Identifying Conserved RNA Structures in Viruses using Rfam and Infernal

Eric P. Nawrocki(1), Ioanna Kalvari(2), Joanna Argasinska(2), Anton I. Petrov(2) and Sean R. Eddy(3)

1: National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA. 2: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. 3: Howard Hughes Medical Institute, FAS Center for Systems Biology, John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

nawrocke@ncbi.nlm.nih.gov

RNA homology searches based on sequence and structure

Functional RNAs do not encode proteins, but rather function directly as RNAs. Many of these RNAs form stable, evolutionarily conserved three-dimensional structures that are crucial to their functions in various fundamental cellular processes including protein synthesis, gene expression, splicing, protein transport, and more.

Functional RNAs, with the exception of transfer RNAs and ribosomal RNAs are rarely annotated in sequences in databases like GenBank and RefSeq, at least partly because finding homologs of structural RNAs is challenging: the sequences are often short (100-200 nt), lack ORFs, and have regions of high sequence variability even while conserving their three-dimensional structure. The Infernal software package is one of the most successful approaches for RNA homology search because it takes advantage of both sequence and secondary structure conservation [1]. Infernal is the software behind the Rfam database [2] which contains information for more than 2500 RNA families.

The most prevalent viral genomes in GenBank

						RefSeq		Rfam	
species	#seqs	RefSeq accn	family	type	host	CDS	RNA	#	types
Influenza	503,115	NC_007370	Orthomyxoviridae	(-)ssRNA	humans+	11	-	1	Cis-reg
Rotavirus A	58,405[a]	NC_011503	Reoviridae	dsRNA	humans	12	-	1	Cis-reg
Hepatitis B	9211	NC_011503	Hepadnaviridae	dsDNA-RT	humans	7	-	1	Cis-reg
Dengue	4853	NC_001477	Flaviviridae	(+)ssRNA	humans	1	-	5	Cis-reg

HIV-1	2597	NC_001802	Retroviridae	ssRNA-RT	humans	10	-	10	Cis-reg(8)
									miRNA(1), FSE(1)

Hepatitis C	2185	NC_004102	Flaviviridae	(+)ssRNA	humans	2	-	6	Cis-reg(5), IRES(1)
Porcine circovirus	1905	NC_005148	Circoviridae	ssDNA	pigs	3	-	-	-
West Nile	1667	NC_009942	Flaviviridae	(+)ssRNA	humans	3	-	6	Cis-reg(5), FSE(1)
Ebola	1384	NC_002549	Flaviviridae	(+)ssRNA	humans	9	-	-	-
Enterovirus A	1222	NC_001612	Picarnoviridae	(+)ssRNA	humans	1	-	3	Cis-reg(2), IRES(1)
RSV	1122	NC_001781	Orthopneumovirus	(-)ssRNA	humans	11	-	-	-
Norwalk virus	1009	NC_029646	Calciviridae	(+)ssRNA	humans	3	-	1	Cis-reg
Maize streak virus	884	NC_001346	Geminiviridae	ssDNA	plants	4	1	-	-
Rabies lyssavirus	826	NC_001542	Rhabdoviridae	(-)ssRNA	humans+	5	-	-	-
Enterovirus C	765	NC_002058	Picarnoviridae	(+)ssRNA	humans	1	13	3	Cis-reg(2), IRES(1)
HPV 16	764	NC_001526	Papillomaviridae	dsDNA	humans	9	-	-	-

Table 1: *Attributes of the 15 viruses with the most genome sequences in GenBank as of March, 2018, and number of annotations in corresponding RefSeq entries.*

[a] sum of 11 segments

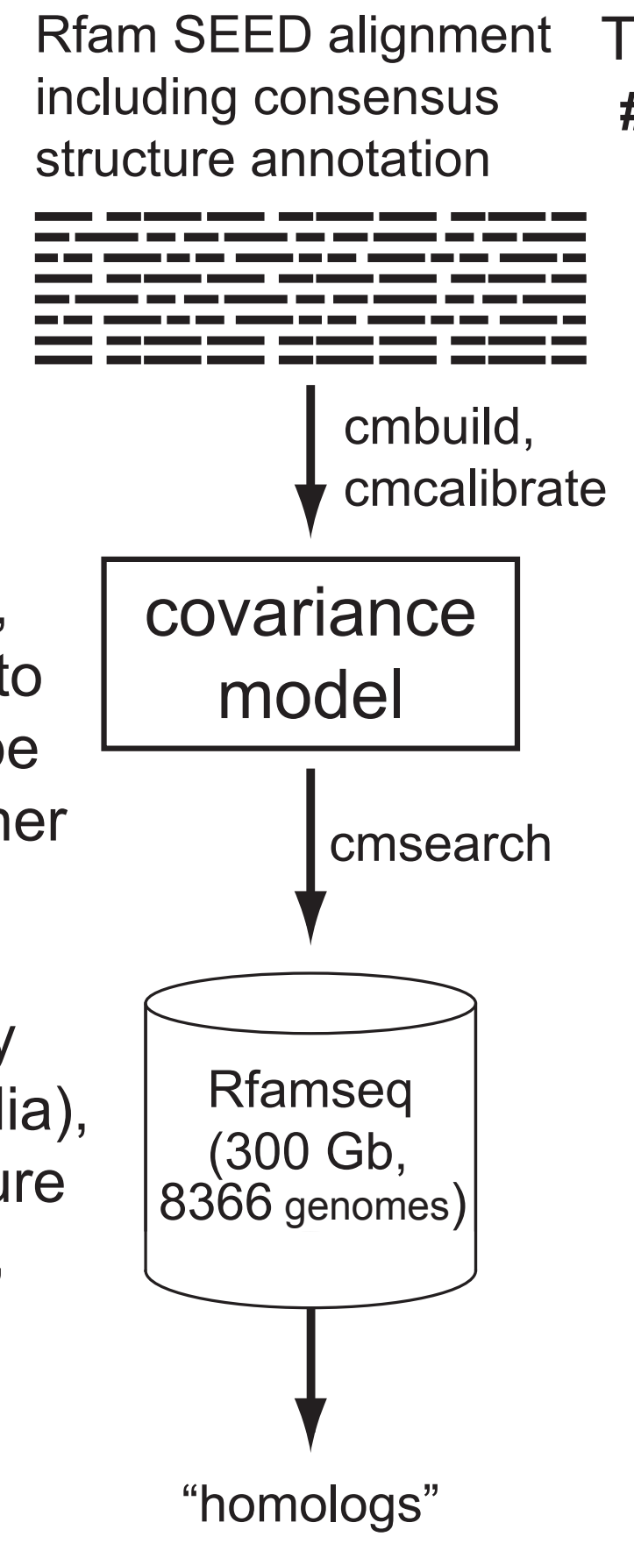
Rfam: the RNA families database [2]

Rfam contains 2686 families, each represented by a
- SEED alignment
- covariance model (CM)
- list of annotated hits in Rfamseq database (2,272,100 total hits)

If you submit a family to Rfam, these tools become available to the community, allowing it to be annotated in genomes and other datasets.

Rfam also includes community annotation of families (wikipedia), RNA motifs, secondary structure diagrams, alignment statistics, taxonomic information, and sequence search capability.

rfam.xfam.org


Rfam SEED alignment including consensus structure annotation
cmbuild, cmcalibrate
covariance model
cmsearch
Rfamseq (300 Gb, 8366 genomes)
“homologs”

Types of RNAs in Rfam

#fams all	#fams viral	#hits viral	type
241	114	757	Cis-reg;
28	21	116	Cis-reg; frameshift-element;
32	10	127	Cis-reg; IRES;
30	2	3	Cis-reg; leader;
33	1	18	Cis-reg; riboswitch;
31	1	19	Cis-reg; thermoregulator;
74	7	155	Gene;
38	2	46	Gene; antisense;
11	0	-	Gene; antitoxin;
64	0	-	Gene; CRISPR;
219	0	-	Gene; lncRNA;
529	15	52	Gene; miRNA;
30	4	60	Gene; ribozyme;
14	0	-	Gene; rRNA;
3	1	4	Gene; snRNA;
470	1	3	Gene; snRNA; snoRNA; CD-box;
269	0	-	Gene; snRNA; snoRNA; HACA-box;
29	0	-	Gene; snRNA; snoRNA; scaRNA;
15	0	-	Gene; snRNA; splicing;
513	3	6	Gene; sRNA;
2	1	5330	Gene; tRNA;
9	1	141	Intron;
2686	184	6837	

Conserved sequence and structure as statistical signals

The conserved sequence and secondary structure of RNAs offers two statistical signals that can be harnessed when searching databases for homologs using CMs. In Figure 1 below, the amount of information, measured in *bits*, inherent in a sequence-only profile (14 bits) and a sequence-and-structure profile (17 bits) is shown for a toy example of an RNA family. We expect a match to a sequence-only profile for this family once in every $2^{14} = 16,384$ random nucleotides. Additionally modeling structure with a sequence-and-structure based profile (like a CM) reduces this probability 8-fold, to once every every $2^{17} = 131,072$ random nucleotides.

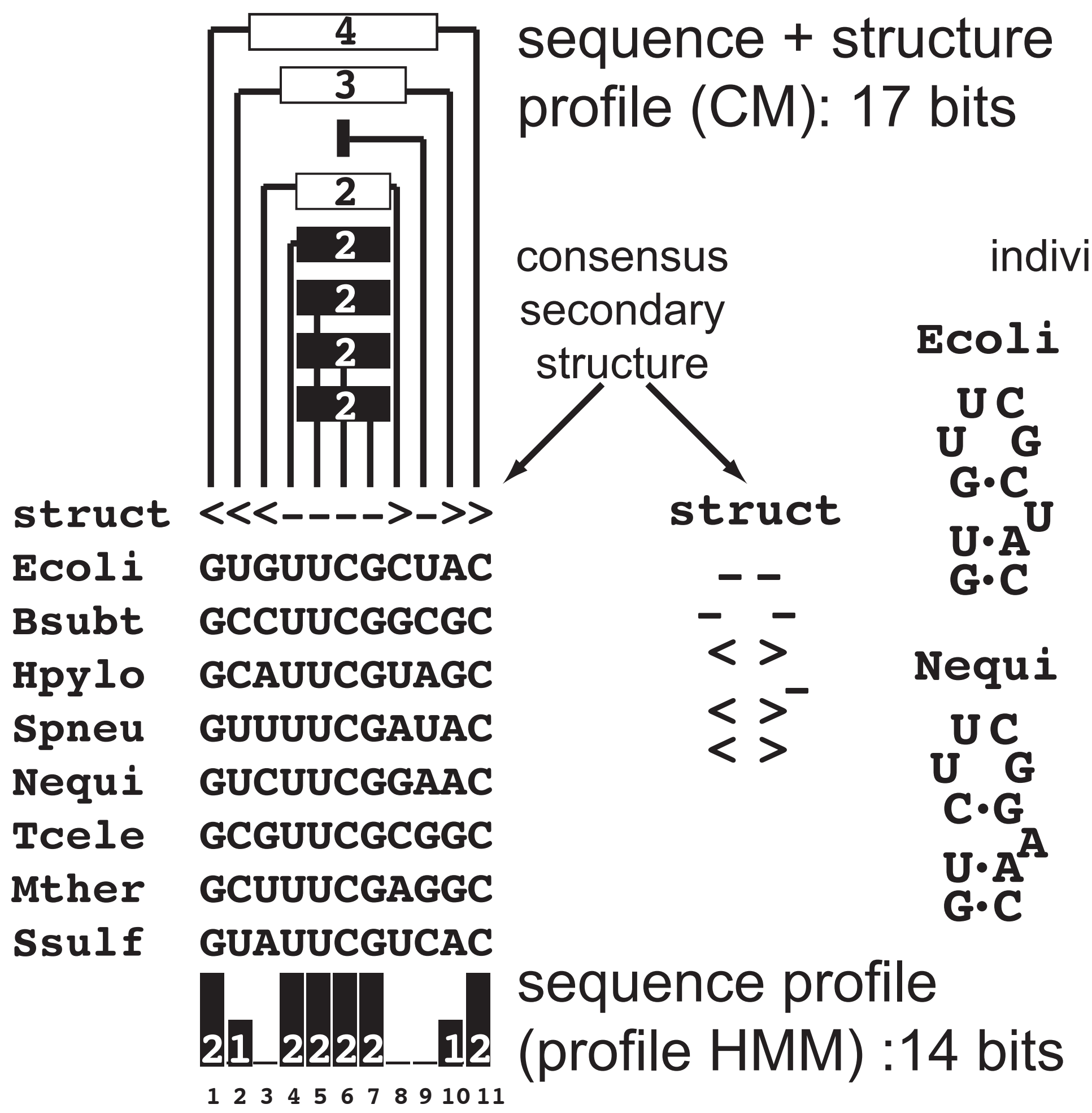


Figure 1: *Information in a sequence-only versus a sequence and structure profile. and > characters and connected by lines at top of figure. Boxes with internal numbers at top and bottom of the alignment indicate the number of bits per position from the sequence (black), or per basepair from the structure (white). This figure is from [3].*

The amount of additional information gained from structure varies widely for real RNA families, as shown for about 160 families in Figure 2 below. Note that for most families, modeling structure contributes at least 10 additional bits of information, which corresponds to lowering the expected chance of a false positive in a random database (i.e. the E-value of a database hit) by three orders of magnitude ($2^{10} = 1024$).

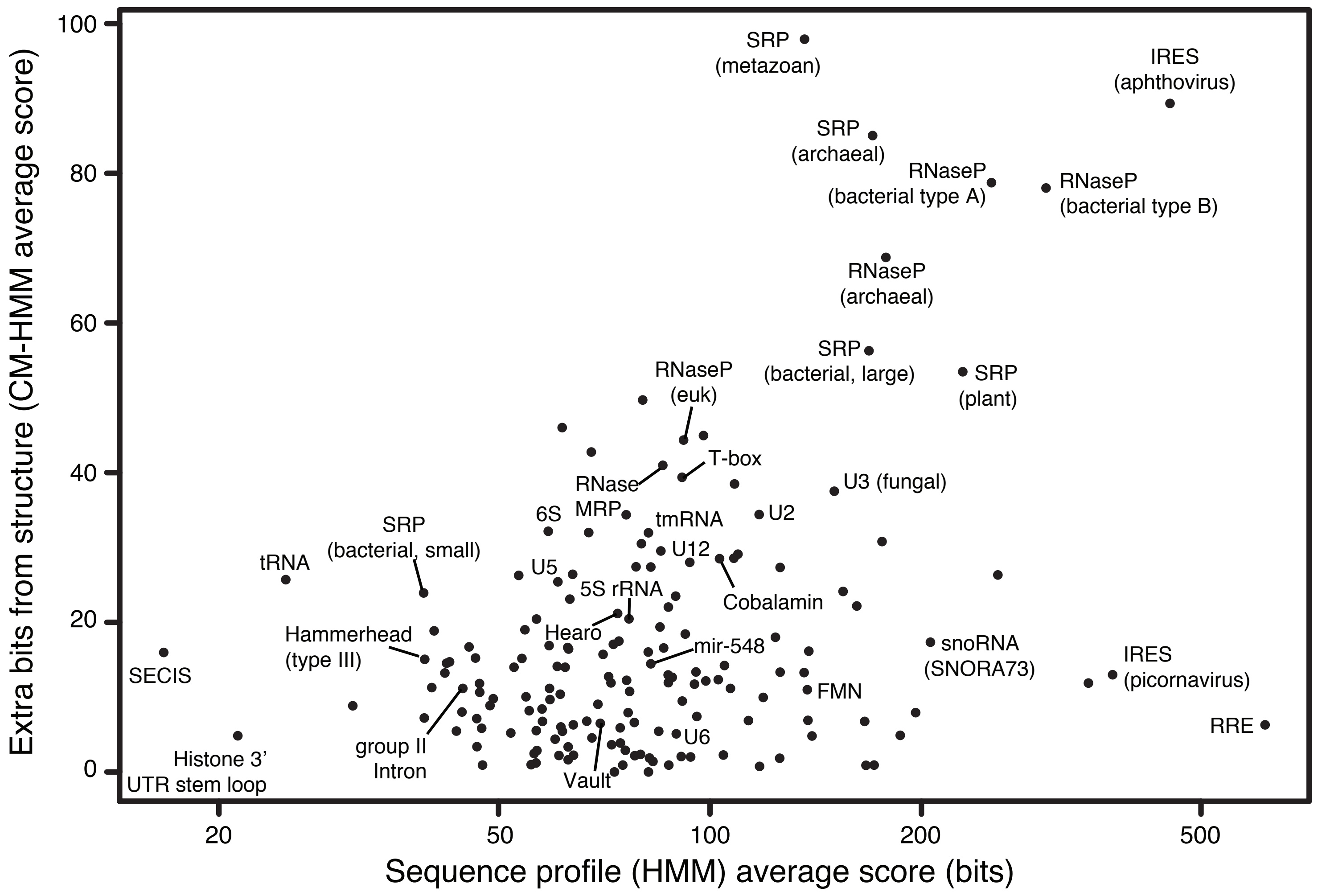


Figure 2: *Additional information (in bits) gained by sequence and structure profiles (CMs) versus sequence-only profiles (HMMs) for various RNA families. Data shown for the 164 Rfam release 11.0 families with 50 or more sequences in the* seed *alignment. For each family, the seed alignment was used to build two profile models, a CM and a profile HMM. From each model, 10,000 sequences were generated and scored, and the average score per sampled sequence was calculated. Infernal version 1.1 was used for all steps. This figure is from [4].*

Internal benchmark shows benefit of modeling sequence and structure conservation

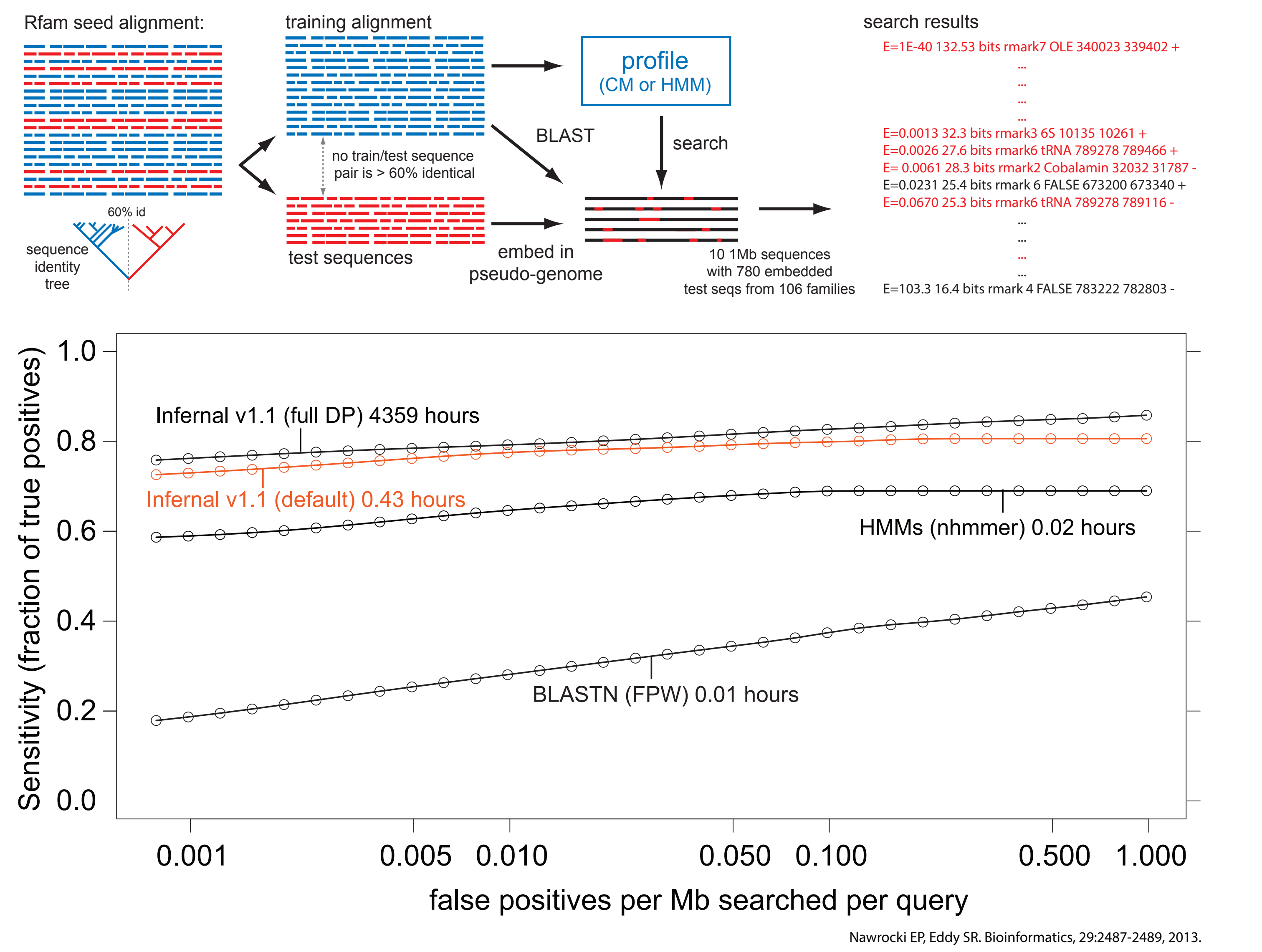
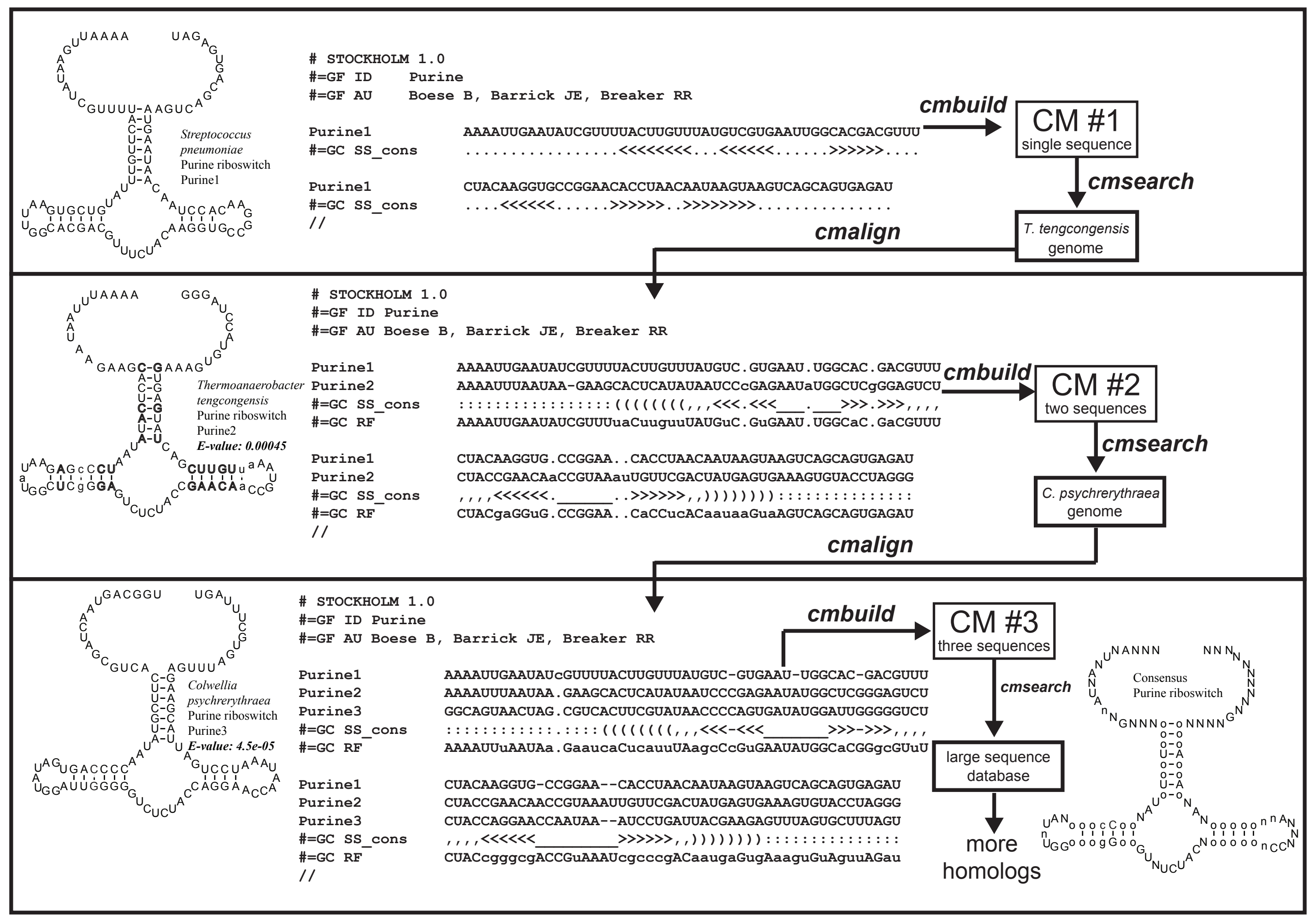


Figure 3: *RMARK benchmark. Top: schematic for benchmark construction. Bottom: Results of benchmark. Plots are shown for the new Infernal 1.1 with and without filters, for profile HMM searches with nhmmer [5] (from the HMMER package included in Infernal 1.1, default parameters) and for family-pairwise-searches with BLASTN (ncbi-blast-2.2.28+ default parameters). The Infernal times do not include time required for model calibration. This figure is from [6].*

Iterative search to expand knowledge of an RNA family



```
# STOCKHOLM 1.0
#=GF ID    Purine
#=GF AU    Boese B, Barrick JE, Breaker RR

Purine1         AAAAUUGAAUAUCGUUUUACUUGUUUAUGUCGUGAAUUGGCACGACGUUU
#=GC SS_cons    ................<<<<<<<<...<<<<<<.....>>>>>>....

Purine1         CUACAAGGUGCCGGAACACCUAACAAUAAGUAAGUCAGCAGUGAGAU
#=GC SS_cons    ....<<<<<<......>>>>>>..>>>>>>>>..............
//
```

```
# STOCKHOLM 1.0
#=GF ID Purine
#=GF AU Boese B, Barrick JE, Breaker RR

Purine1         AAAAUUGAAUAUCGUUUUACUUGUUUAUGUC.GUGAAU.UGGCAC.GACGUUU
Purine2         AAAAUUUAAUAA-GAAGCACUCAUAUAAUCCcGAGAAUaUGGCUCgGGAGUCU
#=GC SS_cons    ::::::::::::::::((((((((,,,<<<.<<<___.___>>>.>>>,,,,
#=GC RF         AAAAUUGAAUAUCGUUUuaCuuguuUAUGuC.GuGAAU.UGGCaC.GaCGUUU

Purine1         CUACAAGGUG.CCGGAA..CACCUAACAAUAAGUAAGUCAGCAGUGAGAU
Purine2         CUACCGAACAaCCGUAAauUGUUCGACUAUGAGUGAAAGUGUACCUAGGG
#=GC SS_cons    ,,,,<<<<<<.______..>>>>>>,,))))))))::::::::::::::
#=GC RF         CUACgaGGuG.CCGGAA..CaCCucACaauaaGuaAGUCAGCAGUGAGAU
//
```

```
# STOCKHOLM 1.0
#=GF ID Purine
#=GF AU Boese B, Barrick JE, Breaker RR

Purine1         AAAAUUGAAUAUcGUUUUACUUGUUUAUGUC-GUGAAU-UGGCAC-GACGUUU
Purine2         AAAAUUUAAUAA.GAAGCACUCAUAUAAUCCCGAGAAUAUGGCUCGGGAGUCU
Purine3         GGCAGUAACUAG.CGUCACUUCGUAUAACCCCAGUGAUAUGGAUUGGGGGUCU
#=GC SS_cons    ::::::::::::.:::((((((((,,,<<<-<<<_______>>>->>>,,,,
#=GC RF         AAAAUUuAAUAa.GaaucaCucauuUAagcCcGuGAAUAUGGCaCGGgcGUuU

Purine1         CUACAAGGUG-CCGGAA--CACCUAACAAUAAGUAAGUCAGCAGUGAGAU
Purine2         CUACCGAACAACCGUAAAUUGUUCGACUAUGAGUGAAAGUGUACCUAGGG
Purine3         CUACCAGGAACCAAUAA--AUCCUGAUUACGAAGAGUUUAGUGCUUUAGU
#=GC SS_cons    ,,,,<<<<<<_________>>>>>>,,))))))))::::::::::::::
#=GC RF         CUACcgggcgACCGuAAAUcgcccgACaaugaGugAaaguGuAguuAGau
//
```

Funding

This work is supported by the Intramural Research Program of the NIH, National Library of Medicine (EPN), by Howard Hughes Medical Institute (EPN (previously), SRE), by the European Bioinformatics Institute (IK, JA, AIP), and by the Biotechnology and Biological Sciences Research Council (IK, JA, AIP).

References

[1] E. K. Freyhult, J. P. Bollback, and P. P. Gardner. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.*, 17:117–125, 2007.

[2] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov. Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. 2017.

[3] E. P. Nawrocki. Annotating functional RNAs in genomes using Infernal. In J. Gorodkin and W. L. Ruzzo, editors, *RNA Sequence, Structure, and Function: Computational and Bioinformatic methods*, pages 163–197. Springer Humana Press, 2014.

[4] E. P. Nawrocki and S. R. Eddy. Computational identification of functional RNA homologs in metagenomic data. *RNA Biol.*, 10:1170–1179, 2013.

[5] T. J. Wheeler and S. R. Eddy. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29:2487–2489, 2013.

[6] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935, 2013.