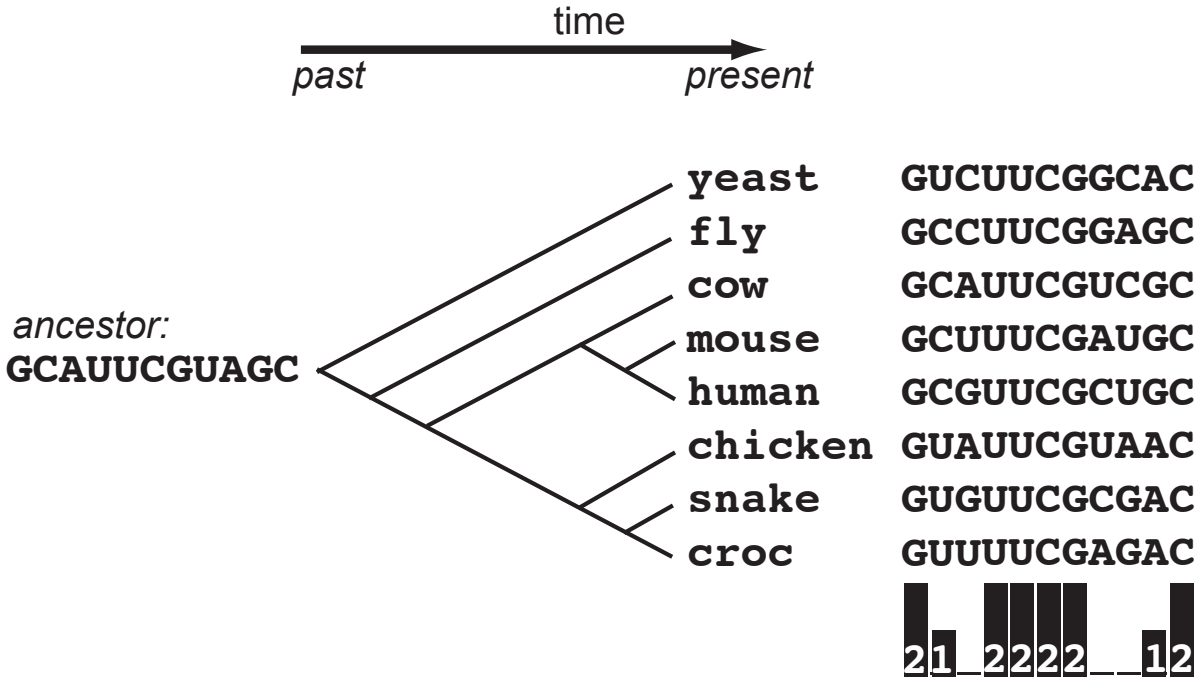
Discovery of Archaeal Group I Introns using Infernal

Eric Nawrocki

National Center for Biotechnology Information
National Institutes of Health

Sequence conservation provides information for homology searches



Each column is independent

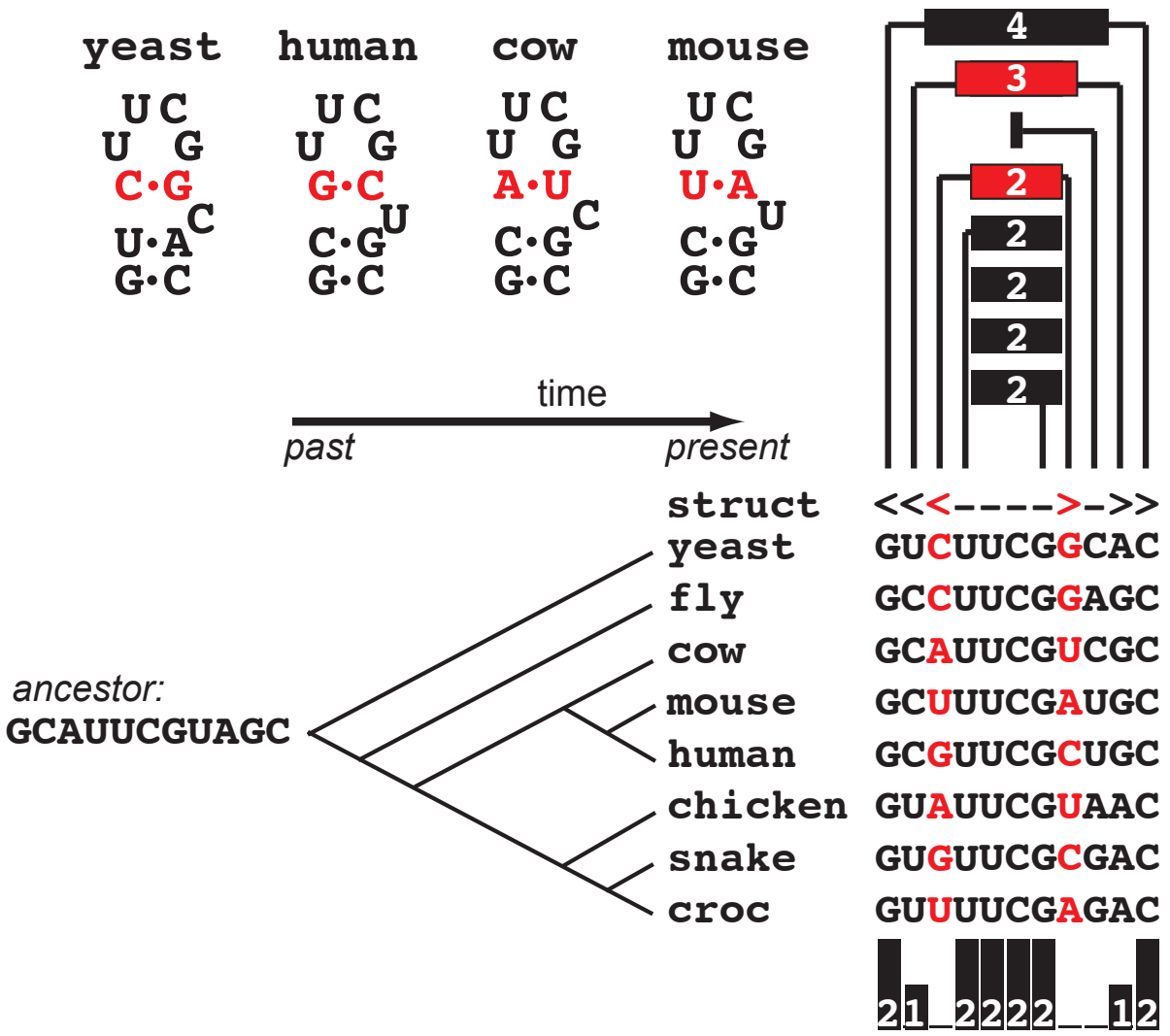
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ bits

sequence profile: 14 bits

expect a match by chance: 1 in 2^{14} nt =~ 16 Kb

Structure contributes additional information from covariation



sequence + structure profile: 17 bits

Base-paired columns are not independent

Maximum ***extra*** info: 2 bits per base pair

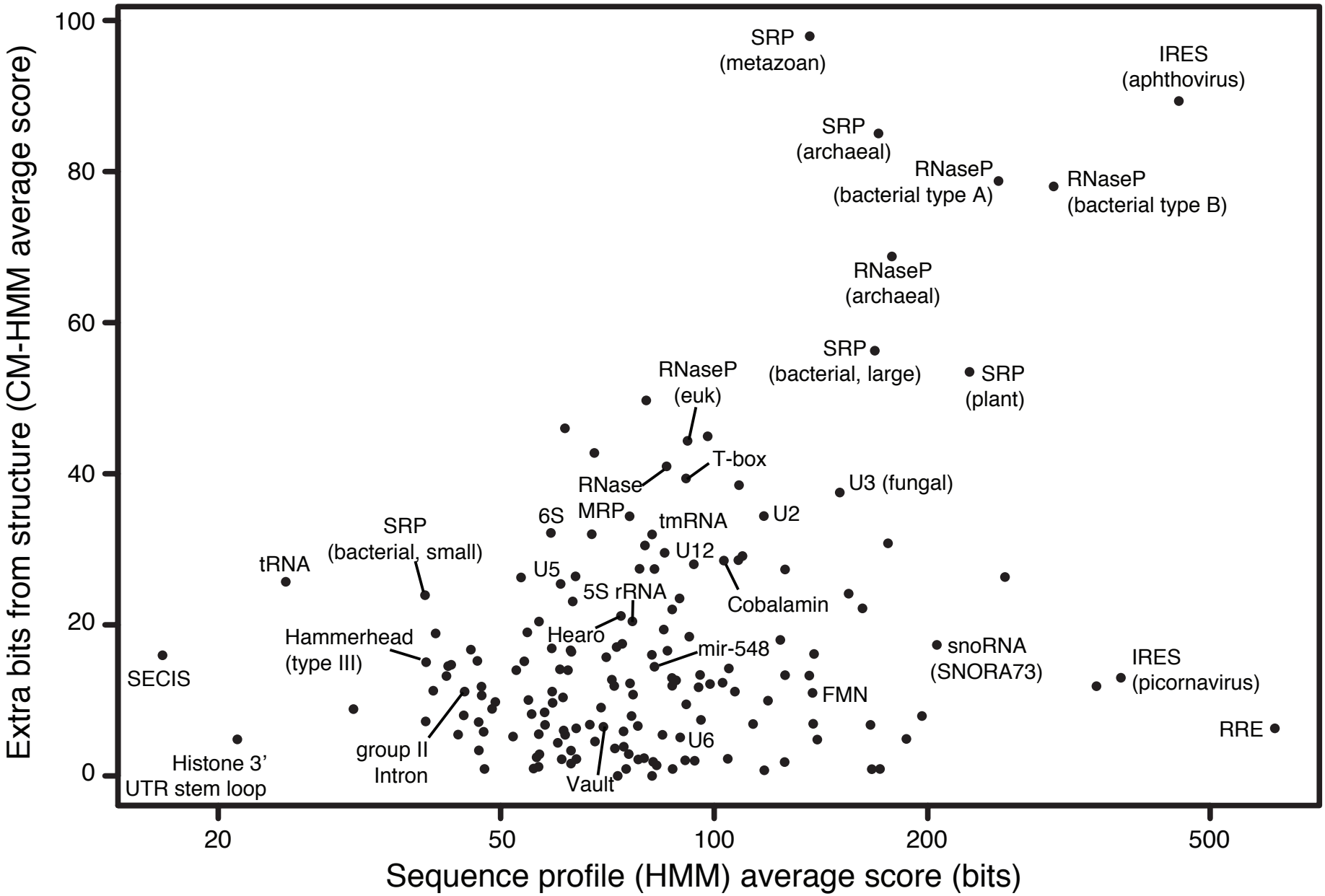
$\log_2 (16/4) = 2$ bits

sequence profile: 14 bits

expect a match by chance: 1 in 2^{17} nt =~ 130 Kb

reducing expected false positives by $2^3 = 8$-fold

Levels of sequence and structure conservation in RNA families



Eddy lab software for profile probabilistic models (since 1994)

	sequence profiles	sequence and structure profiles
models	profile HMMs	covariance models (CMs)
software	HMMER	Infernal
main use	proteins, repetitive DNA elements	structural RNAs
databases	Pfam and Dfam (16712 and 4150 entries)	Rfam (2791 families)
performance for RNAs	faster but less accurate	slower but more accurate

HMMER

http://hmmer.janelia.org

Eddy, SR. PLoS Comp. Biol., 7:e1002195, 2011.

Eddy, SR. PLoS Comp. Biol., 4:e1000069, 2008.

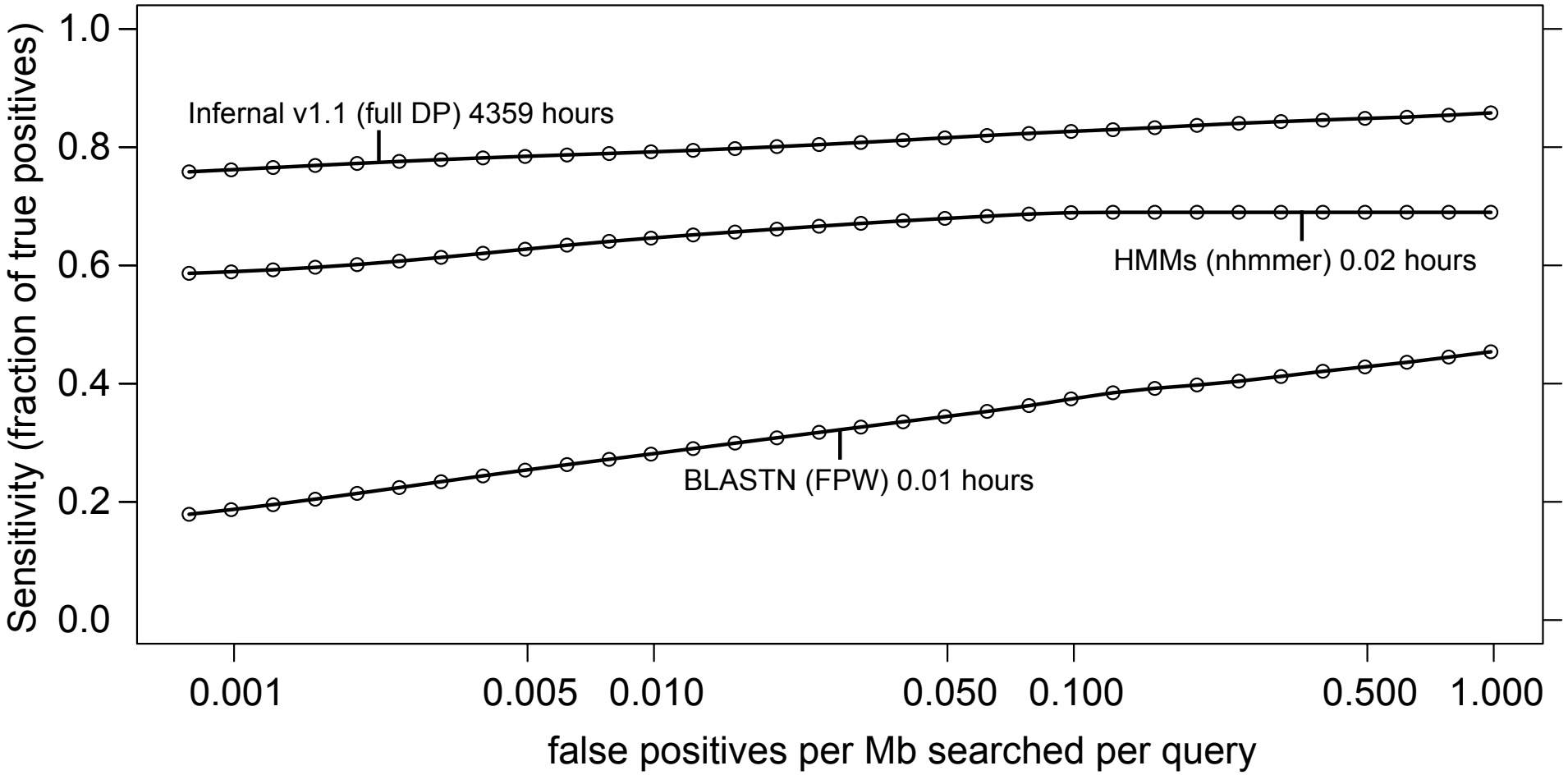
Eddy. SR. Bioinformatics, 14:755-763, 1998.

http://infernal.janelia.org

Nawrocki EP, Eddy SR Bioinformatics, 29:2933-2935, 2013.

Eddy SR, Durbin R. Nucleic Acids Research, 22:2079-2088, 1994.

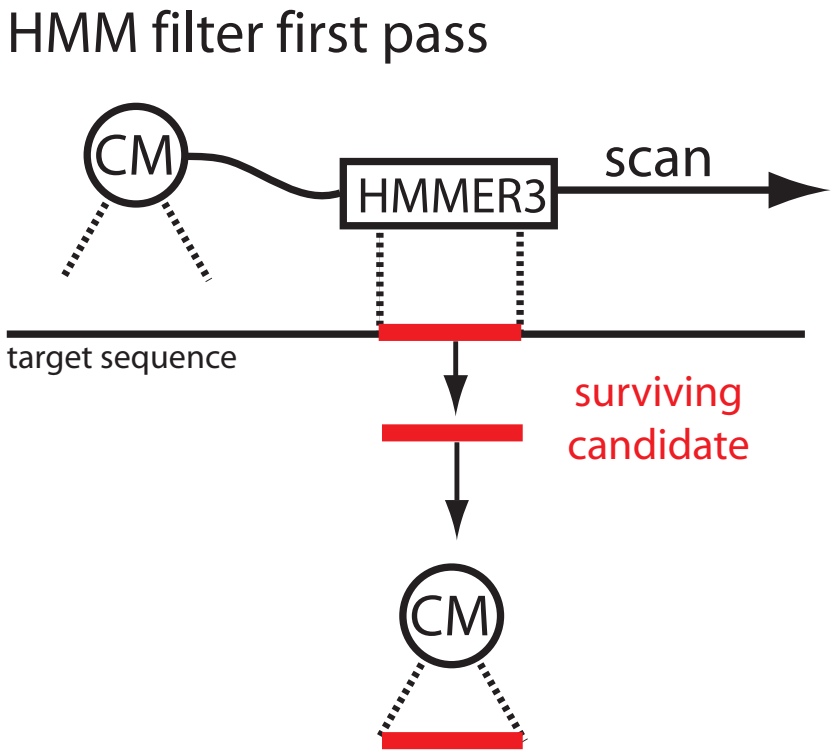
Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)



Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

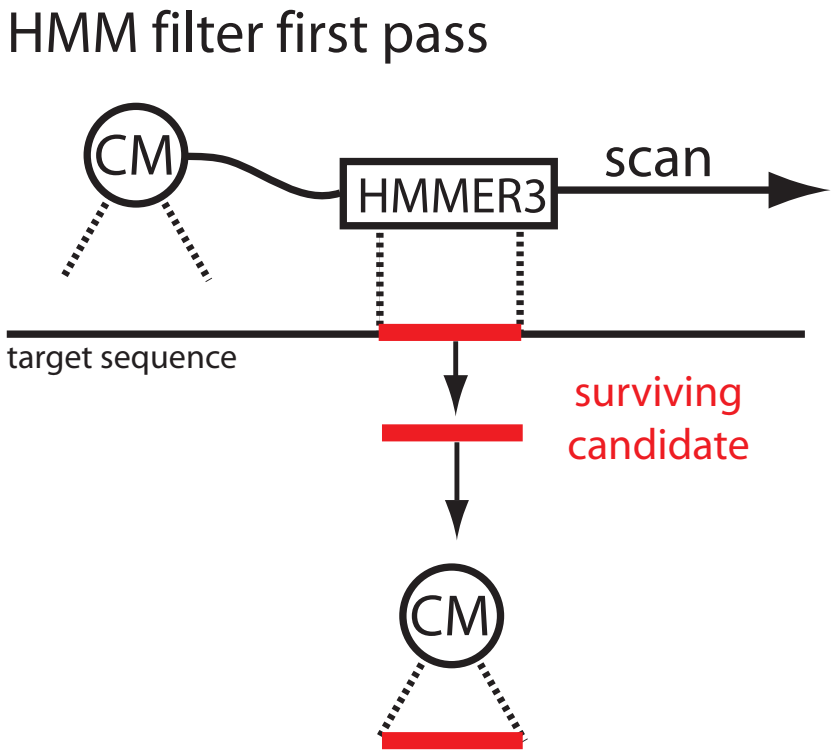
*Freyhult EK, Bollback JP, Gardner PP. Genome Res. 2007 17: 117-125.

Filter target database using profile HMMs*



*Weinberg, Ruzzo, RECOMB, 243-251, 2004; Weinberg, Ruzzo, Bioinformatics, 22(1) 35-39 2006.

Filter target database using profile HMMs*

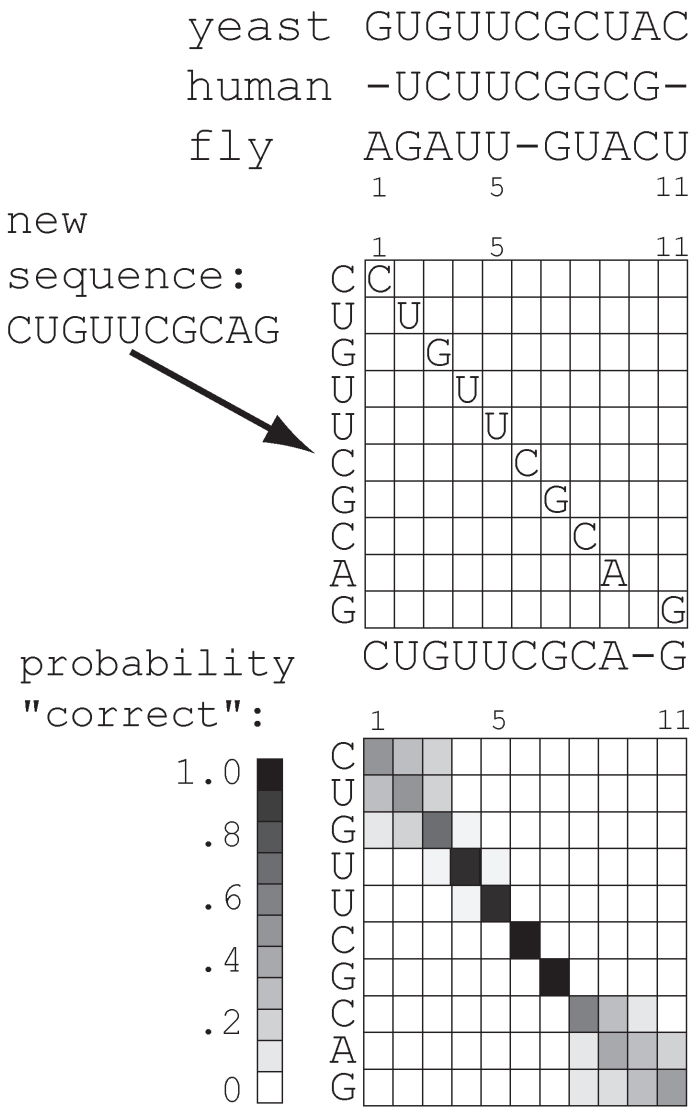


- Even if we filter out 99% of the database (for up to 100X acceleration), searches will still be too slow.

- CM step needs to be accelerated.

*Weinberg, Ruzzo, RECOMB, 243-251, 2004; Weinberg, Ruzzo, Bioinformatics, 22(1) 35-39 2006.

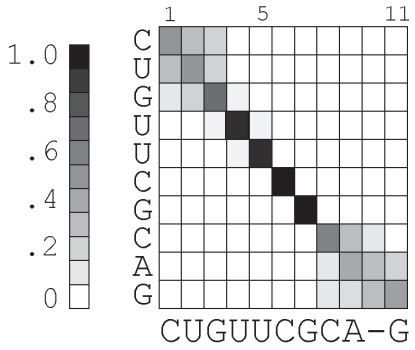
Accelerating CM alignment step 1: HMM posterior decoding to get confidence estimates

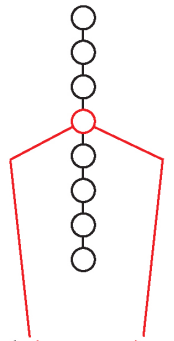


Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*

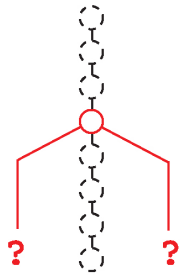
HMMs – Each column of seed alignment corresponds to a column of matrix.

```
yeast GUGUUCGCUAC
human -UCUUCGGCG-
fly   AGAUU-GUACU
```





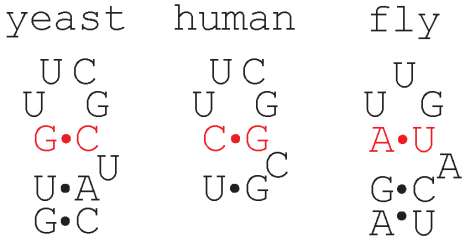
```
struct <<<<----->->>
yeast  GUGUUCGCUAC
human  -UCUUCGGCG-
fly    AGAUU-GUACU
```



CUGUUCGCAG

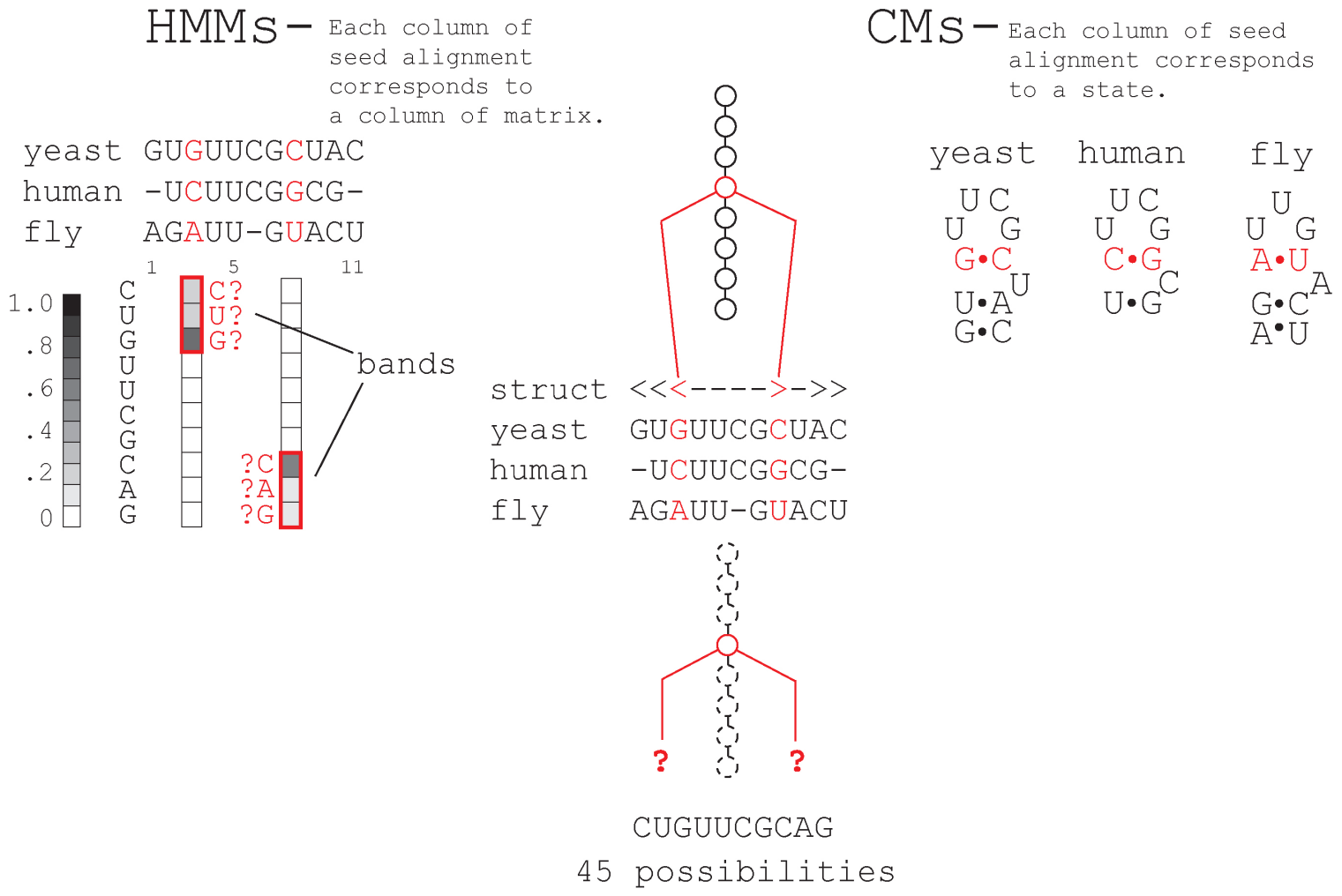
45 possibilities

CMs – Each column of seed alignment corresponds to a state.



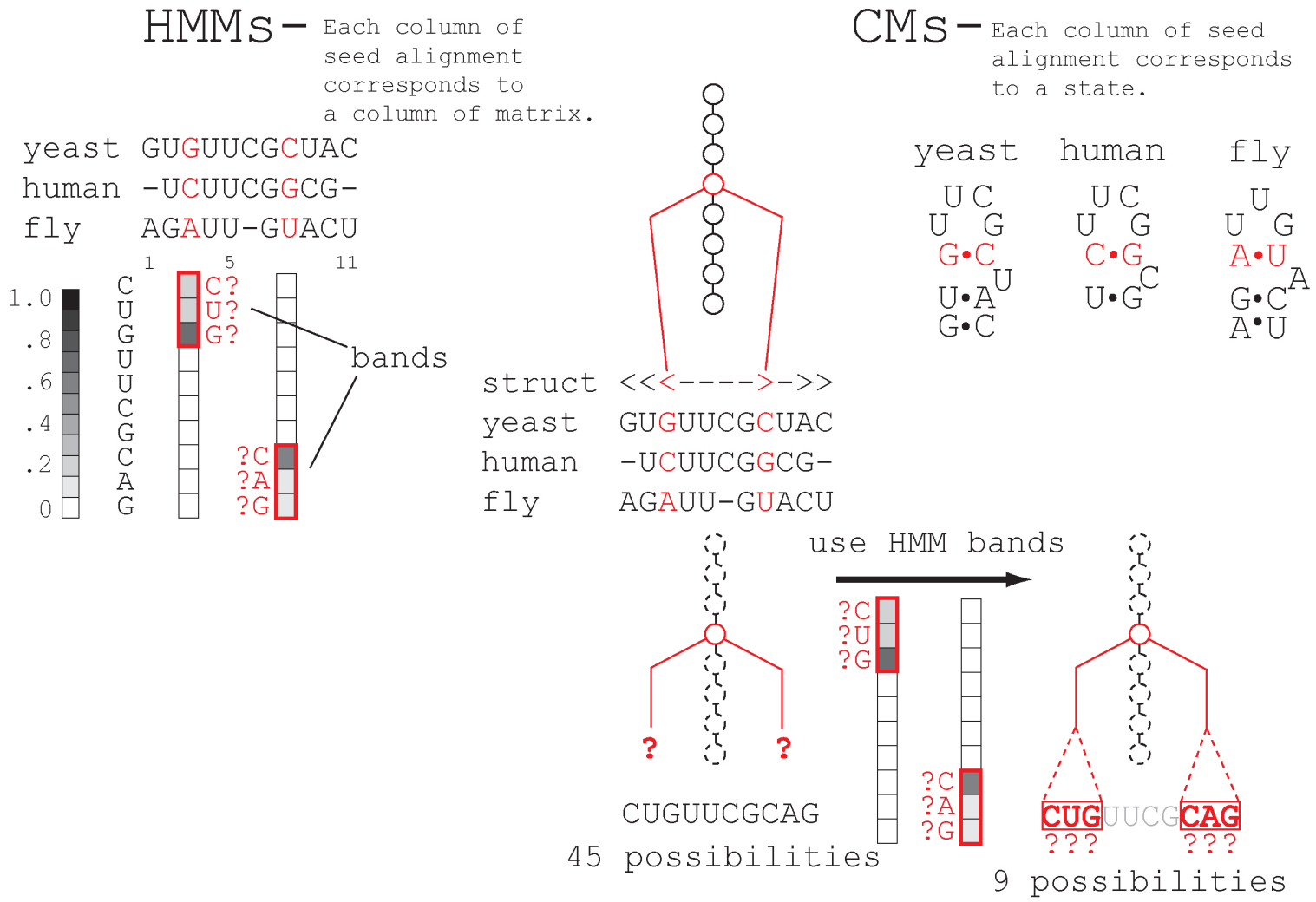
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*



*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

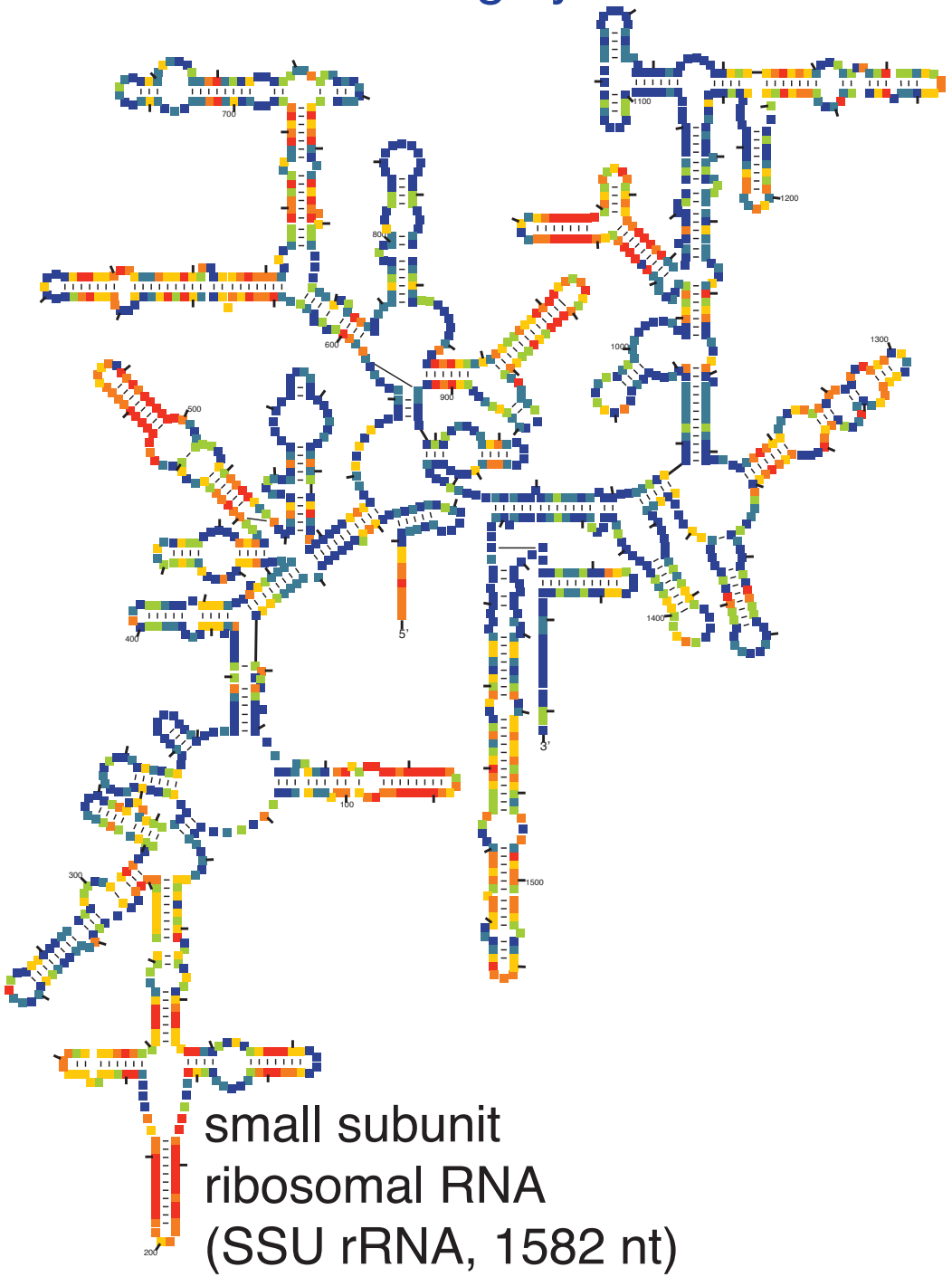
Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*



*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

Sequence conservation per position

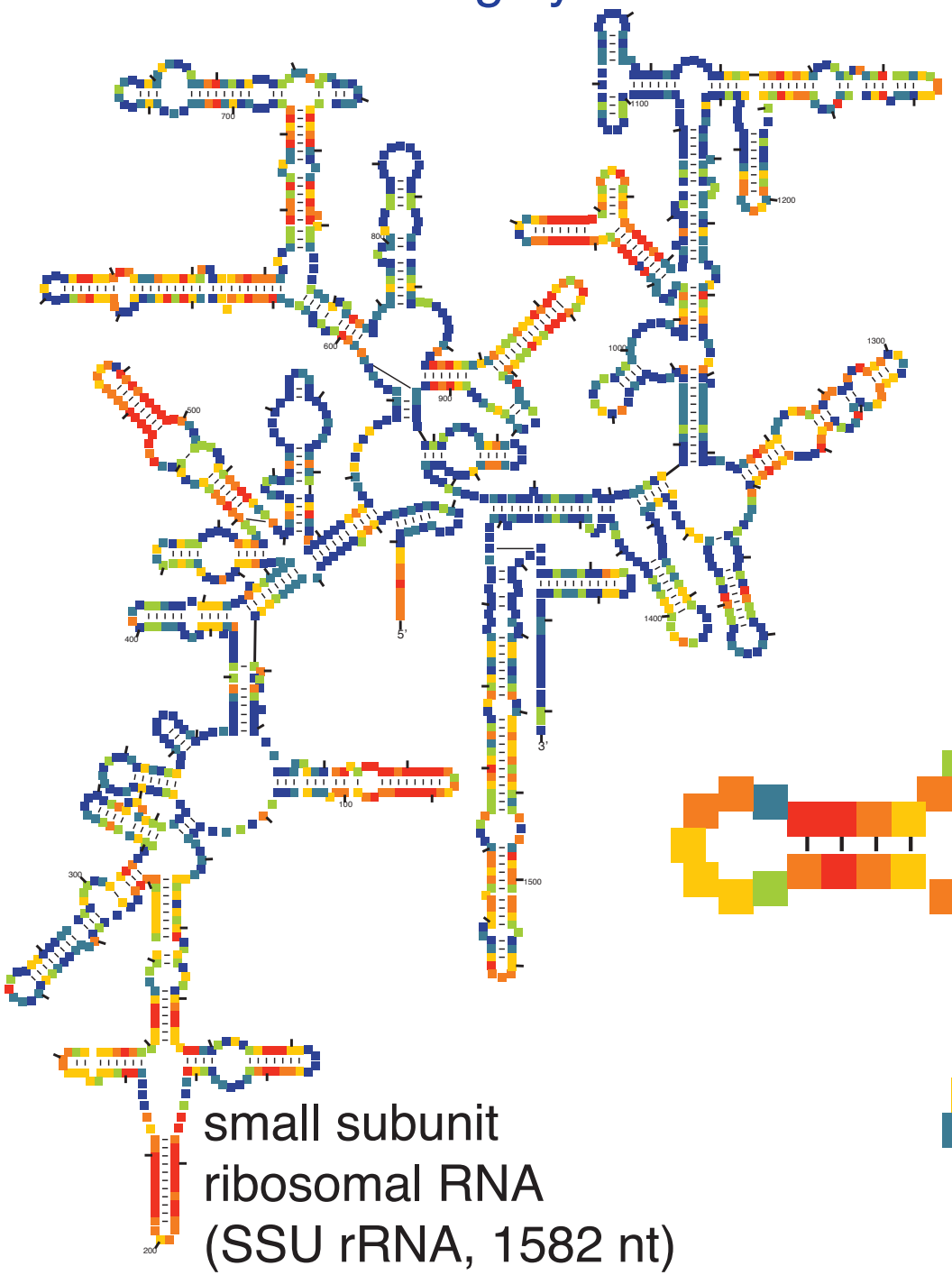
blue:highly conserved ..... red: highly variable



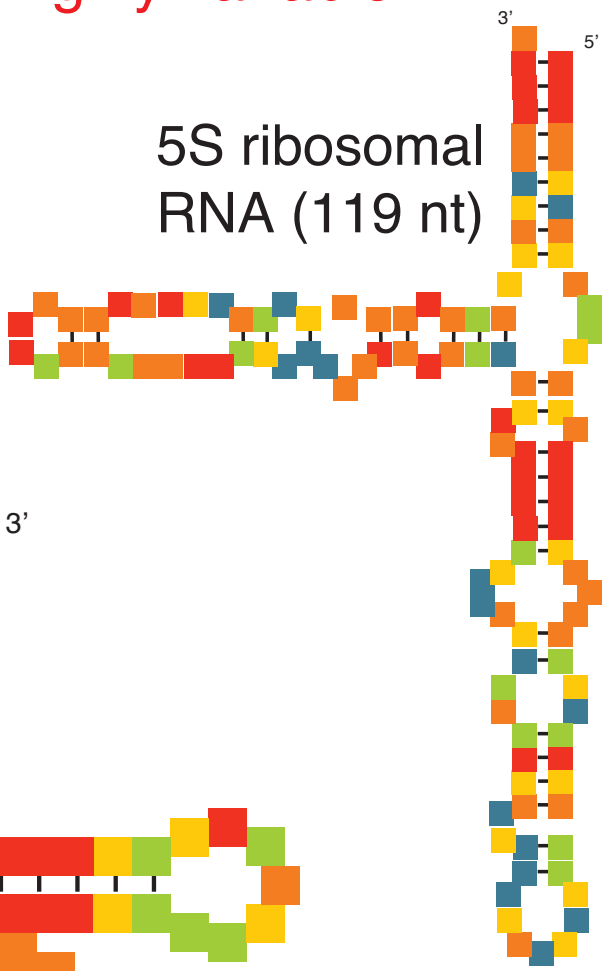
small subunit
ribosomal RNA
(SSU rRNA, 1582 nt)

Sequence conservation per position

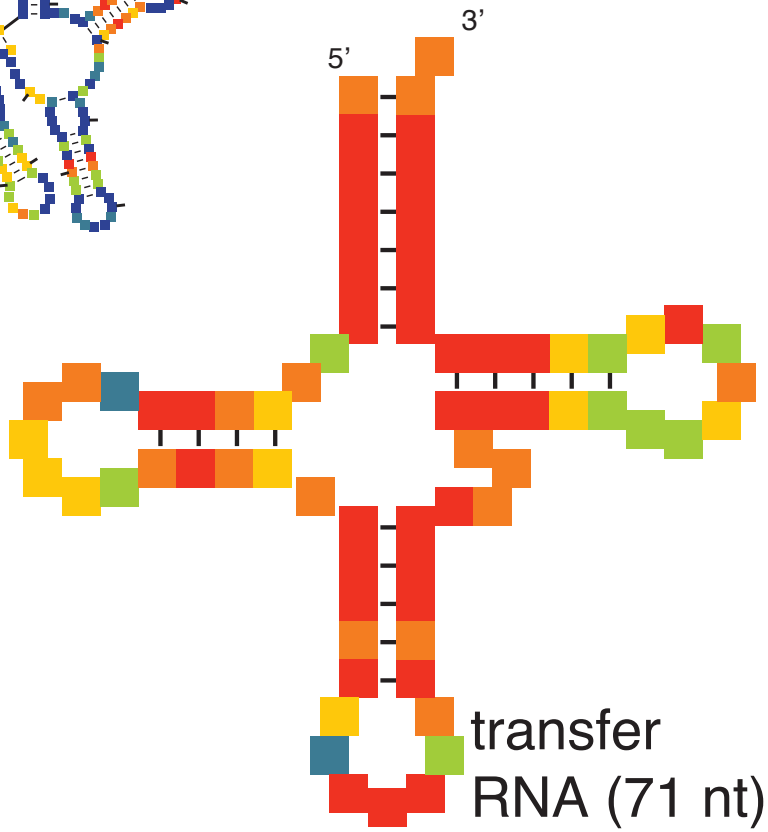
blue:highly conserved ..... red: highly variable



small subunit ribosomal RNA (SSU rRNA, 1582 nt)

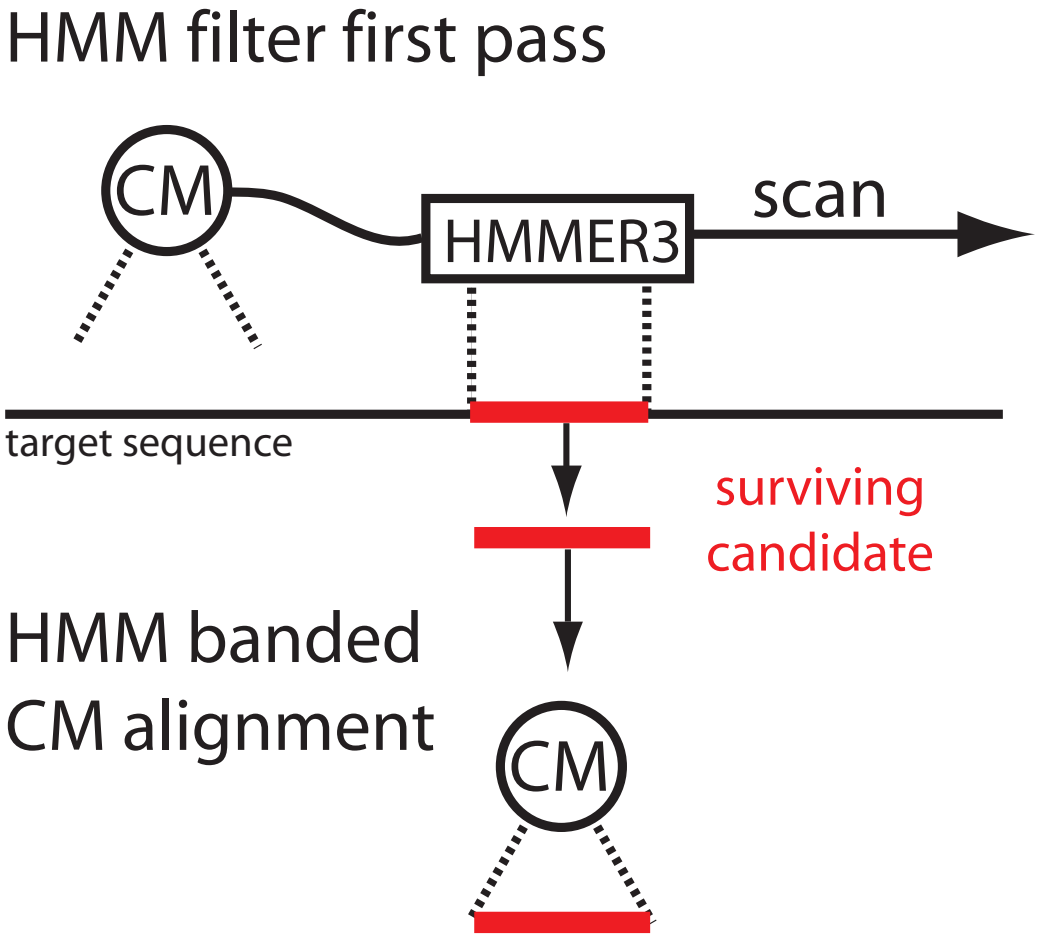


5S ribosomal RNA (119 nt)

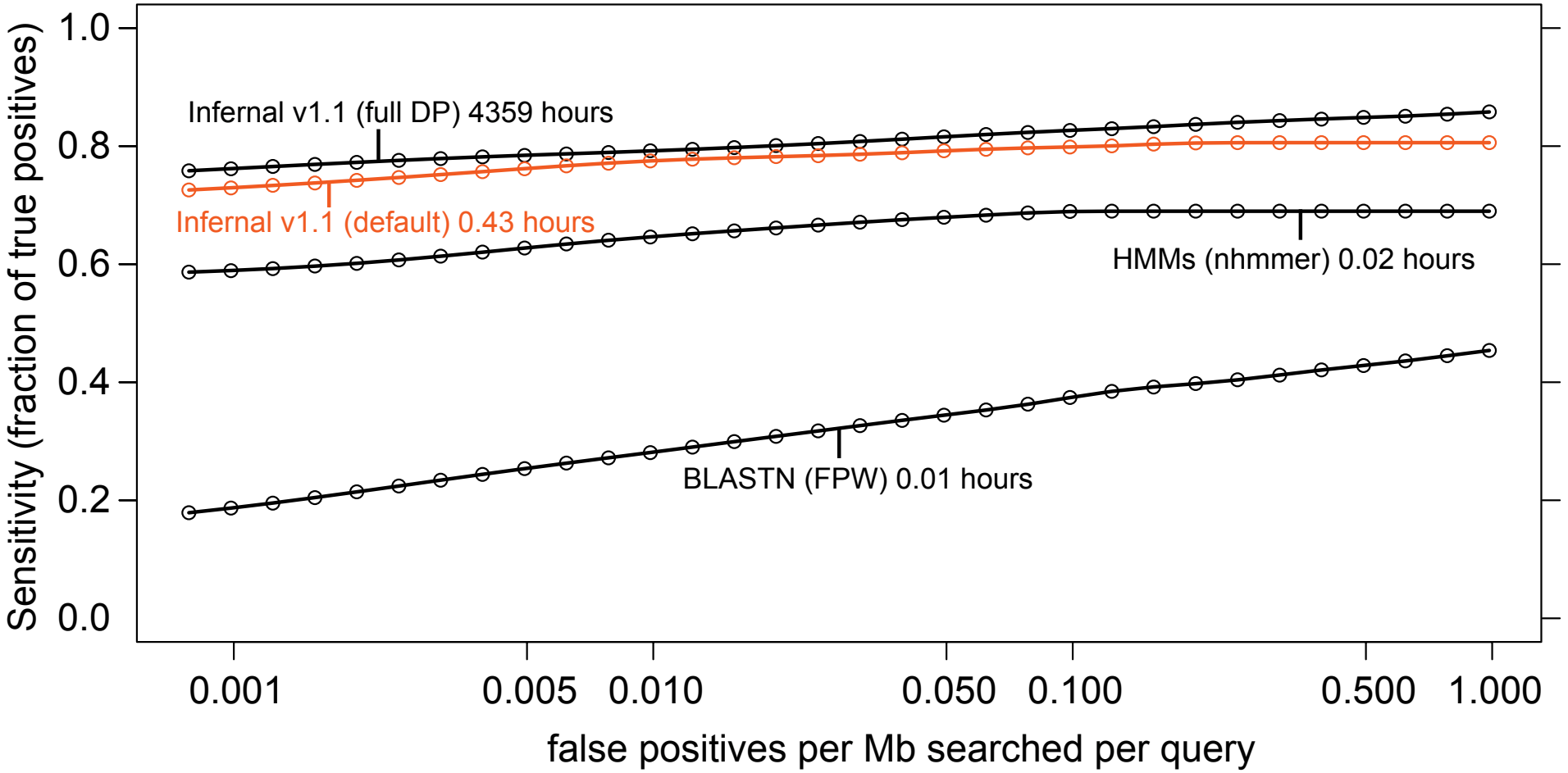


transfer RNA (71 nt)

Use HMMs as filters and to constrain CM alignment



HMM-based acceleration makes Infernal 10,000 times faster



Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

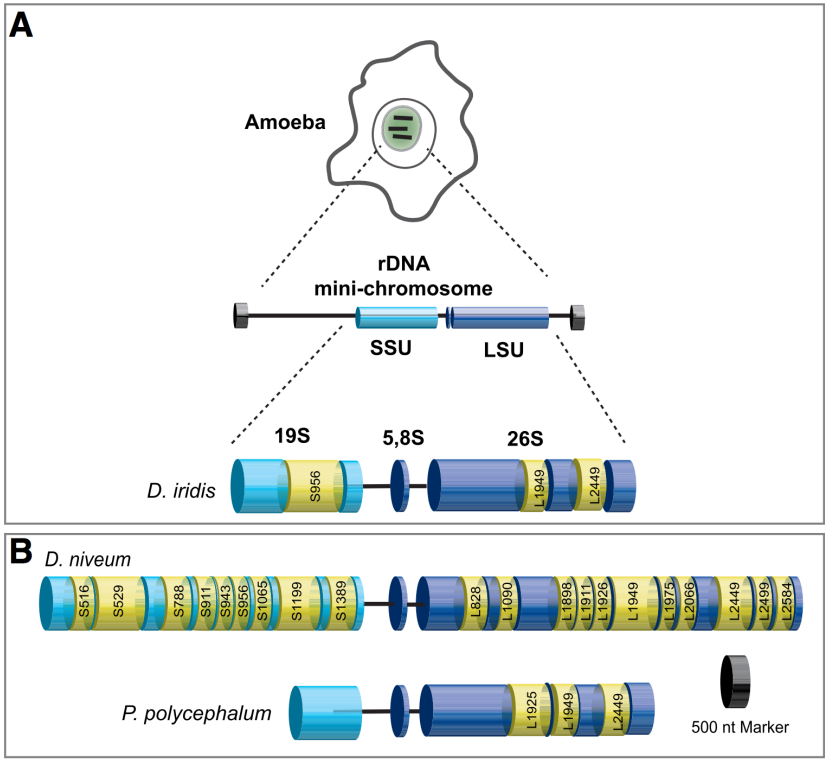
Infernal 1.1 finds 11,000 new group I intron candidates

Table 1. Comparison of the old Rfam 11.0 BLAST and Infernal 1.0 search strategy versus the new Rfam 12.0 Infernal 1.1 search strategy for 15 of 200 randomly chosen families

Accession	Family ID	Length (nt)	#of seed seqs	Time new (h)	Time old (h)	Time (old/new)	New total hits	Old total hits	New unique hits	Old unique hits
Top five families										
RF00028	Intron_gpI	251	12	125.0	357.2	2.8	71 433	60 264	11 175	1
RF00026	U6	104	188	31.2	181.1	5.8	66 517	62 174	4367	14
RF00003	U1	166	100	11.6	64.0	5.5	15 770	14 867	904	1
RF00162	SAM	108	433	8.3	590.0	70.8	4905	4797	108	0
RF00050	FMN	140	144	17.1	169.9	23.9	4381	4306	76	1

Group I catalytic introns

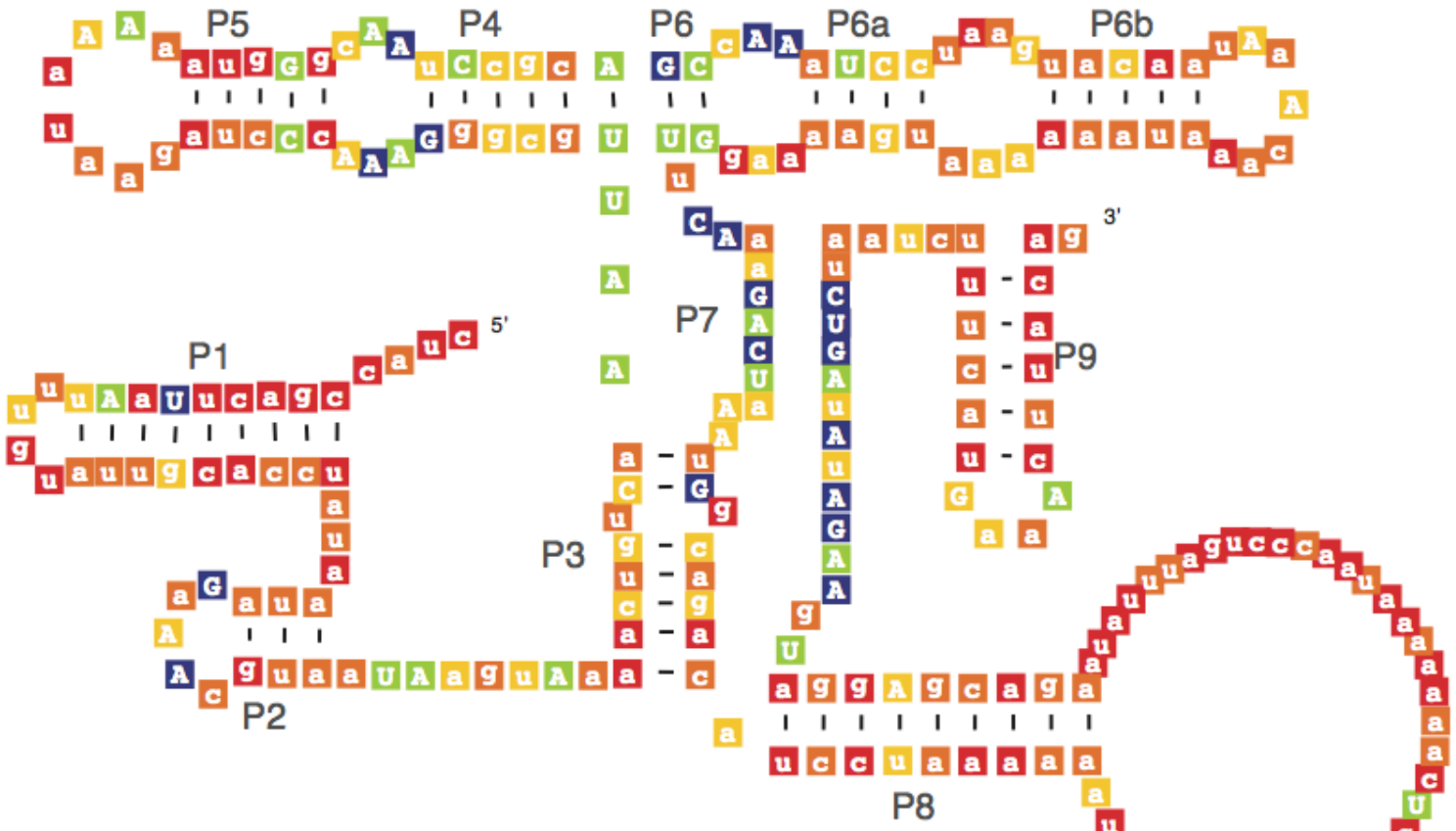
- self splicing ribozymes found in lower eukaryotes, higher plants, bacteria and bacteriophages
- often have ORFs (homing endonucleases) inserted in loop regions
- genes they are found in:
  - bacteria and mitochondria and chloroplast of lower euks: rRNA, mRNA, and tRNAs
  - higher plants mitochondria and chloroplast: a few tRNA and mRNA genes
  - nuclear lower eukaryotic genomes: only rRNA



*

*A. Hedberg and S. D. Johansen, Mobile DNA, 2013 4:17

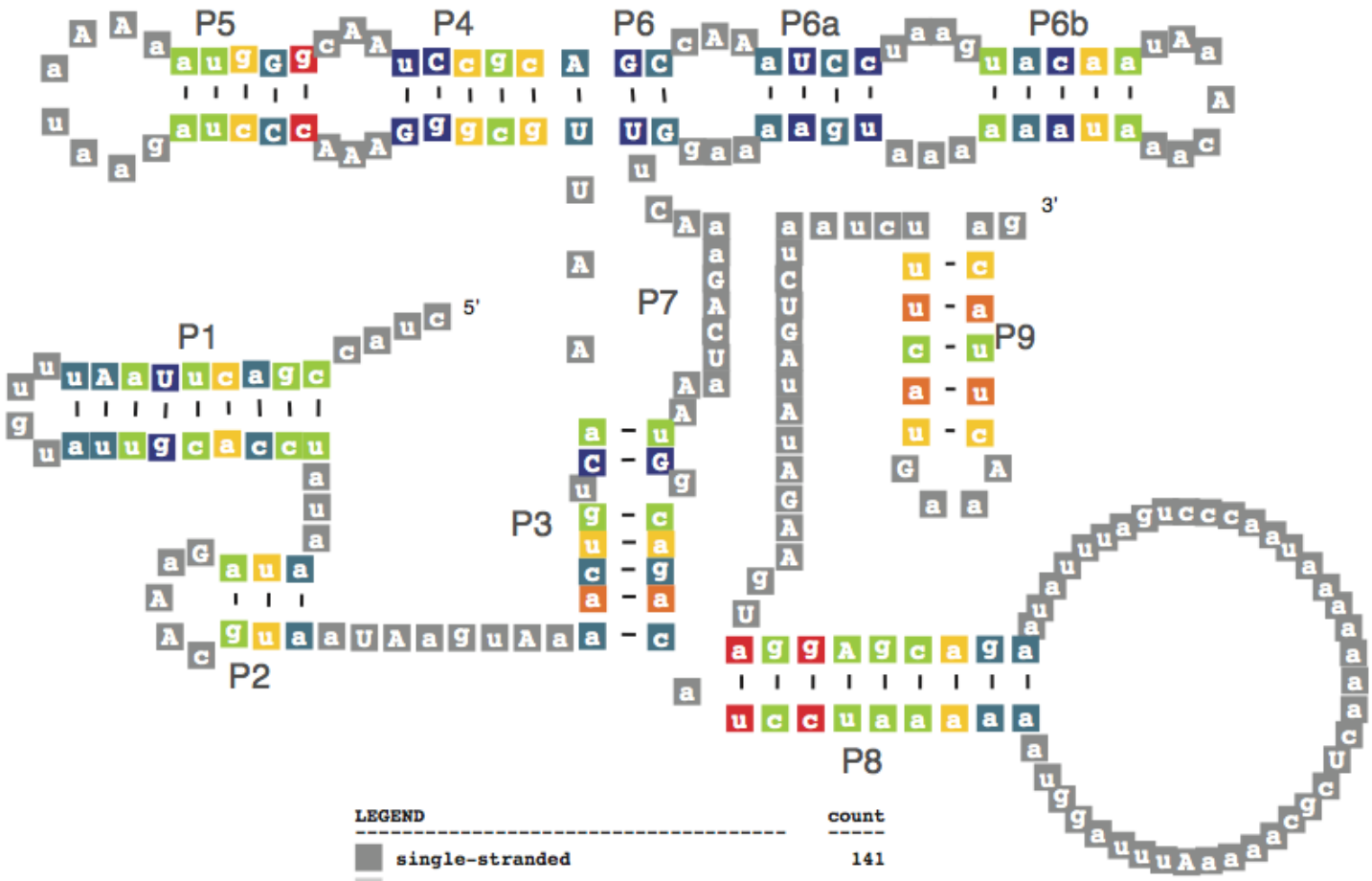
model	#pos	#bps	#seqs	description
Group I Intron	251	55	12	information content per position



LEGEND	count
100% gaps	0
information content (bits):	
[0.000-0.400)	71
[0.400-0.800)	88
[0.800-1.200)	45
[1.200-1.600)	26
[1.600-1.990)	0
[1.990-2.000]	21

Consensus nucleotides (nt) are displayed, defined as the most frequent nt at each position.
Capitalized nts occur in >= 0.75 fraction of sequences that do not have a gap at the position.

model	#pos	#bps	#seqs	description
Group I Intron	251	55	12	mutual information per basepaired position



LEGEND	count
single-stranded	141
0 complete basepairs	0
mutual information per position (bits):	
[0.000-0.167)	16
[0.167-0.333)	28
[0.333-0.500)	34
[0.500-0.667)	20
[0.667-0.833)	6
[0.833-1.000]	6

Consensus nucleotides (nt) are displayed, defined as the most frequent nt at each position.
Capitalized nts occur in >= 0.75 fraction of sequences that do not have a gap at the position.

GISSD*: Group I Intron Sequence and Structure Database

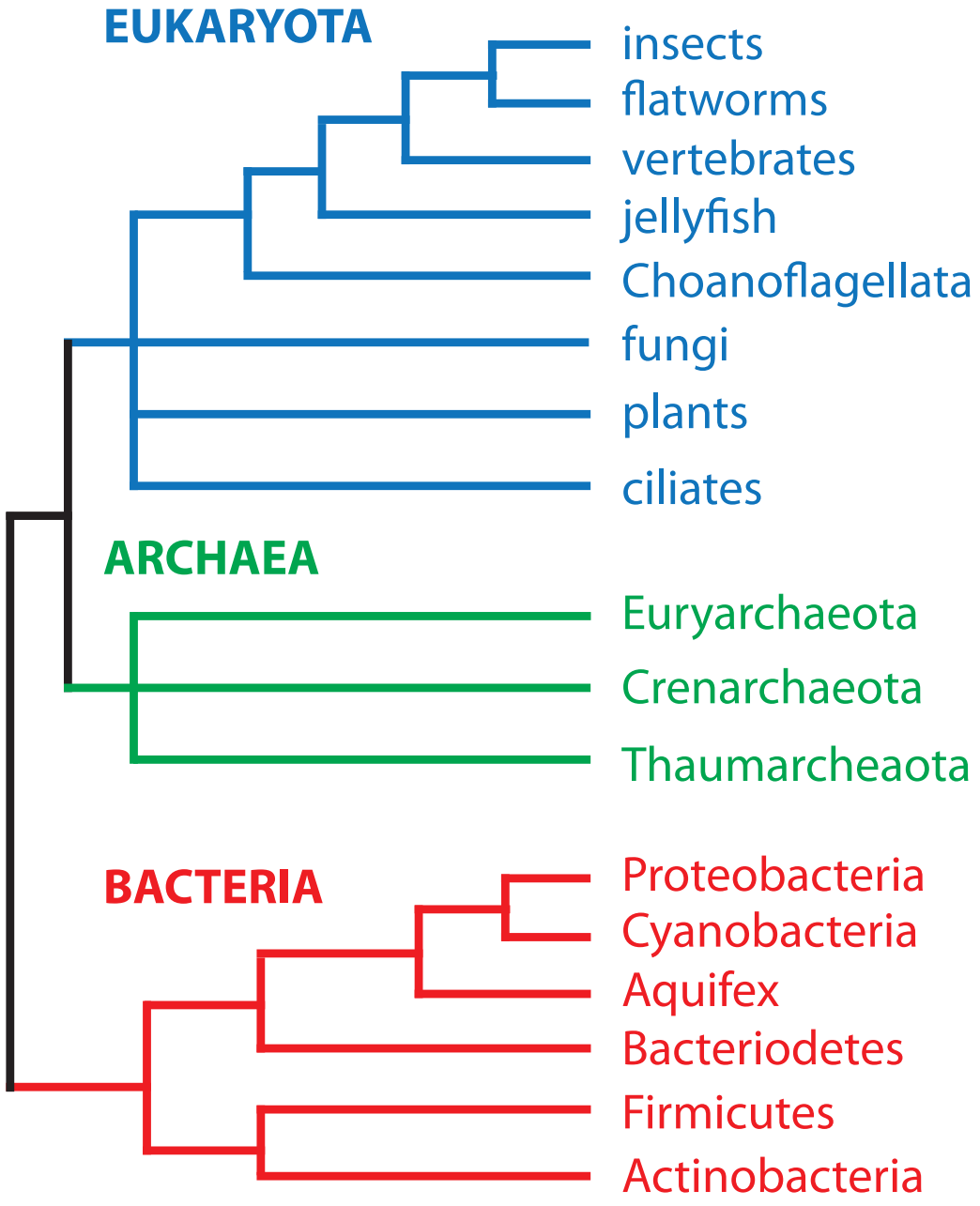
# idx	name	nseq	avlen	%id	nbp	R-scape nsig	bits HMM	bits CM-HMM
1	RF00028	12	364.8	34	61	8	62.6	41.7
2	IA1	76	583.6	45	82	51	166.5	38.0
3	IA2	15	276.8	38	67	24	58.0	47.5
4	IA3	56	282.3	46	81	50	122.4	39.1
5	IB1	42	298.0	72	87	9	320.4	51.8
6	IB2	18	242.2	39	65	27	57.1	39.3
7	IB3	7	277.7	52	72	10	98.5	60.3
8	IB4	89	282.3	44	72	43	108.3	33.2
9	IC1	837	436.0	39	103	84	130.3	40.3
10	IC2	32	320.2	66	86	27	298.1	51.7
11	IC3	328	255.8	67	58	16	244.1	13.2
12	ID	17	242.5	53	66	16	122.8	45.5
13	IE1	38	362.2	60	95	19	268.0	44.8
14	IE2	56	399.9	55	112	38	250.9	47.4
15	IE3	110	405.9	57	119	51	293.4	45.7

*Y. Zhou et. al, NAR, 2008. 36(suppl 1), D31-D37.

Searching Rfamseq with GISSD models

type	# RF00028 seed seqs	# hits total	# hits common	# hits unique
IA1	3	814	385	425
IA2	1	1722	823	899
IA3		958	401	557
IB1		3949	1033	2916
IB2		1861	467	1394
IB3		479	136	343
IB4	1	5717	2400	3317
IC1	3	8475	5385	3090
IC2		4870	3858	1012
IC3	4	72692	66033	6659
ID		572	0	572
IE1		1305	10	1295
IE2		1377	8	1369
IE3		1379	1	1378
total	12	106170*	80940*	16842
RF00028	-	71421	71421	-

* contains overlaps



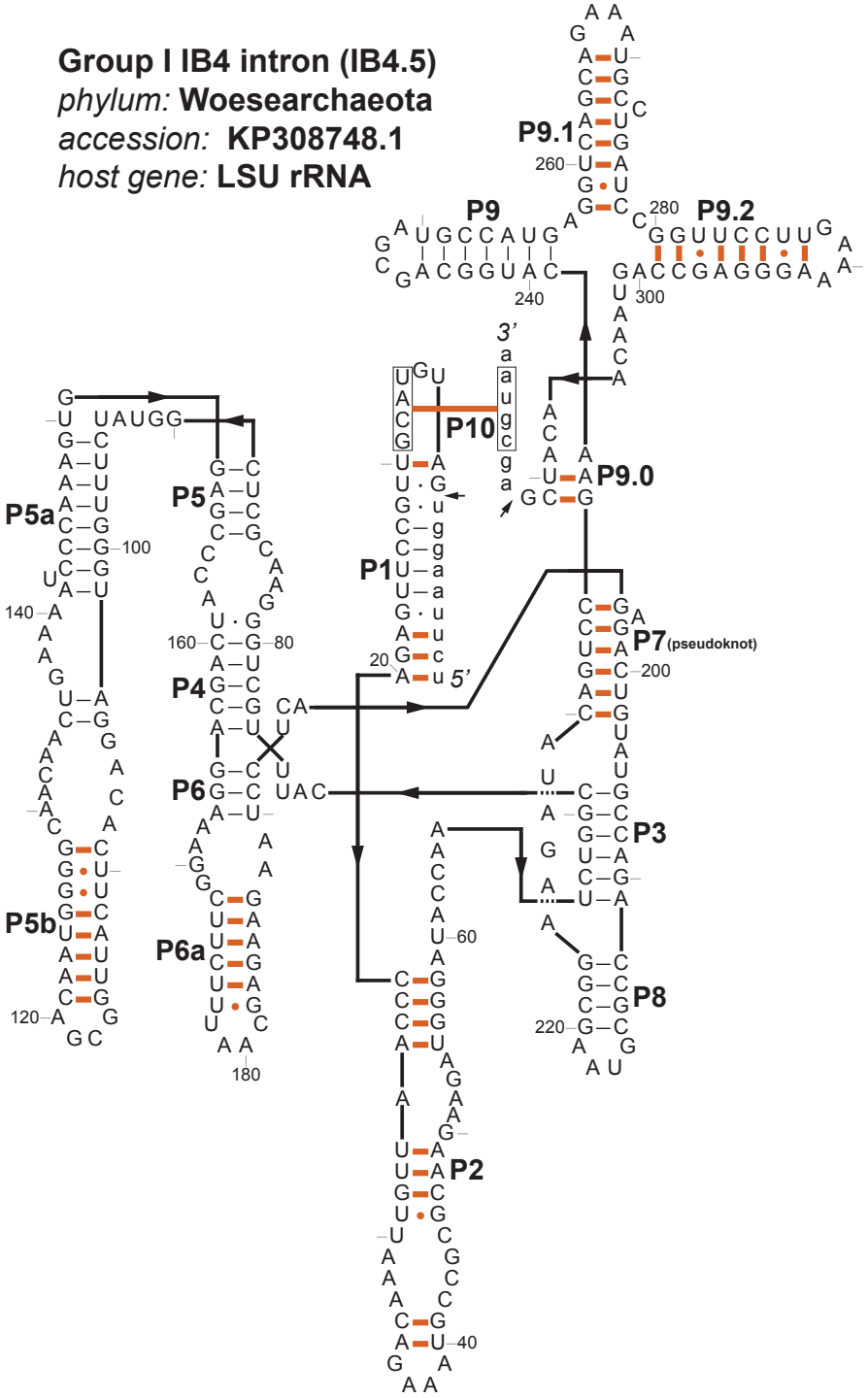
	Group I Introns?	
	previously known	Infernal v1.1 predictions
insects	-	+
flatworms	-	+
vertebrates	-	+
jellyfish	+	+
Choanoflagellata	-	+
fungi	+	+
plants	+	+
ciliates	+	+
Euryarchaeota	-	-
Crenarchaeota	-	+
Thaumarcheaota	-	+
Proteobacteria	+	+
Cyanobacteria	+	+
Aquifex	-	+
Bacteriodetes	-	+
Firmicutes	+	+
Actinobacteria	-	+

Homology searches for group I introns in Archaea

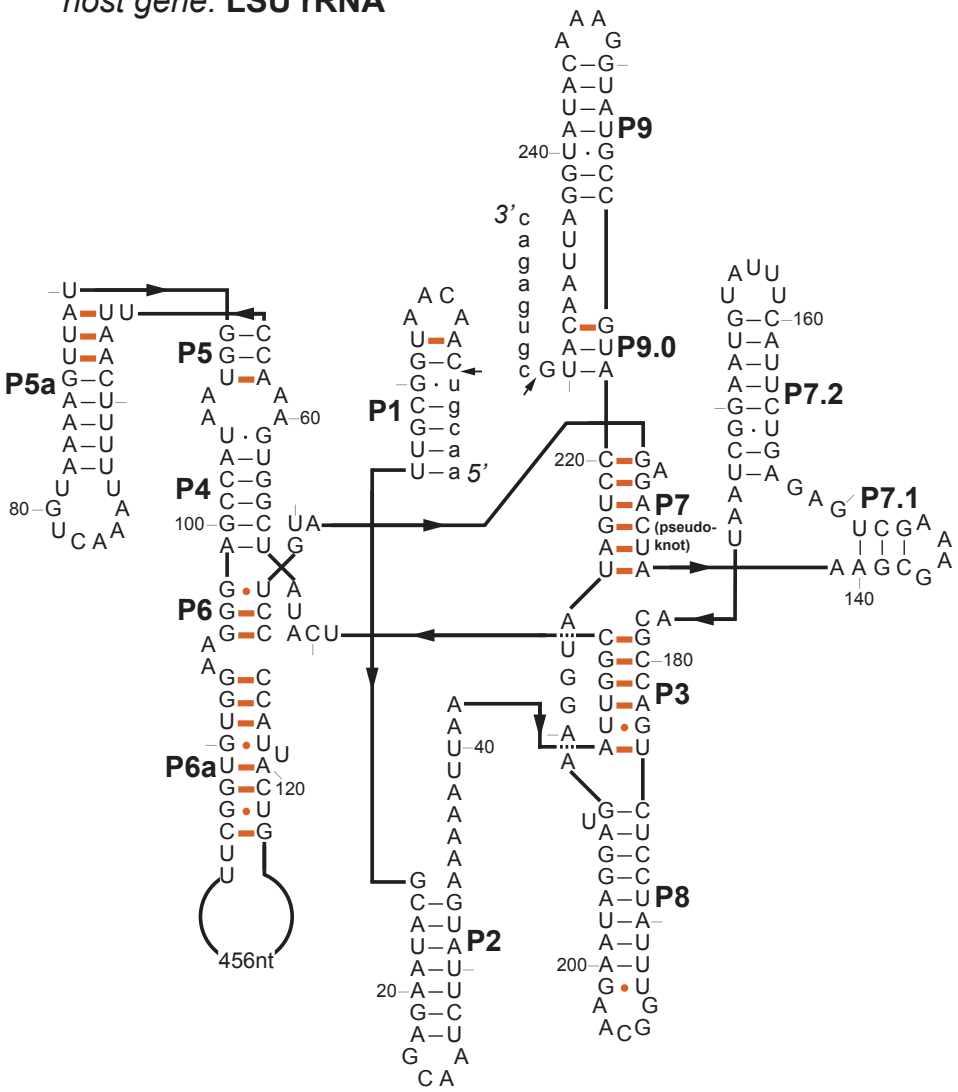
- downloaded all archaeal sequences in GenBank (6.7Gb as of Sept 2017)
- searched archaeal sequences with all GISSD models + RF00028 with default cmsearch parameters and with `--anytrunc`
- 95 non-overlapping hits with $E < 0.01$ corresponding* to 39 group I intron candidates (12 IA3 and 27 IB4)
- 30/39 introns have at least one hit with $E < 10^{-10}$
- 36 within LSU rRNA, 3 within SSU rRNA
- All IA3s are in one of two LSU insertion positions:
  - LSU/2593 (N=10)
  - LSU/2500 (N=2)
- All IB4s are in one of two LSU insertion positions and one SSU position:
  - LSU/1931 (N=15)
  - LSU/1923 (N=9)
  - SSU/1498 (N=3)

*as determined via manual sequence analysis by Tom Jones

Group I IB4 intron (IB4.5)
phylum: **Woesearchaeota**
accession: **KP308748.1**
host gene: **LSU rRNA**



Group I IA3 intron (IA3.1)
phylum: **Woesearchaeota**
accession: **CP010426.1**
host gene: **LSU rRNA**



*

*E. P. Nawrocki, T. A. Jones, and S. R. Eddy, NAR, 2018, gky414

Could archaeal group I introns have evolved into BHB introns?

Evolution of introns in the archaeal world

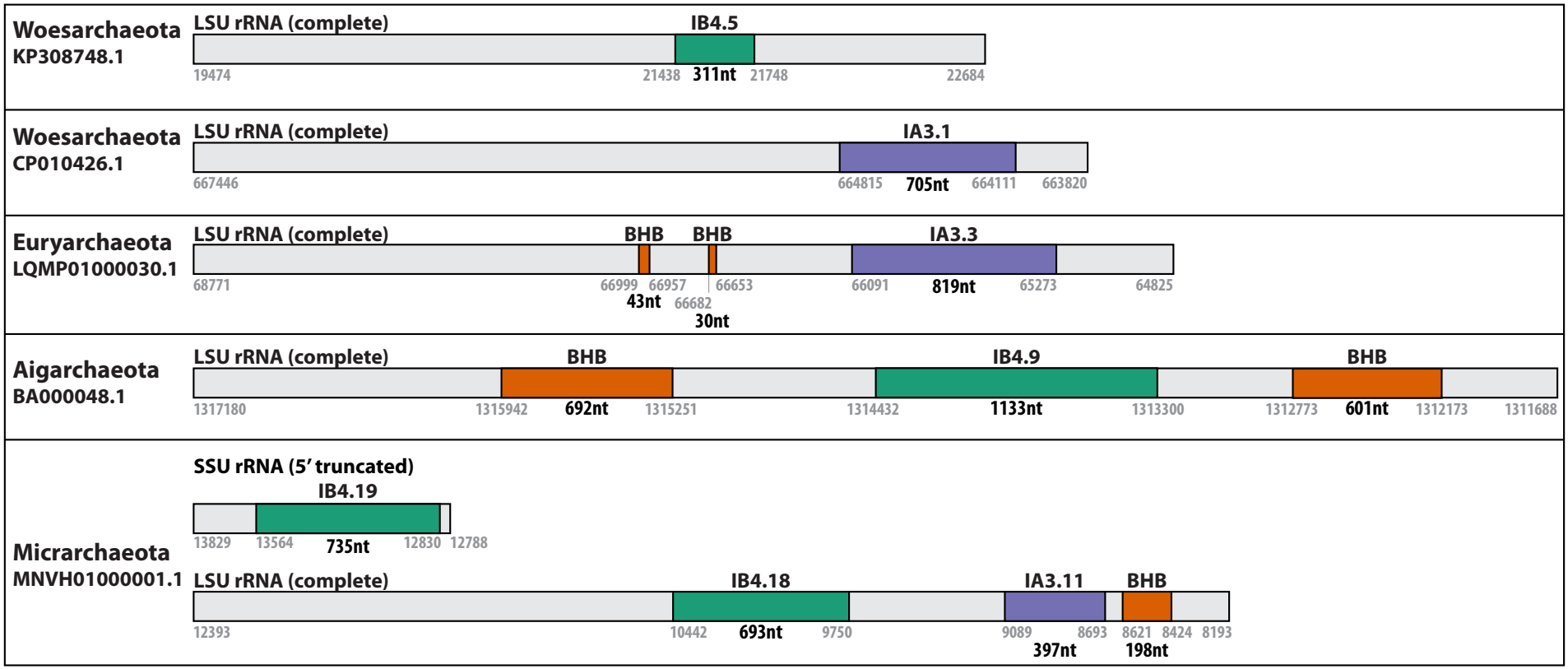
Giuseppe D. Tocchini-Valentini, Paolo Fruscoloni, and Glauco P. Tocchini-Valentini[1]

Istituto di Biologia Cellulare, Consiglio Nazionale delle Ricerche, Campus A, Buzzati-Traverso, Via Ramarini 32, Monterotondo Scalo, 00016 Rome, Italy

Contributed by Glauco P. Tocchini-Valentini, January 24, 2011 (sent for review December 1, 2010)

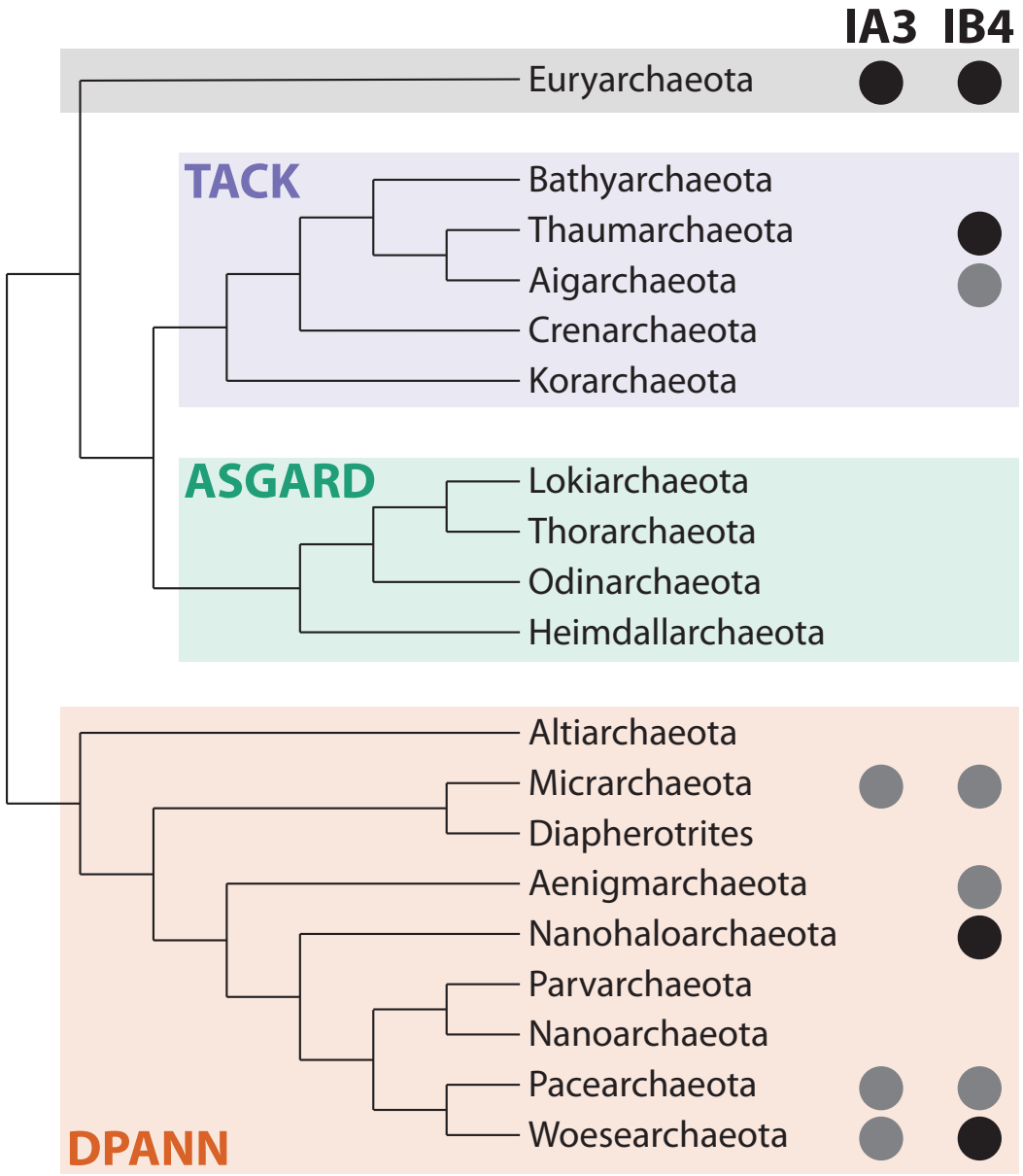
*

Archaeal group I introns can occur in same host gene as BHB introns



*PNAS March 22, 2011. 108 (12) 4782-4787;

Group I introns are widespread in Archaea



Acknowledgements

Harvard/Janelia
Sean Eddy
Tom Jones
Diana Kolbe
Travis Wheeler
Elena Rivas
Michael Farrar

NCBI
Alejandro Schäffer
David Landsman
Jim Ostell
David Lipman

GISSD
Yu Zhou
Chen Lu
Qi-Jia Wu
Yu Wang
Zhi-Tao Sun
Jia-Cong Deng
Yi Zhang

David Haussler's group
Yasu Sakakibara
(CM-like models in 1994)


Thank you
Elena and Eric!