Structural RNA and viral sequence analysis

Eric Nawrocki

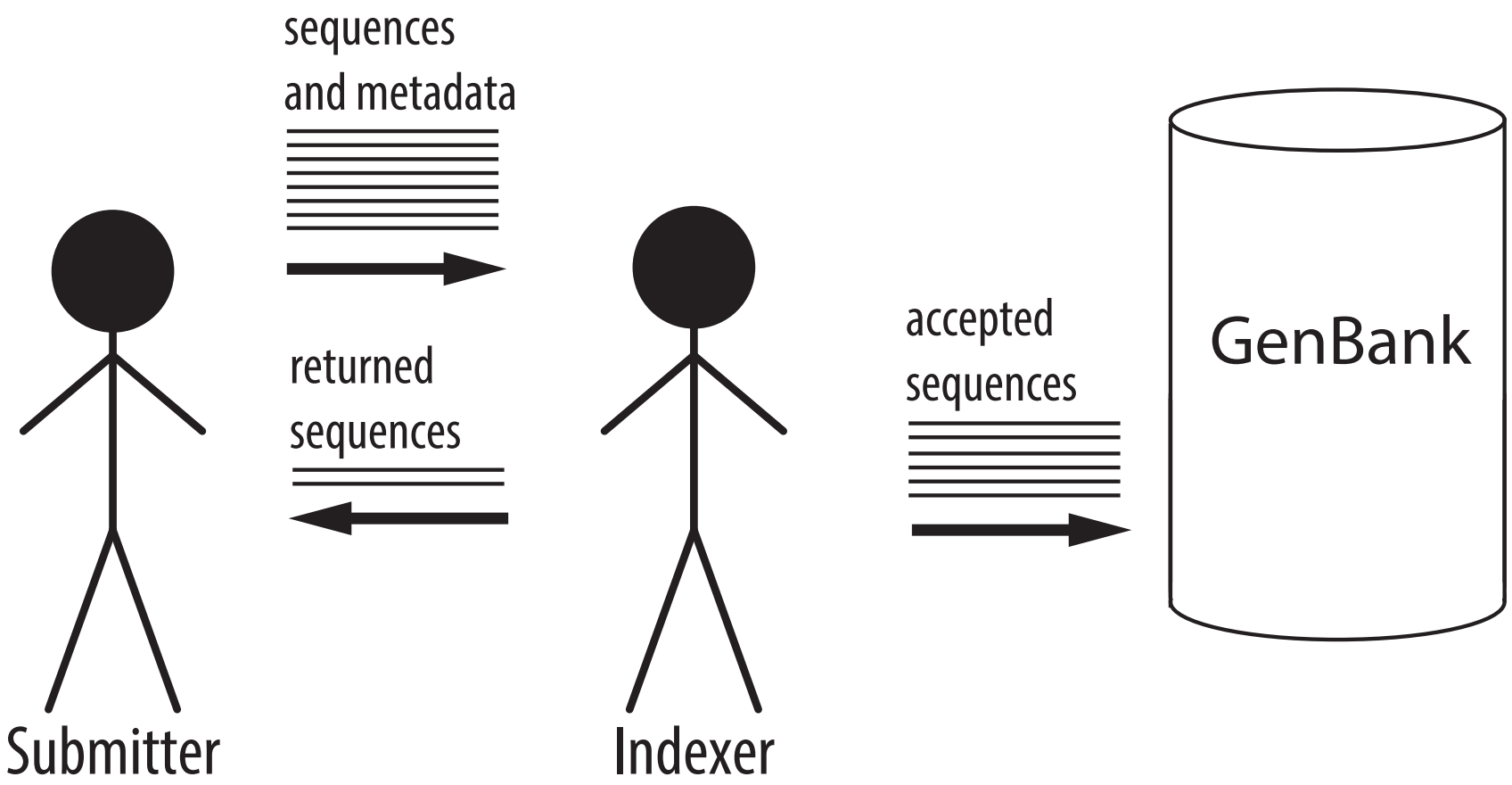
Intramural Research Program
National Library of Medicine
National Institutes of Health

Two main areas of my research:

1. Viral sequence analysis tools, since 2015

2. Structural RNA analysis tools, since 2004

GenBank indexers handle incoming sequence submissions



Schäffer *et al. BMC Bioinformatics* (2020) 21:211
https://doi.org/10.1186/s12859-020-3537-3

BMC Bioinformatics

SOFTWARE Open Access

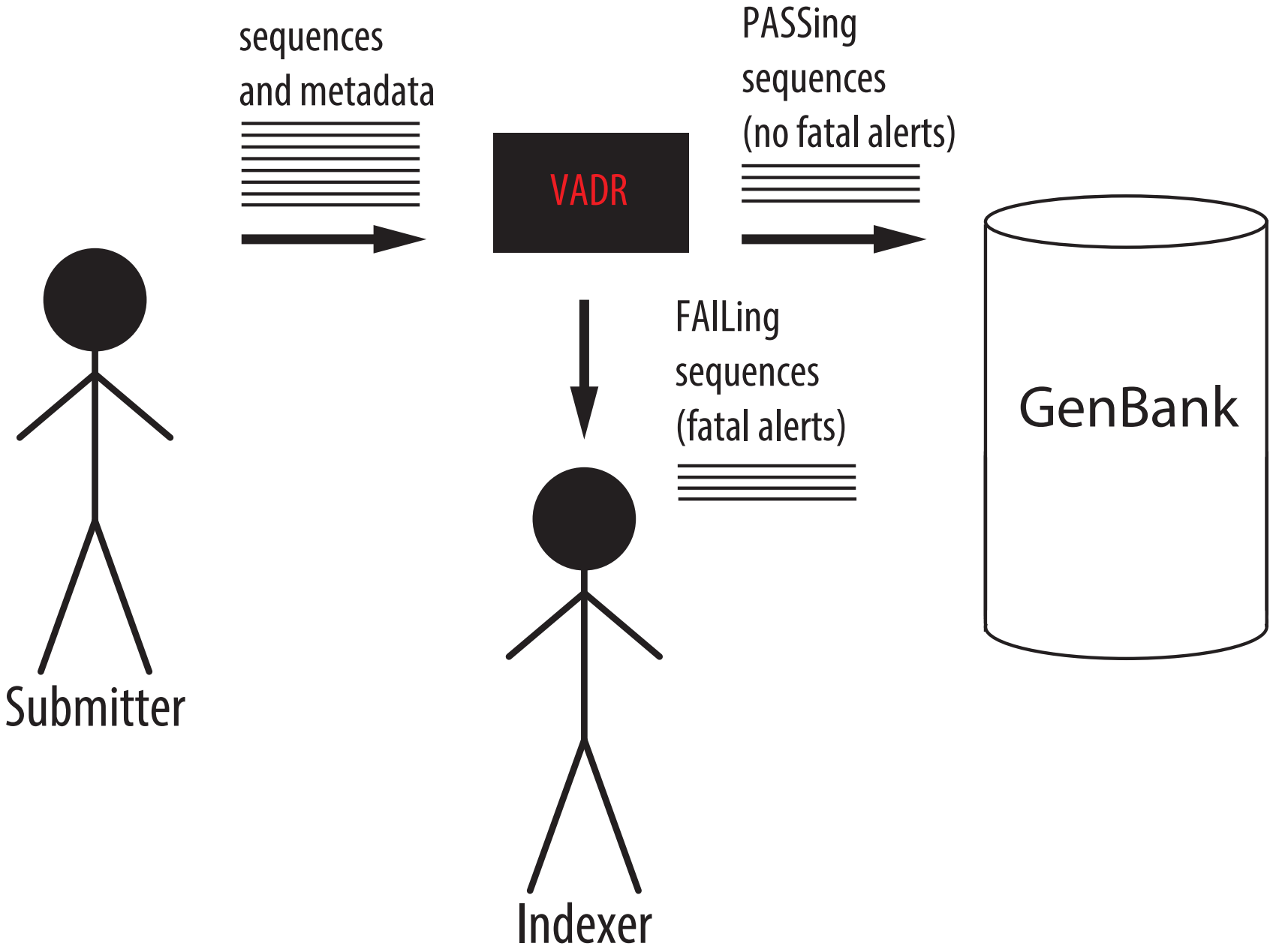
VADR: validation and annotation of virus sequence submissions to GenBank

Alejandro A. Schäffer[1,2], Eneida L. Hatcher[2], Linda Yankie[2], Lara Shonkwiler[2,3], J. Rodney Brister[2], Ilene Karsch-Mizrachi[2] and Eric P. Nawrocki[2*]

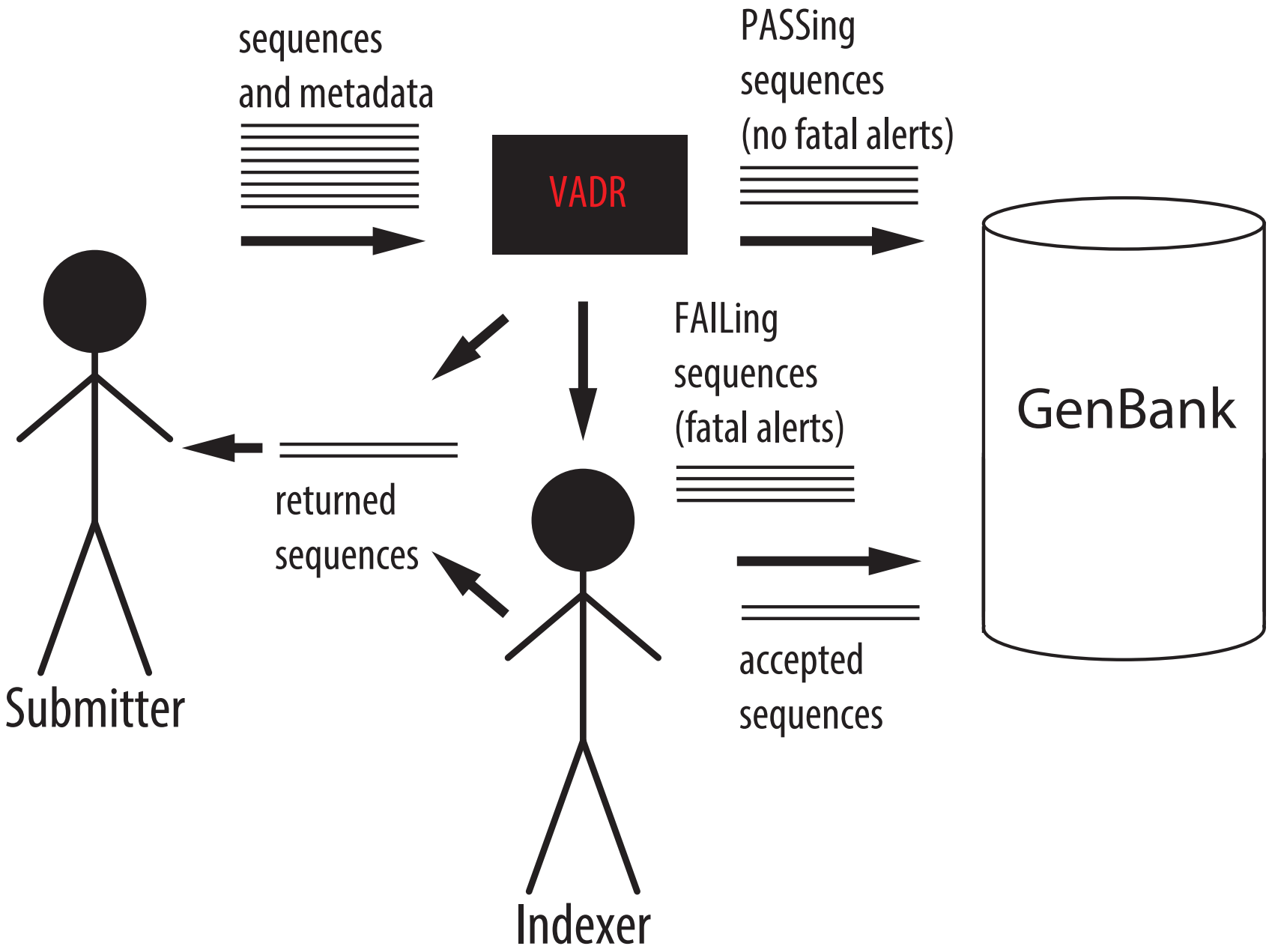
- general tool for reference-based annotation of viral sequences
- used for Norovirus and Dengue virus submissions since 2018
- used for SARS-CoV-2 submissions since March 2020
- currently also used manually for RSV, Mpox, and some Influenza submissions

VADR assists GenBank indexers:

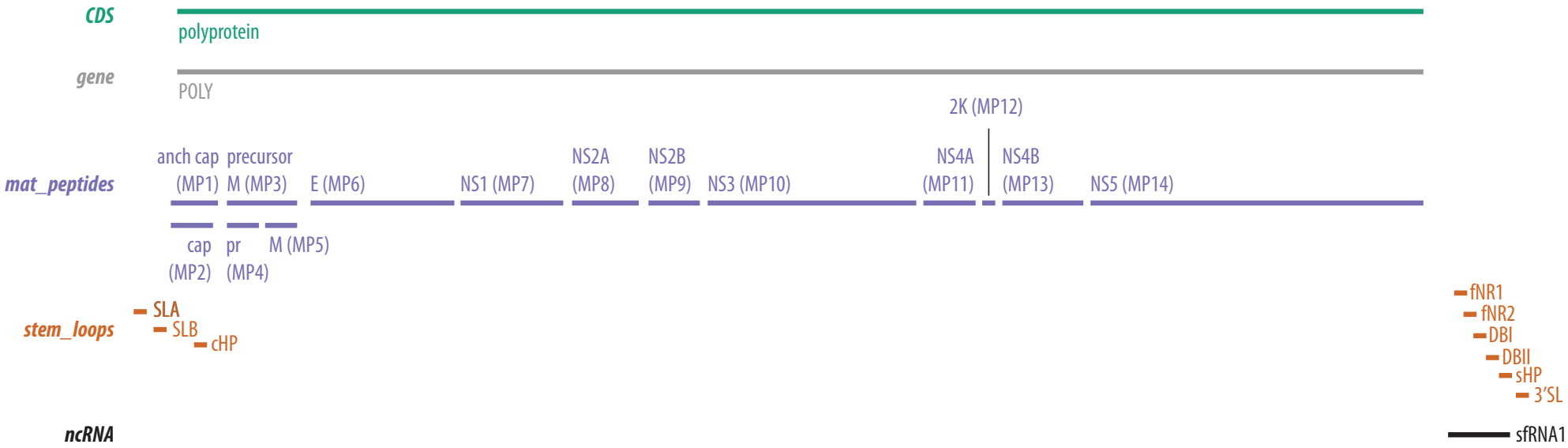
Each sequence PASSes or FAILs

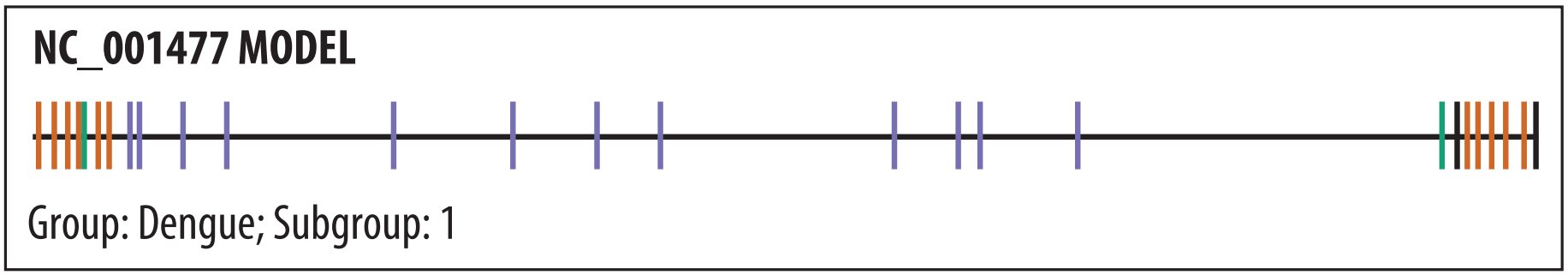


Indexers decide fate of some FAILing sequences
but some are sent directly back to submitter with error reports



VADR builds a reference model of a RefSeq and its features





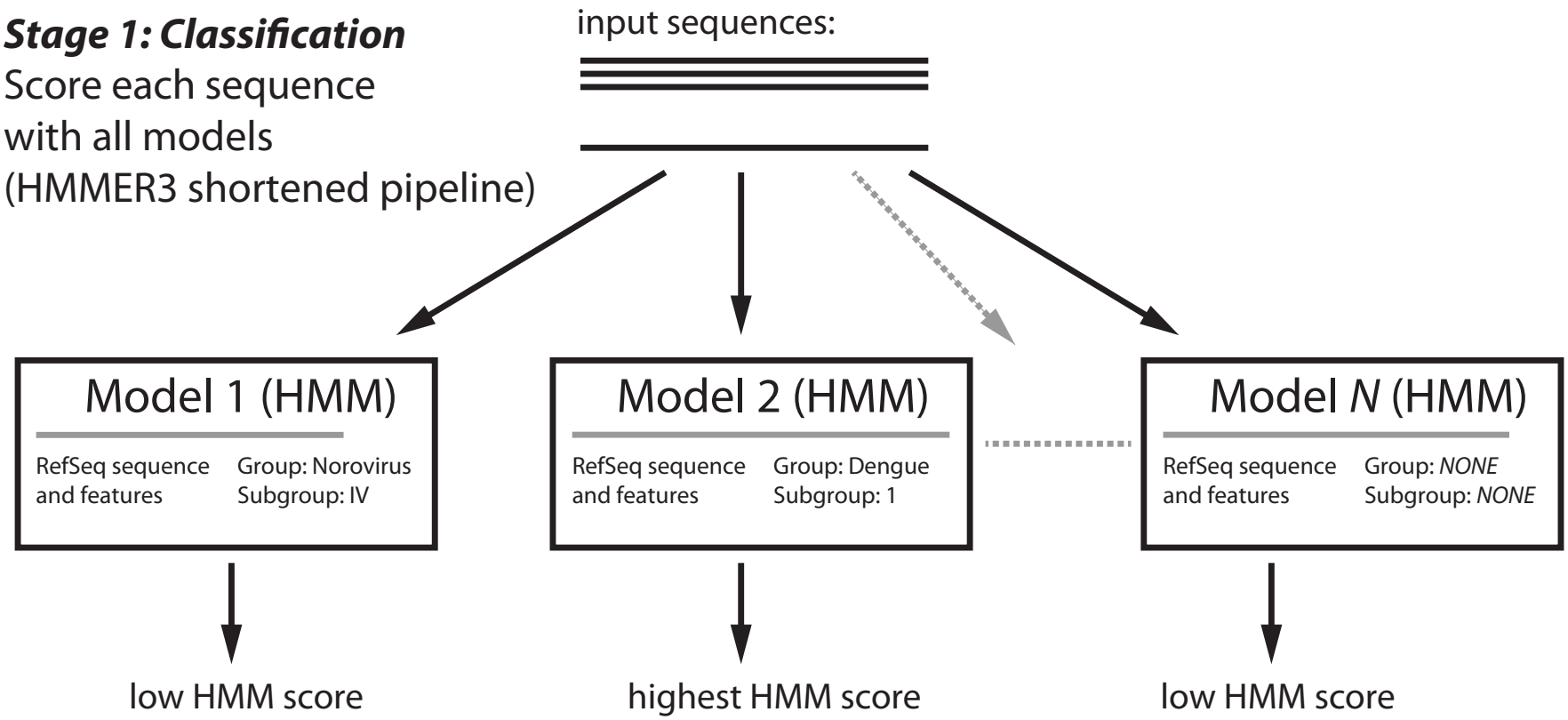
VADR validates and annotates each input sequence using its best-matching model

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

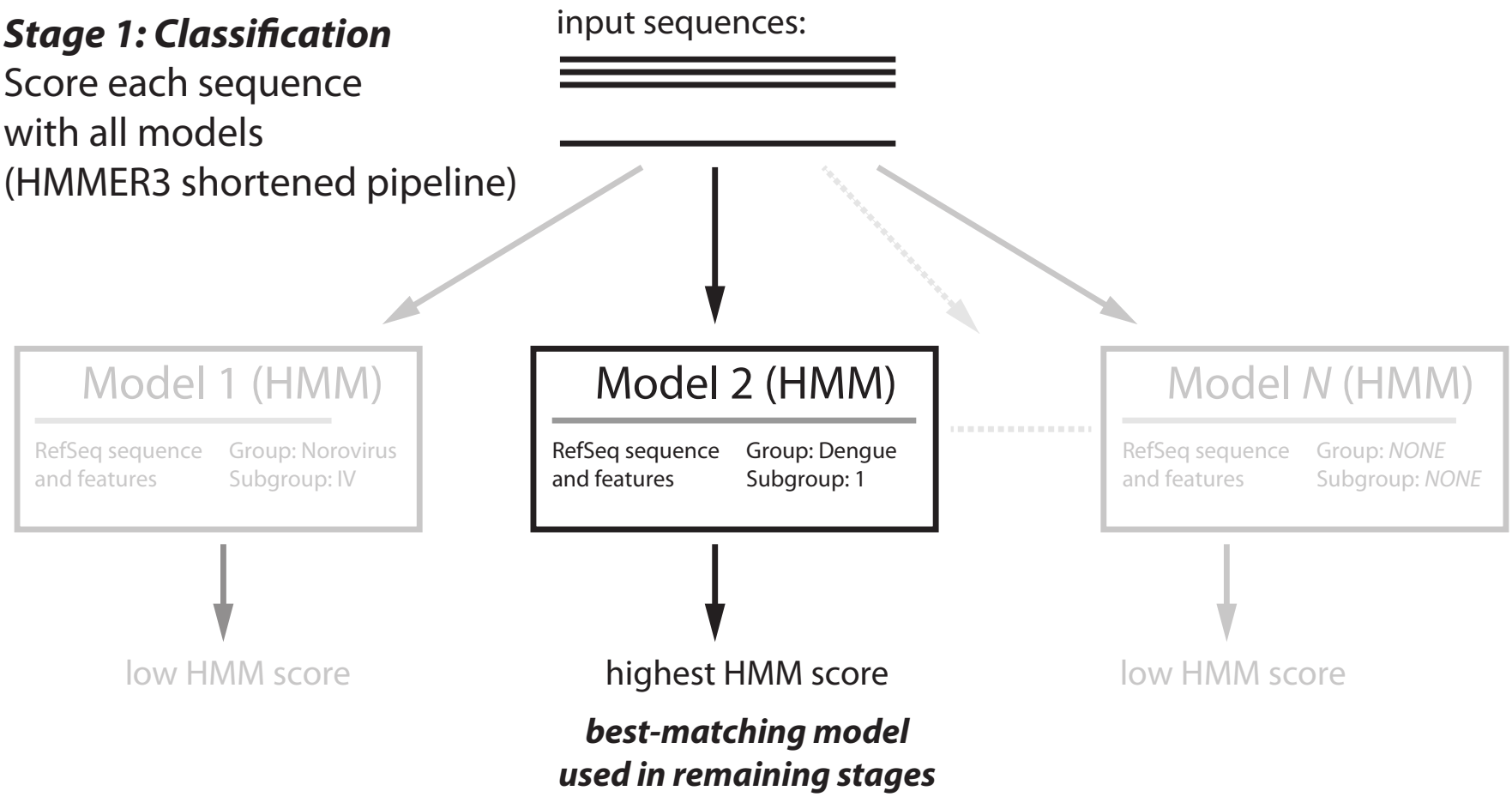
Different types of alerts are identified and reported at each stage

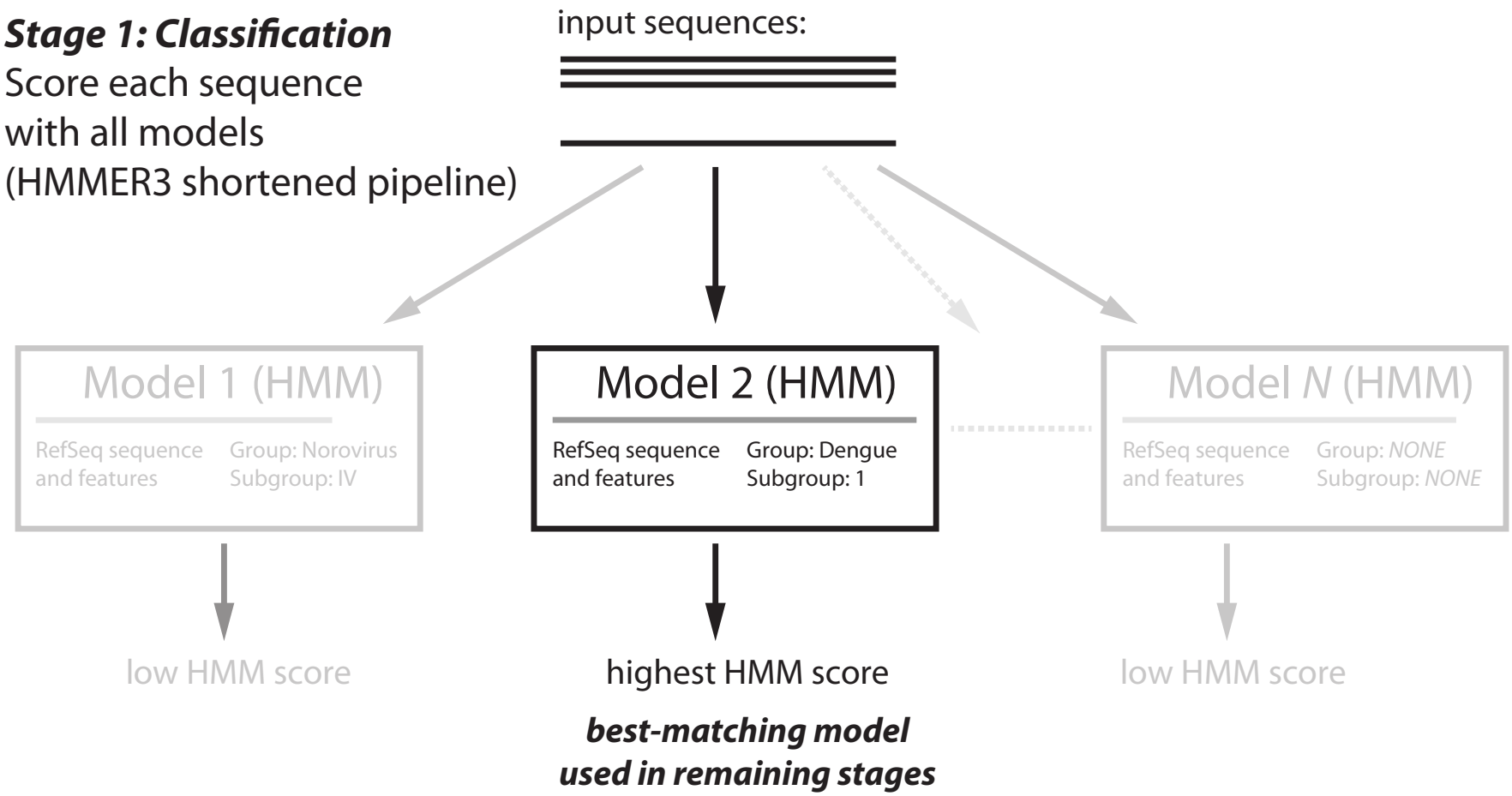
Stage 1: Classification

Score each sequence
with all models
(HMMER3 shortened pipeline)



Stage 1: Classification
Score each sequence
with all models
(HMMER3 shortened pipeline)

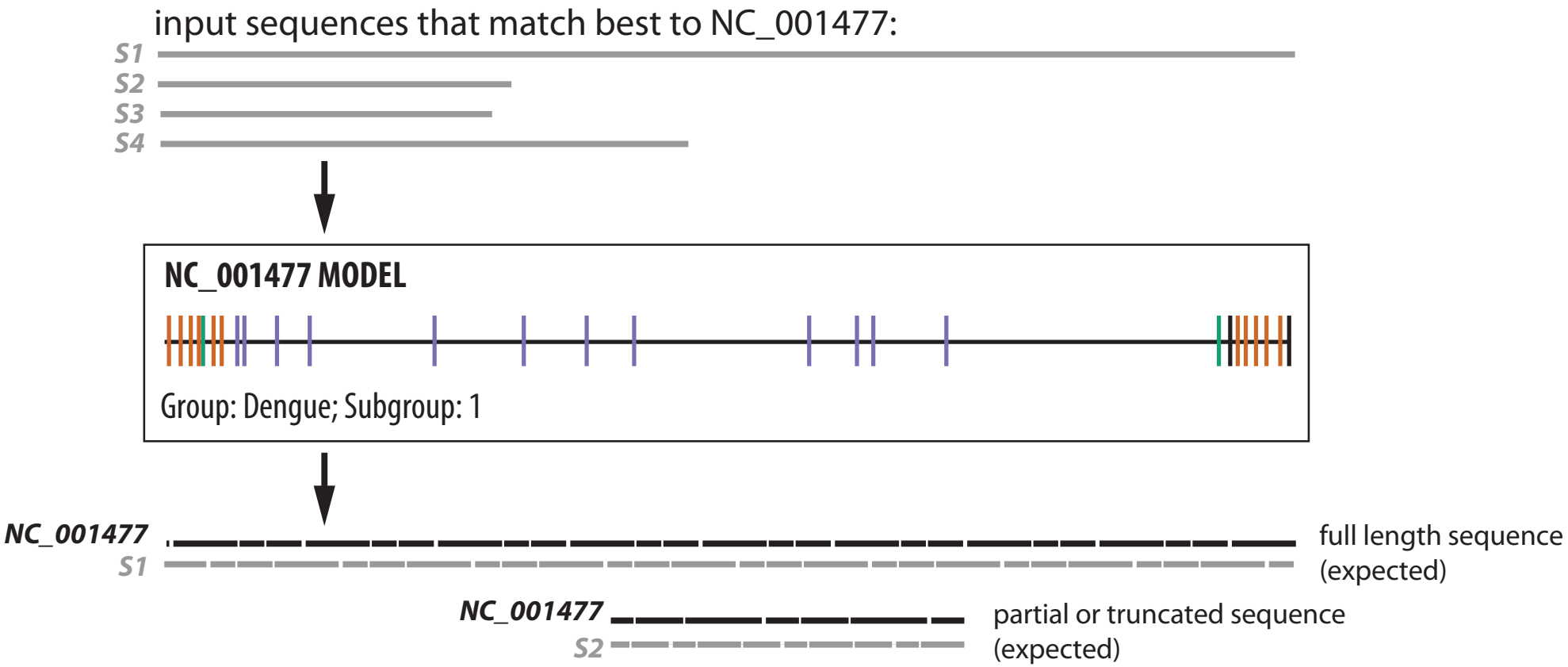




code	S/F	error message	description
Fatal alerts detected in the classification stage			
noannotn*	S	NO_ANNOTATION	no significant similarity detected
revcompl*	S	REVCOMPLEM	sequence appears to be reverse complemented
incsbgrp	S	INCORRECT_SPECIFIED_SUBGROUP	score difference too large between best overall model and best specified subgroup model
incgroup	S	INCORRECT_SPECIFIED_GROUP	score difference too large between best overall model and best specified group model
Non-fatal alerts detected in the classification stage			
qstsbgrp	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

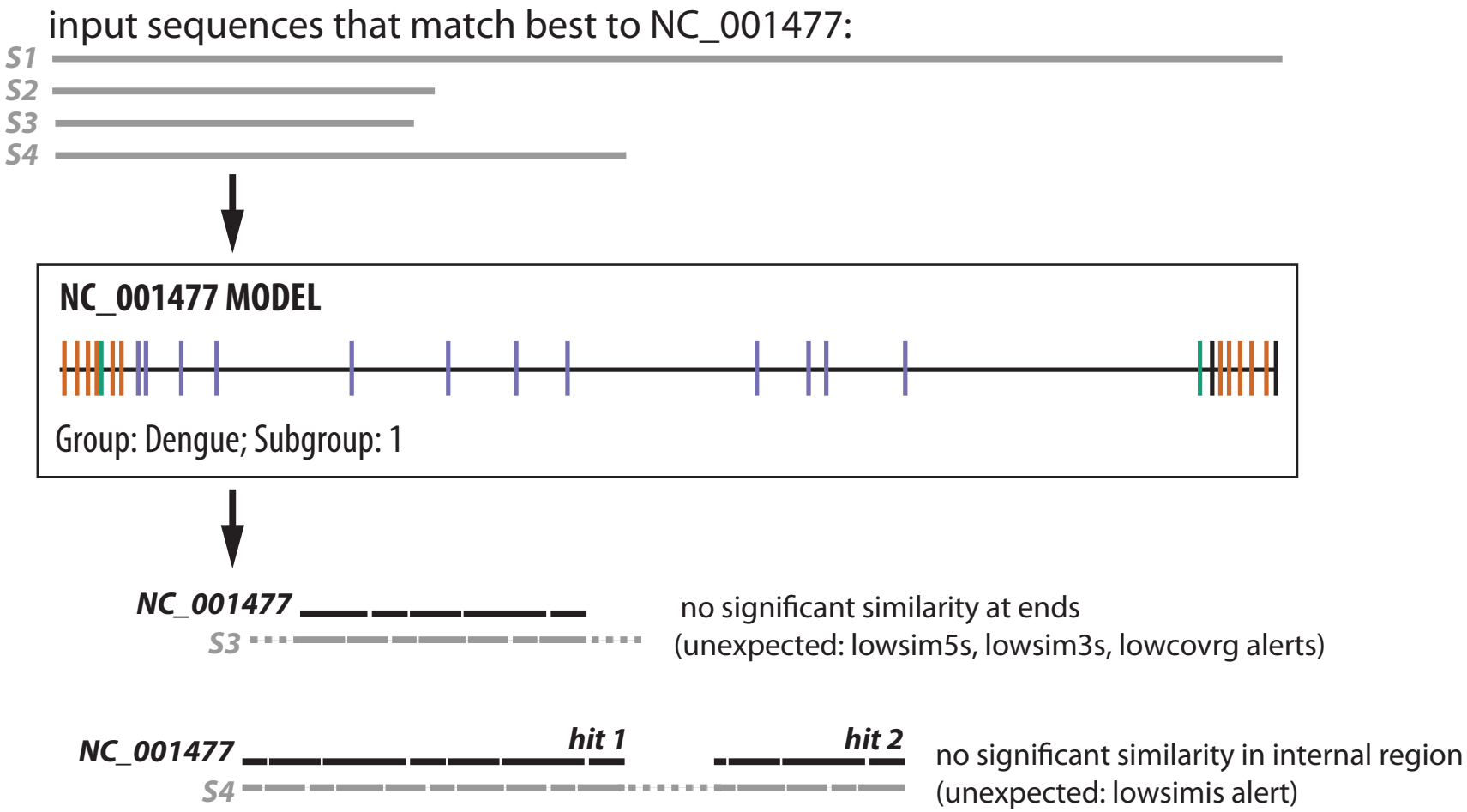
Stage 2: Coverage determination

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



Stage 2: Coverage determination

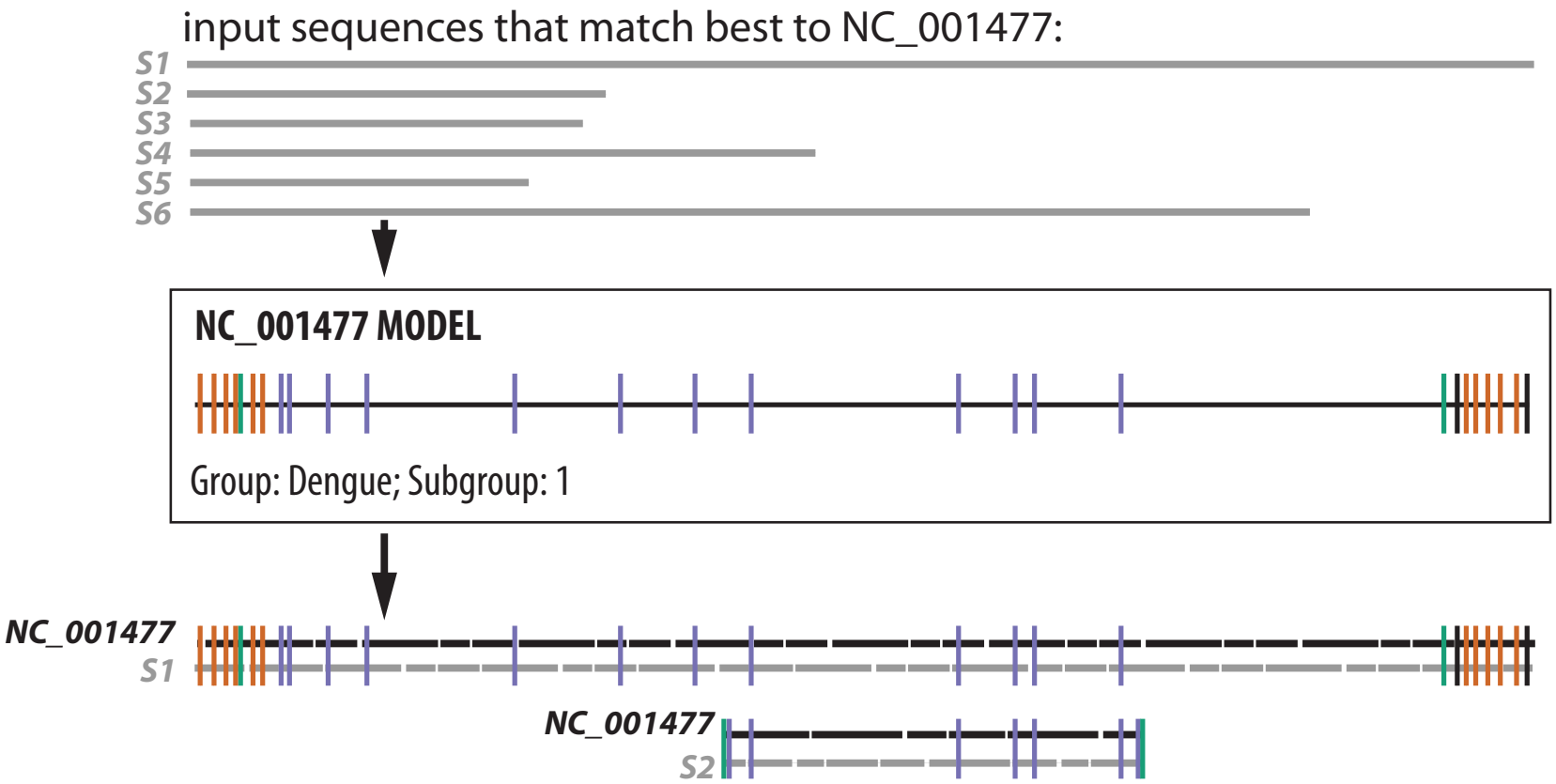
Search each sequence with best-matching model (HMMER3 full pipeline)



code	S/F	error message	description
Fatal alerts detected in the coverage stage			
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
Non-fatal alerts detected in the coverage stage			
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)



Stage 3: Alignment and feature mapping

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

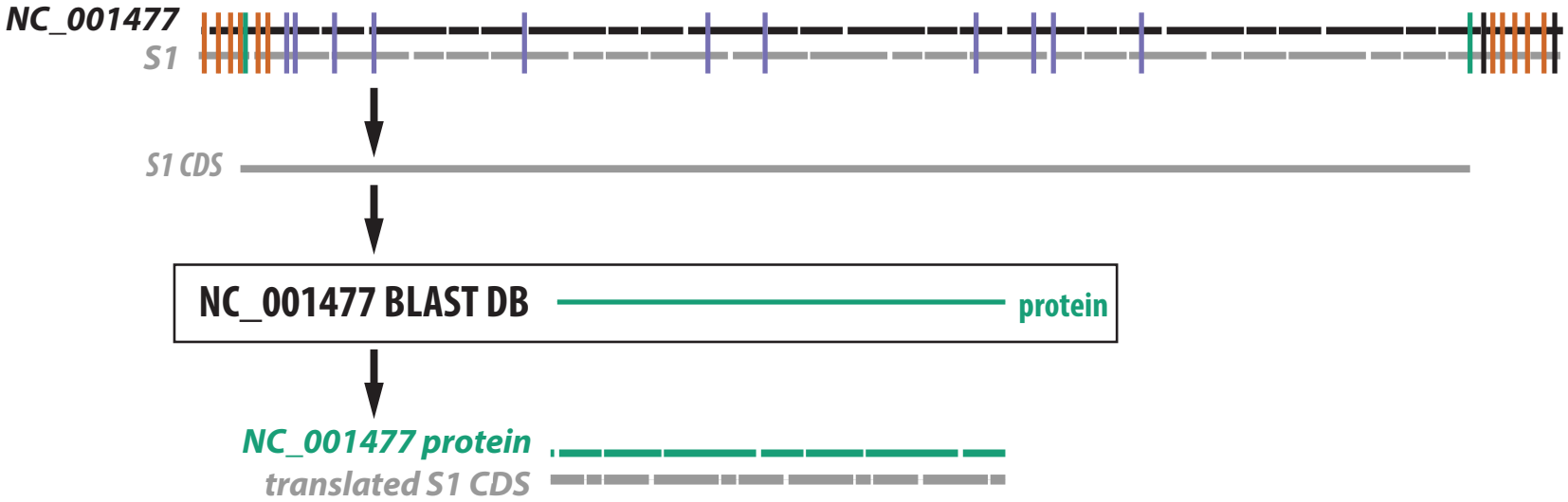
gap or low confidence at feature boundary
(indf5gap, indf3gap, indf5loc, indf3loc alert)

early stop codon in CDS
(cdsstopn alert)

code	S/F	error message	description
Fatal alerts detected in the annotation stage			
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
cdsstopn	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

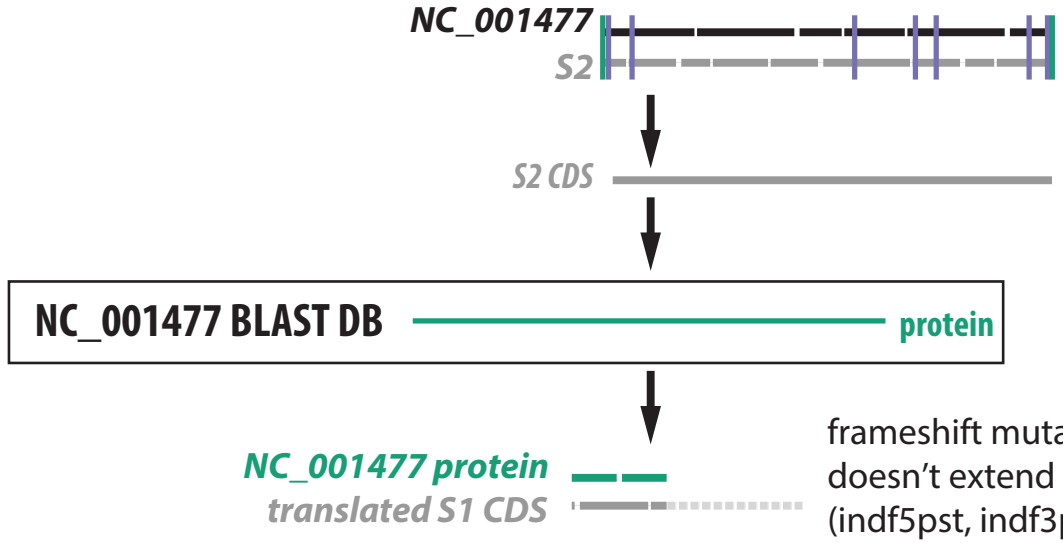
Stage 4: Protein validation (Alejandro Schäffer)

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



frameshift mutation; BLASTX alignment doesn't extend full length of CDS prediction (indf5pst, indf3pst alerts)

code	S/F	error message	description
Fatal alerts detected in the protein validation stage			
cdsstopp	F	CDS_HAS_STOP_CODON	stop codon in protein-based alignment
indfantp	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 >= 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

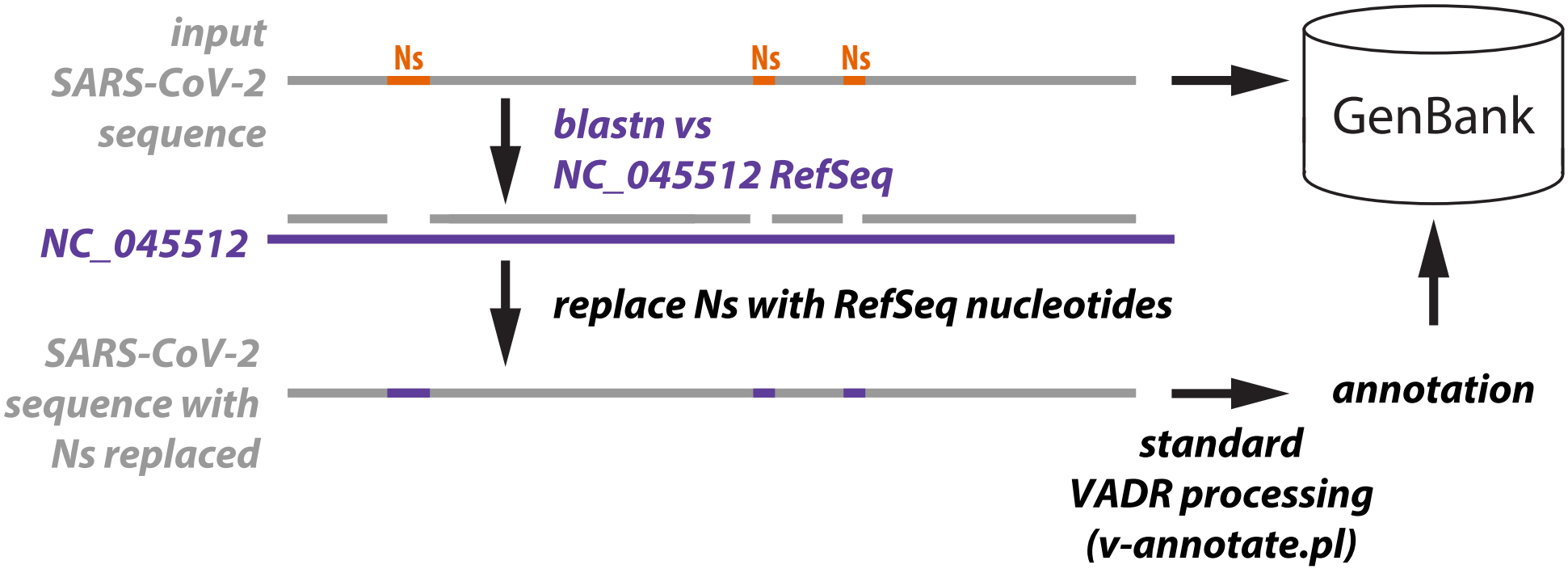
SARS-CoV-2 sequence submissions have increased since early 2020

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

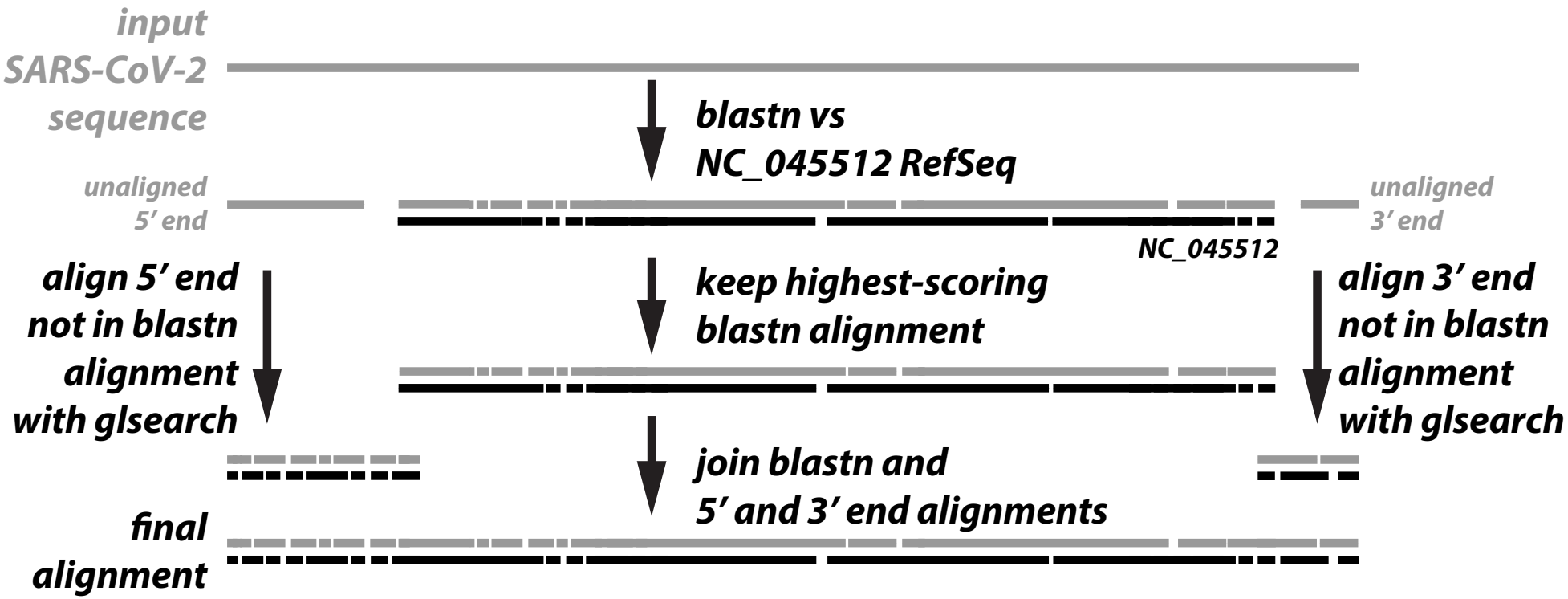
SARS-CoV-2 sequences differ from Norovirus and Dengue virus in several ways that impact VADR processing

	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 >= 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

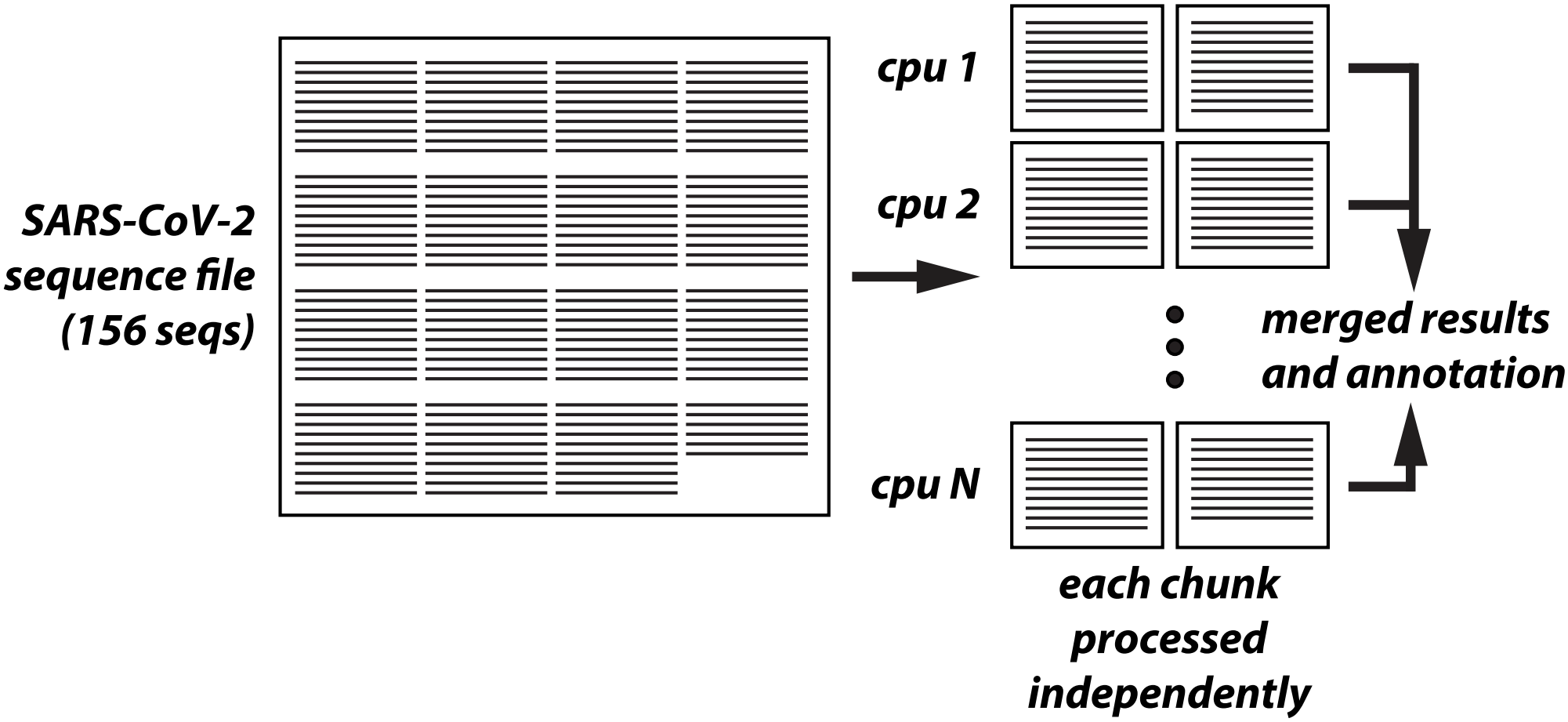


Seeded alignment using blastn makes alignment stage faster



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 align-ment?	N replace-ment?	glsearch?	# cpus	required RAM	secs per seq	hours per 100K seqs	speedup vs v1.0
v1.0	–	–	–	1	64 Gb	329.91	9164.3	-

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

VADR version	seeded align-ment?	N replace-ment?	glsearch?	# cpus	required RAM	secs per seq	hours per 100K seqs	speedup vs v1.0
v1.0	–	–	–	1	64 Gb	329.91	9164.3	-
v1.4.1	+	+	+	1	2 Gb	2.51	69.8	131.4

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

VADR version	seeded align-ment?	N replace-ment?	glsearch?	# cpus	required RAM	secs per seq	hours per 100K seqs	speedup vs v1.0
v1.0	–	–	–	1	64 Gb	329.91	9164.3	-
v1.4.1	+	+	+	1	2 Gb	2.51	69.8	131.4
v1.4.1	+	+	+	**8**	**16 Gb**	**0.33**	**9.3**	**986.8**
v1.4.1	+	+	+	32	64 Gb	0.13	3.7	2462.2

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)

- 14 releases since March 2020
- 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)

Published online 20 January 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 1 **1**
https://doi.org/10.1093/nargab/lqad002

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

- Respiratory Syncitial Virus (RSV) models
- Mpox virus
- Influenza models
  - compare with existing GenBank flu tool FLAN

Additional VADR models and development

- Respiratory Syncitial Virus (RSV) models
- Mpox virus
- Influenza models
  - compare with existing GenBank flu tool FLAN

Additional VADR models and development

- Respiratory Syncitial Virus (RSV) models
- Mpox virus
- Influenza models
- Zika models (in progress, EB Dickinson (postbac))
  - matches or exceeds performance of existing GenBank flu tool FLAN
  - able to annotate highly pathogenic avian influenza (H5N1)

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

- VADR is general, standalone and includes a module for users to build new models

- 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 collaborates with me building VADR models:
    * Herpes Simplex Virus (HSV-1 and HSV-2) models
    * Human metapneumovirus (HMPV) models

VADR and NCBI virus

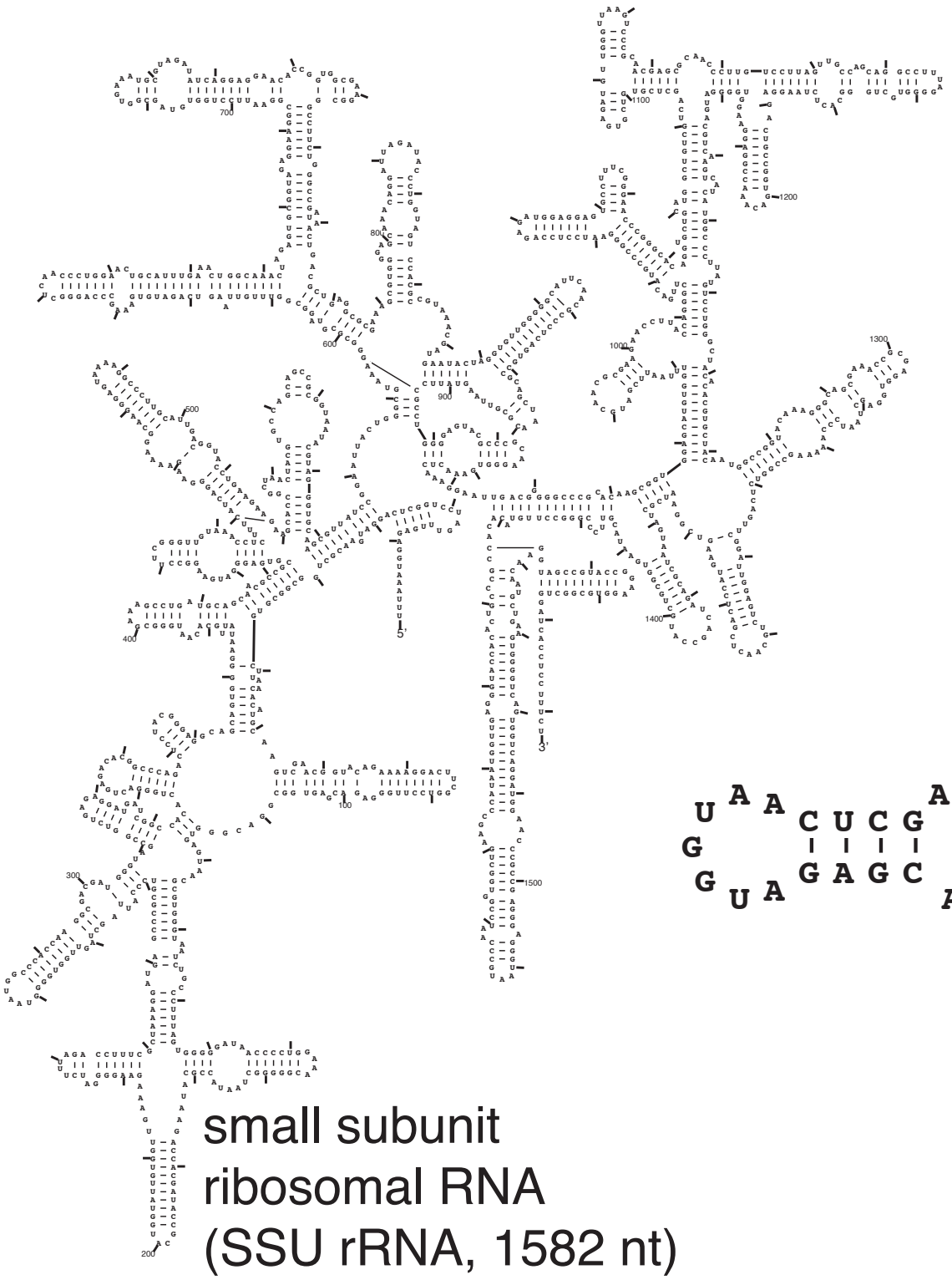
- FY2025 NCBI virus goals related to VADR:
  - VADR web server
  - Replacement of FLAN with VADR

Future directions

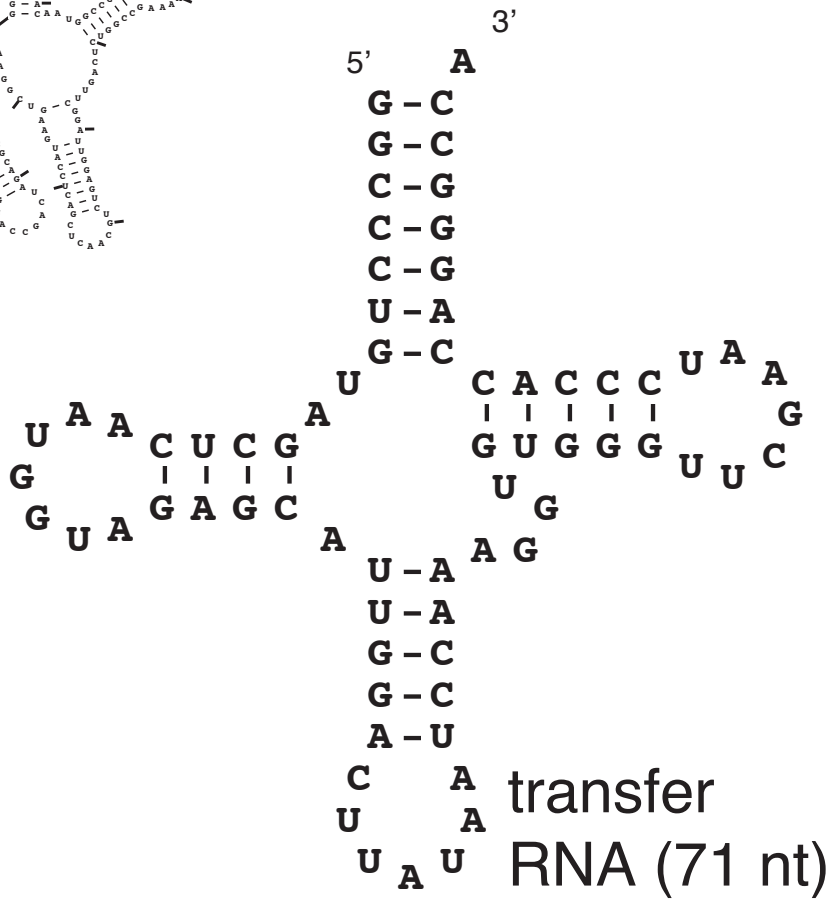
- More models (our group and Greninger lab)
- Alignment-based models
- RNA structure annotation in Flaviviruses (already exists for Dengue)

Structural RNA sequence analysis

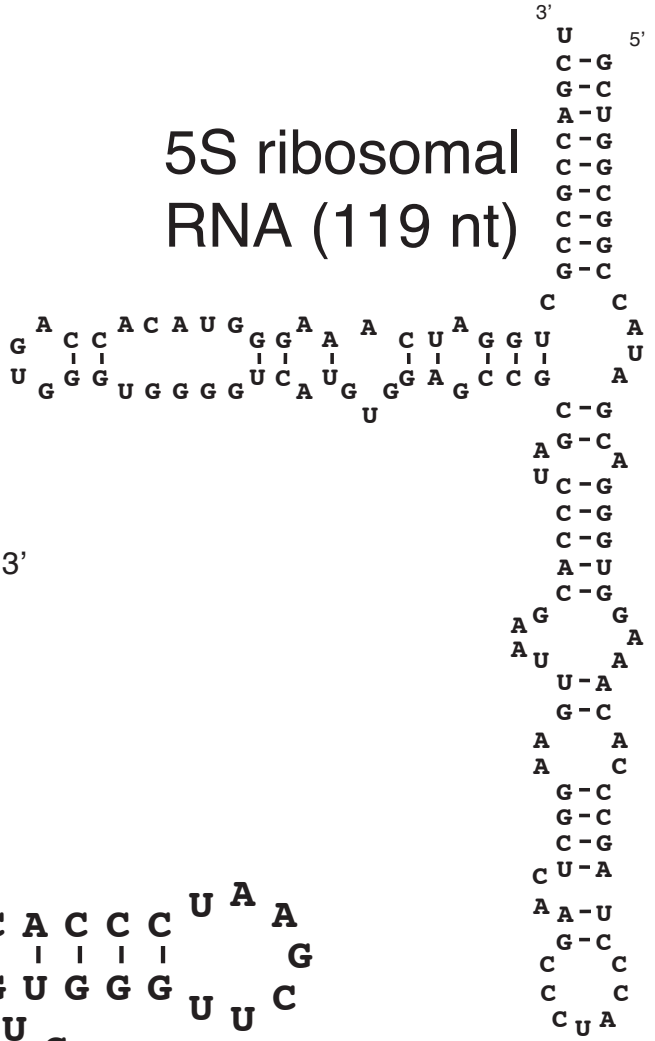
- intro to Infernal
- intro to Rfam
- intro to RNAcentral?
- intro to R2DT?
- RNA annotation: PGAP, Euk genomes (?), Hydractinia genomes
- tmRNA DB



small subunit ribosomal RNA (SSU rRNA, 1582 nt)



transfer RNA (71 nt)



5S ribosomal RNA (119 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...				

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...				

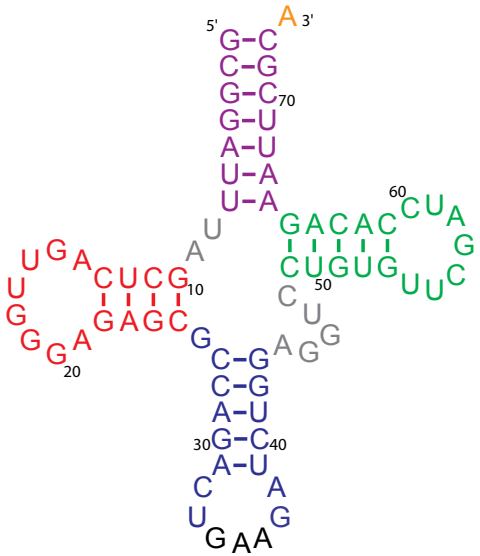
database of more than 2700 non-coding RNA families
each represented by a secondary structure, alignment, and covariance model.

Many functional RNAs adopt a conserved 3-dimensional structure

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure

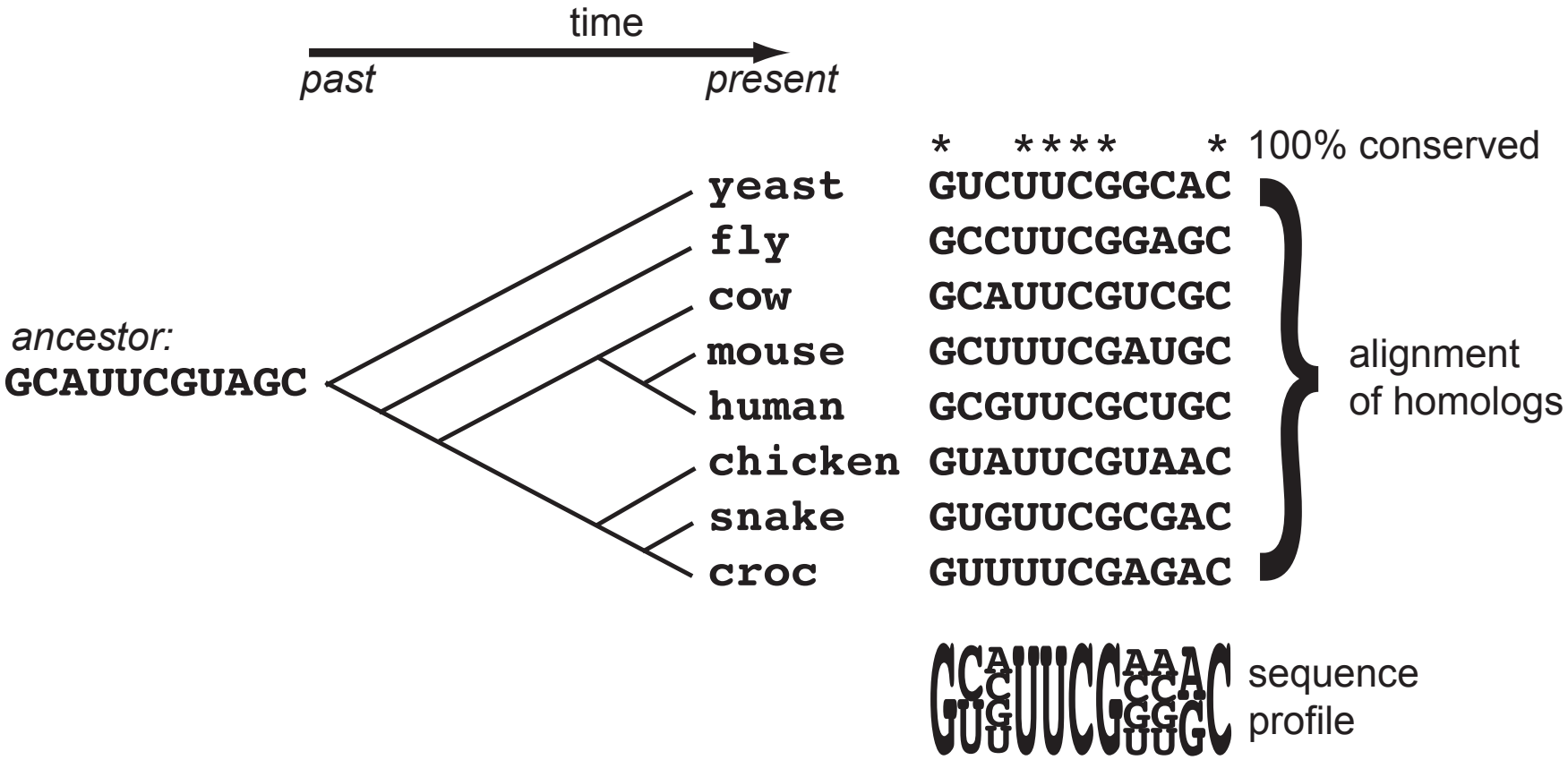


3-dimensional structure

- BLAST: given a single sequence, search genomes for similar sequences.
- Structural RNAs are difficult to find
  - short ($\sim$ 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

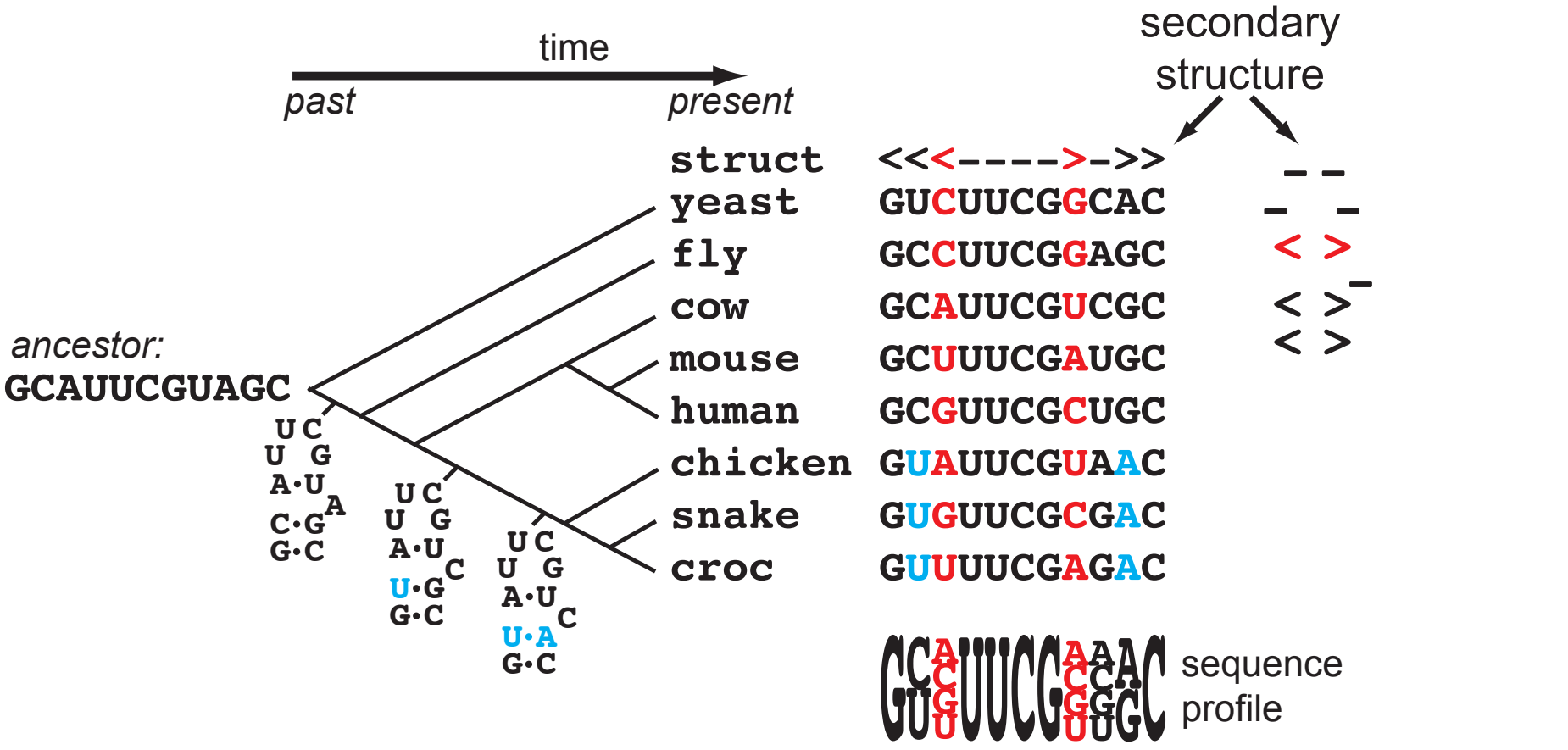
Sequence conservation provides information for homology searches

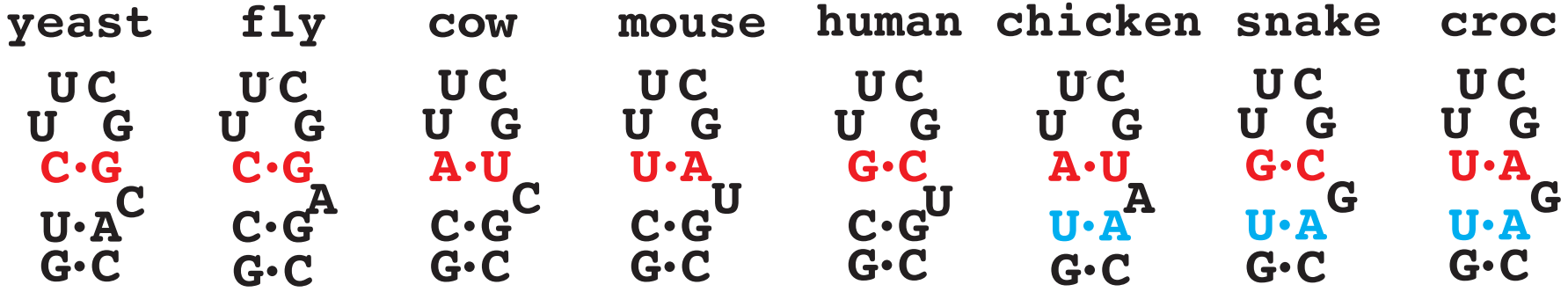
Conservation levels vary across alignment columns.



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 (17929 and 4150 entries)	Rfam (2791 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.

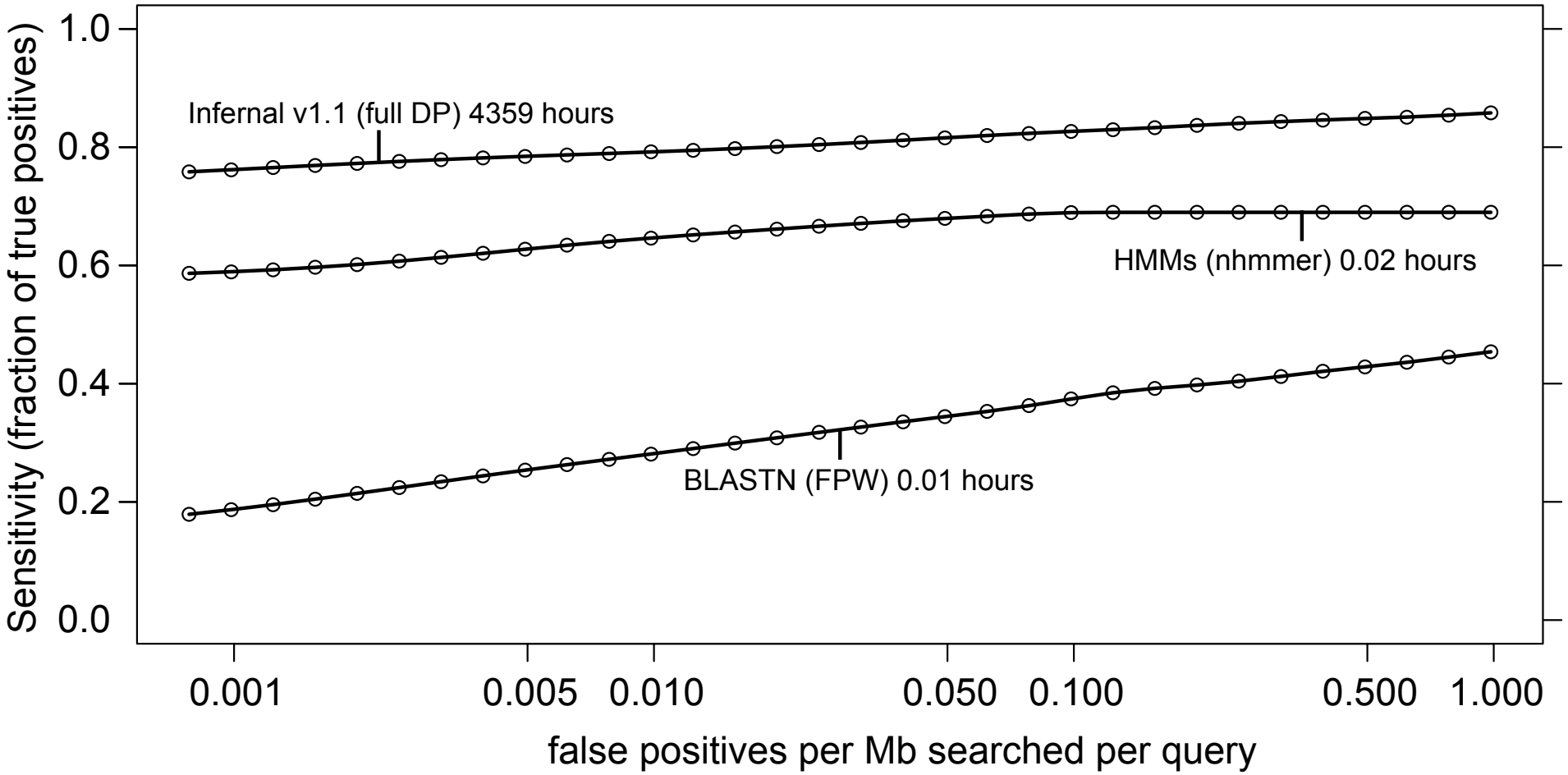
Eddy. SR. Bioinformatics, 14:755-763, 1998.

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)

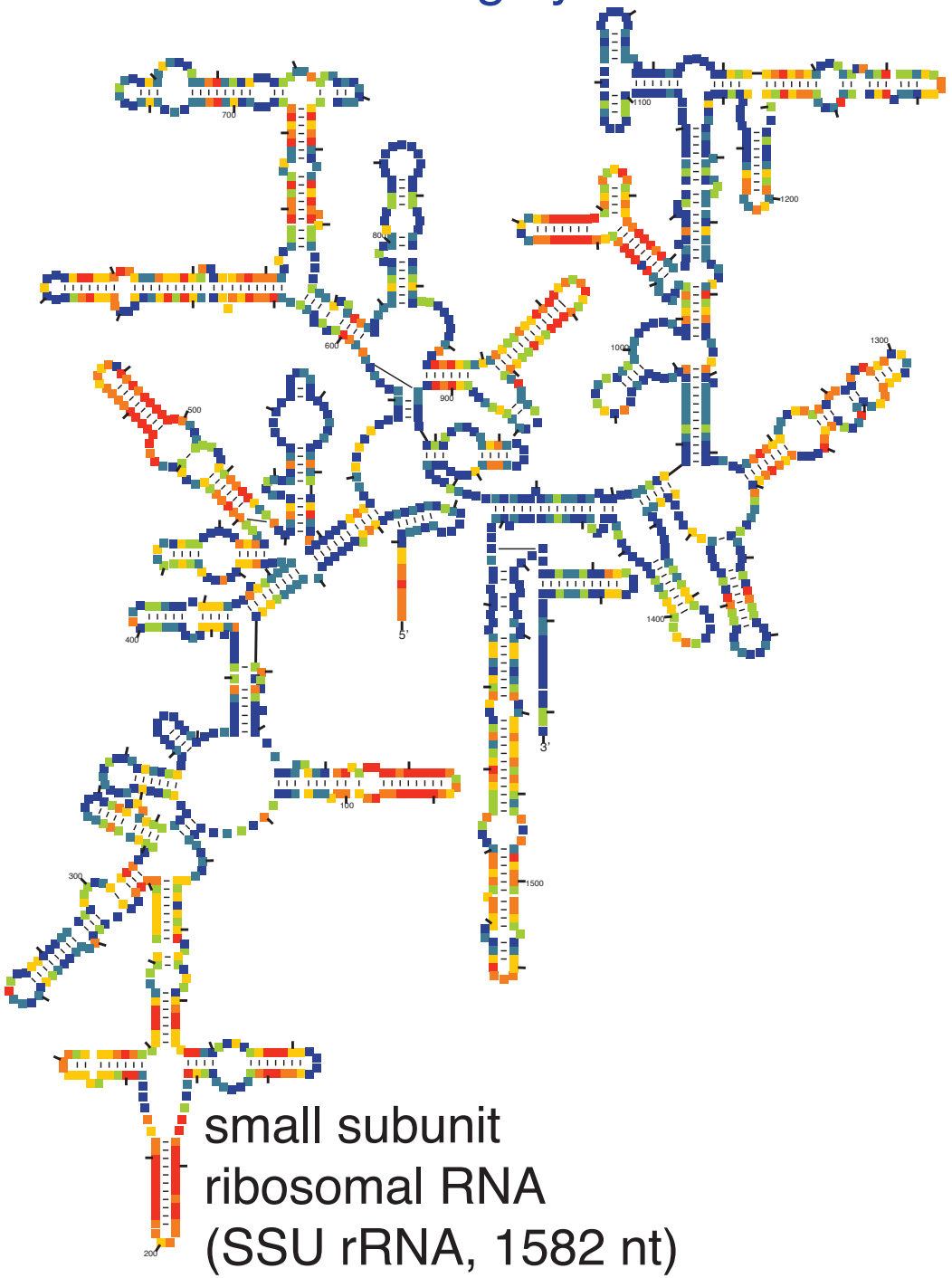


Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

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

Sequence conservation per position

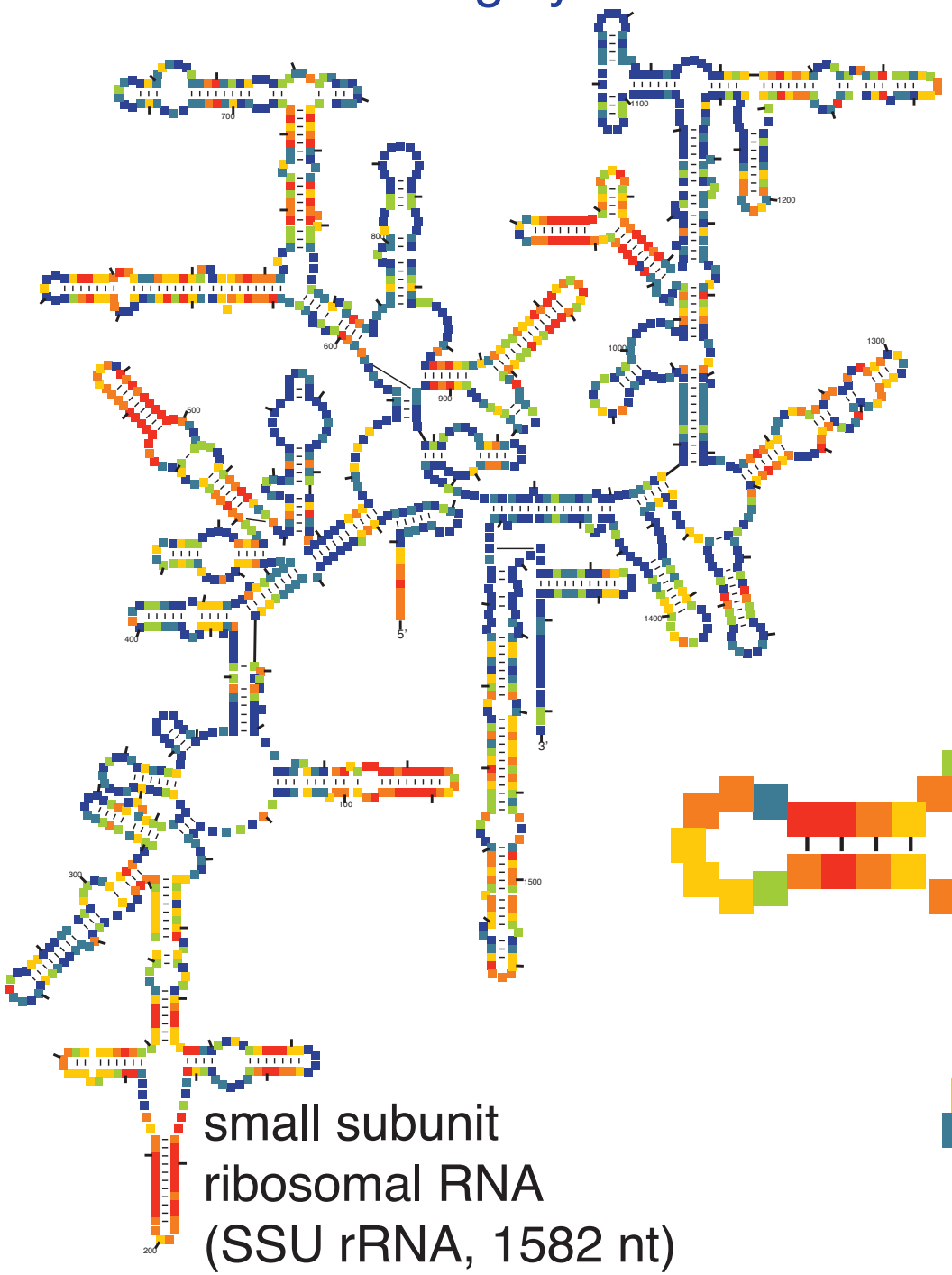
blue:highly conserved ..... red: highly variable



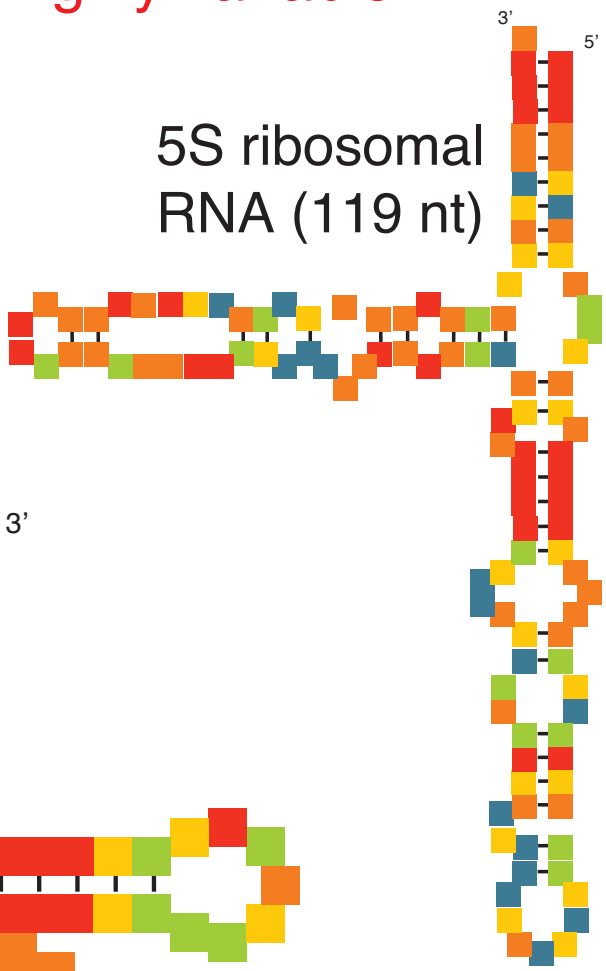
small subunit
ribosomal RNA
(SSU rRNA, 1582 nt)

Sequence conservation per position

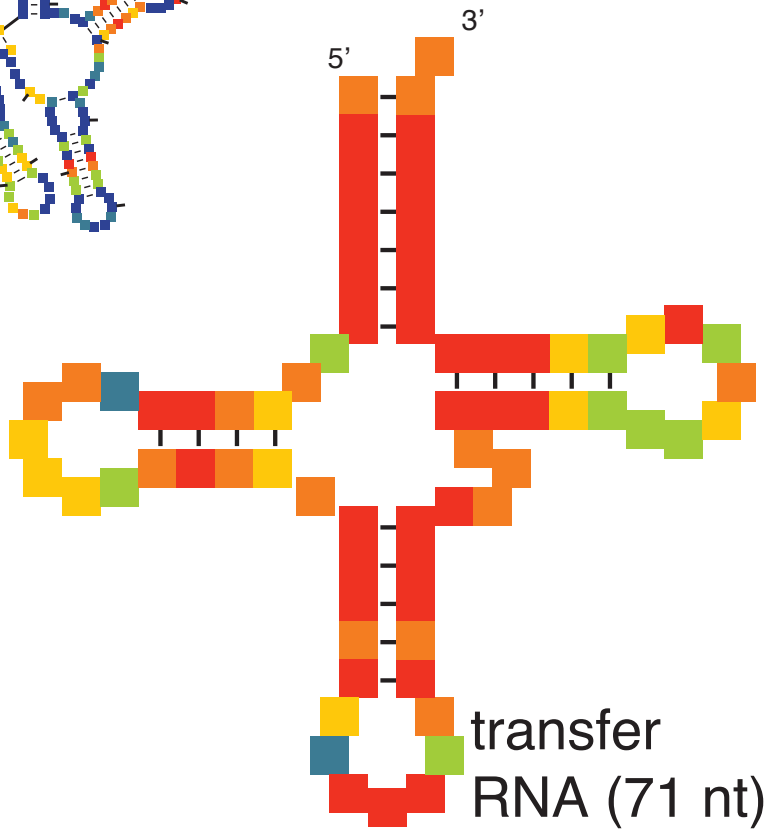
blue:highly conserved ..... red: highly variable



small subunit ribosomal RNA (SSU rRNA, 1582 nt)

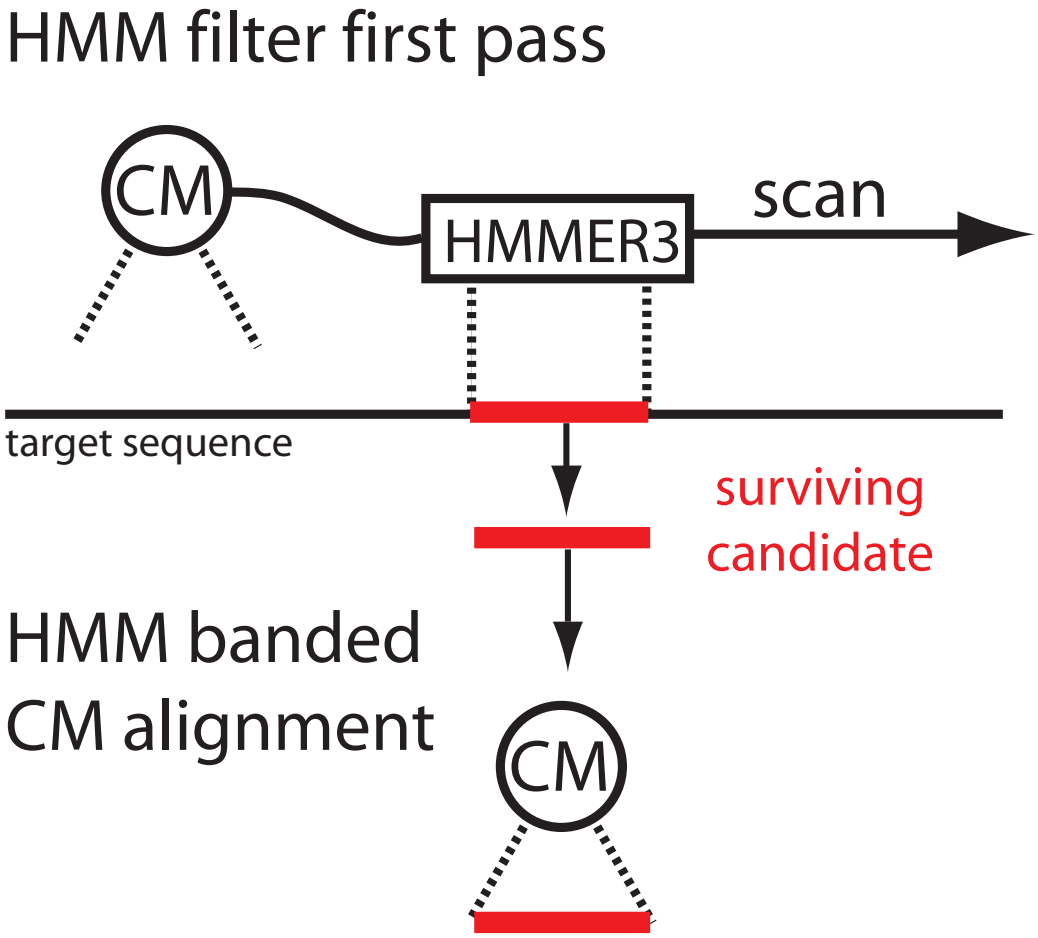


5S ribosomal RNA (119 nt)

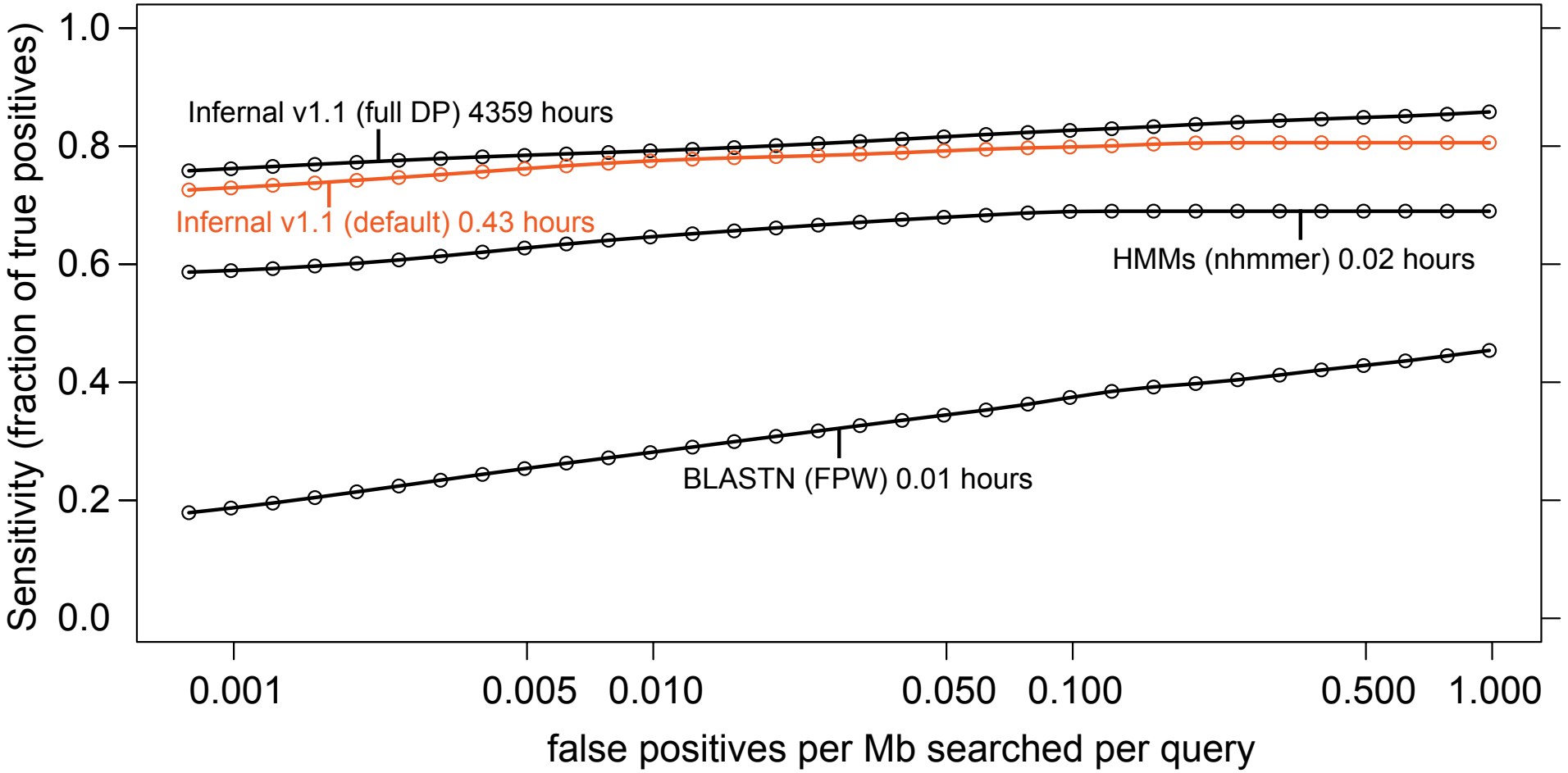


transfer RNA (71 nt)

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.

Ribovore? Rfam? not sure where to go from here