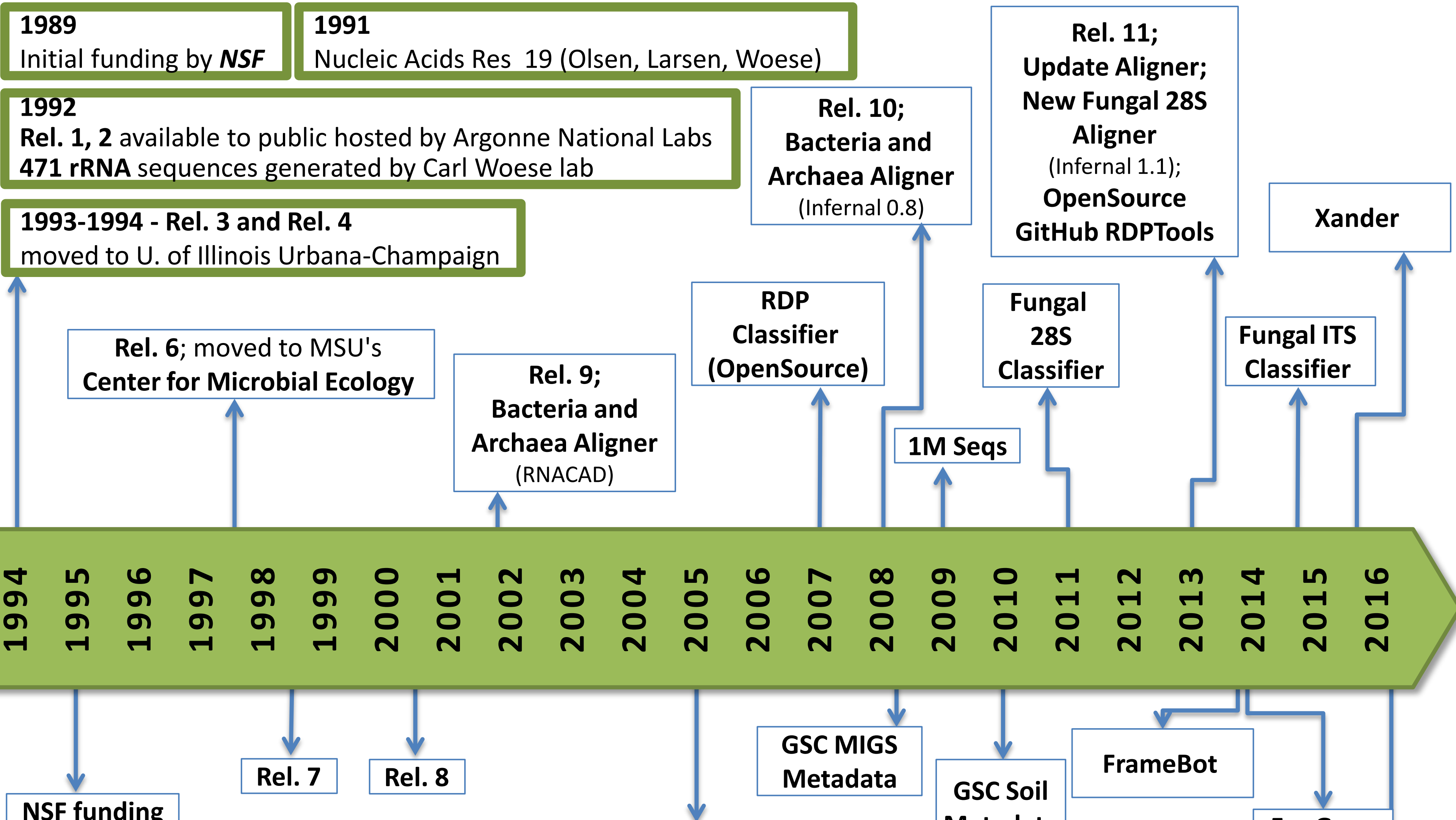
25 March 2019

Collaborations in Molecular Microbial Ecology Bioinformatics

Jim Cole

Center for Microbial Ecology
Dept. of Plant, Soil & Microbial Sciences
Michigan State University
East Lansing, Michigan U.S.A.



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 7

Rel. 8

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

FunGene v. 0.1

RDP Classifier (OpenSource)

GSC MIGS Metadata

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

GSC Soil Metadata

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

FrameBot

FunGene v. 2

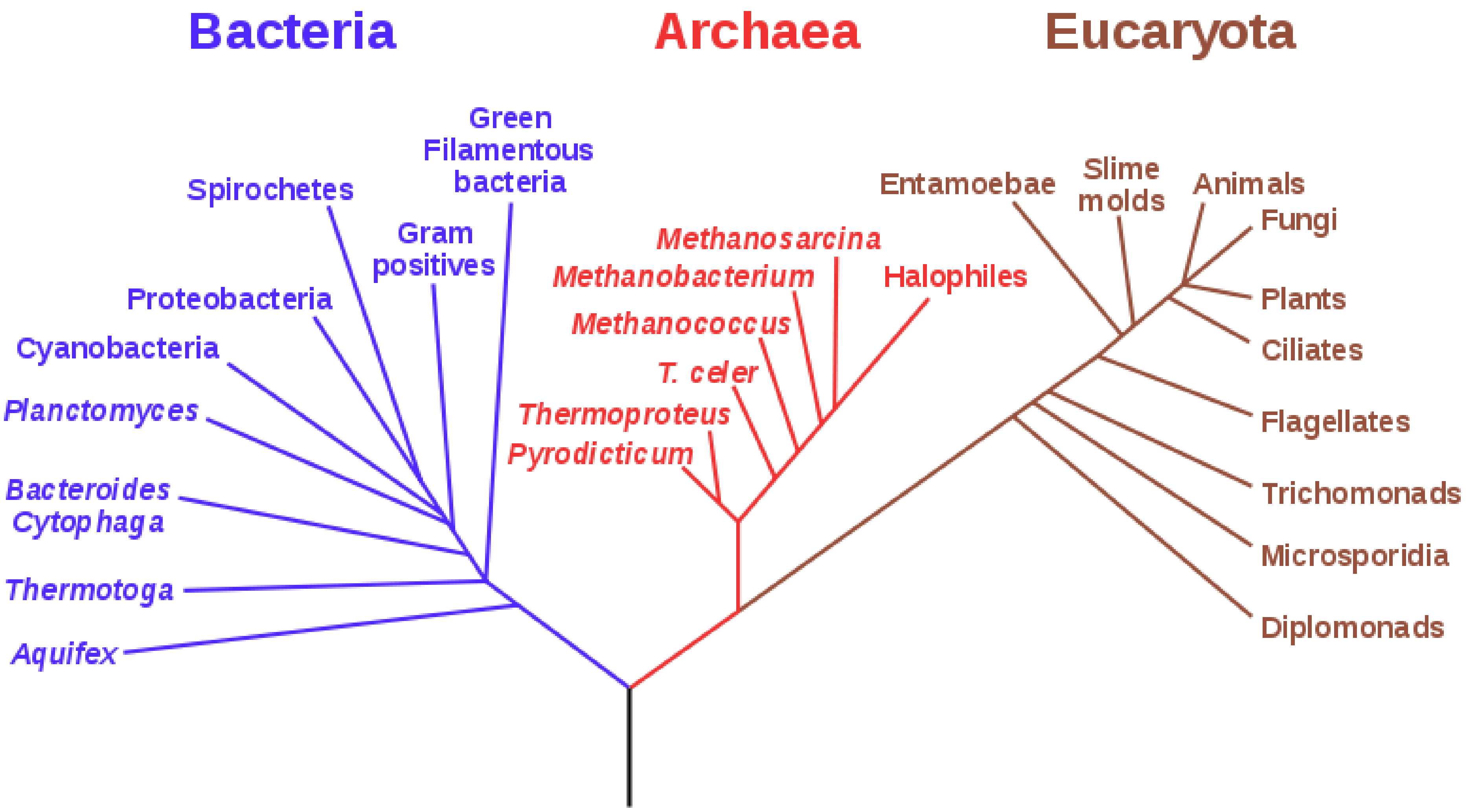
Fungal ITS Classifier

Xander

MiGA

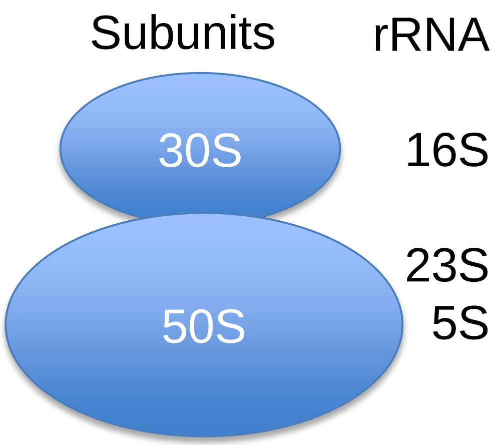
Carl Woese
Gary Olsen

Phylogenetic Tree of Life

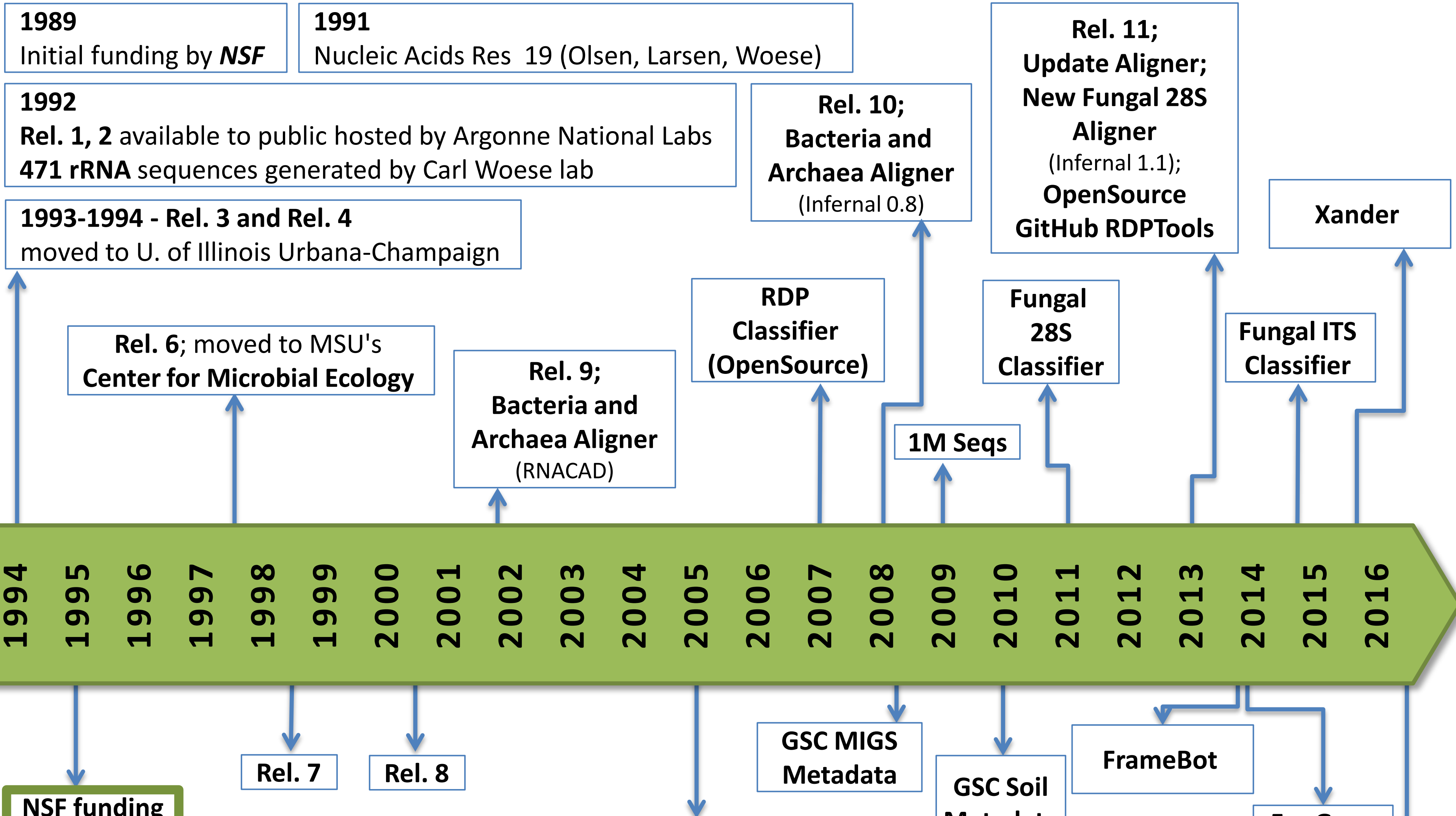


Three domains of life based on the work of Carl Woese and colleagues

Why Ribosomal RNA Sequences



- Ribosomes are the protein synthesis factories.
- Core function present in all cellular organisms.
- Very little evidence of horizontal gene transfer.
- Historically easy to work with.
  - Purify by centrifugation and extract rRNA.
- Now we use PCR to amplify from genomic DNA.
  - rRNA genes have conserved regions interspersed with highly variable regions.
  - Conserved regions used for both PCR primers and sequencing primers.



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

RDP Classifier (OpenSource)

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

Fungal ITS Classifier

Xander

1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

NSF funding
MSU-UIUC

Rel. 7

Rel. 8

FunGene v. 0.1

GSC MIGS Metadata

GSC Soil Metadata

FrameBot

FunGene v. 2

MiGA

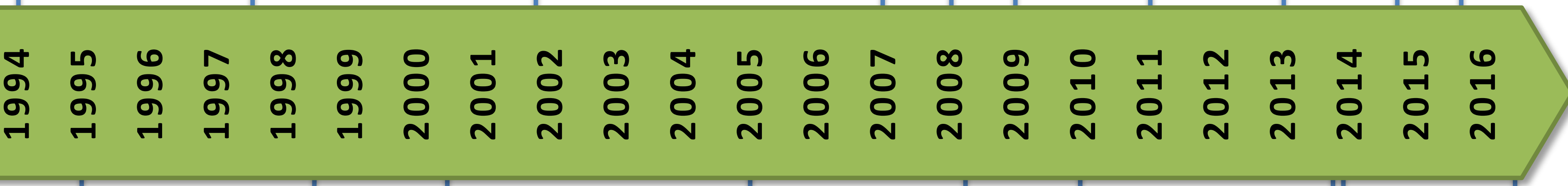
Gary Olsen
Sakti Pramanik

1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign



NSF funding
MSU-UIUC

Rel. 6; moved to MSU's
Center for Microbial Ecology

Rel. 7

Rel. 8

Rel. 9;
Bacteria and
Archaea Aligner
(RNACAD)

FunGene
v. 0.1

RDP
Classifier
(OpenSource)

GSC MIGS
Metadata

Rel. 10;
Bacteria and
Archaea Aligner
(Infernal 0.8)

1M Seqs

GSC Soil
Metadata

Fungal
28S
Classifier

Rel. 11;
Update Aligner;
New Fungal 28S
Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

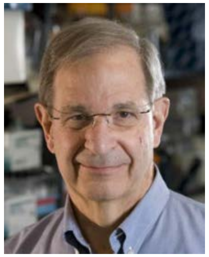
FrameBot

FunGene
v. 2

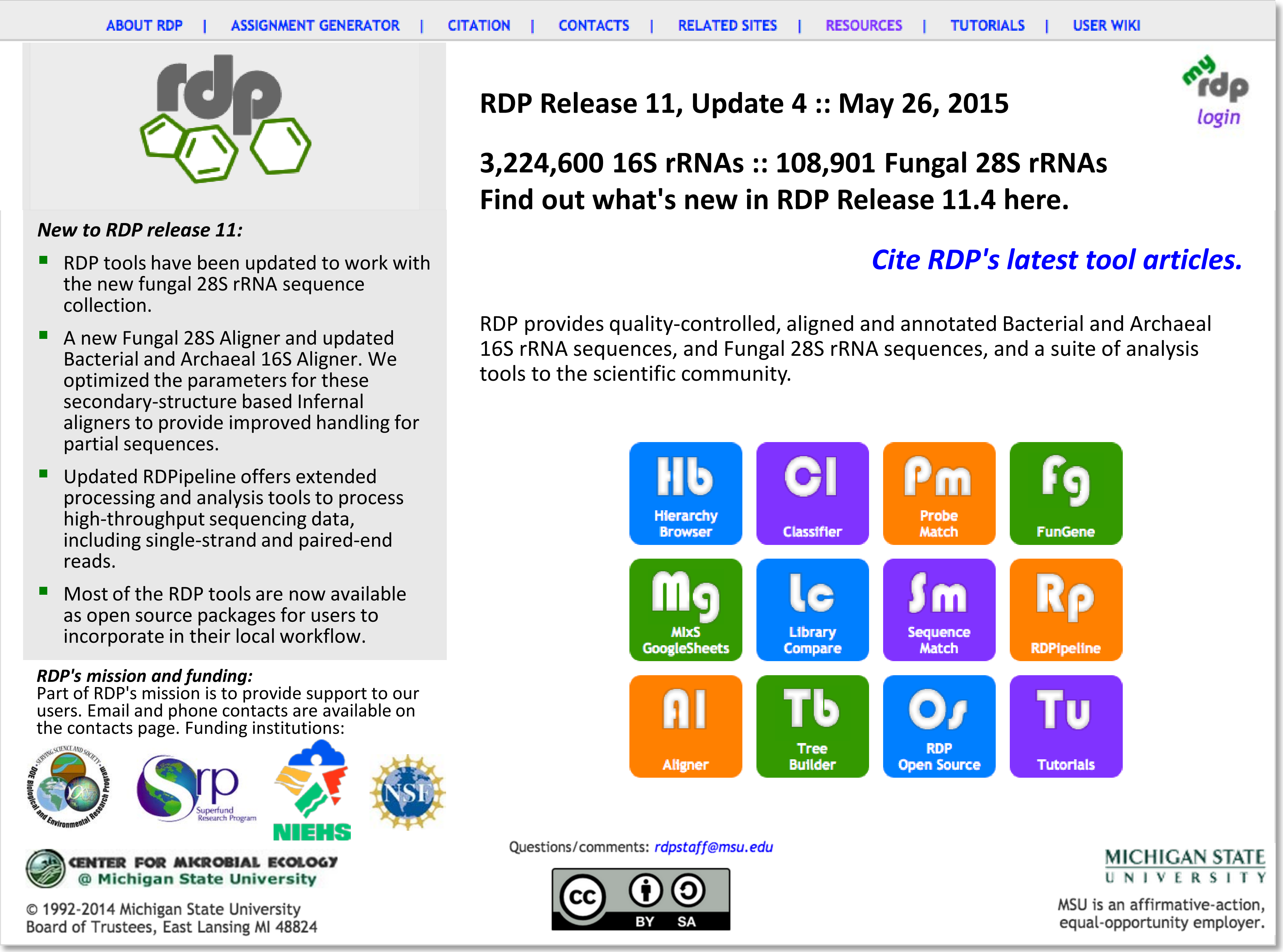
Fungal ITS
Classifier

Xander

MiGA



Jim Tiedje



RDP Release 11, Update 4 :: May 26, 2015

3,224,600 16S rRNAs :: 108,901 Fungal 28S rRNAs
Find out what's new in RDP Release 11.4 here.

Cite RDP's latest tool articles.

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community.

New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.

RDP's mission and funding:
Part of RDP's mission is to provide support to our users. Email and phone contacts are available on the contacts page. Funding institutions:

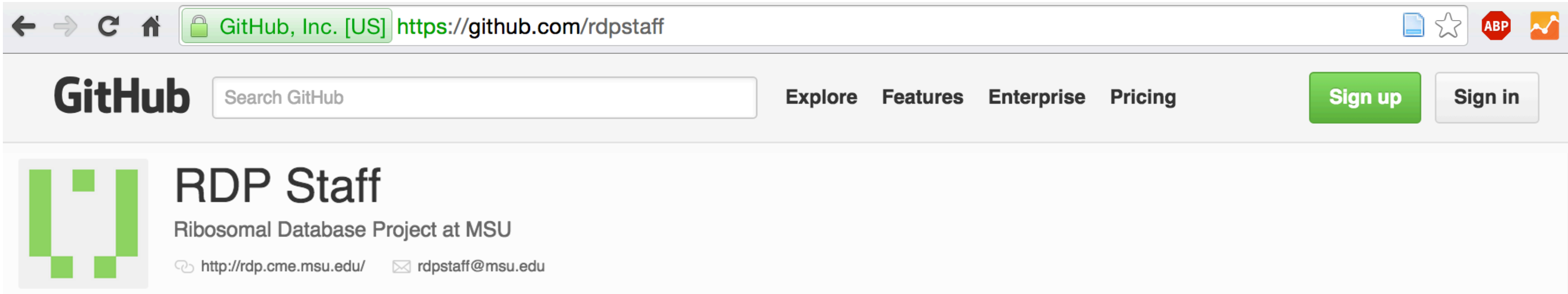
CENTER FOR MICROBIAL ECOLOGY
@ Michigan State University

Questions/comments: *rdpstaff@msu.edu*

MICHIGAN STATE UNIVERSITY

MSU is an affirmative-action, equal-opportunity employer.

http://rdp.cme.msu.edu/



RDP Staff

Ribosomal Database Project at MSU

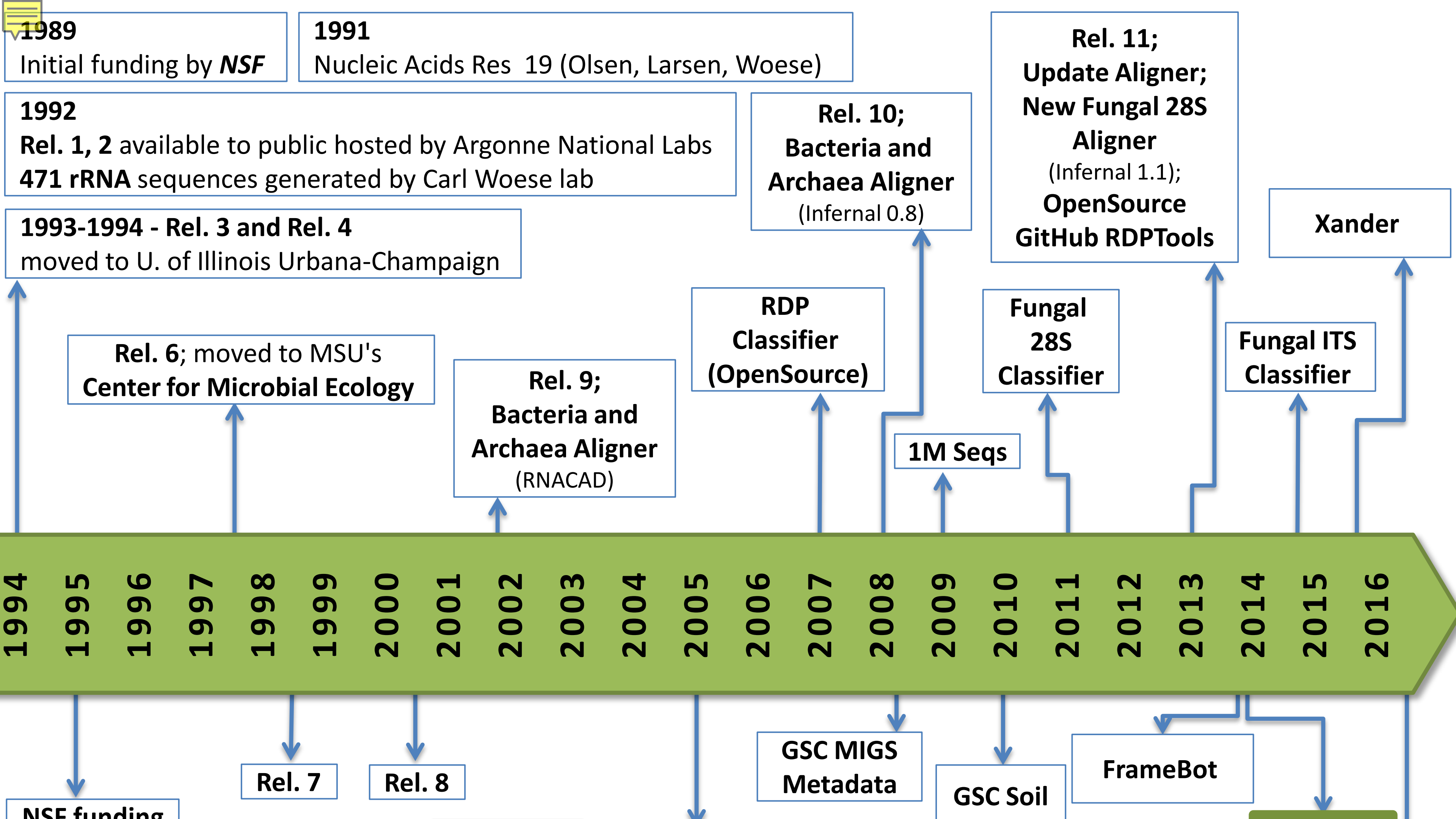
http://rdp.cme.msu.edu/ rdpstaff@msu.edu

RDPTools	Collection of commonly used RDP Tools for high throughput sequence processing and analysis.	Package
classifier	RDP extensible sequence classifier for fungal large subunit rRNA, bacterial & archaeal 16S rRNA.	Java
ReadSeq	Sequence file reader and format converter.	Java
Xander_assembler	A gene-targeted assembler tool for metagenomic sequences.	Shell
AlignmentTools	Tools for pairwise sequence comparison, distance calculation, and hidden markov model sequence scoring (using HMMER3 models).	Java
Framebot	Dynamic programming based frameshift detection and correction tool with nearest neighbor classification.	Java
Clustering	RDP memory-constrained hierarchical clustering tools.	Java
fungene_pipeline	Scripts and resources for analyzing sequence data for select eco-functional genes.	Python
KmerFilter	Tool for kmer analysis.	Java
TaxonomyTree	Taxonomy tree building and traversal utility tool.	Java
SeqFilters	Tool for sorting and selecting nucleotide sequences according to given filters and tags.	Java
SequenceMatch	K-mer based sequence matching tool to calculate nearest neighbors of sequences.	Java
ProbeMatch	Tool for finding (and removing) DNA/RNA primers in sequence reads.	Java
AbundanceStats	Tool for generating various ecological abundance statistics.	Java
FungeneUtils	Package of tools for protein sequence analysis.	Java
SOAP-examples	Code samples from various languages for interacting with RDP soap services.	Perl
gfclassify	A gene family classifier that allows for fast and accurate classification of amplicons (or open reading frame) nucleotide sequences.	C and Biopython

Genes Beyond rRNA

- Faster evolving and single copy phylogenetic markers

- Genes encoding important ecological functions often not phylogenetically coherent.



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's
Center for Microbial Ecology

Rel. 7

Rel. 8

Rel. 9;
Bacteria and
Archaea Aligner
(RNACAD)

FunGene
v. 0.1

RDP
Classifier
(OpenSource)

GSC MIGS
Metadata

Rel. 10;
Bacteria and
Archaea Aligner
(Infernal 0.8)

1M Seqs

GSC Soil
Metadata

Fungal
28S
Classifier

Rel. 11;
Update Aligner;
New Fungal 28S
Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

FrameBot

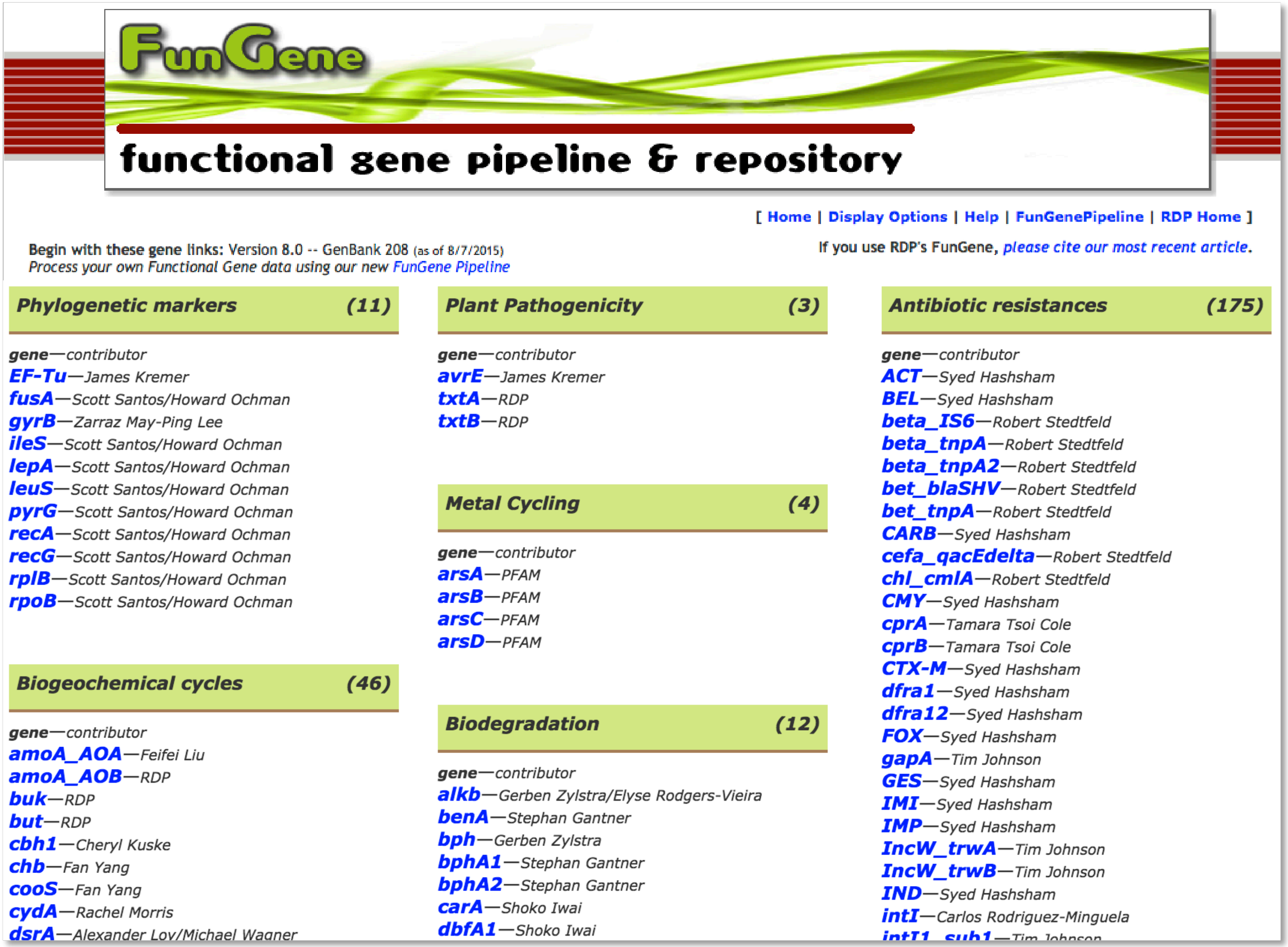
FunGene
v. 2

Fungal ITS
Classifier

Xander

MiGA

Jordan Fish
Benli Chai



FunGene

functional gene pipeline & repository

[Home | Display Options | Help | FunGenePipeline | RDP Home]

If you use RDP's FunGene, *please cite our most recent article*.

Begin with these gene links: Version 8.0 -- GenBank 208 (as of 8/7/2015)
Process your own Functional Gene data using our new FunGene Pipeline

Phylogenetic markers (11)

gene—*contributor*
EF-Tu—*James Kremer*
fusA—*Scott Santos/Howard Ochman*
gyrB—*Zarraz May-Ping Lee*
ileS—*Scott Santos/Howard Ochman*
lepA—*Scott Santos/Howard Ochman*
leuS—*Scott Santos/Howard Ochman*
pyrG—*Scott Santos/Howard Ochman*
recA—*Scott Santos/Howard Ochman*
recG—*Scott Santos/Howard Ochman*
rplB—*Scott Santos/Howard Ochman*
rpoB—*Scott Santos/Howard Ochman*

Biogeochemical cycles (46)

gene—*contributor*
amoA_AOA—*Feifei Liu*
amoA_AOB—*RDP*
buk—*RDP*
but—*RDP*
cbh1—*Cheryl Kuske*
chb—*Fan Yang*
cooS—*Fan Yang*
cydA—*Rachel Morris*
dsrA—*Alexander Loy/Michael Wagner*

Plant Pathogenicity (3)

gene—*contributor*
avrE—*James Kremer*
txtA—*RDP*
txtB—*RDP*

Metal Cycling (4)

gene—*contributor*
arsA—*PFAM*
arsB—*PFAM*
arsC—*PFAM*
arsD—*PFAM*

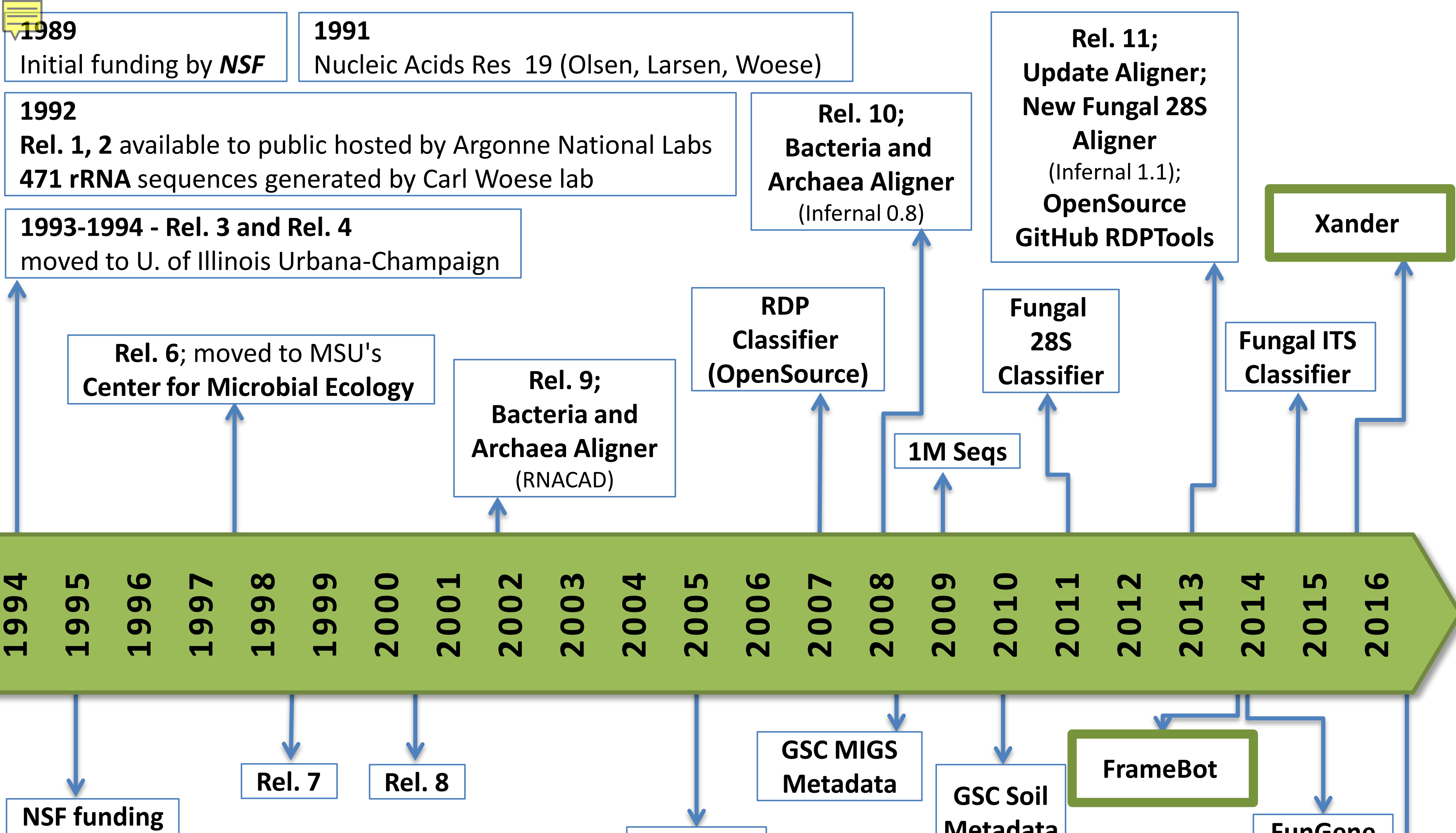
Biodegradation (12)

gene—*contributor*
alkb—*Gerben Zylstra/Elyse Rodgers-Vieira*
benA—*Stephan Gantner*
bph—*Gerben Zylstra*
bphA1—*Stephan Gantner*
bphA2—*Stephan Gantner*
carA—*Shoko Iwai*
dbfA1—*Shoko Iwai*

Antibiotic resistances (175)

gene—*contributor*
ACT—*Syed Hashsham*
BEL—*Syed Hashsham*
beta_IS6—*Robert Stedtfeld*
beta_tnpA—*Robert Stedtfeld*
beta_tnpA2—*Robert Stedtfeld*
bet_blaSHV—*Robert Stedtfeld*
bet_tnpA—*Robert Stedtfeld*
CARB—*Syed Hashsham*
cefa_qacEdelta—*Robert Stedtfeld*
chl_cmlA—*Robert Stedtfeld*
CMY—*Syed Hashsham*
cprA—*Tamara Tsoi Cole*
cprB—*Tamara Tsoi Cole*
CTX-M—*Syed Hashsham*
dfra1—*Syed Hashsham*
dfra12—*Syed Hashsham*
FOX—*Syed Hashsham*
gapA—*Tim Johnson*
GES—*Syed Hashsham*
IMI—*Syed Hashsham*
IMP—*Syed Hashsham*
IncW_trwA—*Tim Johnson*
IncW_trwB—*Tim Johnson*
IND—*Syed Hashsham*
intI—*Carlos Rodriguez-Minguela*
intI1_sul1—*Tim Johnson*

htpp://fungene.cme.msu.edu



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

RDP Classifier (OpenSource)

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

Fungal ITS Classifier

Xander

NSF funding MSU-UIUC

Rel. 7

Rel. 8

FunGene v. 0.1

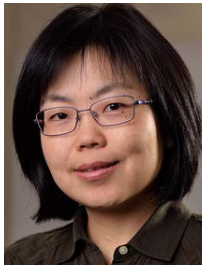
GSC MIGS Metadata

GSC Soil Metadata

FrameBot

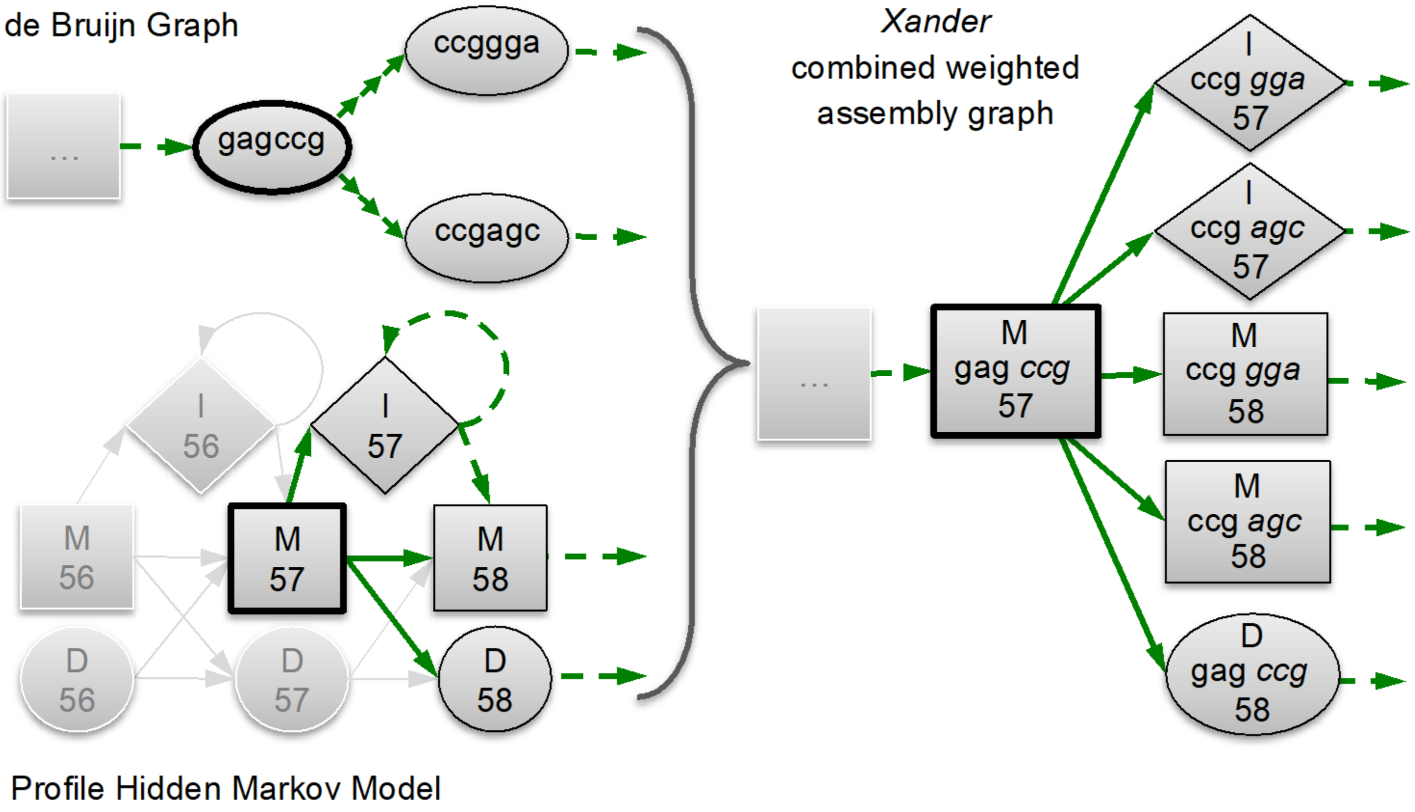
FunGene v. 2

MiGA

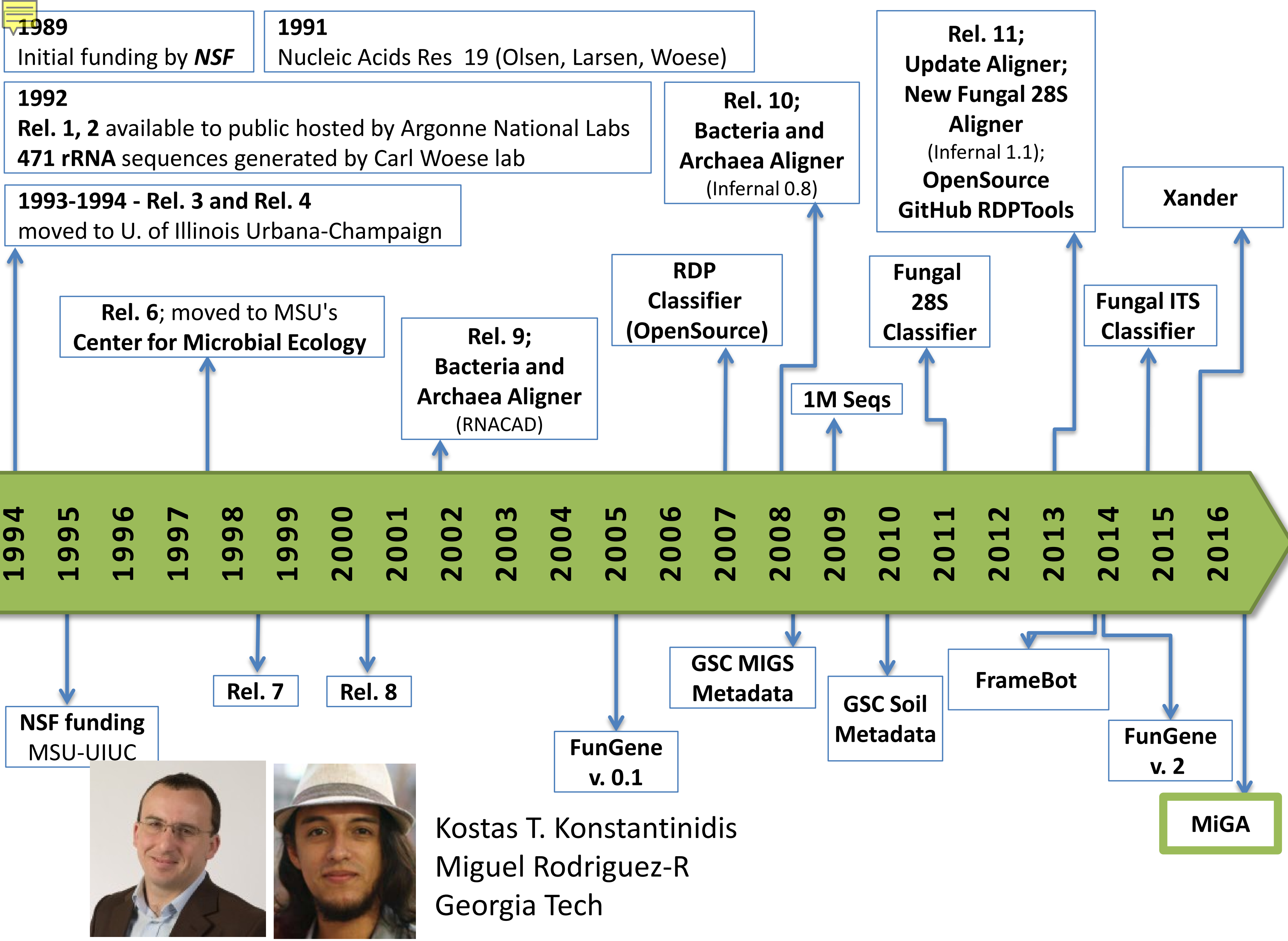


Yanni Sun
Qiong Wang

Xander: Gene-Targeted Assembler Combining de Bruijn Graph and HMM



Wang et al., (2015) Microbiome3:32



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 7

Rel. 8

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

FunGene v. 0.1

RDP Classifier (OpenSource)

GSC MIGS Metadata

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

GSC Soil Metadata

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

FrameBot

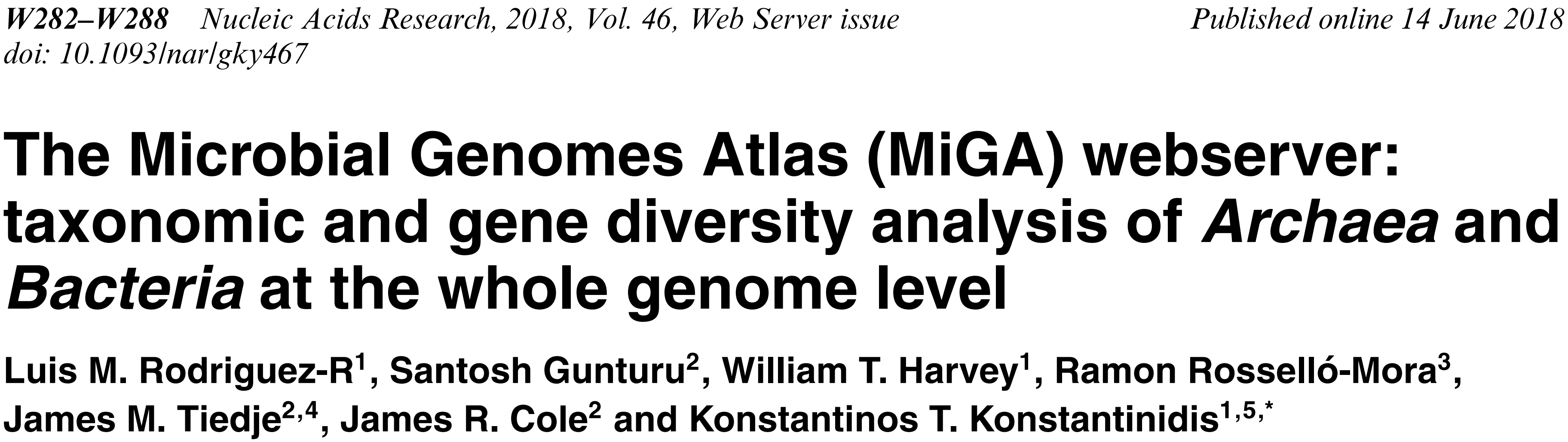
Fungal ITS Classifier

FunGene v. 2

Xander

MiGA

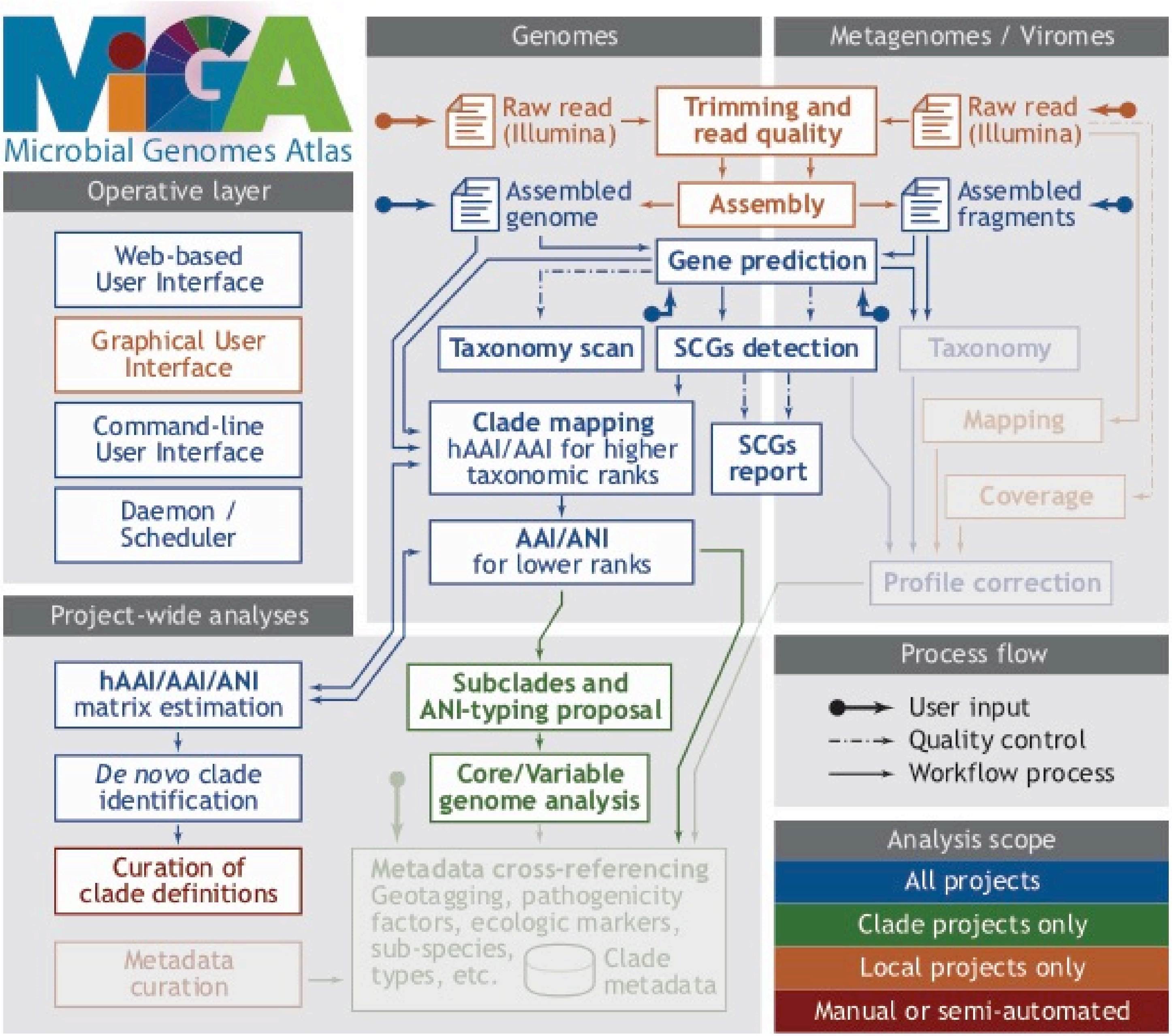
Kostas T. Konstantinidis
Miguel Rodriguez-R
Georgia Tech



W282–W288 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 14 June 2018*
doi: 10.1093/nar/gky467

The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level

Luis M. Rodriguez-R[1], Santosh Gunturu[2], William T. Harvey[1], Ramon Rosselló-Mora[3], James M. Tiedje[2,4], James R. Cole[2] and Konstantinos T. Konstantinidis[1,5,*]



Microbial Genomes Atlas

Operative layer

Web-based User Interface

Graphical User Interface

Command-line User Interface

Daemon / Scheduler

Project-wide analyses

hAAI/AAI/ANI matrix estimation

De novo clade identification

Curation of clade definitions

Metadata curation

Genomes

Raw read (Illumina)

Trimming and read quality

Assembled genome

Assembly

Gene prediction

Taxonomy scan

SCGs detection

Clade mapping hAAI/AAI for higher taxonomic ranks

SCGs report

AAI/ANI for lower ranks

Subclades and ANI-typing proposal

Core/Variable genome analysis

Metadata cross-referencing
Geotagging, pathogenicity factors, ecologic markers, sub-species, types, etc.

Clade metadata

Metagenomes / Viromes

Raw read (Illumina)

Assembled fragments

Taxonomy

Mapping

Coverage

Profile correction

Process flow

●→ User input
-·-·→ Quality control
——→ Workflow process

Analysis scope

All projects

Clade projects only

Local projects only

Manual or semi-automated

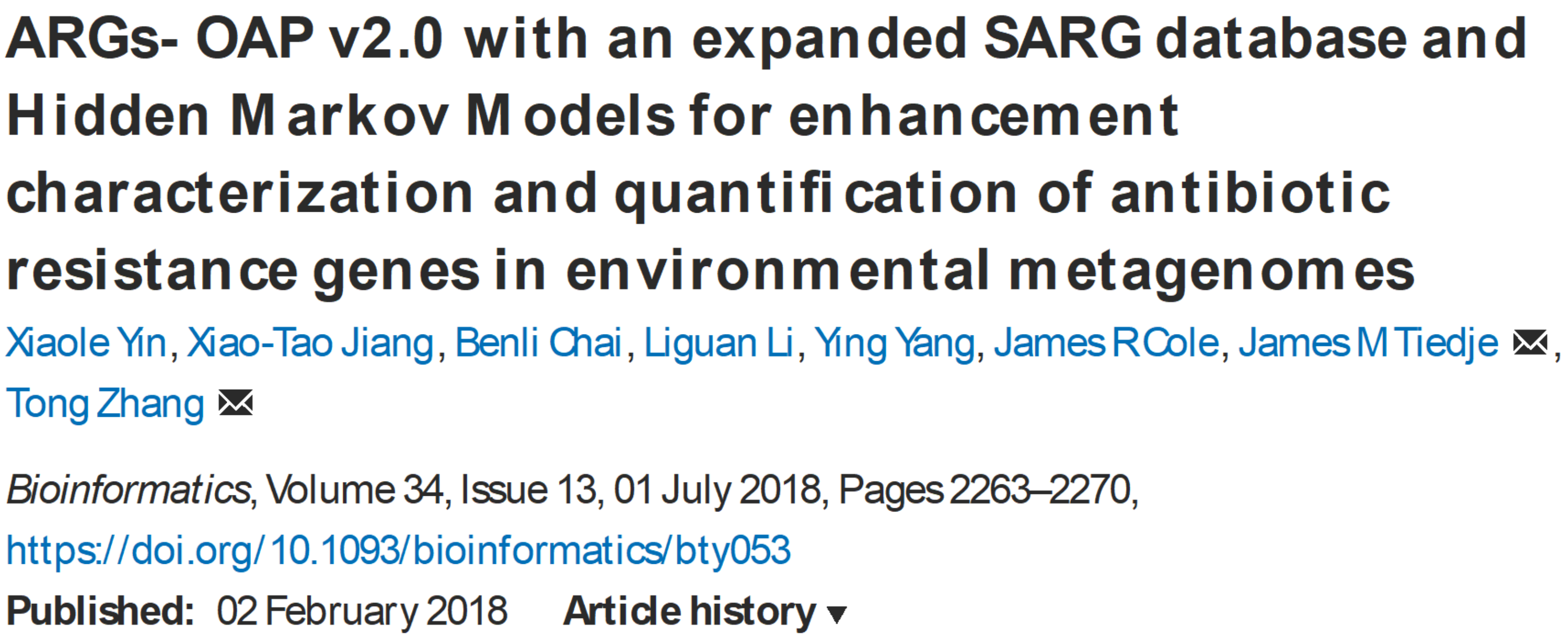
Chief workflow (v4, Jul 2015)

Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity

Nonpareil Collaboration

- GaTech
  - Initial Implementation
  - Applications
- MSU
  - Speed up 300-fold
  - Harden for release

Environmental Antibiotic Resistance



ARGs- OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes

Xiaole Yin, Xiao-Tao Jiang, Benli Chai, Liguan Li, Ying Yang, James R Cole, James M Tiedje ✉, Tong Zhang ✉

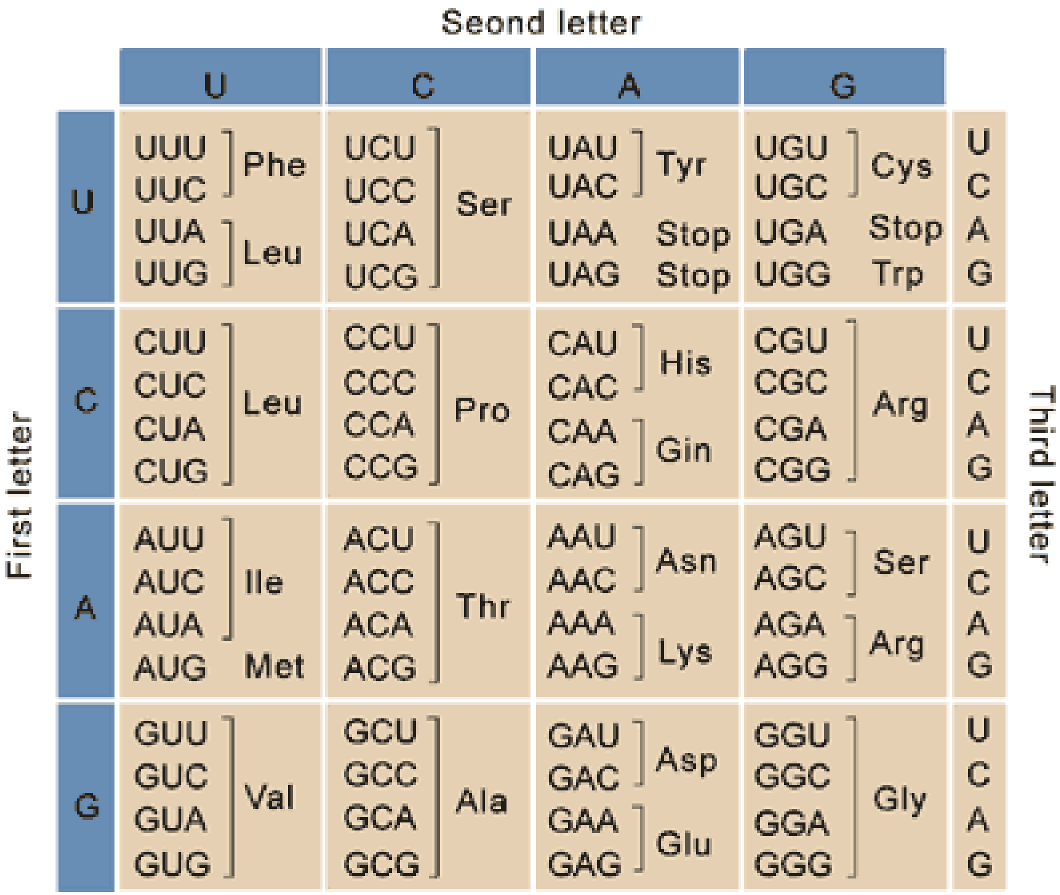
Bioinformatics, Volume 34, Issue 13, 01 July 2018, Pages 2263–2270, https://doi.org/10.1093/bioinformatics/bty053

Published: 02 February 2018 **Article history** ▾

Environmental Antibiotic Resistance Our Role:

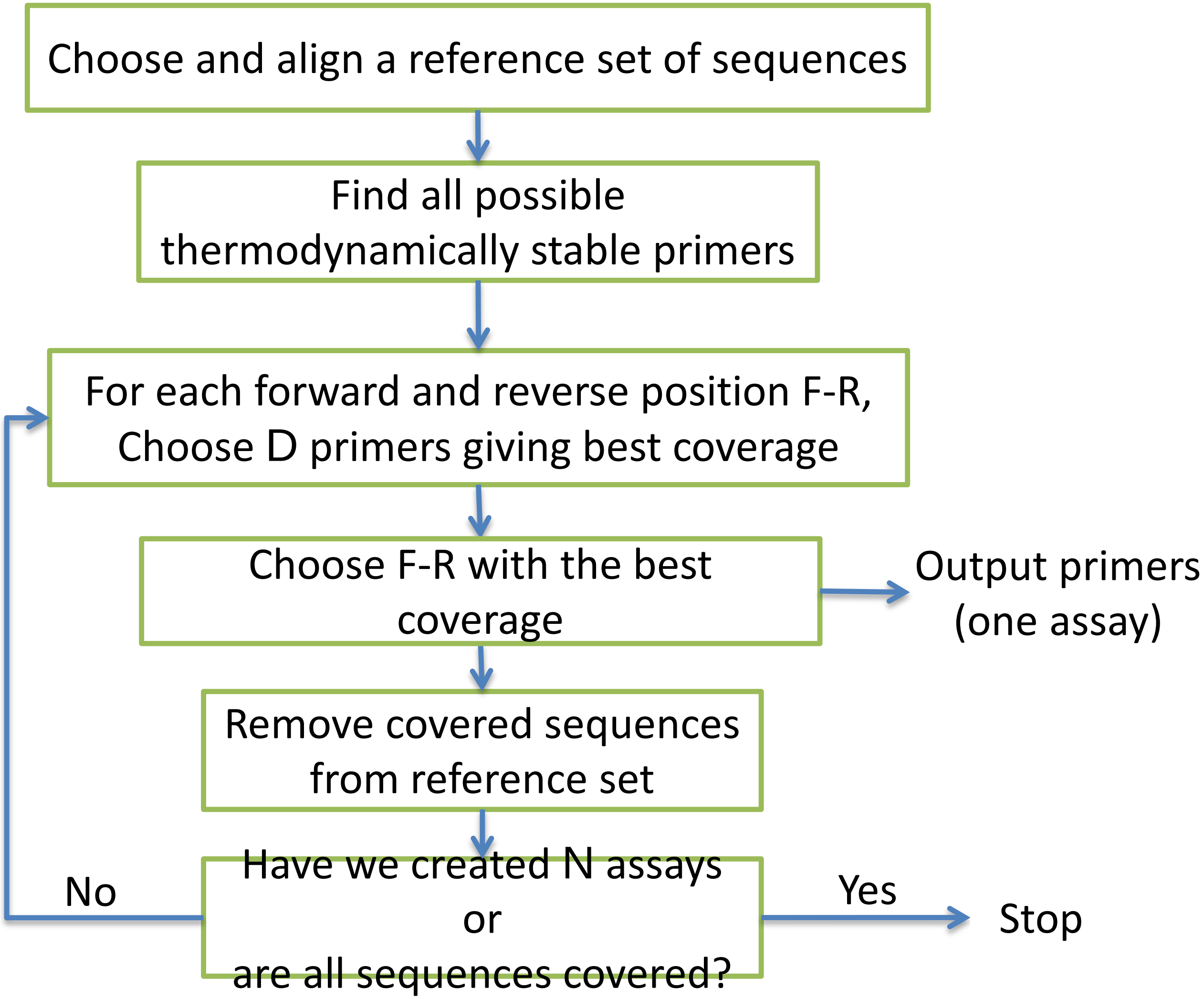
- Develop tools to test consistency of sequences in antibiotic types and subtypes and to assure clear demarcation subtypes.
- Help develop strategy and tools for detecting antibiotic ARGS in environmental metagenomes.

Reminder:
DNA less conserved than Protein



First letter	Seond letter								Third letter
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

EcoFunPrimer – Primer Design



Choose and align a reference set of sequences

Find all possible thermodynamically stable primers

For each forward and reverse position F-R, Choose D primers giving best coverage

Choose F-R with the best coverage

Output primers (one assay)

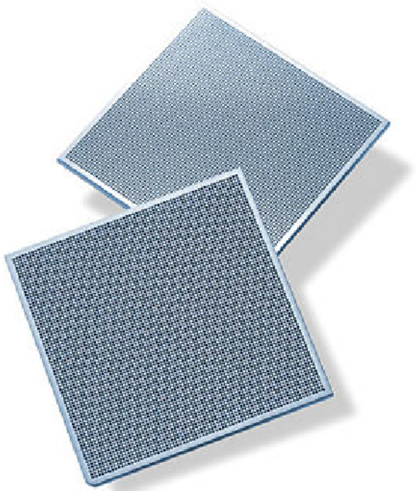
Remove covered sequences from reference set

Have we created N assays or are all sequences covered?

No

Yes

Stop



Smart Chip: 5184 well qPCR platform

Acknowledgements

http://rdp.cme.msu.edu http://fungene.cme.msu.edu

Collaborators:

C. Titus Brown (UCD)
George Garrity (MSU)
John Quensen (MSU)
Sakti Pramanik (MSU)
Tom Schmidt (UM)
Yanni Sun (CityU)
Jim Tiedje (MSU)
Kostas Konstantinidis (GaTech)
Miguel Rodrigues R. (GaTech)
Tong Zhang (HKU)

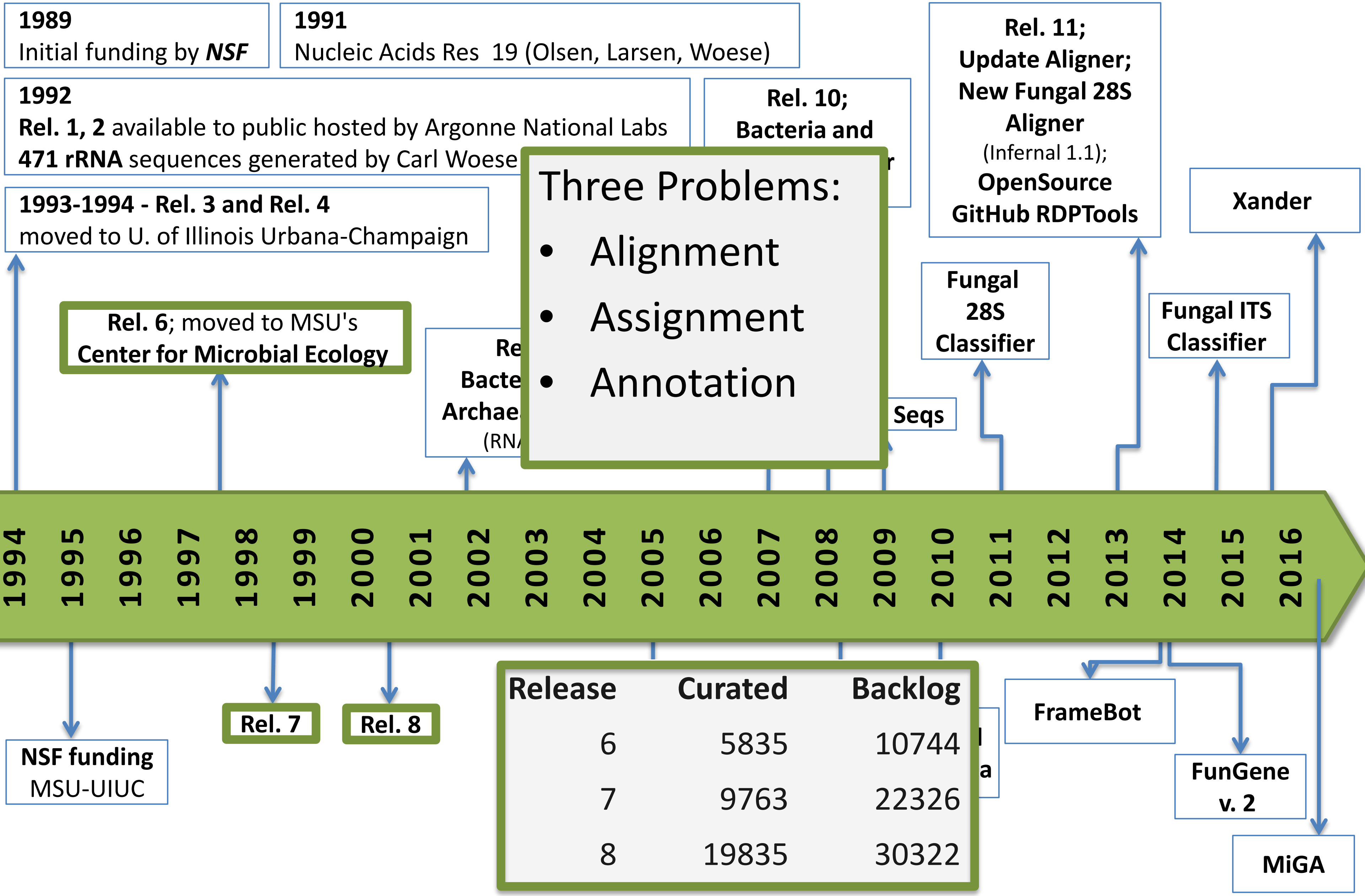
RDP Bioinformatics Group:

Benli Chai (Swift Biosci)
Alex Fall
Jordan Fish (Google)
Mariah Gilman (Google)
Santosh Gunturu
Donna McGarrell
Michael Okeefe (MSU BCH)
Leo Tift
Qiong Wang (DuPont/IB)
Ziye Xing (UCLA)
Tae-Kwon Lee (U Vienna)
All Past RDP Group members

All Current and Past Collaborators

Support from DOE, NSF, NIH, USDA, NIEHS

The End



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

Rel. 6; moved to MSU's
Center for Microbial Ecology

Re
Bacte
Archae
(RN

Rel. 10;
Bacteria and
r

Seqs

Rel. 11;
Update Aligner;
New Fungal 28S Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

Fungal 28S Classifier

Fungal ITS Classifier

Xander

Three Problems:

- Alignment
- Assignment
- Annotation

NSF funding
MSU-UIUC

Rel. 7

Rel. 8

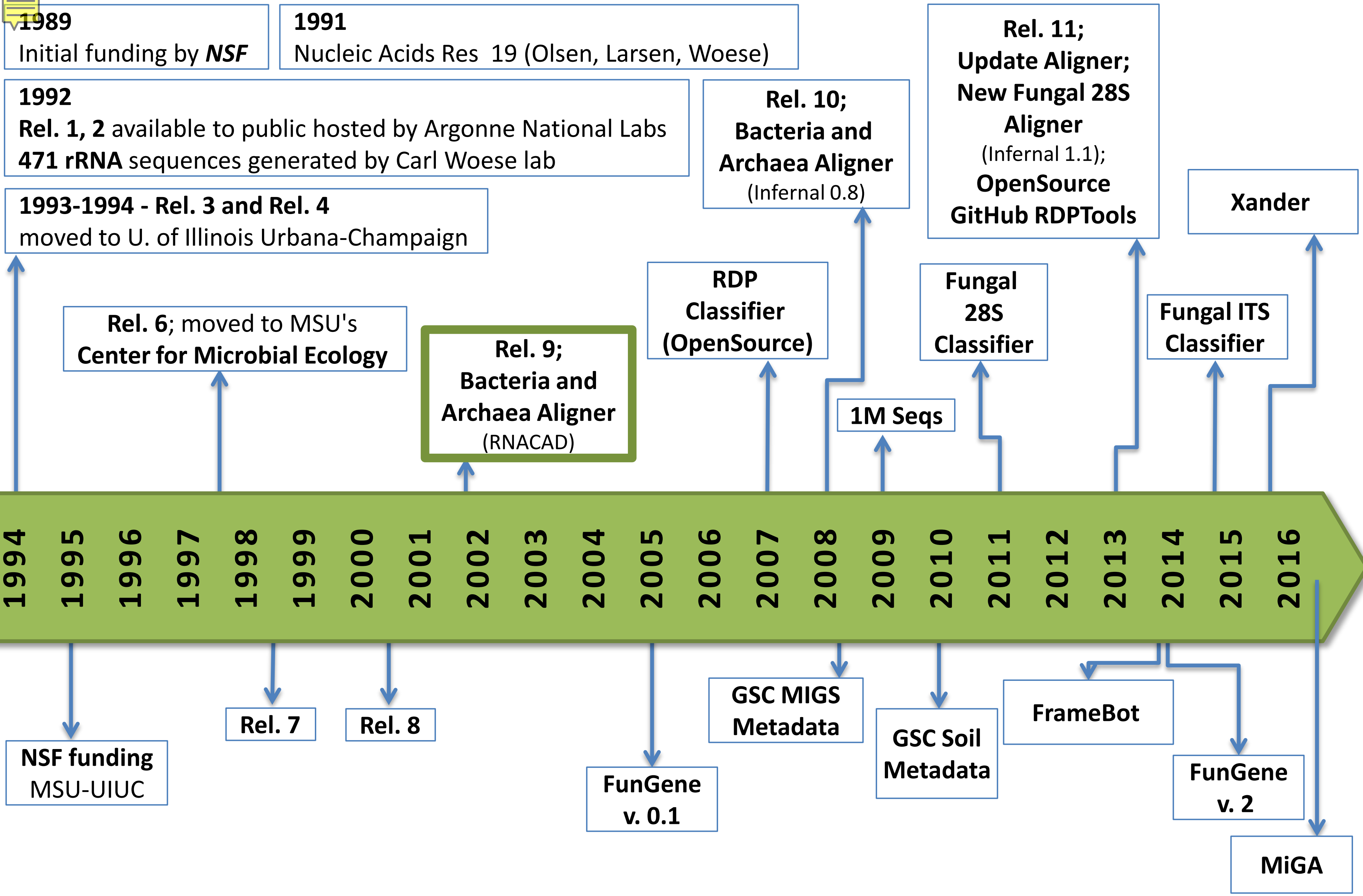
Release	Curated	Backlog
6	5835	10744
7	9763	22326
8	19835	30322

l
a

FrameBot

FunGene v. 2

MiGA



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

RDP Classifier (OpenSource)

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

Fungal ITS Classifier

Xander

NSF funding
MSU-UIUC

Rel. 7

Rel. 8

FunGene v. 0.1

GSC MIGS Metadata

GSC Soil Metadata

FrameBot

FunGene v. 2

MiGA

M.S.P. Brown (2000) ISMB Proceedings

Secondary structure of small-subunit ribosomal RNA

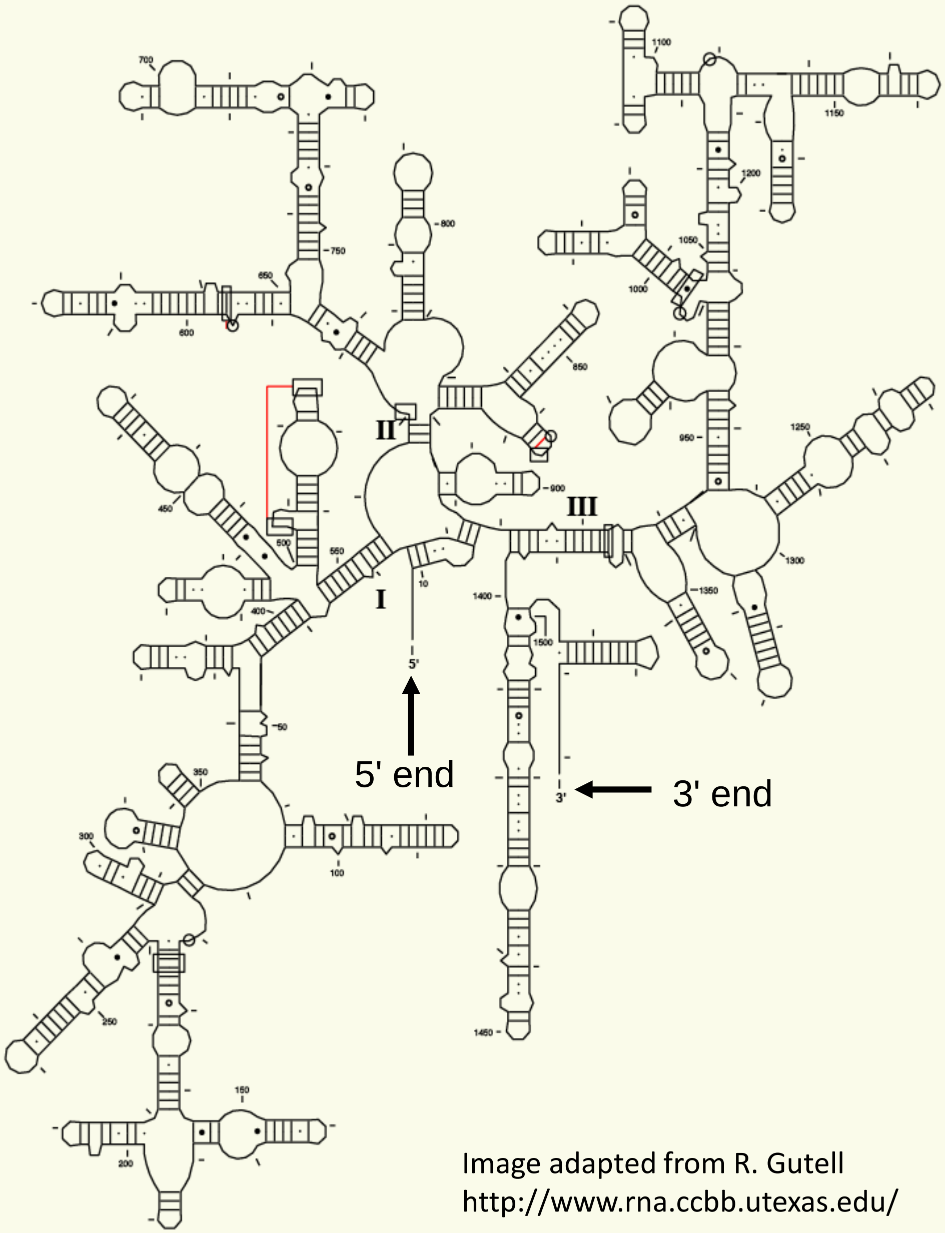


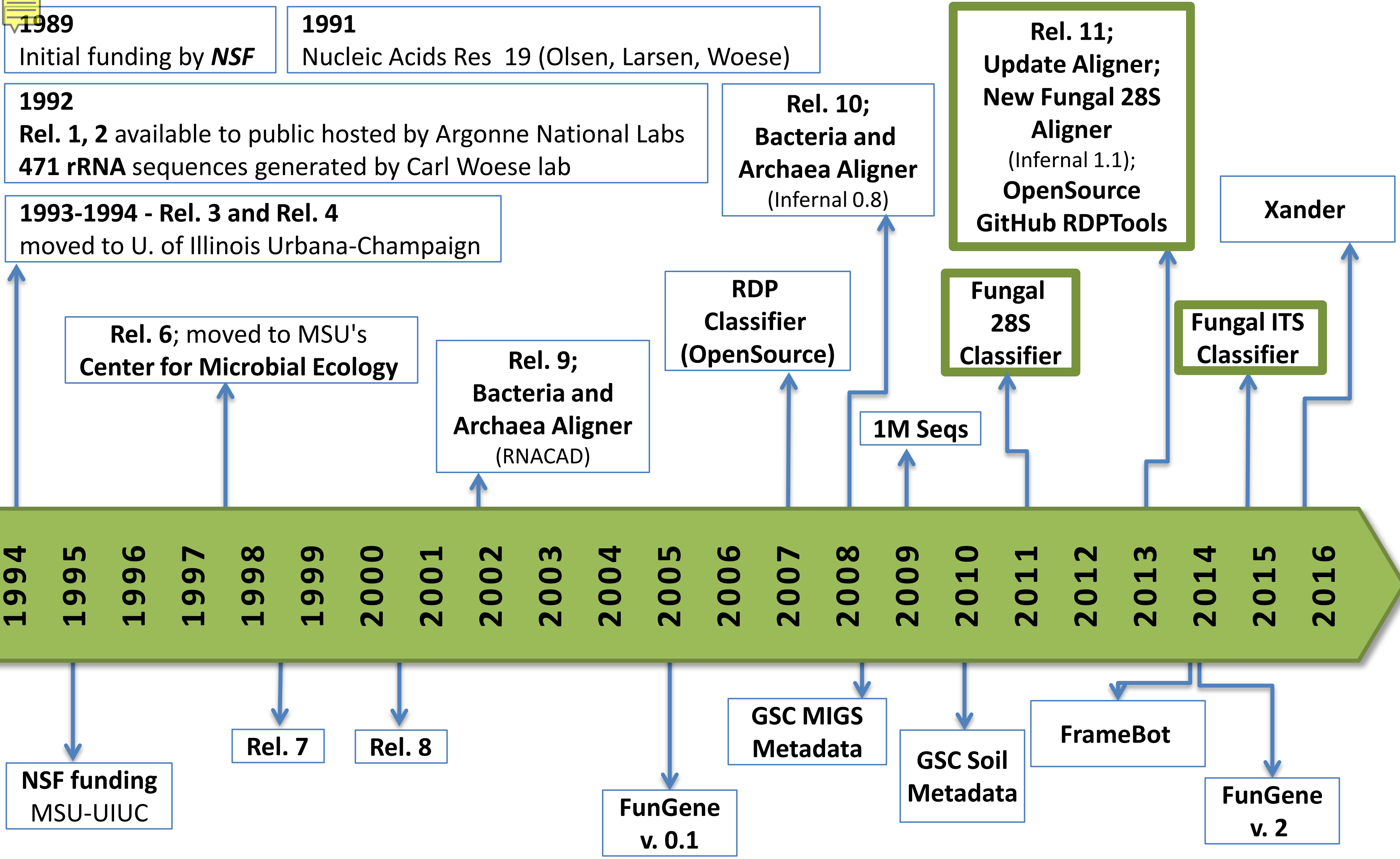
Image adapted from R. Gutell
http://www.rna.ccbb.utexas.edu/

Microbiome

Wang et al. (2015) 3:32

Xander

Gene-Targeted Metagenome Assembly



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's
Center for Microbial Ecology

Rel. 7

Rel. 8

Rel. 9;
Bacteria and
Archaea Aligner
(RNACAD)

FunGene
v. 0.1

RDP
Classifier
(OpenSource)

GSC MIGS
Metadata

Rel. 10;
Bacteria and
Archaea Aligner
(Infernal 0.8)

1M Seqs

GSC Soil
Metadata

Fungal
28S
Classifier

Rel. 11;
Update Aligner;
New Fungal 28S
Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

FrameBot

Fungal ITS
Classifier

FunGene
v. 2

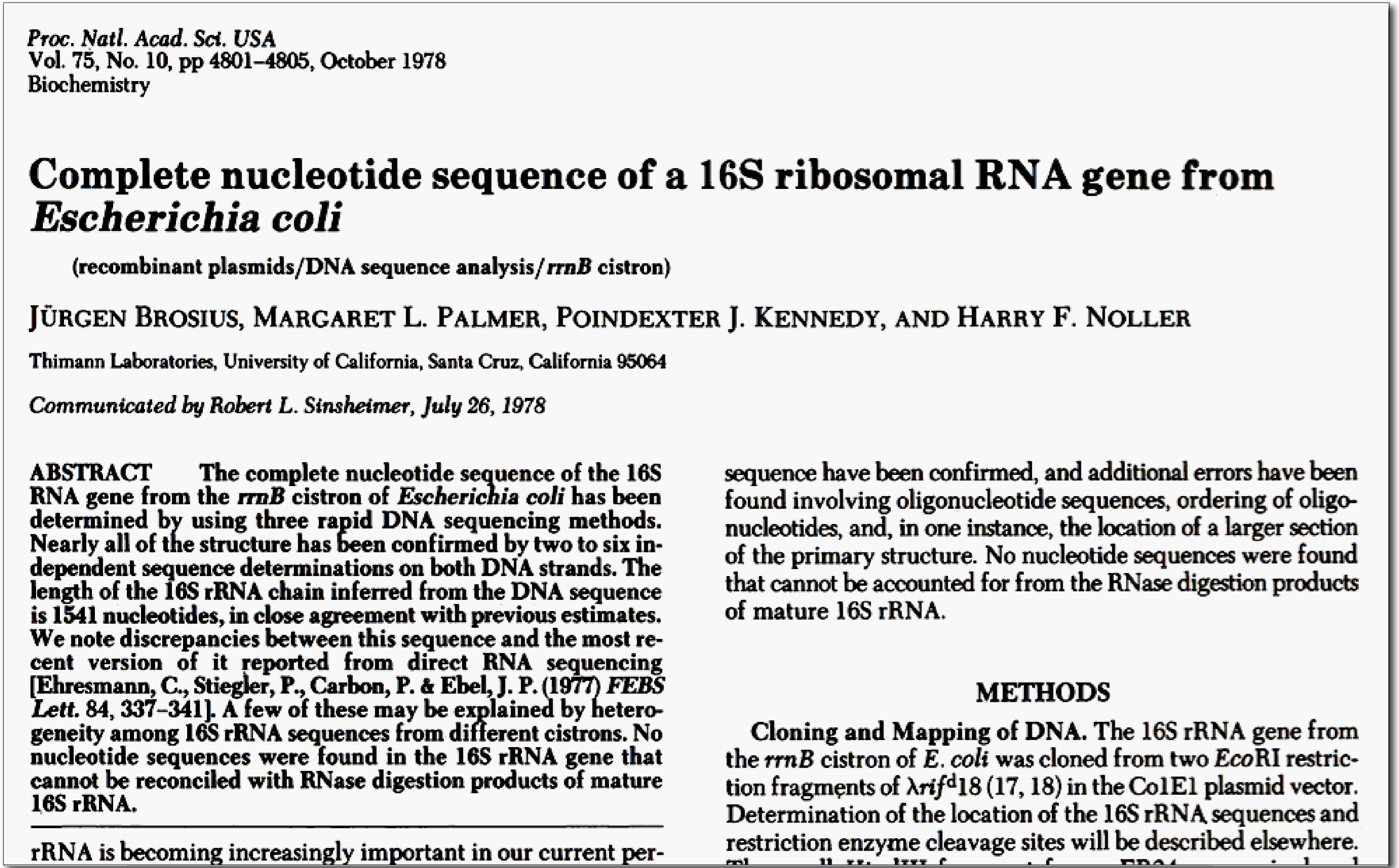
Xander

http://rdp.cme.msu.edu/

Major rRNA Databases

- Ribosomal Database Project
  - Dept. of Plant Soil and Microbial Sciences
- Arb/Silva
  - Max Planck Inst. For Oceanography
- GreenGenes
  - UCSD Biomedical Sciences
- Human Oral Microbiome Database
  - Forsyth Institute

36 Years of rRNA gene Sequencing



Proc. Natl. Acad. Sci. USA
Vol. 75, No. 10, pp 4801–4805, October 1978
Biochemistry

Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*

(recombinant plasmids/DNA sequence analysis/*rrnB* cistron)

JÜRGEN BROSIUS, MARGARET L. PALMER, POINDEXTER J. KENNEDY, AND HARRY F. NOLLER

Thimann Laboratories, University of California, Santa Cruz, California 95064

Communicated by Robert L. Sinsheimer, July 26, 1978

ABSTRACT The complete nucleotide sequence of the 16S RNA gene from the *rrnB* cistron of *Escherichia coli* has been determined by using three rapid DNA sequencing methods. Nearly all of the structure has been confirmed by two to six independent sequence determinations on both DNA strands. The length of the 16S rRNA chain inferred from the DNA sequence is 1541 nucleotides, in close agreement with previous estimates. We note discrepancies between this sequence and the most recent version of it reported from direct RNA sequencing [Ehresmann, C., Stiegler, P., Carbon, P. & Ebel, J. P. (1977) *FEBS Lett.* 84, 337–341]. A few of these may be explained by heterogeneity among 16S rRNA sequences from different cistrons. No nucleotide sequences were found in the 16S rRNA gene that cannot be reconciled with RNase digestion products of mature 16S rRNA.

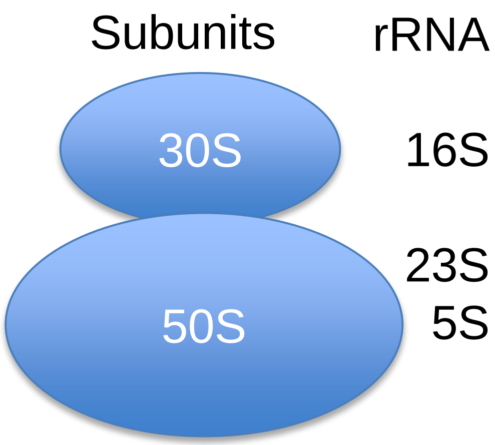
rRNA is becoming increasingly important in our current per-

sequence have been confirmed, and additional errors have been found involving oligonucleotide sequences, ordering of oligonucleotides, and, in one instance, the location of a larger section of the primary structure. No nucleotide sequences were found that cannot be accounted for from the RNase digestion products of mature 16S rRNA.

METHODS

Cloning and Mapping of DNA. The 16S rRNA gene from the *rrnB* cistron of *E. coli* was cloned from two *Eco*RI restriction fragments of $\lambda rif^{d}18$ (17, 18) in the ColE1 plasmid vector. Determination of the location of the 16S rRNA sequences and restriction enzyme cleavage sites will be described elsewhere.

Why Ribosomal RNA Sequences



Subunits rRNA

30S 16S

50S 23S 5S

- Easy to purify ribosomes and rRNA species by sedimentation.
- Now we use PCR with primers targeting conserved regions to amplify rRNA genes.

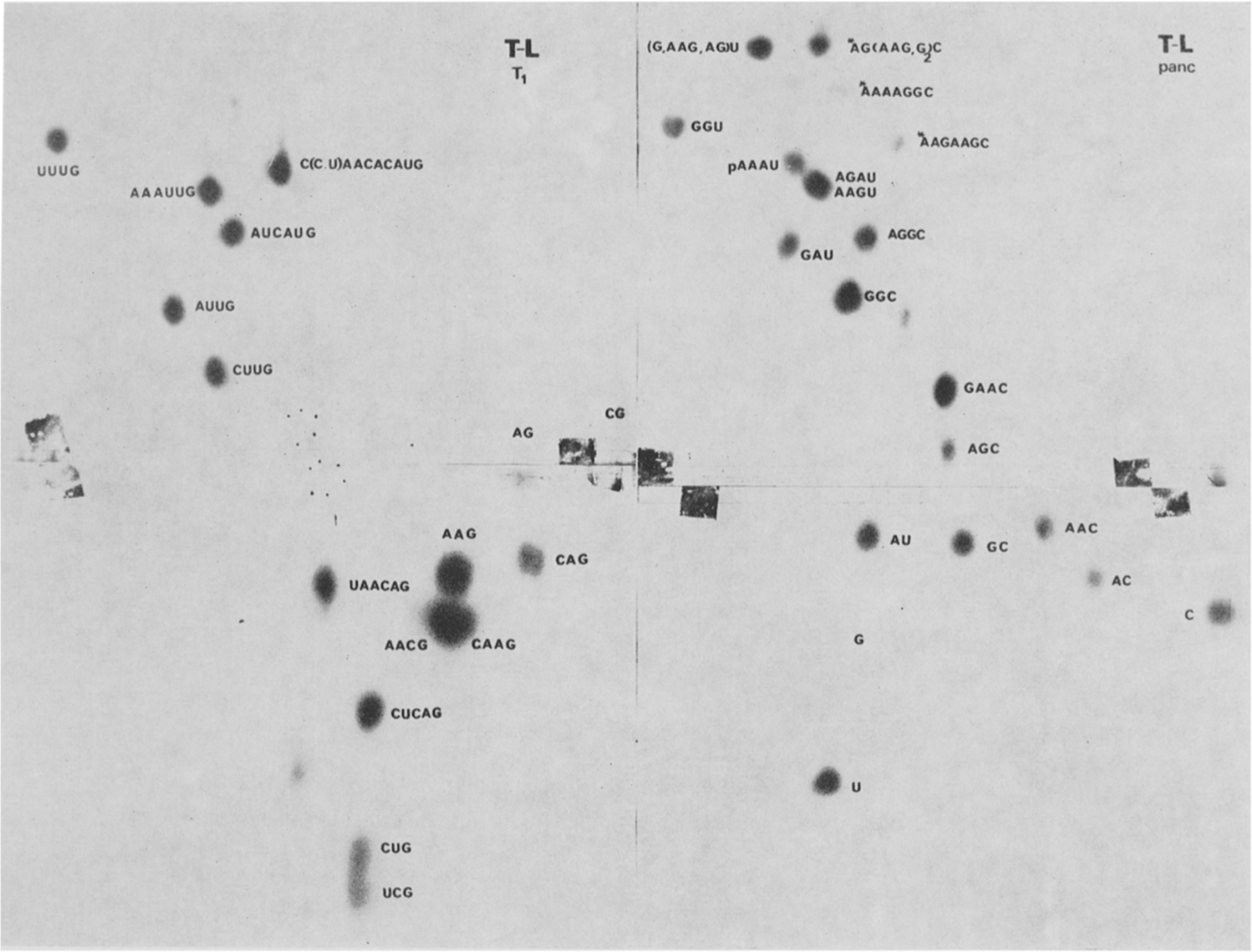


FIG. 2. — *T*. (+ alkaline phosphatase) *and pancreatic ribonuclease fingerprints of section L*. This fragment was obtained from the digestion of intact 30 S subunits (see figure 1). Some degradation products (AUG and C(C. U)AACAC) arising from spot C(C. U)AACACAUG can be detected in the sample.

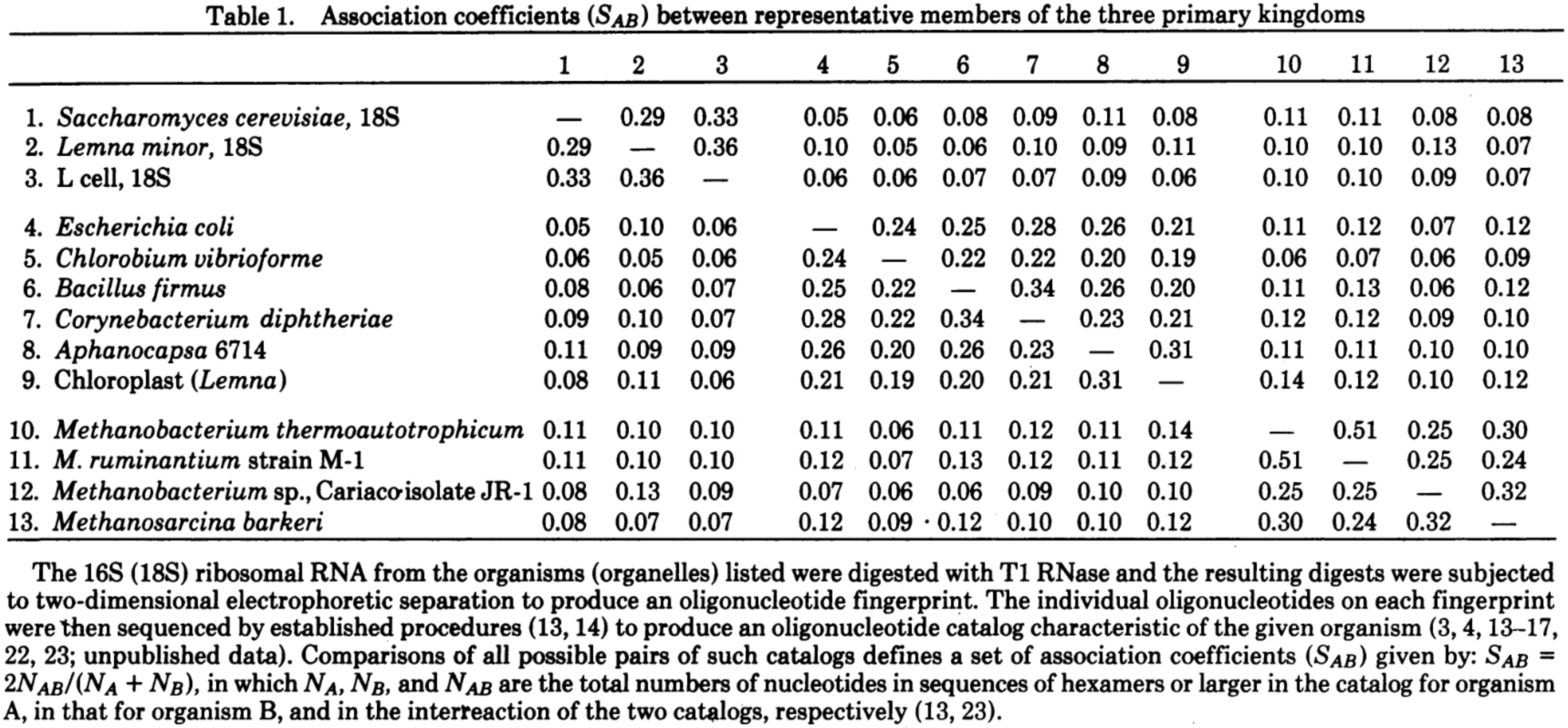
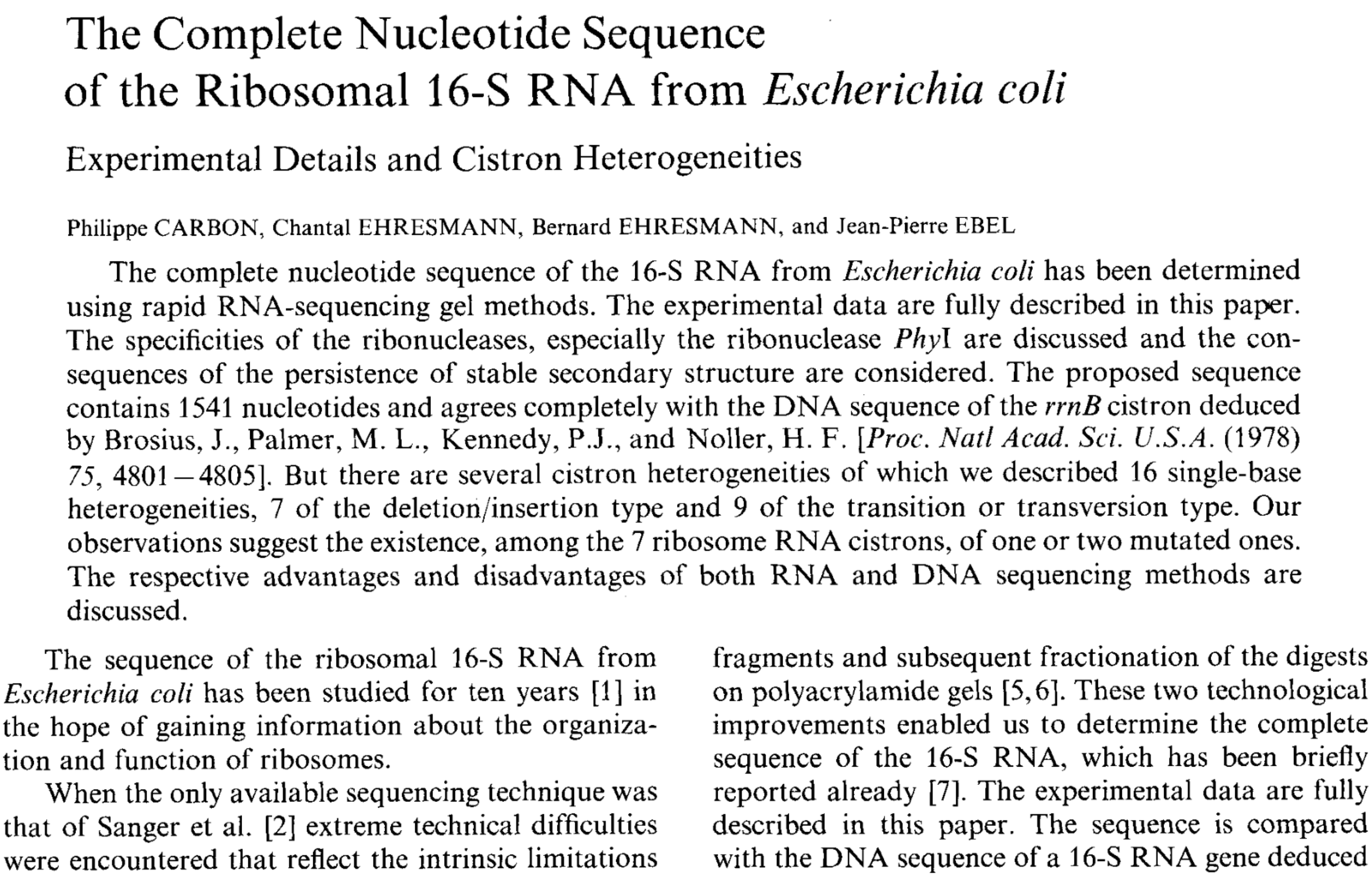


Table 1. Association coefficients (S_{AB}) between representative members of the three primary kingdoms

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *Saccharomyces cerevisiae*, 18S	—	0.29	0.33	0.05	0.06	0.08	0.09	0.11	0.08	0.11	0.11	0.08	0.08
2. *Lemna minor*, 18S	0.29	—	0.36	0.10	0.05	0.06	0.10	0.09	0.11	0.10	0.10	0.13	0.07
3. L cell, 18S	0.33	0.36	—	0.06	0.06	0.07	0.07	0.09	0.06	0.10	0.10	0.09	0.07
4. *Escherichia coli*	0.05	0.10	0.06	—	0.24	0.25	0.28	0.26	0.21	0.11	0.12	0.07	0.12
5. *Chlorobium vibrioforme*	0.06	0.05	0.06	0.24	—	0.22	0.22	0.20	0.19	0.06	0.07	0.06	0.09
6. *Bacillus firmus*	0.08	0.06	0.07	0.25	0.22	—	0.34	0.26	0.20	0.11	0.13	0.06	0.12
7. *Corynebacterium diphtheriae*	0.09	0.10	0.07	0.28	0.22	0.34	—	0.23	0.21	0.12	0.12	0.09	0.10
8. *Aphanocapsa* 6714	0.11	0.09	0.09	0.26	0.20	0.26	0.23	—	0.31	0.11	0.11	0.10	0.10
9. Chloroplast (*Lemna*)	0.08	0.11	0.06	0.21	0.19	0.20	0.21	0.31	—	0.14	0.12	0.10	0.12
10. *Methanobacterium thermoautotrophicum*	0.11	0.10	0.10	0.11	0.06	0.11	0.12	0.11	0.14	—	0.51	0.25	0.30
11. *M. ruminantium* strain M-1	0.11	0.10	0.10	0.12	0.07	0.13	0.12	0.11	0.12	0.51	—	0.25	0.24
12. *Methanobacterium* sp., Cariaco isolate JR-1	0.08	0.13	0.09	0.07	0.06	0.06	0.09	0.10	0.10	0.25	0.25	—	0.32
13. *Methanosarcina barkeri*	0.08	0.07	0.07	0.12	0.09	0.12	0.10	0.10	0.12	0.30	0.24	0.32	—

The 16S (18S) ribosomal RNA from the organisms (organelles) listed were digested with T1 RNase and the resulting digests were subjected to two-dimensional electrophoretic separation to produce an oligonucleotide fingerprint. The individual oligonucleotides on each fingerprint were then sequenced by established procedures (13, 14) to produce an oligonucleotide catalog characteristic of the given organism (3, 4, 13–17, 22, 23; unpublished data). Comparisons of all possible pairs of such catalogs defines a set of association coefficients (S_{AB}) given by: $S_{AB} = 2N_{AB}/(N_A + N_B)$, in which N_A, N_B, and N_{AB} are the total numbers of nucleotides in sequences of hexamers or larger in the catalog for organism A, in that for organism B, and in the interreaction of the two catalogs, respectively (13, 23).

Woese and Fox. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*. 74(11): 5088–5090. PMCID: PMC432104



The Complete Nucleotide Sequence of the Ribosomal 16-S RNA from *Escherichia coli*

Experimental Details and Cistron Heterogeneities

Philippe CARBON, Chantal EHRESMANN, Bernard EHRESMANN, and Jean-Pierre EBEL

The complete nucleotide sequence of the 16-S RNA from *Escherichia coli* has been determined using rapid RNA-sequencing gel methods. The experimental data are fully described in this paper. The specificities of the ribonucleases, especially the ribonuclease *Phy*I are discussed and the consequences of the persistence of stable secondary structure are considered. The proposed sequence contains 1541 nucleotides and agrees completely with the DNA sequence of the *rrnB* cistron deduced by Brosius, J., Palmer, M. L., Kennedy, P.J., and Noller, H. F. [*Proc. Natl Acad. Sci. U.S.A.* (1978) *75*, 4801 – 4805]. But there are several cistron heterogeneities of which we described 16 single-base heterogeneities, 7 of the deletion/insertion type and 9 of the transition or transversion type. Our observations suggest the existence, among the 7 ribosome RNA cistrons, of one or two mutated ones. The respective advantages and disadvantages of both RNA and DNA sequencing methods are discussed.

The sequence of the ribosomal 16-S RNA from *Escherichia coli* has been studied for ten years [1] in the hope of gaining information about the organization and function of ribosomes.

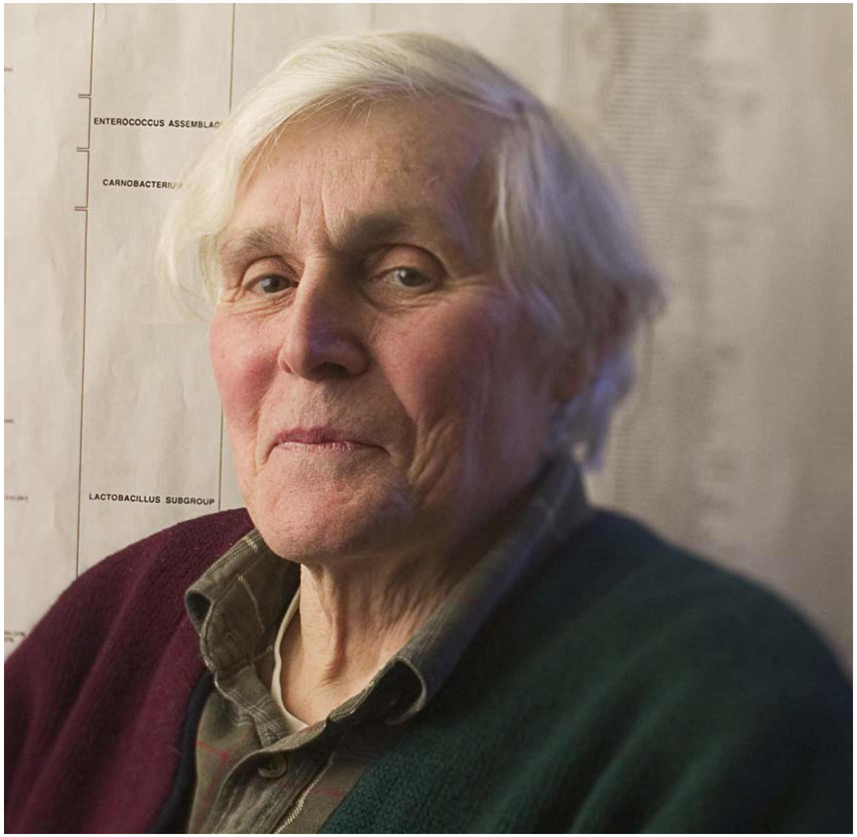
When the only available sequencing technique was that of Sanger et al. [2] extreme technical difficulties were encountered that reflect the intrinsic limitations

fragments and subsequent fractionation of the digests on polyacrylamide gels [5,6]. These two technological improvements enabled us to determine the complete sequence of the 16-S RNA, which has been briefly reported already [7]. The experimental data are fully described in this paper. The sequence is compared with the DNA sequence of a 16-S RNA gene deduced

P. Carbon, C. Ehresmann, B. Ehresmann, and J.-P. Ebel, Laboratoire de Biochimie, Institut de Biologie Moléculaire et Cellulaire du C.N.R.S., 15 Rue René Descartes, Esplanade, F-67084 Strasbourg-Cedex, France

Elucidation of the three domains of life

Carl Woese
(1929 – 2012)

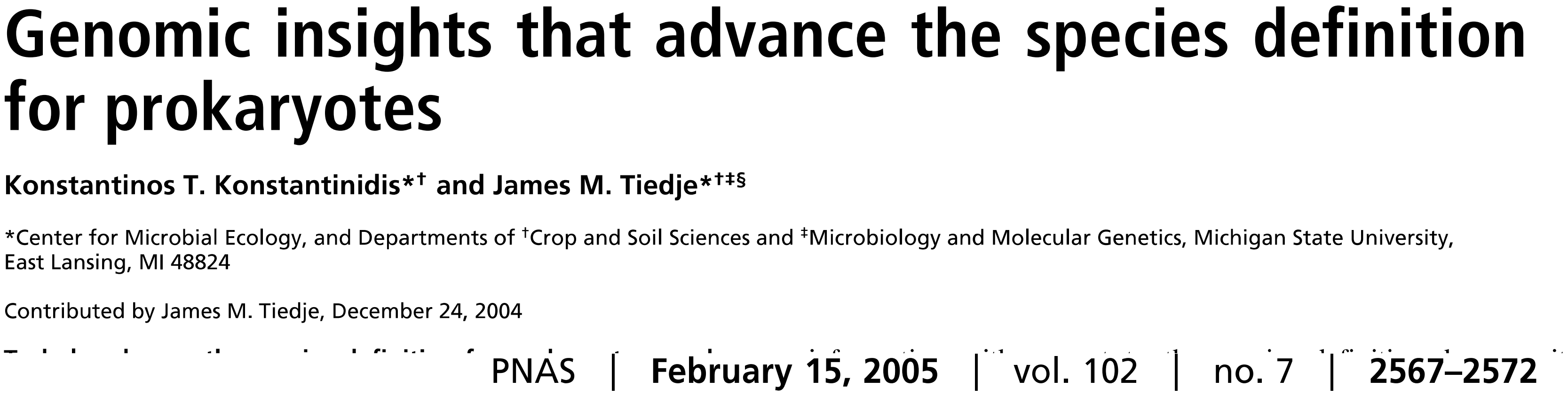


Ribosomal RNA sequence as phylogenetic marker

- Discovered "3rd kingdom"
- Archaea and Bacteria separate domains

Average Nucleotide Identity and Average Amino Acid Identity

Whole Genome Comparison



Genomic insights that advance the species definition for prokaryotes

Konstantinos T. Konstantinidis*[†] and James M. Tiedje*[†‡§]

*Center for Microbial Ecology, and Departments of [†]Crop and Soil Sciences and [‡]Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824

Contributed by James M. Tiedje, December 24, 2004

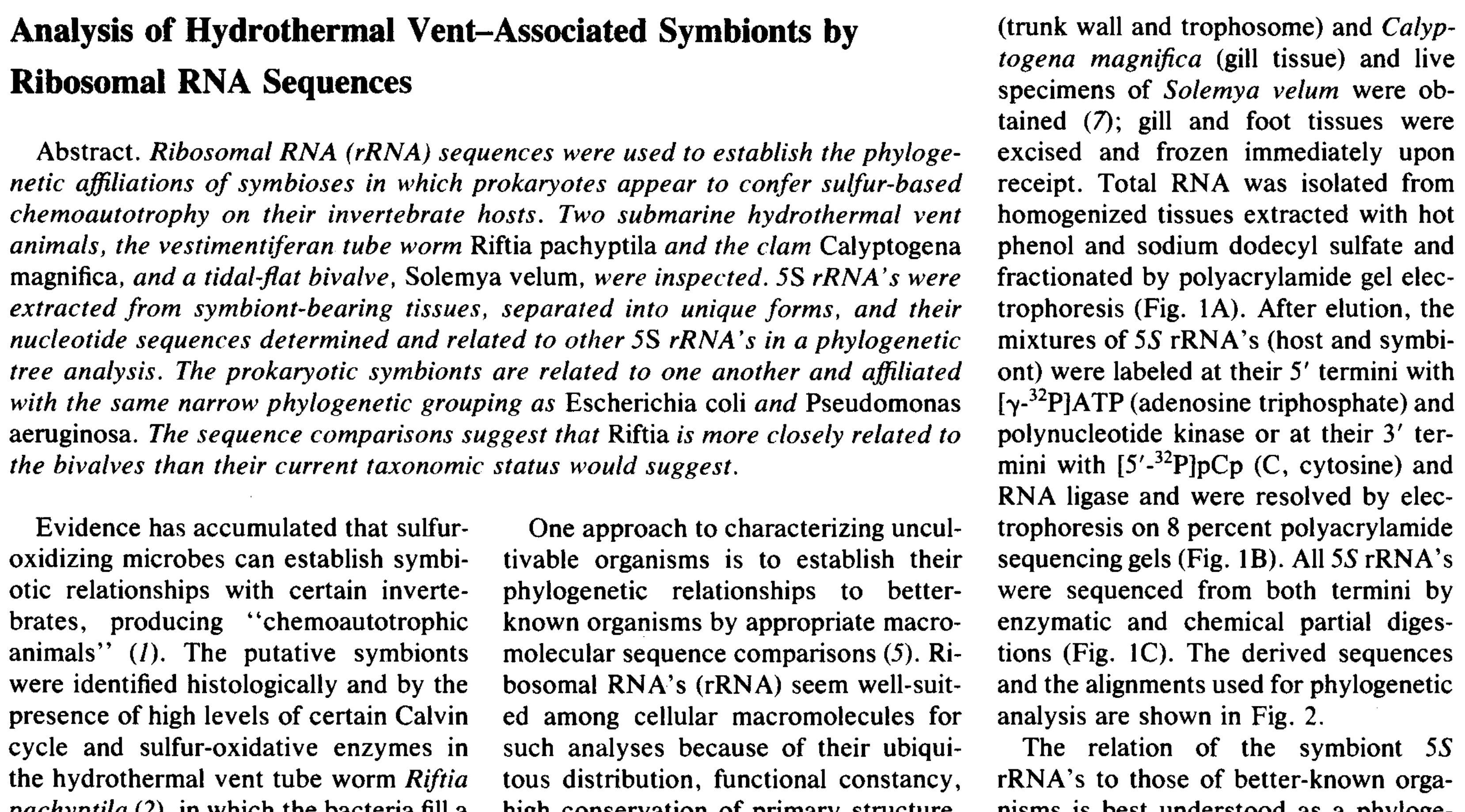
PNAS | **February 15, 2005** | vol. 102 | no. 7 | **2567–2572**

Kostas T. Konstantinidis
Georgia Tech

http://rdp.cme.msu.edu/

Diversity of uncultured organisms explored by rRNA sequencing

David A. Stahl, David J. Lane, Gary J. Olsen and Norman R. Pace
Science, New Series, Vol. 224, No. 4647 (Apr. 27, 1984), pp. 409-411
Published by: American Association for the Advancement of Science



Analysis of Hydrothermal Vent–Associated Symbionts by Ribosomal RNA Sequences

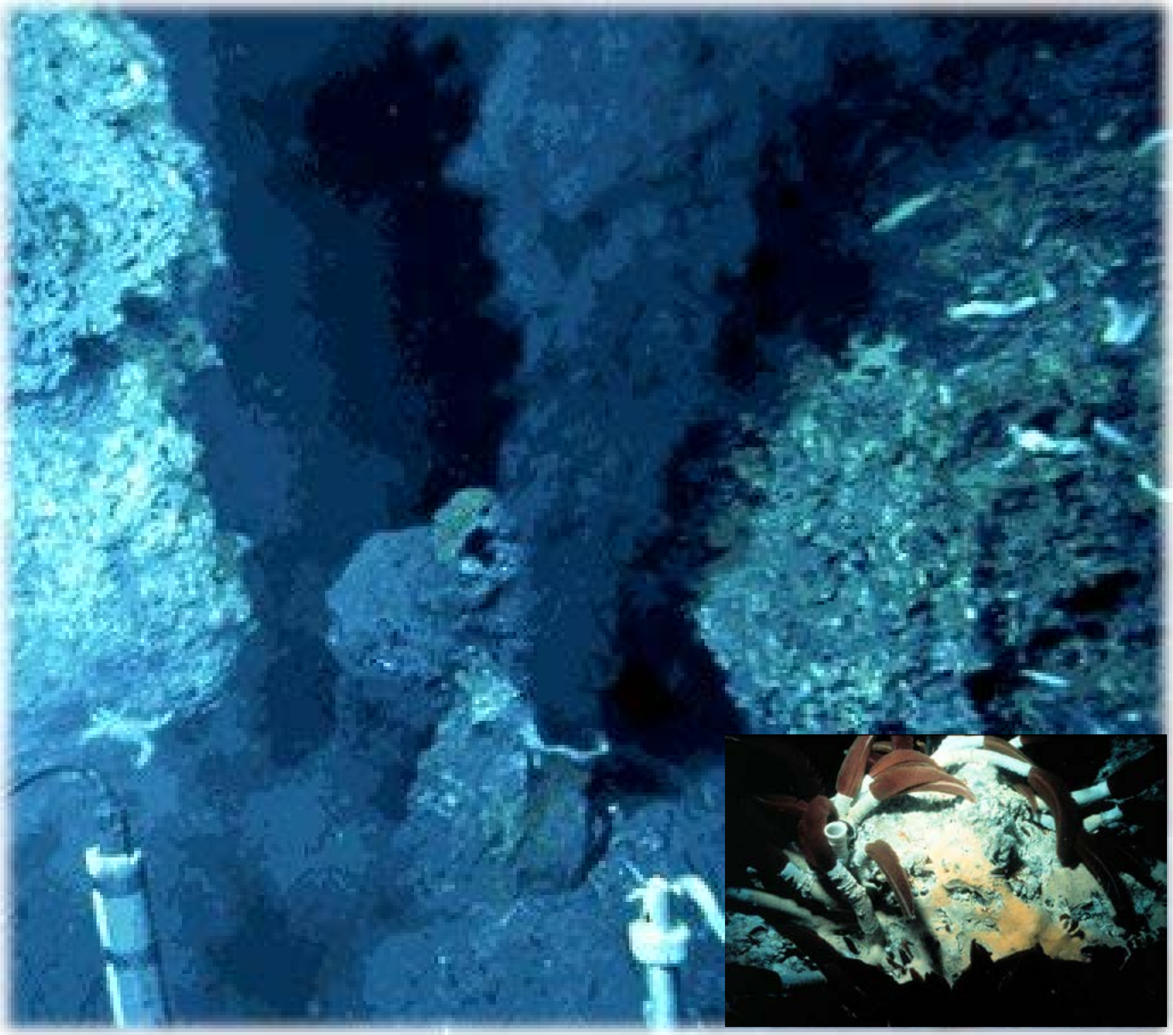
Abstract. *Ribosomal RNA (rRNA) sequences were used to establish the phylogenetic affiliations of symbioses in which prokaryotes appear to confer sulfur-based chemoautotrophy on their invertebrate hosts. Two submarine hydrothermal vent animals, the vestimentiferan tube worm* Riftia pachyptila *and the clam* Calyptogena magnifica, *and a tidal-flat bivalve,* Solemya velum, *were inspected. 5S rRNA's were extracted from symbiont-bearing tissues, separated into unique forms, and their nucleotide sequences determined and related to other 5S rRNA's in a phylogenetic tree analysis. The prokaryotic symbionts are related to one another and affiliated with the same narrow phylogenetic grouping as* Escherichia coli *and* Pseudomonas aeruginosa. *The sequence comparisons suggest that* Riftia *is more closely related to the bivalves than their current taxonomic status would suggest.*

Evidence has accumulated that sulfur-oxidizing microbes can establish symbiotic relationships with certain invertebrates, producing "chemoautotrophic animals" (*1*). The putative symbionts were identified histologically and by the presence of high levels of certain Calvin cycle and sulfur-oxidative enzymes in the hydrothermal vent tube worm *Riftia pachyptila* (*2*), in which the bacteria fill a

One approach to characterizing uncultivable organisms is to establish their phylogenetic relationships to better-known organisms by appropriate macromolecular sequence comparisons (*5*). Ribosomal RNA's (rRNA) seem well-suited among cellular macromolecules for such analyses because of their ubiquitous distribution, functional constancy, high conservation of primary structure,

(trunk wall and trophosome) and *Calyptogena magnifica* (gill tissue) and live specimens of *Solemya velum* were obtained (*7*); gill and foot tissues were excised and frozen immediately upon receipt. Total RNA was isolated from homogenized tissues extracted with hot phenol and sodium dodecyl sulfate and fractionated by polyacrylamide gel electrophoresis (Fig. 1A). After elution, the mixtures of 5*S* rRNA's (host and symbiont) were labeled at their 5′ termini with [γ-^{32}P]ATP (adenosine triphosphate) and polynucleotide kinase or at their 3′ termini with [5′-^{32}P]pCp (C, cytosine) and RNA ligase and were resolved by electrophoresis on 8 percent polyacrylamide sequencing gels (Fig. 1B). All 5*S* rRNA's were sequenced from both termini by enzymatic and chemical partial digestions (Fig. 1C). The derived sequences and the alignments used for phylogenetic analysis are shown in Fig. 2.

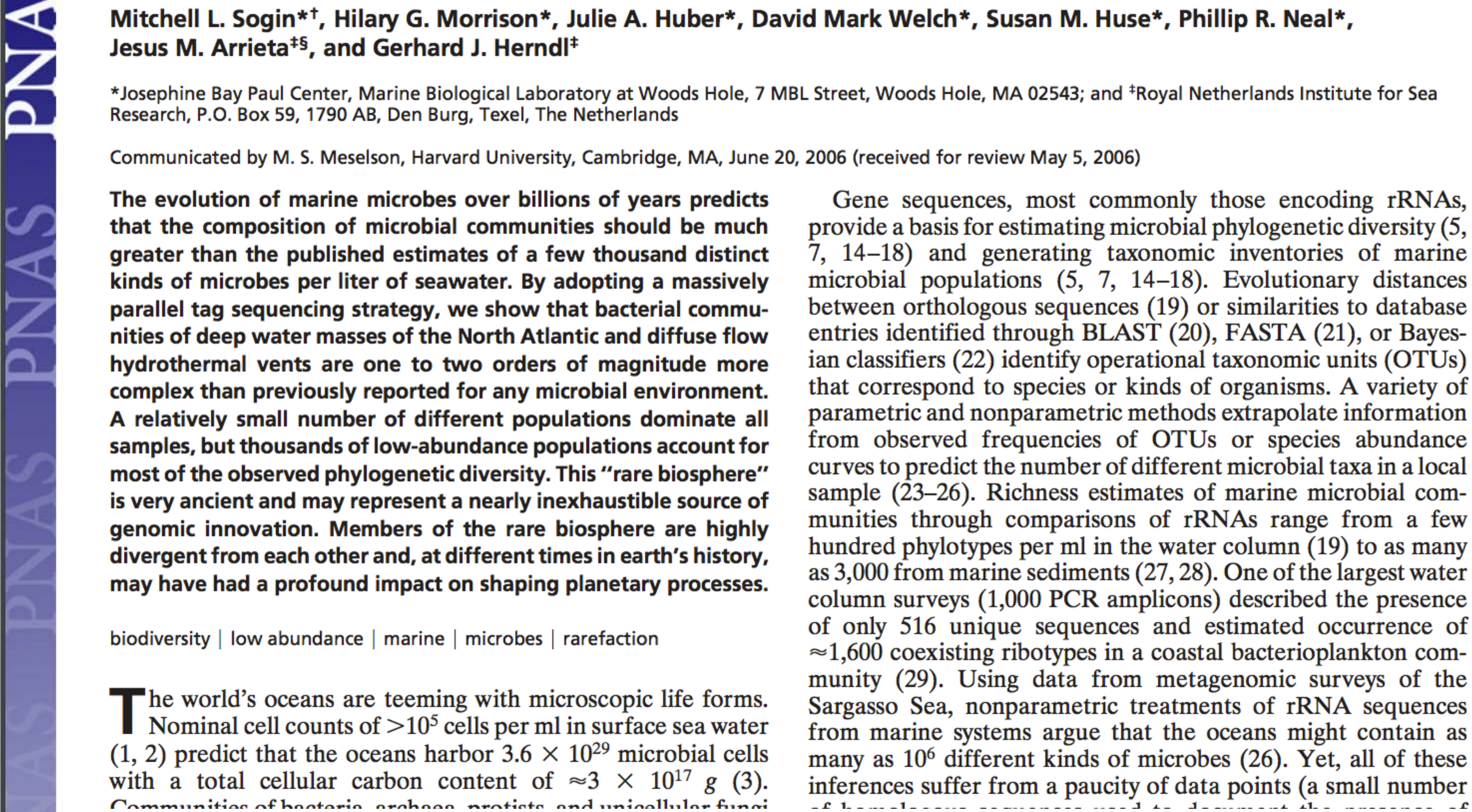
The relation of the symbiont 5*S* rRNA's to those of better-known organisms is best understood as a phyloge-



Hydrothermal Vent Black Smoker

Sogin *et al.*, 2006. *PNAS* 103(32): 12115–12120
doi: 10.1073pnas.0605127103

Microbial diversity in the deep sea and the underexplored "rare biosphere"



Mitchell L. Sogin*†, Hilary G. Morrison*, Julie A. Huber*, David Mark Welch*, Susan M. Huse*, Phillip R. Neal*, Jesus M. Arrieta‡§, and Gerhard J. Herndl‡

*Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and ‡Royal Netherlands Institute for Sea Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

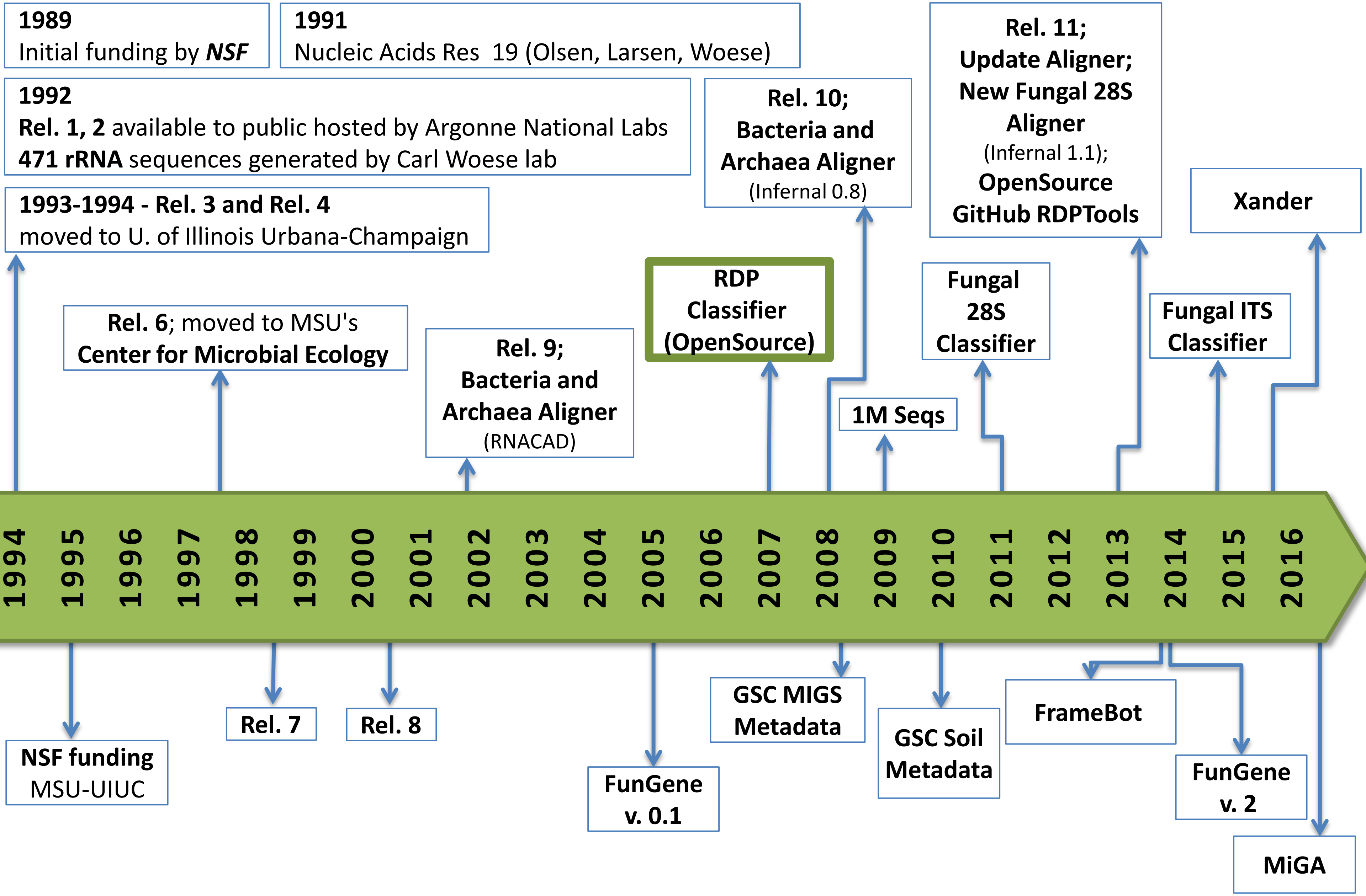
Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

The evolution of marine microbes over billions of years predicts that the composition of microbial communities should be much greater than the published estimates of a few thousand distinct kinds of microbes per liter of seawater. By adopting a massively parallel tag sequencing strategy, we show that bacterial communities of deep water masses of the North Atlantic and diffuse flow hydrothermal vents are one to two orders of magnitude more complex than previously reported for any microbial environment. A relatively small number of different populations dominate all samples, but thousands of low-abundance populations account for most of the observed phylogenetic diversity. This "rare biosphere" is very ancient and may represent a nearly inexhaustible source of genomic innovation. Members of the rare biosphere are highly divergent from each other and, at different times in earth's history, may have had a profound impact on shaping planetary processes.

biodiversity | low abundance | marine | microbes | rarefaction

The world's oceans are teeming with microscopic life forms. Nominal cell counts of >10^5 cells per ml in surface sea water (1, 2) predict that the oceans harbor 3.6 × 10^{29} microbial cells with a total cellular carbon content of ≈3 × 10^{17} *g* (3). Communities of bacteria, archaea, protists, and unicellular fungi

Gene sequences, most commonly those encoding rRNAs, provide a basis for estimating microbial phylogenetic diversity (5, 7, 14–18) and generating taxonomic inventories of marine microbial populations (5, 7, 14–18). Evolutionary distances between orthologous sequences (19) or similarities to database entries identified through BLAST (20), FASTA (21), or Bayesian classifiers (22) identify operational taxonomic units (OTUs) that correspond to species or kinds of organisms. A variety of parametric and nonparametric methods extrapolate information from observed frequencies of OTUs or species abundance curves to predict the number of different microbial taxa in a local sample (23–26). Richness estimates of marine microbial communities through comparisons of rRNAs range from a few hundred phylotypes per ml in the water column (19) to as many as 3,000 from marine sediments (27, 28). One of the largest water column surveys (1,000 PCR amplicons) described the presence of only 516 unique sequences and estimated occurrence of ≈1,600 coexisting ribotypes in a coastal bacterioplankton community (29). Using data from metagenomic surveys of the Sargasso Sea, nonparametric treatments of rRNA sequences from marine systems argue that the oceans might contain as many as 10^6 different kinds of microbes (26). Yet, all of these inferences suffer from a paucity of data points (a small number of homologous sequences used to document the presence of



1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's
Center for Microbial Ecology

Rel. 7

Rel. 8

Rel. 9;
Bacteria and
Archaea Aligner
(RNACAD)

FunGene
v. 0.1

RDP
Classifier
(OpenSource)

GSC MIGS
Metadata

Rel. 10;
Bacteria and
Archaea Aligner
(Infernal 0.8)

1M Seqs

GSC Soil
Metadata

Fungal
28S
Classifier

Rel. 11;
Update Aligner;
New Fungal 28S
Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

FrameBot

FunGene
v. 2

Fungal ITS
Classifier

Xander

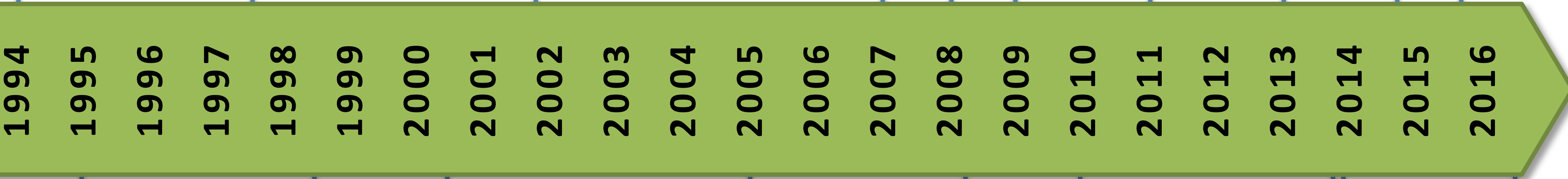
MiGA

1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign



NSF funding
MSU-UIUC

Rel. 6; moved to MSU's
Center for Microbial Ecology

Rel. 7

Rel. 8

Rel. 9;
Bacteria and
Archaea Aligner
(RNACAD)

FunGene
v. 0.1

RDP
Classifier
(OpenSource)

Rel. 10;
Bacteria and
Archaea Aligner
(Infernal 0.8)

GSC MIGS
Metadata

1M Seqs

GSC Soil
Metadata

Fungal
28S
Classifier

Rel. 11;
Update Aligner;
New Fungal 28S
Aligner
(Infernal 1.1);
OpenSource
GitHub RDPTools

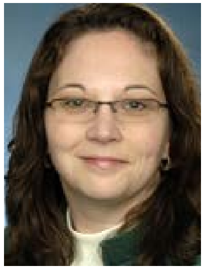
FrameBot

Fungal ITS
Classifier

FunGene
v. 2

Xander

MiGA



Dawn Field
Lynn Schrml

gsc genomic STANDARDS consortium

The mission of the GSC is to work with the wider community towards:

- the implementation of new genomic standards
- methods of capturing and exchanging metadata
- harmonization of metadata collection and analysis efforts across the wider genomics community

Other Ribosomal RNA Databases

silva

high quality ribosomal RNA databases

http://www.arb-silva.de/

GREENGENES

The 16S rRNA Gene Database and Tools

http://greengenes.secondgenome.com/

HOMD

Human Oral Microbiome Database

http://www.homd.org/

THE UNIVERSITY OF TEXAS AT AUSTIN

COMPARATIVE RNA WEB SITE

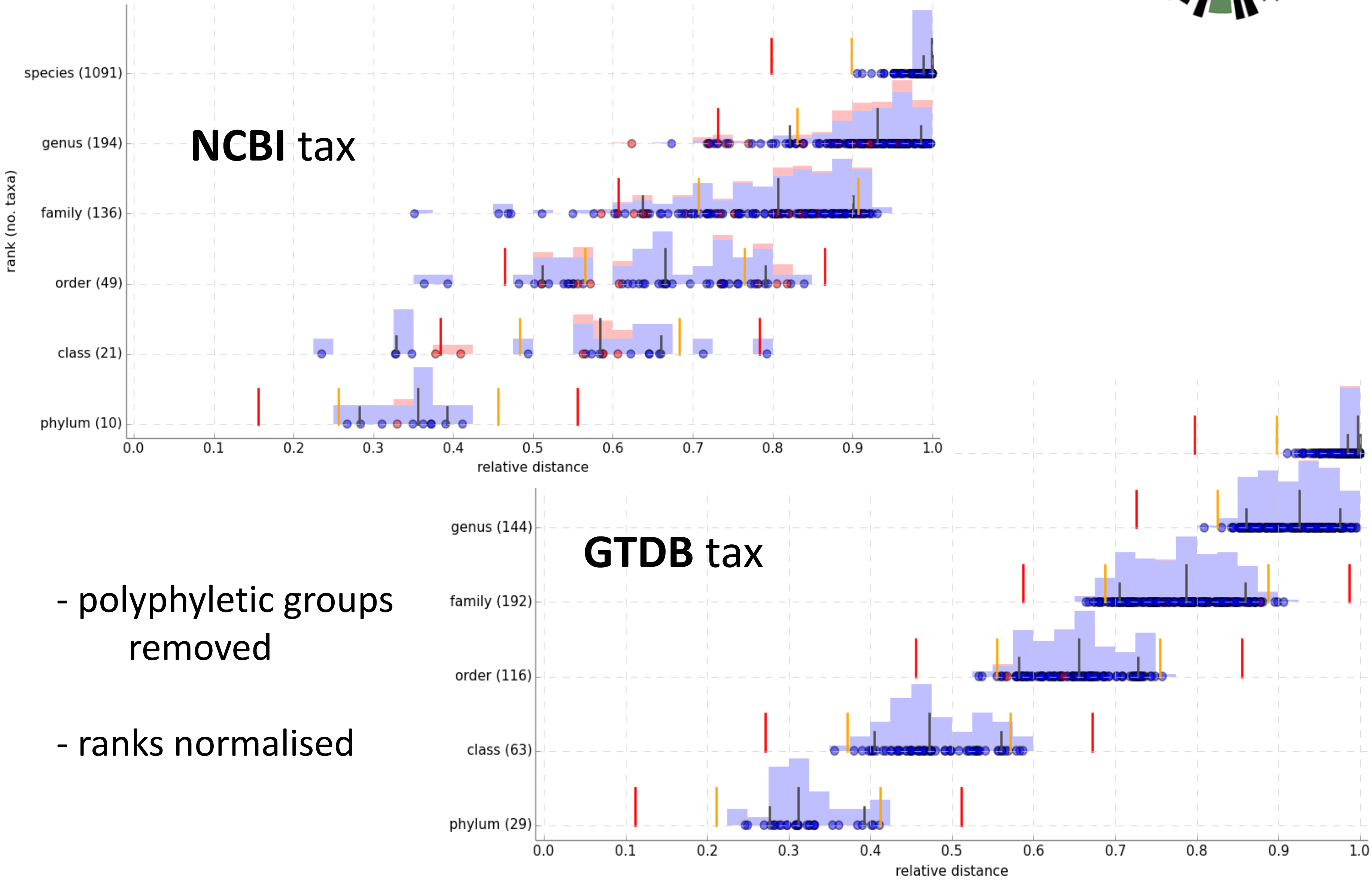
http://www.rna.icmb.utexas.edu/

Current Trends

- Migration back to high-fidelity near-full-length environmental amplicons (e.g. PacBio)

- Integration of higher resolution genomic data into 16S based phylogeny and taxonomy

Genome Taxonomy Database



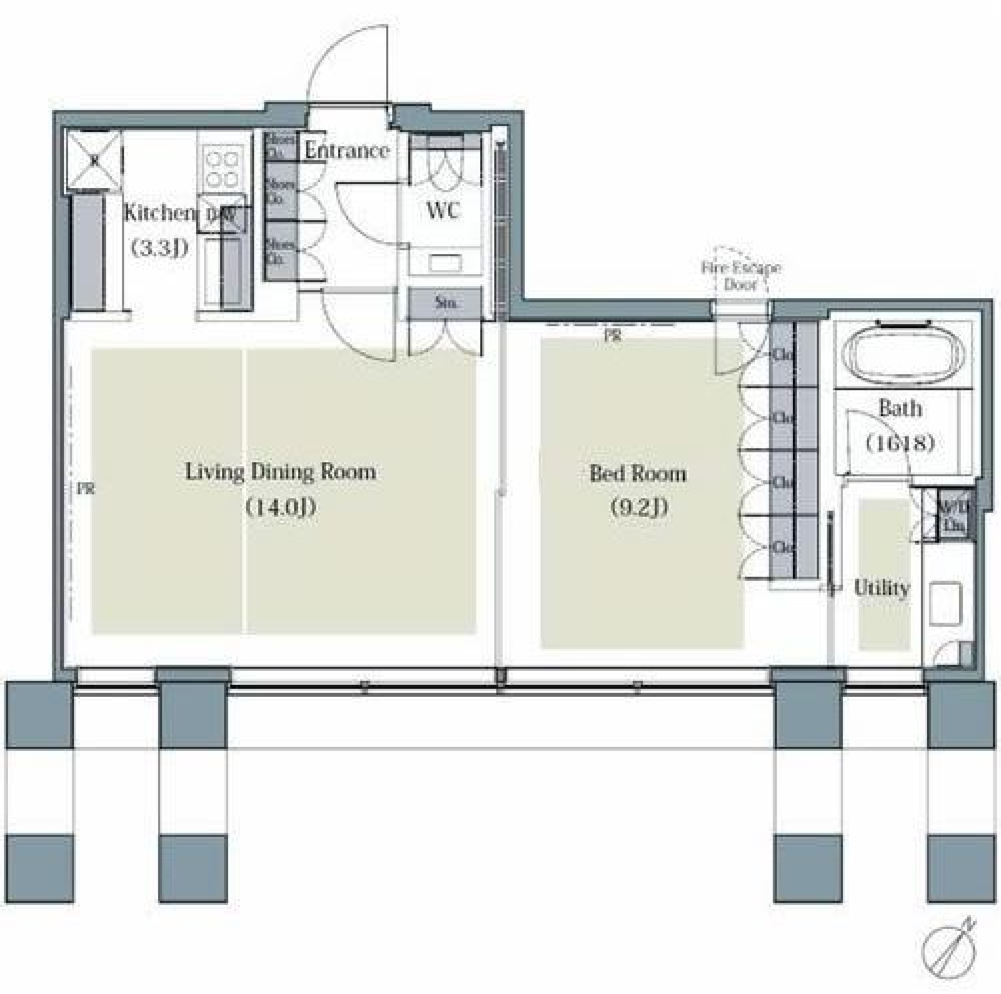
- polyphyletic groups removed

- ranks normalised

Ribosomal RNA Shows the Framework

http://rdp.cme.msu.edu

Functional Genes Show the Details



Akasaka K-Tower Residence

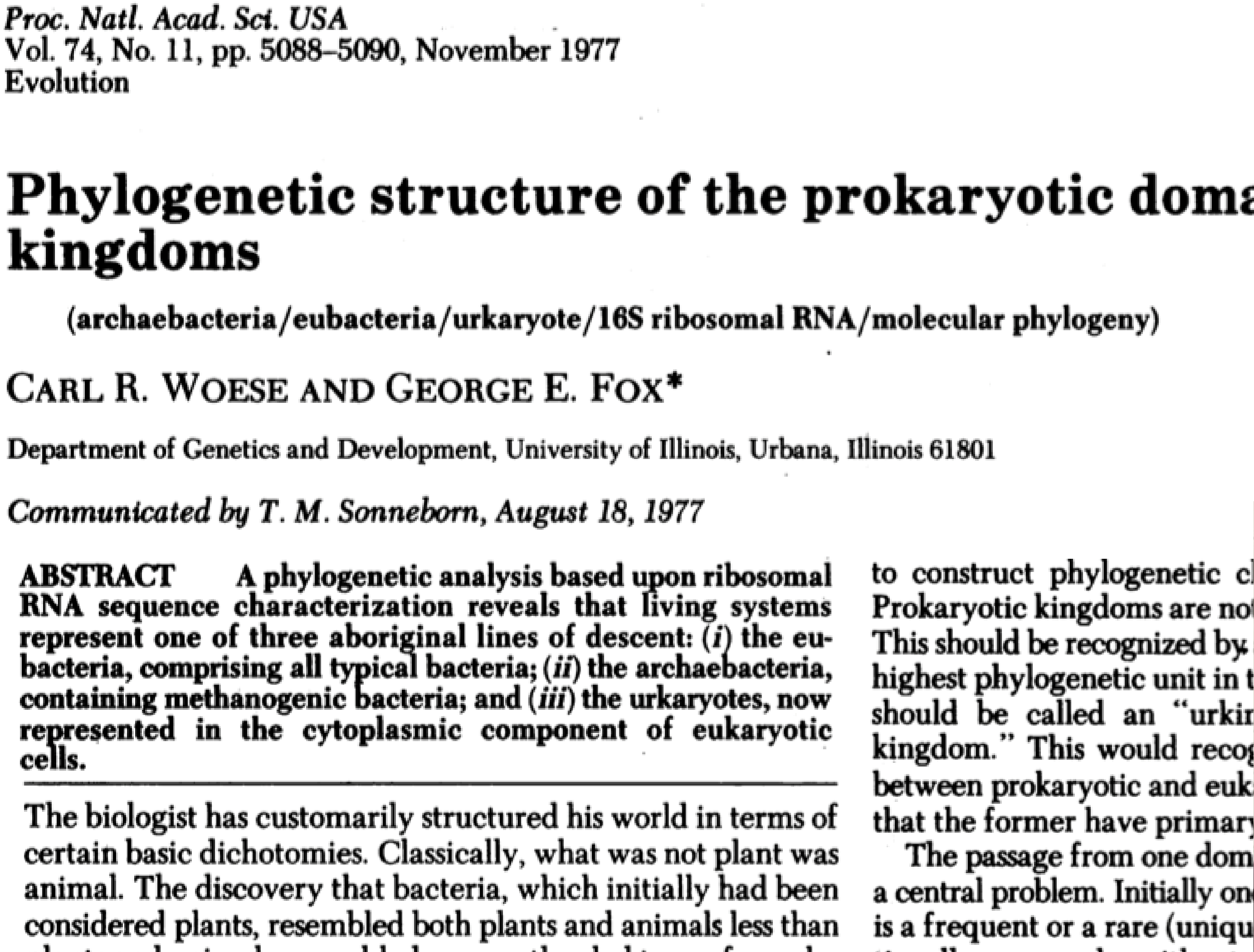
from http://real.tokyoapartment81.com/en/rent/view/231423

Rhizosphere Soil Data, Xander Assembly

Gene	***nirK***			***nifH***			***rplB***		
Crop	C	M	S	C	M	S	C	M	S
# chimeric clusters	16	207	11	0	1	0	14	28	44
# protein contig clusters	1993	1807	1581	39	57	41	19287	20463	17334
# OTUs at 95% aa identity	741	674	582	14	24	17	6100	6887	6004
Median (aa)	215	230	208	294	256	255	274	274	274
Longest (aa)	380	372	370	296	296	296	285	285	284
Median % aa identity	88.3	84.7							76.3
Max % aa identity	100	99.4							100
# reads covering kmers	27404	19815	16661	411	534	461	225985	179867	149661
Gene Abundance	0.121	0.11	0.111	0.002	0.003	0.003			

Gene abundance calculated from the ratio of *nirK* or *nifH* reads to *rplB* reads, corrected for gene length.

Elucidation of the three primary lineages



Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

Communicated by T. M. Sonneborn, August 18, 1977

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent: (*i*) the eubacteria, comprising all typical bacteria; (*ii*) the archaebacteria, containing methanogenic bacteria; and (*iii*) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

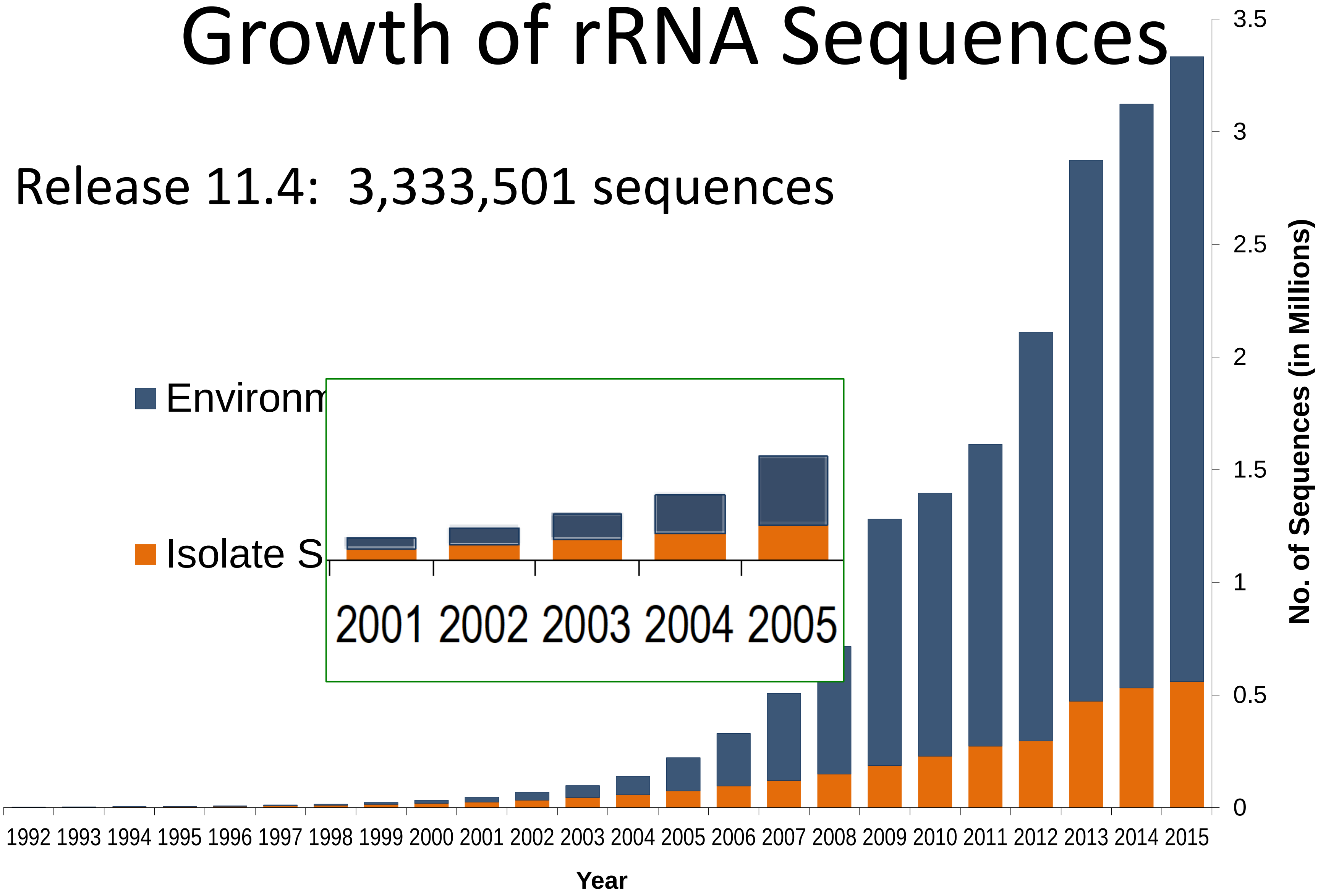
The biologist has customarily structured his world in terms of certain basic dichotomies. Classically, what was not plant was animal. The discovery that bacteria, which initially had been considered plants, resembled both plants and animals less than

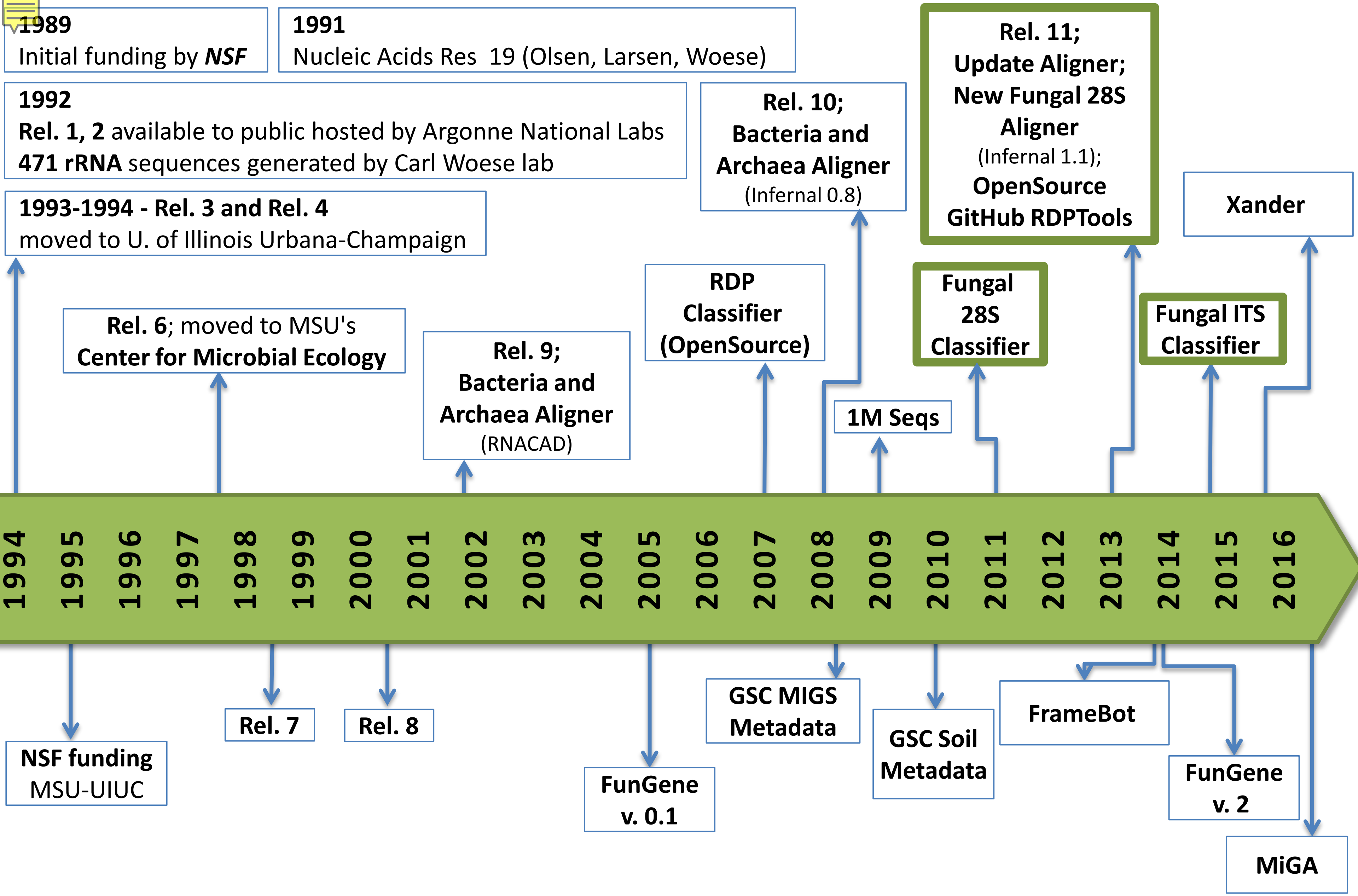
to construct phylogenetic c
Prokaryotic kingdoms are no
This should be recognized by
highest phylogenetic unit in t
should be called an "urkin
kingdom." This would recog
between prokaryotic and euk
that the former have primary

The passage from one dom
a central problem. Initially on
is a frequent or a rare (uniqu

Growth of rRNA Sequences

Release 11.4: 3,333,501 sequences





1989
Initial funding by ***NSF***

1991
Nucleic Acids Res 19 (Olsen, Larsen, Woese)

1992
Rel. 1, 2 available to public hosted by Argonne National Labs
471 rRNA sequences generated by Carl Woese lab

1993-1994 - Rel. 3 and Rel. 4
moved to U. of Illinois Urbana-Champaign

NSF funding
MSU-UIUC

Rel. 6; moved to MSU's **Center for Microbial Ecology**

Rel. 7

Rel. 8

Rel. 9; Bacteria and Archaea Aligner (RNACAD)

FunGene v. 0.1

RDP Classifier (OpenSource)

GSC MIGS Metadata

Rel. 10; Bacteria and Archaea Aligner (Infernal 0.8)

1M Seqs

GSC Soil Metadata

Fungal 28S Classifier

Rel. 11; Update Aligner; New Fungal 28S Aligner (Infernal 1.1); **OpenSource GitHub RDPTools**

FrameBot

Fungal ITS Classifier

FunGene v. 2

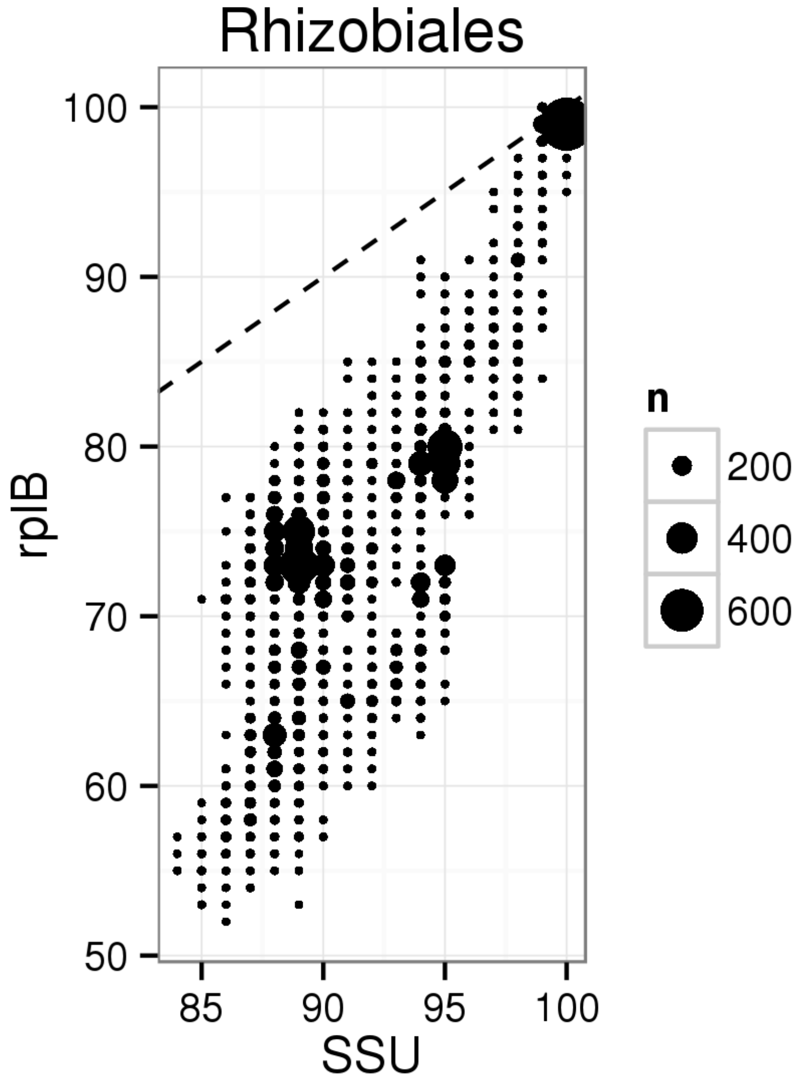
Xander

MiGA

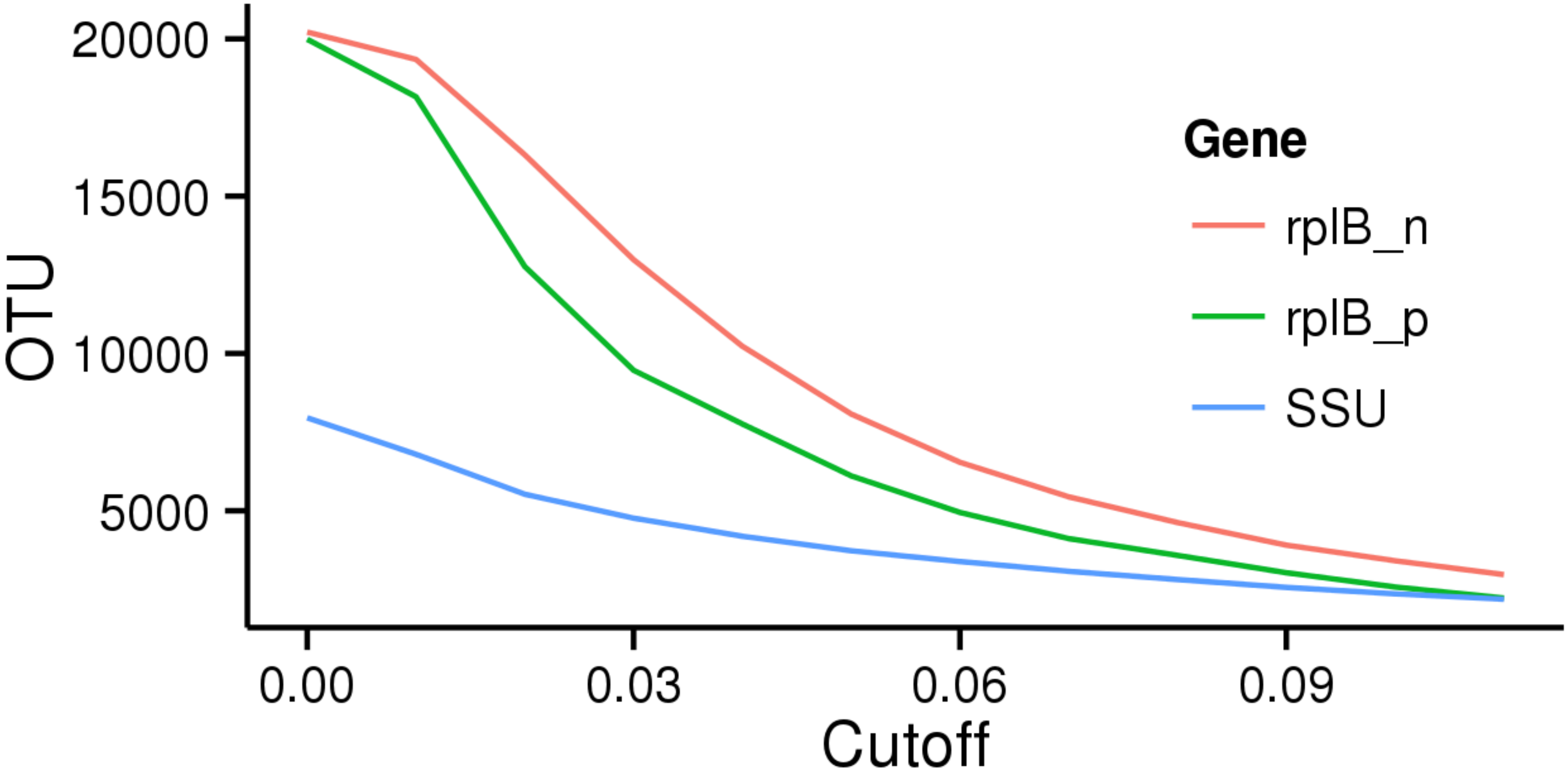
http://rdp.cme.msu.edu/

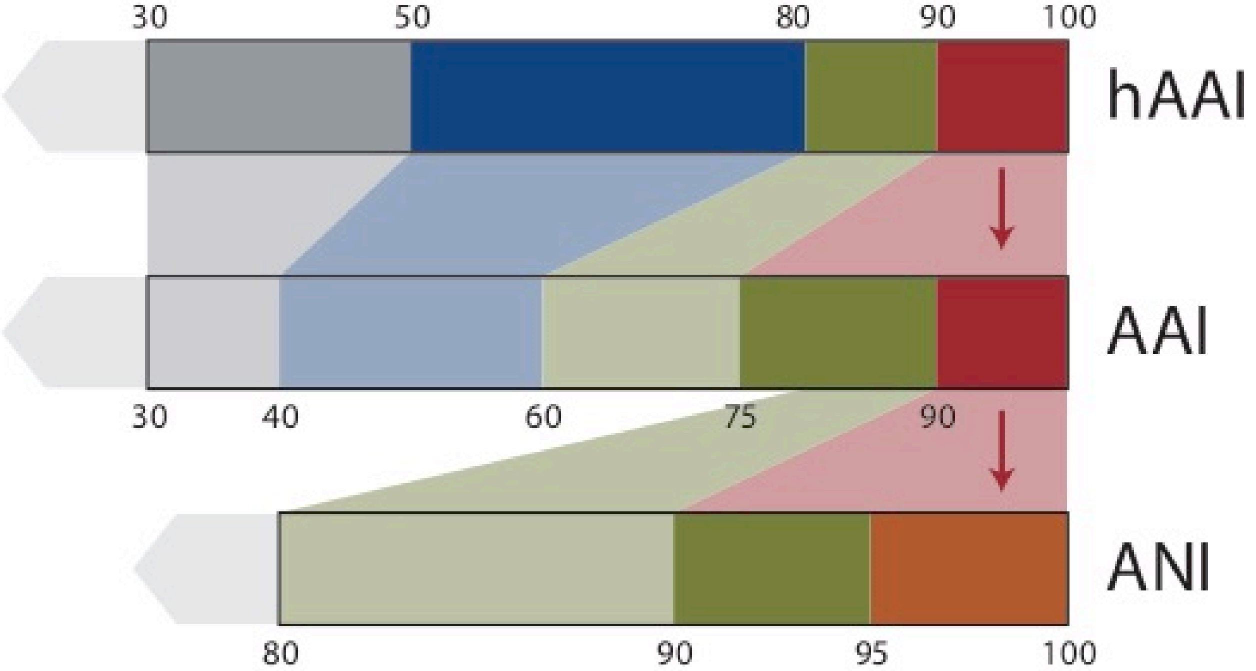
rplB vs 16S Parwise Distances in one Order (RefSeq Genomes)

Jiarong Guo



Number of OTUs in Agricultural Rhizosphere Soil Metagenome Sample





Novel phylum from known domain

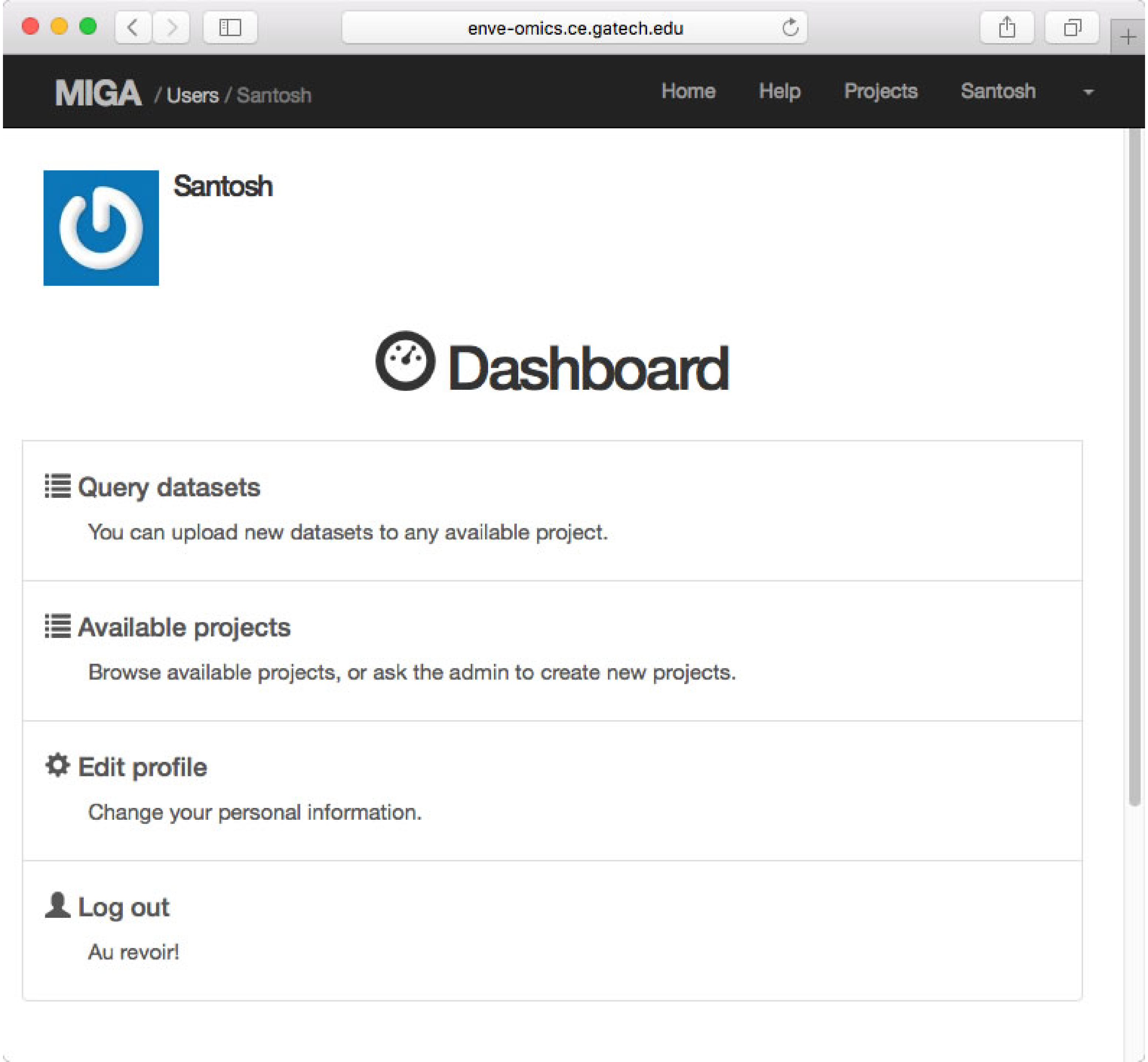
Novel genus from known phylum

Novel species from known genus

Known species

Unreliable value

Hierarchical approach to genome classification



Santosh

Dashboard

Query datasets

You can upload new datasets to any available project.

Available projects

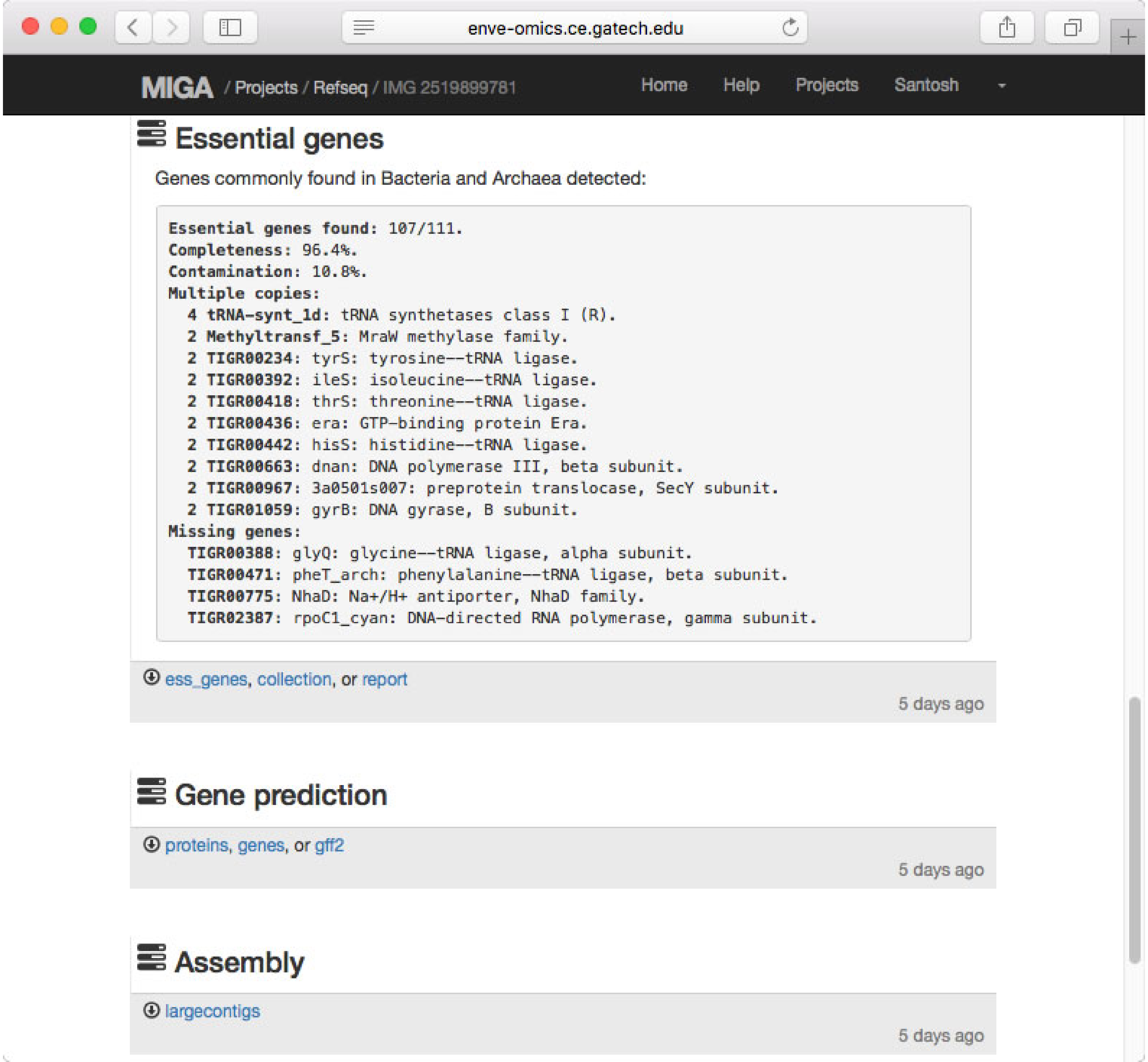
Browse available projects, or ask the admin to create new projects.

Edit profile

Change your personal information.

Log out

Au revoir!



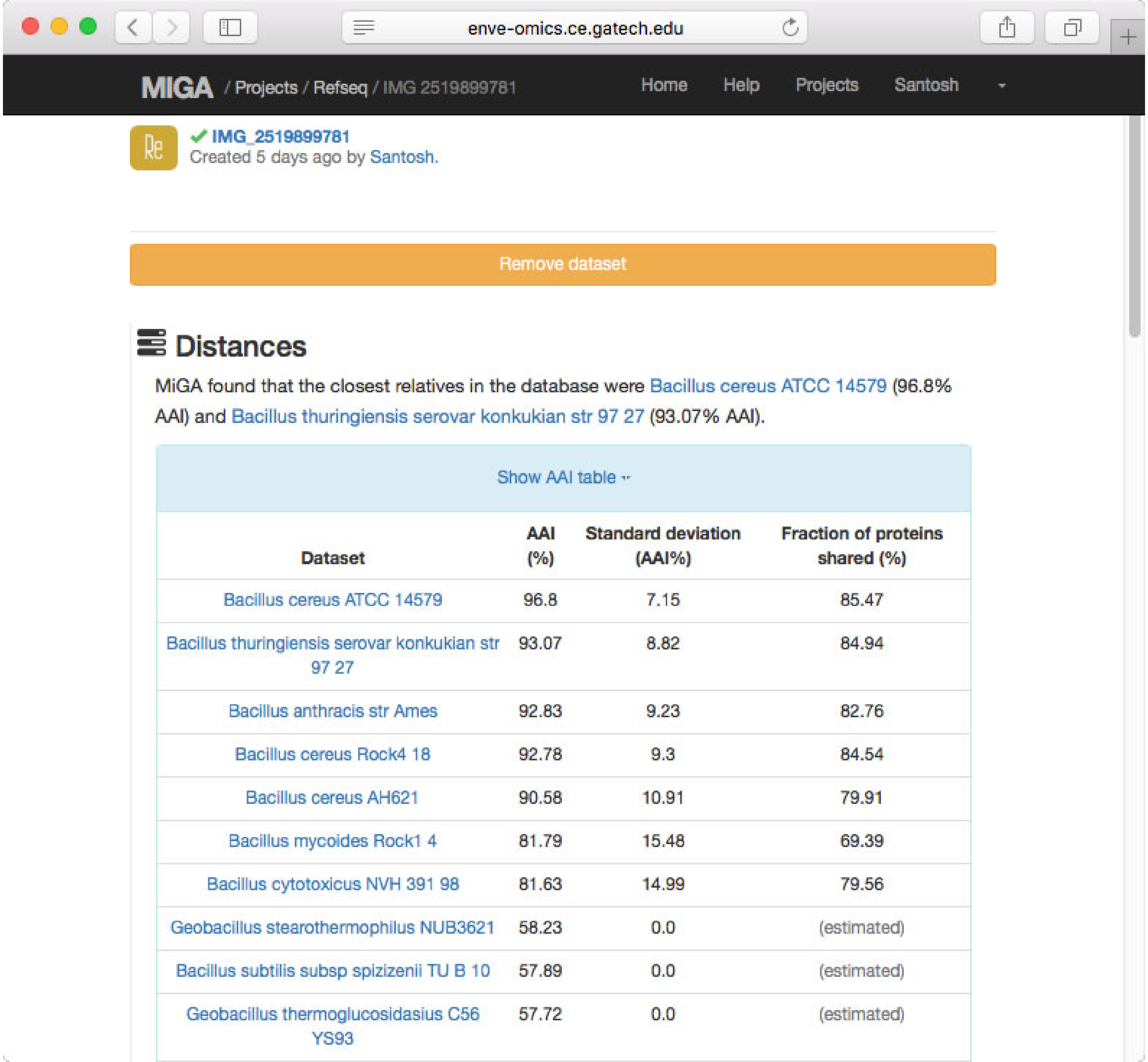
Essential genes

Genes commonly found in Bacteria and Archaea detected:

```
Essential genes found: 107/111.
Completeness: 96.4%.
Contamination: 10.8%.
Multiple copies:
  4 tRNA-synt_1d: tRNA synthetases class I (R).
  2 Methyltransf_5: MraW methylase family.
  2 TIGR00234: tyrS: tyrosine--tRNA ligase.
  2 TIGR00392: ileS: isoleucine--tRNA ligase.
  2 TIGR00418: thrS: threonine--tRNA ligase.
  2 TIGR00436: era: GTP-binding protein Era.
  2 TIGR00442: hisS: histidine--tRNA ligase.
  2 TIGR00663: dnan: DNA polymerase III, beta subunit.
  2 TIGR00967: 3a0501s007: preprotein translocase, SecY subunit.
  2 TIGR01059: gyrB: DNA gyrase, B subunit.
Missing genes:
  TIGR00388: glyQ: glycine--tRNA ligase, alpha subunit.
  TIGR00471: pheT_arch: phenylalanine--tRNA ligase, beta subunit.
  TIGR00775: NhaD: Na+/H+ antiporter, NhaD family.
  TIGR02387: rpoC1_cyan: DNA-directed RNA polymerase, gamma subunit.
```

Gene prediction

Assembly



✓ IMG_2519899781

Created 5 days ago by Santosh.

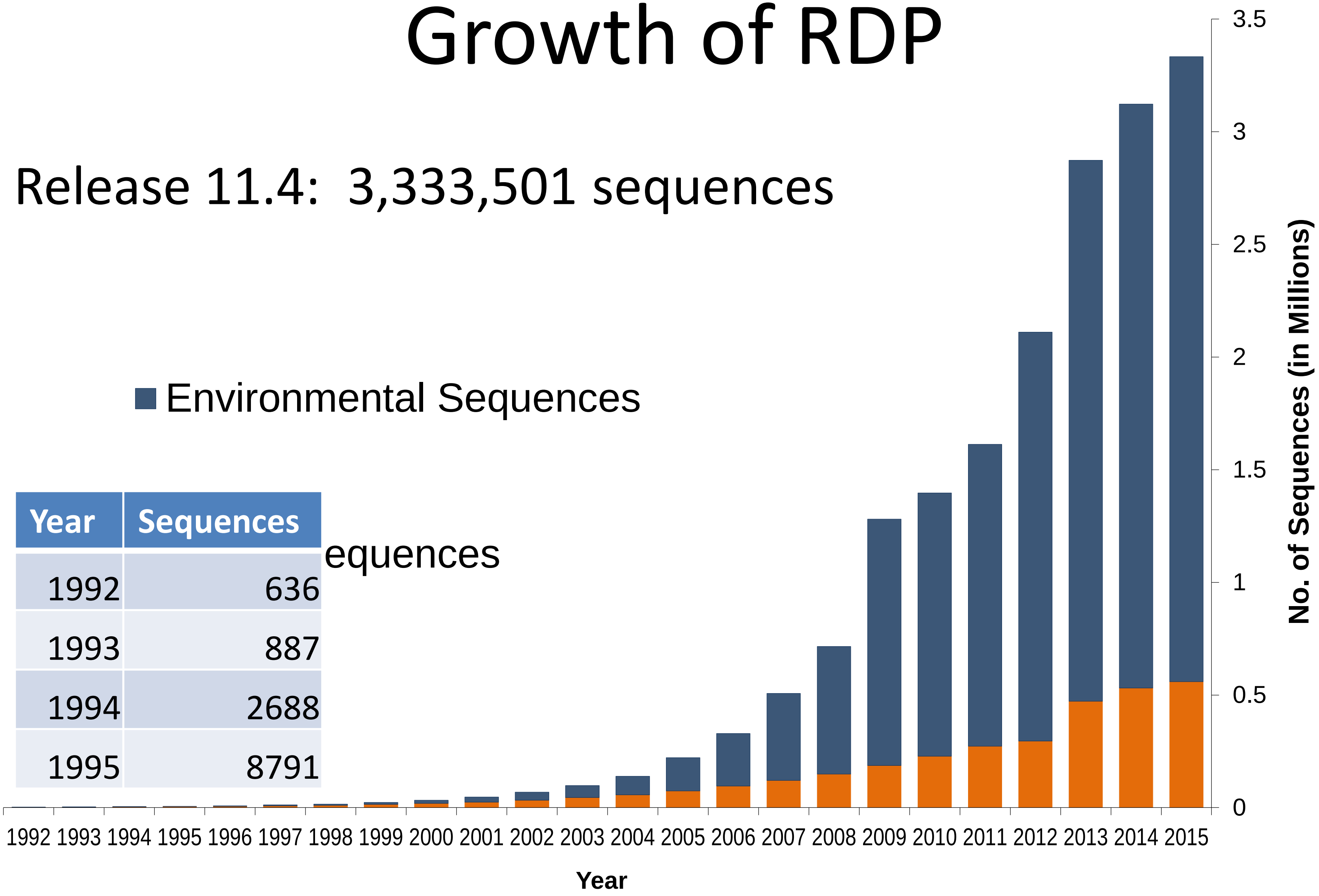
Distances

MiGA found that the closest relatives in the database were Bacillus cereus ATCC 14579 (96.8% AAI) and Bacillus thuringiensis serovar konkukian str 97 27 (93.07% AAI).

Dataset	AAI (%)	Standard deviation (AAI%)	Fraction of proteins shared (%)
Bacillus cereus ATCC 14579	96.8	7.15	85.47
Bacillus thuringiensis serovar konkukian str 97 27	93.07	8.82	84.94
Bacillus anthracis str Ames	92.83	9.23	82.76
Bacillus cereus Rock4 18	92.78	9.3	84.54
Bacillus cereus AH621	90.58	10.91	79.91
Bacillus mycoides Rock1 4	81.79	15.48	69.39
Bacillus cytotoxicus NVH 391 98	81.63	14.99	79.56
Geobacillus stearothermophilus NUB3621	58.23	0.0	(estimated)
Bacillus subtilis subsp spizizenii TU B 10	57.89	0.0	(estimated)
Geobacillus thermoglucosidasius C56 YS93	57.72	0.0	(estimated)

Growth of RDP

Release 11.4: 3,333,501 sequences



Year	Sequences
1992	636
1993	887
1994	2688
1995	8791