

# **Structural RNA and viral sequence analysis**

Eric Nawrocki

Intramural Research Program  
National Library of Medicine  
National Institutes of Health

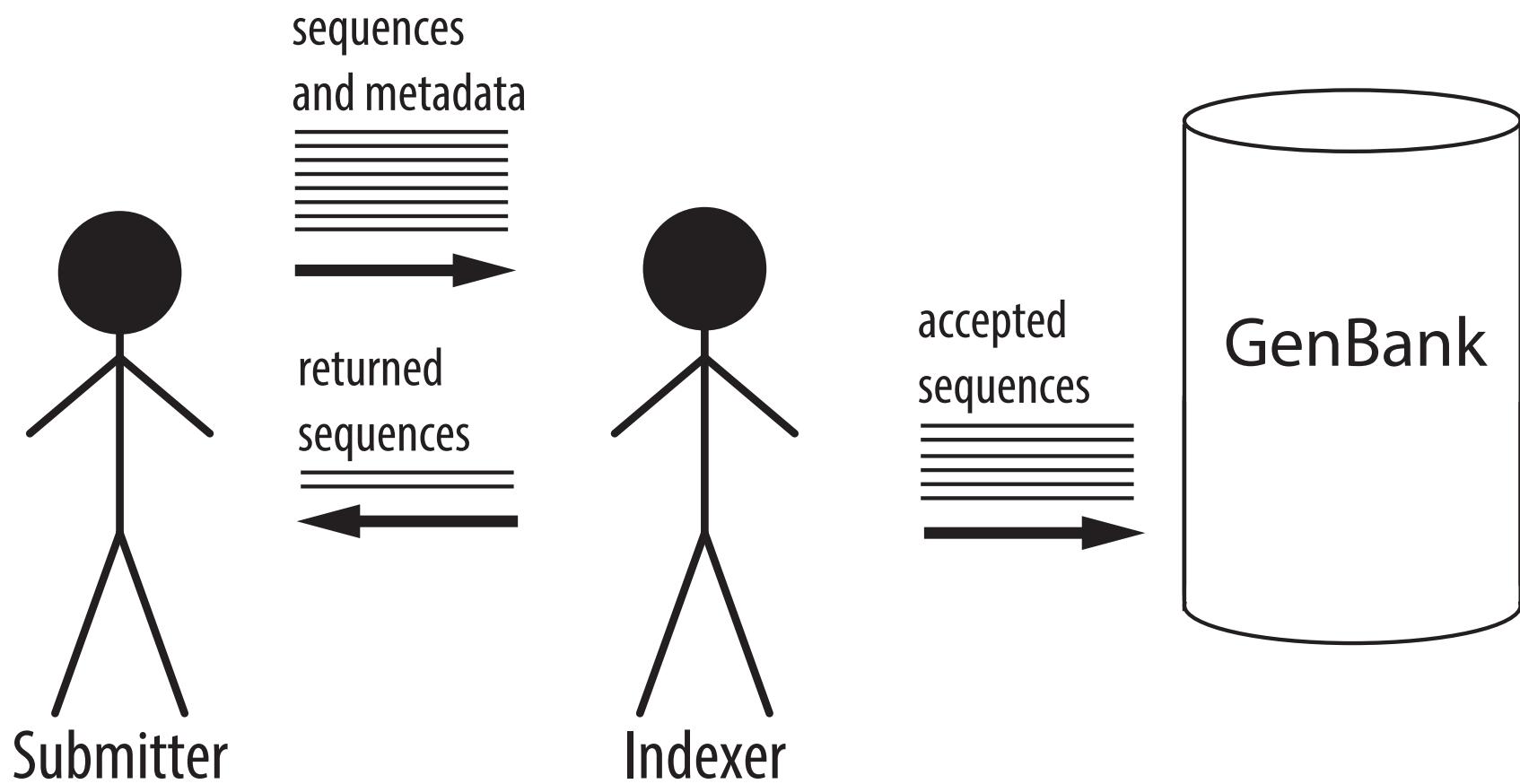


## **Two main areas of my research:**

- 1. Viral sequence analysis tools, since 2015**
  
- 2. Structural RNA analysis tools, since 2004**

add logos?

# GenBank indexers handle incoming sequence submissions



SOFTWARE

Open Access

# VADR: validation and annotation of virus sequence submissions to GenBank

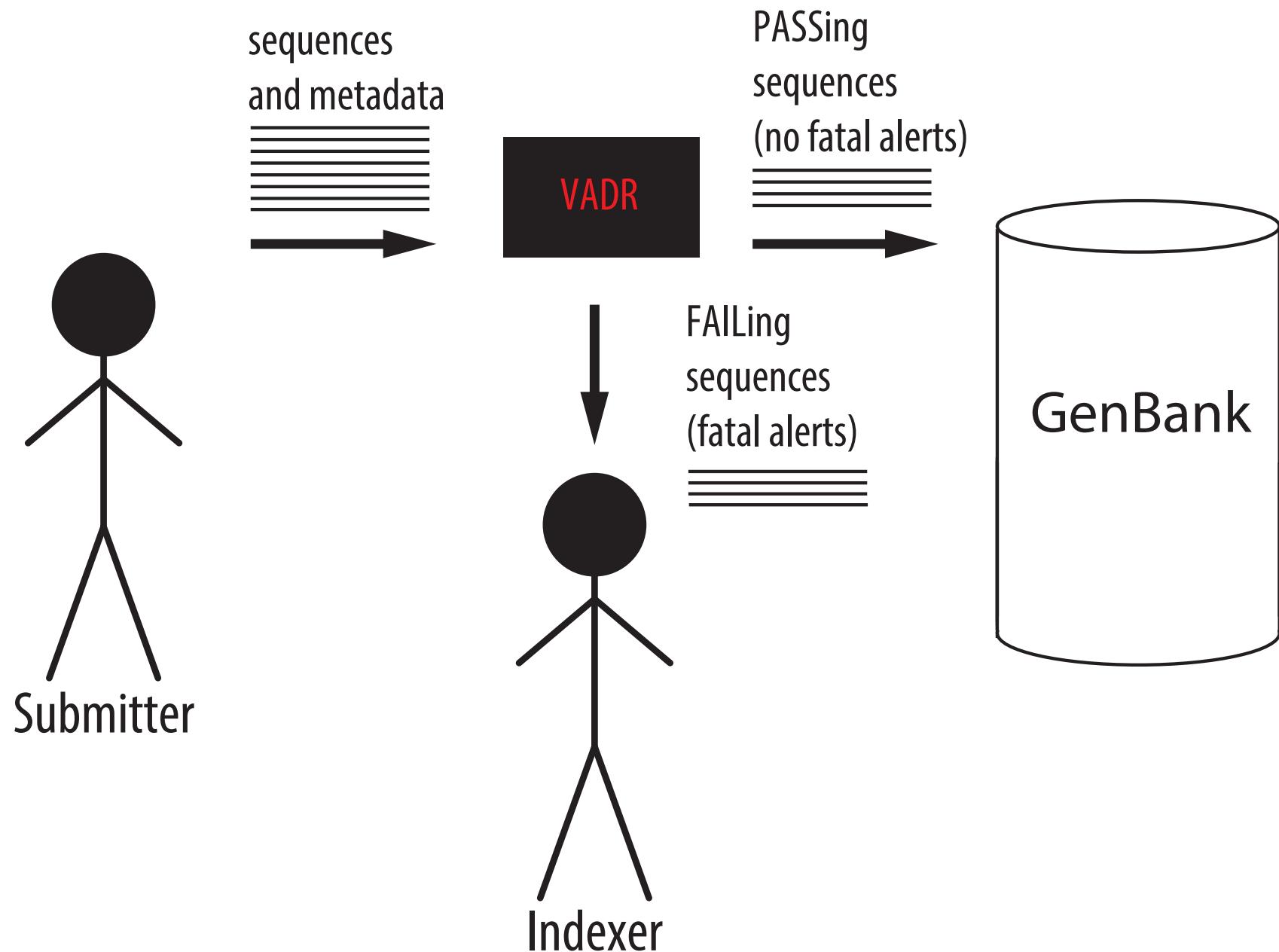


Alejandro A. Schäffer<sup>1,2</sup>, Eneida L. Hatcher<sup>2</sup>, Linda Yankie<sup>2</sup>, Lara Shonkwiler<sup>2,3</sup>, J. Rodney Brister<sup>2</sup>, Ilene Karsch-Mizrachi<sup>2</sup> and Eric P. Nawrocki<sup>2\*</sup> 

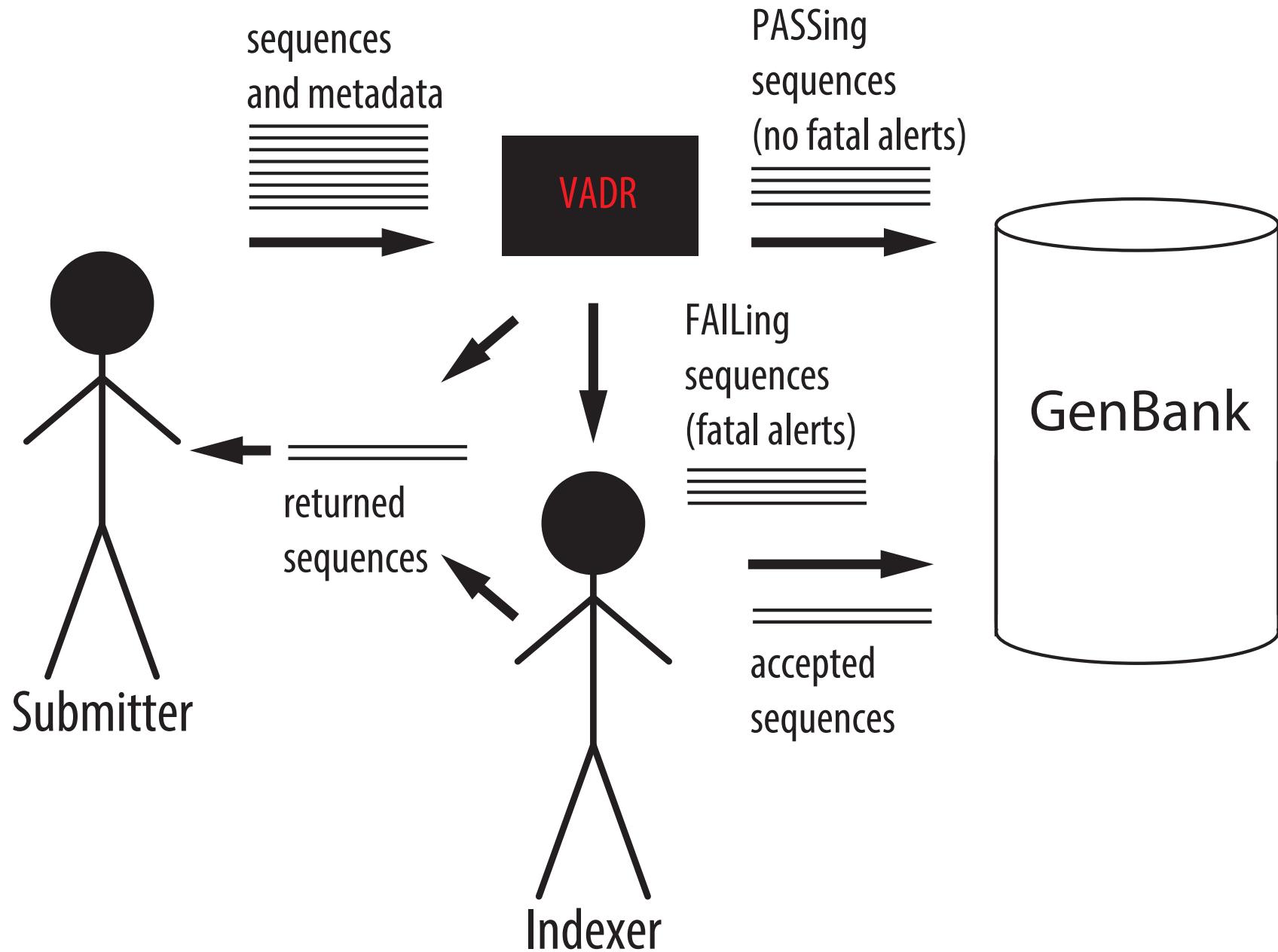
- general tool for reference-based annotation of viral sequences
- used for Norovirus and Dengue virus submissions since 2018
- used for SARS-CoV-2 submissions since March 2020
- currently also used manually for RSV, MpoX, and some Influenza submissions

# VADR assists GenBank indexers:

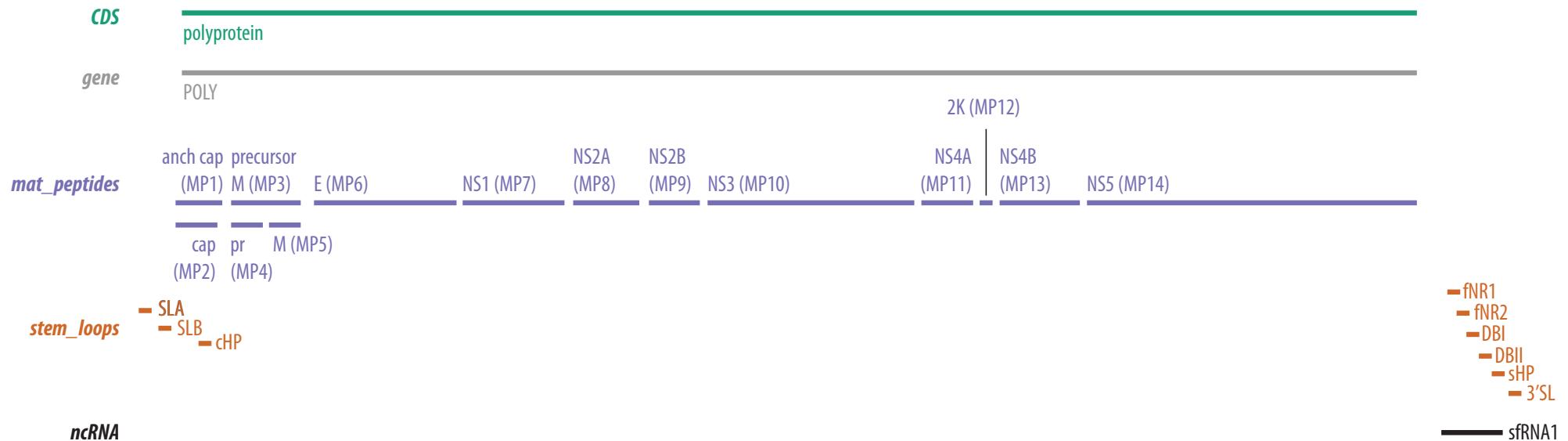
## Each sequence **PASSes** or **FAILs**



**Indexers decide fate of some FAILing sequences  
but some are sent directly back to submitter with error reports**



# VADR builds a reference model of a RefSeq and its features



## NC\_001477 MODEL



Group: Dengue; Subgroup: 1

# VADR validates and annotates each input sequence using its best-matching model

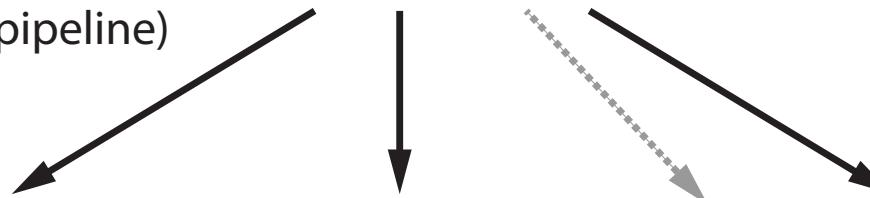
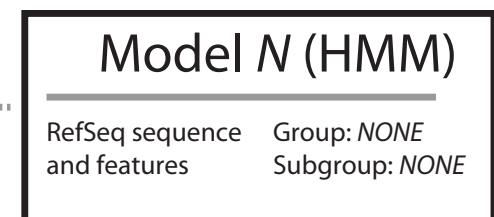
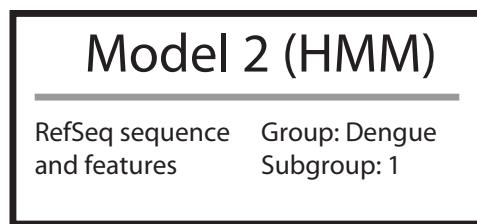
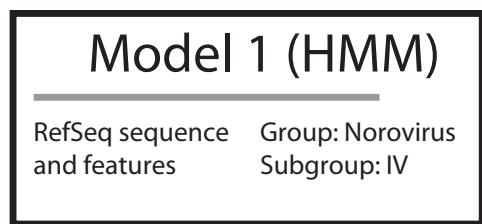
- Each sequence  $S$  proceeds through 4 stages:
  1. **Classification**
  2. **Coverage determination**
  3. **Alignment**
  4. **Protein validation**

*Different types of alerts are identified and reported at each stage*

## **Stage 1: Classification**

Score each sequence  
with all models  
(HMMER3 shortened pipeline)

input sequences:



low HMM score

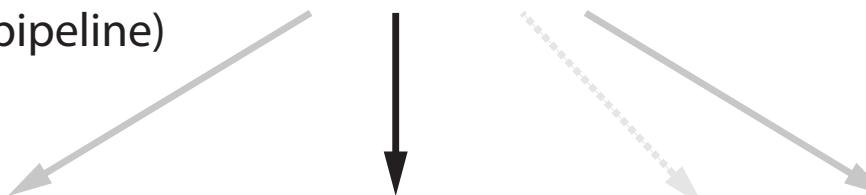
highest HMM score

low HMM score

## **Stage 1: Classification**

Score each sequence  
with all models  
(HMMER3 shortened pipeline)

input sequences:



### Model 1 (HMM)

RefSeq sequence  
and features      Group: Norovirus  
Subgroup: IV

### Model 2 (HMM)

RefSeq sequence  
and features      Group: Dengue  
Subgroup: 1

### Model N (HMM)

RefSeq sequence  
and features      Group: NONE  
Subgroup: NONE

low HMM score

highest HMM score

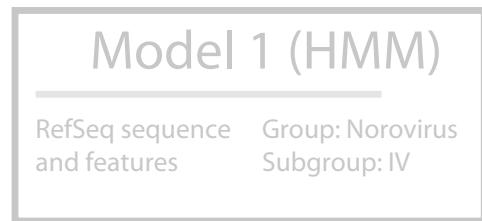
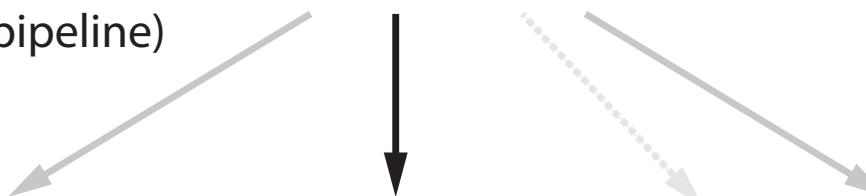
low HMM score

***best-matching model  
used in remaining stages***

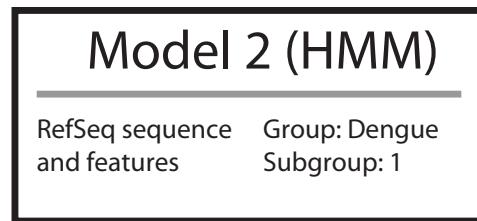
## Stage 1: Classification

Score each sequence  
with all models  
(HMMER3 shortened pipeline)

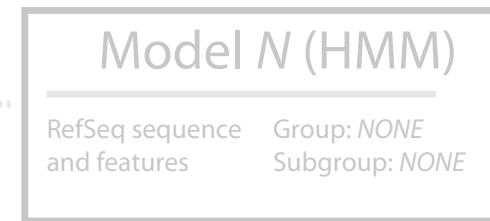
input sequences:



low HMM score



highest HMM score



low HMM score

***best-matching model  
used in remaining stages***

code	S/F	error message	description
<b>Fatal alerts detected in the classification stage</b>			
noannotn*	S	NO_ANNOTATION	no significant similarity detected
revcompl*	S	REVCOMPLEM	sequence appears to be reverse complemented
incsbgrp	S	INCORRECT_SPECIFIED_SUBGROUP	score difference too large between best overall model and best specified subgroup model
incgroup	S	INCORRECT_SPECIFIED_GROUP	score difference too large between best overall model and best specified group model
<b>Non-fatal alerts detected in the classification stage</b>			
qstsbgp	S	QUESTIONABLE_SPECIFIED_SUBGROUP	best overall model is not from specified subgroup
qstgroup	S	QUESTIONABLE_SPECIFIED_GROUP	best overall model is not from specified group
indfclas	S	INDEFINITE_CLASSIFICATION	low score difference between best overall model and second best model (not in best model's subgroup)
lowscore	S	LOW_SCORE	score to homology model below low threshold

## **Stage 2: Coverage determination**

Search each sequence with best-matching model (HMMER3 full pipeline)

input sequences that match best to NC\_001477:

S1 \_\_\_\_\_  
S2 \_\_\_\_\_  
S3 \_\_\_\_\_  
S4 \_\_\_\_\_



**NC\_001477 MODEL**



Group: Dengue; Subgroup: 1



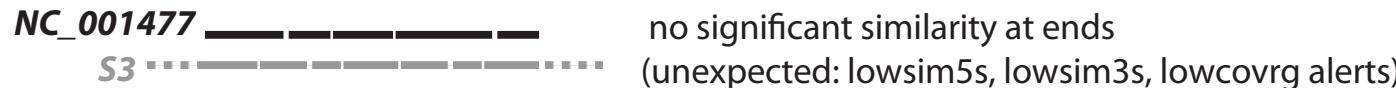
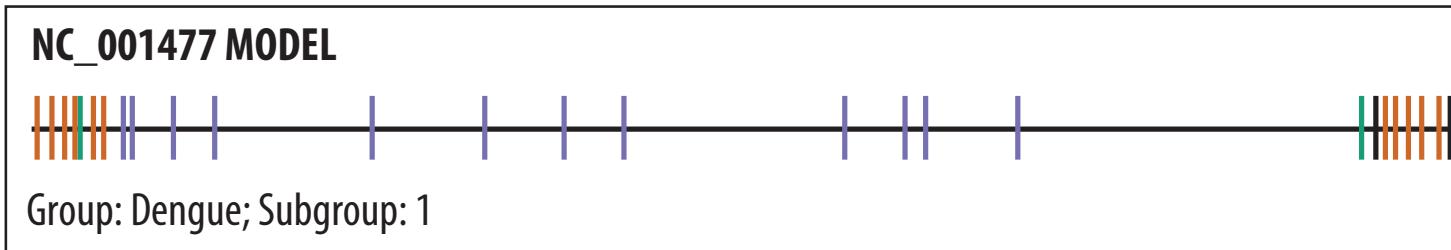
**NC\_001477** . . . . . full length sequence  
S1 . . . . . (expected)  
**NC\_001477** . . . . . partial or truncated sequence  
S2 . . . . . (expected)

## Stage 2: Coverage determination

Search each sequence with best-matching model (HMMER3 full pipeline)

input sequences that match best to NC\_001477:

S1 \_\_\_\_\_  
S2 \_\_\_\_\_  
S3 \_\_\_\_\_  
S4 \_\_\_\_\_

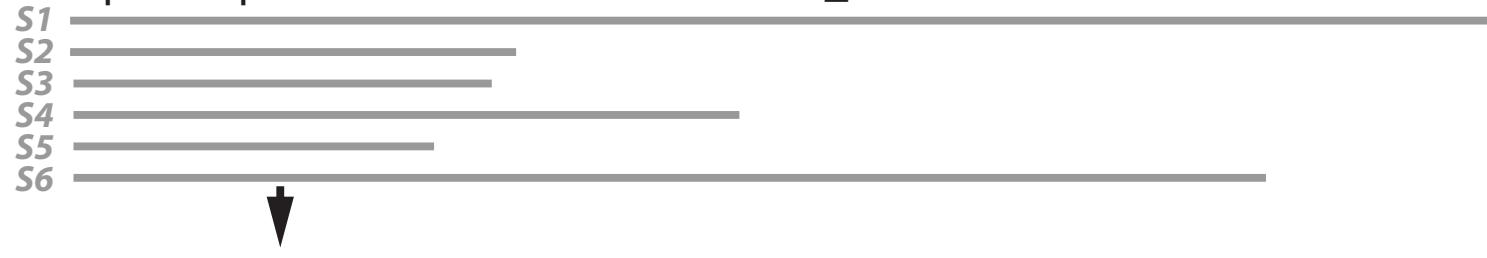


code	S/F	error message	description
<b>Fatal alerts detected in the coverage stage</b>			
lowcovrg	S	LOW_COVERAGE	low sequence fraction with significant similarity to homology model
dupregin	S	DUPLICATE_REGIONS	similarity to a model region occurs more than once
discontn	S	DISCONTINUOUS_SIMILARITY	not all hits are in the same order in the sequence and the homology model
indfstrn	S	INDEFINITE_STRAND	significant similarity detected on both strands
lowsim5s	S	LOW_SIMILARITY_START	significant similarity not detected at 5' end of the sequence
lowsim3s	S	LOW_SIMILARITY_END	significant similarity not detected at 3' end of the sequence
lowsimis	S	LOW_SIMILARITY	internal region without significant similarity
<b>Non-fatal alerts detected in the coverage stage</b>			
biasdseq	S	BIASED_SEQUENCE	high fraction of score attributed to biased sequence composition

### **Stage 3: Alignment and feature mapping**

Align each sequence to its best-matching model (Infernal's cmalign)

input sequences that match best to NC\_001477:



**NC\_001477 MODEL**

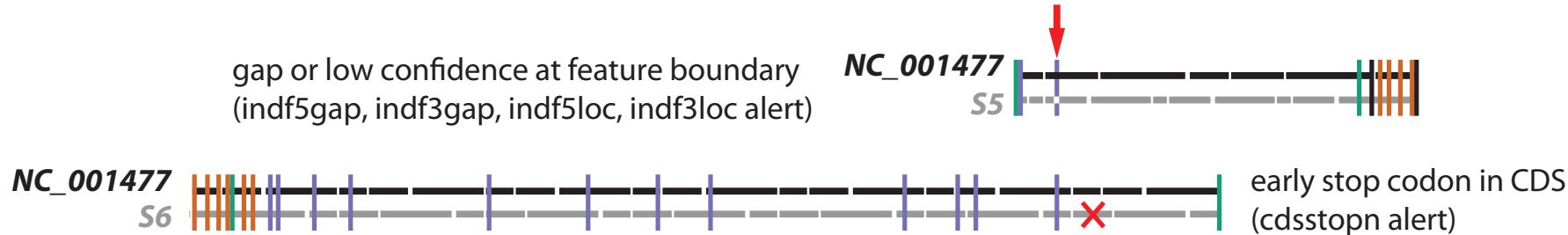


Group: Dengue; Subgroup: 1



## Stage 3: Alignment and feature mapping

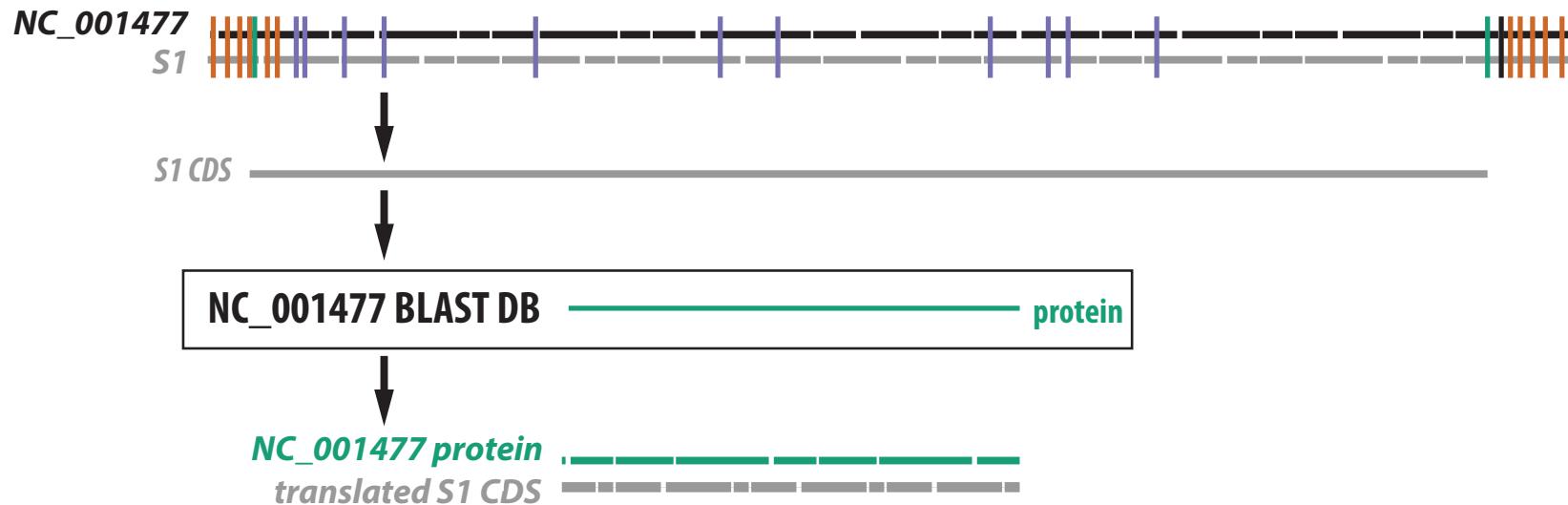
Align each sequence to its best-matching model (Infernal's cmalign)



code	S/F	error message	description
<b>Fatal alerts detected in the annotation stage</b>			
unexdivg*	S	UNEXPECTED_DIVERGENCE	sequence is too divergent to confidently assign nucleotide-based annotation
noftrann*	S	NO_FEATURES_ANNOTATED	sequence similarity to homology model does not overlap with any features
mutstart	F	MUTATION_AT_START	expected start codon could not be identified
mutendcd	F	MUTATION_AT_END	expected stop codon could not be identified, predicted CDS stop by homology is invalid
mutendns	F	MUTATION_AT_END	expected stop codon could not be identified, no in-frame stop codon exists 3' of predicted valid start codon
mutendex	F	MUTATION_AT_END	expected stop codon could not be identified, first in-frame stop codon exists 3' of predicted stop position
unexleng	F	UNEXPECTED_LENGTH	length of complete coding (CDS or mat_peptide) feature is not a multiple of 3
cdsstoppn	F	CDS_HAS_STOP_CODON	in-frame stop codon exists 5' of stop position predicted by homology to reference
peptrans	F	PEPTIDE_TRANSLATION_PROBLEM	mat_peptide may not be translated because its parent CDS has a problem
pepadjcy	F	PEPTIDE_ADJACENCY_PROBLEM	predictions of two mat_peptides expected to be adjacent are not adjacent
indfantn	F	INDEFINITE_ANNOTATION	nucleotide-based search identifies CDS not identified in protein-based search
indf5gap	F	INDEFINITE_ANNOTATION_START	alignment to homology model is a gap at 5' boundary
indf5loc	F	INDEFINITE_ANNOTATION_START	alignment to homology model has low confidence at 5' boundary
indf3gap	F	INDEFINITE_ANNOTATION_END	alignment to homology model is a gap at 3' boundary
indf3loc	F	INDEFINITE_ANNOTATION_END	alignment to homology model has low confidence at 3' boundary
lowsim5f	F	LOW FEATURE SIMILARITY_START	region within annotated feature at 5' end of sequence lacks significant similarity
lowsim3f	F	LOW FEATURE SIMILARITY_END	region within annotated feature at 3' end of sequence lacks significant similarity
lowsimif	F	LOW FEATURE SIMILARITY	region within annotated feature lacks significant similarity

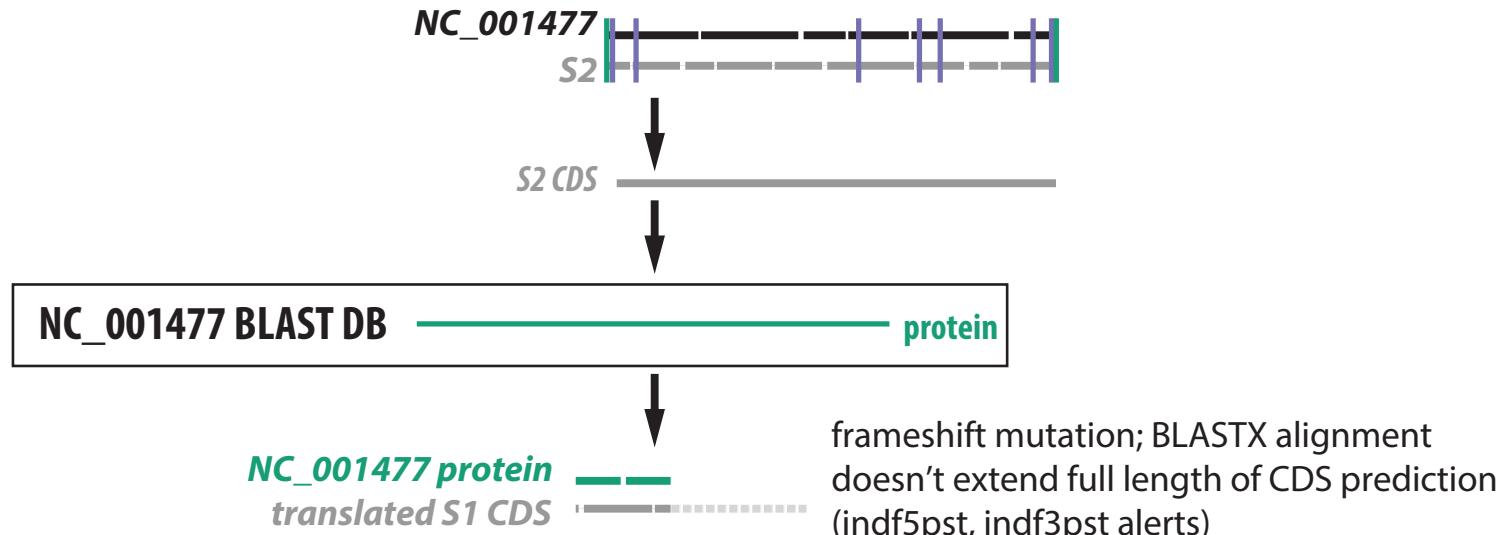
## **Stage 4: Protein validation (Alejandro Schäffer)**

Compare each predicted CDS to model (RefSeq) proteins with BLASTX



## **Stage 4: Protein validation (Alejandro Schäffer)**

Compare each predicted CDS to model (RefSeq) proteins with BLASTX



code	S/F	error message	description
<b>Fatal alerts detected in the protein validation stage</b>			
cdsstopp	F	CDS_HAS_STOP_CODON	stop codon in protein-based alignment
indf5antp	F	INDEFINITE_ANNOTATION	protein-based search identifies CDS not identified in nucleotide-based search
indf5plg	F	INDEFINITE_ANNOTATION_START	protein-based alignment extends past nucleotide-based alignment at 5' end
indf5pst	F	INDEFINITE_ANNOTATION_START	protein-based alignment does not extend close enough to nucleotide-based alignment 5' endpoint
indf3plg	F	INDEFINITE_ANNOTATION_END	protein-based alignment extends past nucleotide-based alignment at 3' end
indf3pst	F	INDEFINITE_ANNOTATION_END	protein-based alignment does not extend close enough to nucleotide-based alignment 3' endpoint
indfstrp	F	INDEFINITE_STRAND	strand mismatch between protein-based and nucleotide-based predictions
insertnp	F	INSERTION_OF_NT	too large of an insertion in protein-based alignment
deletinp	F	DELETION_OF_NT	too large of a deletion in protein-based alignment

## VADR used for Norovirus and Dengue virus sequences since 2018

	Norovirus	Dengue virus
length	7.6Kb	10.7Kb
# seqs	44,936	113,211
% seqs full length	5.1%	8.4%
% Ns	0.5%	0.2%
% seqs with stretch of $\geq$ 50 Ns	1.0%	0.4%
average % identity	81.6%	94.4%

### VADR v1.0 performance

seconds per sequence	42.4	92.6
required RAM	8Gb	8Gb
total running time, CPU days	1.1	10.2

# SARS-CoV-2 sequence submissions have increased since early 2020

month	year	#new seqs	#cumulative seqs
Jan	2020	32	32
Feb	2020	58	90
Mar	2020	332	422
Apr	2020	1541	1963
May	2020	2974	4937
Jun	2020	3394	8331
Jul	2020	3604	11,935
Aug	2020	3818	15,753
Sep	2020	6731	22,484
Oct	2020	11,939	34,423
Nov	2020	4274	38,697
Dec	2020	4530	43,227
Jan	2021	8775	52,002
Feb	2021	26,078	78,080
Mar	2021	42,607	120,687
Apr	2021	97,095	217,782
May	2021	104,729	322,511
Jun	2021	46,187	368,698
Jul	2021	43,336	412,034
Aug	2021	141,958	553,992
Sep	2021	267,562	821,554
Oct	2021	239,296	1,060,850
Nov	2021	267,270	1,328,120
Dec	2021	288,771	1,616,891
Jan	2022	258,522	1,875,413
Feb	2022	230,185	2,105,598

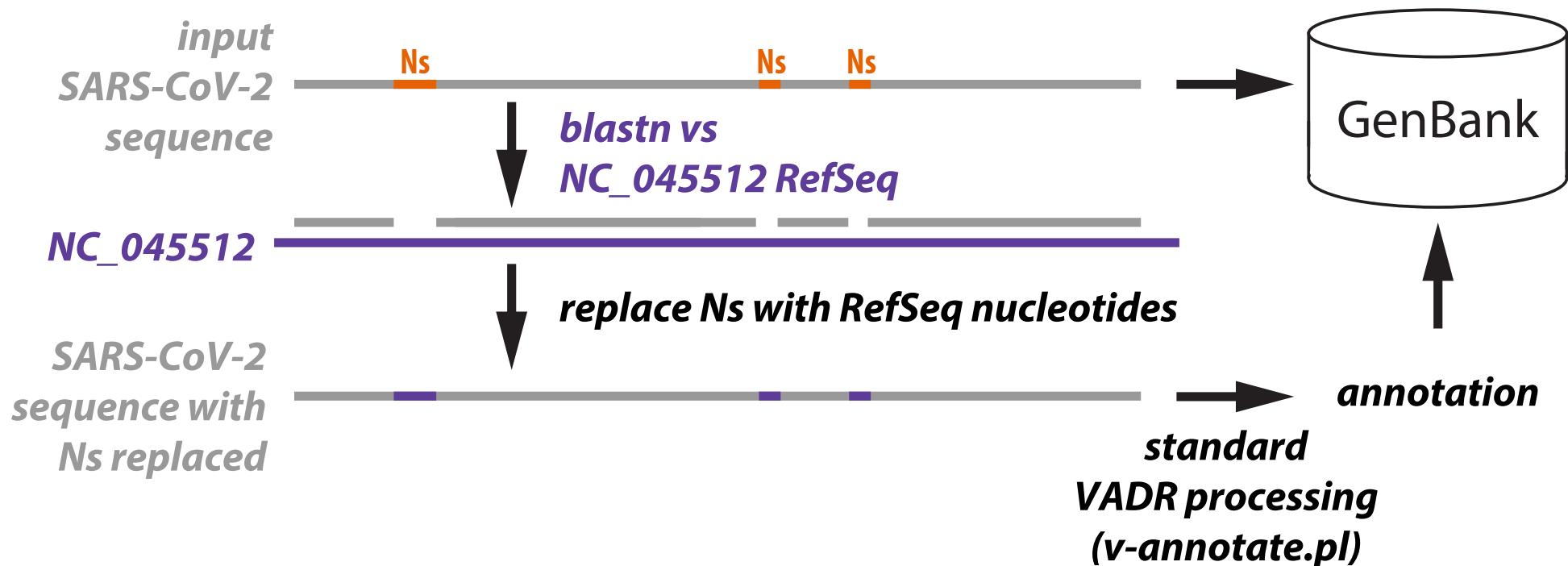
# SARS-CoV-2 sequences differ from Norovirus and Dengue virus in several ways that impact VADR processing

	Norovirus	Dengue virus	SARS-CoV-2
length	7.6Kb	10.7Kb	29.9Kb
# seqs	44,936	113,211	1,616,891
% seqs full length	5.1%	8.4%	99.7%
% Ns	0.5%	0.2%	1.4%
% seqs with stretch of $\geq$ 50 Ns	1.0%	0.4%	38.7%
average % identity	81.6%	94.4%	99.4%

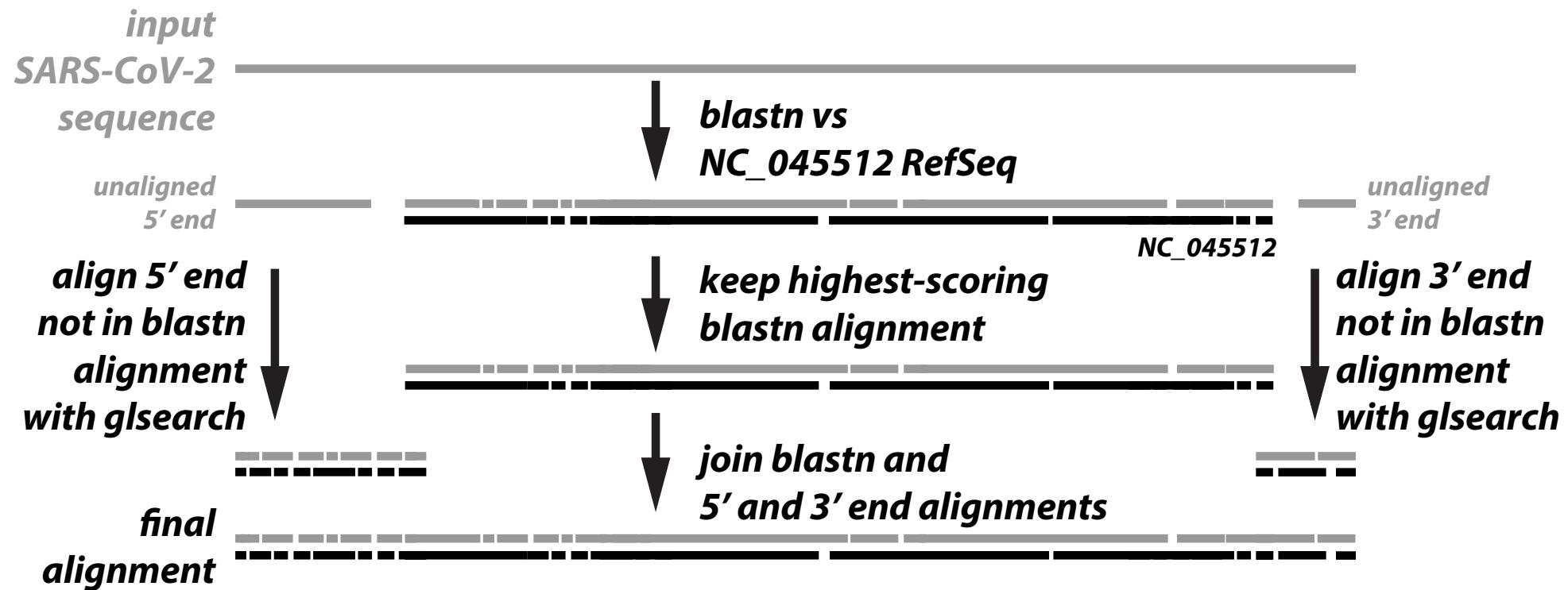
## VADR v1.0 performance

seconds per sequence	42.4	92.6	331.8
required RAM	8Gb	8Gb	64Gb
total running time, CPU days	1.1	10.2	6187.6

## Replacing Ns with expected nucleotides allows many 'good' sequences to pass

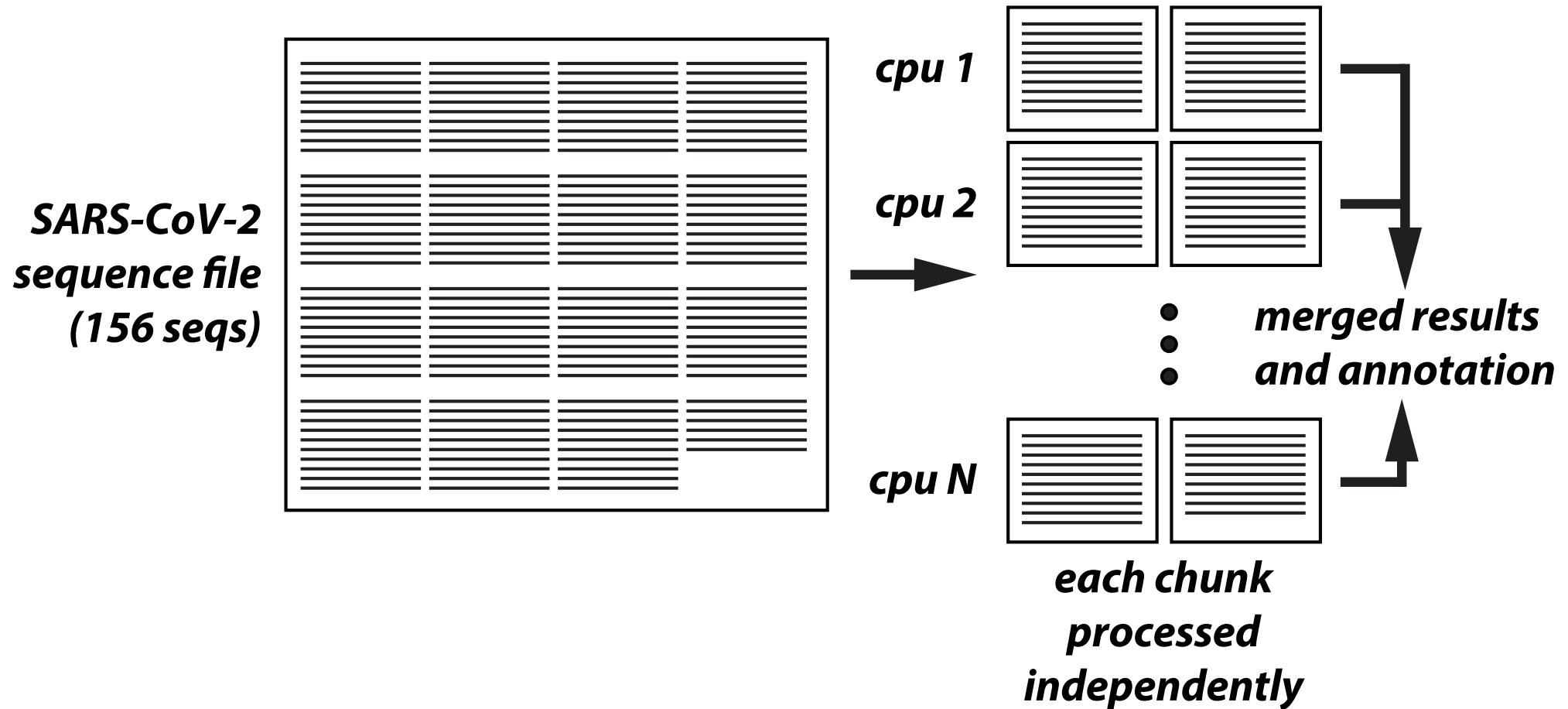


# Seeded alignment using blastn makes alignment stage faster



# Using glsearch instead of cmalign reduces memory requirement

- lower memory requirement (2Gb max) allows for multi-threading



# VADR is now 1000-fold faster in practice for SARS-CoV-2 processing

VADR version	seeded alignment?	N replacement?	glsearch?	# cpus	required RAM	secs per seq	hours per 100K seqs	speedup vs v1.0
v1.0	-	-	-	1	64 Gb	329.91	9164.3	-
<b>v1.4.1</b>	<b>+</b>	<b>+</b>	<b>+</b>	<b>8</b>	<b>16 Gb</b>	<b>0.33</b>	<b>9.3</b>	<b>986.8</b>

# VADR is now fast enough to handle hundreds of thousands of sequences per month

month	year	#new seqs	#cumulative seqs
Jan	2020	32	32
Feb	2020	58	90
Mar	2020	332	422
Apr	2020	1541	1963
May	2020	2974	4937
Jun	2020	3394	8331
Jul	2020	3604	11,935
Aug	2020	3818	15,753
Sep	2020	6731	22,484
Oct	2020	11,939	34,423
Nov	2020	4274	38,697
Dec	2020	4530	43,227
Jan	2021	8775	52,002
Feb	2021	26,078	78,080
Mar	2021	42,607	120,687
Apr	2021	97,095	217,782
May	2021	104,729	322,511
Jun	2021	46,187	368,698
Jul	2021	43,336	412,034
Aug	2021	141,958	553,992
Sep	2021	267,562	821,554
Oct	2021	239,296	1,060,850
Nov	2021	267,270	1,328,120
Dec	2021	288,771	1,616,891
Jan	2022	258,522	1,875,413
Feb	2022	230,185	2,105,598

**Besides getting faster, VADR has changed in other ways  
(work with Linda Yankie and Vince Calhoun and GenBank team)**

- 13 releases between March 2020 and January 2022
- 3 additional models (all eventually dropped):
  - B.1.1.7 (alpha)
  - B.1.525
  - 28254-deletion
- allow some alerts for non-essential ORFs without failing sequence  
(they become a `misc_feature` instead)

# Faster SARS-CoV-2 sequence validation and annotation for GenBank using VADR

Eric P. Nawrocki \*

National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

Received September 08, 2022; Revised November 28, 2022; Editorial Decision December 14, 2022; Accepted January 03, 2023

## Additional VADR models and development

	length	num models	new feature(s)	author
RSV	15Kb	2	alignment-based models	EPN
COX-1	1.5Kb	86	protein-coding gene	EPN
Mpox	197Kb	1	minimap alignment	EPN
Influenza	1-2Kb	70	segmented virus	EPN
Zika	11Kb	?	?	EBD

## Additional VADR models and development

	length	num models	new feature(s)	author
RSV	15Kb	2	alignment-based models	EPN
COX-1	1.5Kb	86	protein-coding gene	EPN
Mpox	197Kb	1	minimap alignment	EPN
Influenza	1-2Kb	70	segmented virus	EPN
Zika	11Kb	?	?	EBD

Database, 2024, baae091

DOI: <https://doi.org/10.1093/database/baae091>

Original article



## Influenza sequence validation and annotation using VADR

Vincent C. Calhoun, Eneida L. Hatcher, Linda Yankie, Eric P. Nawrocki \*

National Center for Biotechnology Information, U.S. National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, MD 20894, United States

## **Additional VADR models and development**

- Alex Greninger's lab at Univ of Washington:
  - sequences a wide variety of human pathogenic viruses
  - previously developed the VAPiD software tool for validating and annotating viral sequences
  - now a collaborator that builds VADR models

## Additional VADR models and development

	length	num models	new feature(s)	author
RSV	15Kb	2	alignment-based models	
Mpox	197Kb	1	minimap alignment	
COX-1	1.5Kb	86	protein-coding gene	
Influenza	1-2Kb	70	segmented virus	
Zika	11Kb	?	?	
Herpes Simplex Virus (HSV)	150Kb	2	-	Grenier et al.
Human metapneumovirus (HMPV)	13Kb	6	-	Grenier et al.

## **VADR and NCBI virus**

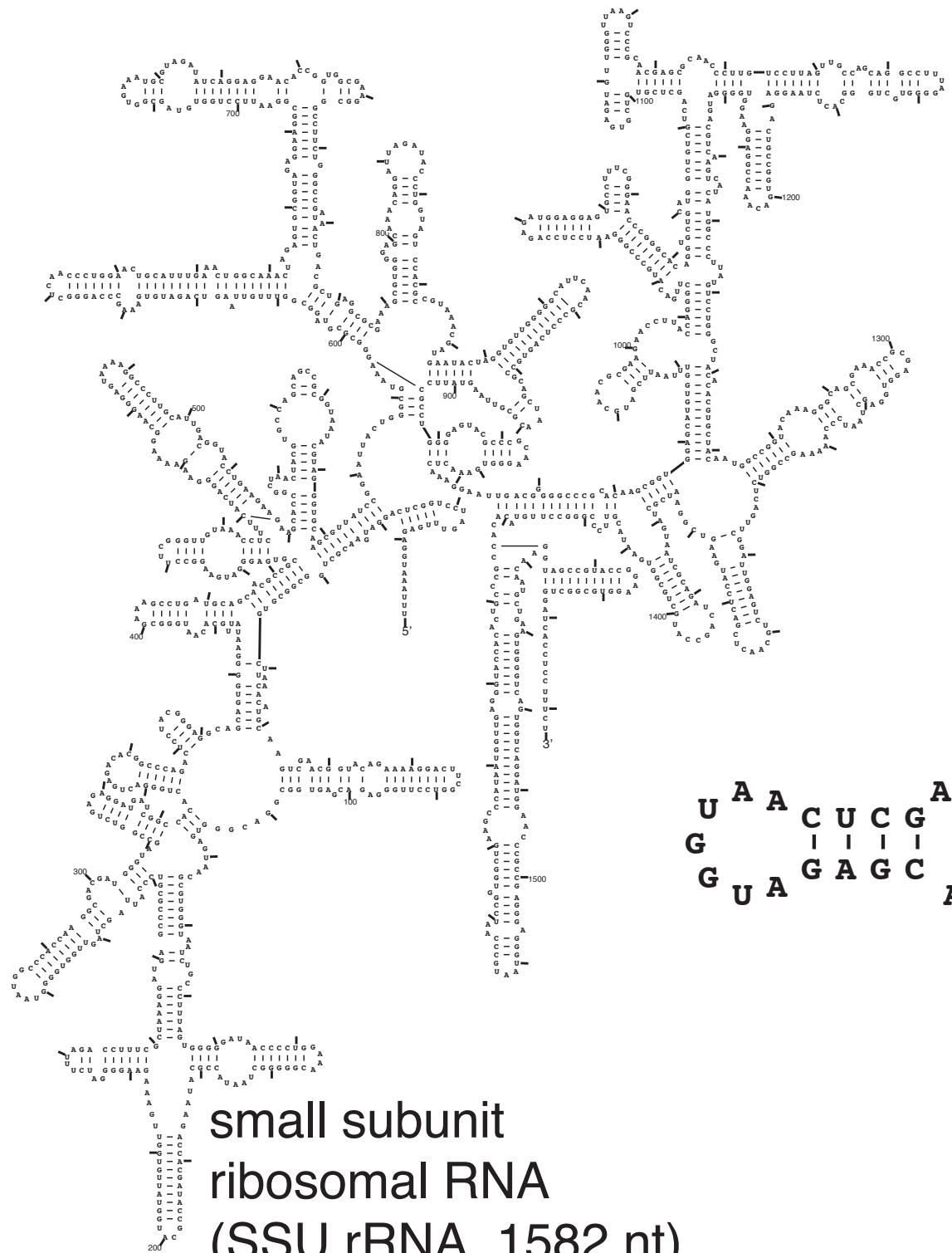
- FY2025 NCBI virus goals related to VADR:
  - VADR web server
  - Replacement of FLAN with VADR

## **Future directions**

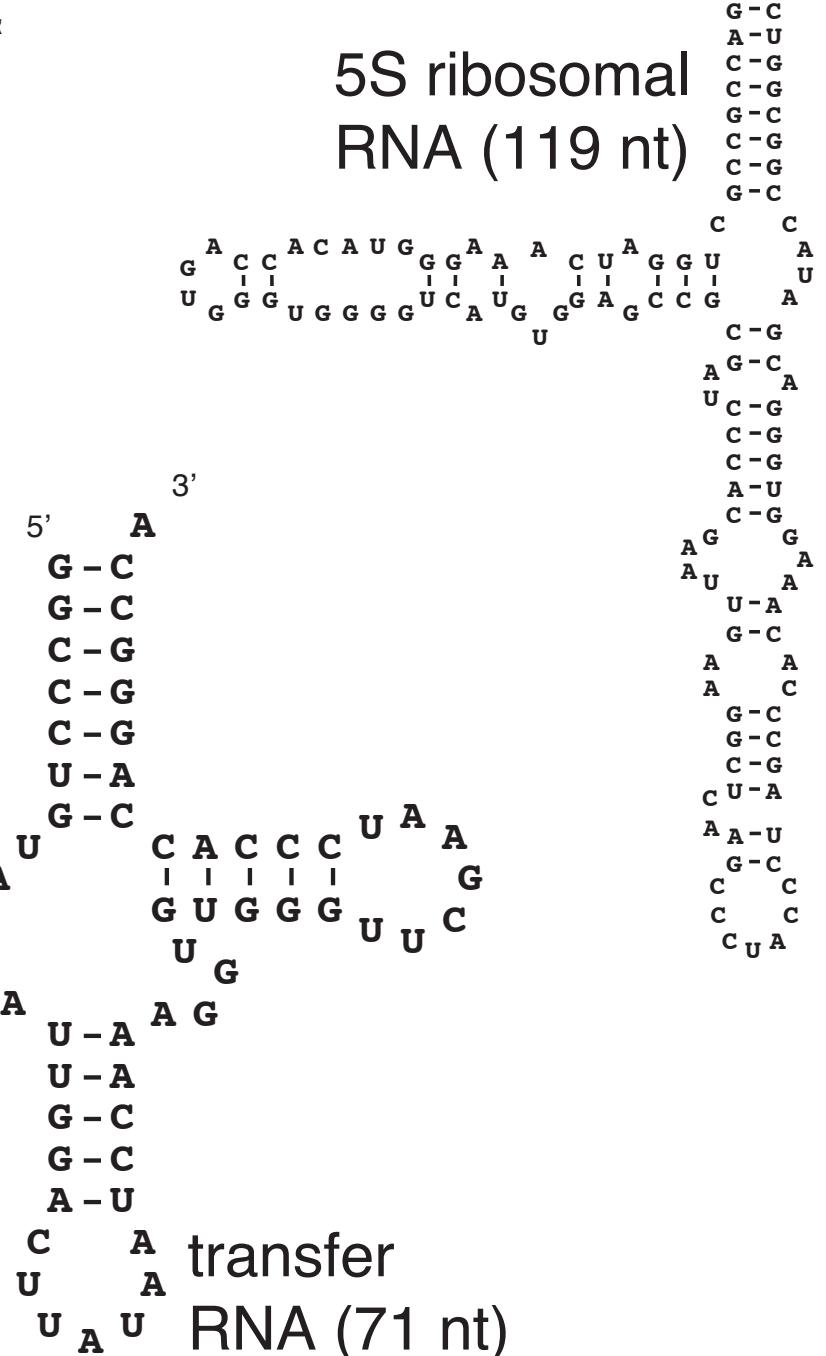
- More models (our group and Greninger lab)
- Alignment-based models
- RNA structure annotation in Flaviviruses (already exists for Dengue)

# Structural RNA sequence analysis

- intro to Infernal
- intro to Rfam
- intro to RNACentral?
- intro to R2DT?
- RNA annotation: PGAP, Euk genomes (?), Hydractinia genomes
- tmRNA DB



small subunit  
ribosomal RNA  
(SSU rRNA, 1582 nt)



# Functional RNAs play many vital roles in the cell

	key RNAs involved	archaea	bacteria	eukarya	viruses
translation	ribosomal RNAs	x	x	x	
	transfer RNAs	x	x	x	
	RNase P RNA	x	x	x	
	snoRNAs	x		x	
	SRP RNA	x	x	x	
	tmRNA		x		
	RNaseMRP			x	
gene expression	riboswitches	?	x	?	
	microRNAs			x	x
	6S RNA		x	x	
splicing	U1, U2, U4, U5, U6			x	
other	tracrRNA	x	x		
	telomerase RNA			x	
	group I introns	x	x	x	x
	sfRNAs				x
	many more...				x

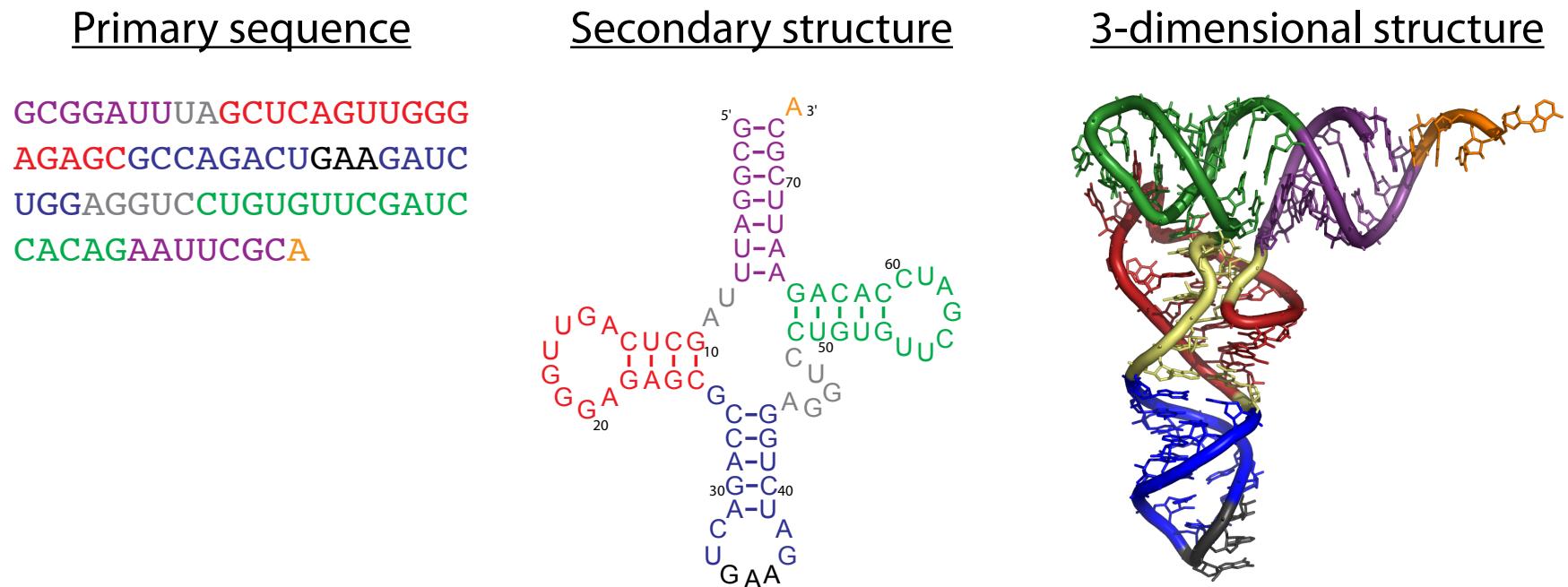
## Functional RNAs play many vital roles in the cell

	key RNAs involved	archaea	bacteria	eukarya	viruses
translation	ribosomal RNAs	x	x	x	
	transfer RNAs	x	x	x	
	RNase P RNA	x	x	x	
	snoRNAs	x		x	
	SRP RNA	x	x	x	
	tmRNA		x		
	RNaseMRP			x	
gene expression	riboswitches	?	x	?	
	microRNAs			x	x
	6S RNA		x	x	
splicing	U1, U2, U4, U5, U6			x	
other	tracrRNA	x	x		
	telomerase RNA			x	
	group I introns	x	x	x	x
	sfRNAs				x
	many more...				x



database of more than 4100 non-coding RNA families  
each represented by a secondary structure, alignment, and covariance model.

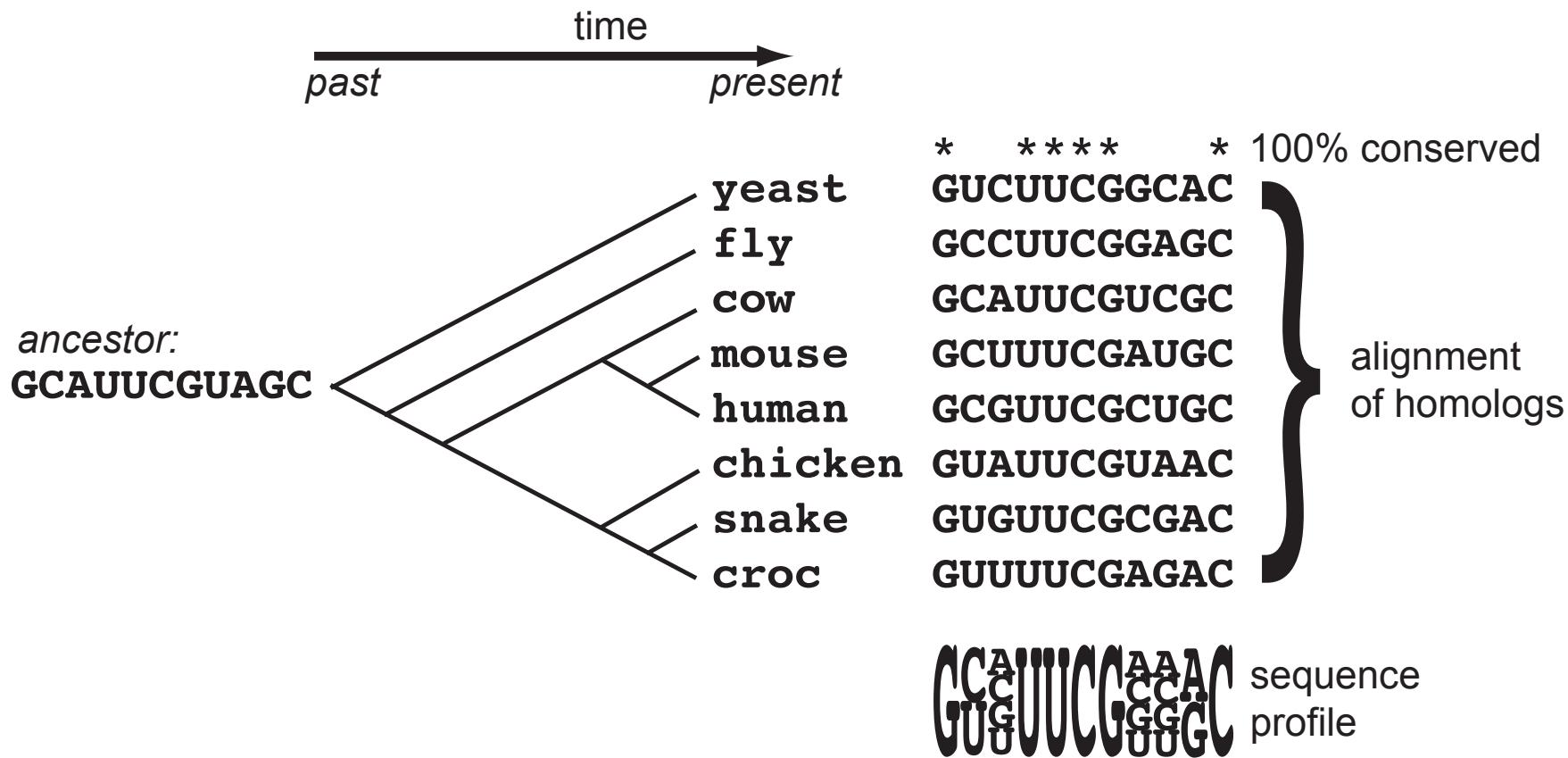
# Many functional RNAs adopt a conserved 3-dimensional structure



- BLAST: given a single sequence, search genomes for similar sequences.
- Structural RNAs are difficult to find
  - short (~ 100 nt) and evolve rapidly at sequence level
  - lack open reading frames
  - small, 4 letter alphabet
- BLAST cannot take advantage of:
  - sequence conservation, which varies across the gene
  - secondary structure

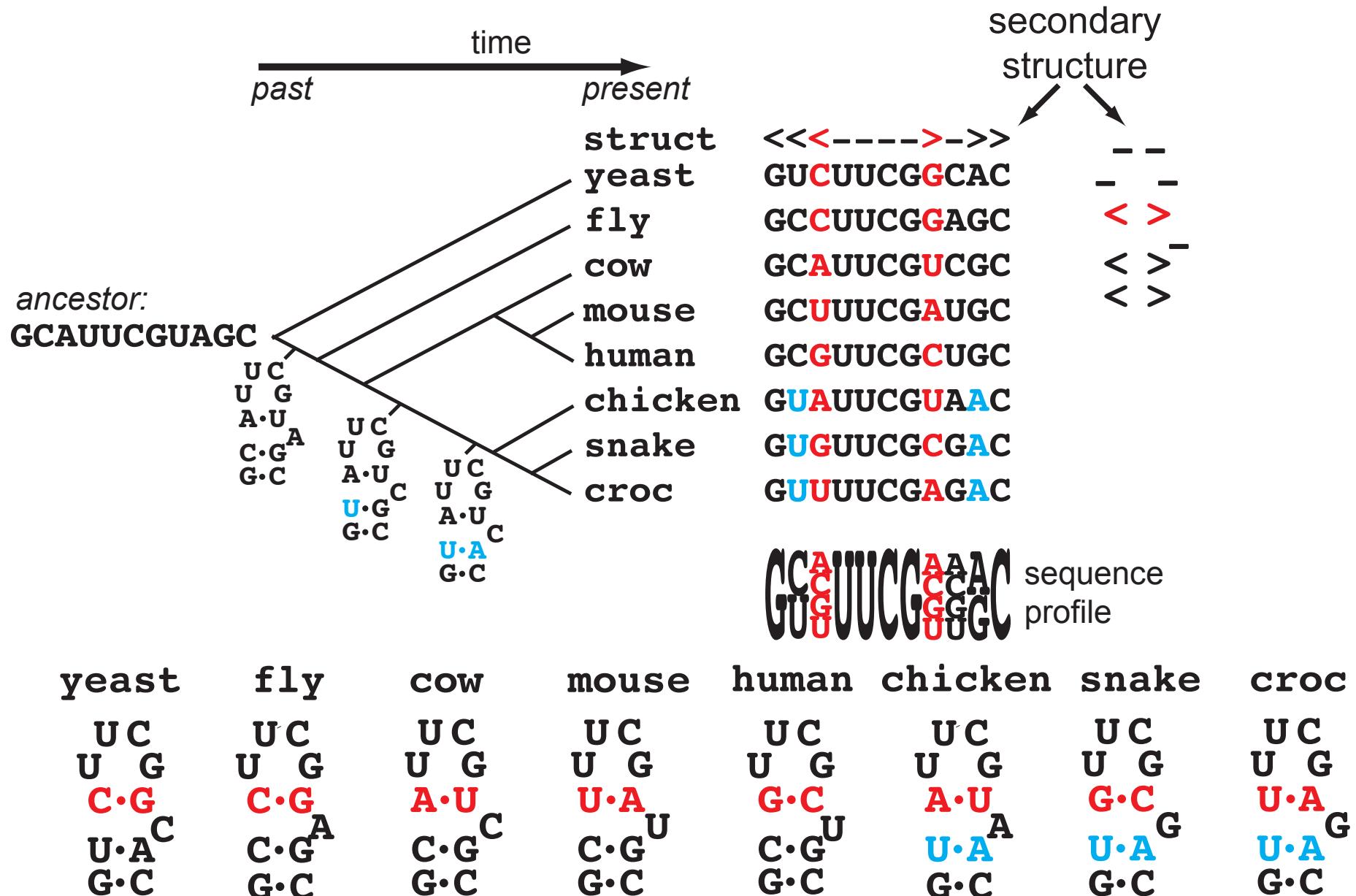
# Sequence conservation provides information for homology searches

Conservation levels vary across alignment columns.



# Structure conservation provides additional information

Base-paired positions covary  
to maintain Watson-Crick complementarity.



# profile HMMs and covariance models

	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 (23794 and 4150 entries)	Rfam (4178 families)
performance for RNAs	faster but less accurate	slower but more accurate

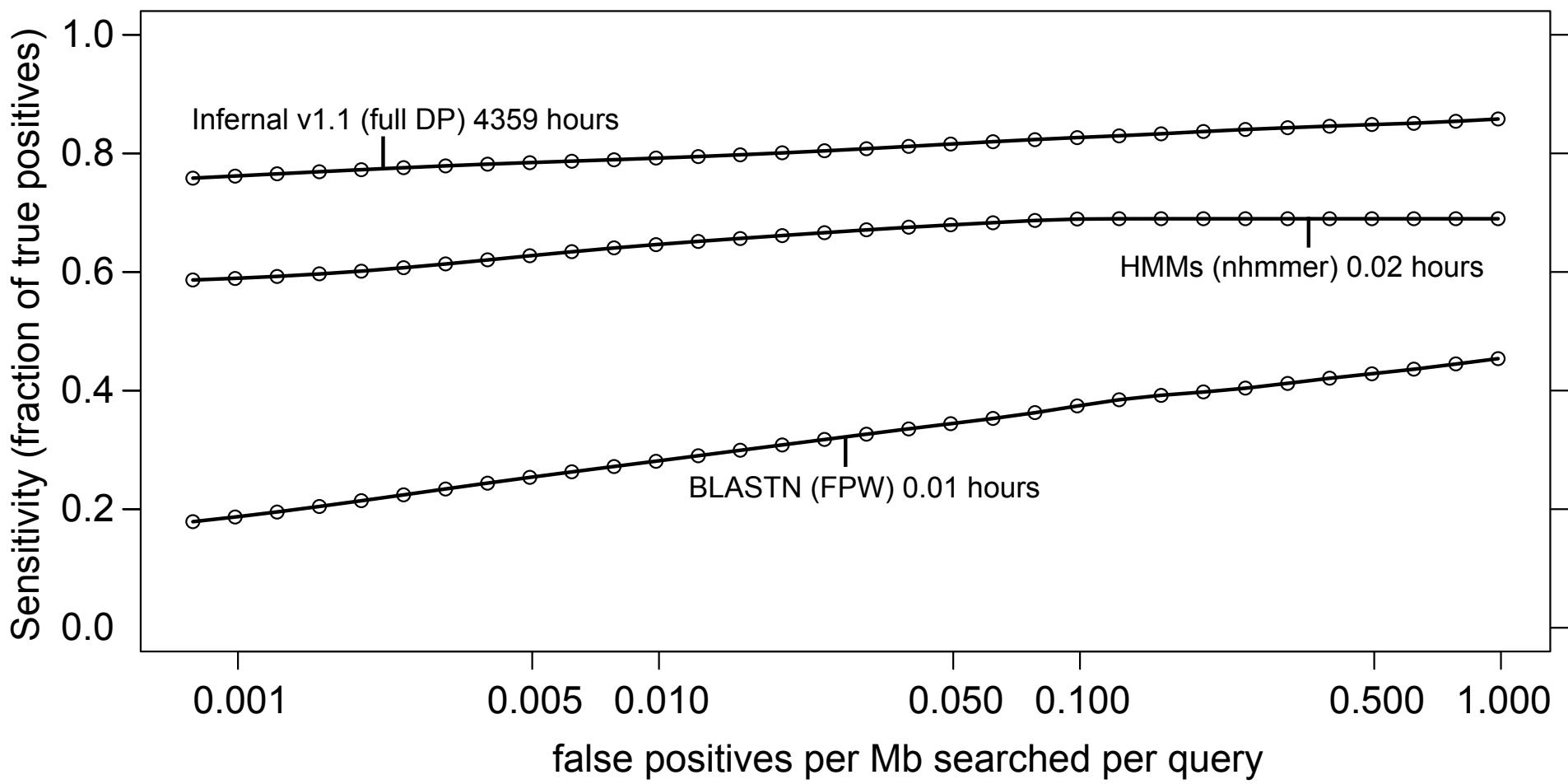


<http://hmmer.org>  
Potter et al. NAR  
46:W200-204  
Wheeler, TJ, Eddy SR.  
Bioinformatics, 29:2487-89, 2013.  
Eddy, SR. PLoS Comp. Biol.,  
7:e1002195, 2011.  
Eddy, SR. Bioinformatics,  
14:755-763, 1998.



<http://eddylab.org/infernal/>  
Nawrocki EP, Eddy SR.  
Bioinformatics, 29:  
2487-2489, 2013.  
Eddy SR, Durbin R.  
Nucleic Acids Research,  
22:2079-2088, 1994.

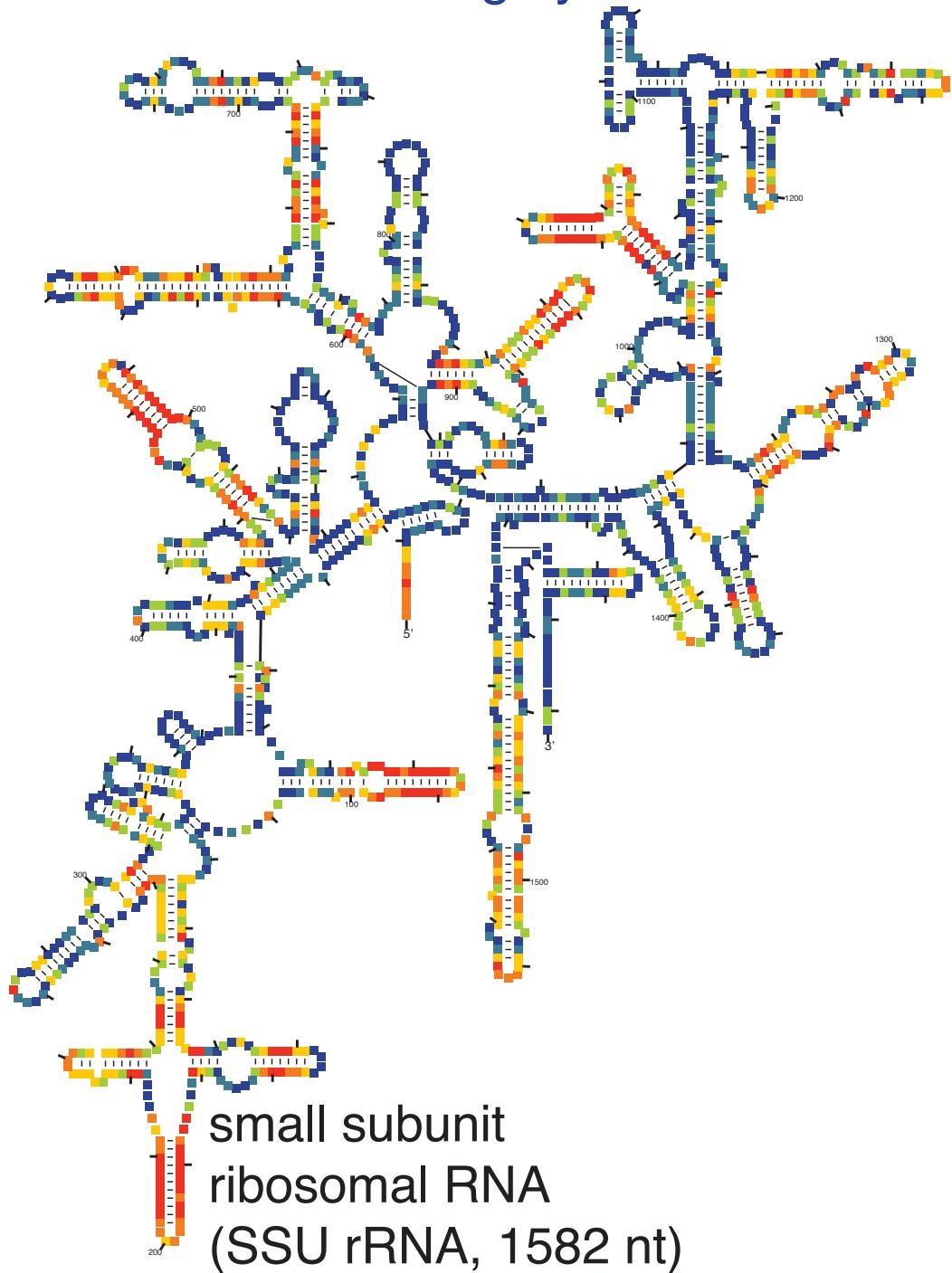
# Infernal outperforms primary-sequence based methods on our benchmark (and others\*, not shown)



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

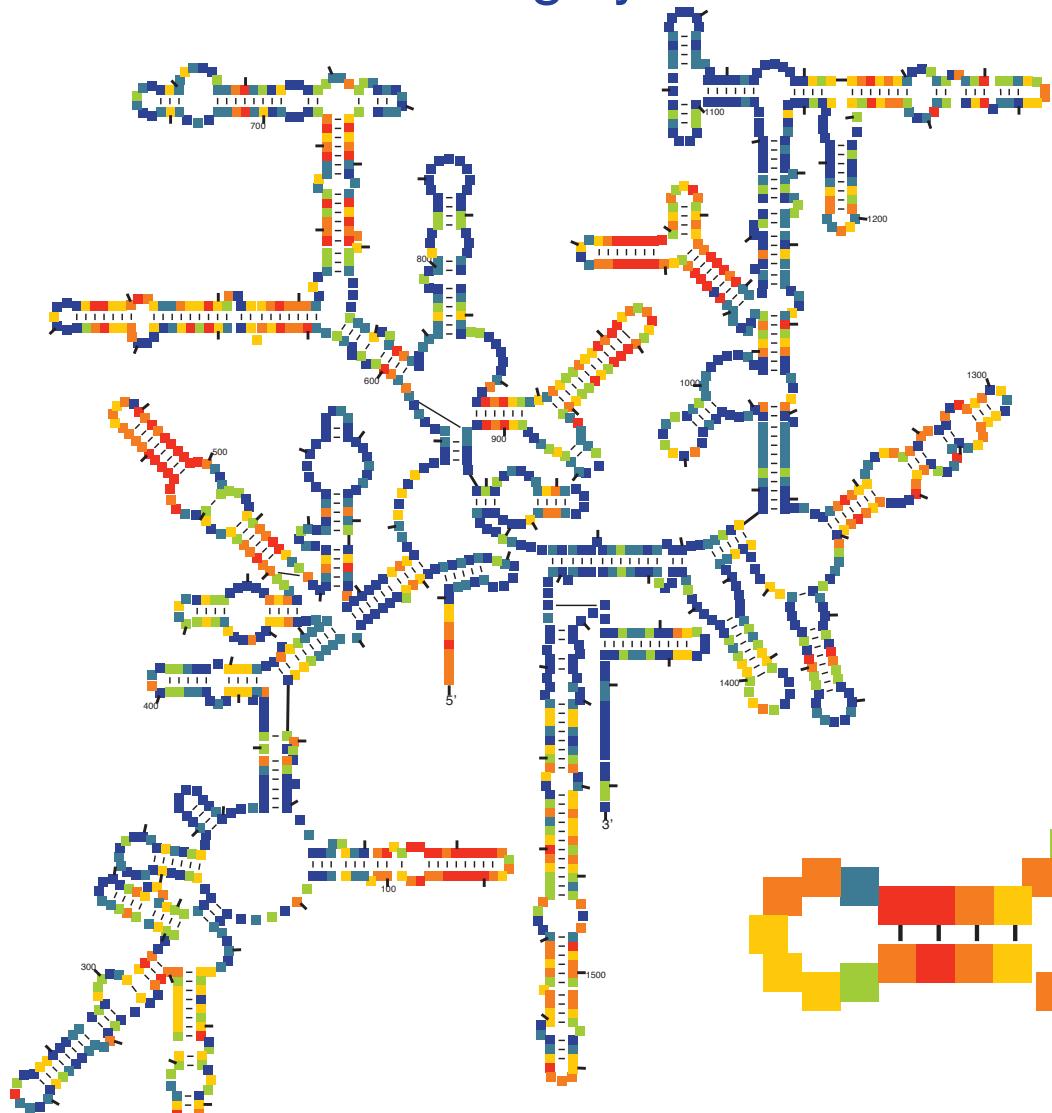
# Sequence conservation per position

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

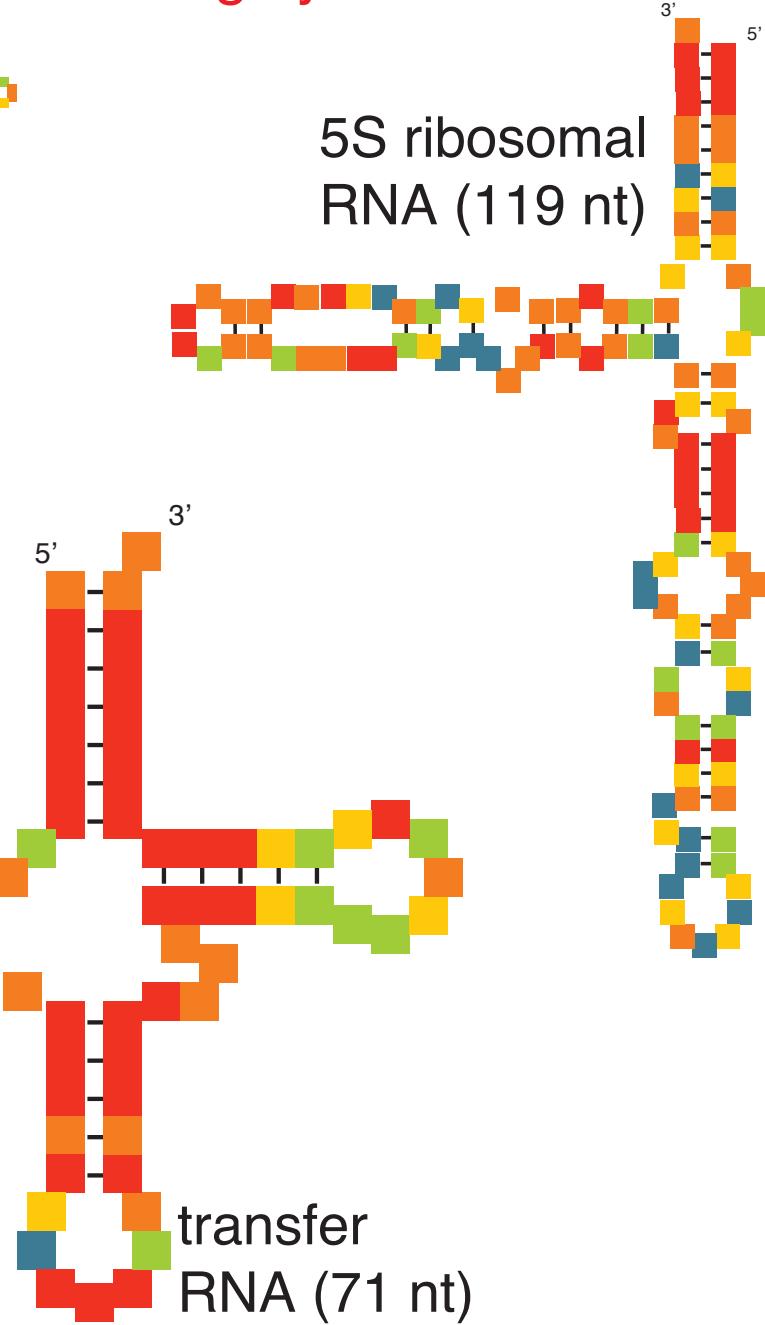


# Sequence conservation per position

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



small subunit  
ribosomal RNA  
(SSU rRNA, 1582 nt)

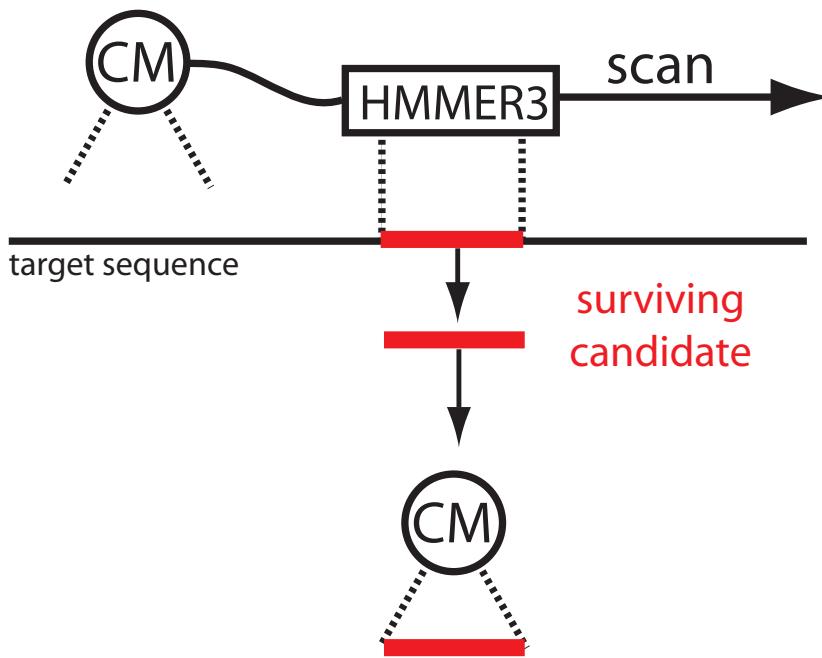


5S ribosomal  
RNA (119 nt)

transfer  
RNA (71 nt)

# Filter target database using profile HMMs\*

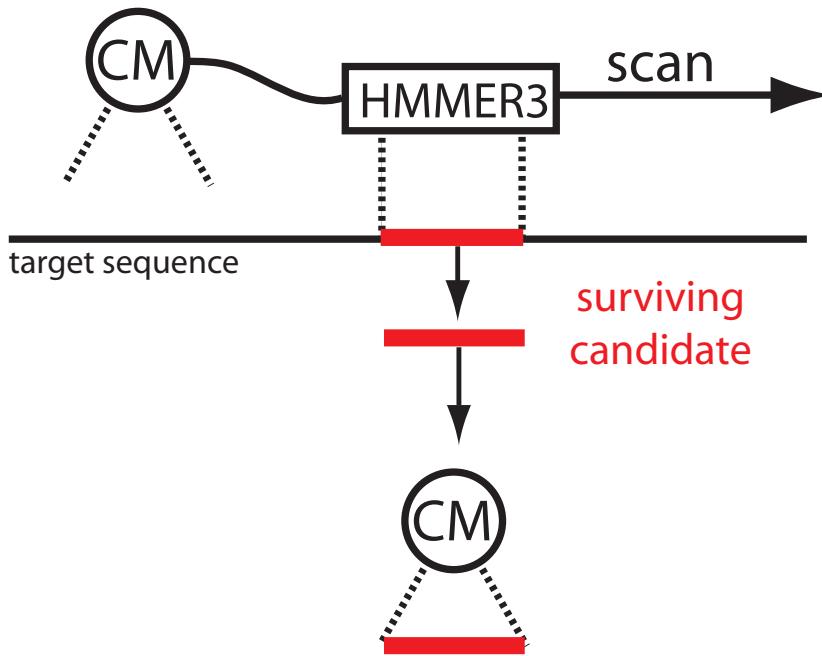
HMM filter first pass



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

# Filter target database using profile HMMs\*

HMM filter first pass



- 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

yeast    GUGUUCGCUAC  
 human    -UCUUCGGCG-  
 fly       AGAUU-GUACU

1              5              11

new  
 sequence:  
 CUGUUCGCAAG



probability  
 "correct":

	1	5	11
C			
U			
G			
U			
U			
C			
G			
C			
A			
G			

CUGUUCGCA-G

1.0  
 .8  
 .6  
 .4  
 .2  
 0



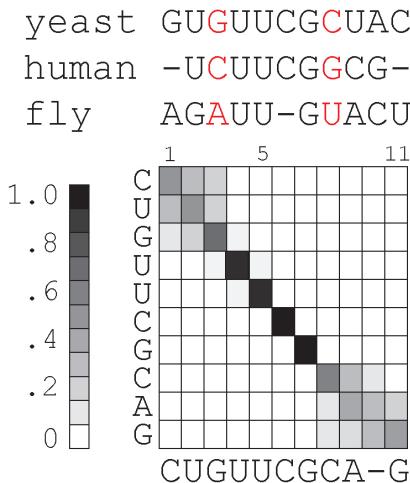
1              5              11

	1	5	11
C			
U			
G			
U			
U			
C			
G			
C			
A			
G			

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

**HMMs -**

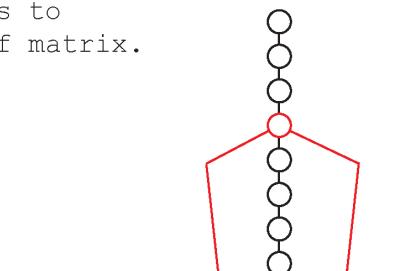
Each column of seed alignment corresponds to a column of matrix.



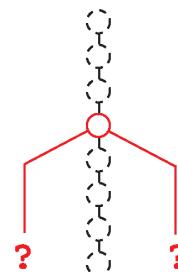
**CMs -**

Each column of seed alignment corresponds to a state.

yeast	human	fly
U C	U C	U
U G	U G	U G
<b>G•C</b>	<b>C•G</b>	<b>A•U</b>
U•A U	U•G C	G•C A
G•C		A•U



struct <<----->->  
 yeast GUGUUCG**C**UAC  
 human -UCUUCGG**G**CG-  
 fly AG**A**UU-G**U**ACU



CUGUUCGCAG  
 45 possibilities

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

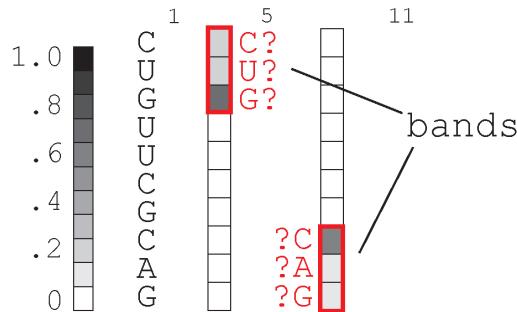
**HMMs -**

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

yeast GUGUUCGCUAC

human -UCUUCGGCG-

fly AGAUU-GUACU



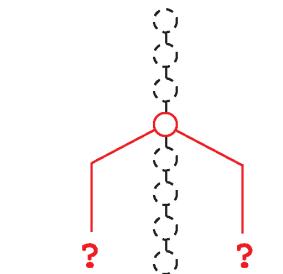
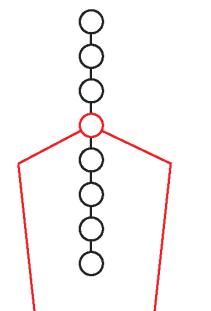
struct <<----->->>  
yeast GUGUUCGCUAC  
human -UCUUCGGCG-  
fly AGAUU-GUACU

**CMs -**

Each column of seed alignment corresponds to a state.

yeast      human      fly

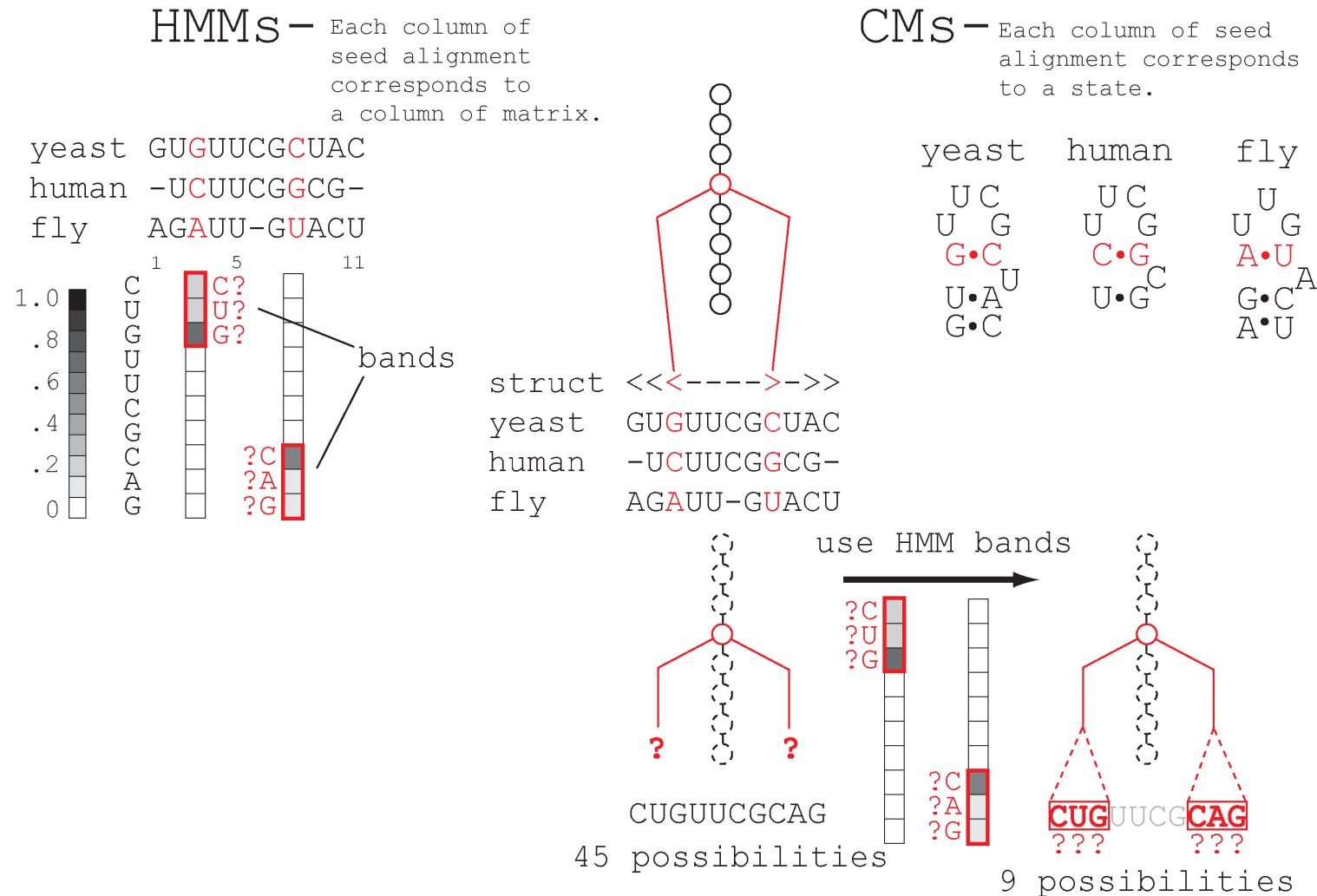
U C	U C	U
U G	U G	U G
G•C	C•G	A•U
U•A U	U•G C	G•C A
G•C		A•U



CUGUUCGCAG

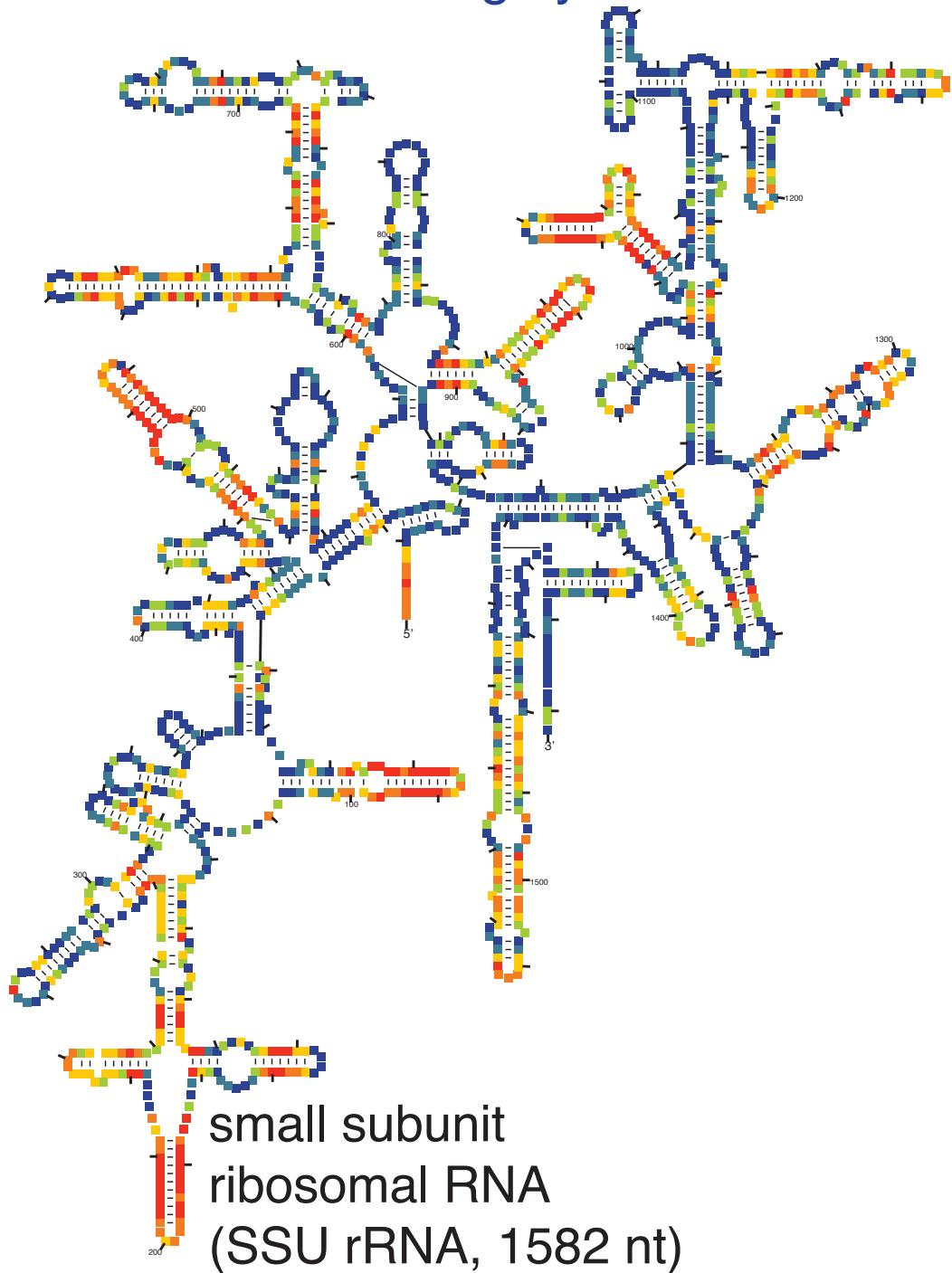
45 possibilities

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



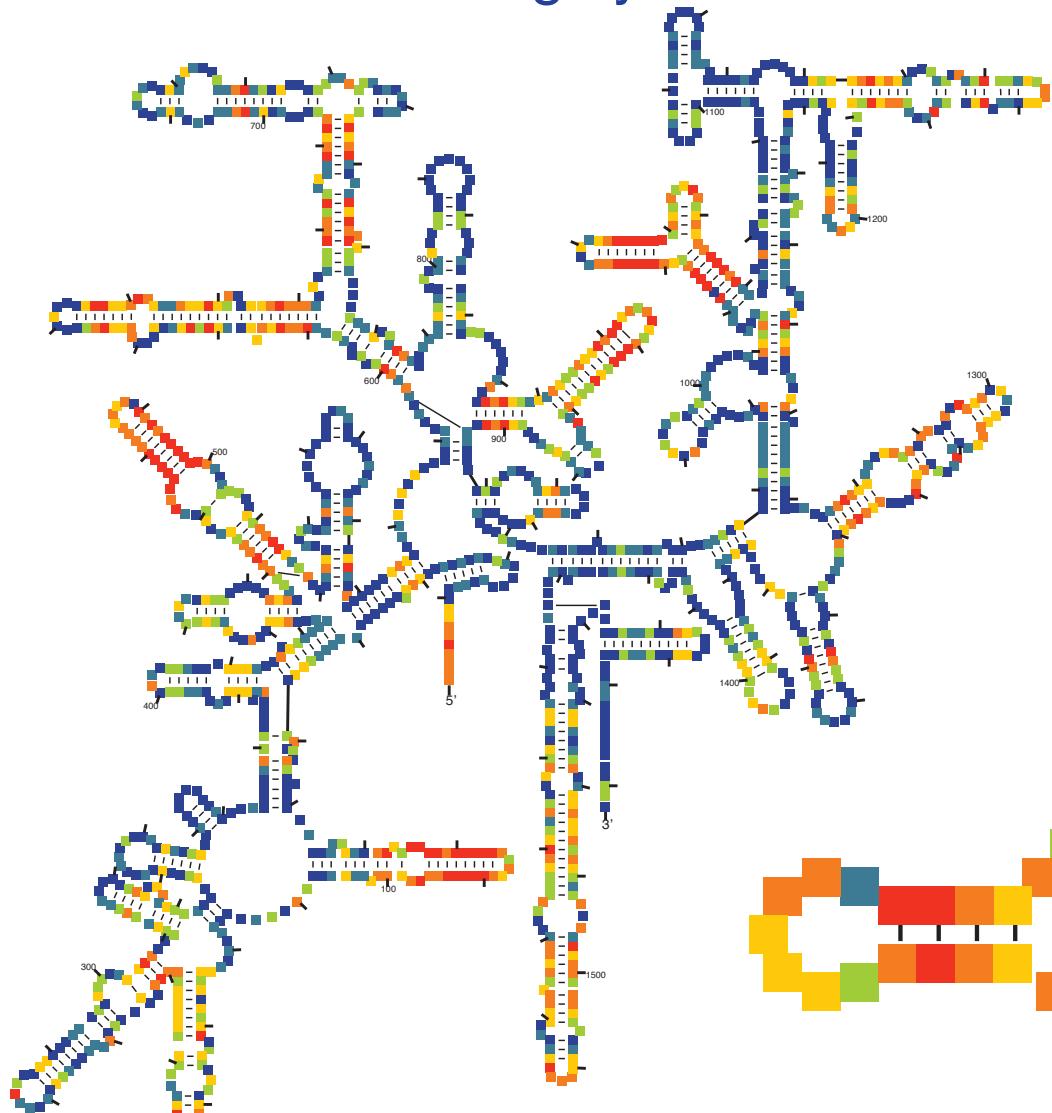
# Sequence conservation per position

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

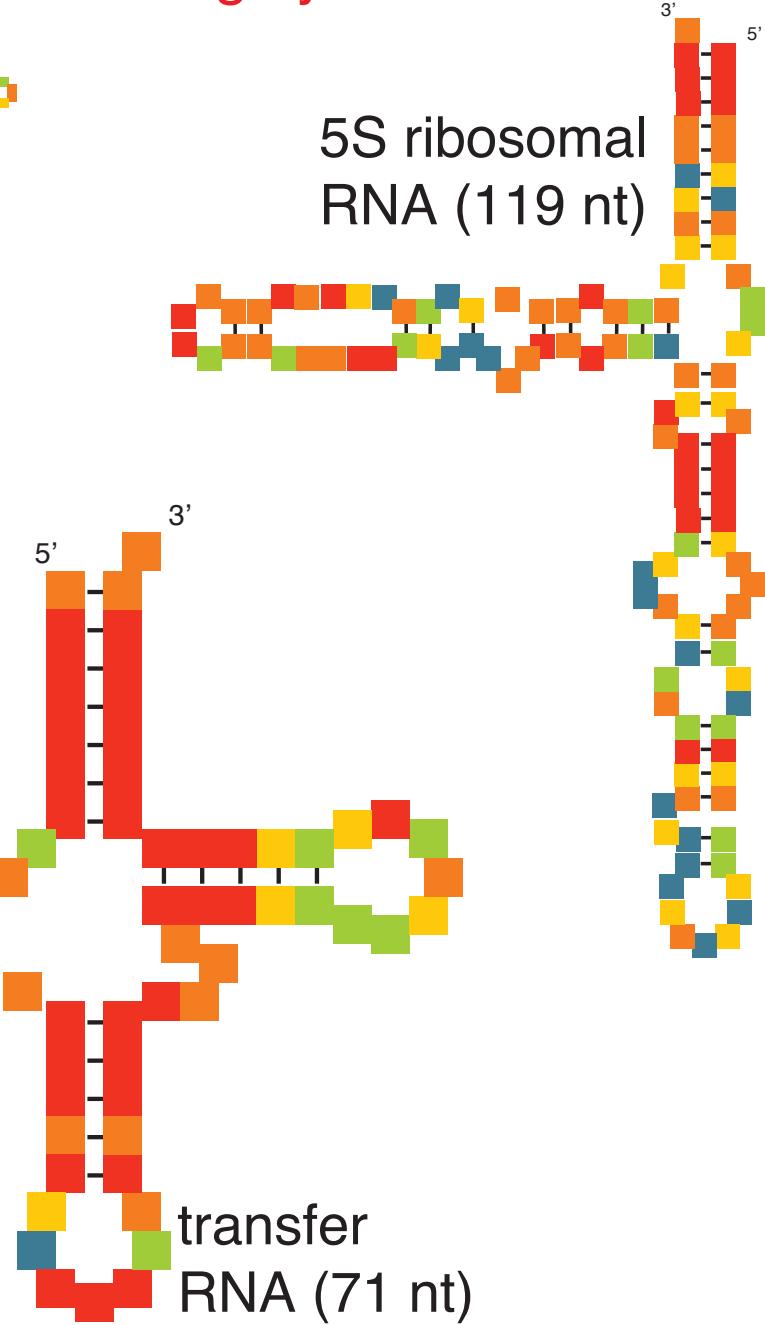


# Sequence conservation per position

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



small subunit  
ribosomal RNA  
(SSU rRNA, 1582 nt)

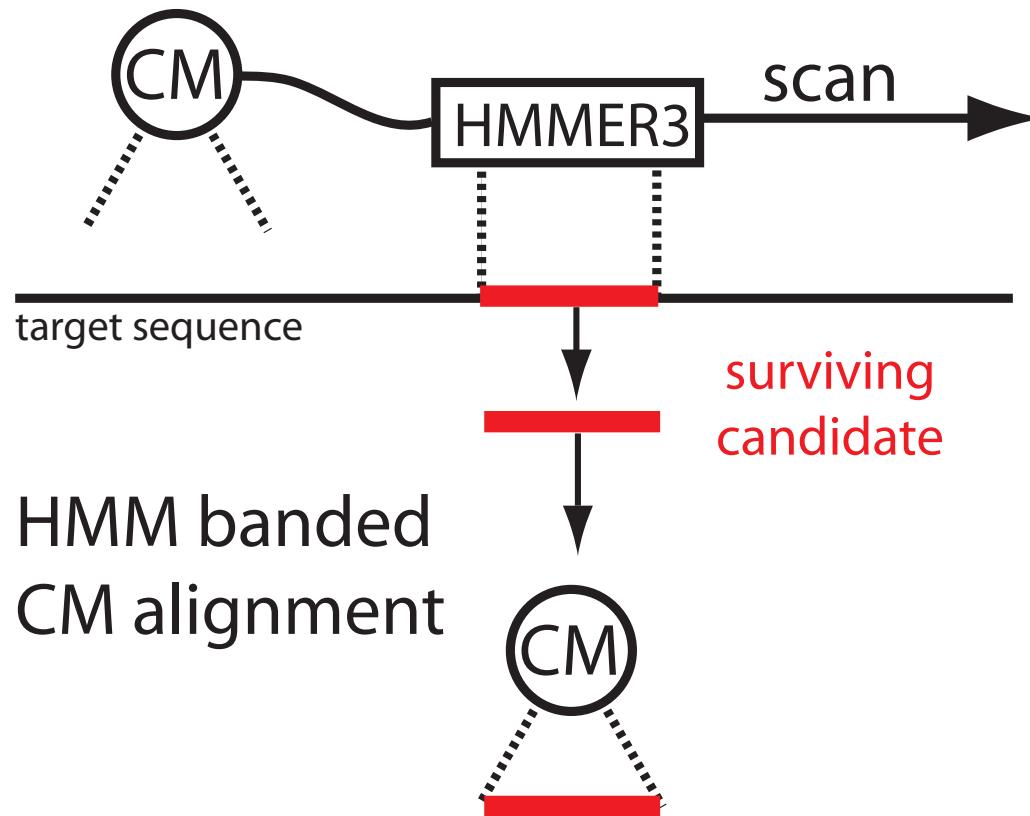


5S ribosomal  
RNA (119 nt)

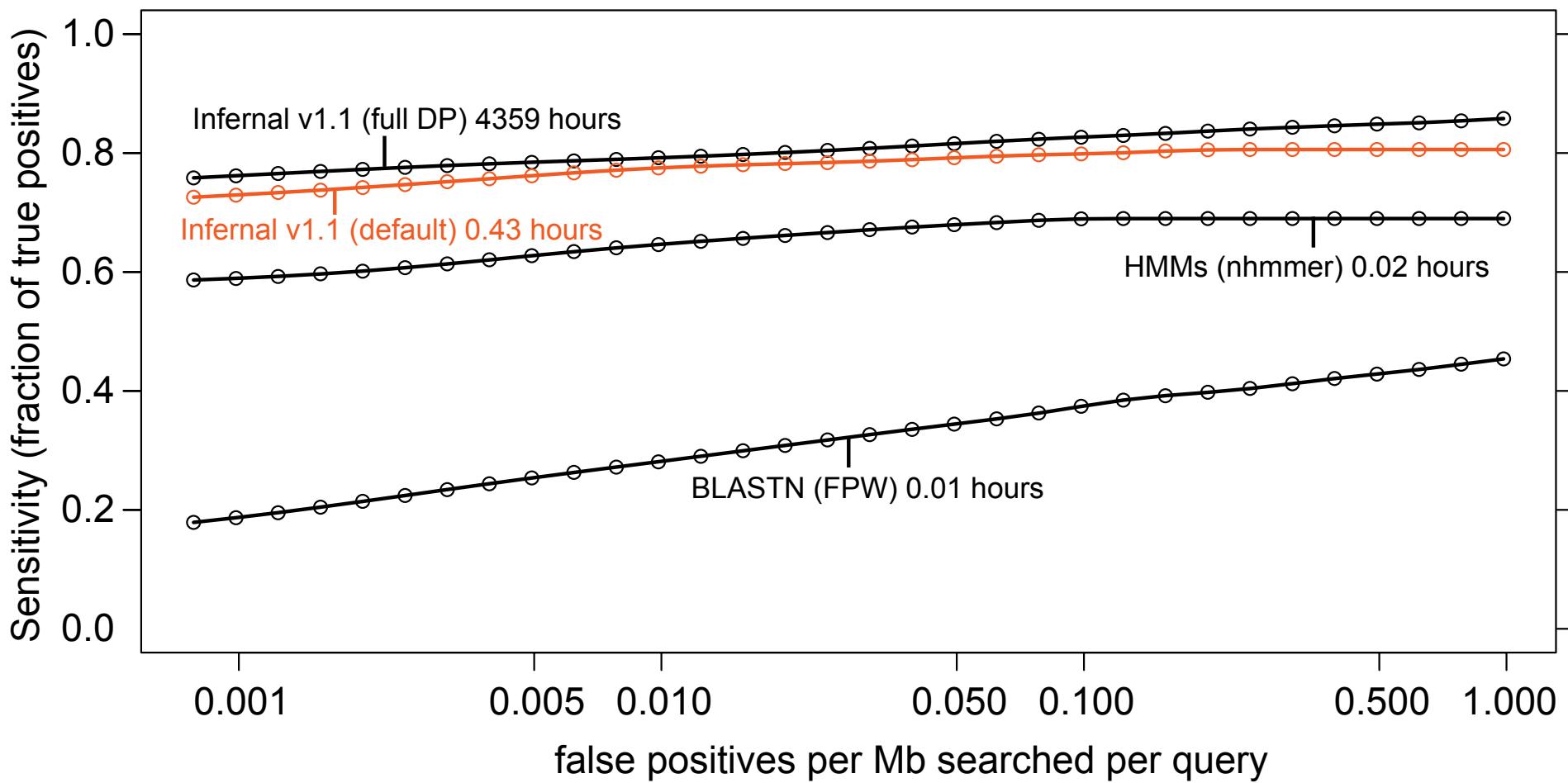
transfer  
RNA (71 nt)

# Use HMMs as filters and to constrain CM alignment

## HMM filter first pass



# HMM-based acceleration makes Infernal 10,000 times faster



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

# Practical structural RNA genome annotation

- Faster Infernal integrated into:
  - NCBI prokaryotic genome annotation pipeline PGAP
  - NCBI eukaryotic genome annotation (Françoise Thibaud-Nissen)

6614–6624 *Nucleic Acids Research*, 2016, Vol. 44, No. 14  
doi: 10.1093/nar/gkw569

Published online 24 June 2016

## NCBI prokaryotic genome annotation pipeline

Tatiana Tatusova<sup>1,†</sup>, Michael DiCuccio<sup>1,†</sup>, Azat Badretdin<sup>1</sup>, Vyacheslav Chetvernin<sup>1</sup>, Eric P. Nawrocki<sup>1</sup>, Leonid Zaslavsky<sup>1</sup>, Alexandre Lomsadze<sup>2</sup>, Kim D. Pruitt<sup>1</sup>, Mark Borodovsky<sup>2,3,\*‡</sup> and James Ostell<sup>1,‡</sup>

<sup>1</sup>National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA,

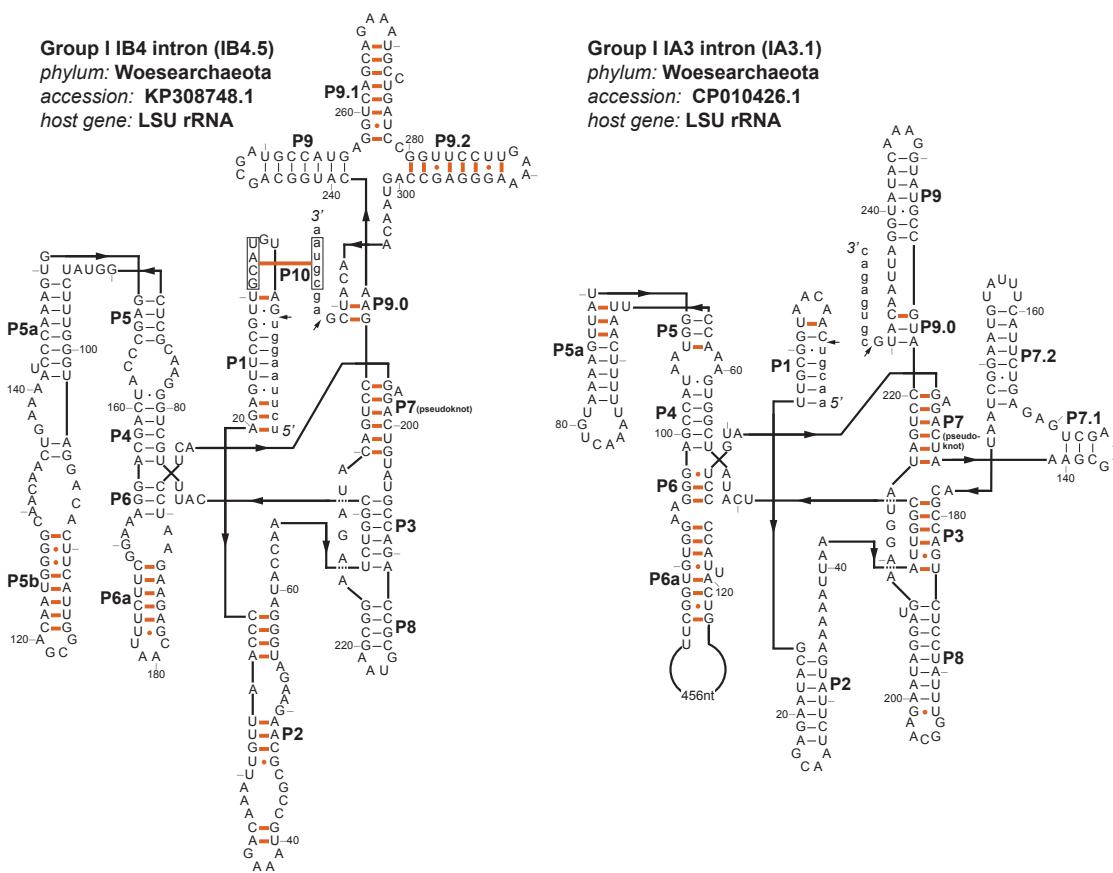
<sup>2</sup>Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332, USA and <sup>3</sup>School of Computational Science and Engineering, Georgia Tech, Atlanta, GA 30332, USA

# Group I introns are widespread in archaea

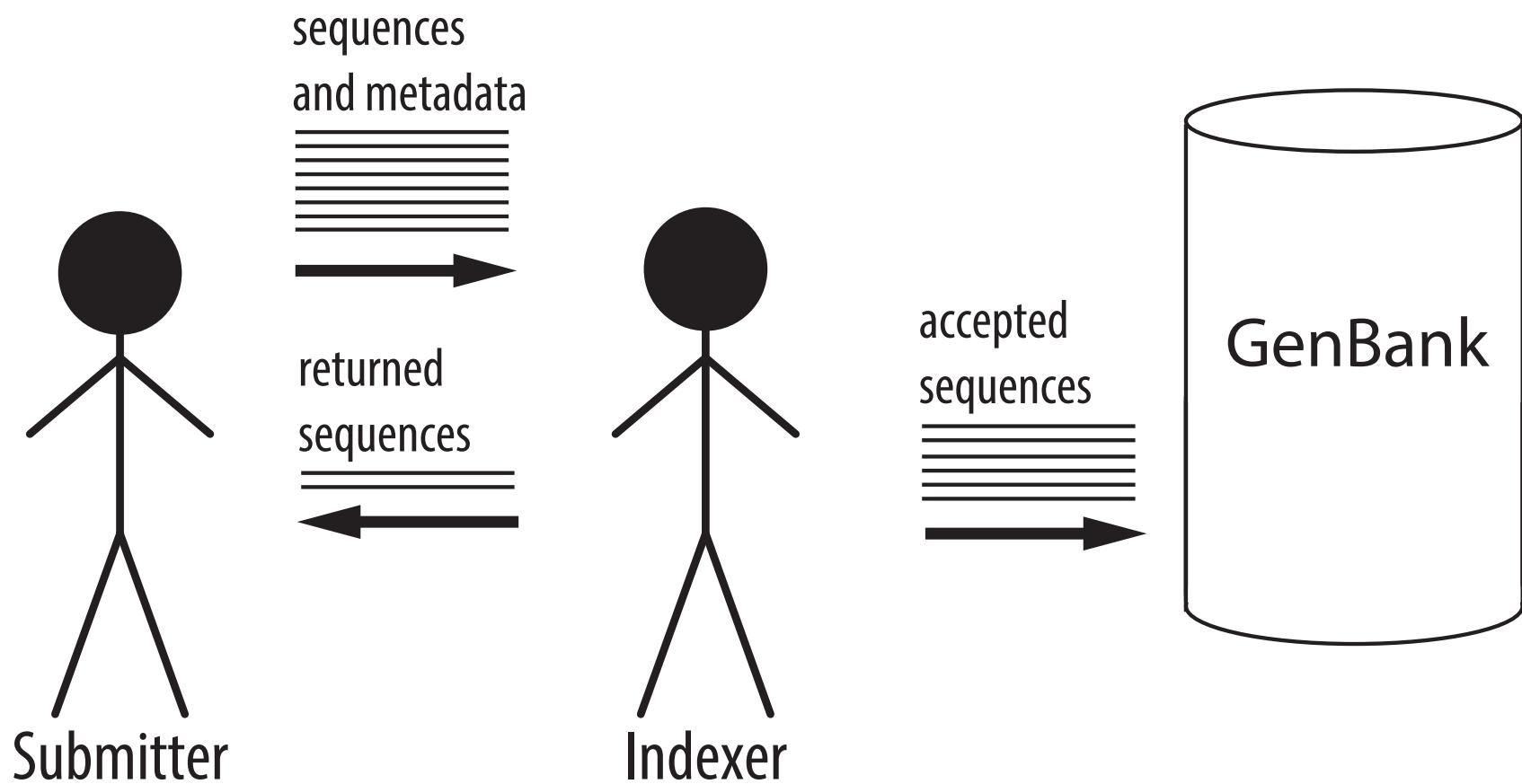
Eric P. Nawrocki<sup>1,\*</sup>, Thomas A. Jones<sup>2,3</sup> and Sean R. Eddy<sup>2,3,4,\*</sup>

<sup>1</sup>National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA,

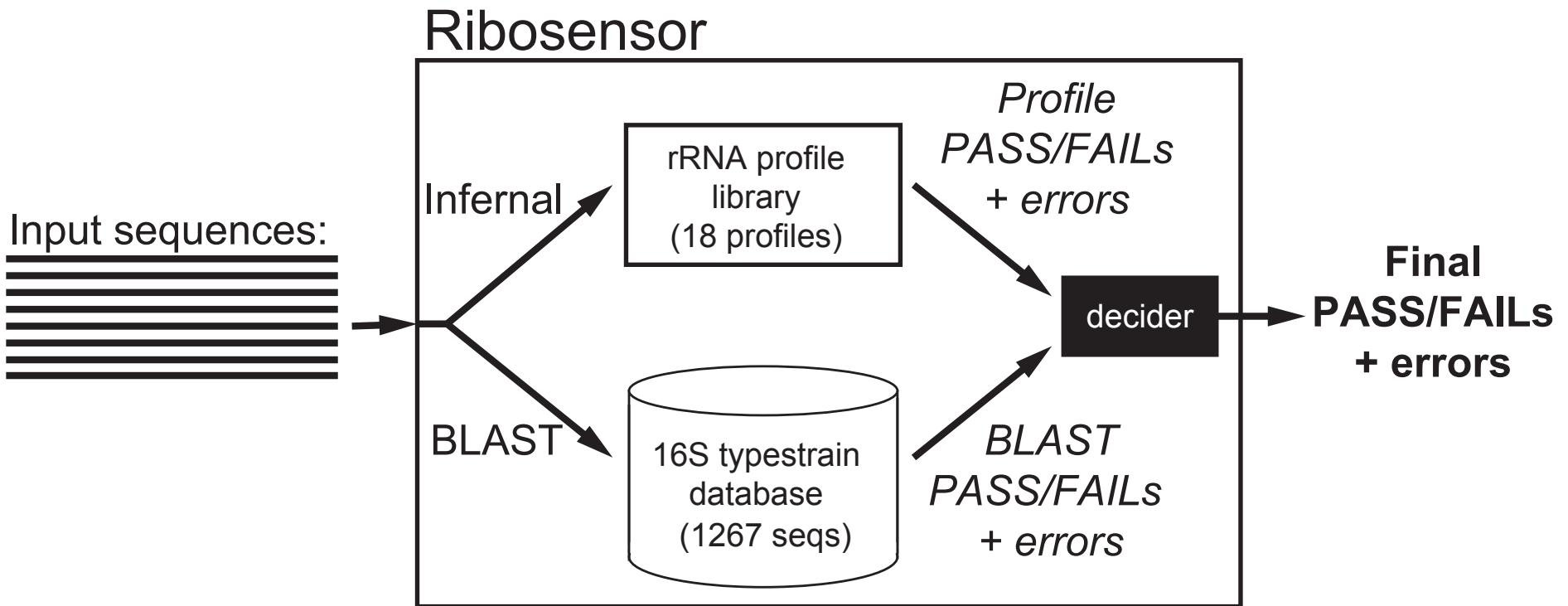
<sup>2</sup>Howard Hughes Medical Institute, Harvard University, Cambridge, USA, <sup>3</sup>Department of Molecular and Cellular Biology, Harvard University, Cambridge, USA and <sup>4</sup>School of Engineering and Applied Sciences, Harvard University, Cambridge, USA



# GenBank indexers handle incoming sequence submissions



# Ribosensor: a tool for evaluating ribosomal RNA datasets using profiles and BLAST\*



- Profile-based analysis:
  - 18 ribosomal RNA models (15 SSU rRNA, 3 LSU rRNA); 8 from Rfam
  - Detects unexpected features ("UnacceptableModel", "DuplicatedRegions", etc.)
- Profile and BLAST results considered together to determine PASS/FAIL

\*Alejandro Schäffer developed the BLAST-based scheme

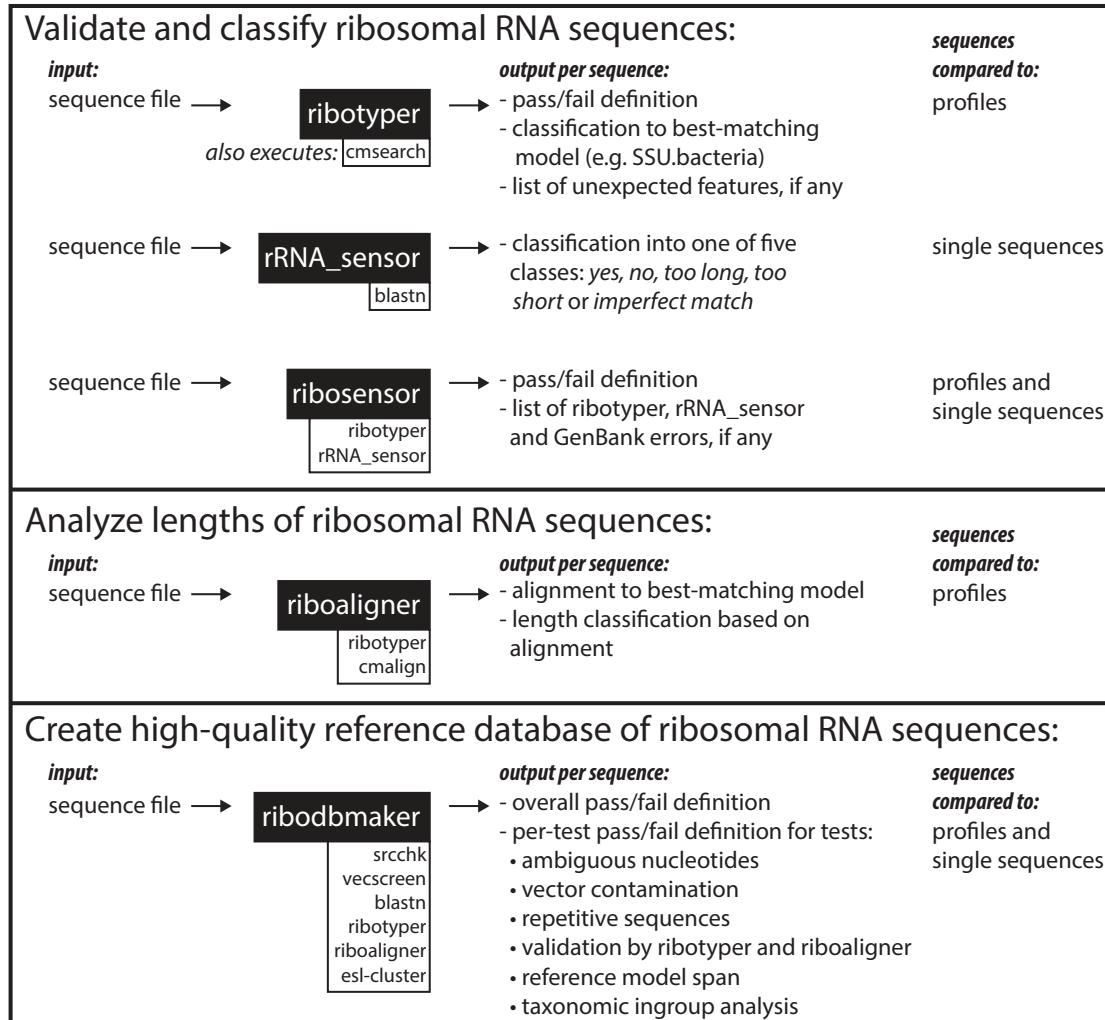
SOFTWARE

Open Access



# Ribovore: ribosomal RNA sequence analysis for GenBank submissions and database curation

Alejandro A. Schäffer<sup>1,2</sup>, Richard McVeigh<sup>2</sup>, Barbara Robbertse<sup>2</sup>, Conrad L. Schoch<sup>2</sup>, Anjanette Johnston<sup>2</sup>, Beverly A. Underwood<sup>2</sup>, Ilene Karsch-Mizrachi<sup>2</sup> and Eric P. Nawrocki<sup>2\*</sup> 



# Rfam 15: RNA families database in 2025

**Nancy Ontiveros-Palacios**  <sup>1</sup>, **Emma Cooke**  <sup>2</sup>, **Eric P. Nawrocki**  <sup>3</sup>, **Sandra Triebel**  <sup>4,5</sup>,  
**Manja Marz**  <sup>4,5</sup>, **Elena Rivas**  <sup>6</sup>, **Sam Griffiths-Jones**  <sup>7</sup>, **Anton I. Petrov**  <sup>8</sup>, **Alex Bateman**  <sup>1</sup> and  
**Blake Sweeney**  <sup>1,\*</sup>

<sup>1</sup>European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

<sup>2</sup>SciBite Limited, BioData Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK

<sup>3</sup>National Center for Biotechnology Information, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

<sup>4</sup>RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany

<sup>5</sup>European Virus Bioinformatics Center, Friedrich Schiller University Jena, 07743 Jena, Germany

<sup>6</sup>Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

<sup>7</sup>School of Biological Sciences, Faculty of Medicine, Biology and Health, Michael Smith Building, The University of Manchester, Dover St, Manchester M13 9NT, UK

<sup>8</sup>Riboscope Ltd, Cambridge CB1 1AH, UK

\*To whole correspondence should be addressed. Tel: +44 1223 494359; Email: bsweeney@ebi.ac.uk

# RNAcentral: a hub of information for non-coding RNA sequences

The RNAcentral Consortium<sup>1–38,\*</sup>

D212–D220 Nucleic Acids Research, 2021, Vol. 49, Database issue  
doi: 10.1093/nar/gkaa921

Published online 27 October 2020

# RNAcentral 2021: secondary structure integration, improved sequence search and new member databases

RNAcentral Consortium<sup>1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,\*</sup>

ARTICLE



<https://doi.org/10.1038/s41467-021-23555-5>

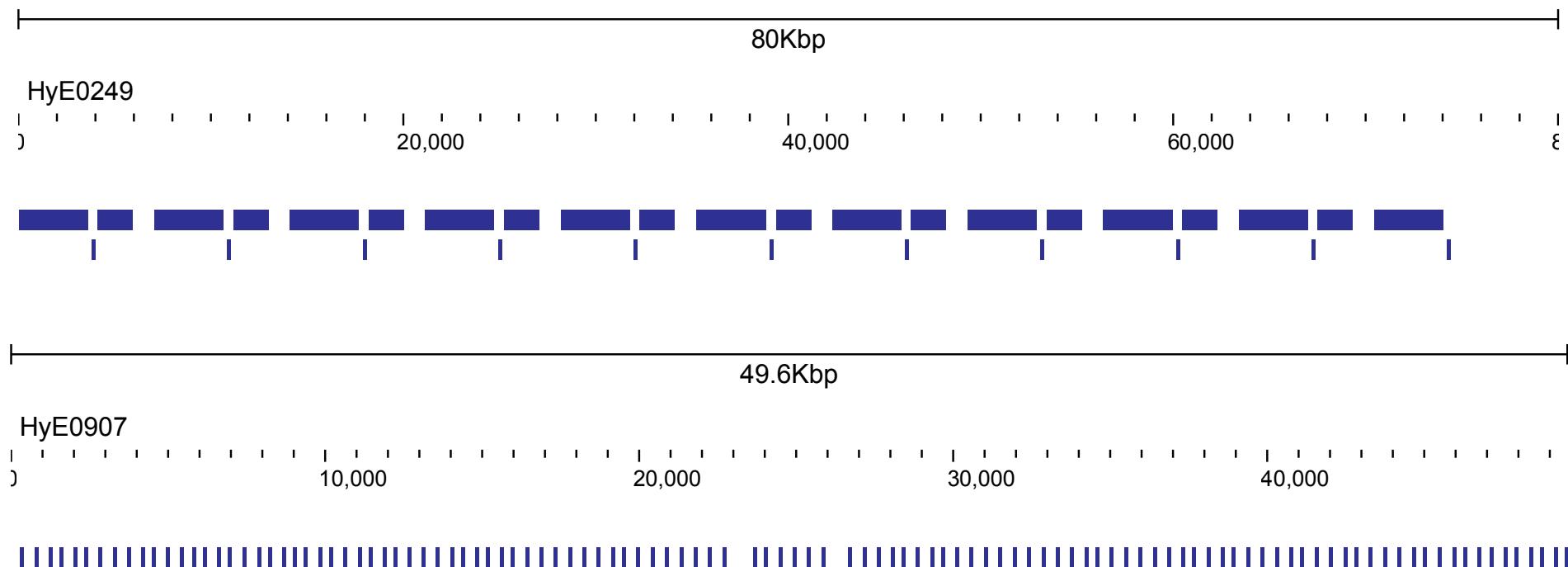
OPEN

# R2DT is a framework for predicting and visualising RNA secondary structure using templates

Blake A. Sweeney<sup>1,7</sup>, David HokszaID<sup>2,7</sup>, Eric P. Nawrocki<sup>3</sup>, Carlos Eduardo RibasID<sup>1</sup>, Fábio MadeiraID<sup>1</sup>, Jamie J. Cannone<sup>4</sup>, Robin Gutell<sup>4</sup>, Aparna Maddala<sup>5</sup>, Caeden D. Meade<sup>5</sup>, Loren Dean WilliamsID<sup>5</sup>, Anton S. PetrovID<sup>5</sup>, Patricia P. ChanID<sup>6</sup>, Todd M. Lowe<sup>6</sup>, Robert D. FinnID<sup>1,8</sup> & Anton I. PetrovID<sup>1,8</sup>✉

# The genome of the colonial hydroid *Hydractinia* reveals that their stem cells use a toolkit of evolutionarily shared genes with all animals

Christine E. Schnitzler,<sup>1,2</sup> E. Sally Chang,<sup>3,4</sup> Justin Waletich,<sup>1,2</sup>  
Gonzalo Quiroga-Artigas,<sup>1,2,5</sup> Wai Yee Wong,<sup>6</sup> Anh-Dao Nguyen,<sup>3</sup> Sofia N. Barreira,<sup>3</sup>  
Liam B. Doonan,<sup>7</sup> Paul Gonzalez,<sup>3</sup> Sergey Koren,<sup>3</sup> James M. Gahan,<sup>7,8</sup>  
Steven M. Sanders,<sup>9,10</sup> Brian Bradshaw,<sup>7</sup> Timothy Q. DuBuc,<sup>7,11</sup> Febrimarsa,<sup>7,12</sup>  
Danielle de Jong,<sup>1,2</sup> Eric P. Nawrocki,<sup>4</sup> Alexandra Larson,<sup>1</sup> Samantha Klasfeld,<sup>3</sup>  
Sebastian G. Gornik,<sup>7,13</sup> R. Travis Moreland,<sup>3</sup> Tyra G. Wolfsberg,<sup>3</sup> Adam M. Phillippy,<sup>3</sup>  
James C. Mullikin,<sup>3,14</sup> Oleg Simakov,<sup>6</sup> Pauly Cartwright,<sup>15</sup> Matthew Nicotra,<sup>9,10</sup>  
Uri Frank,<sup>7</sup> and Andreas D. Baxevanis<sup>3</sup>



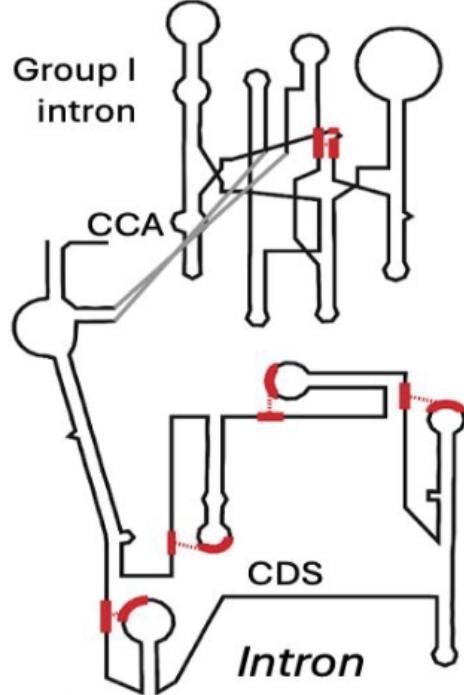
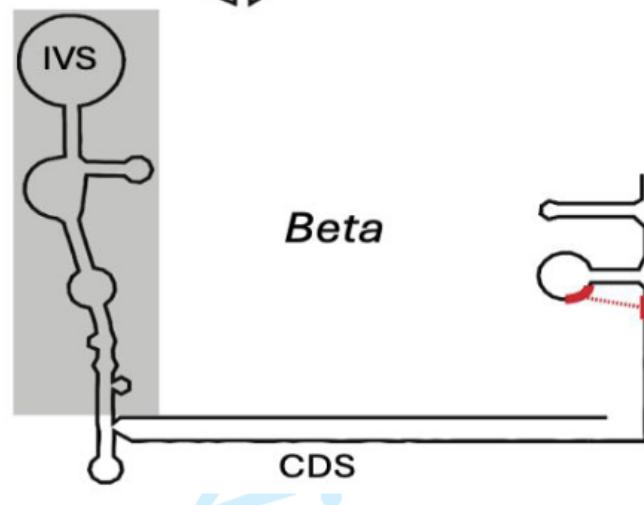
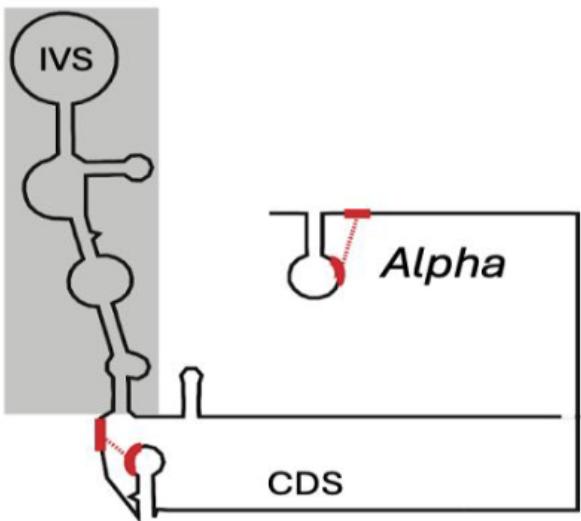
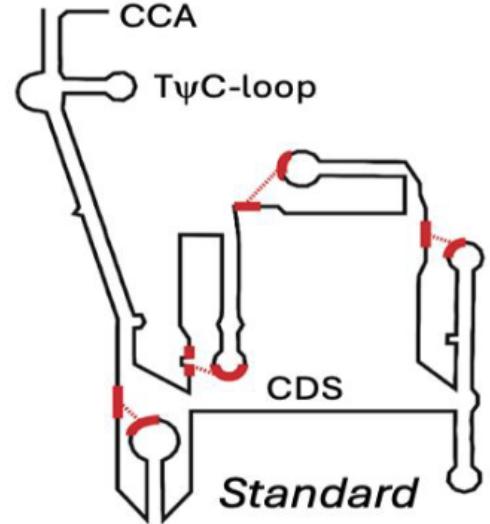
# Expansion of the tmRNA sequence database and new tools for search and visualization

## Authors

Eric P. Nawrocki<sup>1</sup>, Anton I. Petrov<sup>2</sup>, Kelly P. Williams<sup>3</sup>

## Affiliations

1. Division of Intramural Research, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA  
0000-0002-2497-3427  
[eric.nawrocki@nih.gov](mailto:eric.nawrocki@nih.gov)
2. Riboscope Ltd, 23 King Street, Cambridge, CB1 1AH, UK  
0000-0001-7279-2682  
[apetrov@riboscope.com](mailto:apetrov@riboscope.com)
3. Sandia National Laboratories, Livermore CA 94550, USA  
0000-0002-2606-9562  
[kwilli@sandia.gov](mailto:kwilli@sandia.gov)



Model	Length	Bps	Stems	Pknots
<i>Standard</i>	358	117	12	4
<i>Intron</i>	616	180	22	5
<i>Alpha</i>	355	55	8	2
<i>Beta</i>	331	55	7	1
<i>Cyano</i>	288	72	9	1
<i>Mito</i>	77	19	3	0

