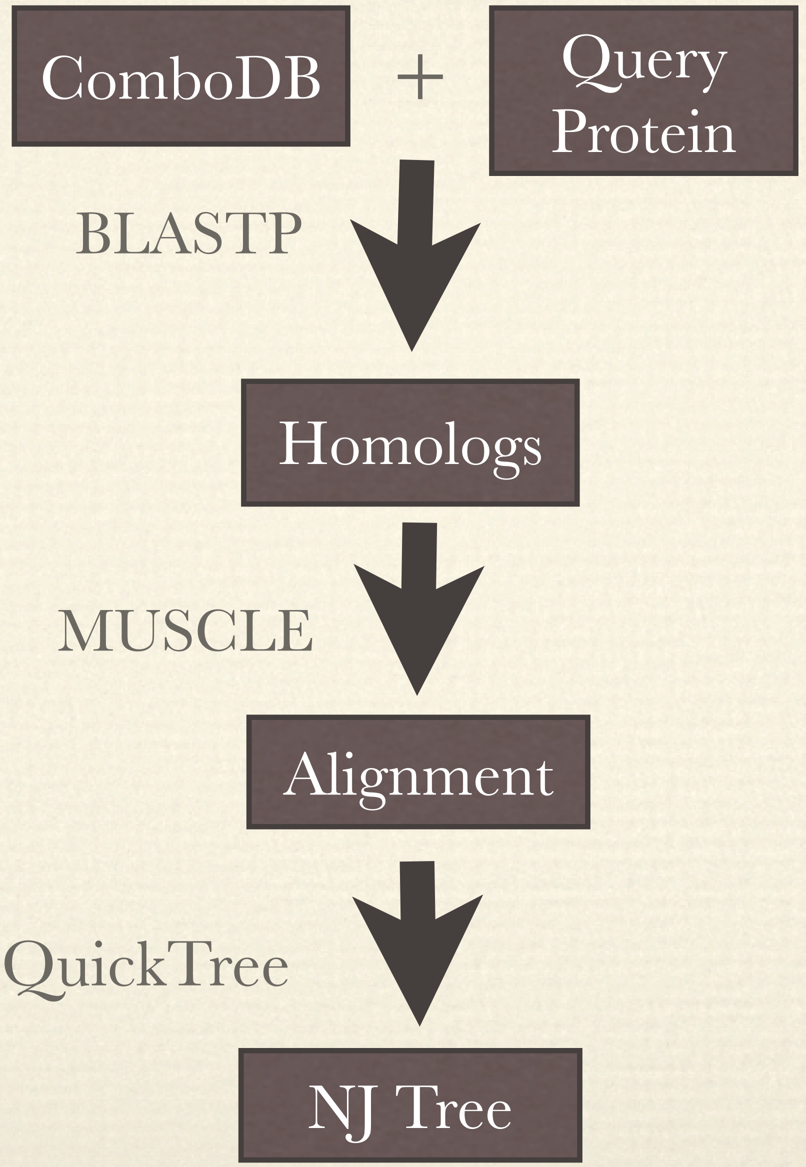
APIS

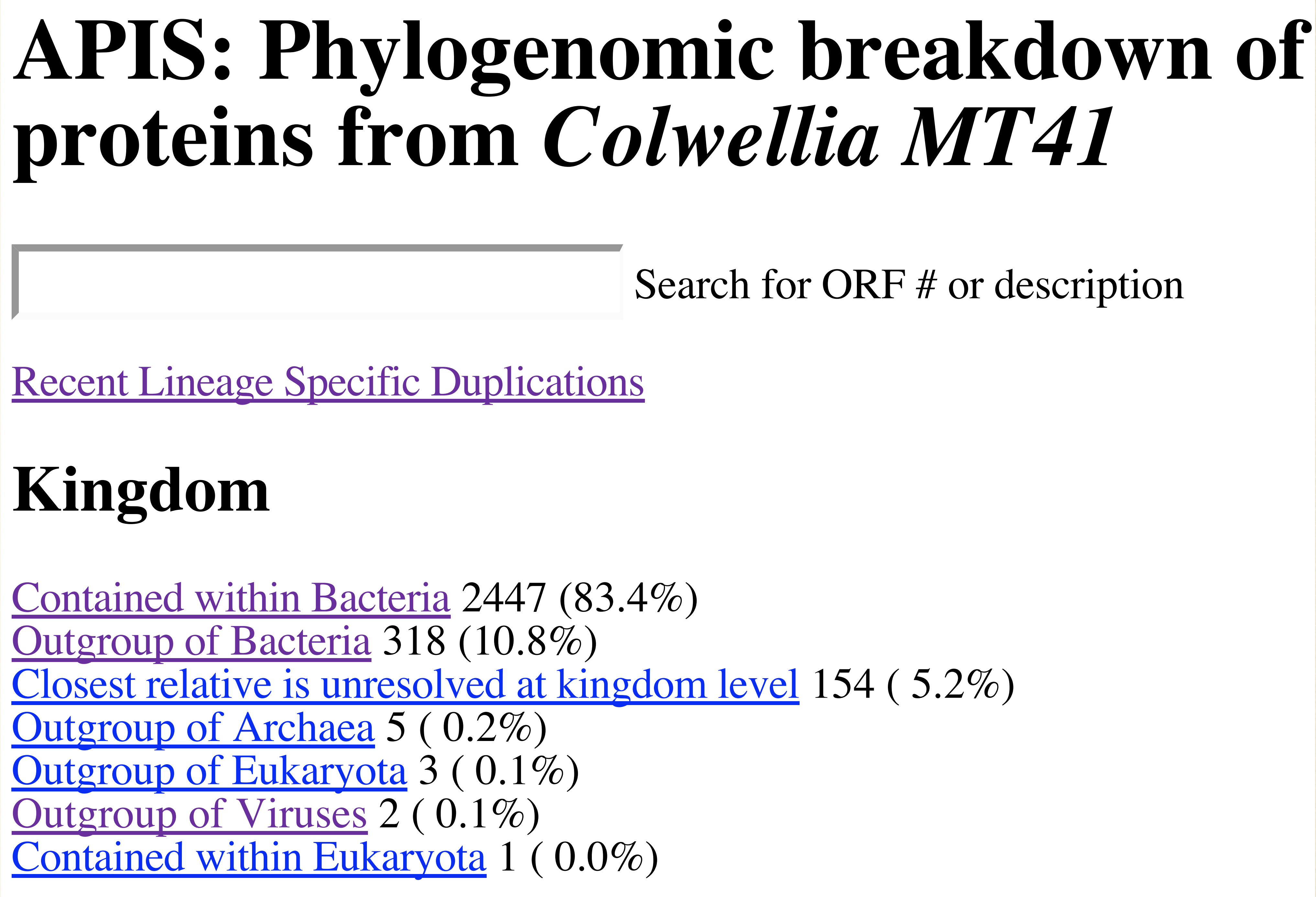
Rationale for APIS

- Traditional taxonomy based on one gene - 16S rRNA
- Sometimes other markers such as HSP70 used, and sometimes this conflicts with 16S
- But systematic phylogeny of all proteins in genome individually not done -- would take too much manual effort
- APIS makes this possible; it makes trees for each protein and summarizes results.
- Helps find LGT, misclassified organisms

APIS Outline

For each query protein in genome:





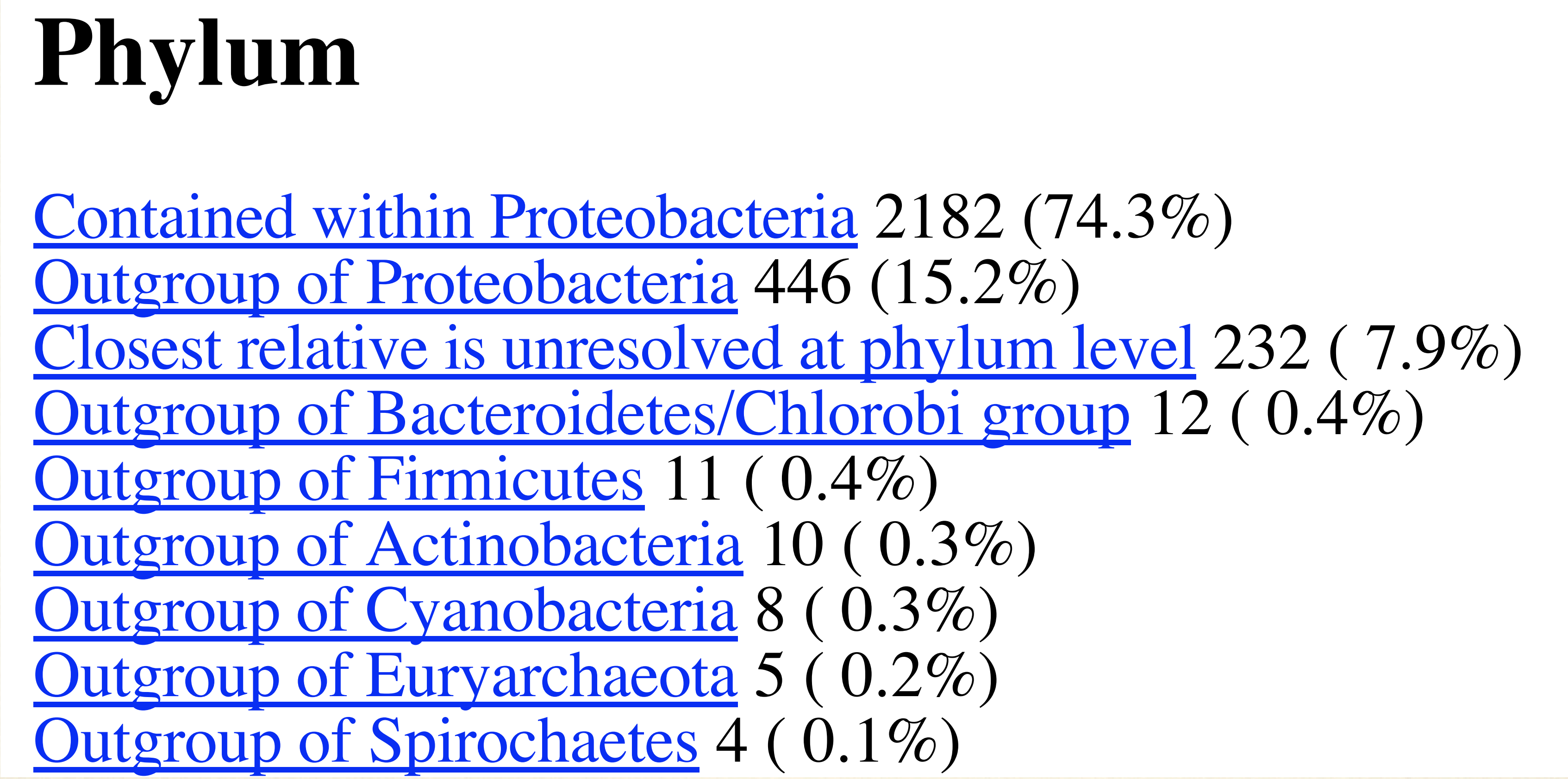
APIS: Phylogenomic breakdown of proteins from *Colwellia MT41*

Search for ORF # or description

Recent Lineage Specific Duplications

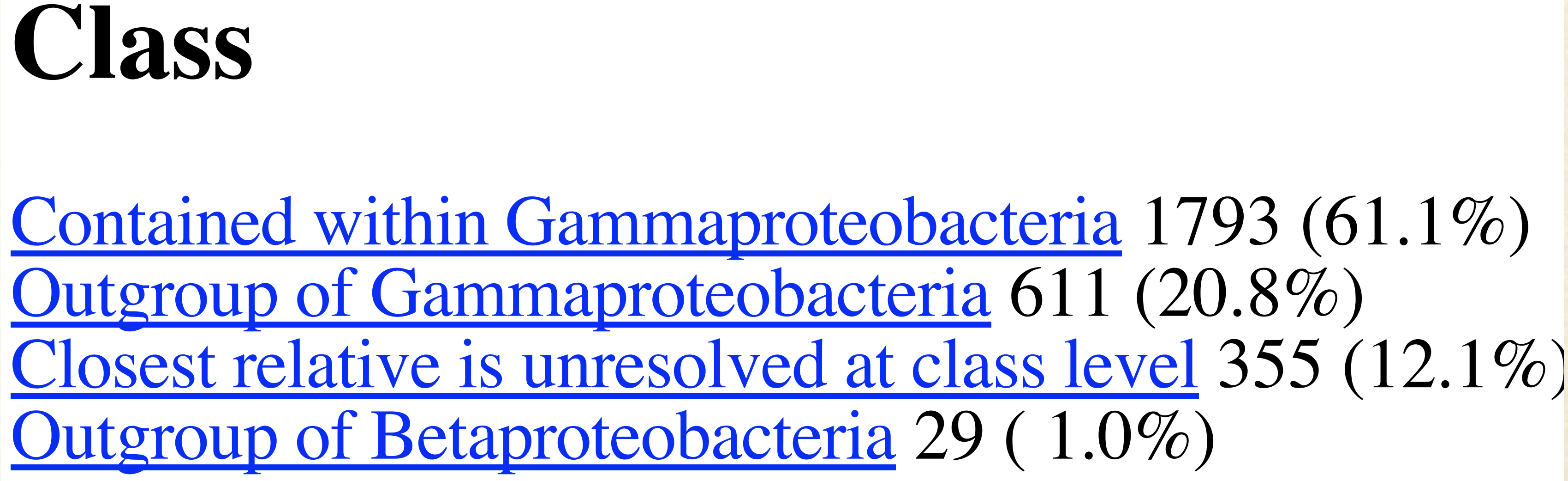
Kingdom

Contained within Bacteria 2447 (83.4%)
Outgroup of Bacteria 318 (10.8%)
Closest relative is unresolved at kingdom level 154 (5.2%)
Outgroup of Archaea 5 (0.2%)
Outgroup of Eukaryota 3 (0.1%)
Outgroup of Viruses 2 (0.1%)
Contained within Eukaryota 1 (0.0%)



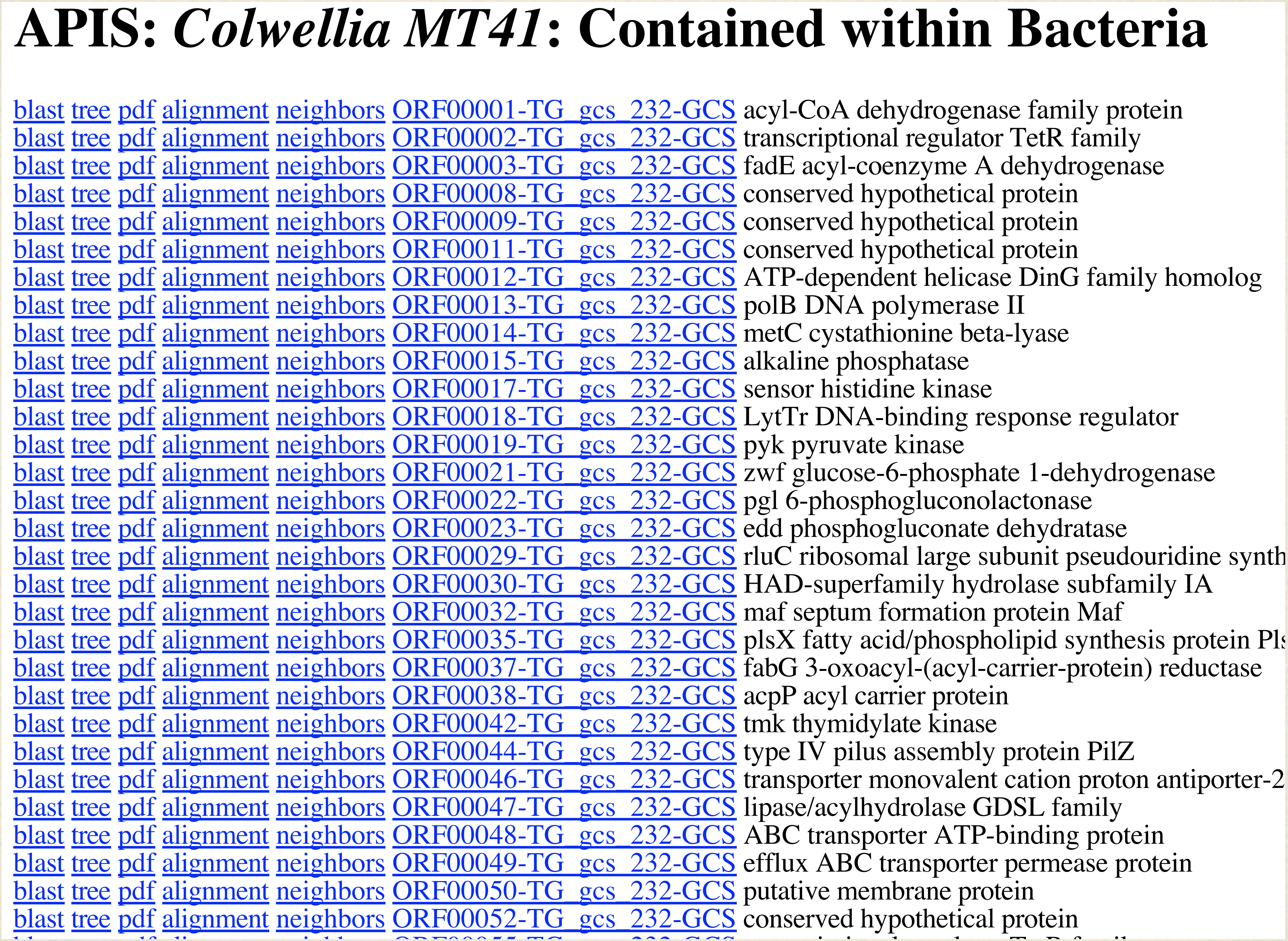
Phylum

Contained within Proteobacteria 2182 (74.3%)
Outgroup of Proteobacteria 446 (15.2%)
Closest relative is unresolved at phylum level 232 (7.9%)
Outgroup of Bacteroidetes/Chlorobi group 12 (0.4%)
Outgroup of Firmicutes 11 (0.4%)
Outgroup of Actinobacteria 10 (0.3%)
Outgroup of Cyanobacteria 8 (0.3%)
Outgroup of Euryarchaeota 5 (0.2%)
Outgroup of Spirochaetes 4 (0.1%)



Class

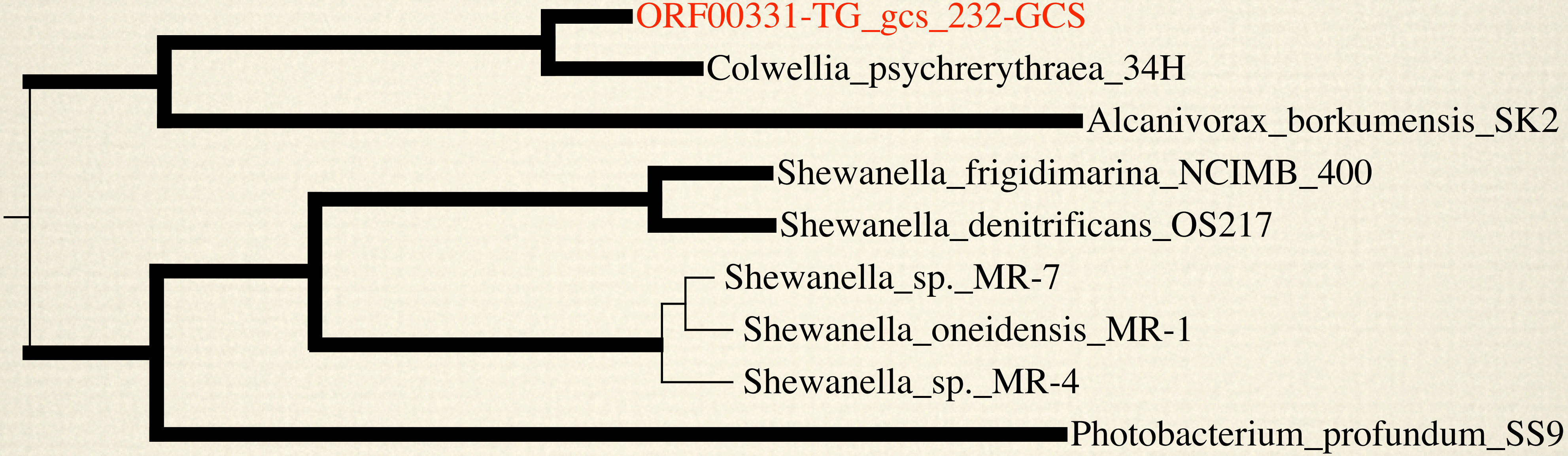
Contained within Gammaproteobacteria 1793 (61.1%)
Outgroup of Gammaproteobacteria 611 (20.8%)
Closest relative is unresolved at class level 355 (12.1%)
Outgroup of Betaproteobacteria 29 (1.0%)



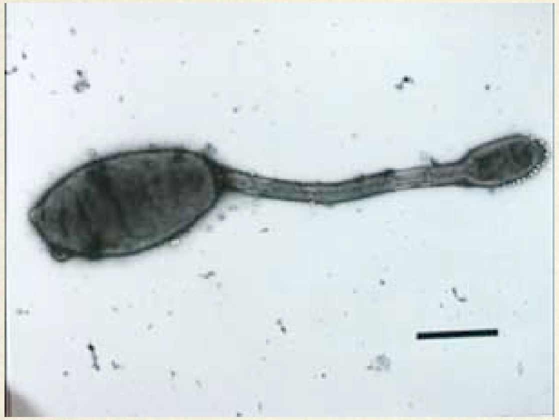
APIS: *Colwellia MT41*: Contained within Bacteria

blast tree pdf alignment neighbors ORF00001-TG_gcs_232-GCS acyl-CoA dehydrogenase family protein
blast tree pdf alignment neighbors ORF00002-TG_gcs_232-GCS transcriptional regulator TetR family
blast tree pdf alignment neighbors ORF00003-TG_gcs_232-GCS fadE acyl-coenzyme A dehydrogenase
blast tree pdf alignment neighbors ORF00008-TG_gcs_232-GCS conserved hypothetical protein
blast tree pdf alignment neighbors ORF00009-TG_gcs_232-GCS conserved hypothetical protein
blast tree pdf alignment neighbors ORF00011-TG_gcs_232-GCS conserved hypothetical protein
blast tree pdf alignment neighbors ORF00012-TG_gcs_232-GCS ATP-dependent helicase DinG family homolog
blast tree pdf alignment neighbors ORF00013-TG_gcs_232-GCS polB DNA polymerase II
blast tree pdf alignment neighbors ORF00014-TG_gcs_232-GCS metC cystathionine beta-lyase
blast tree pdf alignment neighbors ORF00015-TG_gcs_232-GCS alkaline phosphatase
blast tree pdf alignment neighbors ORF00017-TG_gcs_232-GCS sensor histidine kinase
blast tree pdf alignment neighbors ORF00018-TG_gcs_232-GCS LytTr DNA-binding response regulator
blast tree pdf alignment neighbors ORF00019-TG_gcs_232-GCS pyk pyruvate kinase
blast tree pdf alignment neighbors ORF00021-TG_gcs_232-GCS zwf glucose-6-phosphate 1-dehydrogenase
blast tree pdf alignment neighbors ORF00022-TG_gcs_232-GCS pgl 6-phosphogluconolactonase
blast tree pdf alignment neighbors ORF00023-TG_gcs_232-GCS edd phosphogluconate dehydratase
blast tree pdf alignment neighbors ORF00029-TG_gcs_232-GCS rluC ribosomal large subunit pseudouridine synth
blast tree pdf alignment neighbors ORF00030-TG_gcs_232-GCS HAD-superfamily hydrolase subfamily IA
blast tree pdf alignment neighbors ORF00032-TG_gcs_232-GCS maf septum formation protein Maf
blast tree pdf alignment neighbors ORF00035-TG_gcs_232-GCS plsX fatty acid/phospholipid synthesis protein Pls
blast tree pdf alignment neighbors ORF00037-TG_gcs_232-GCS fabG 3-oxoacyl-(acyl-carrier-protein) reductase
blast tree pdf alignment neighbors ORF00038-TG_gcs_232-GCS acpP acyl carrier protein
blast tree pdf alignment neighbors ORF00042-TG_gcs_232-GCS tmk thymidylate kinase
blast tree pdf alignment neighbors ORF00044-TG_gcs_232-GCS type IV pilus assembly protein PilZ
blast tree pdf alignment neighbors ORF00046-TG_gcs_232-GCS transporter monovalent cation proton antiporter-2
blast tree pdf alignment neighbors ORF00047-TG_gcs_232-GCS lipase/acylhydrolase GDSL family
blast tree pdf alignment neighbors ORF00048-TG_gcs_232-GCS ABC transporter ATP-binding protein
blast tree pdf alignment neighbors ORF00049-TG_gcs_232-GCS efflux ABC transporter permease protein
blast tree pdf alignment neighbors ORF00050-TG_gcs_232-GCS putative membrane protein
blast tree pdf alignment neighbors ORF00052-TG_gcs_232-GCS conserved hypothetical protein
blast tree pdf alignment neighbors ORF00055-TG_gcs_232-GCS transcriptional regulator TetR family

APIS Tree

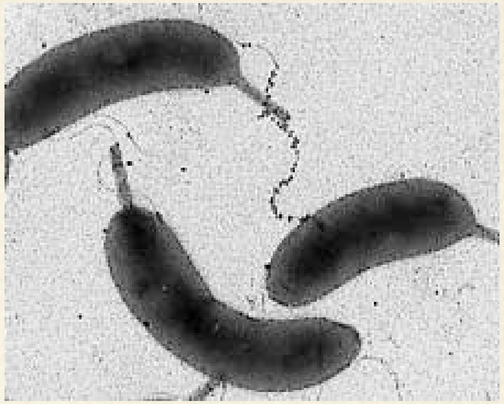

ORF00331-TG_gcs_232-GCS
Colwellia_psychrerythraea_34H
Alcanivorax_borkumensis_SK2
Shewanella_frigidimarina_NCIMB_400
Shewanella_denitrificans_OS217
Shewanella_sp._MR-7
Shewanella_oneidensis_MR-1
Shewanella_sp._MR-4
Photobacterium_profundum_SS9

What Can APIS Tell Us?



H. neptunium (Order Rhodobacterales)

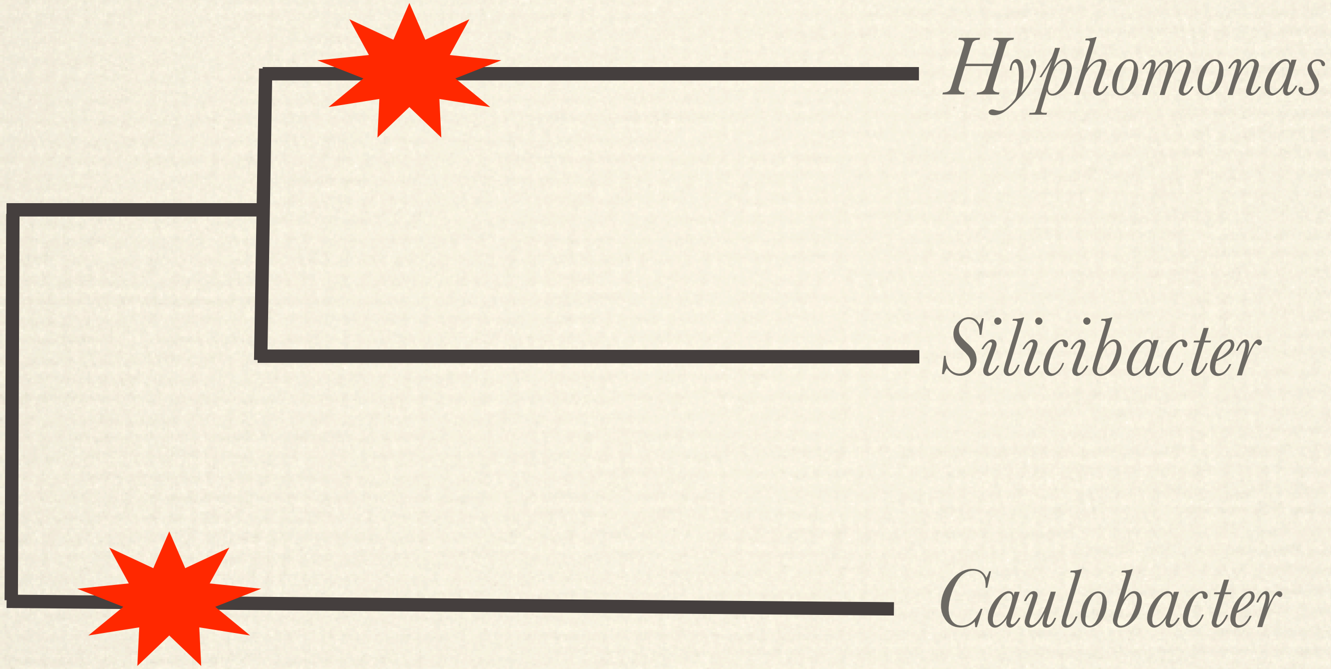
According to APIS, 30% of *Hyphomonas* proteins cluster with *Caulobacter*; only 6% with *Silicibacter*. Why? That goes against the current classification! What is going on?



C. crescentus
(Order Caulobacterales)

S. pomeroyi
(Order Rhodobacterales)

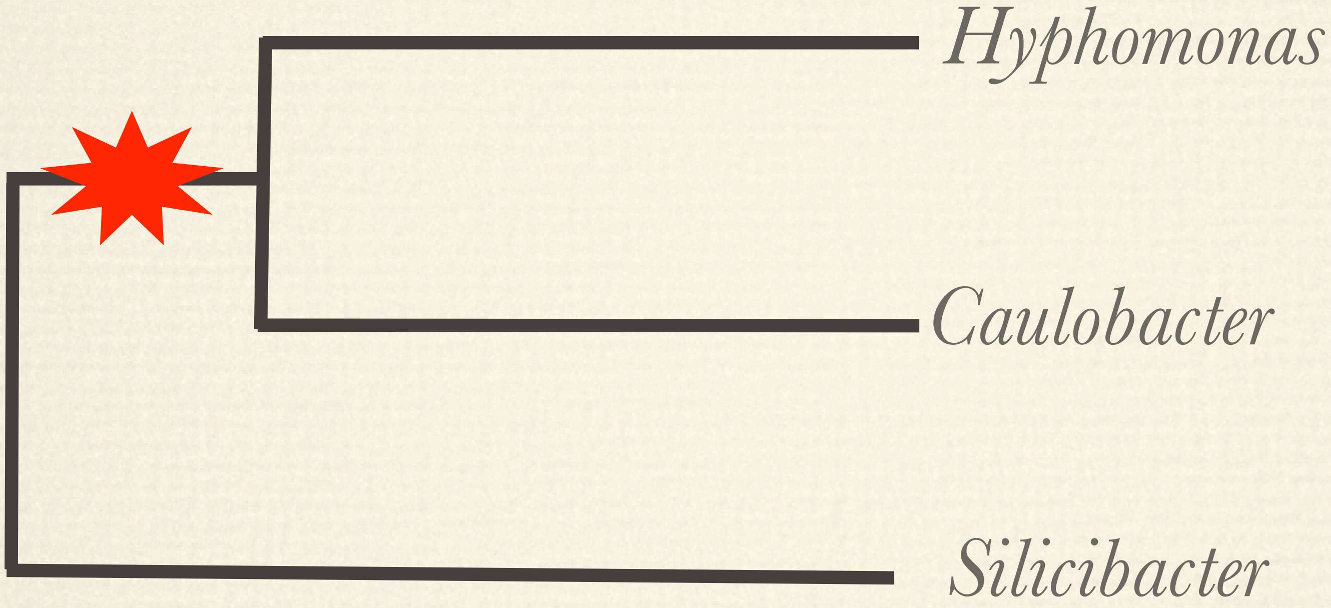
Scenario I



Current Classification according to 16S

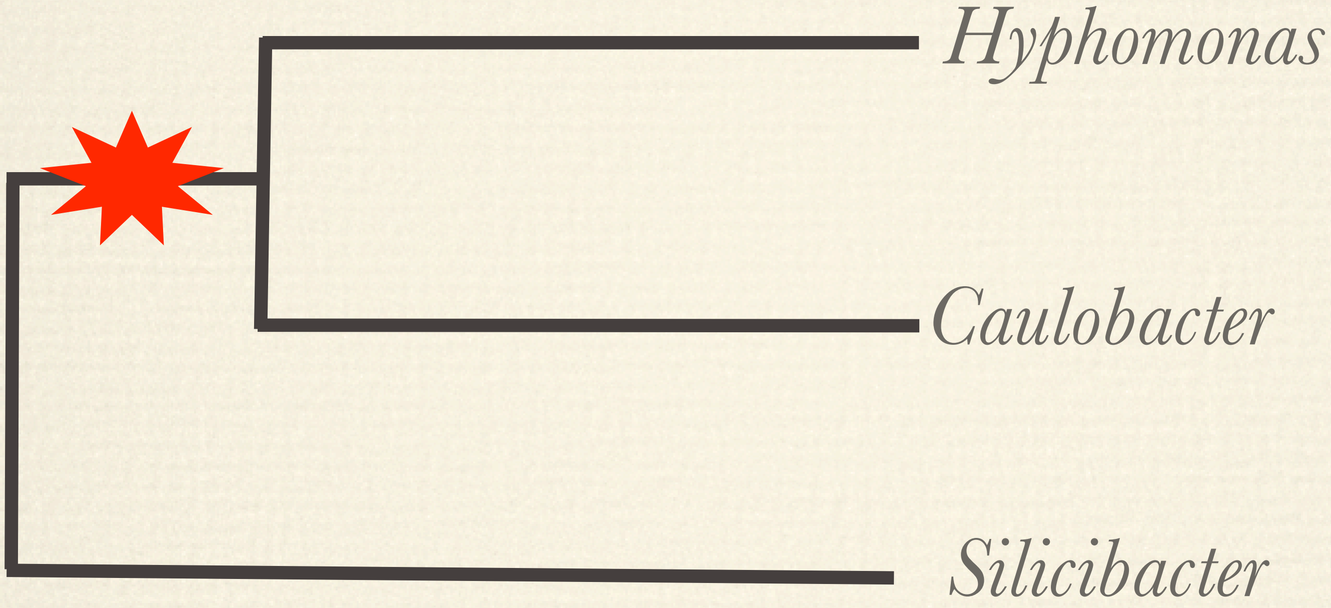
Would require independent evolution of prosthecate lifestyle in *Hyphomonas* & *Caulobacter* lineages (stars)

Scenario II



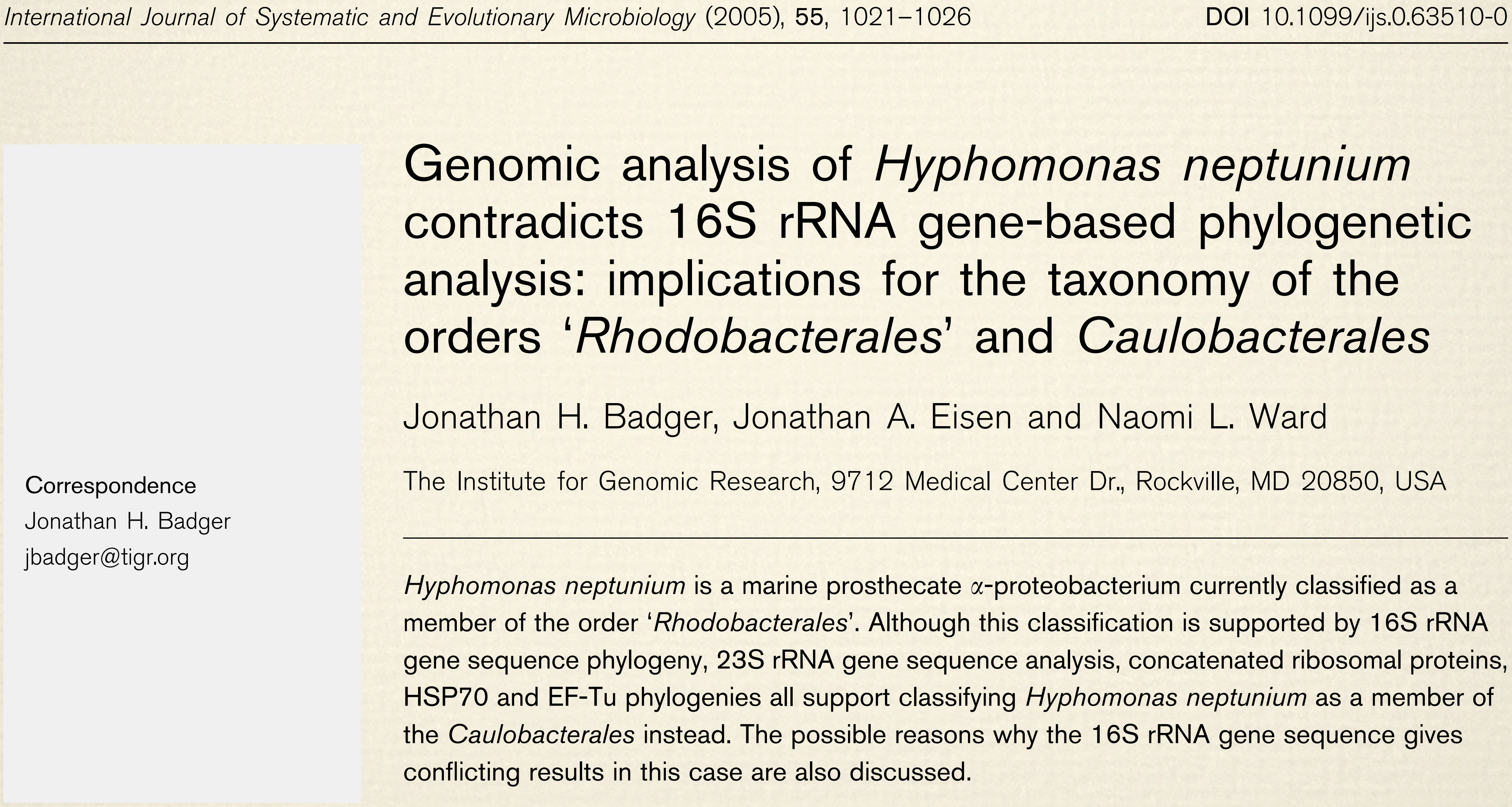
Scenario Supported by APIS - and by morphology
Prosthecate lifestyle evolved once, in shared lineage
(star)

Additional Support for Scenario II



This scenario also supported by other accepted markers, 23S rRNA, EF-Tu, HSP70, concatenated ribosomal proteins

IJSEM Publication



International Journal of Systematic and Evolutionary Microbiology (2005), **55**, 1021–1026 DOI 10.1099/ijs.0.63510-0

Genomic analysis of *Hyphomonas neptunium* contradicts 16S rRNA gene-based phylogenetic analysis: implications for the taxonomy of the orders '*Rhodobacterales*' and *Caulobacterales*

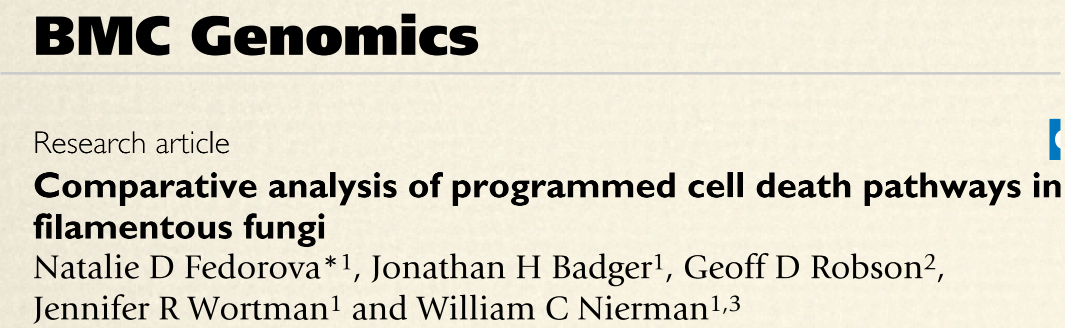
Jonathan H. Badger, Jonathan A. Eisen and Naomi L. Ward

The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, MD 20850, USA

Correspondence
Jonathan H. Badger
jbadger@tigr.org

Hyphomonas neptunium is a marine prosthecate α-proteobacterium currently classified as a member of the order '*Rhodobacterales*'. Although this classification is supported by 16S rRNA gene sequence phylogeny, 23S rRNA gene sequence analysis, concatenated ribosomal proteins, HSP70 and EF-Tu phylogenies all support classifying *Hyphomonas neptunium* as a member of the *Caulobacterales* instead. The possible reasons why the 16S rRNA gene sequence gives conflicting results in this case are also discussed.

Other Publications using APIS

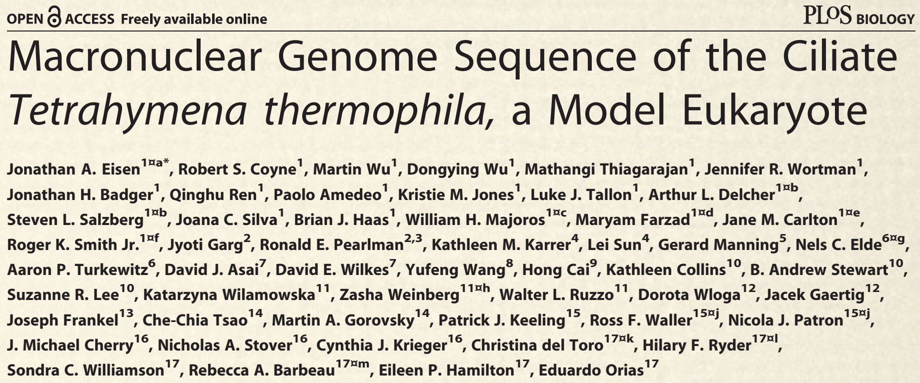


BMC Genomics

Research article

Comparative analysis of programmed cell death pathways in filamentous fungi

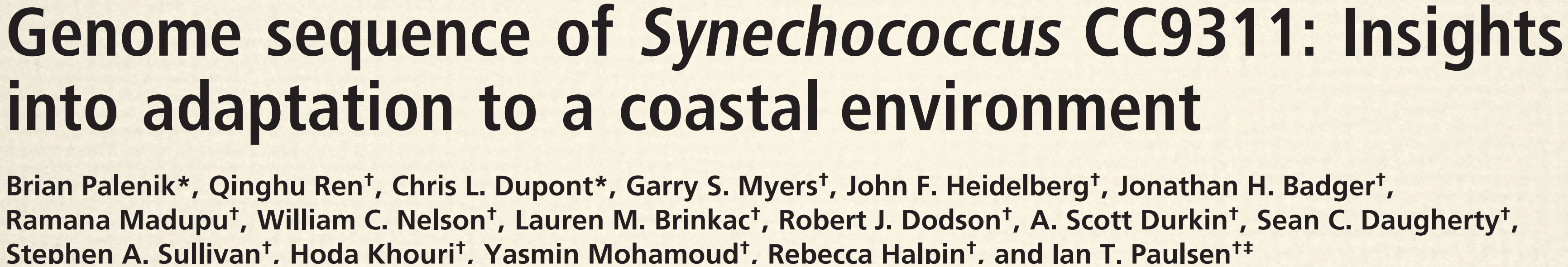
Natalie D Fedorova*[1], Jonathan H Badger[1], Geoff D Robson[2], Jennifer R Wortman[1] and William C Nierman[1,3]



OPEN ACCESS Freely available online PLoS BIOLOGY

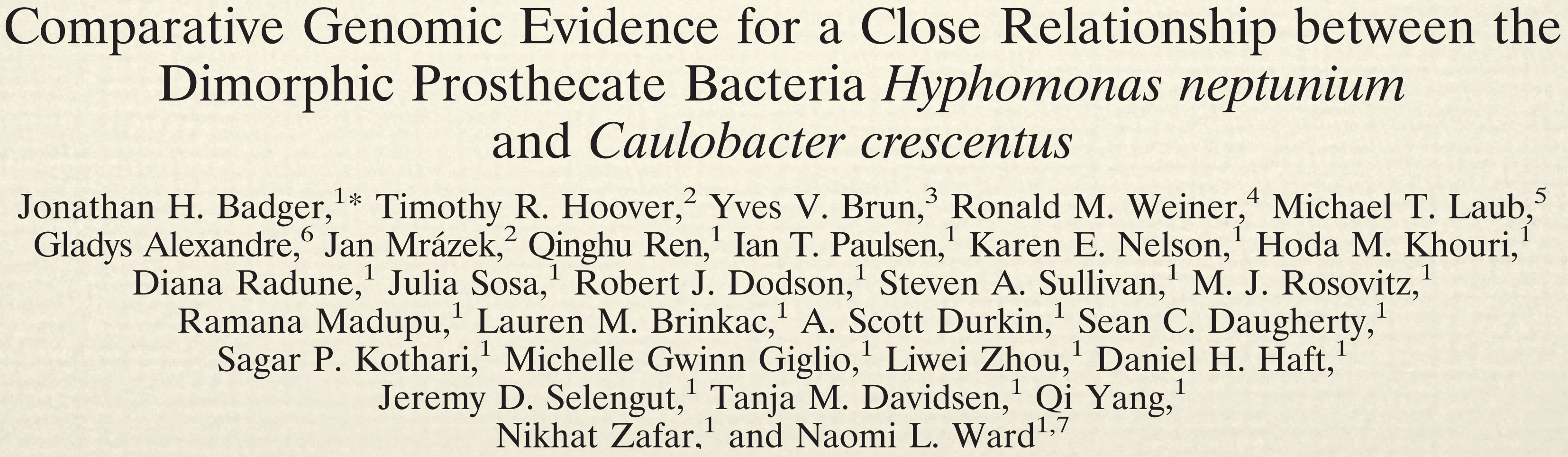
Macronuclear Genome Sequence of the Ciliate *Tetrahymena thermophila,* a Model Eukaryote

Jonathan A. Eisen[1¤a*], Robert S. Coyne[1], Martin Wu[1], Dongying Wu[1], Mathangi Thiagarajan[1], Jennifer R. Wortman[1], Jonathan H. Badger[1], Qinghu Ren[1], Paolo Amedeo[1], Kristie M. Jones[1], Luke J. Tallon[1], Arthur L. Delcher[1¤b], Steven L. Salzberg[1¤b], Joana C. Silva[1], Brian J. Haas[1], William H. Majoros[1¤c], Maryam Farzad[1¤d], Jane M. Carlton[1¤e], Roger K. Smith Jr.[1¤f], Jyoti Garg[2], Ronald E. Pearlman[2,3], Kathleen M. Karrer[4], Lei Sun[4], Gerard Manning[5], Nels C. Elde[6¤g], Aaron P. Turkewitz[6], David J. Asai[7], David E. Wilkes[7], Yufeng Wang[8], Hong Cai[9], Kathleen Collins[10], B. Andrew Stewart[10], Suzanne R. Lee[10], Katarzyna Wilamowska[11], Zasha Weinberg[11¤h], Walter L. Ruzzo[11], Dorota Wloga[12], Jacek Gaertig[12], Joseph Frankel[13], Che-Chia Tsao[14], Martin A. Gorovsky[14], Patrick J. Keeling[15], Ross F. Waller[15¤j], Nicola J. Patron[15¤j], J. Michael Cherry[16], Nicholas A. Stover[16], Cynthia J. Krieger[16], Christina del Toro[17¤k], Hilary F. Ryder[17¤l], Sondra C. Williamson[17], Rebecca A. Barbeau[17¤m], Eileen P. Hamilton[17], Eduardo Orias[17]



Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment

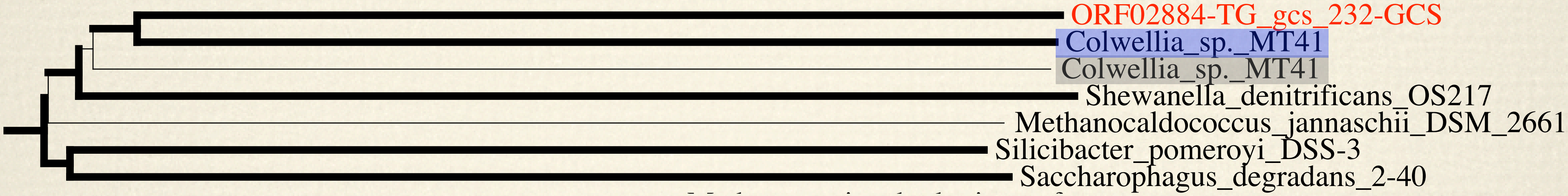
Brian Palenik*, Qinghu Ren†, Chris L. Dupont*, Garry S. Myers†, John F. Heidelberg†, Jonathan H. Badger†, Ramana Madupu†, William C. Nelson†, Lauren M. Brinkac†, Robert J. Dodson†, A. Scott Durkin†, Sean C. Daugherty†, Stephen A. Sullivan†, Hoda Khouri†, Yasmin Mohamoud†, Rebecca Halpin†, and Ian T. Paulsen†‡



Comparative Genomic Evidence for a Close Relationship between the Dimorphic Prosthecate Bacteria *Hyphomonas neptunium* and *Caulobacter crescentus*

Jonathan H. Badger,[1*] Timothy R. Hoover,[2] Yves V. Brun,[3] Ronald M. Weiner,[4] Michael T. Laub,[5] Gladys Alexandre,[6] Jan Mrázek,[2] Qinghu Ren,[1] Ian T. Paulsen,[1] Karen E. Nelson,[1] Hoda M. Khouri,[1] Diana Radune,[1] Julia Sosa,[1] Robert J. Dodson,[1] Steven A. Sullivan,[1] M. J. Rosovitz,[1] Ramana Madupu,[1] Lauren M. Brinkac,[1] A. Scott Durkin,[1] Sean C. Daugherty,[1] Sagar P. Kothari,[1] Michelle Gwinn Giglio,[1] Liwei Zhou,[1] Daniel H. Haft,[1] Jeremy D. Selengut,[1] Tanja M. Davidsen,[1] Qi Yang,[1] Nikhat Zafar,[1] and Naomi L. Ward[1,7]

Recent Duplications


ORF02884-TG_gcs_232-GCS
Colwellia_sp._MT41
Colwellia_sp._MT41
Shewanella_denitrificans_OS217
Methanocaldococcus_jannaschii_DSM_2661
Silicibacter_pomeroyi_DSS-3
Saccharophagus_degradans_2-40

Duplication Cluster #40

ORF02884 peptidase M9A family

ORF03332 putative alkaline serine protease

ORF03839 microbial collagenase

A duplication cluster is a set of proteins which all group with other proteins from the same family -- evidence of expansion in lineage as other species have only one copy

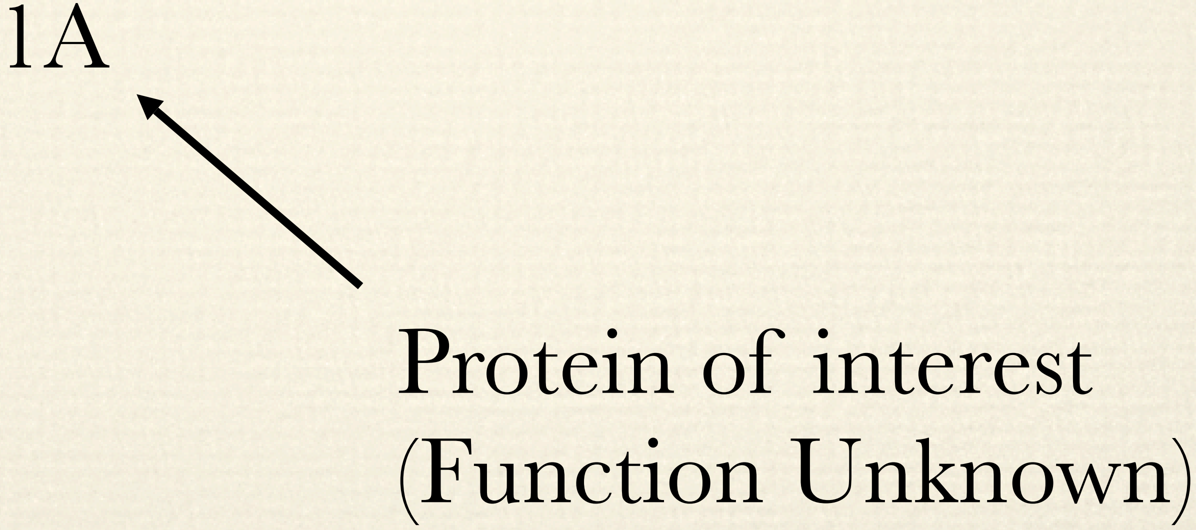
ECFinder

KEGG Pathways

- KEGG - Kyoto Encyclopedia of Genes & Genomes (Web site run by Kyoto University)
- Allows reconstruction of metabolic pathways for novel genome based on known pathways of a related organism.
- Uses EC numbers as input.
- Therefore, there needs to be some way of automating the assignment of EC numbers to genes in a genome - can phylogeny help us?

Phylogenomics

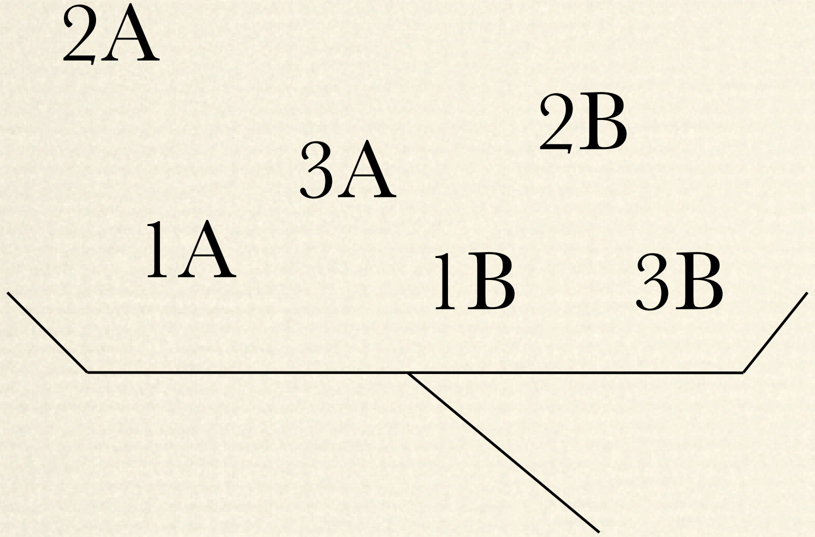
Jonathan Eisen's (1998) concept of assigning function by means of phylogenetic trees



1A

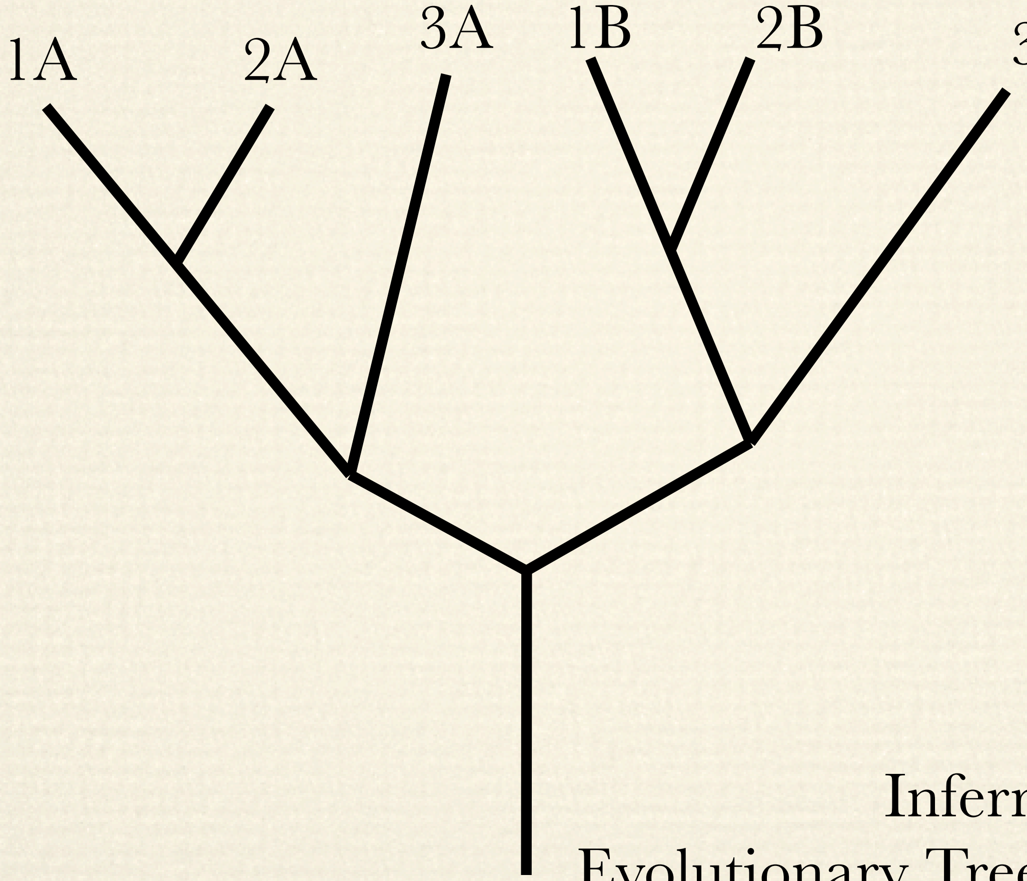
Protein of interest
(Function Unknown)

Phylogenomics



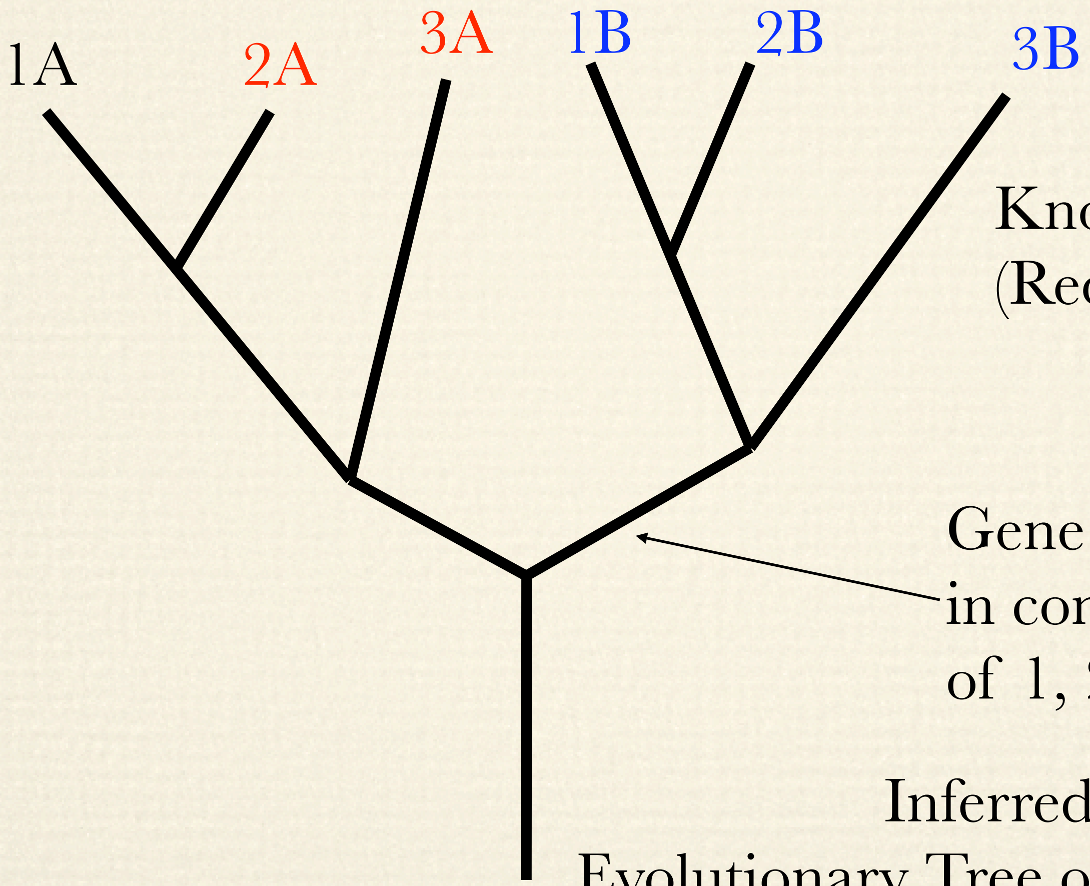
Homologs to 1A found by BLAST (Functions mostly known), from three species (1, 2, and 3)

Phylogenomics



Inferred
Evolutionary Tree of Sequences

Phylogenomics



Known Functions
(Red and Blue)

Gene Duplication
in common ancestor
of 1, 2, 3

Inferred
Evolutionary Tree of Sequences

EC Numbers

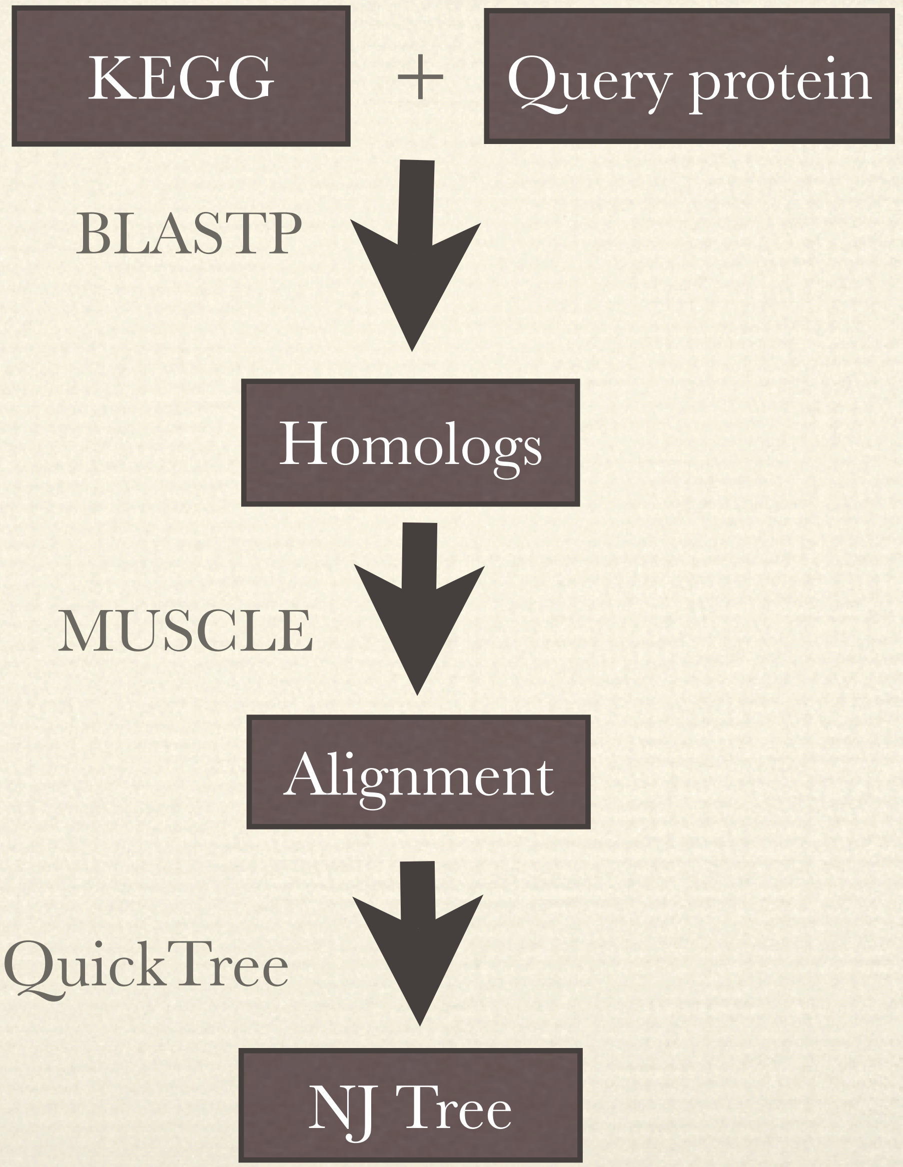
1. Oxidoreductases
    - 1.1 Acting on the CH-OH group of donors
        - 1.1.1 With NAD+ or NADP+ as acceptor
            - 1.1.1.1 alcohol dehydrogenase (NAD+)
            - 1.1.1.2 alcohol dehydrogenase (NADP+)
            - 1.1.1.3 homoserine dehydrogenase

Like Call Numbers in a Library;
Classification not Phylogeny

Allow automated comparison of enzyme annotation - no spelling issues, etc.

ECFinder Outline

For each query protein in genome:



KEGG + Query protein

BLASTP

Homologs

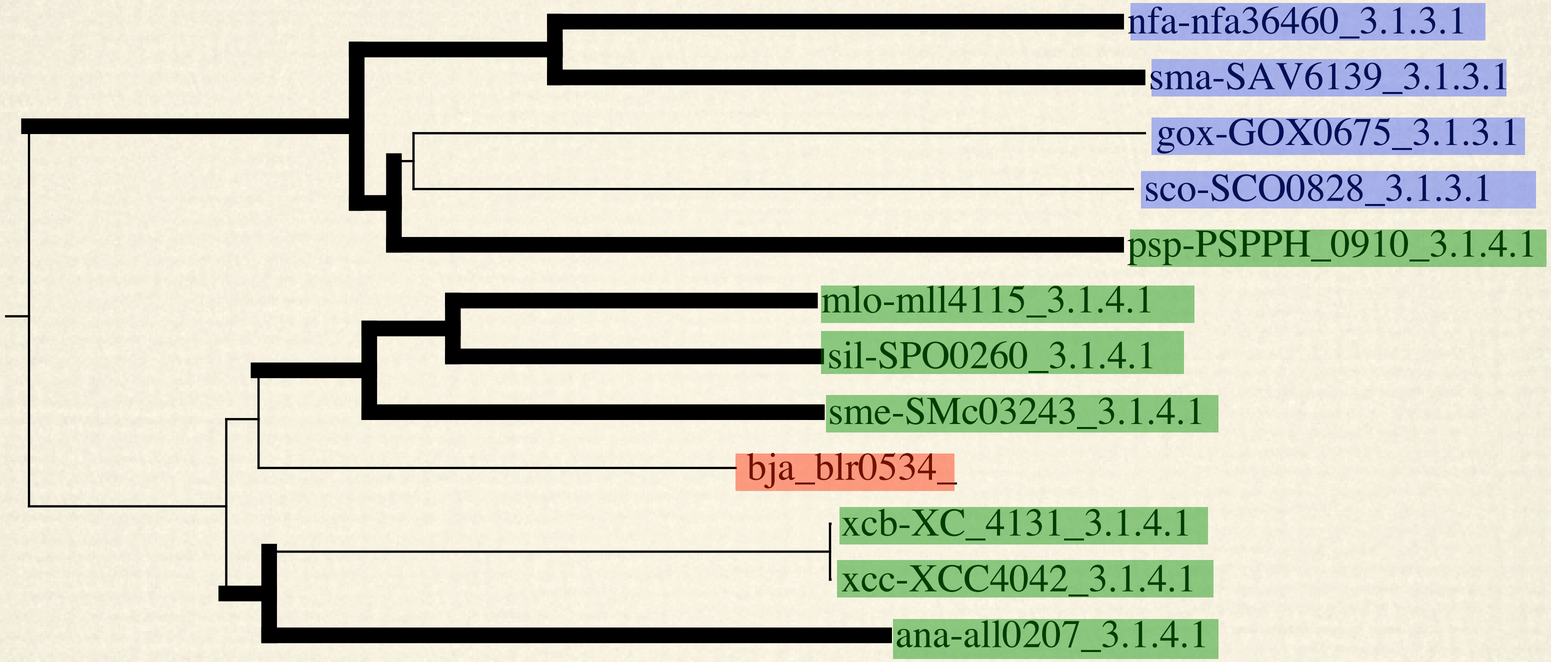
MUSCLE

Alignment

QuickTree

NJ Tree

ECFinder: Example



3.1.3.1: Alkaline phosphatase

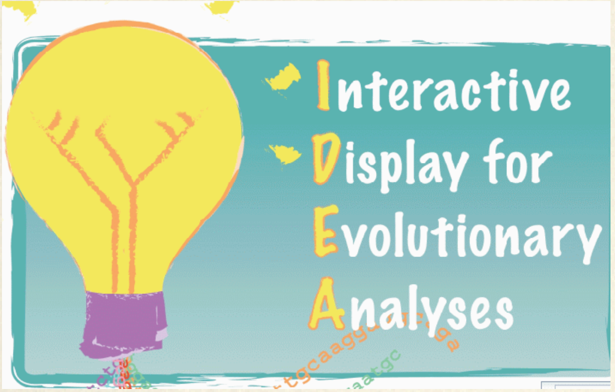
3.1.4.1: Phosphodiesterase

How well does it work?

- Test case: *B. japonicum* proteins using version of KEGG database with those removed
- 92% of the EC numbers the same
- Obviously, KEGG not perfect (combination of experimental & computational annotation) -- but shows that idea works.

IDEA

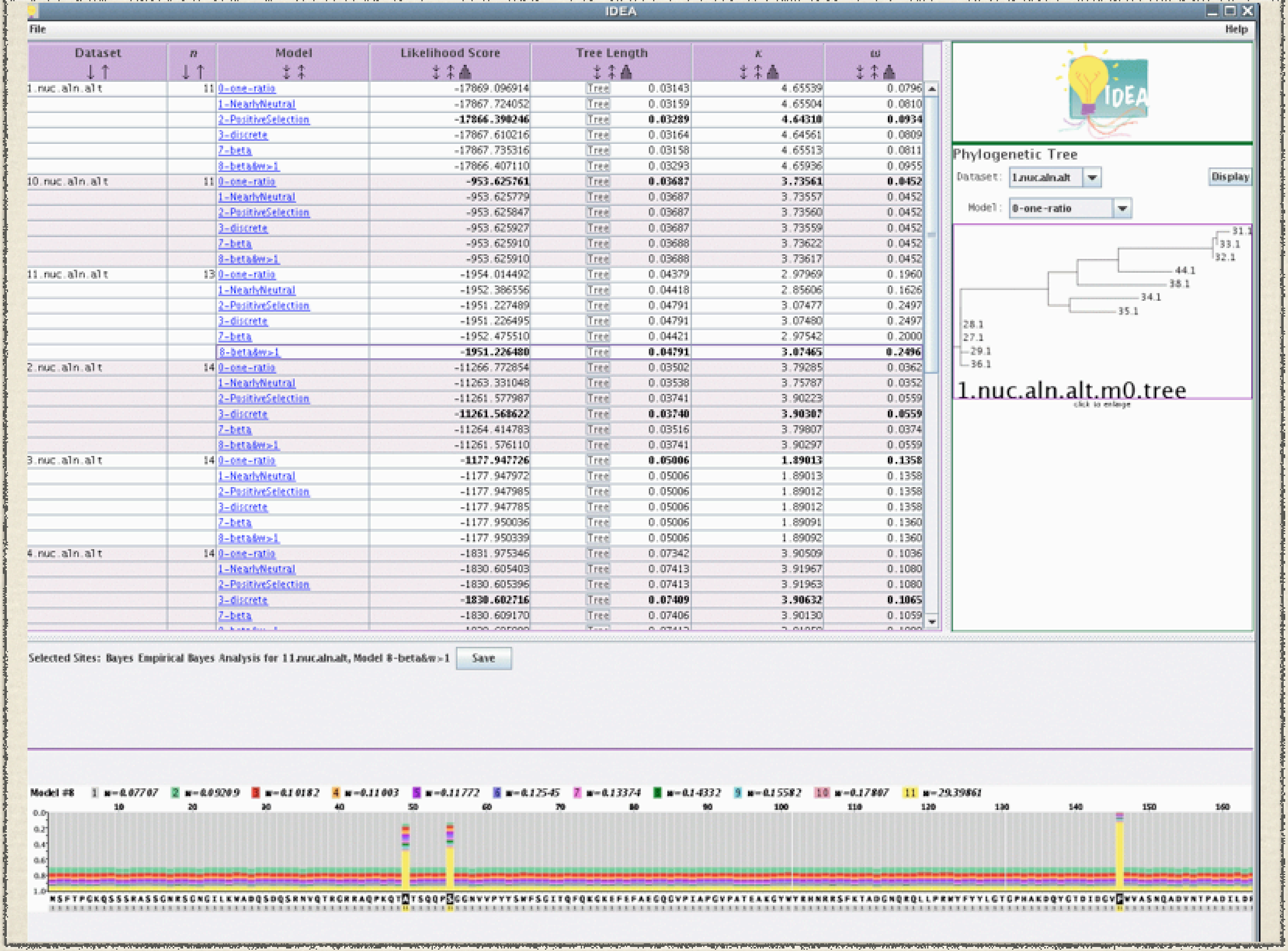
IDEA - an Interface for PAML



- Project of Joana Silva & Amy Egan
- Provides a friendly interface to scary PAML program
- Finds sites under positive selection - sites where amino acids changing faster than randomly expected for nucleotide changes.
- Needs phylogeny of sequences -- uses phylipFasta routines

Why Study Positive Selection?

- Besides interest in evolution itself, understanding which residues are evolving quickly can help us:
  - Avoid such sites in designing vaccines
  - Understand what sites are responsible for specificity - protein engineering



Dataset	n	Model	Likelihood Score	Tree Length	κ	ω
1.nuc.aln.alt	11	0-one-ratio	-17869.096914	Tree 0.03143	4.65539	0.0796
		1-NearlyNeutral	-17867.724052	Tree 0.03159	4.65504	0.0810
		2-PositiveSelection	**-17866.390246**	Tree **0.03289**	**4.64310**	**0.0934**
		3-discrete	-17867.610216	Tree 0.03164	4.64561	0.0809
		7-beta	-17867.735316	Tree 0.03158	4.65513	0.0811
		8-beta&w>1	-17866.407110	Tree 0.03293	4.65936	0.0955
10.nuc.aln.alt	11	0-one-ratio	**-953.625761**	Tree **0.03687**	**3.73561**	**0.0452**
		1-NearlyNeutral	-953.625779	Tree 0.03687	3.73557	0.0452
		2-PositiveSelection	-953.625847	Tree 0.03687	3.73560	0.0452
		3-discrete	-953.625927	Tree 0.03687	3.73559	0.0452
		7-beta	-953.625910	Tree 0.03688	3.73622	0.0452
		8-beta&w>1	-953.625910	Tree 0.03688	3.73617	0.0452
11.nuc.aln.alt	13	0-one-ratio	-1954.014492	Tree 0.04379	2.97969	0.1960
		1-NearlyNeutral	-1952.386556	Tree 0.04418	2.85606	0.1626
		2-PositiveSelection	-1951.227489	Tree 0.04791	3.07477	0.2497
		3-discrete	-1951.226495	Tree 0.04791	3.07480	0.2497
		7-beta	-1952.475510	Tree 0.04421	2.97542	0.2000
		8-beta&w>1	**-1951.226480**	Tree **0.04791**	**3.07465**	**0.2496**
2.nuc.aln.alt	14	0-one-ratio	-11266.772854	Tree 0.03502	3.79285	0.0362
		1-NearlyNeutral	-11263.331048	Tree 0.03538	3.75787	0.0352
		2-PositiveSelection	-11261.577987	Tree 0.03741	3.90223	0.0559
		3-discrete	**-11261.568622**	Tree **0.03740**	**3.90307**	**0.0559**
		7-beta	-11264.414783	Tree 0.03516	3.79807	0.0374
		8-beta&w>1	-11261.576110	Tree 0.03741	3.90297	0.0559
3.nuc.aln.alt	14	0-one-ratio	**-1177.947726**	Tree **0.05006**	**1.89013**	**0.1358**
		1-NearlyNeutral	-1177.947972	Tree 0.05006	1.89013	0.1358
		2-PositiveSelection	-1177.947985	Tree 0.05006	1.89012	0.1358
		3-discrete	-1177.947785	Tree 0.05006	1.89012	0.1358
		7-beta	-1177.950036	Tree 0.05006	1.89091	0.1360
		8-beta&w>1	-1177.950339	Tree 0.05006	1.89092	0.1360
4.nuc.aln.alt	14	0-one-ratio	-1831.975346	Tree 0.07342	3.90509	0.1036
		1-NearlyNeutral	-1830.605403	Tree 0.07413	3.91967	0.1080
		2-PositiveSelection	-1830.605396	Tree 0.07413	3.91963	0.1080
		3-discrete	**-1830.602716**	Tree **0.07409**	**3.90632**	**0.1065**
		7-beta	-1830.609170	Tree 0.07406	3.90130	0.1059

Phylogenetic Tree

Dataset: 1.nuc.aln.alt

Model: 0-one-ratio

Selected Sites: Bayes Empirical Bayes Analysis for 11.nuc.aln.alt, Model 8-beta&w>1

Model #8 1 w=0.07707 2 w=0.09209 3 w=0.10182 4 w=0.11003 5 w=0.11772 6 w=0.12545 7 w=0.13374 8 w=0.14332 9 w=0.15582 10 w=0.17807 11 w=29.39861

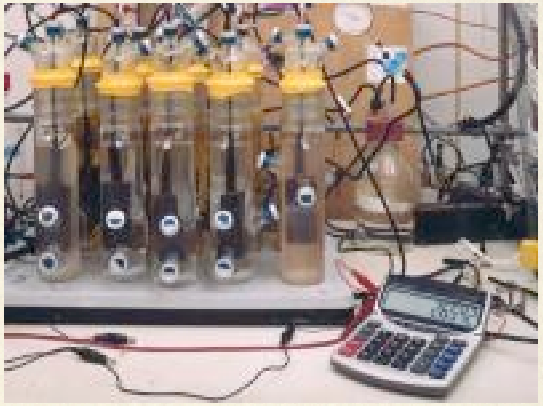
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
0.0
0.2
0.4
0.6
0.8
1.0
MSFTPGKQSSSSRASSGNRSGNGILKWADQSDQSRNVQTRGRRAQPKQTATSQQPSGGNVVPYYSWFSGITQFQKGKEFEFAEGQGVPIAPGVPATEAKGYWYRHNRRSFKTADGNQRQLLPRWYFYYLGTGPHAKDQYGTDIDGVFWVASNQADVNTPADILDR

Part II: Case Studies

Positive Selection in *Geobacter*

Positive Selection in *Geobacter*

DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA

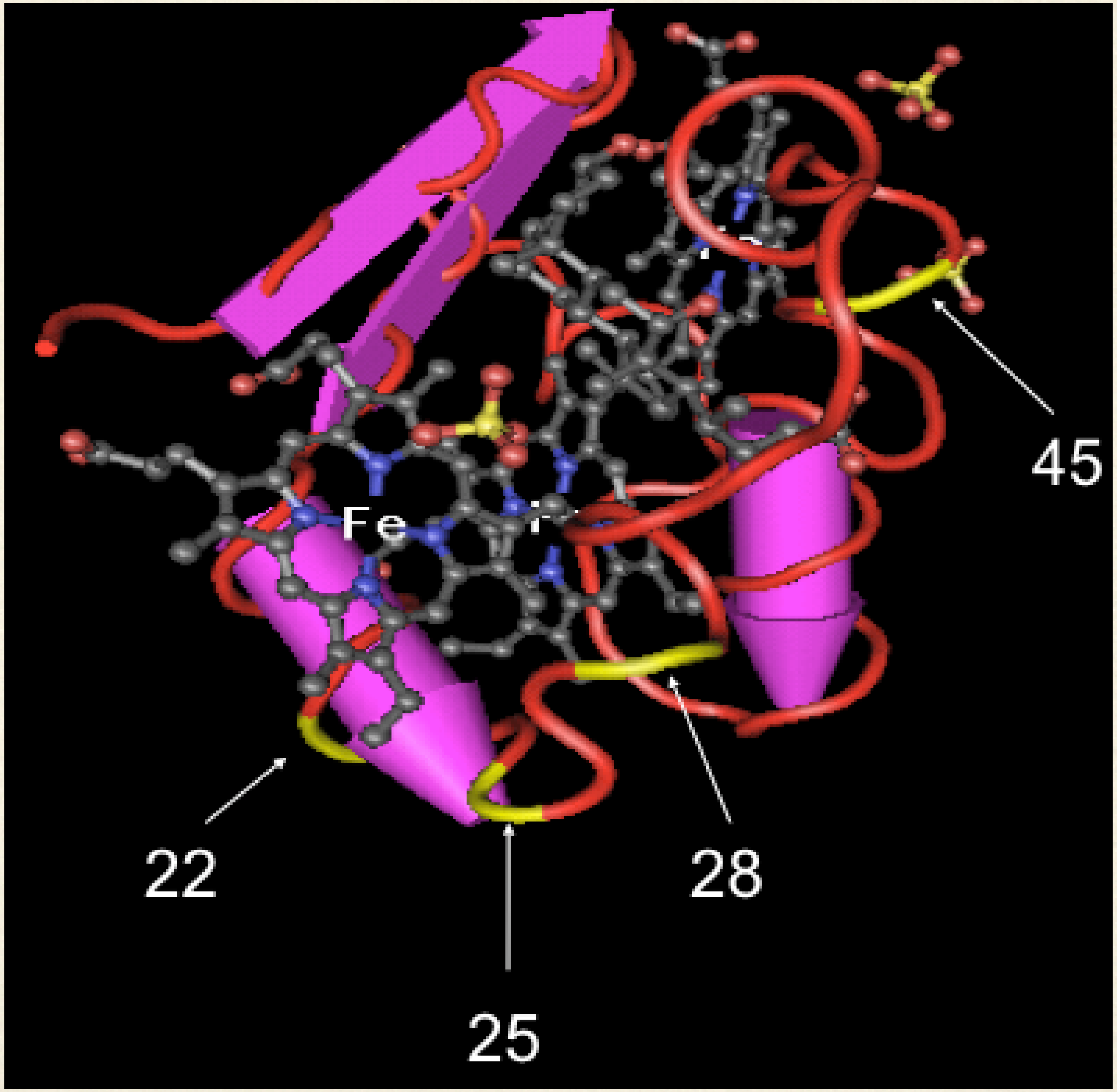


- *Geobacter* -- reduces metals such as iron and uranium -- useful to bioremediation (precipitates metals out of solution) & as biological fuel cells
- *Geobacter* genomes have many cytochromes; are the cytochromes from Geobacters that reduce different metals adapted to help that task?

Orthologous Clusters in *Geobacteraceae*

- Four complete *Geobacter* genomes + 6 other complete genomes from related species
- 26 orthologous clusters of cytochrome c genes that are found in at least 5 of the following 10 genomes: *Geobacter sulfurreducens*, *Geobacter* sp. FRC-32, *Geobacter metallireducens*, *Geobacter uraniumreducens*, *Pelobacter carbinolicus*, *Pelobacter propionicus*, *Desulfuromonas acetoxidans*, *Desulfovibrio desulfuricans*, *Desulfovibrio vulgaris*, *Rhodoferax ferrireducens.*
- Most clusters had no sites under positive selection
- One had 4 sites under positive selection - and structure available.

Positive Selected Sites Overlaid on Structure



Conclusions so far - *Geobacter*

- Most *Geobacter* cytochromes not specialized to help reduce specific metal of host.
- Aim to expand study to include all orthologous clusters in *Geobacter* and relatives -- not just cytochromes.

Reverse Gyrase Phylogeny

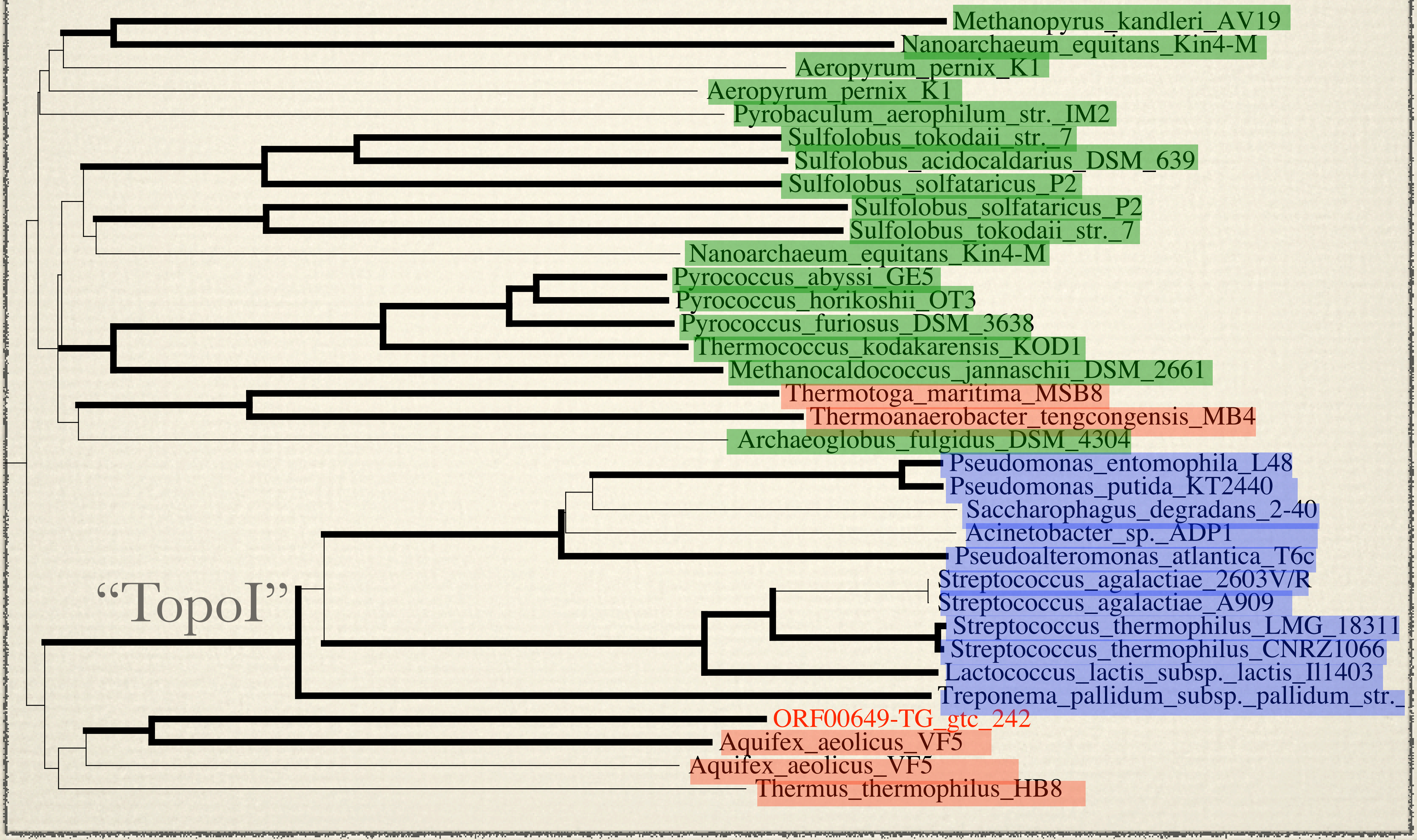
Thermodesulfobacterium commune

- Lives in hot springs in Yellowstone park.
- 70C optimal growth temperature
- Thought to be in own phylum by 16S, now looks like a deeply branching proteobacterium in light of genome
- Encodes a “reverse gyrase”

Reverse Gyrase

- Introduces positive supercoils into DNA
- Most prokaryotes abhor supercoils -- encode Topo I and gyrase to get rid of them.
- Reverse gyrase thought to prevent DNA from melting
- Found in most hyperthermophilic Archaea, some hyperthermophilic bacteria

Reverse Gyrase Phylogeny


Methanopyrus_kandleri_AV19
Nanoarchaeum_equitans_Kin4-M
Aeropyrum_pernix_K1
Aeropyrum_pernix_K1
Pyrobaculum_aerophilum_str._IM2
Sulfolobus_tokodaii_str._7
Sulfolobus_acidocaldarius_DSM_639
Sulfolobus_solfataricus_P2
Sulfolobus_solfataricus_P2
Sulfolobus_tokodaii_str._7
Nanoarchaeum_equitans_Kin4-M
Pyrococcus_abyssi_GE5
Pyrococcus_horikoshii_OT3
Pyrococcus_furiosus_DSM_3638
Thermococcus_kodakarensis_KOD1
Methanocaldococcus_jannaschii_DSM_2661
Thermotoga_maritima_MSB8
Thermoanaerobacter_tengcongensis_MB4
Archaeoglobus_fulgidus_DSM_4304
Pseudomonas_entomophila_L48
Pseudomonas_putida_KT2440
Saccharophagus_degradans_2-40
Acinetobacter_sp._ADP1
Pseudoalteromonas_atlantica_T6c
Streptococcus_agalactiae_2603V/R
Streptococcus_agalactiae_A909
Streptococcus_thermophilus_LMG_18311
Streptococcus_thermophilus_CNRZ1066
Lactococcus_lactis_subsp._lactis_Il1403
Treponema_pallidum_subsp._pallidum_str._
ORF00649-TG_gtc_242
Aquifex_aeolicus_VF5
Aquifex_aeolicus_VF5
Thermus_thermophilus_HB8
"TopoI"

Conclusions So Far - Reverse Gyrase

- Some non hyperthermophilic bacteria have a rgy homolog annotated as Topo I.
- Not closely related to other Topo I proteins
  - In fact, can’t even align them
- Perhaps not Topo I at all
  - Organisms all seem to have another protein annotated as Topo I that might be the real one.

Influenza Phylogeny

Influenza B

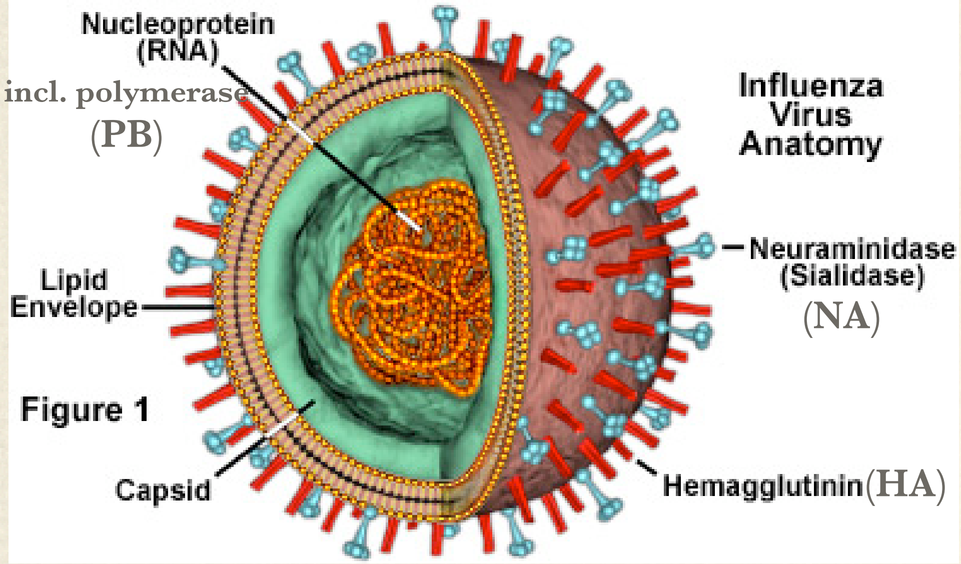


Figure by Michael Davidson, FSU

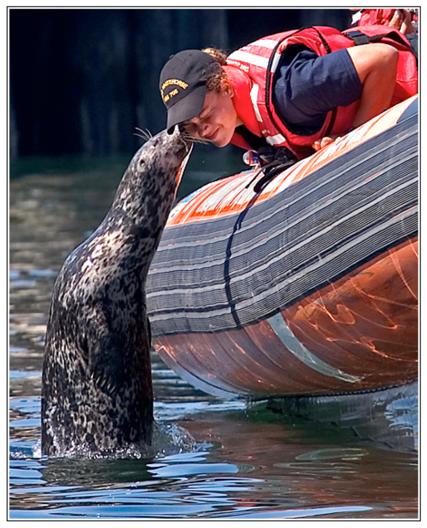
Genome in 8 pieces of -RNA

exterior glycoproteins, HA & NA targets for antivirals

Only known hosts: seals & humans

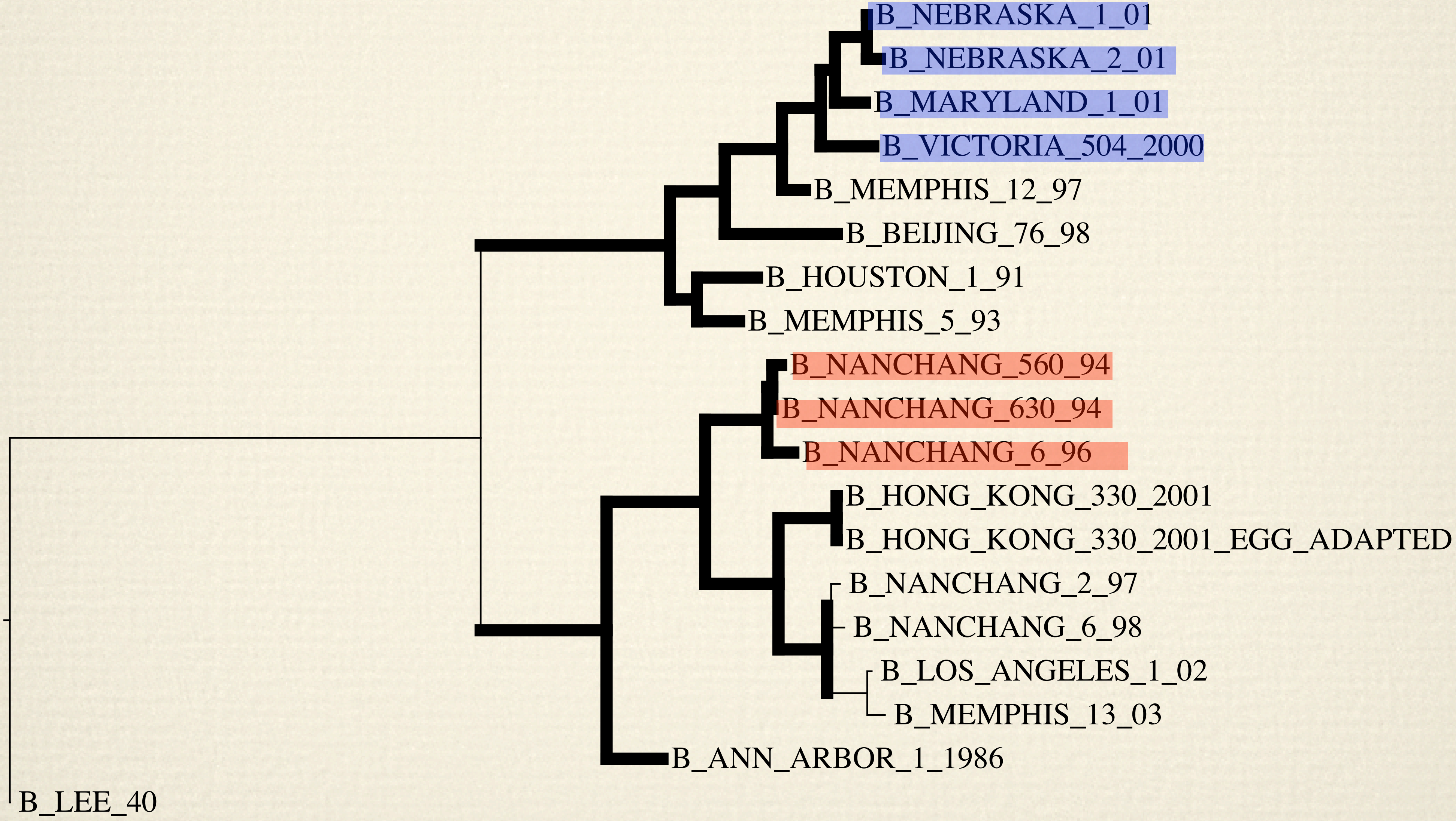
Normally only causes minor illness in heathy people

But, wide open for study, much less studied than Influenza A

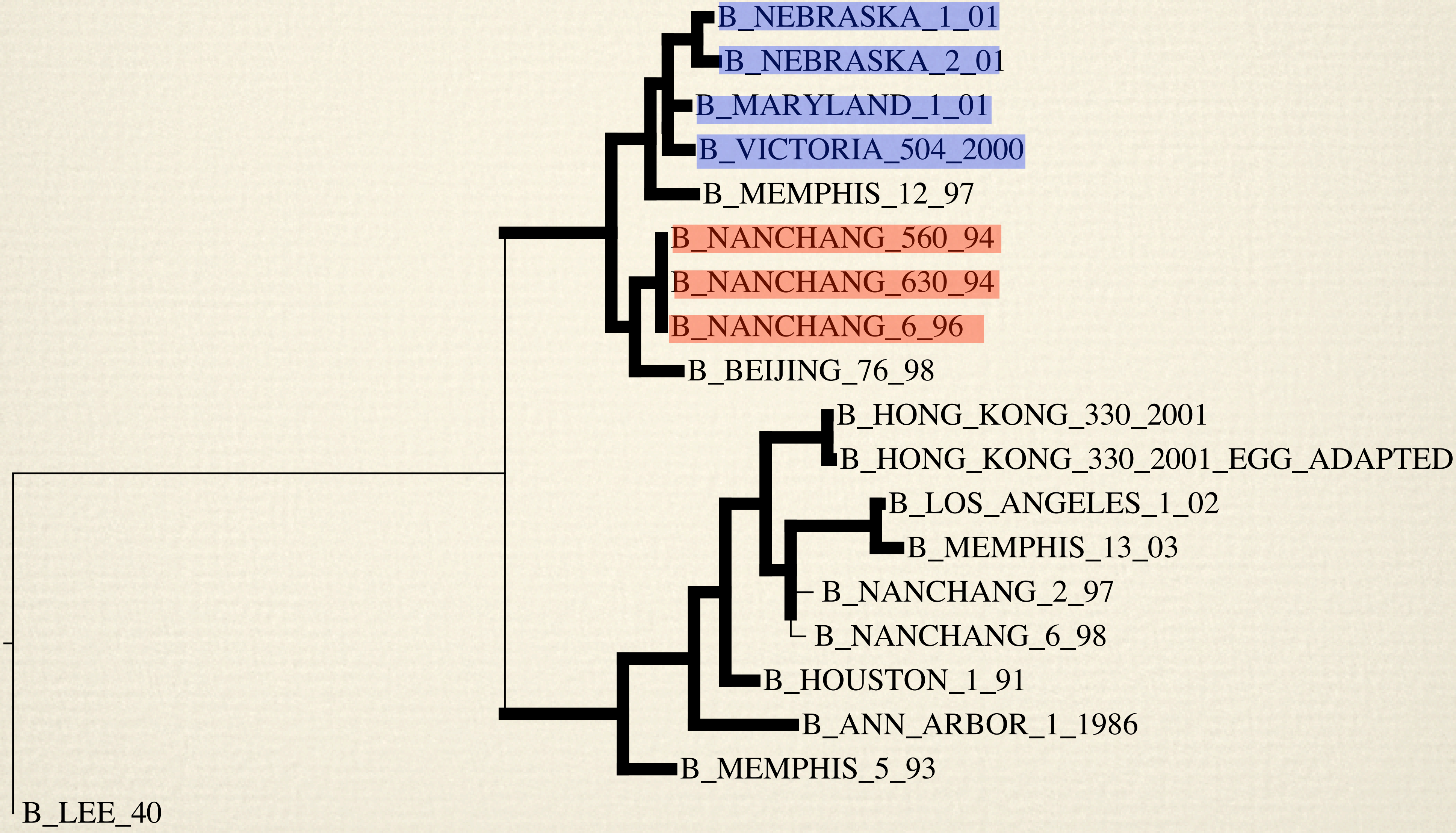


Influenza B vector infecting unidentified human

Influenza B - HA

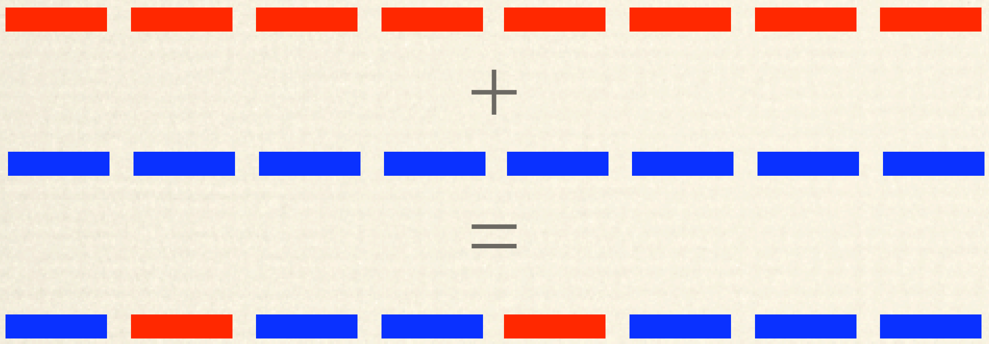


Influenza B - PB2



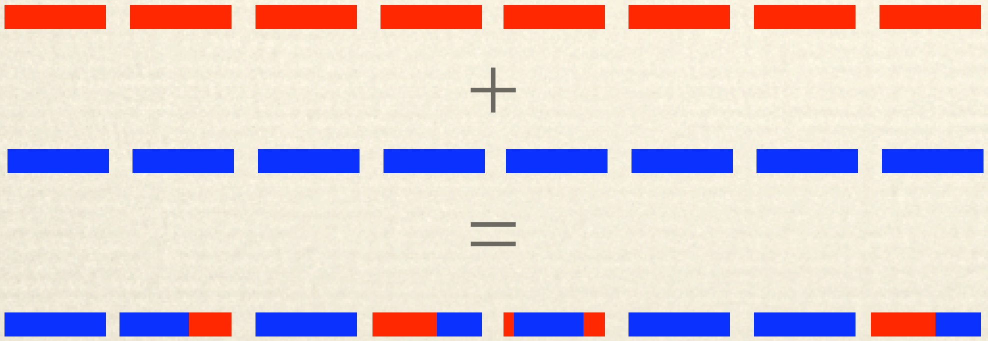
Reassortment or Recombination?

Given co-infection by two different strains:



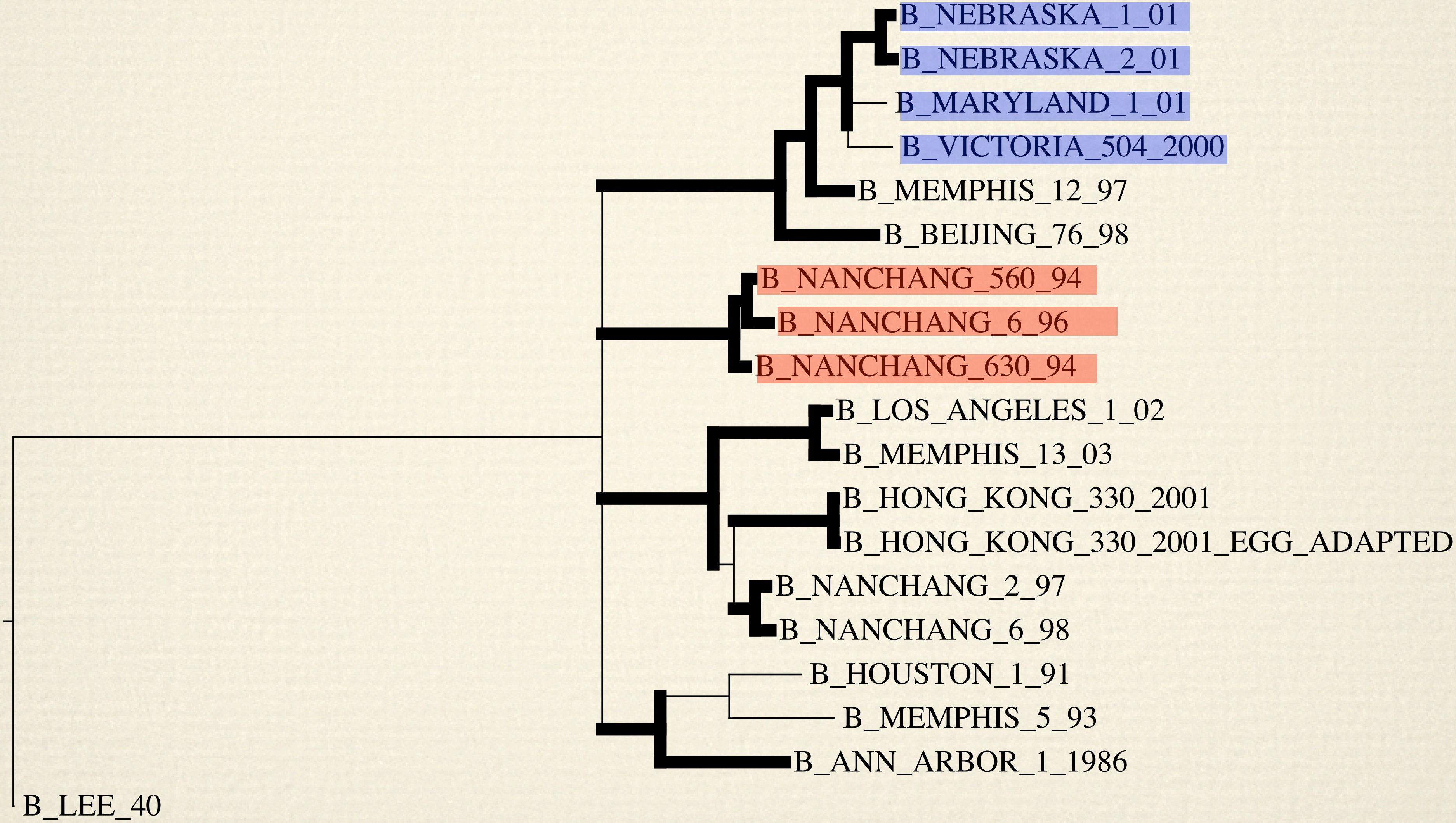
Reassortment

or



Recombination

Influenza B - Concatenated



Conclusions on Influenza B

- Trees compatible with either recombination or reassortment.
- Based on previous co-infection studies, we are leaning towards reassortment.
- Phylogenies of partial sequences may help resolve question

Future Directions

- Setting up APIS server for all prokaryotic genomes
- Making APIS better for eukaryotes -- breaking up multi-domain proteins
- Grid Support for APIS & ECFinder
- Better visualizations

Acknowledgments

Funding

DOE GTL

PIs

Naomi Ward
Jonathan Eisen
David Spiro
Barbara Methé
Joana Silva

ComboDB

Martin Wu

Users

Naomi Ward
Amber Hartman
Garry Myers
David Spiro
Naomi Sengamalay
Michael Montague
Derrick Fouts
Lis Caler
Brian Haas
Sean Daugherty
Kevin Penn
Amy Egan
Joana Silva