More Questions Than Answers:

Comparative genomics, evolutionary analysis, and knowing what we do not know about DNA repair.

Jonathan A. Eisen

September 20, 2005

TIGR

More Questions Outline

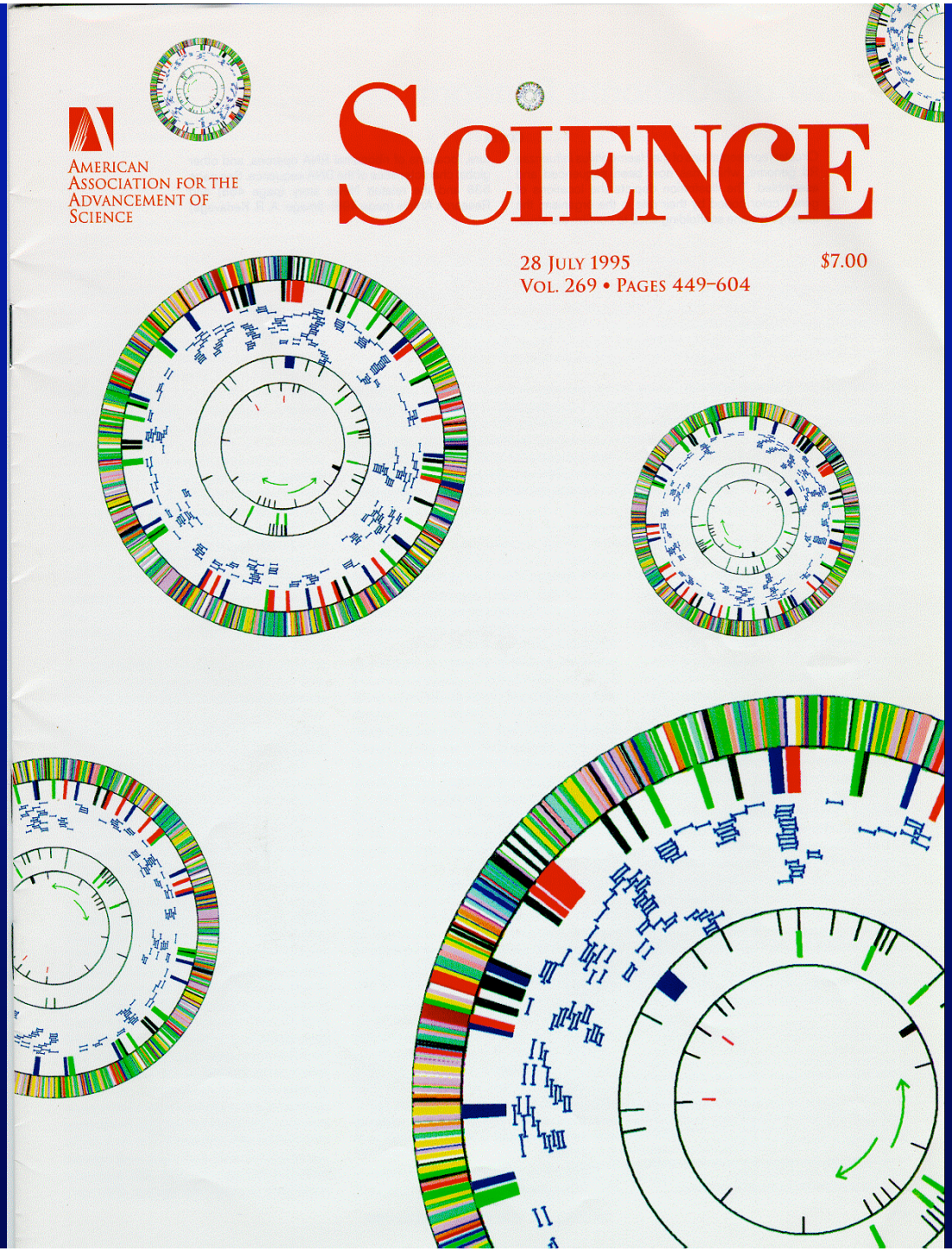
- Background
- More questions
  - I. Case study - *H. pylori*
  - II. New subfamilies and domain patterns
  - III. Recent duplications
  - IV. Missing pathways
  - V. Unusual patterns
  - VI. Even more questions
- Answers
  - Better informatics
  - Experiments
- The end

TIGR

Background

Genome sequencing and evolutionary analysis

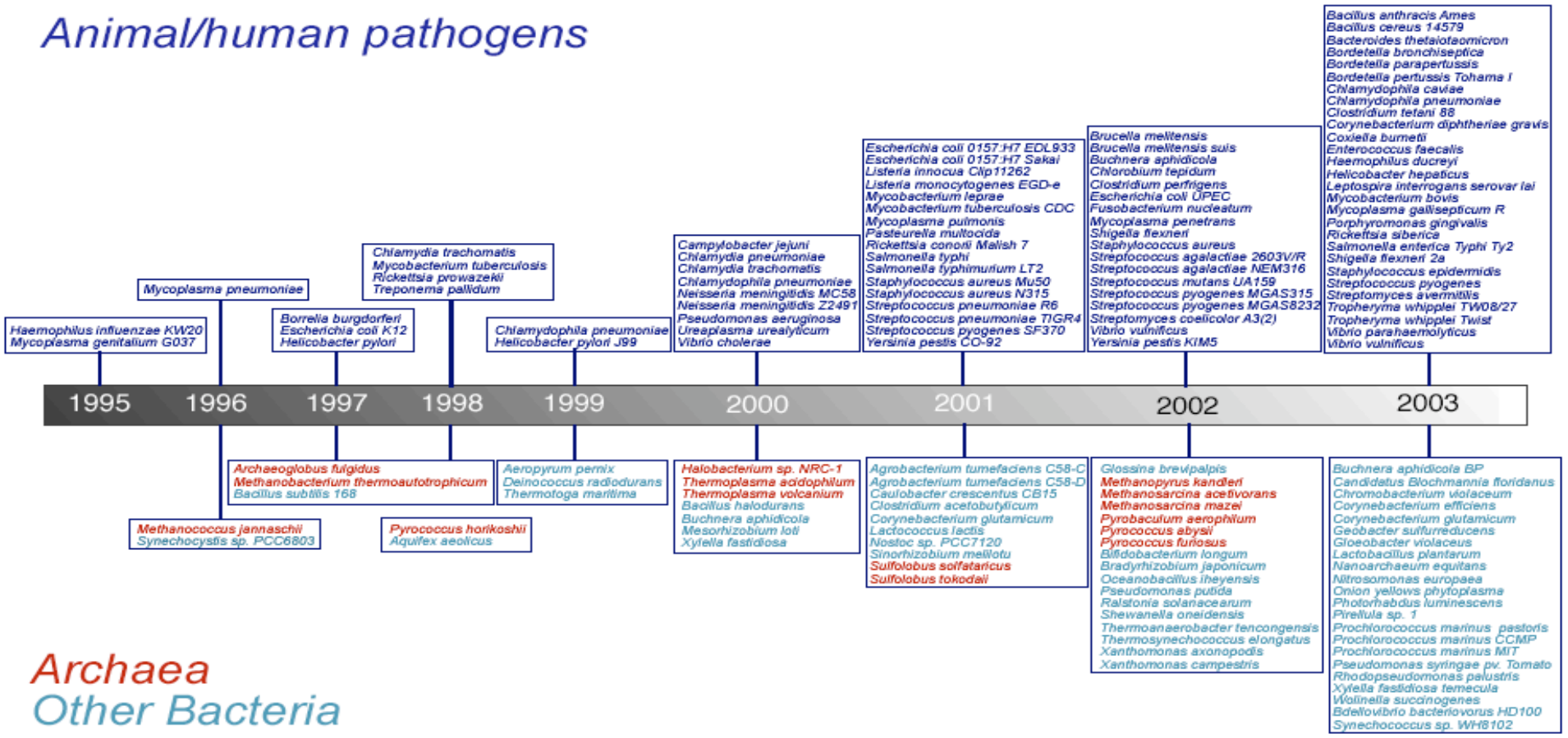
TIGR

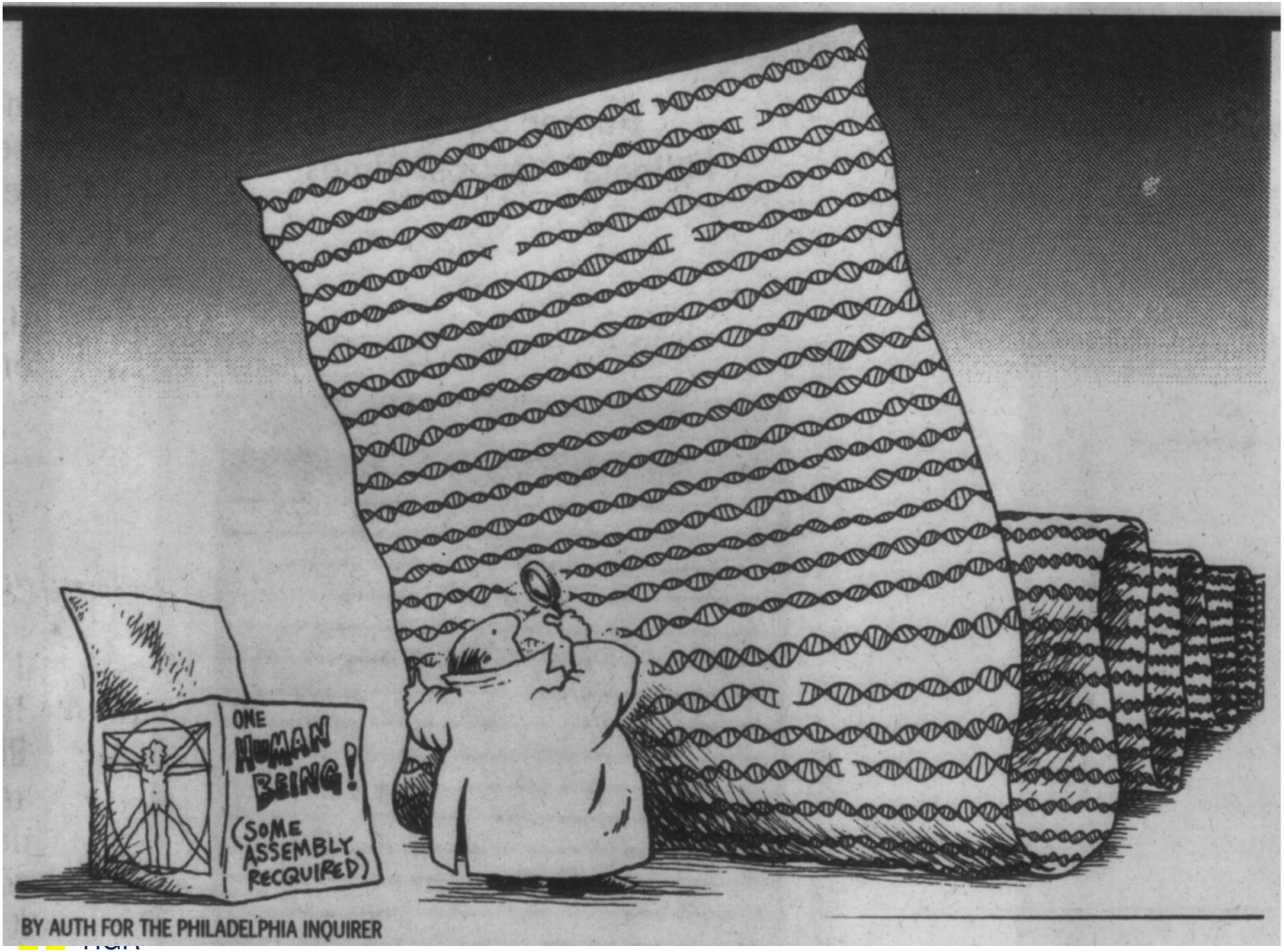


Fleischmann et al. 1995

TIGR

Prokaryotic Genomes




ONE
HUMAN
BEING!
(SOME
ASSEMBLY
RECQUIRED)
BY AUTH FOR THE PHILADELPHIA INQUIRER

Why Completeness is Important

- Improves characterization of genome features
- Better comparative genomics
- Presence/absence is less subjective
- Missing sequence might be important (e.g., centromere)
- Allows researchers to focus on biology not sequencing
- Facilitates large scale correlation studies
- Controls for contamination

see Fraser et al. (2002). J. Bacteriol. 184: 6403-5

Analysis of Complete Genomes

- Identification/prediction of genes
- Characterization of gene features
- Characterization of genome features
- Prediction of gene function
- Prediction of pathways
- Integration with known biological data
- Comparative genomics

“Nothing in biology makes sense except in the light of evolution.”

T. H. Dobzhansky (1973)

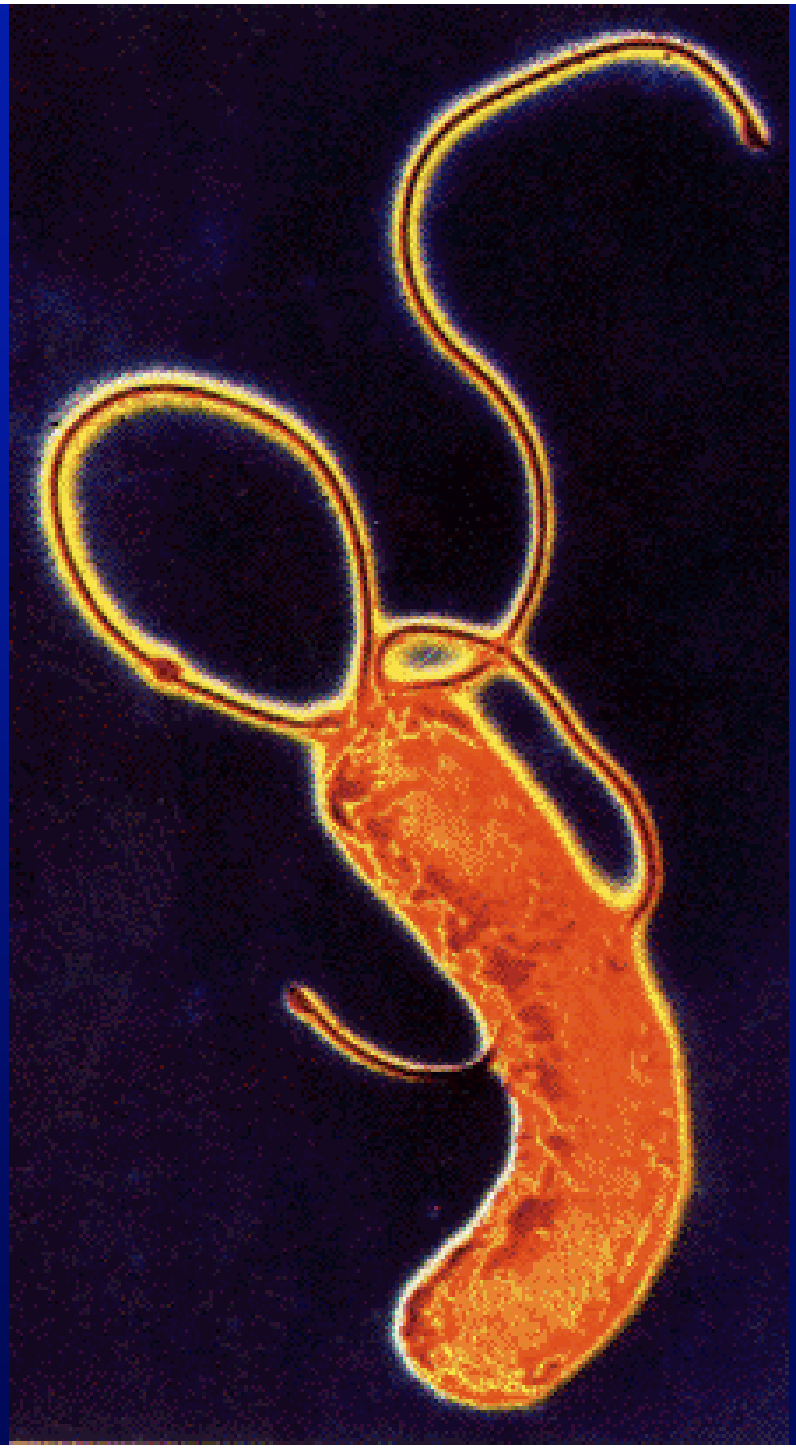
Evolutionary Perspective and Comparative Biology

- Comparative biology is the analysis of differences and similarities between species.
- An evolutionary perspective is useful in such studies because this allows one to focus not just on the levels and degrees of similarity or difference but on how and why similarities and differences came to be.

More Questions I: A Case Study

The Unbearable Lightness of *Helicobacter pylori's* DNA Repair Gene List

TIGR



TIGR

articles

The complete genome sequence of the gastric pathogen *Helicobacter pylori*

Jean-F. Tomb*, Owen White*, Anthony R. Kerlavage*, Rebecca A. Clayton*, Granger G. Sutton*, Robert D. Fleischmann*, Karen A. Ketchum*, Hans Peter Klenk*, Steven Gill*, Brian A. Dougherty*, Karen Nelson*, John Quackenbush*, Lixin Zhou*, Ewen F. Kirkness*, Scott Peterson*, Brendan Loftus*, Delwood Richardson*, Robert Dodson*, Hanif G. Khalak*, Anna Glodek*, Keith McKenney*, Lisa M. Fitzegerald*, Norman Lee*, Mark D. Adams*, Erin K. Hickey*, Douglas E. Berg†, Jeanine D. Gocayne*, Teresa R. Utterback*, Jeremy D. Peterson*, Jenny M. Kelley*, Matthew D. Cotton*, Janice M. Weidman*, Claire Fujii*, Cheryl Bowman*, Larry Watthey*, Erik Wallin‡, William S. Hayes§, Mark Borodovsky§, Peter D. Karp‖, Hamilton O. Smith¶, Claire M. Fraser* & J. Craig Venter*

* *The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
† *Department of Molecular Biology, School of Medicine, Washington University St Louis, 660 S. Euclid Avenue, St Louis, Missouri 63110, USA*
‡ *Department of Biochemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden*
§ *School of Biology, Georgia Tech, Atlanta, Georgia 30332, USA*
‖ *SRI International, Artificial Intelligence Center, 333 Ravenswood Avenue, Menlo Park, California 94025, USA*
¶ *Department of Molecular Biology and Genetics, School of Medicine, Johns Hopkins University, 725 N. Wolfe Street, Baltimore, Maryland 21205, USA*

to be absent. The ability of *H. pylori* to perform mismatch repair is suggested by the presence of methyl transferases, mutS and uvrD. However, orthologues of MutH and MutL were not identified. Components of an SOS system also appear to be absent.

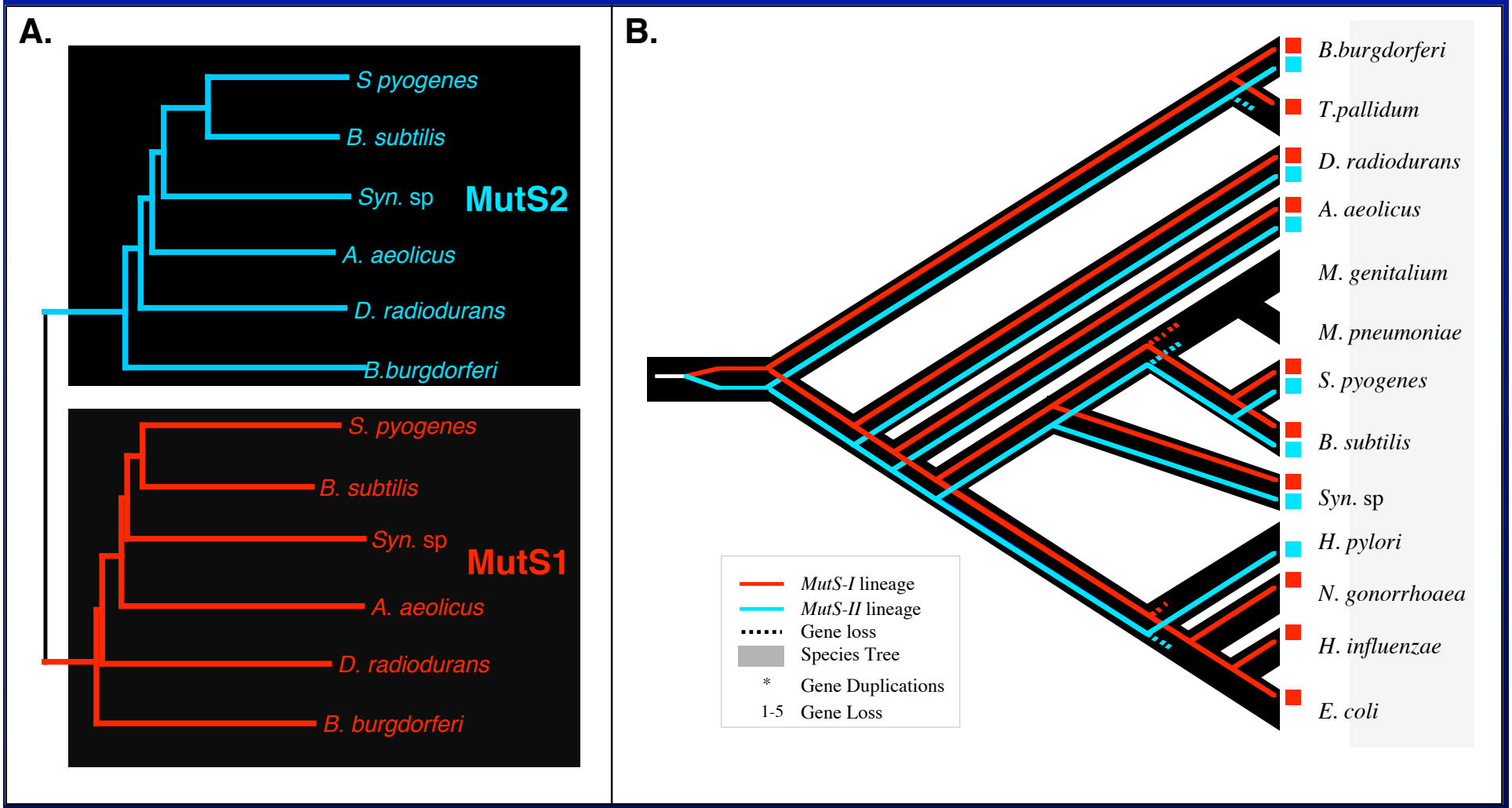
Tomb et al. 1997

H. pylori and MutS

- *H. pylori* has a MutS-like gene but not MutL
- Prior to this genome, all species that encoded a MutS homolog also encoded a MutL homolog
- Experimental studies have shown MutS and MutL always work together in mismatch repair
- Problem: what do we conclude about *H. pylori* mismatch repair

See Eisen et al, 1997

Two Prokaryotic MutS Subfamilies



Eisen, 1998, NAR 26: 4291-4300.

H. pylori No MMR

Helicobacter pylori genetic diversity and risk of human disease

Martin J. Blaser[1] and Douglas E. Berg[2]

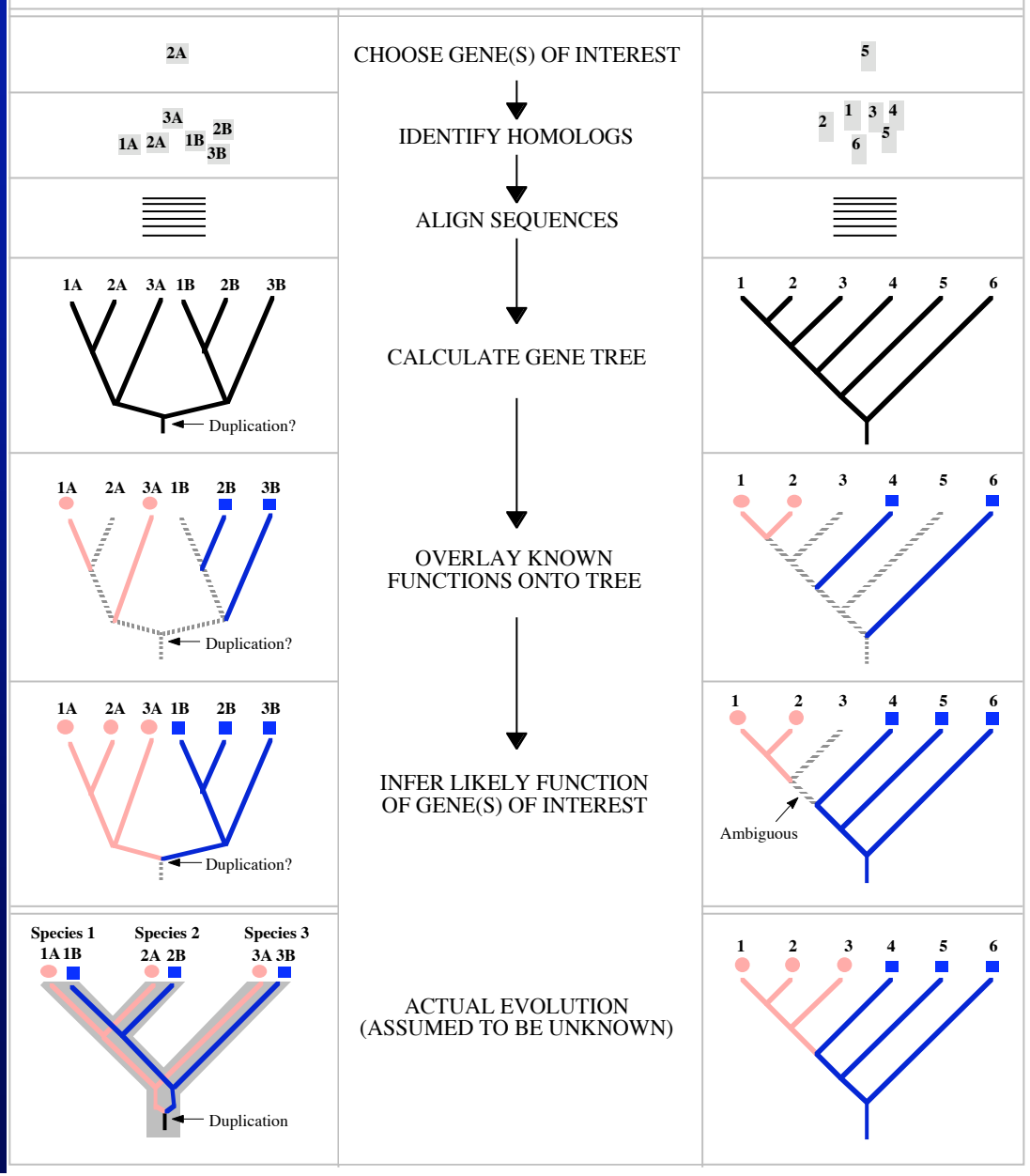
Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*

Britta Björkholm*†‡, Maria Sjölund‡§, Per G. Falk¶, Otto G. Berg‖, Lars Engstrand*§, and Dan I. Andersson*,**

Among the several factors that affect the appearance and spread of acquired antibiotic resistance, the mutation frequency and the biological cost of resistance are of special importance. Measurements of the mutation frequency to rifampicin resistance in *Helicobacter pylori* strains isolated from dyspeptic patients showed that ≈1/4 of the isolates had higher mutation frequencies than *Enterobacteriaceae* mismatch-repair defective mutants. This high mutation frequency could explain why resistance is so frequently acquired during antibiotic treatment of *H. pylori* infections. Inactivation of the *mutS* gene had no substantial effect on the mutation frequency, suggesting that MutS-dependent mismatch repair is absent in this bacterium. Furthermore, clarithromycin resistance conferred a biological cost, as measured by a decreased competitive ability of the resistant mutants in mice. In clinical isolates this cost could be reduced, indicating that compensation is a clinically relevant phenomenon that could act to stabilize resistant bacteria in a population.

TIGR

PHYLOGENENETIC PREDICTION OF GENE FUNCTION



Eisen, 1998, Genome Res. 8: 163-167.

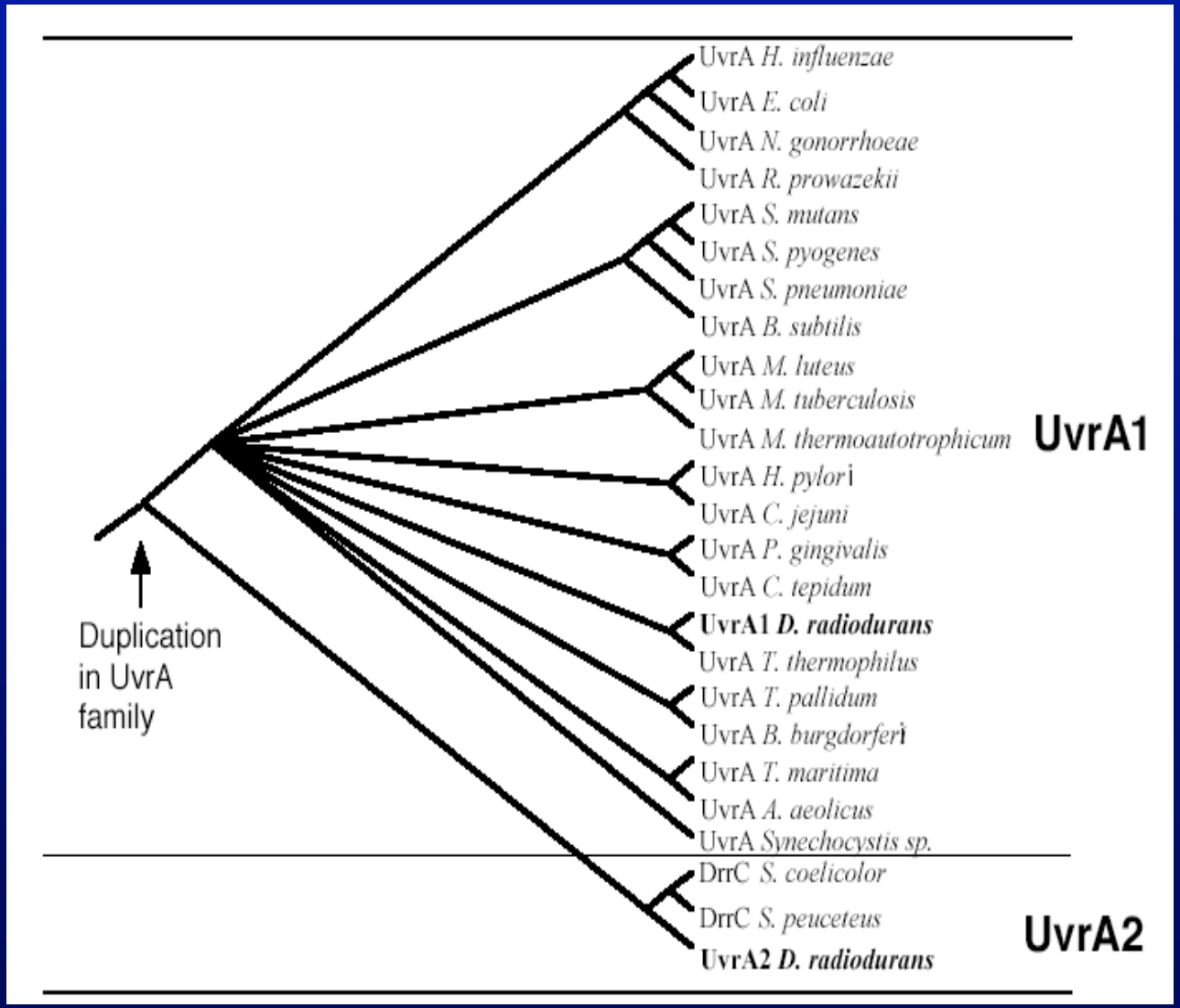
TIGR

More Questions II:

Long Lost Cousins

TIGR

Two UvrA Subfamilies

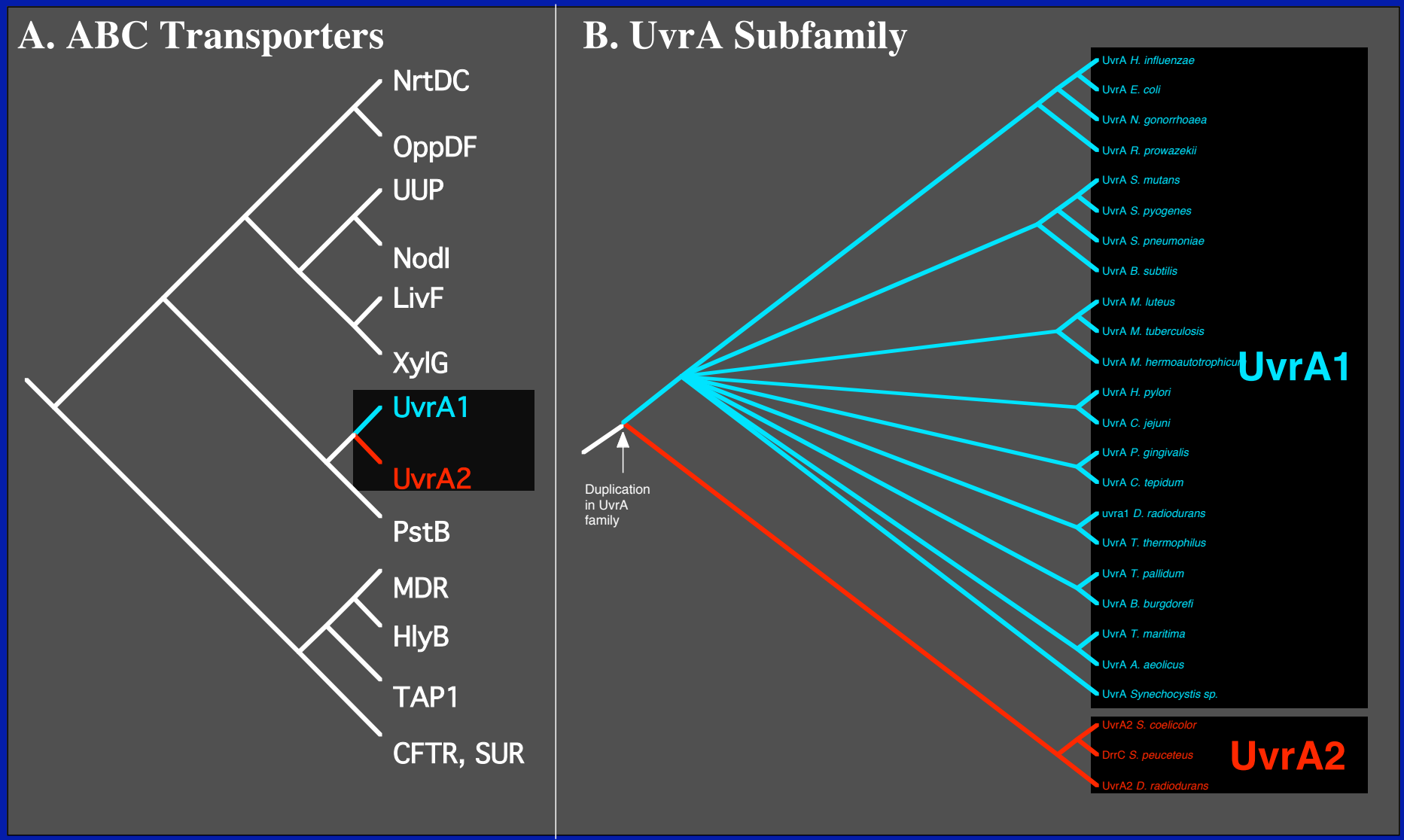


- Some species have a second UvrA-like gene
- Some results on UvrA2 in *D. radiodurans*
- Unclear what UvrA2s do

TIGR

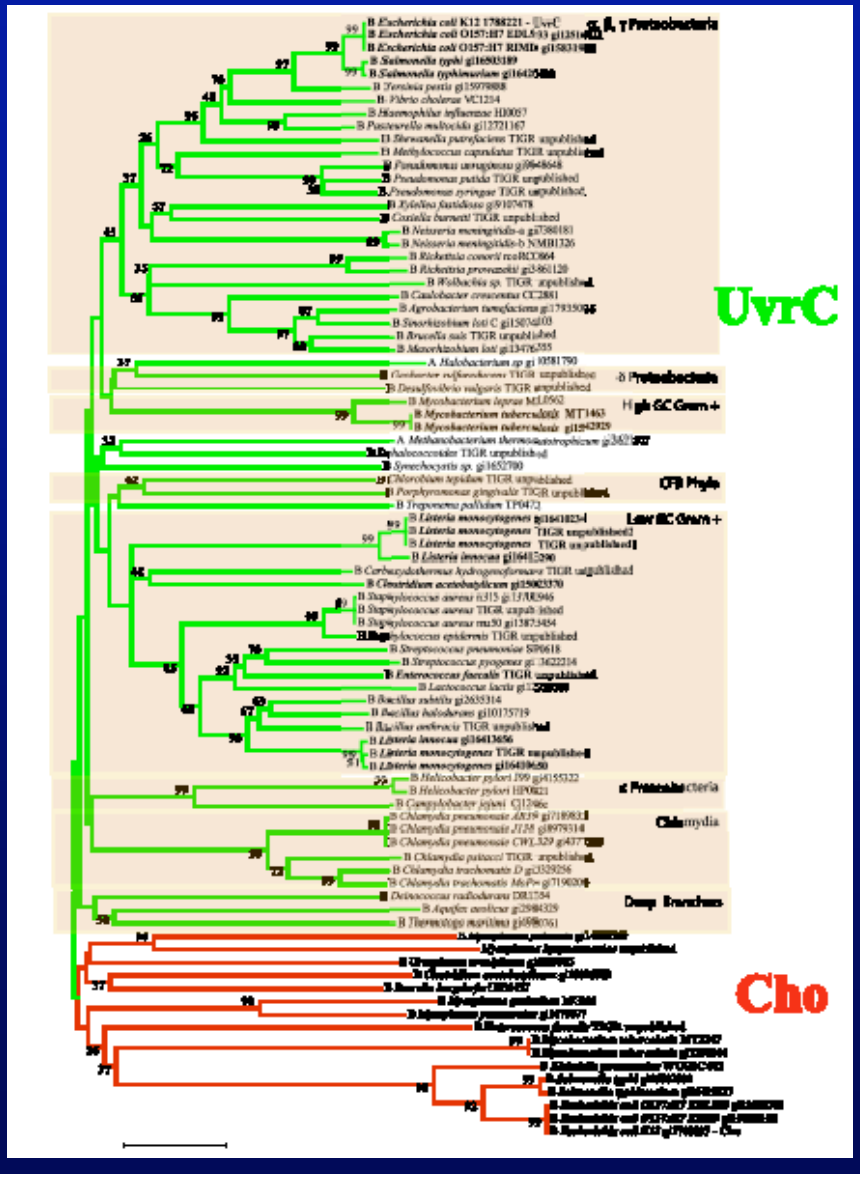
Eisen and Wu, 2002, Theor. Pop. Biol. 61: 481-487.

Evolution of UvrA Family



TIGR

UvrC vs. Cho

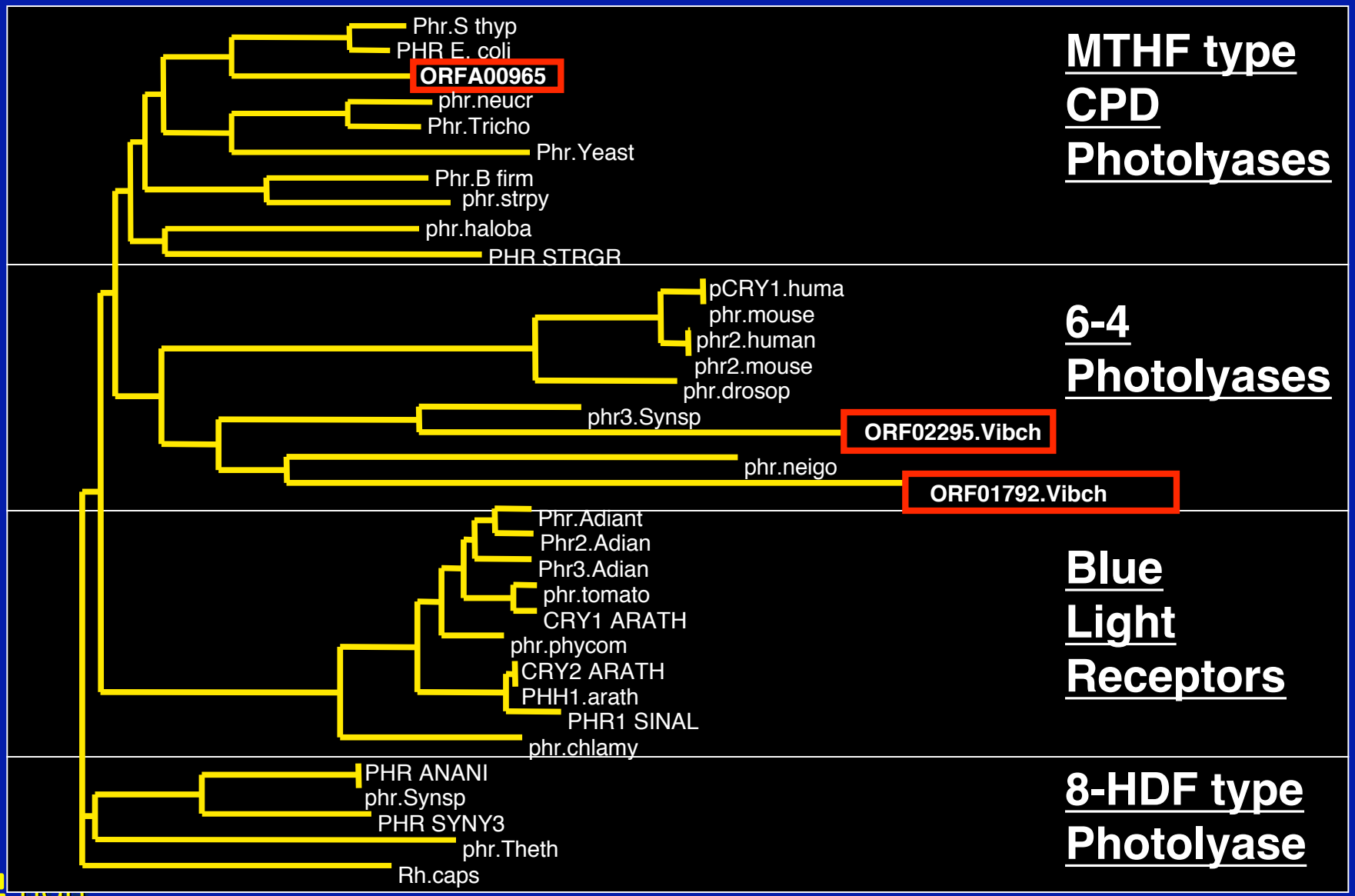


- Cho in *E. coli* is involved in an alternative NER pathway (see Moolenar et al. 2002)
- What do the other relatives of Cho do?

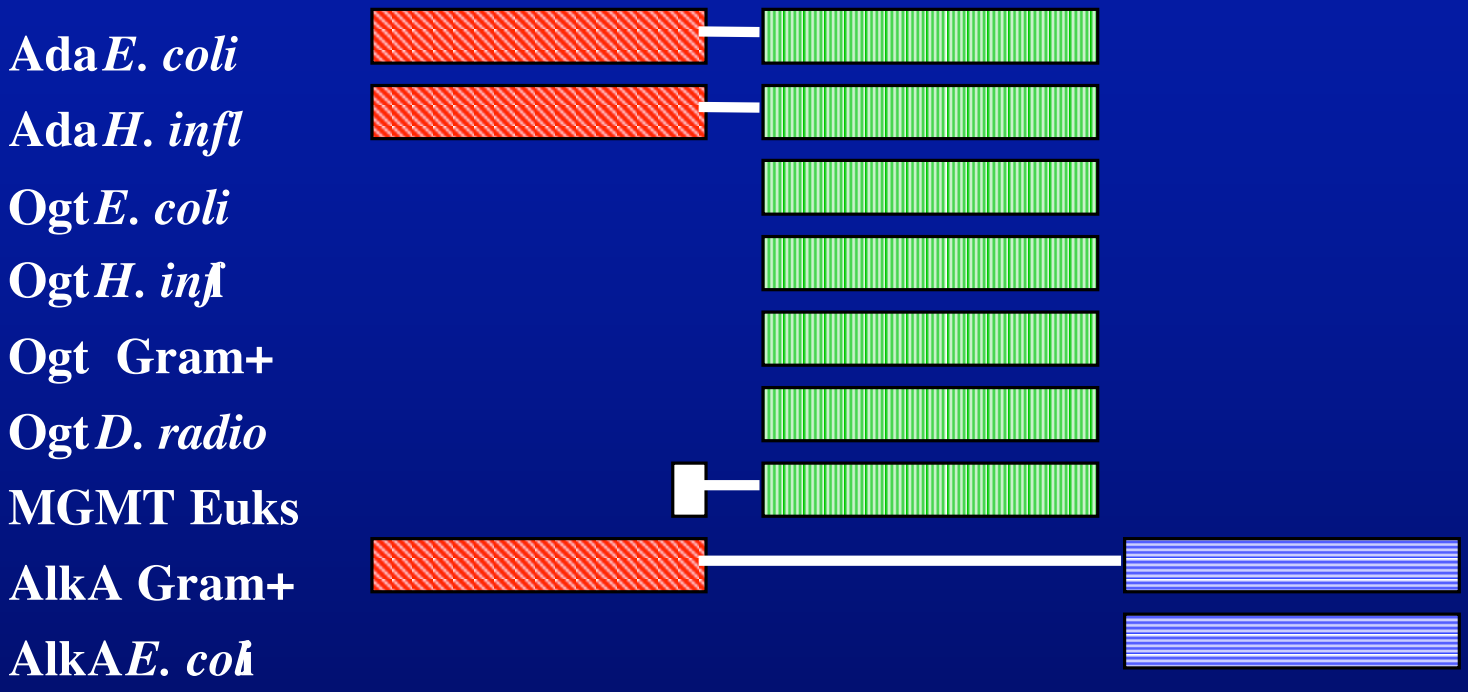
Van Houten et al, 2002, PNAS 99: 2581-2583.

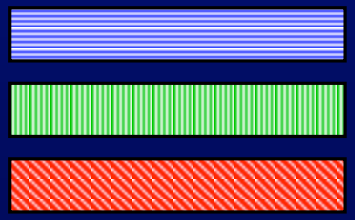
TIGR

Three *V. cholerae* Phr Homologs



Alkylation Repair Genes





AlkA Domain (O6-Me-G glycosylase)

Ogt Domain (O6-Me-G alkyltransferase)

Ada Domain (transcriptions regulator)

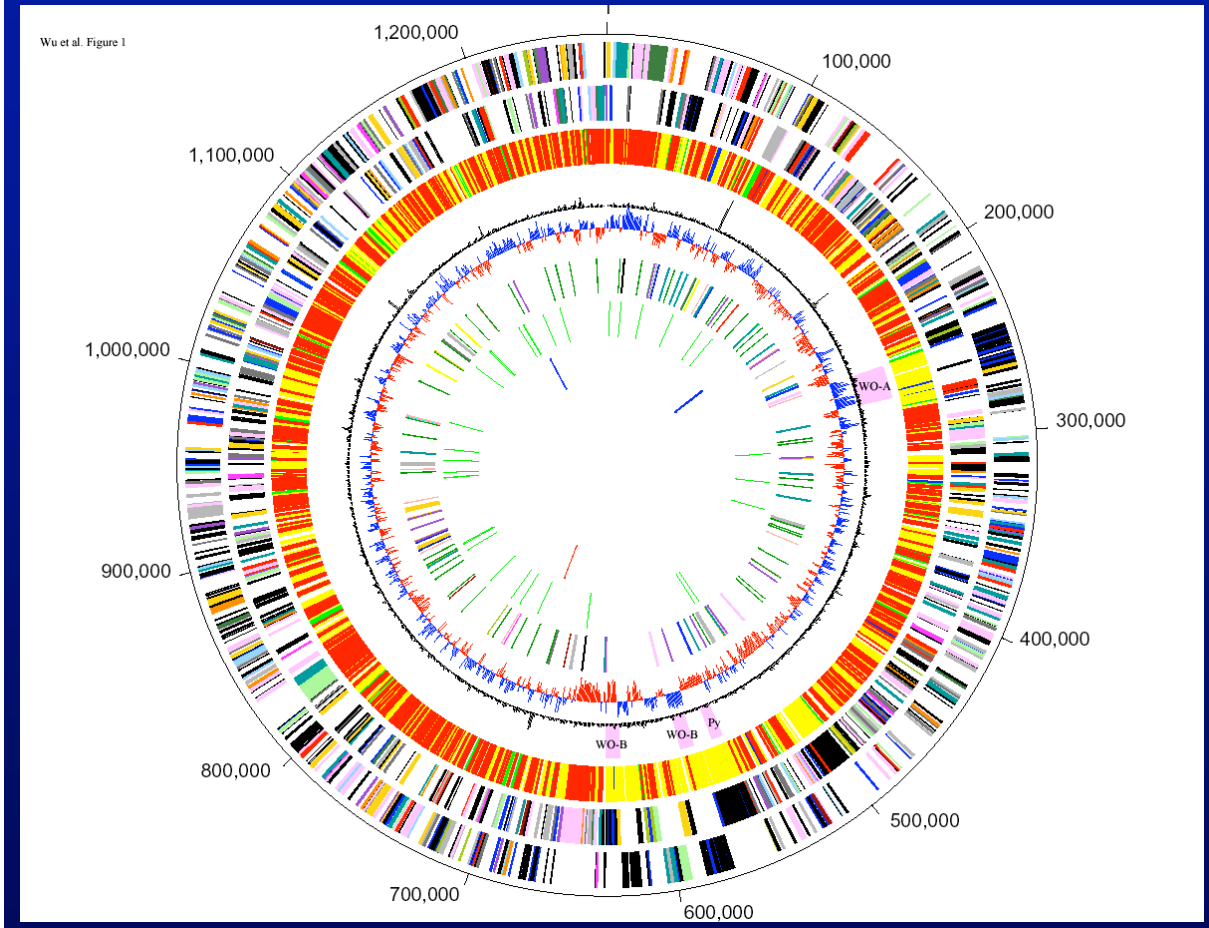
TIGR

Eisen and Hanawalt, 1999, Mut. Res. 435: 171-213.

More Questions III:

Does 1+1 = 1? Or 2?

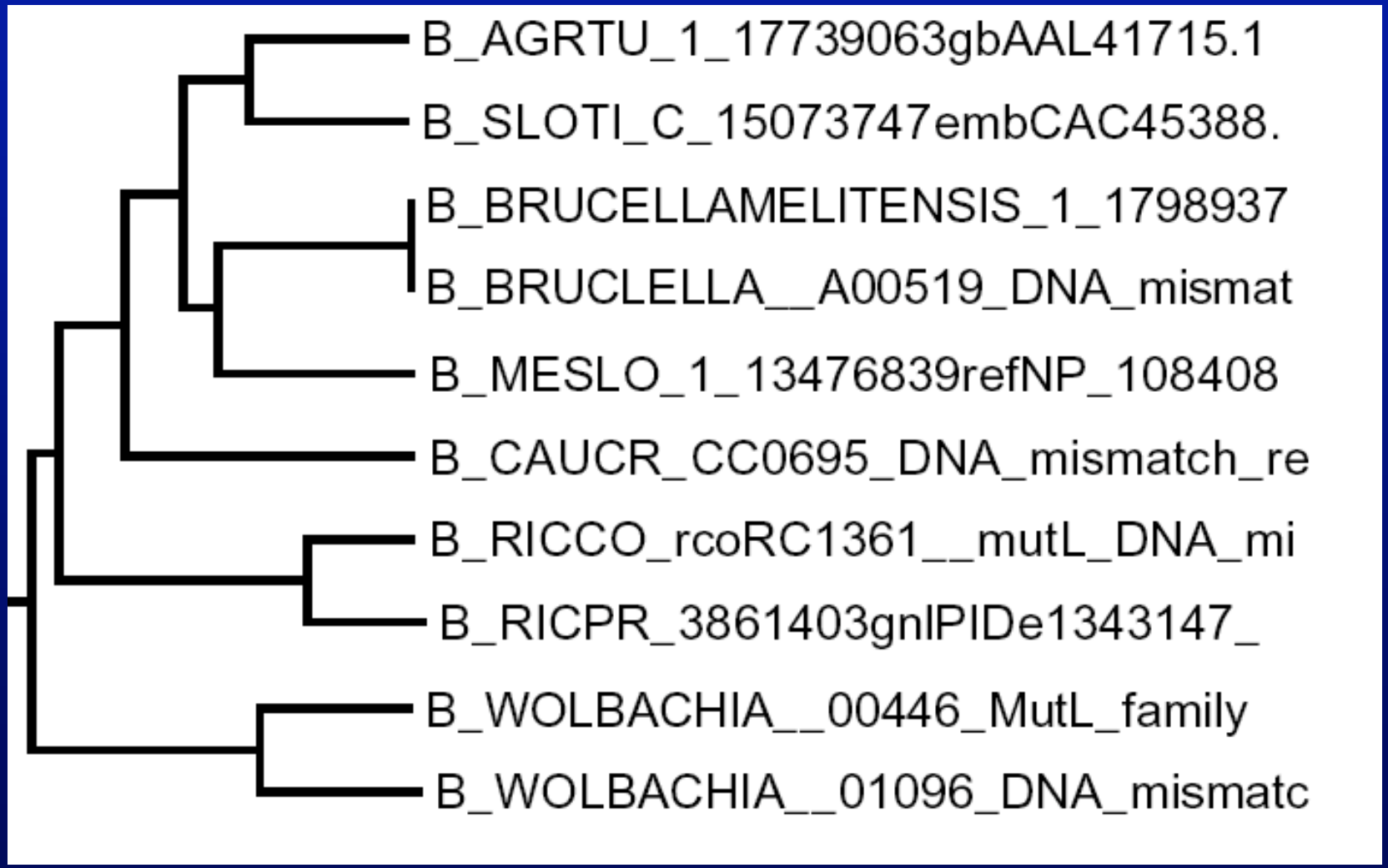
Wolbachia pipientis wMel



TIGR

Wu et al, 2004, PLOS Biology 2: 327-341

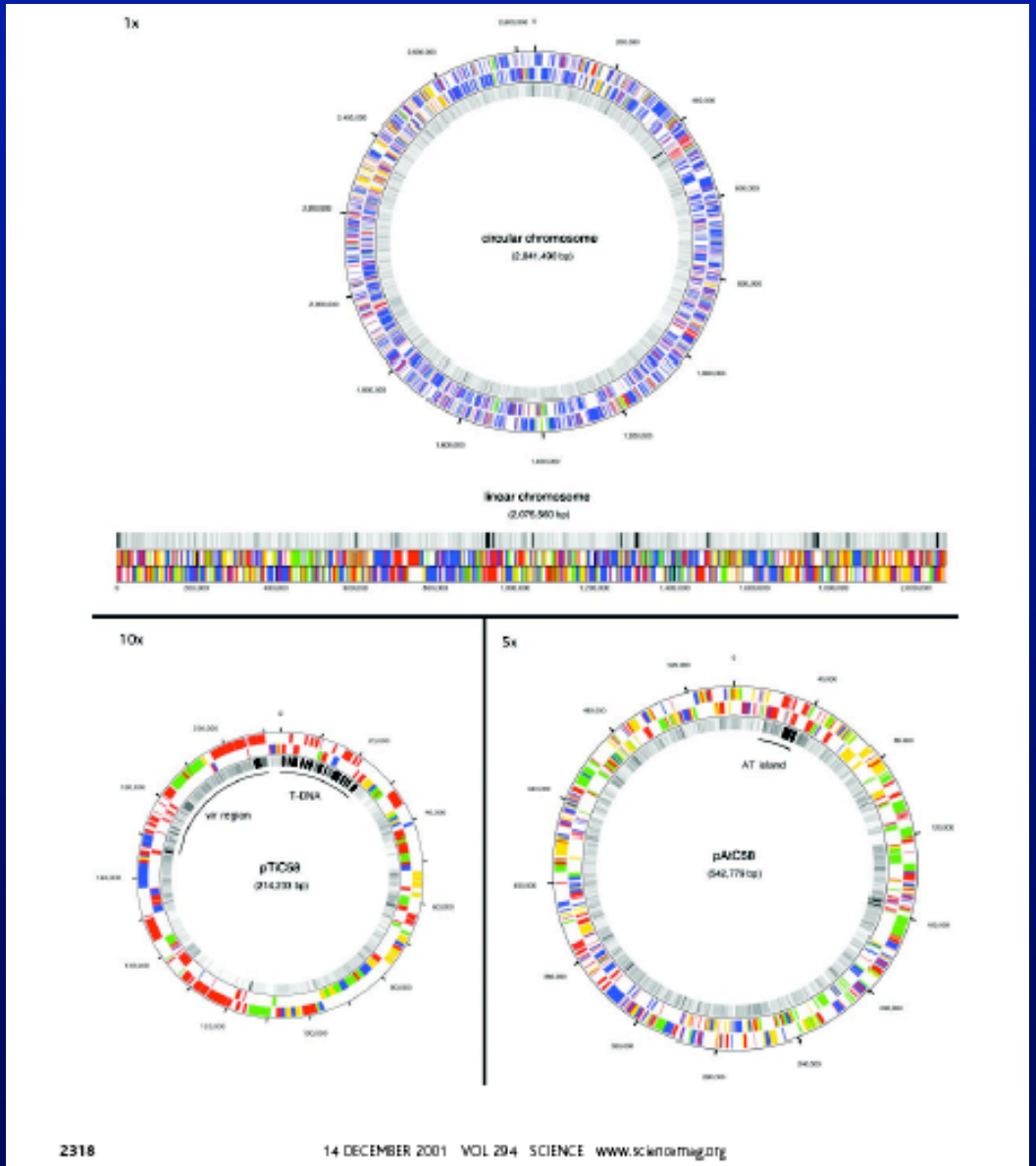
MutL Duplication in *Wolbachia*



TIGR

Eisen, unpublished.

Agrobacterium tumefaciens



Wood et al., 2001, Science 294: 2317-2323

TIGR

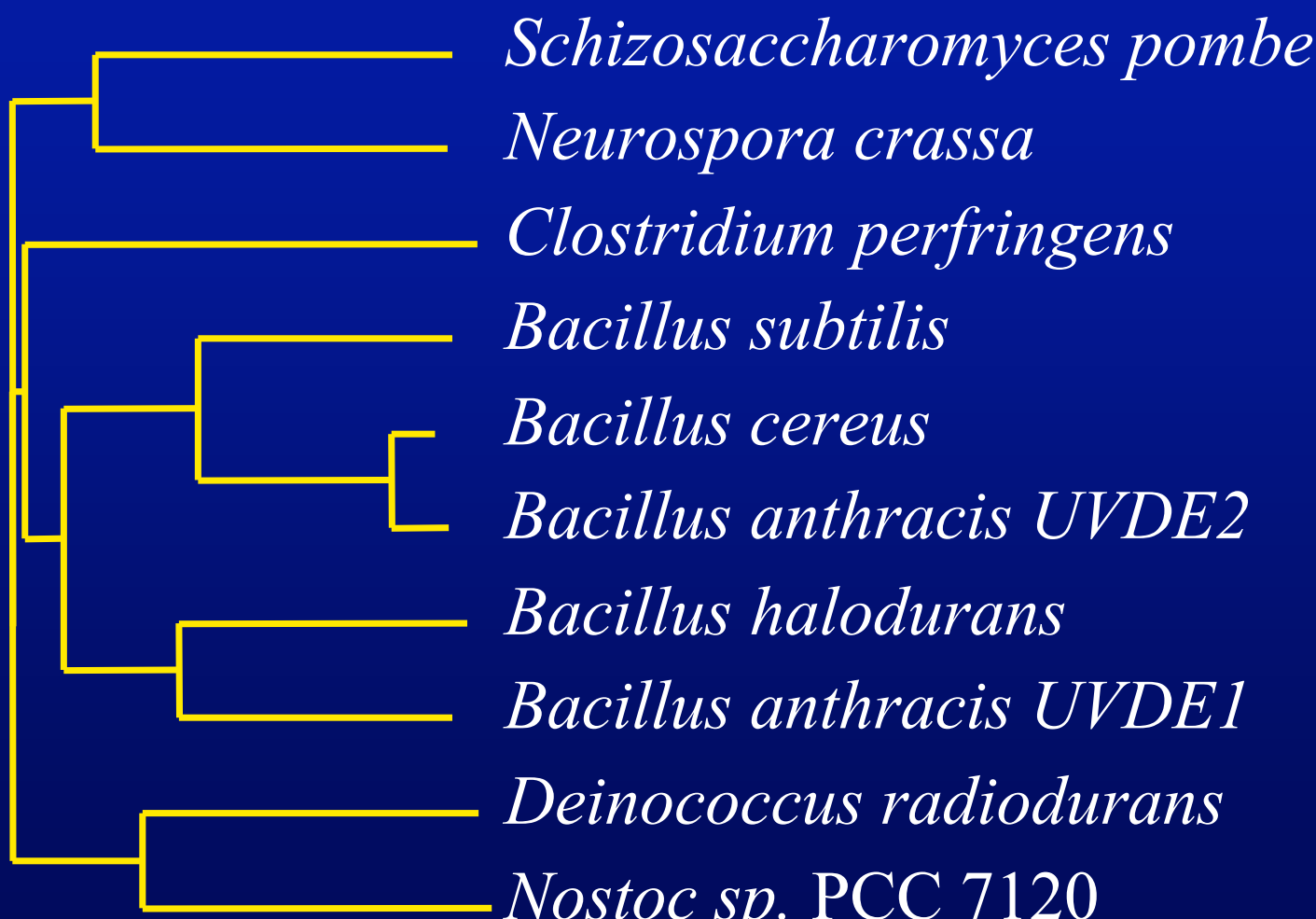
Agrobacterium tumefaciens has multiple duplications

Chromosome1	DnaE, LigA, Xth
Chromosome2	DnaE, LigA
Plasmid1	DnaE, LigA, Xth
Plasmid2	DnaE?, LigA

TIGR

Wood et al., 2001, Science 294: 2317-2323

Duplication of UVDE in Anthrax



TIGR

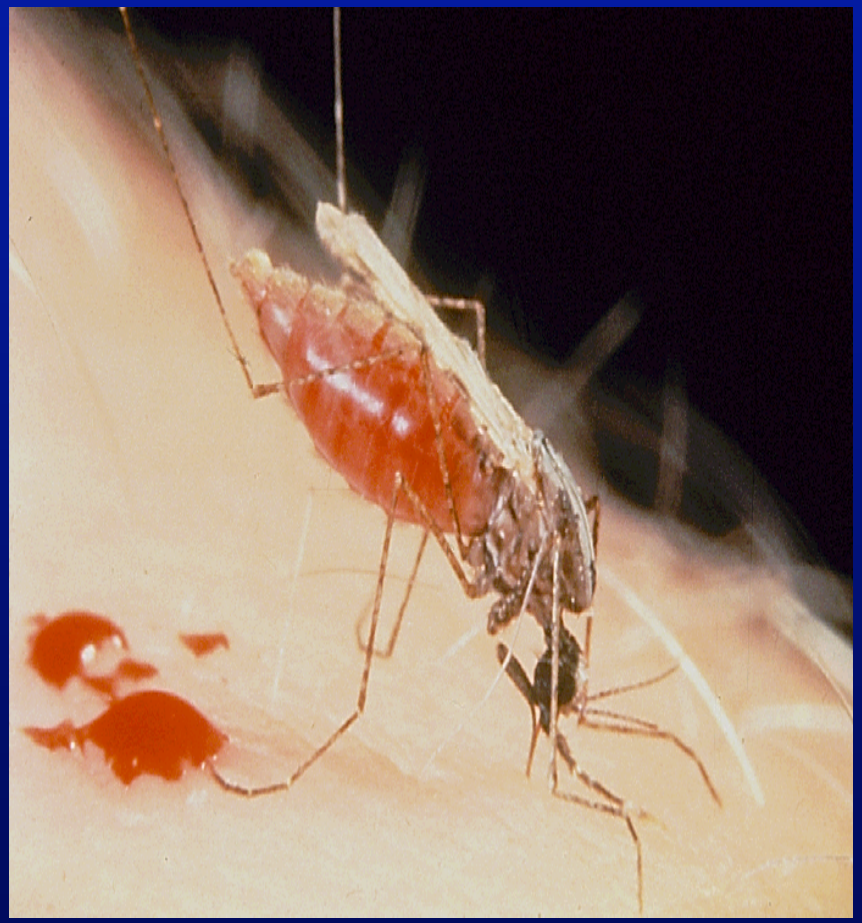
Eisen,unpublished.

More Questions IV:

When a good repair pathway
is hard to find

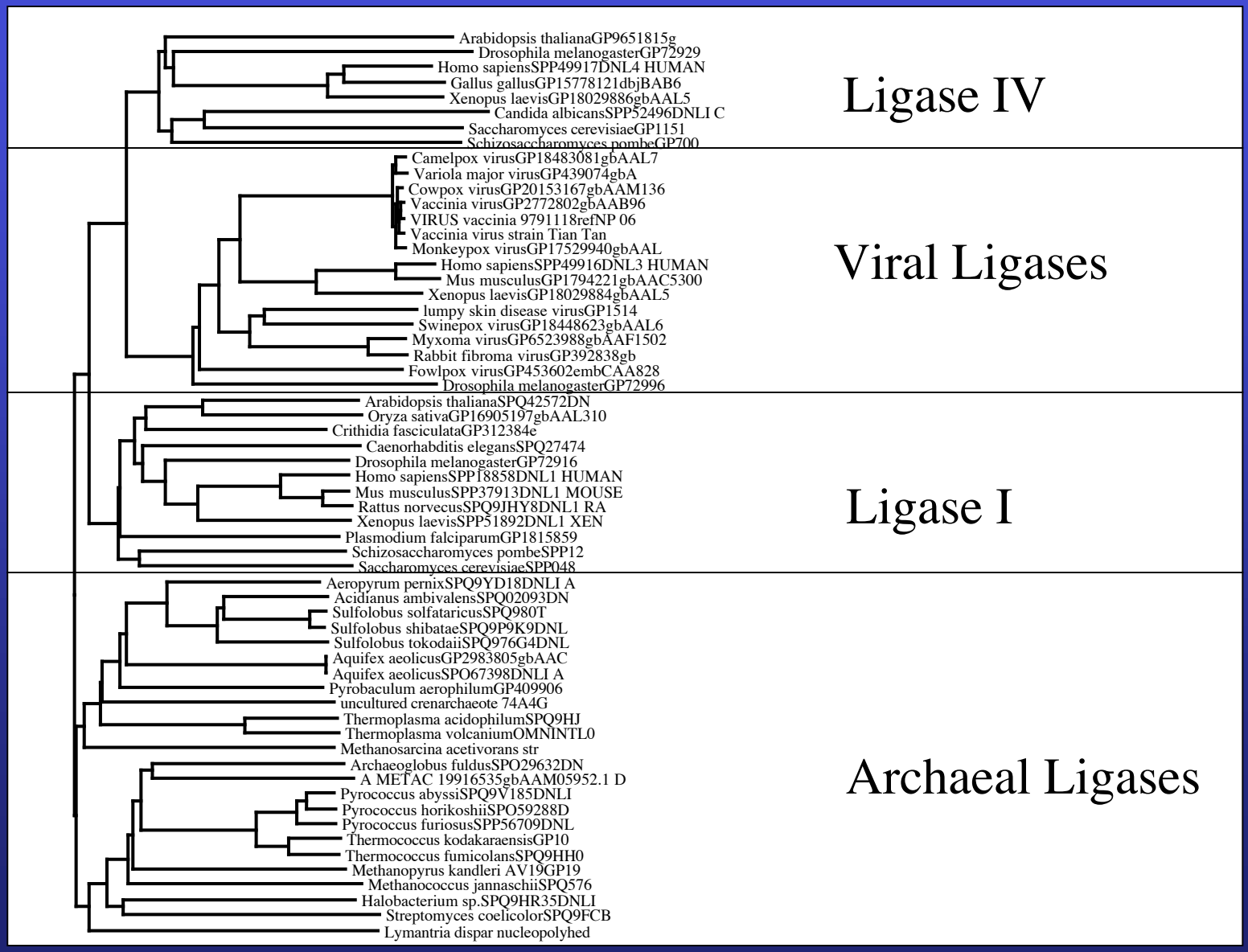
TIGR

P. falciparum DNA Repair



- *P. falciparum* initial annotation showed the presence of a DNA Ligase IV homolog but not a DNA Ligase I.
- Yet Ligase I has more of a core function in other eukaryotes, while Ligase IV is involved in NHEJ

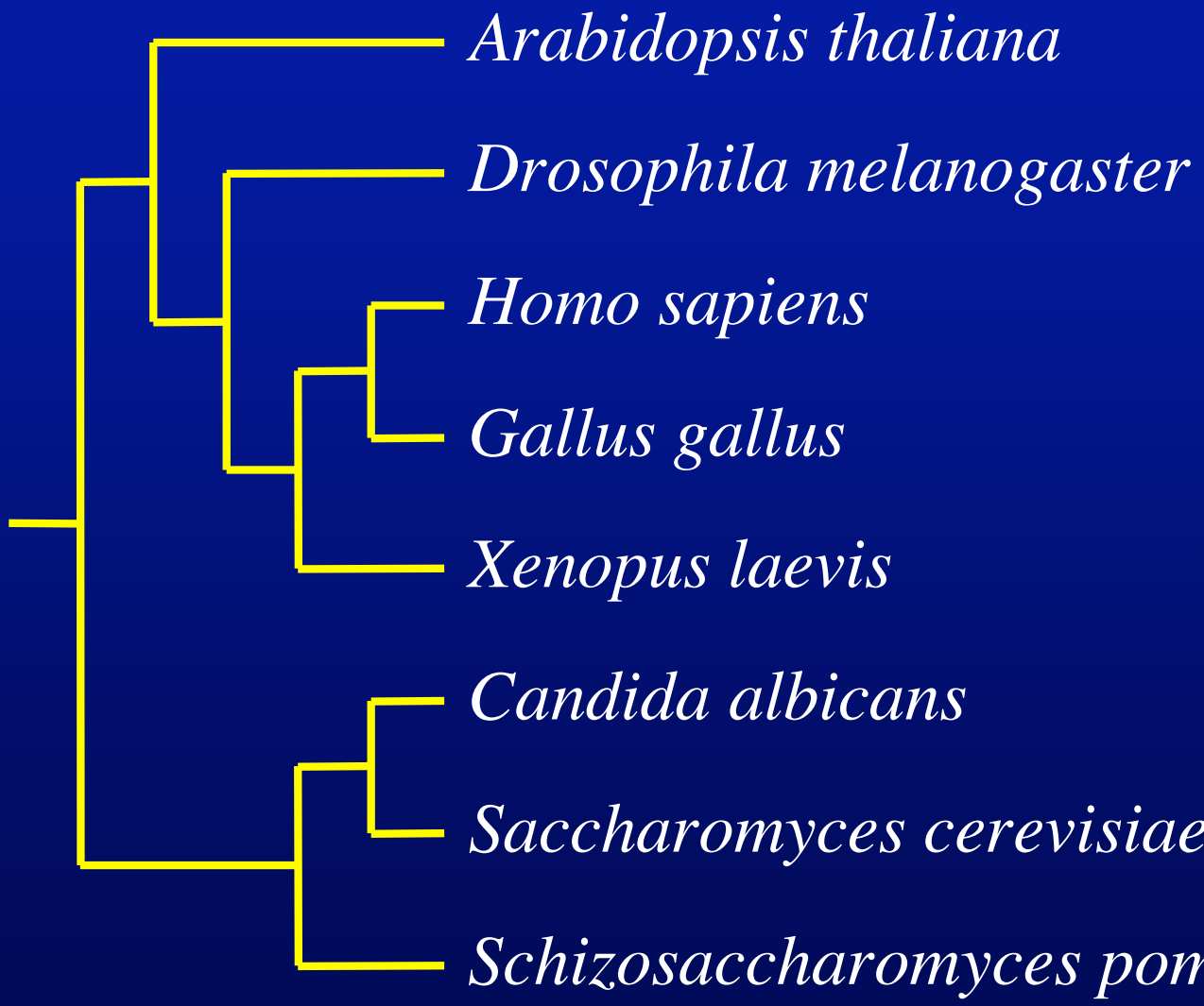
DNA Ligase Tree



TIGR

Eisen, unpublished

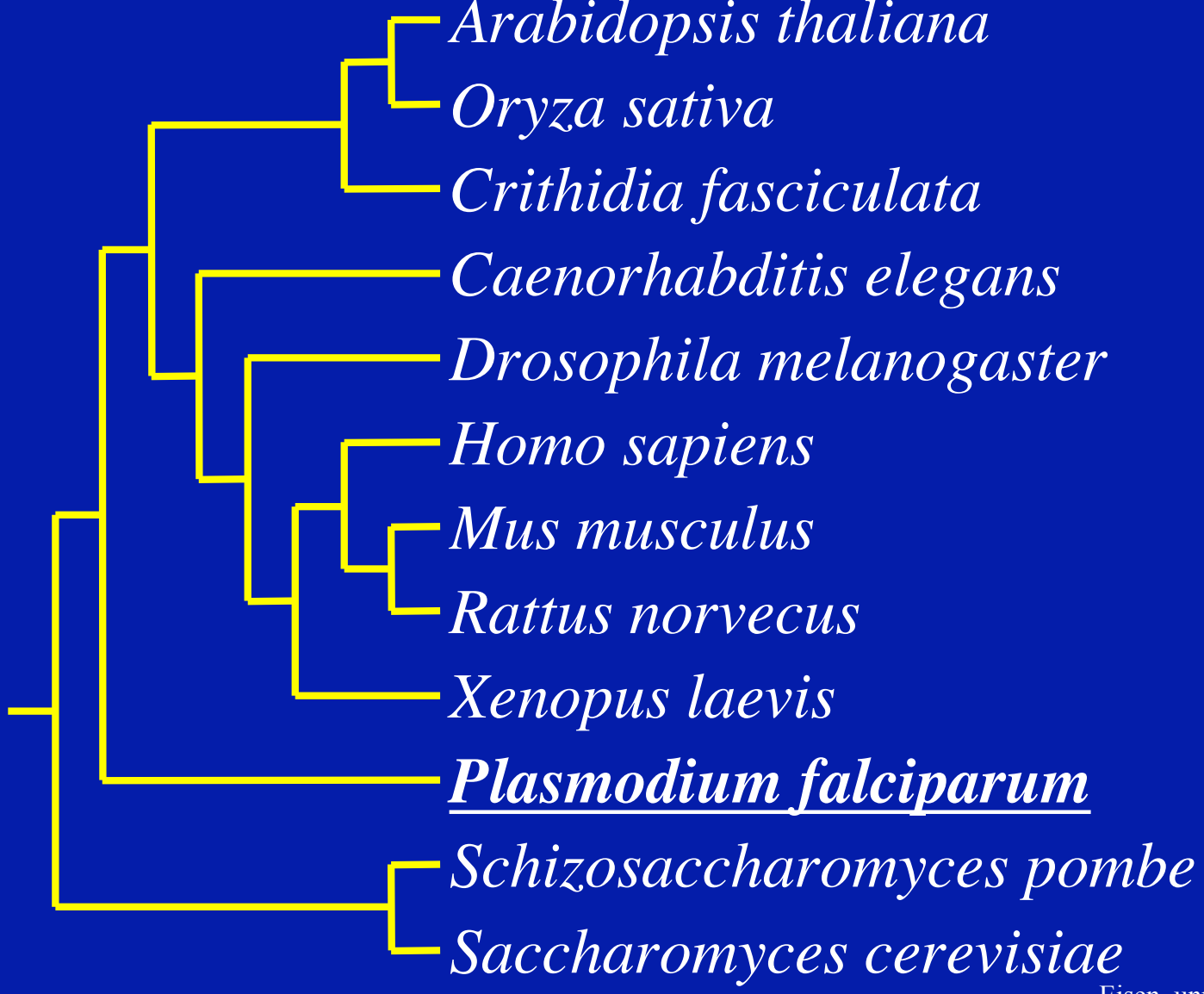
DNA Ligase IV SubTree



TIGR

Eisen, unpublished

DNA Ligase I SubTree



TIGR

Eisen, unpublished

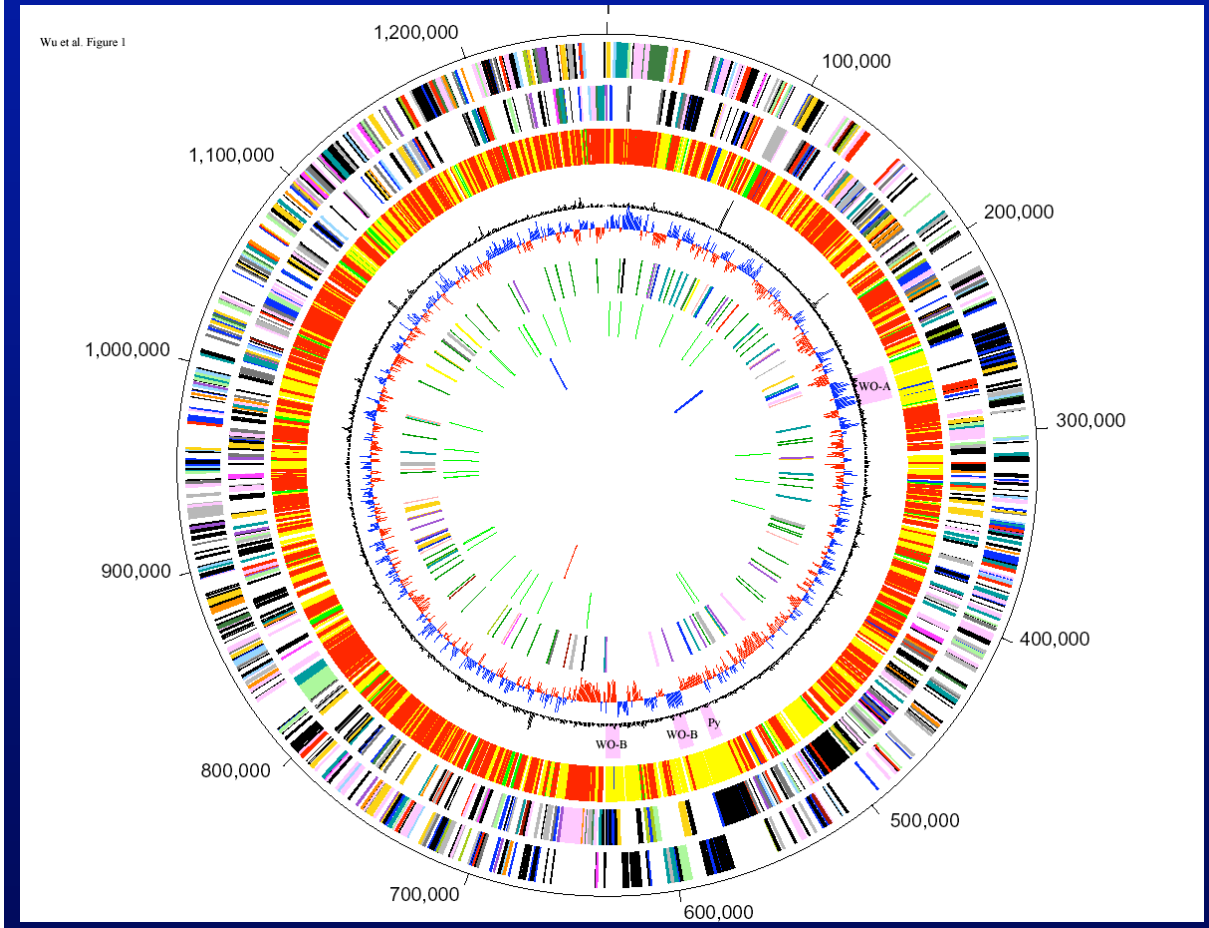
No NHEJ in *P. falciparum*?

- No Ligase IV
- None of the other "normal" NHEJ genes for eukaryotes
- Does it use only homologous recombination or is there another NHEJ pathway?

TIGR

See Gardner et al 2002

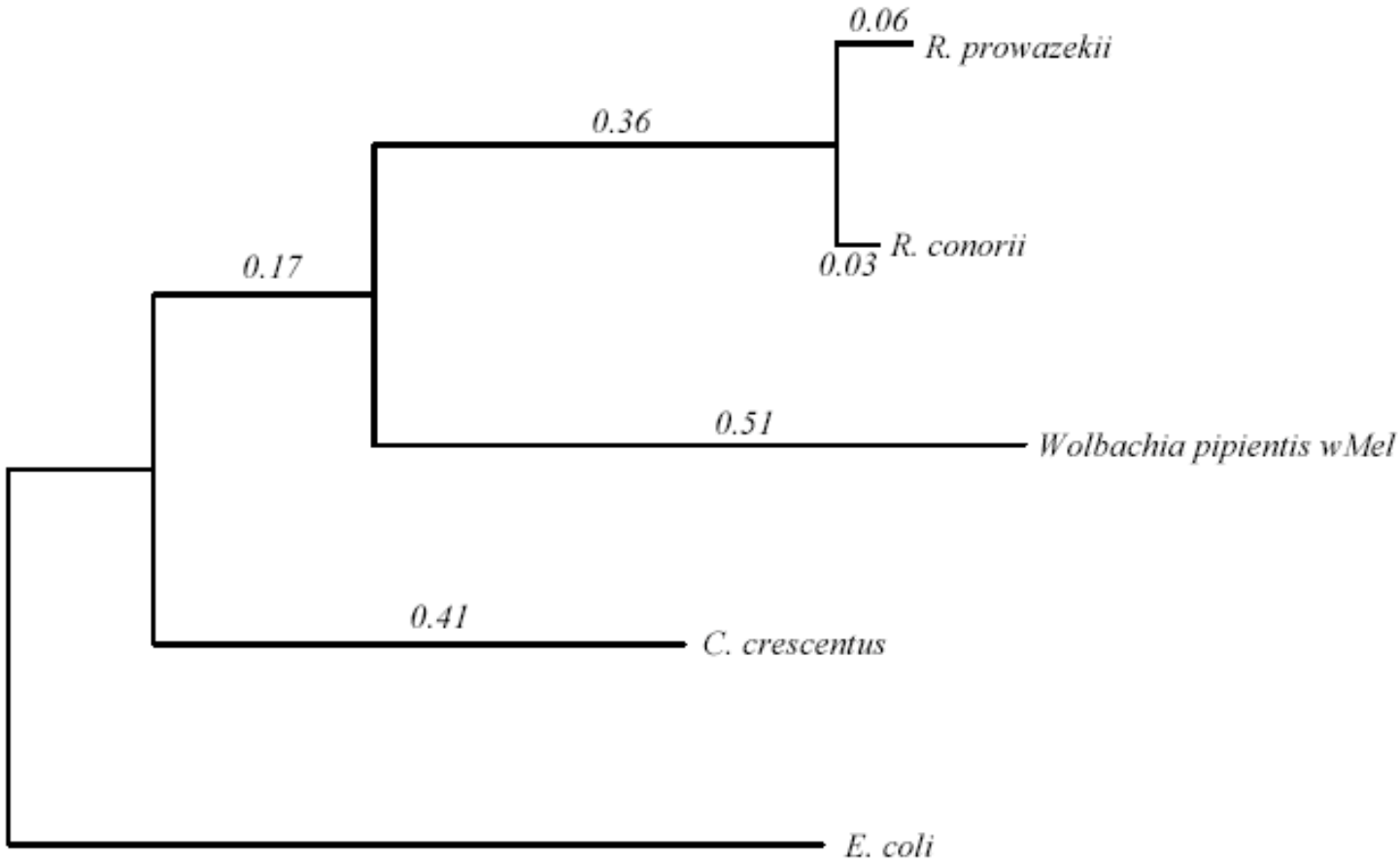
Wolbachia pipientis wMel



TIGR

Wu et al, 2004, PLOS Biology 2: 327-341

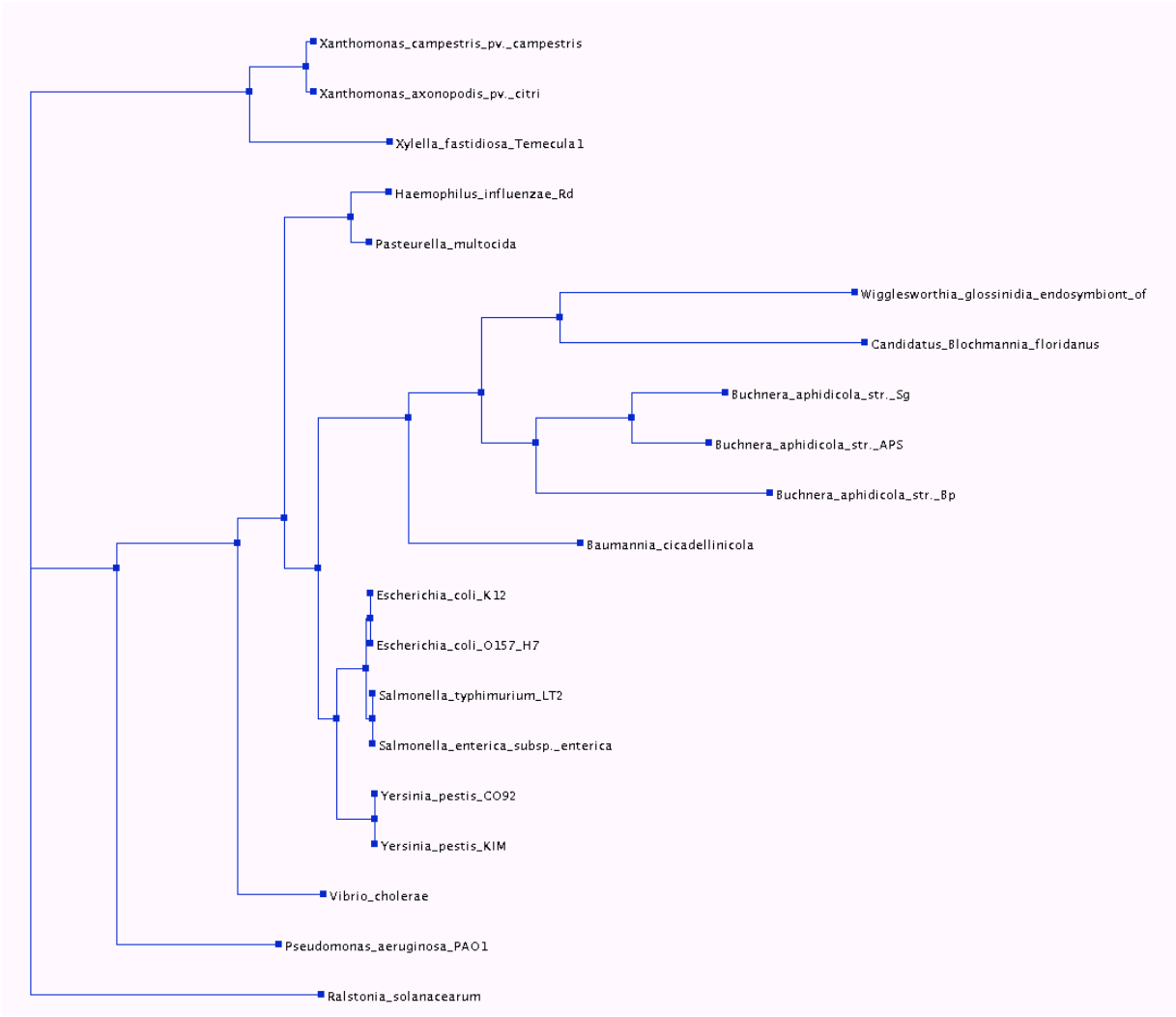
Wolbachia Evolves Rapidly



Wu et al., 2004

TIGR

Insect Symbiont Tree



TIGR

Wu et al. In preparation

Intracellular Organisms and Evolutionary Rate

- Many intracellular bacteria have higher evolutionary rates than free-living relatives
- Two main theories to explain this:
  - loss of DNA repair
  - altered population genetics (e.g., bottlenecks)

Wolbachia Has Similar Repair Genes to More Slowly Evolving Species

Repair Category	Gene Name
MMR	*mutS1, mutL1, mutL2*
NER	*uvrABCD*
BER	*fpg, mpg, nth*
AP Endo	*xth*
Rec	*recA, ruvABC, recG, recFJR*
Other	*ligA, ssb, radA*

TIGR

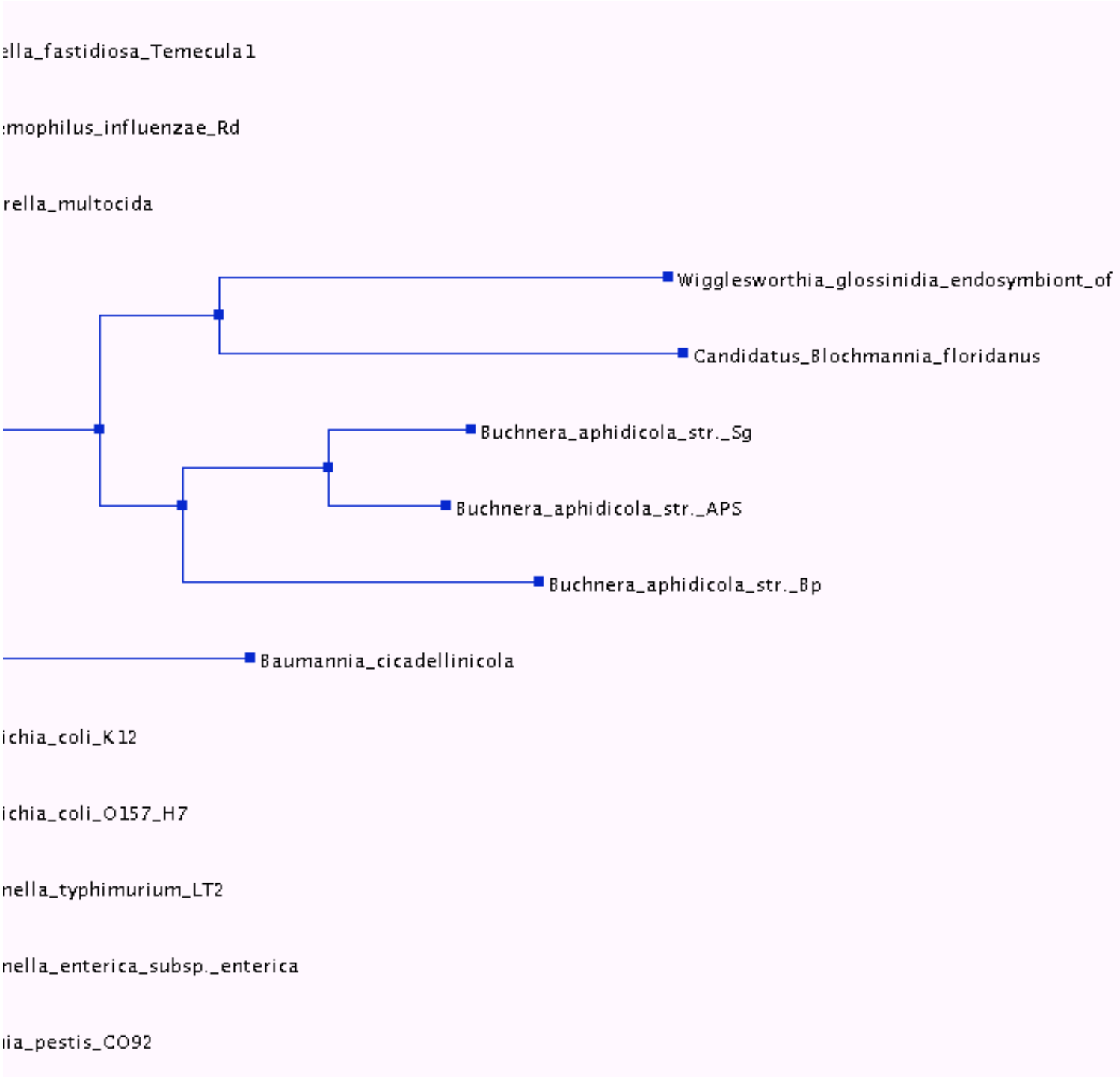
Wu et al., 2004

Insect Symbiont Repair

	Buchnera (APS)	Buchnera (BP)	Buchnera (SG)	Wigglesworthia	Blochmannia	Baumannia
Base Excision Repair						
Ung	gi15616802	gi27904674	-	gi32491032	gi33519990	GEBCORF00107
MutY	gi15617145	gi27904974	gi21672797	-	gi33519714	-
Fpg	-	-	-	-	-	GEBCORF00321
Nfo	gi15616757	gi27904631	gi21672420	-	-	GEBCORF00528
Xth	-	-	-	gi32491120	gi33519889	-
Nucleotide Excision Repair						
UvrD	-	-	-	gi32491021	-	GEBCORF00274
Transcription-Repair Coupling Factor						
Mfd	gi15616905	gi27904769	-	gi32490848	-	-
Mismatch Repair						
MutL	gi15617161	gi27904987	gi21672810	-	-	GEBCORF00079
MutS	gi15617028	gi27904861	gi21672681	-	-	GEBCORF00349
MutH	-	gi27904531	-	-	-	GEBCORF00038
8-oxo-dGTP Hydrolysis						
MutT	gi15616821	-	gi21672482	-	-	-
Pyrimidine Dimer Repair						
Phr	gi15616910	gi27904772	-	gi32490931	-	-
Homologous Recombination						
RecG	-	-	-	-	-	GEBCORF00259
RuvA	-	-	-	-	-	GEBCORF00452
RuvB	-	-	-	-	-	GEBCORF00451
RuvC	-	-	-	gi32490863	-	GEBCORF00453
RecB	gi15617053	gi27904885	gi21672704	gi32491015	gi33519733	GEBCORF00043
RecC	gi15617052	gi27904884	gi21672703	gi32491014	gi33519731	GEBCORF00042
RecD	gi15617054	gi27904886	gi21672705	gi32491016	gi33519732	GEBCORF00044
RecA	-	-	-	gi32490984	-	GEBCORF00346
RecJ	-	-	-	gi32491188	-	GEBCORF00298
RmuC	-	-	-	-	gi33520059	GEBCORF00163

TIGR

Wu et al. In preparation



ree

Mut LS	RecA
-	+
-	-
+	-
+	-
+	+
+	+

Wu et al. In preparation

Organelle Derived Genes in the *A. thaliana* Nuclear Genome

Pathway	Genes
Mismatch Repair	MutS2
Base Excision Repair	Fpg, Tag?
Nucleotide Excision Repair	Mfd
Recombination	RecA, RecG
Direct Repair	Phr
Other	Lon

Eisen and Britt, 2000

More Questions V:

One of these things just does not belong here

TIGR

Other Unusual Patterns

- RecBCD but no RecA
- Mfd but no UvrABCD
- No Ung in thermophiles
- Rad25 in some bacteria
- No MSH4, 5 in many eukaryotes
- No CSB in some eukaryotes
- Tons of repair genes in Mimivirus

Archaea

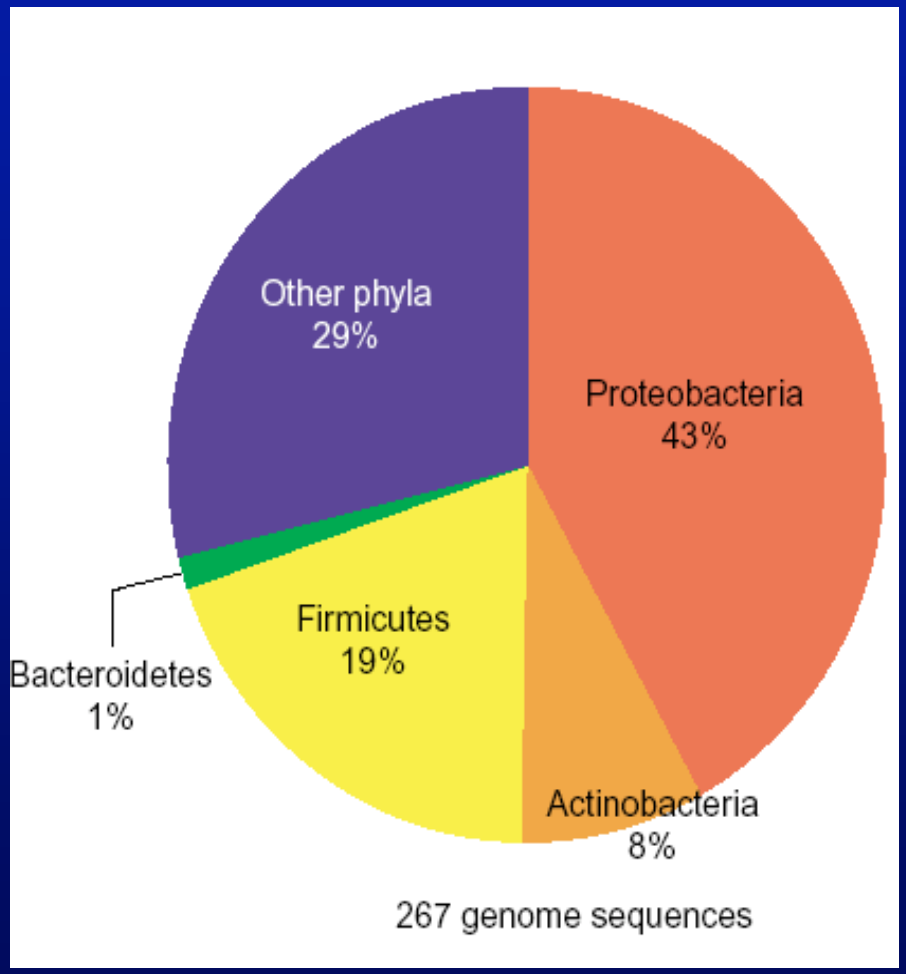
- NER
  - UvrABC in some
  - Rad1, Rad2, Rad25 etc in others
  - Some species with both
- MMR
  - Many species with MutS2 only
  - Some with multiple MutLs, MutS1s
  - Three MutSs in Halophiles

Even More Questions?

You ain’t seen nothing yet

TIGR

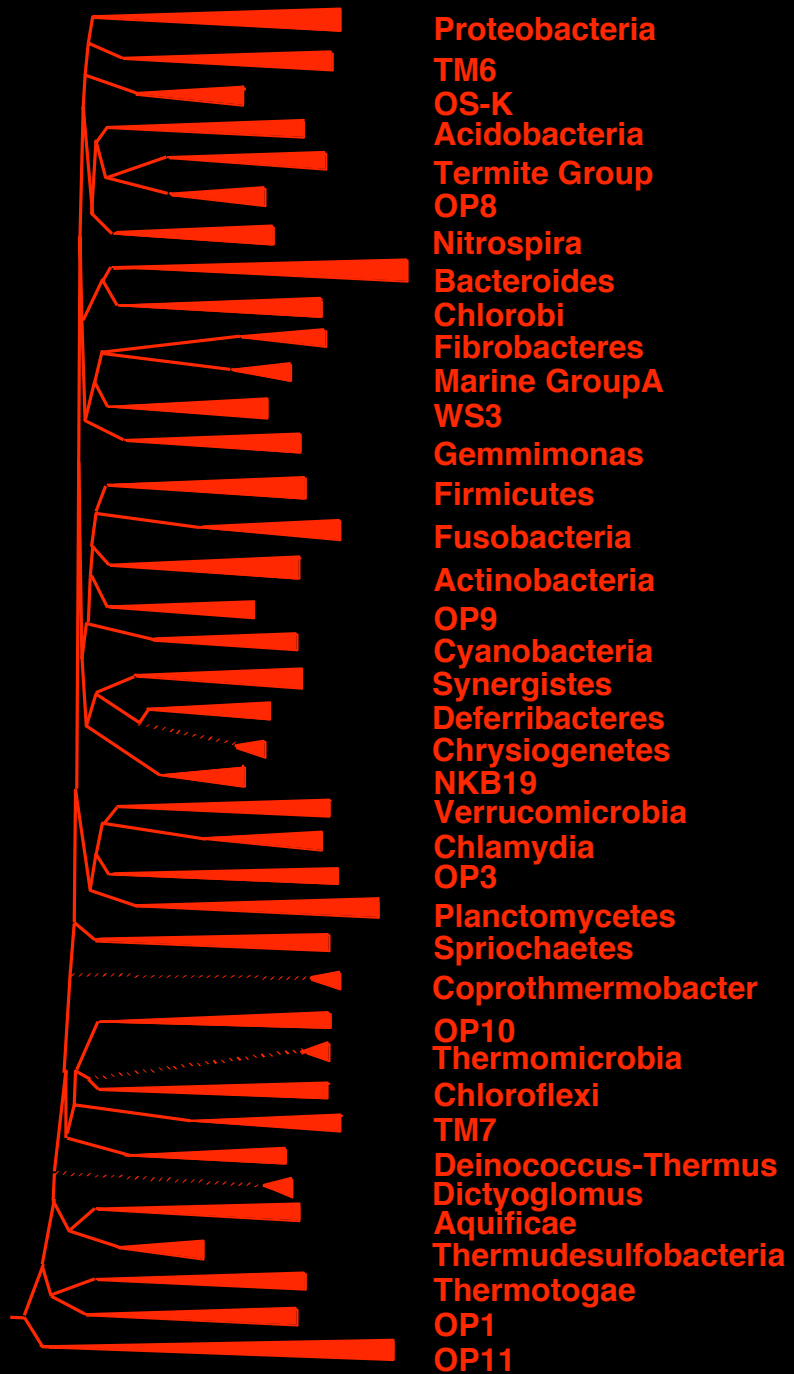
Biased Sampling of Genomes



- Most bacterial genomes come from three phyla
- Most phyla have no genome sequences
- Even worse sampling for eukaryotes and archaea

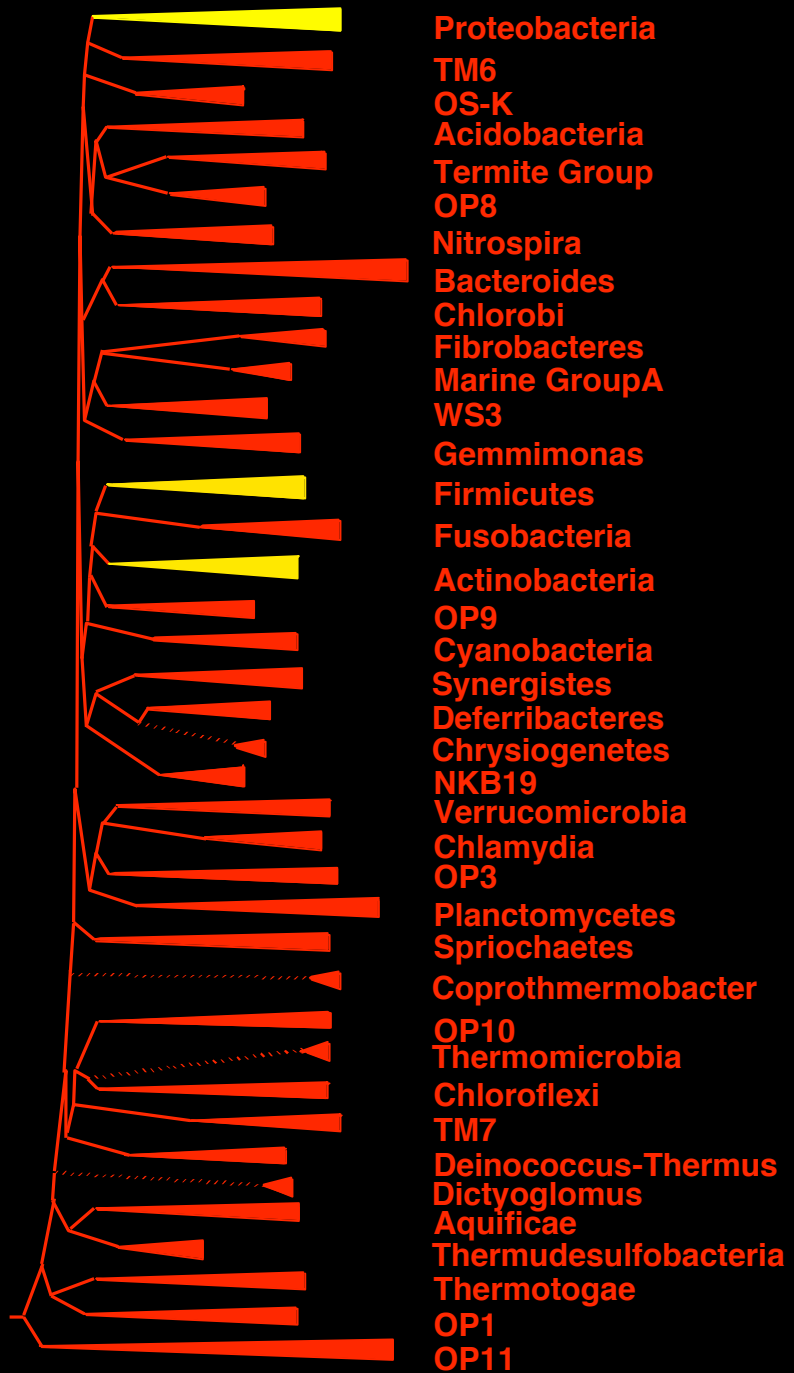
Hugenholtz 2002

TIGR



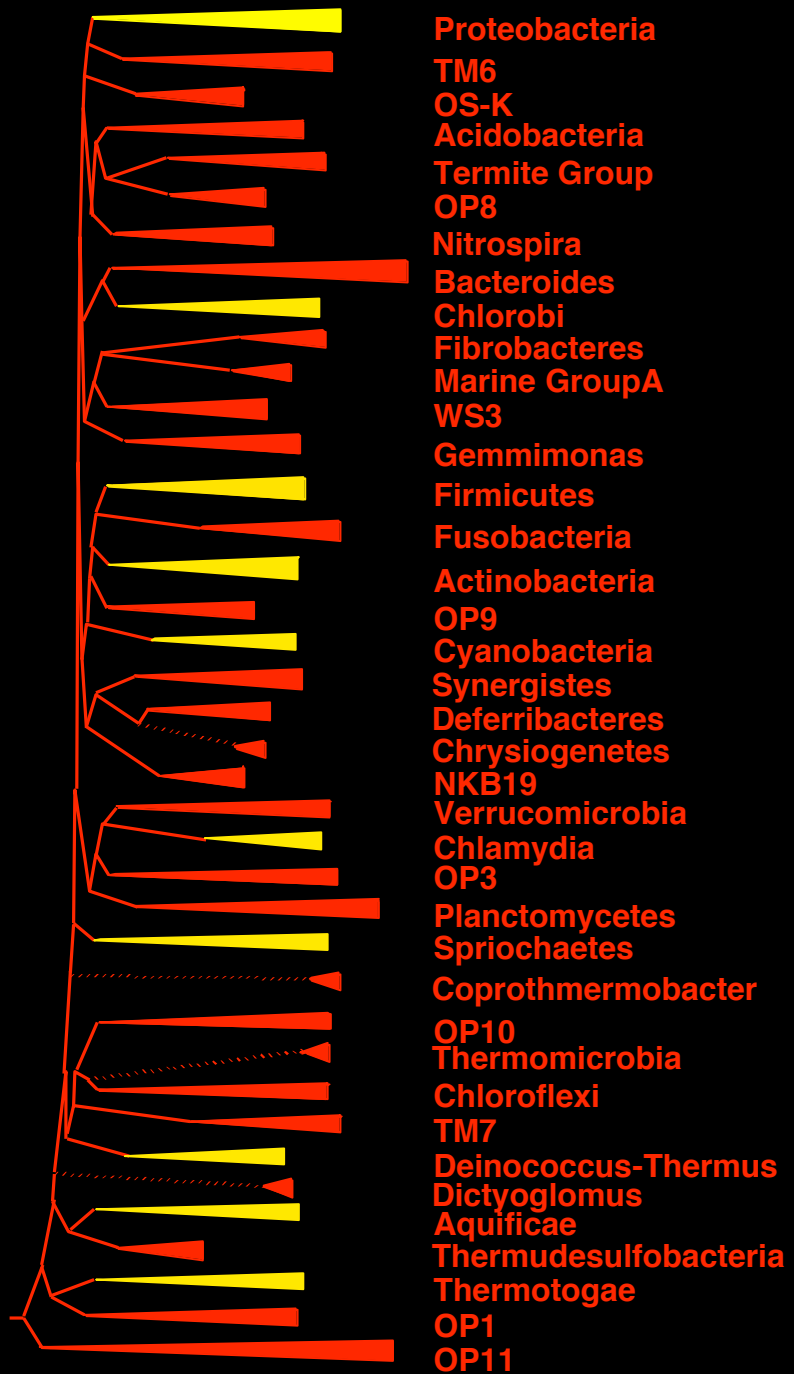
- At least 40 phyla of bacteria

TIGR



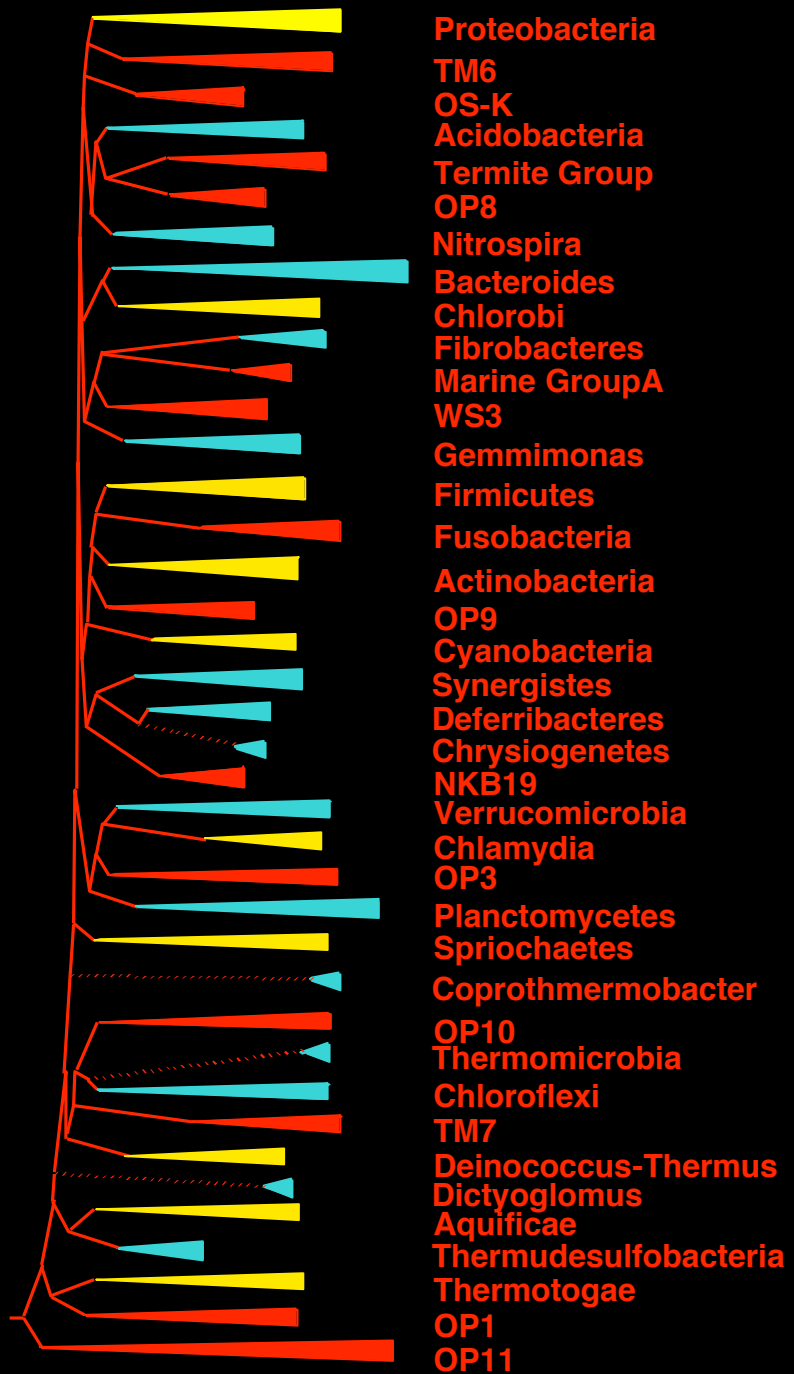
- At least 40 phyla of bacteria
- Genome sequences are mostly from three phyla

TIGR



- At least 40 phyla of bacteria
- Genome sequences are mostly from three phyla
- Some other phyla are only sparsely sampled

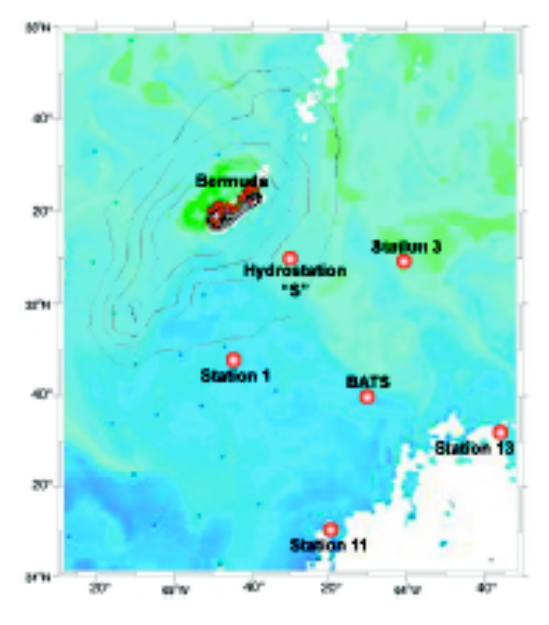
TIGR



- At least 40 phyla of bacteria
- Genome sequences are mostly from three phyla
- Some other phyla are only sparsely sampled
- Solution: sequence more phyla

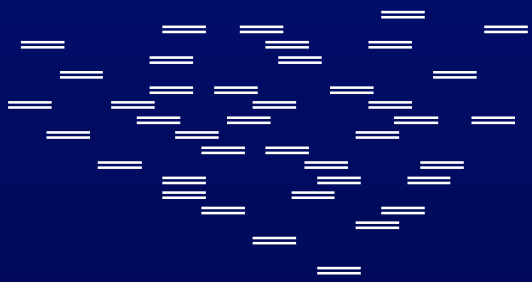
TIGR

Solution II: Environmental Sequencing

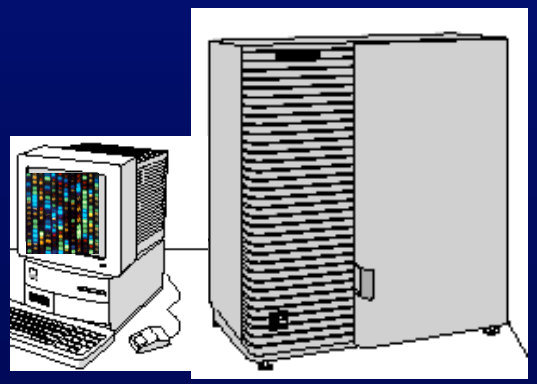


Warner Brothers, Inc.

shotgun

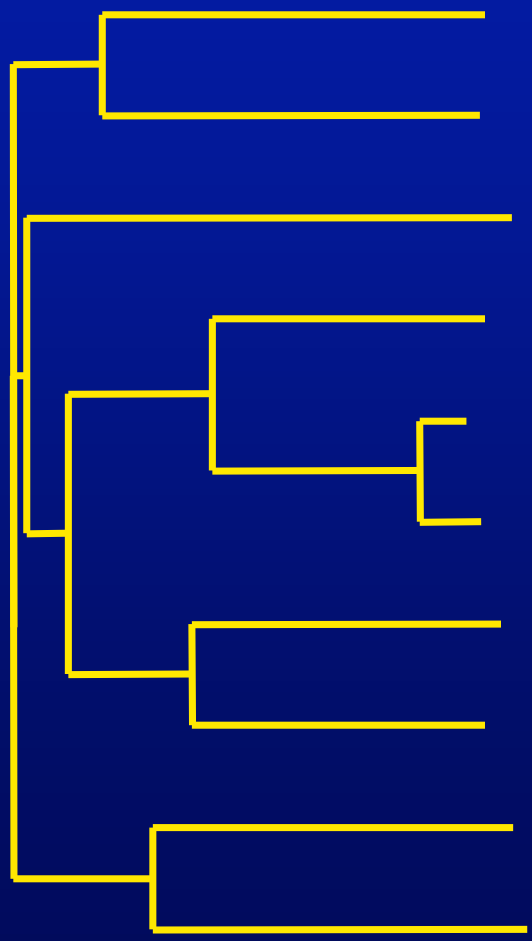


sequence

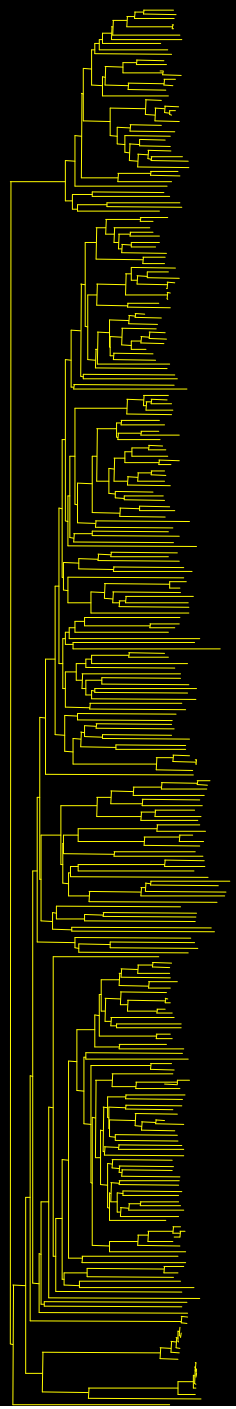


TIGR

Duplication and Loss of UVDE

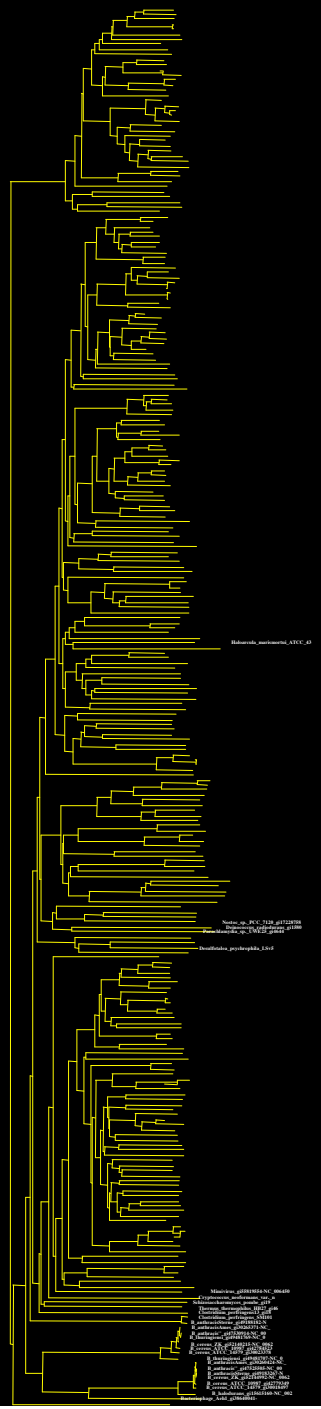


Schizosaccharomyces pombe
Neurospora crassa
Clostridium perfringens
Bacillus subtilis
Bacillus cereus
Bacillus anthracis UVDE2
Bacillus halodurans
Bacillus anthracis UVDE1
Deinococcus radiodurans
Nostoc sp. PCC 7120



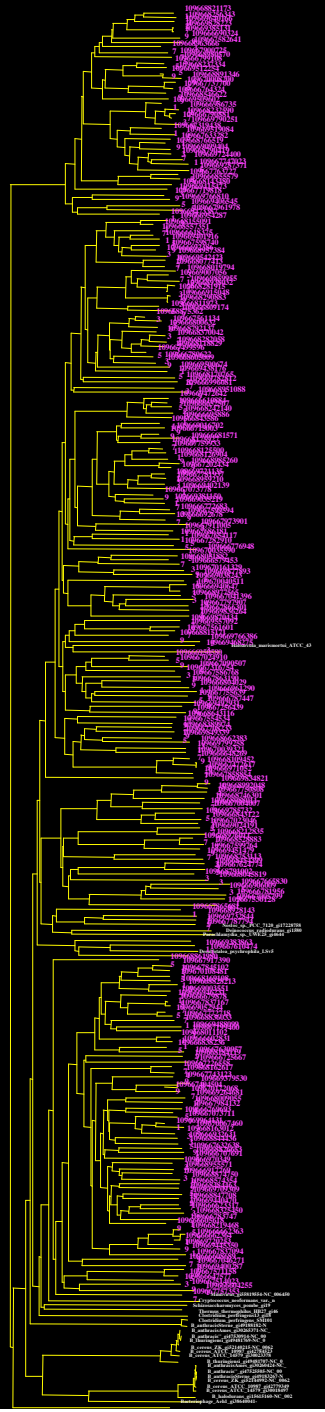
TIGR

Eisen, unpublished



TIGR

Eisen, unpublished



TIGR

Eisen, unpublished

Answers I:

Better Bioinformatics

TIGR

Non Homology Based Functional Prediction: Phylogenetic Profiles

Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 4285–4288, April 1999
Biochemistry

Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles

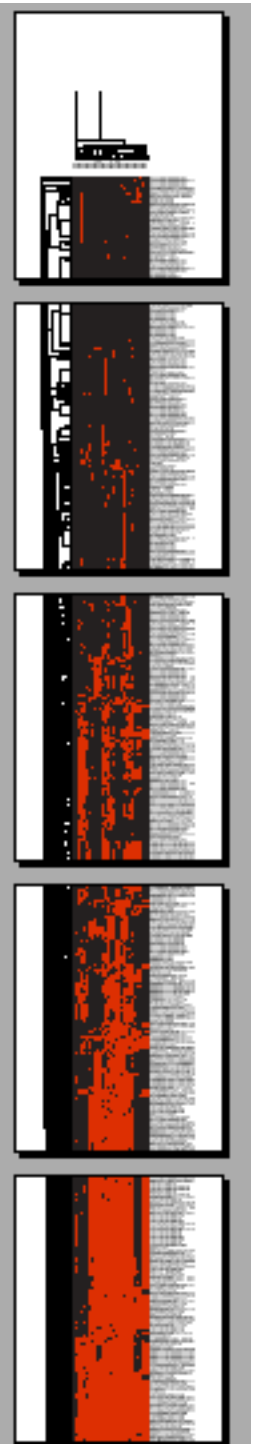
(genomic/bioinformatics/metabolic pathways/structural complexes)

MATTEO PELLEGRINI*, EDWARD M. MARCOTTE*, MICHAEL J. THOMPSON, DAVID EISENBERG, AND TODD O. YEATES†

TIGR

Non-Homology Prediction: Phylogenetic Profiles

- Step 1: Search all genes in organisms of interest against all other genomes
- Ask: Yes or No, is each gene found in each other species
- Cluster genes by distribution patterns (profiles)



TIGR

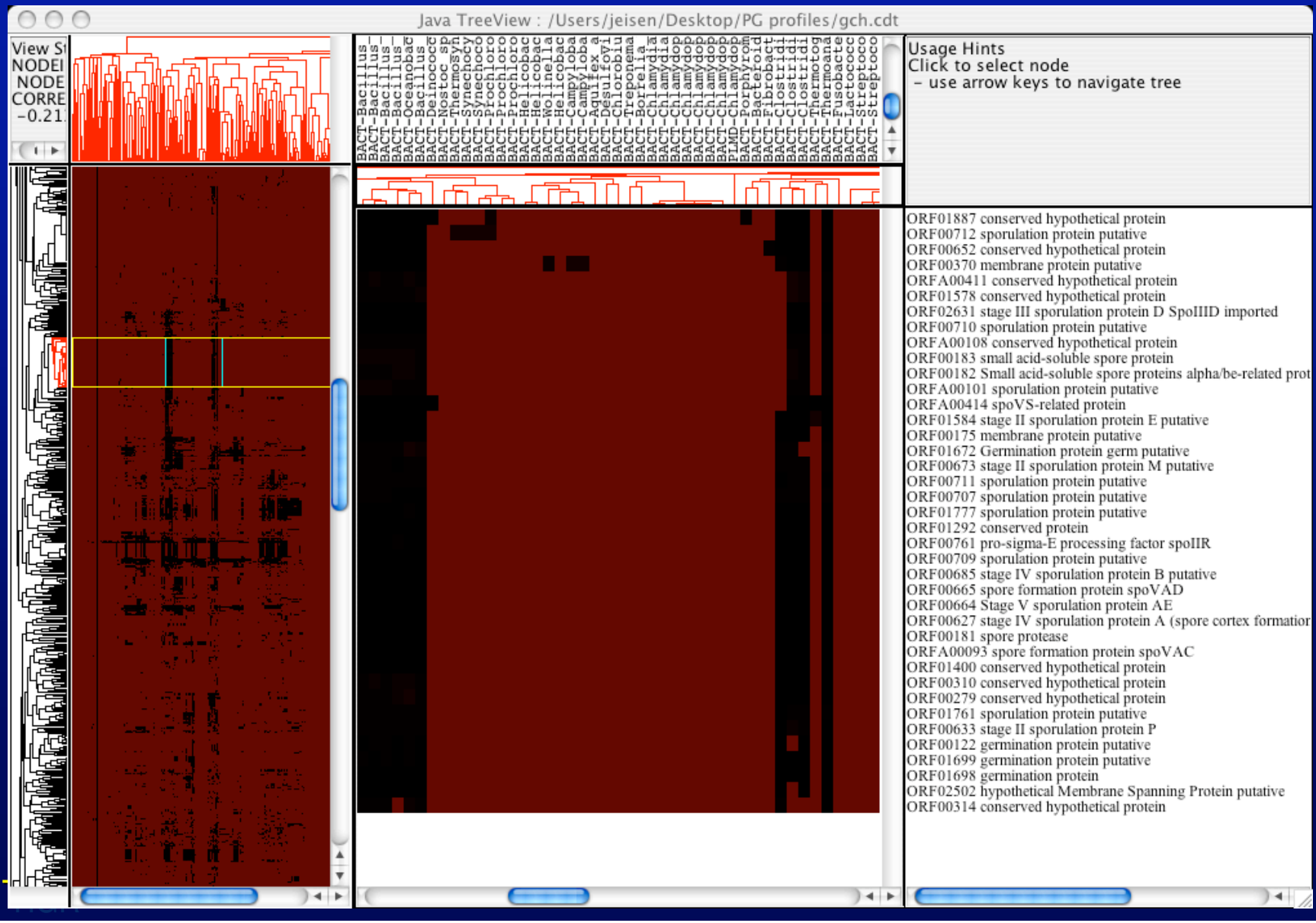
MutL and MutS1 Always Together

Table 3. Presence of MutS Homologs in Complete Genomes Sequences

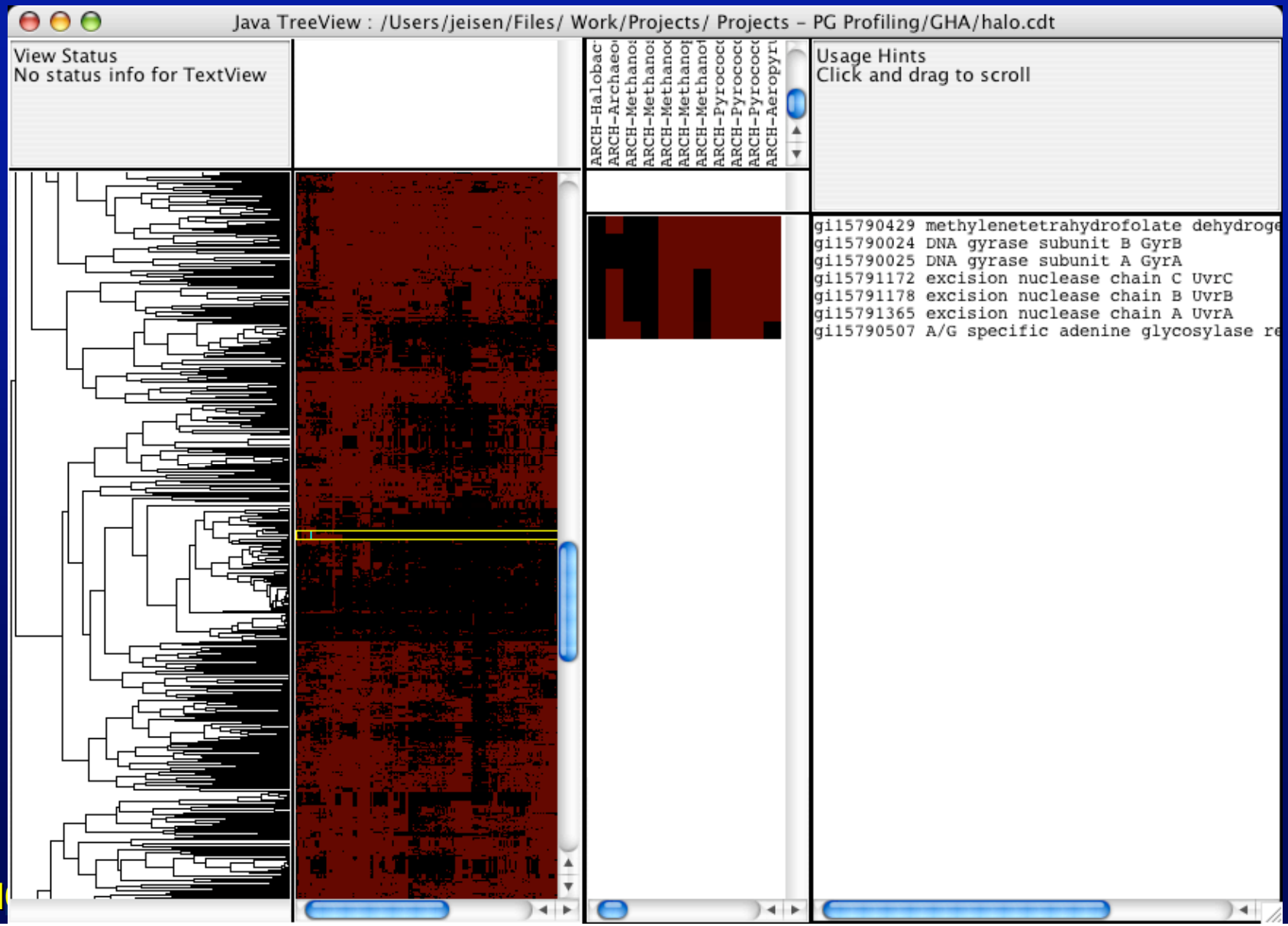
Species	# of MutS Homologs	Which Subfamilies?	MutL Homologs
Bacteria			
Escherichia coli K12	1	*MutS1*	1
Haemophilus influenzae Rd KW20	1	*MutS1*	1
Neisseria gonorrhoeae	1	*MutS1*	1
Helicobacter pylori 26695	1	*MutS2*	-
Mycoplasma genitalium G-37	-	-	-
Mycoplasma pneumoniae M129	-	-	-
Bacillus subtilis 169	2	*MutS1,MutS2*	1
Streptococcus pyogenes	2	*MutS1,MutS2*	1
Mycobacterium tuberculosis	-	-	-
Synechocystis sp. PCC6803	2	*MutS1,MutS2*	1
Treponema pallidum Nichols	1	*MutS1*	1
Borrelia burgdorferi B31	2	*MutS1,MutS2*	1
Aquifex aeolicus	2	*MutS1,MutS2*	1
Deinococcus radiodurans R1	2	*MutS1,MutS2*	1
Archaea			
Archaeoglobus fulgidus VC-16, DSM4304	-	-	-
Methanococcus janasscii DSM 2661	-	-	-
Methanobacterium thermoautotrophicum ΔH	1	*MutS2*	-
Eukaryotes			
Saccharomyces cerevisiae	6	*MSH1-6*	3+
Homo sapiens	5	*MSH2-6*	3+

TIGR

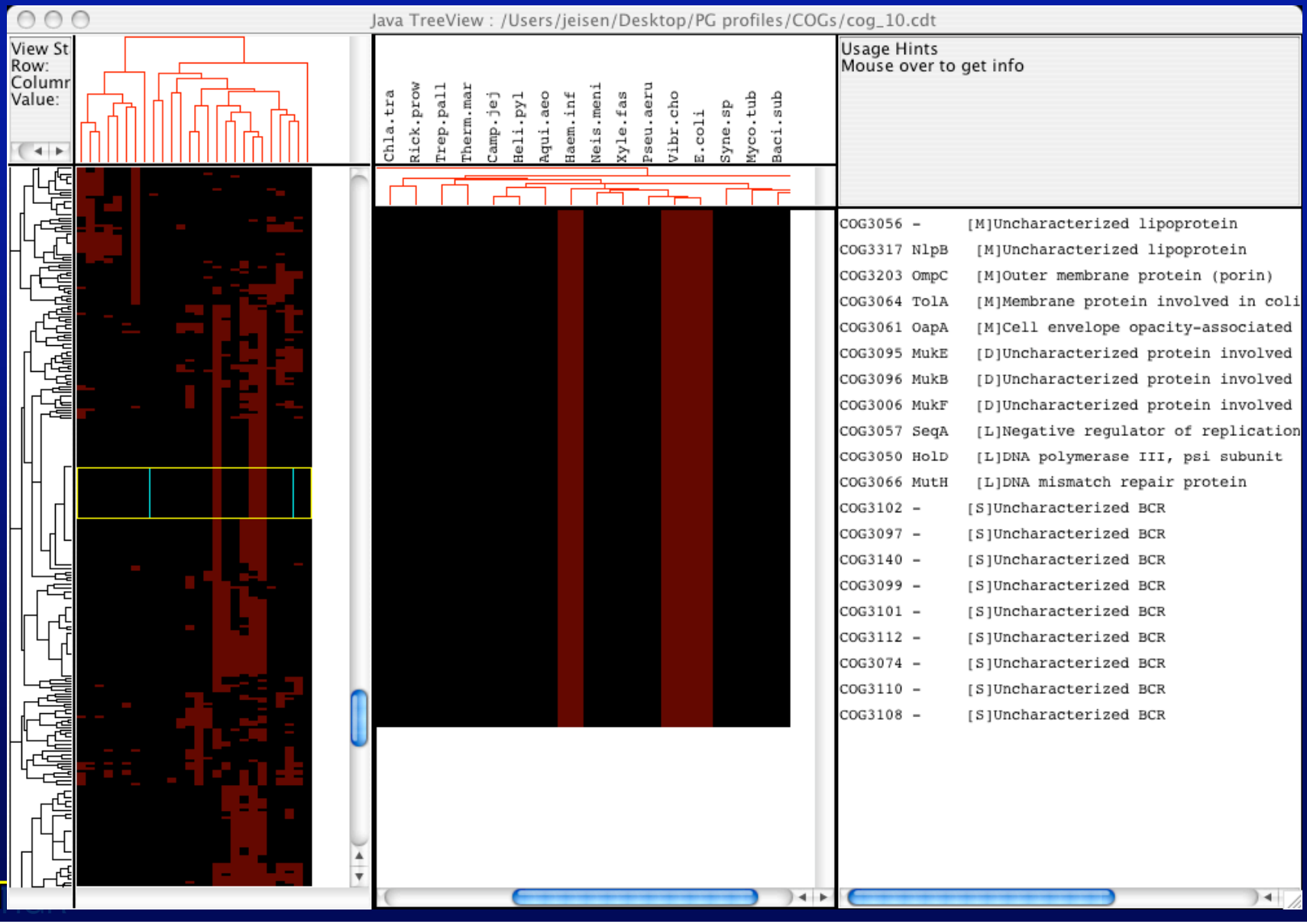
Co-Occurrence of Spo Genes



Co Occurrence of UvrABC



Co-Occurrence of Muk, SeqA, MutH

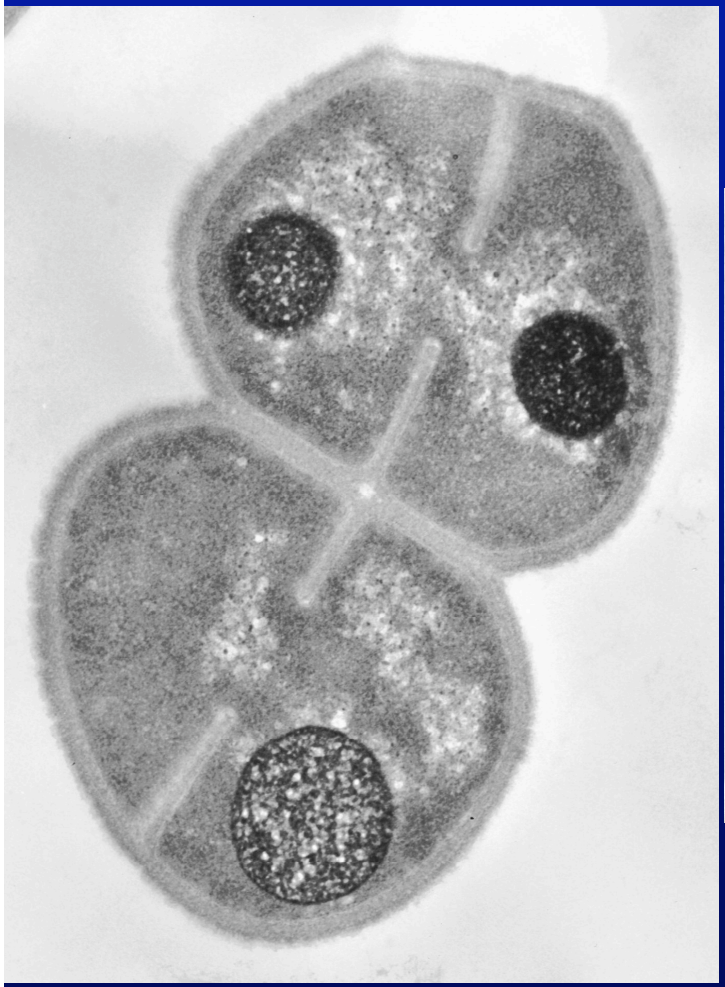


Answers II:

Novelty is good and bad

TIGR

Deinococcus radiodurans



Odd Microbe Survives Vast Dose of Radiation

By MALCOLM W. BROWNE

NO ONE knows for sure whether any creature could maintain even a spark of life while drifting through the killing environment of interstellar space for thousands of

A bacterium can reassemble broken chromosomes.

TIGR

DNA Repair Genes in *D. radiodurans* Complete Genome

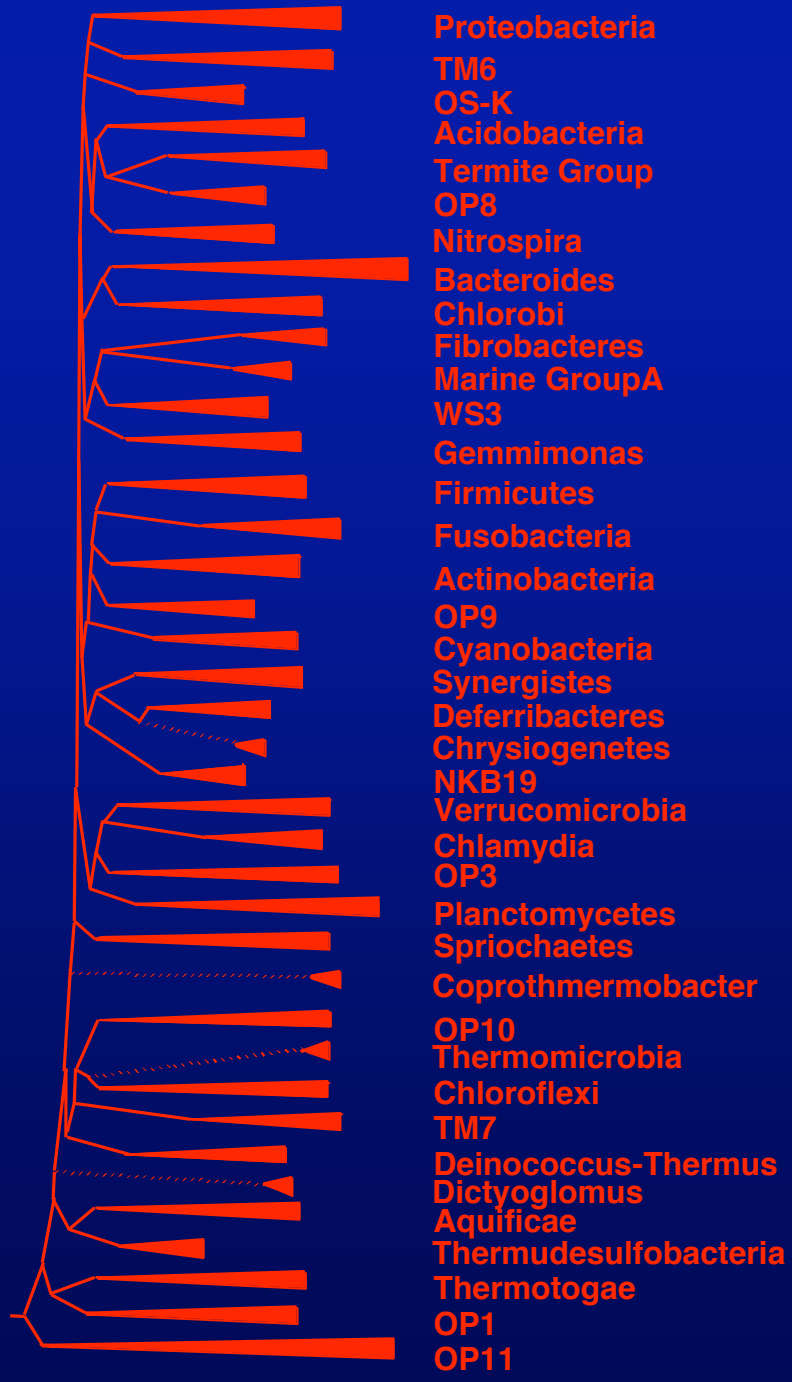
Process	Genes in *D. radiodurans*
Nucleotide Excision Repair	UvrABCD, UvrA2
Base Excision Repair	AlkA, Ung, Ung2, GT, MutM, MutY-Nths, MPG
AP Endonuclease	Xth
Mismatch Excision Repair	MutS, MutL
Recombination	
Initiation	RecFJNRQ, SbcCD, RecD
Recombinase	RecA
Migration and resolution	RuvABC, RecG
Replication	PolA, PolC, PolX, phage Pol
Ligation	DnlJ
dNTP pools, cleanup	MutTs, RRase
Other	LexA, RadA, HepA, UVDE, MutS2

TIGR

White et al, 1999

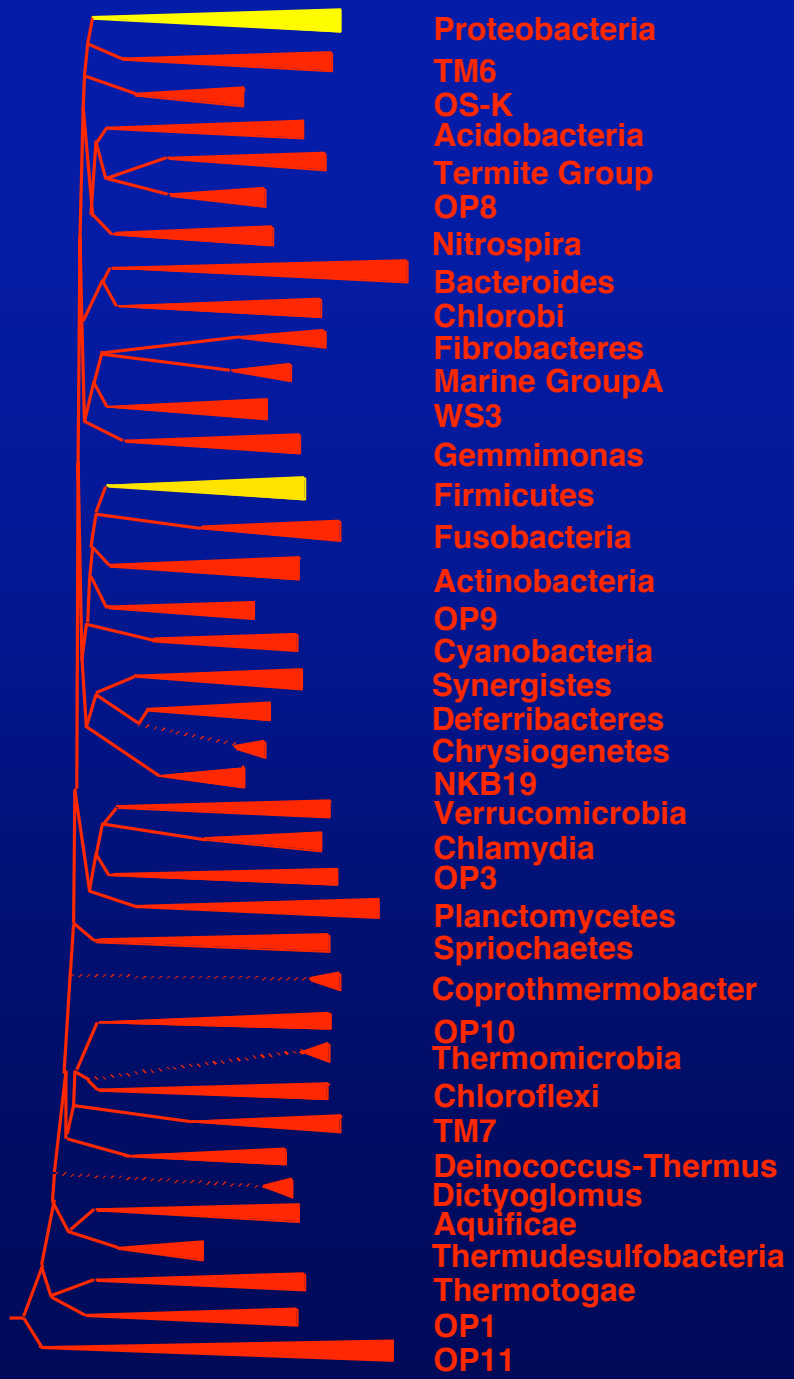
Problem:

List of DNA repair gene homologs in *D. radiodurans* genome is not significantly different from other bacterial genomes of the similar size



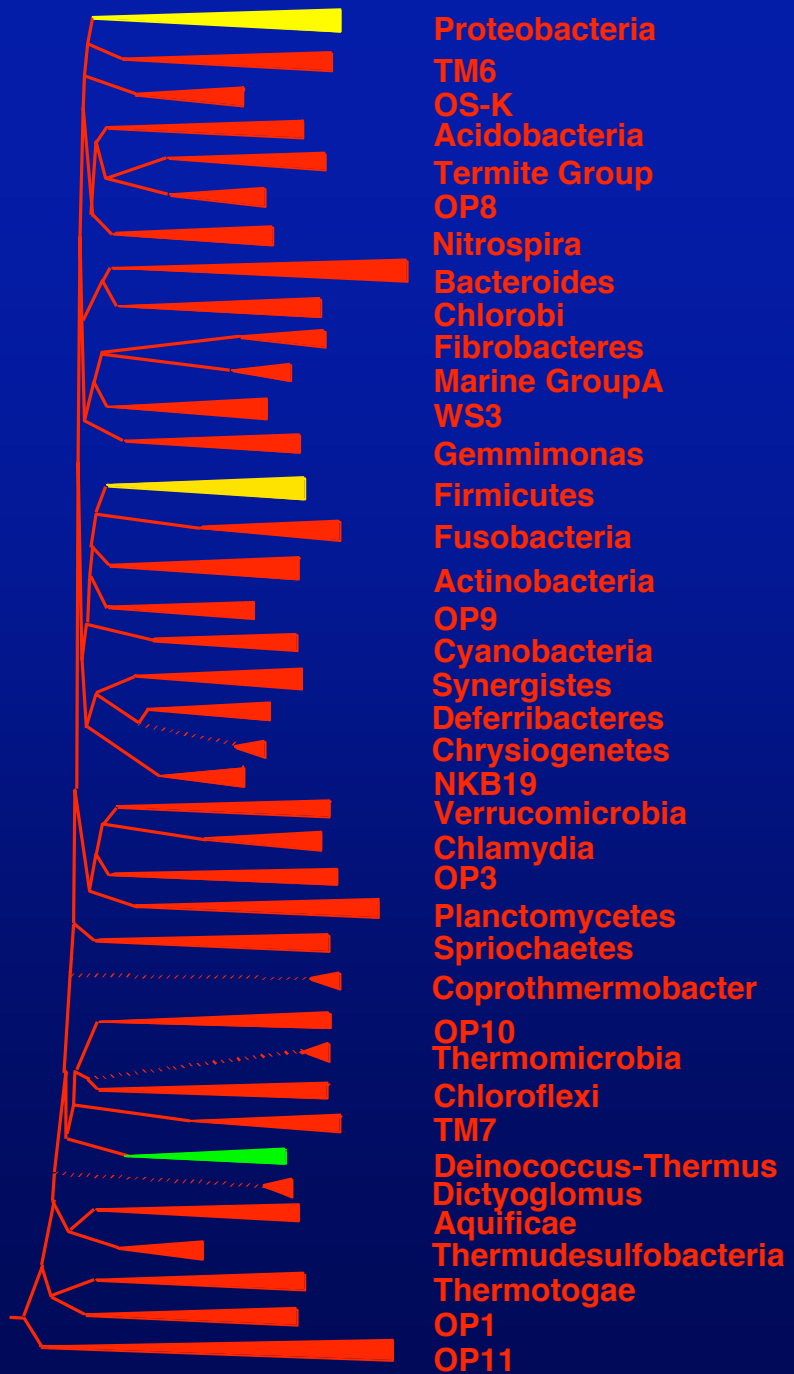
- At least 40 phyla of bacteria

TIGR



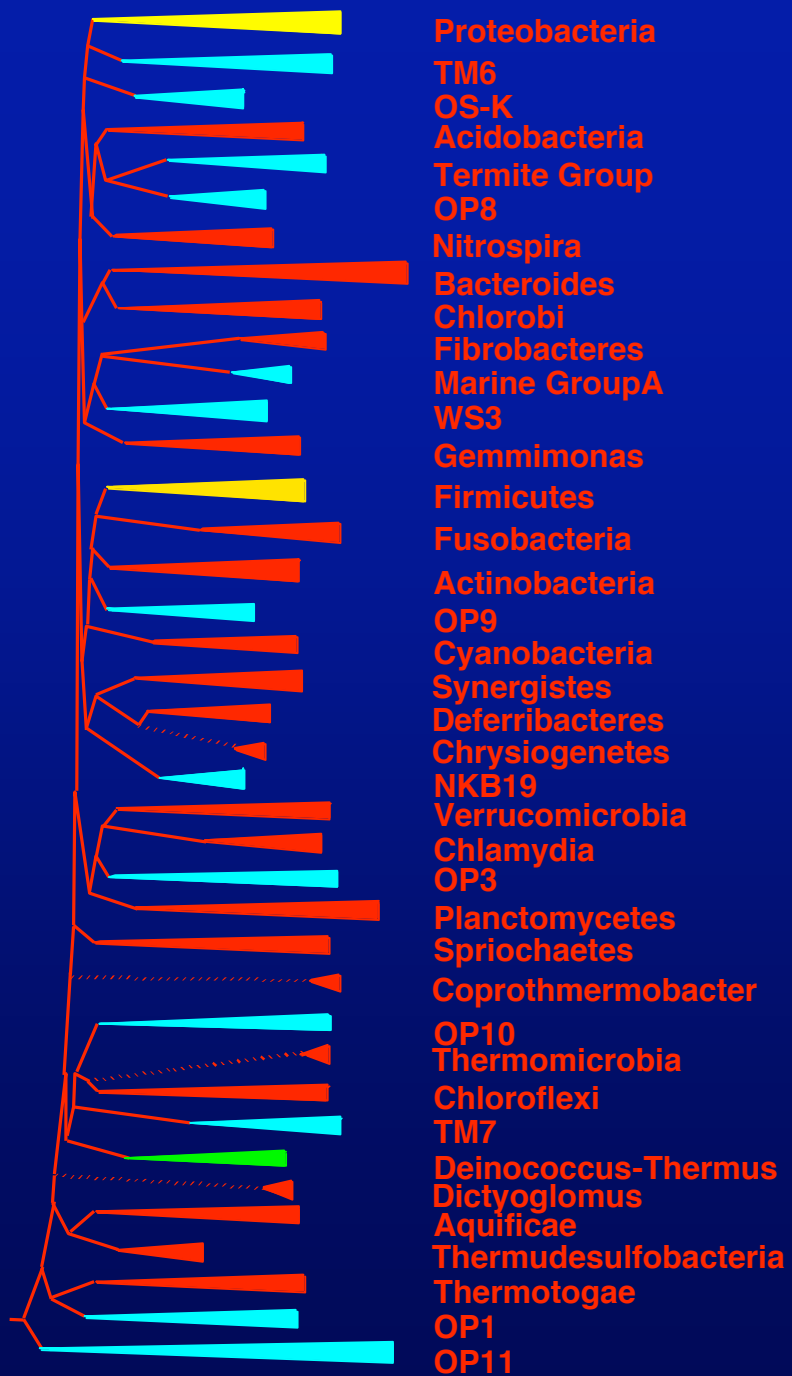
- At least 40 phyla of bacteria
- Most DNA metabolism studies in two phyla

TIGR



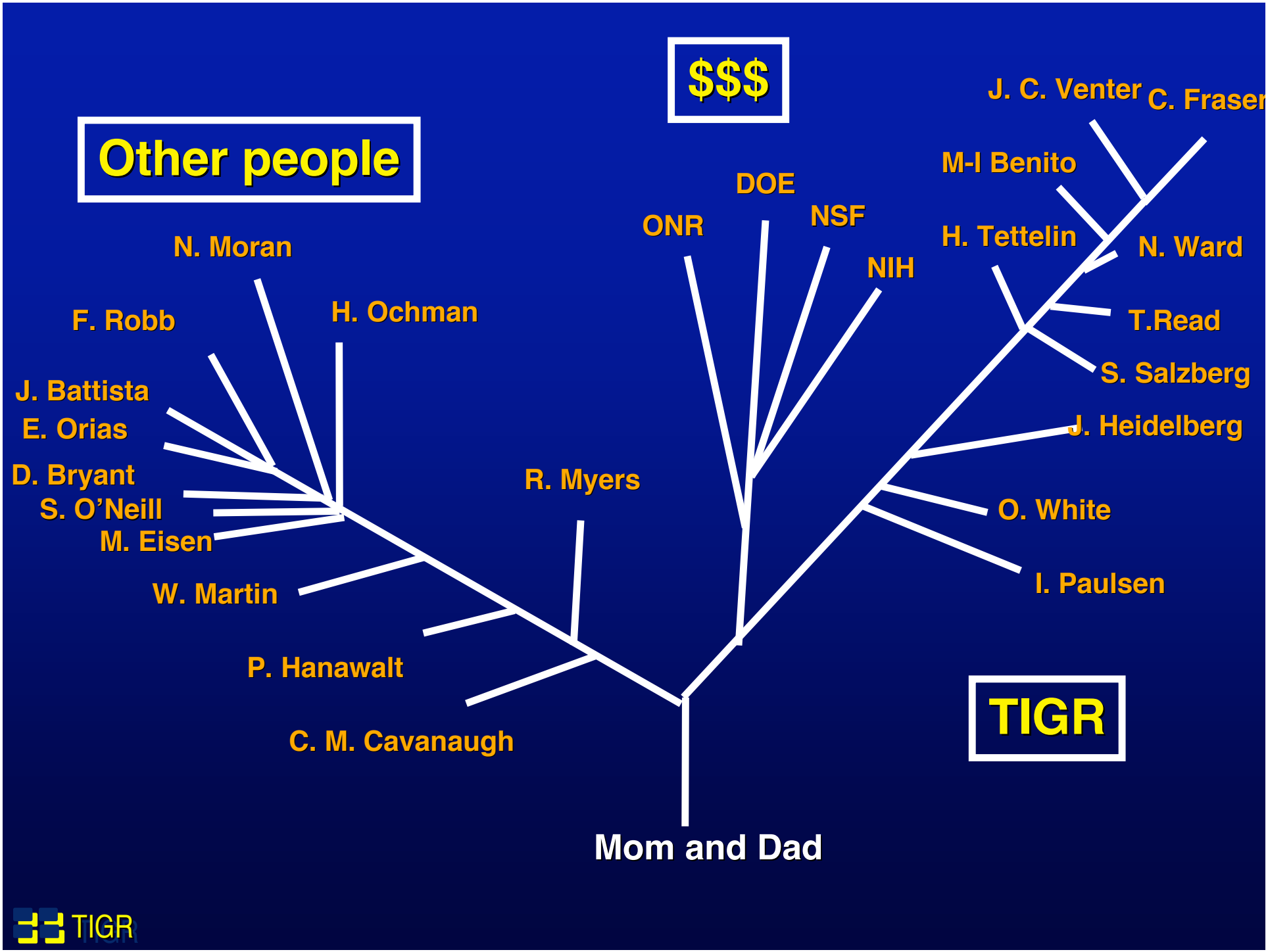
- At least 40 phyla of bacteria
- Most DNA metabolism studies in two phyla
- Deinococcus is distant from any well studied organism

TIGR



- At least 40 phyla of bacteria
- Most DNA metabolism studies in two phyla
- *Deinococcus* is distant from any well studied organism
- Solution: more studies

TIGR


Other people
$$$
J. C. Venter
C. Fraser
M-I Benito
DOE
NSF
ONR
H. Tettelin
N. Ward
NIH
N. Moran
F. Robb
H. Ochman
T.Read
S. Salzberg
J. Battista
E. Orias
J. Heidelberg
D. Bryant
R. Myers
S. O'Neill
O. White
M. Eisen
I. Paulsen
W. Martin
P. Hanawalt
TIGR
C. M. Cavanaugh
Mom and Dad
TIGR