Reference-based viral sequence validation and annotation using VADR

Eric Nawrocki

National Center for Biotechnology Information
National Institutes of Health

INSDC (GenBank/ENA/DDBJ) has a lot of sequence data

D92–D96 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 16 November 2020*
doi: 10.1093/nar/gkaa1023

GenBank

Eric W. Sayers*, Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Conrad L. Schoch, Stephen T. Sherry and Ilene Karsch-Mizrachi

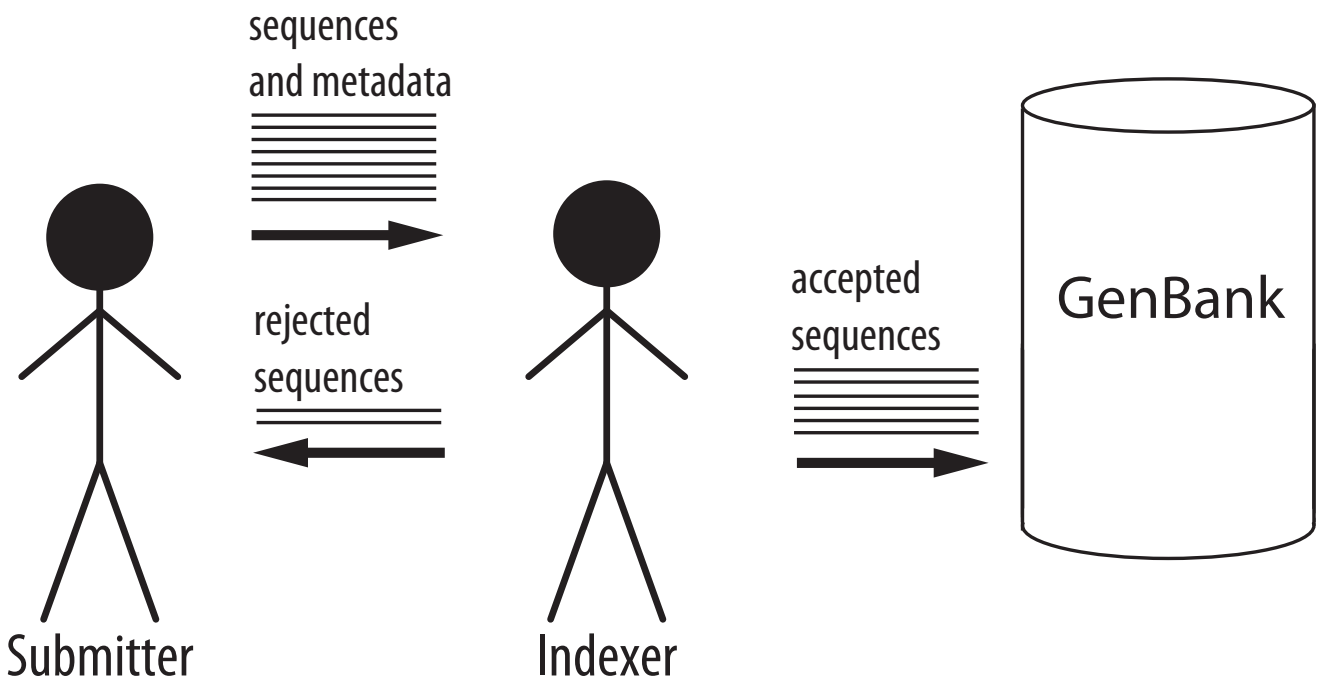
Table 1. GenBank divisions

Division	Description	Base pairs[a]
WGS	Whole genome shotgun data	8 841 649 410 652
TSA	Transcriptome shotgun data	381 148 464 834
PLN	Plants	269 438 877 546
BCT	Bacteria	98 827 135 660
VRT	Other vertebrates	63 565 835 430
EST	Expressed sequence tags	43 301 109 577
TLS	Targeted Loci Studies	27 825 059 498
HTG	High-throughput genomic	27 781 778 663
PAT	Patent sequences	26 452 787 091
GSS	Genome survey sequences	26 378 695 300
MAM	Other mammals	20 844 388 122
INV	Invertebrates	19 759 935 222
ROD	Rodents	12 090 011 771
PRI	Primates	8 767 435 622
SYN	Synthetic	7 932 542 985
ENV	Environmental samples	6 755 612 180
VRL	Viruses	5 824 026 918
PHG	Phages	782 571 323
HTC	High-throughput cDNA	733 210 026
STS	Sequence tagged sites	640 923 137
UNA	Unannotated	679 302
TOTAL	All GenBank sequences	9 890 500 490 859

[a]Release 239 (8/2020).

Sequence submissions are handled by expert NCBI indexers

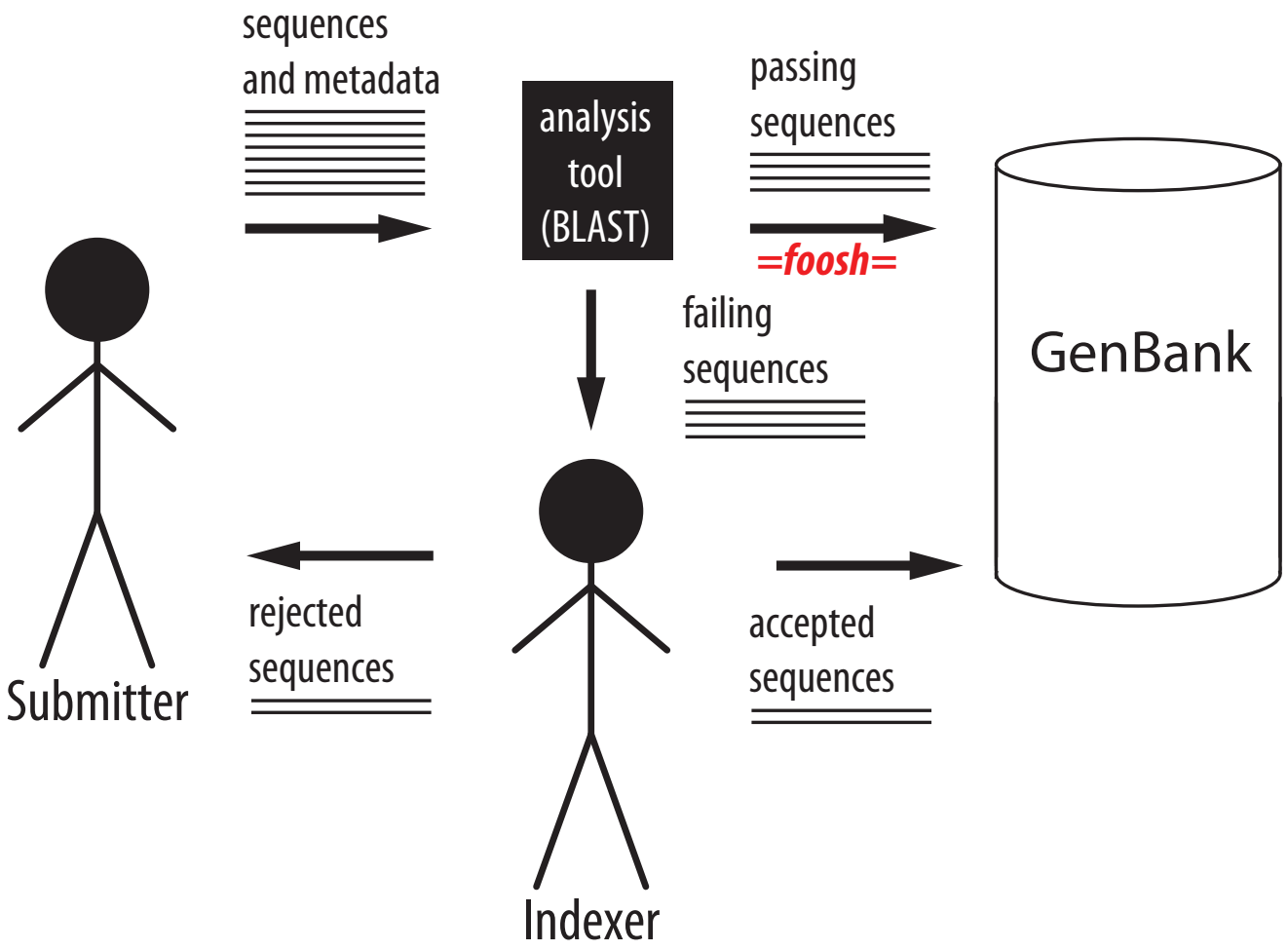
but manual indexing does not scale



Many submitted sequences are marker genes or viruses

marker gene/ virus	2021 # seqs	total # seqs
SARS-CoV-2	3,026,073	6,052,165
16S rRNA	258,194	10,294,372
23S rRNA	59,191	1,236,112
COX1	86,248	541,630
HIV-1	44,359	1,035,342
Influenza	36,037	833,540
ITS2	26,630	260,245
ITS1	16,002	427,675
ITS1+ITS2	13,326	513,077

NCBI GenBank Indexers use BLAST



- Foosh pipelines exist for 16S, 23S, ITS (BLAST-based) and Influenza (FLAN)
- False negatives are better than false positives because indexer or submitter can manually examine them

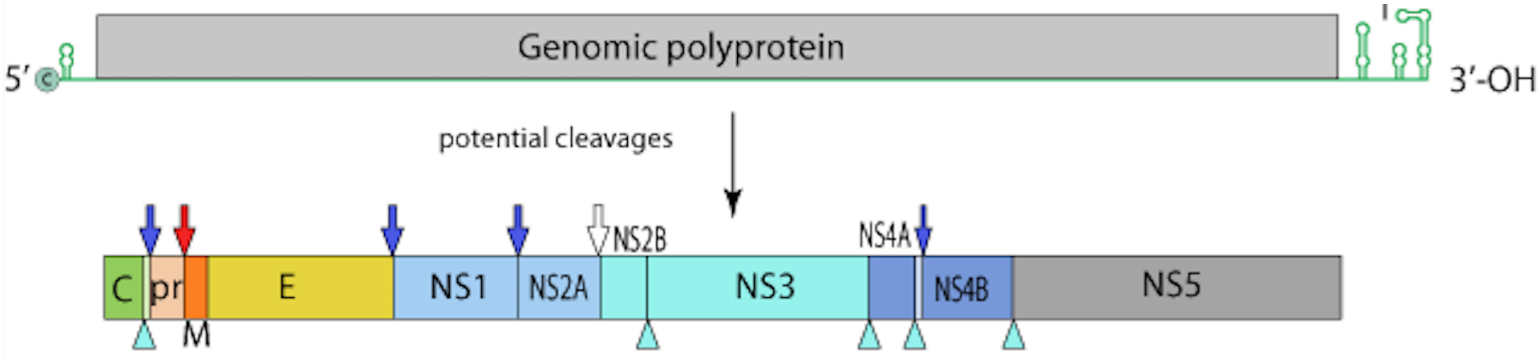
Viruses with highest number of sequences in INSDC*

species	#seqs	family
SARS-CoV-2	6,052,165	*Coronaviridae*
HIV-1	1,035,342	*Retroviridae*
Influenza A virus	833,505	*Orthomyxoviridae*
Hepacivirus C	259,870	*Flaviviridae*
Hepatitis B virus	124,490	*Hepadnaviridae*
Influenza B virus	118,799	*Orthomyxoviridae*
Rotavirus A	96,690	*Reoviridae*
Norovirus (Norwalk virus)	51,748	*Caliciviridae*
SIV	50,454	*Retroviridae*
West Nile virus	49,579	*Flaviviridae*
Dengue virus	39,830	*Flaviviridae*
Enterovirus A	39,527	*Picornaviridae*
PRRSV	38,538	*Arteriviridae*
Human orthopneumovirus	32,835	*Pneumoviridae*
Enterovirus B	28,494	*Picornaviridae*
Lyssavirus rabies	26,798	*Rhabdoviridae*

*as of August, 2022.

Viral sequences are not systematically or thoroughly annotated

- Genome annotation of the Zika virus:

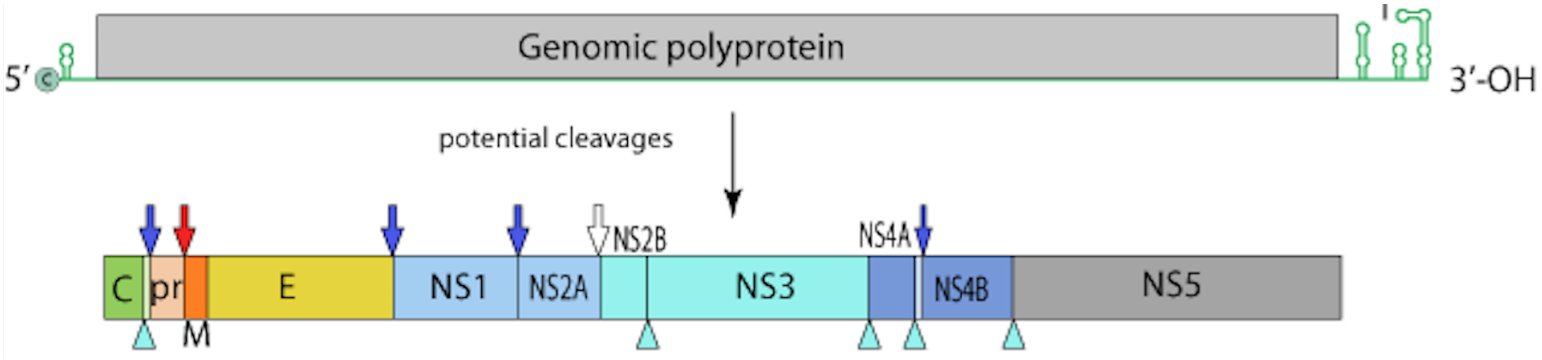


- Zika's genome encodes a single polyprotein that is cleaved into 14 mature peptides.
- Zika RefSeq annotation (NC_012532) includes CDS and mature peptide annotation.

figure from ViralZone, Hulo et al. NAR. 2011 Jan;39:D576-82

Viral sequences are not systematically or thoroughly annotated

- Genome annotation of the Zika virus:

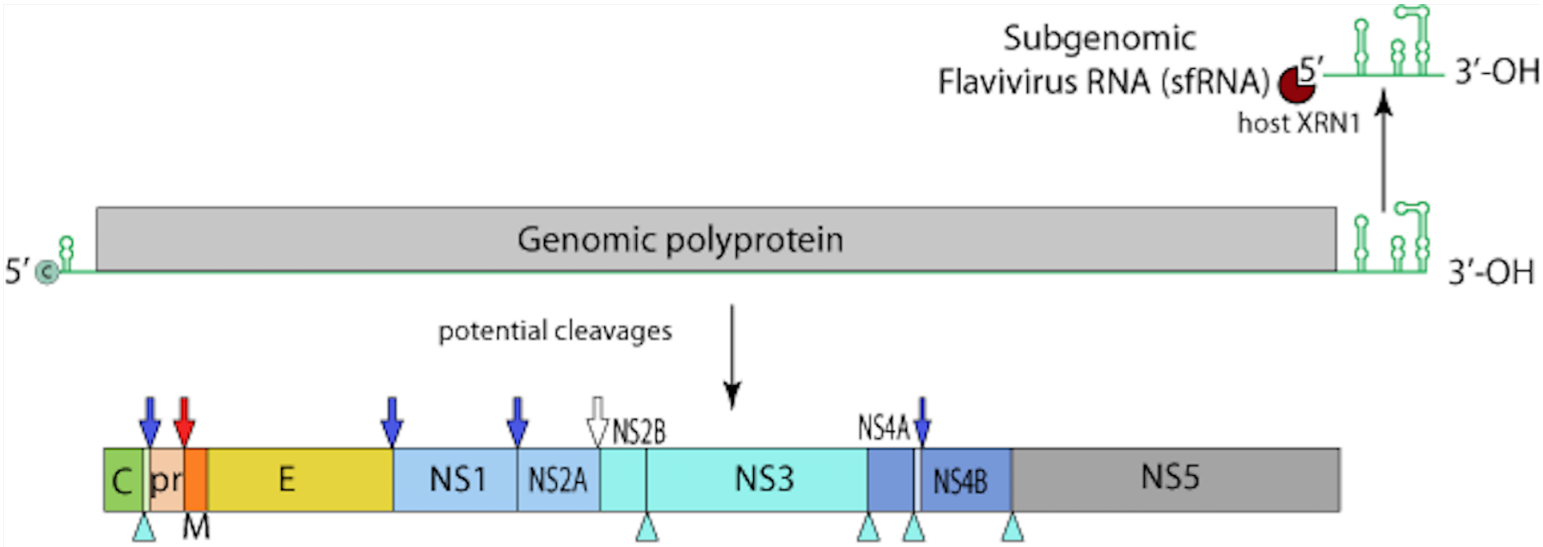


- Zika's genome encodes a single polyprotein that is cleaved into 14 mature peptides.
- Zika RefSeq annotation (NC_012532) includes CDS and mature peptide annotation.
- About 84% of Zika virus sequences have CDS annotation.
- Less than 25% of Zika virus sequences have mature peptide annotation.
- Less than 7% of Dengue virus sequences have mature peptide annotation.
- Less than 2% of Norovirus sequences have mature peptide annotation.

figure from ViralZone, Hulo et al. NAR. 2011 Jan;39:D576-82

Viral sequences are not systematically or thoroughly annotated

- Genome annotation of the Zika virus:



- RNA structures in the 3' UTR halt host exonuclease leading to an accumulation of 300-500nt subgenomic flavivirus RNAs (sfRNAs) are related to pathogenicity.

These RNA structures are not annotated in the Zika genome RefSeq (NC_012532)

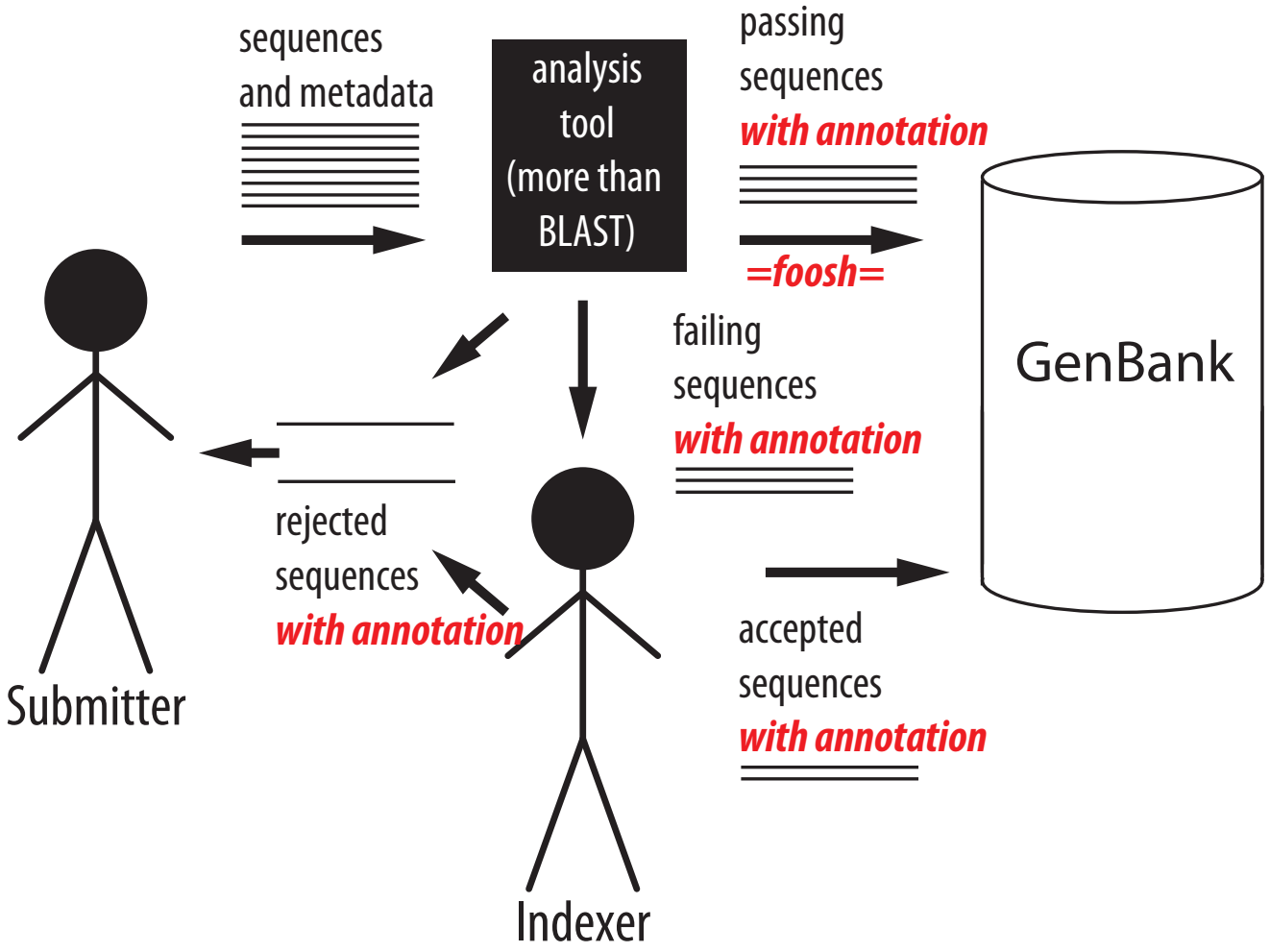
figure from ViralZone, Hulo et al. NAR. 2011 Jan;39:D576-82

Viral sequences are not systematically or thoroughly annotated

- CDS are not always annotated
- Mature peptides are rarely annotated
- Rfam families are rarely to never annotated in viral genomes (more than 200 families)

Systematic and complete annotation would benefit viral researchers (facilitate comparative analyses)

Annotation and validation should be coupled



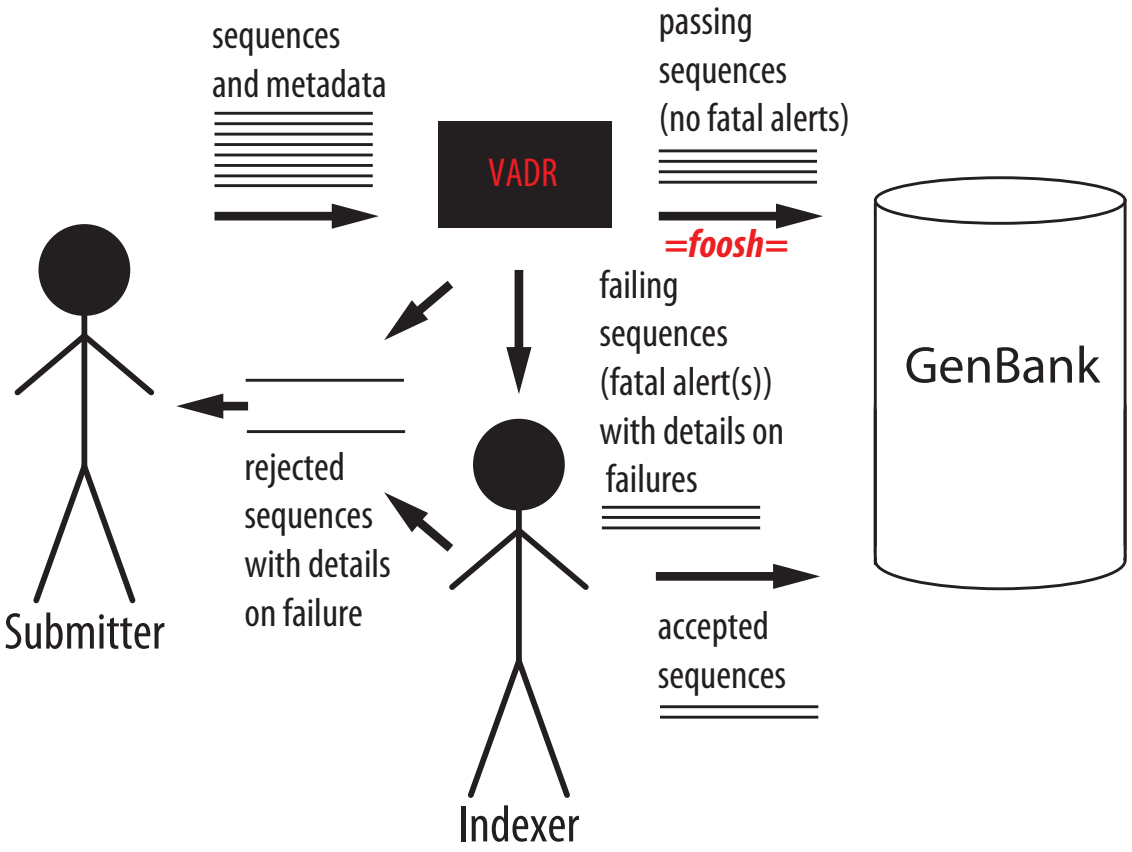
Schäffer *et al. BMC Bioinformatics* (2020) 21:211
https://doi.org/10.1186/s12859-020-3537-3

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*]

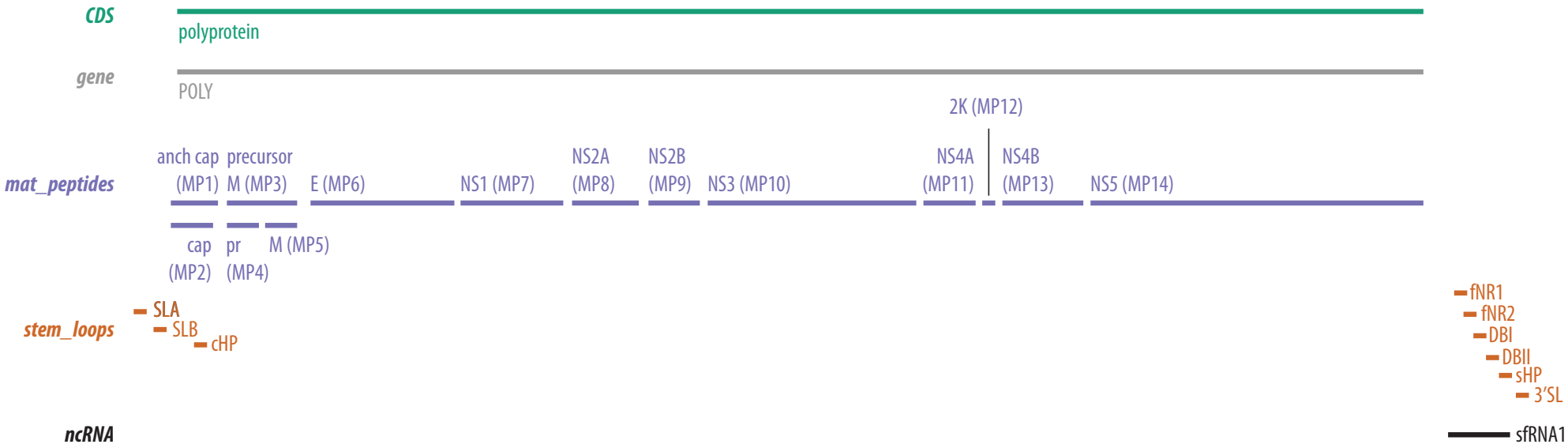


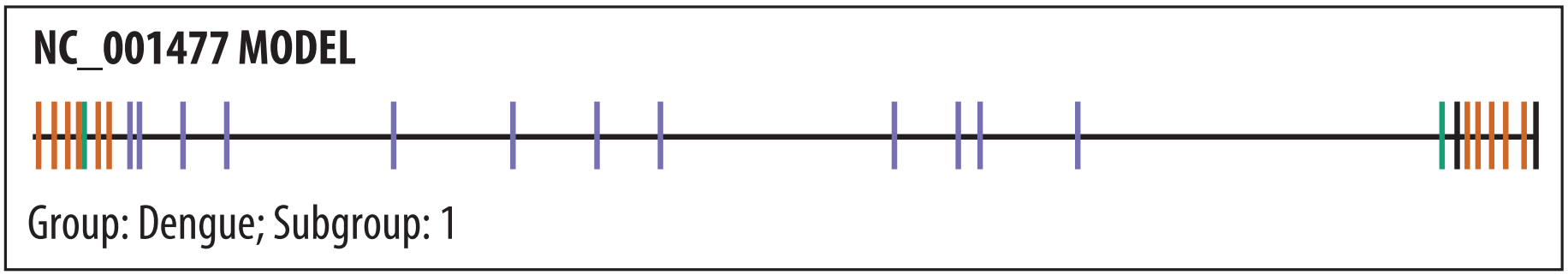
- Unexpected characteristics are reported as *alerts* (e.g. early stop codon), some of which are *fatal* and cause sequences to *fail*

Norovirus and Dengue virus chosen as first viruses for VADR testing

species	#seqs	family
SARS-CoV-2	6,045,832	*Coronaviridae*
HIV-1	1,033,995	*Retroviridae*
Influenza A virus	833,505	*Orthomyxoviridae*
Hepacivirus C	259,870	*Flaviviridae*
Hepatitis B virus	124,490	*Hepadnaviridae*
Influenza B virus	118,799	*Orthomyxoviridae*
Rotavirus A	96,690	*Reoviridae*
Norovirus (Norwalk virus)	51,748	*Caliciviridae*
SIV	50,454	*Retroviridae*
West Nile virus	49,579	*Flaviviridae*
Dengue virus	39,830	*Flaviviridae*
Enterovirus A	39,527	*Picornaviridae*
PRRSV	38,538	*Arteriviridae*
Human orthopneumovirus	32,835	*Pneumoviridae*
Enterovirus B	28,494	*Picornaviridae*
Lyssavirus rabies	26,798	*Rhabdoviridae*

VADR builds a reference model of a RefSeq and its features





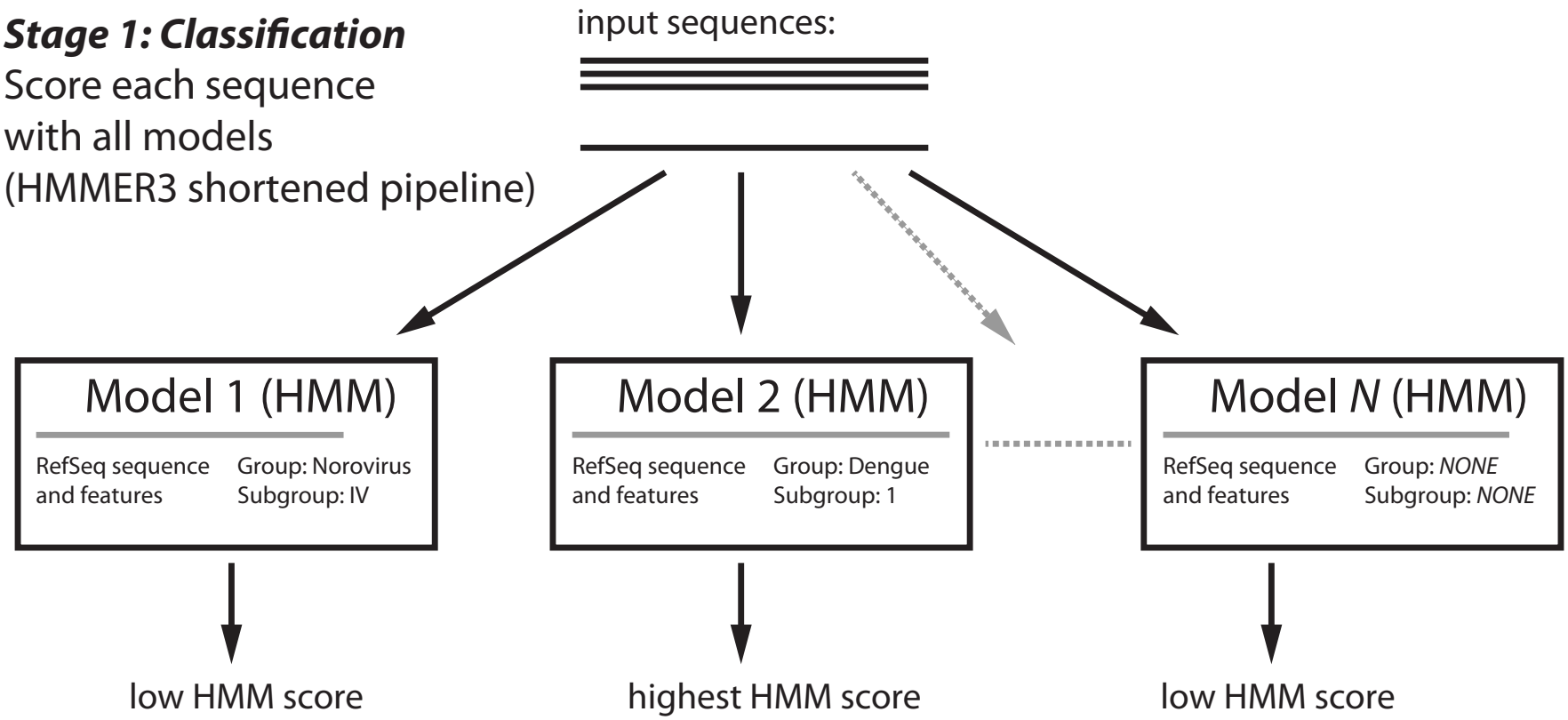
VADR validates and annotates each input sequence using its best-matching model

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

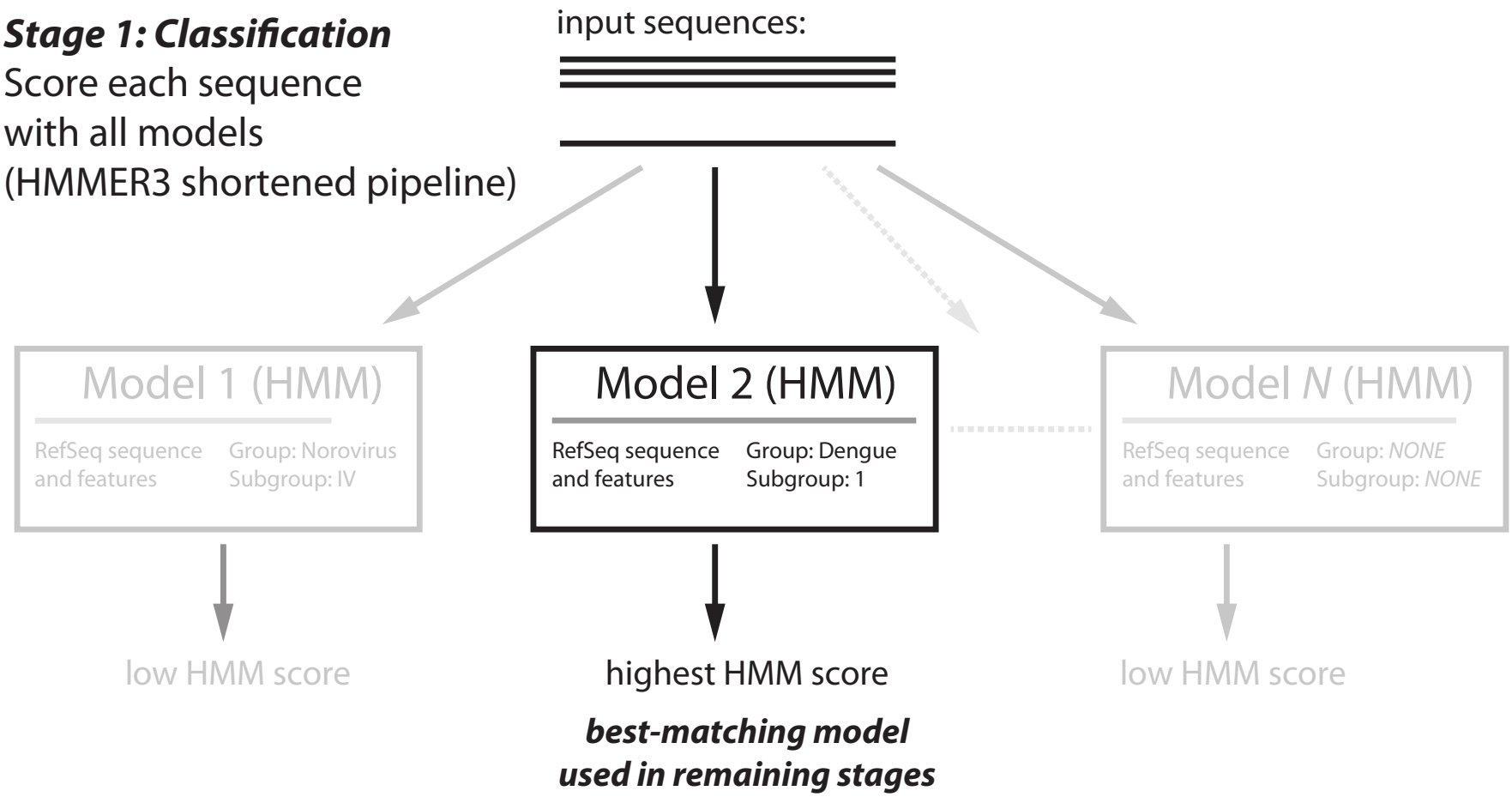
Different types of alerts are identified and reported at each stage

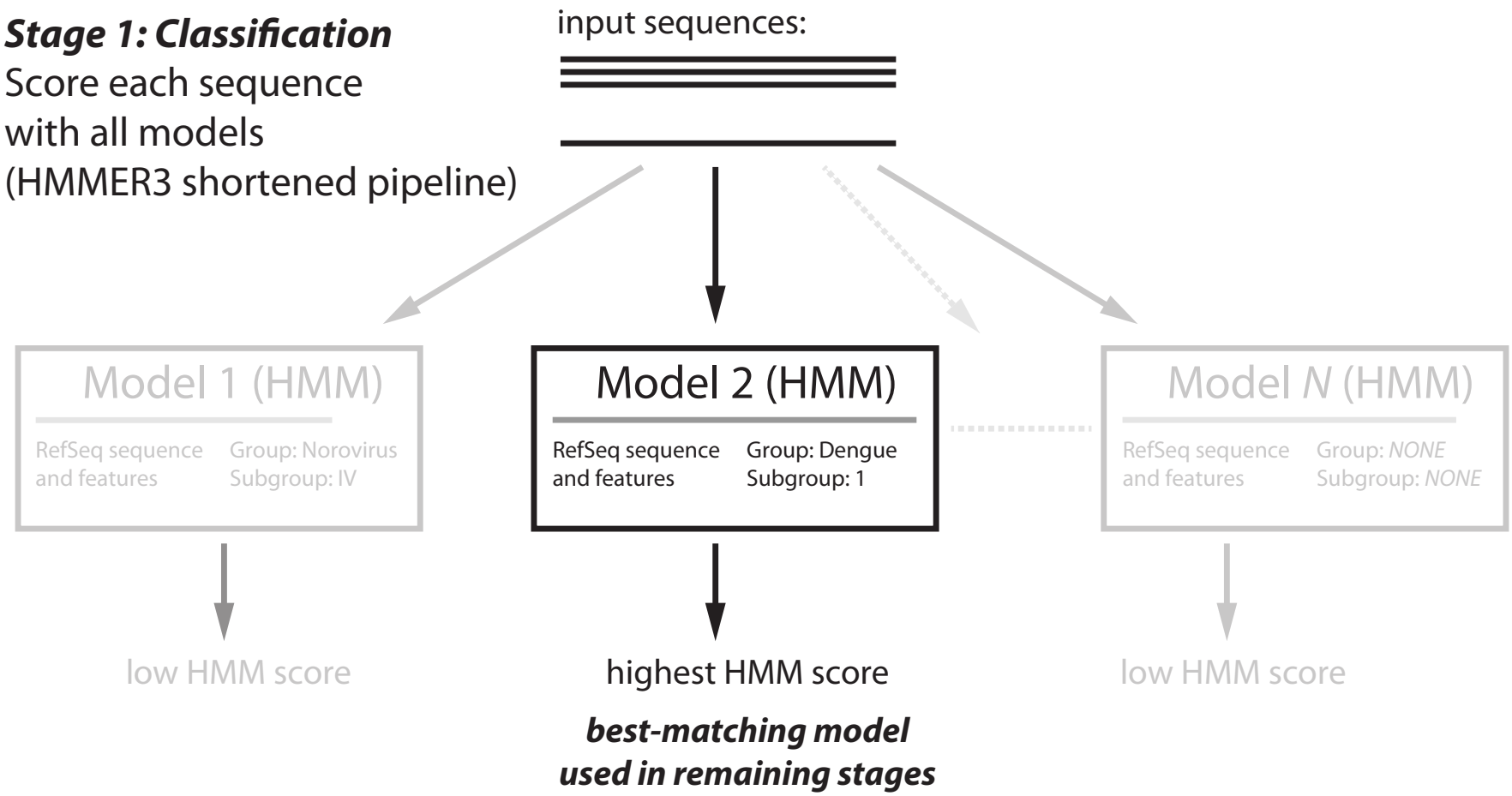
Stage 1: Classification

Score each sequence
with all models
(HMMER3 shortened pipeline)



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

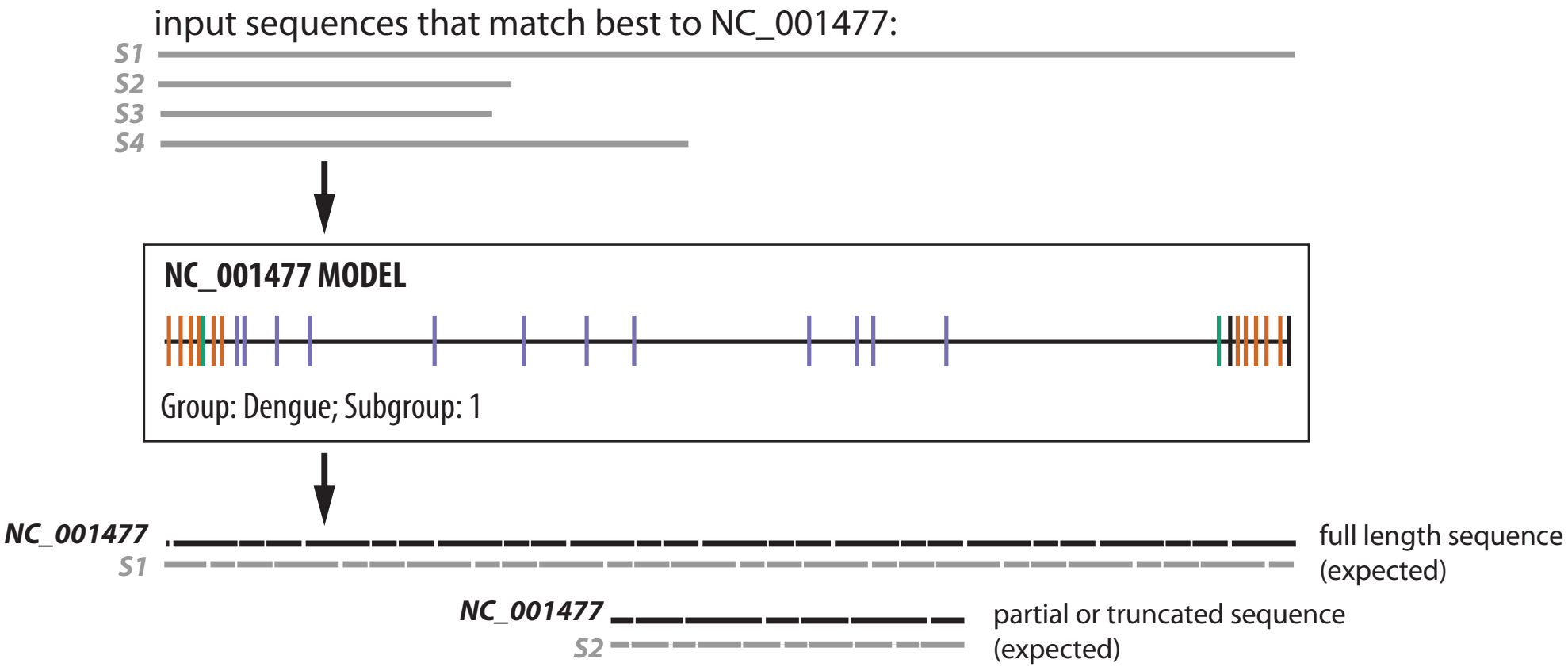




code	S/F	error message	description
Fatal alerts detected in the classification stage			
noannotn*	S	NO_ANNOTATION	no significant similarity detected
revcompl*	S	REVCOMPLEM	sequence appears to be reverse complemented
incsbgrp	S	INCORRECT_SPECIFIED_SUBGROUP	score difference too large between best overall 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

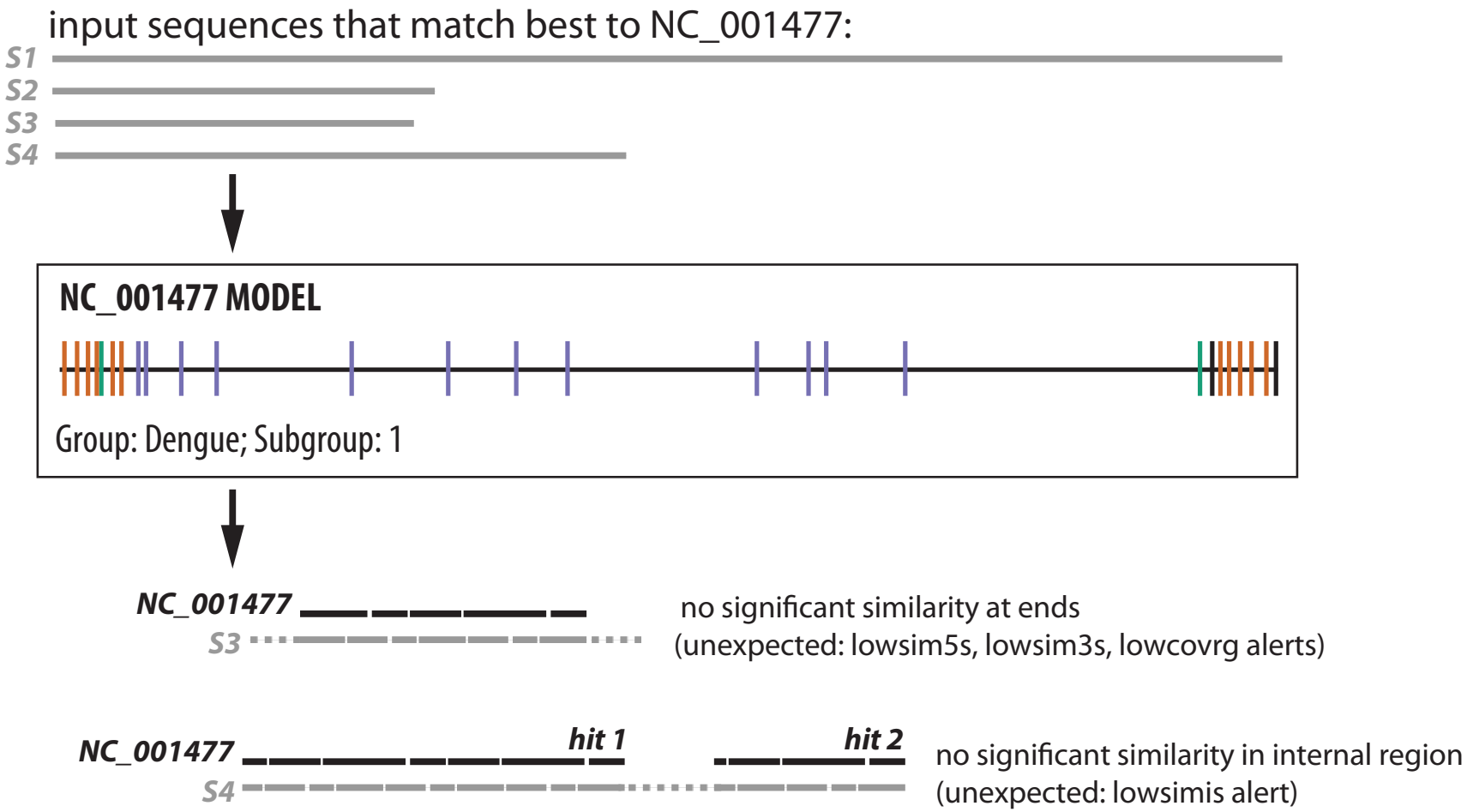
Stage 2: Coverage determination

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



Stage 2: Coverage determination

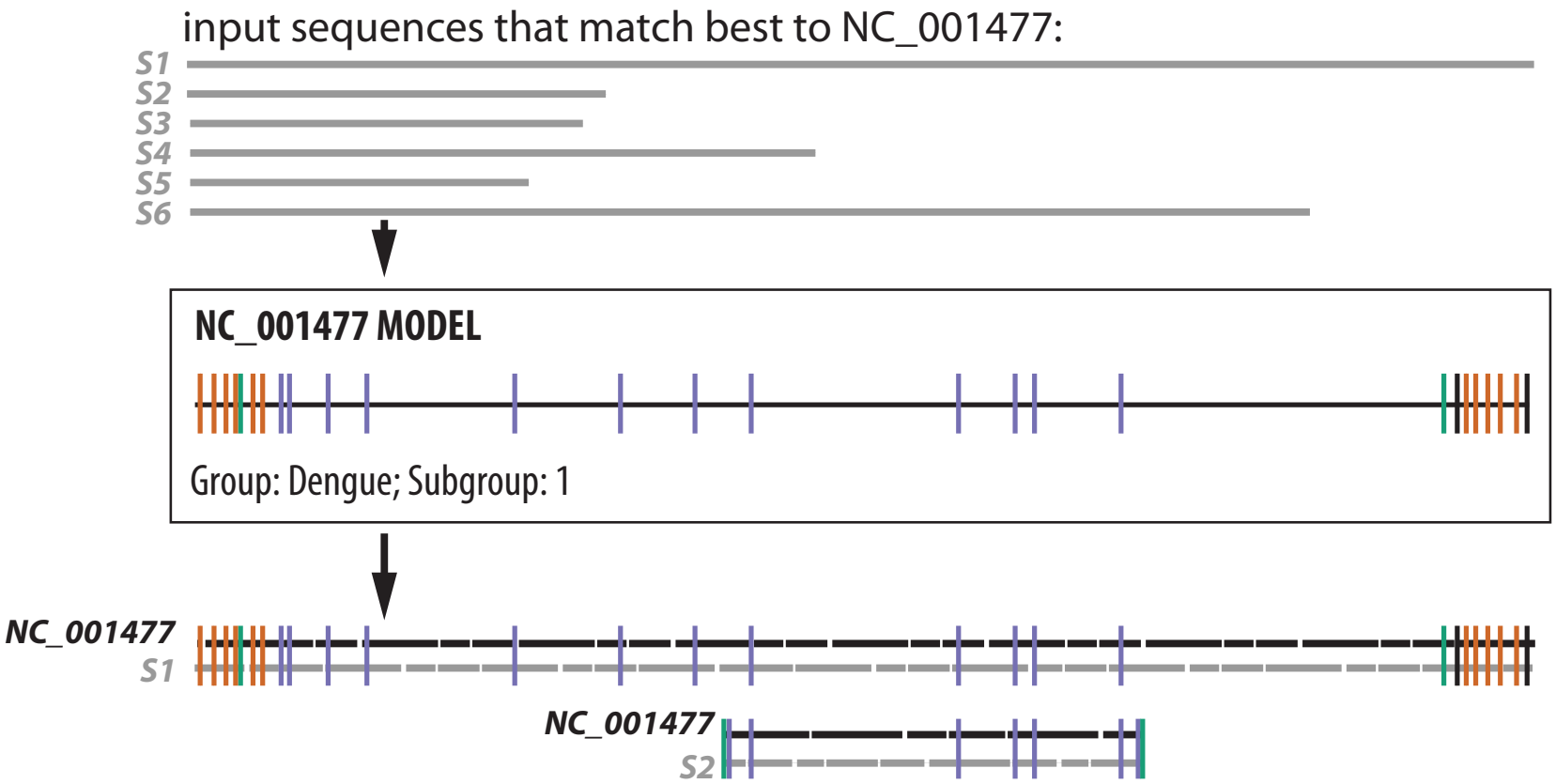
Search each sequence with best-matching model (HMMER3 full pipeline)



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)

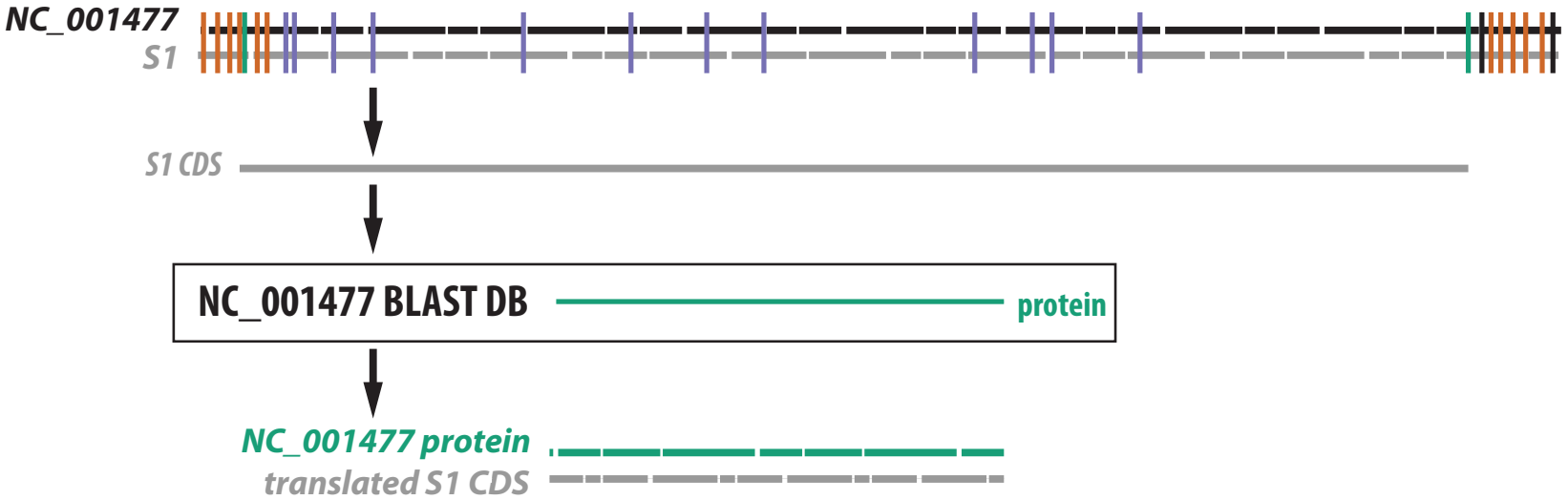
gap or low confidence at feature boundary
(indf5gap, indf3gap, indf5loc, indf3loc alert)

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

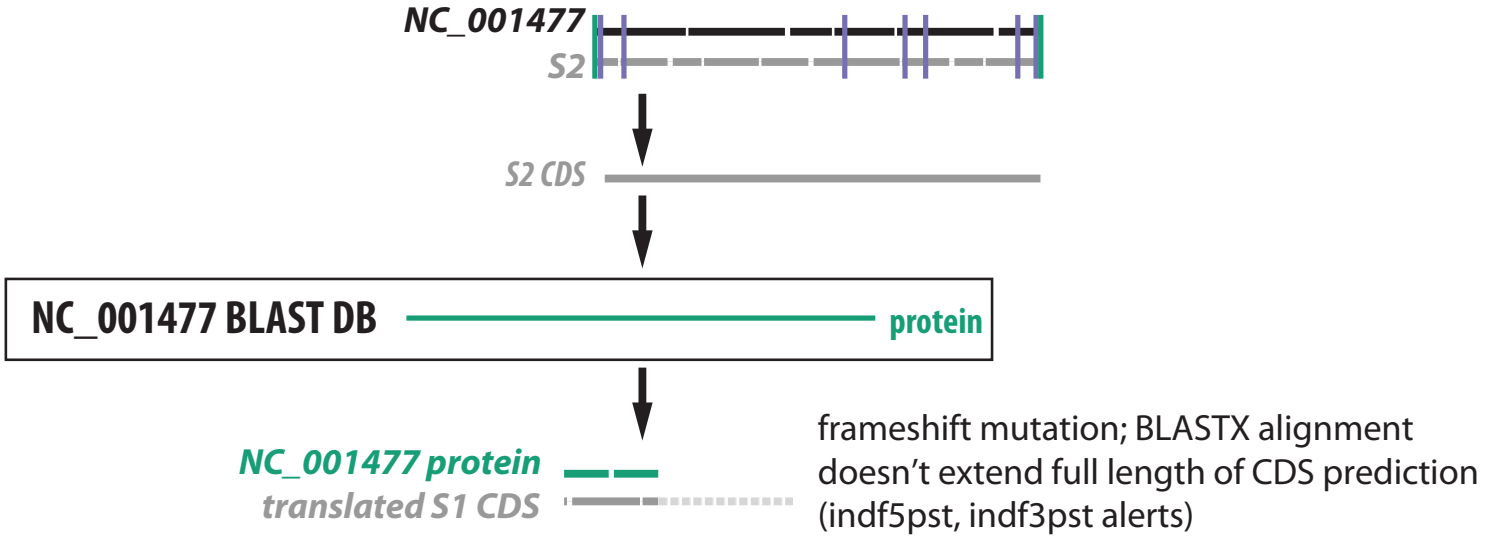
Stage 4: Protein validation

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



Stage 4: Protein validation

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 results on all Norovirus and Dengue sequences

dataset	# seqs	min length	max length	# pass	# fail	fraction pass
Norovirus complete (NC)	1,384	7380	7839	1,157	227	0.836
Dengue complete (DC)	4,580	10372	16254	4,171	409	0.911
Norovirus partial (NP)	32,190	50	7376	29,488	2,702	0.916
Dengue partial (DP)	20,973	50	10370	17,276	3,697	0.824

VADR results on all Norovirus and Dengue sequences

dataset	# seqs	min length	max length	# pass	# fail	fraction pass
Norovirus complete (NC)	1,384	7380	7839	1,157	227	0.836
Dengue complete (DC)	4,580	10372	16254	4,171	409	0.911
Norovirus partial (NP)	32,190	50	7376	29,488	2,702	0.916
Dengue partial (DP)	20,973	50	10370	17,276	3,697	0.824

VADR is portable so submitters can run on their data prior to submission to save time

Sequences processed with VADR will include annotations of CDS, mature peptides, and RNAs (`stem_loop` and `ncRNA`) from models

SARS-CoV-2 sequence submissions in 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

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
Mar	2022	141,333	2,246,931
Apr	2022	148,545	2,395,476
May	2022	164,276	2,559,752
Jun	2022	129,236	2,688,988
Jul	2022	101,737	2,790,725

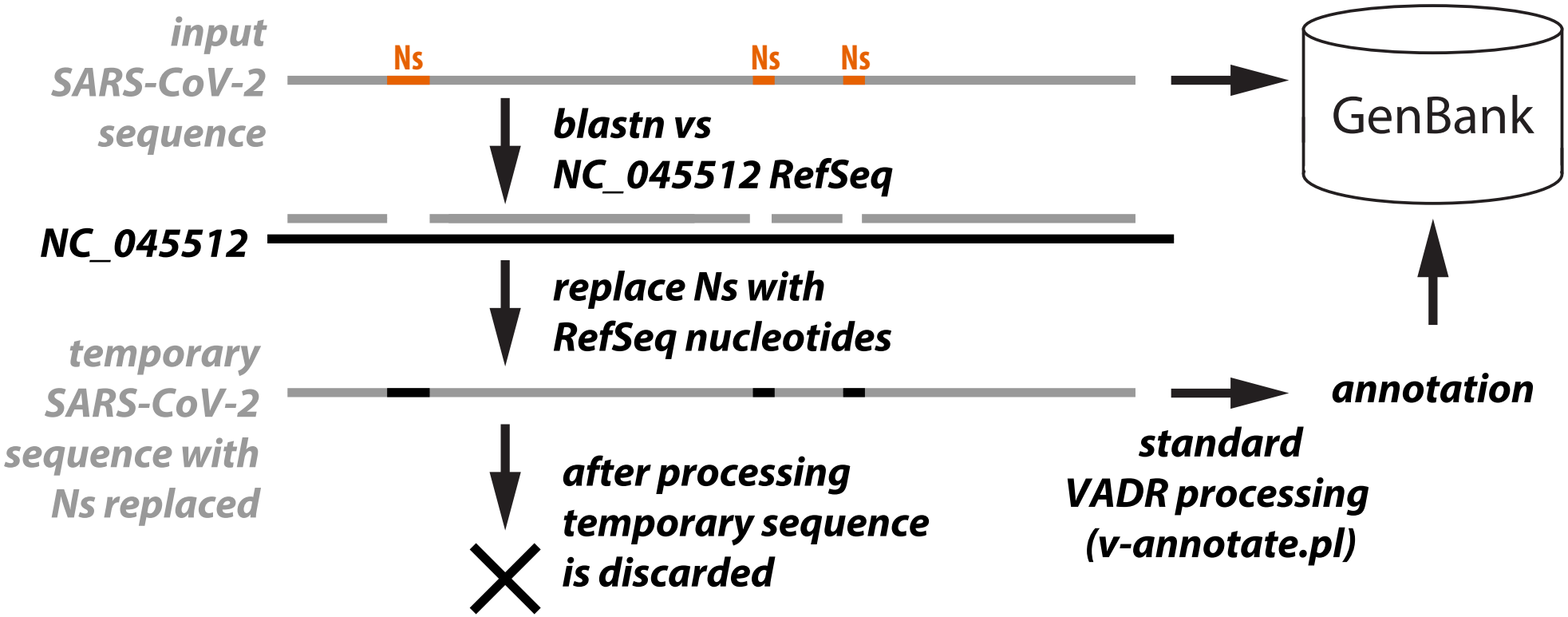
SARS-CoV-2 sequences differ from Norovirus and Dengue virus in several ways that impact VADR processing

	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

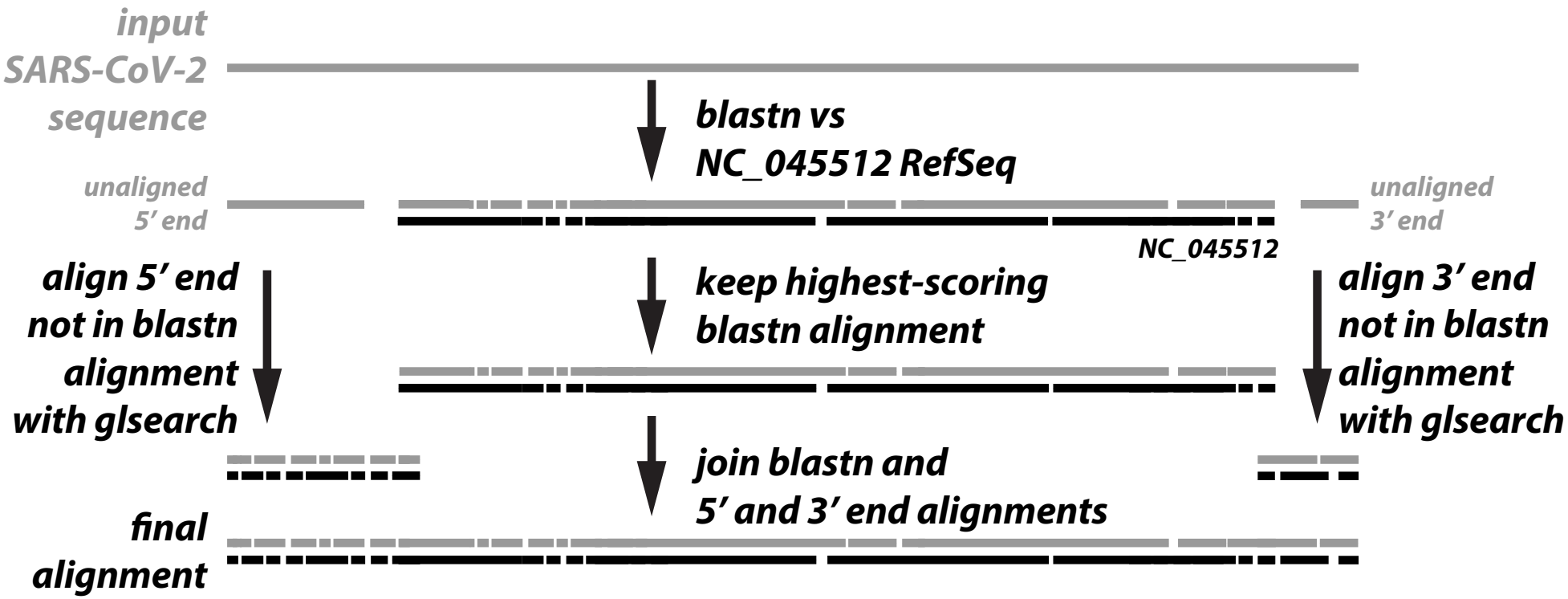
SARS-CoV-2 sequences differ from Norovirus and Dengue virus in several ways that impact VADR processing

	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

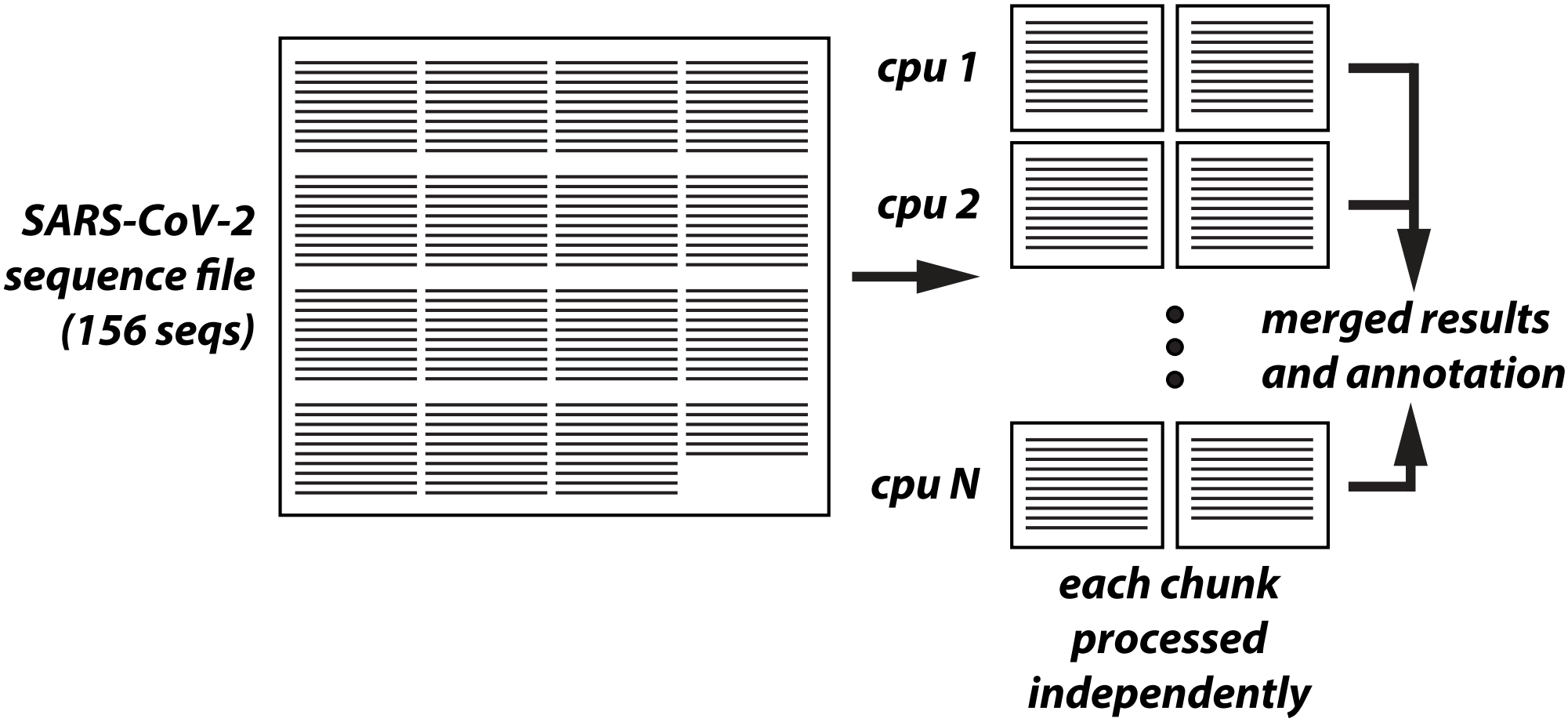


Seeded alignment using blastn makes alignment stage faster



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
Mar	2022	141,333	2,246,931
Apr	2022	148,545	2,395,476
May	2022	164,276	2,559,752
Jun	2022	129,236	2,688,988
Jul	2022	101,737	2,790,725

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

- 14 releases since March 2020 (thanks to "git flow"*)
- 3 additional SARS-CoV-2 models (all eventually dropped):
  - B.1.1.7 (alpha)
  - B.1.525
  - 28254-deletion
- allow some alerts for non-essential ORFs and s2m RNA element without failing sequence (they become a `misc_feature` instead)

*https://nvie.com/posts/a-successful-git-branching-model/

VADR is a general tool

- Also used for COX1 (cytochrome C oxidase subunit I) sequences using 43 class or order specific *profile*-based models covering 5 genetic codes

Limitations

- nucleotide space, not protein space
- RefSeq or alignment must be 'representative' and conserved along full length
  - divergent sequences, regions, repeats, introns, gene order are problematic
- slow (SARS-CoV-2 speedups are not general)
- SARS-CoV-2 sequences that fail are kept out of the database

Future directions

- extend to more viruses
- alignment-based models for viruses

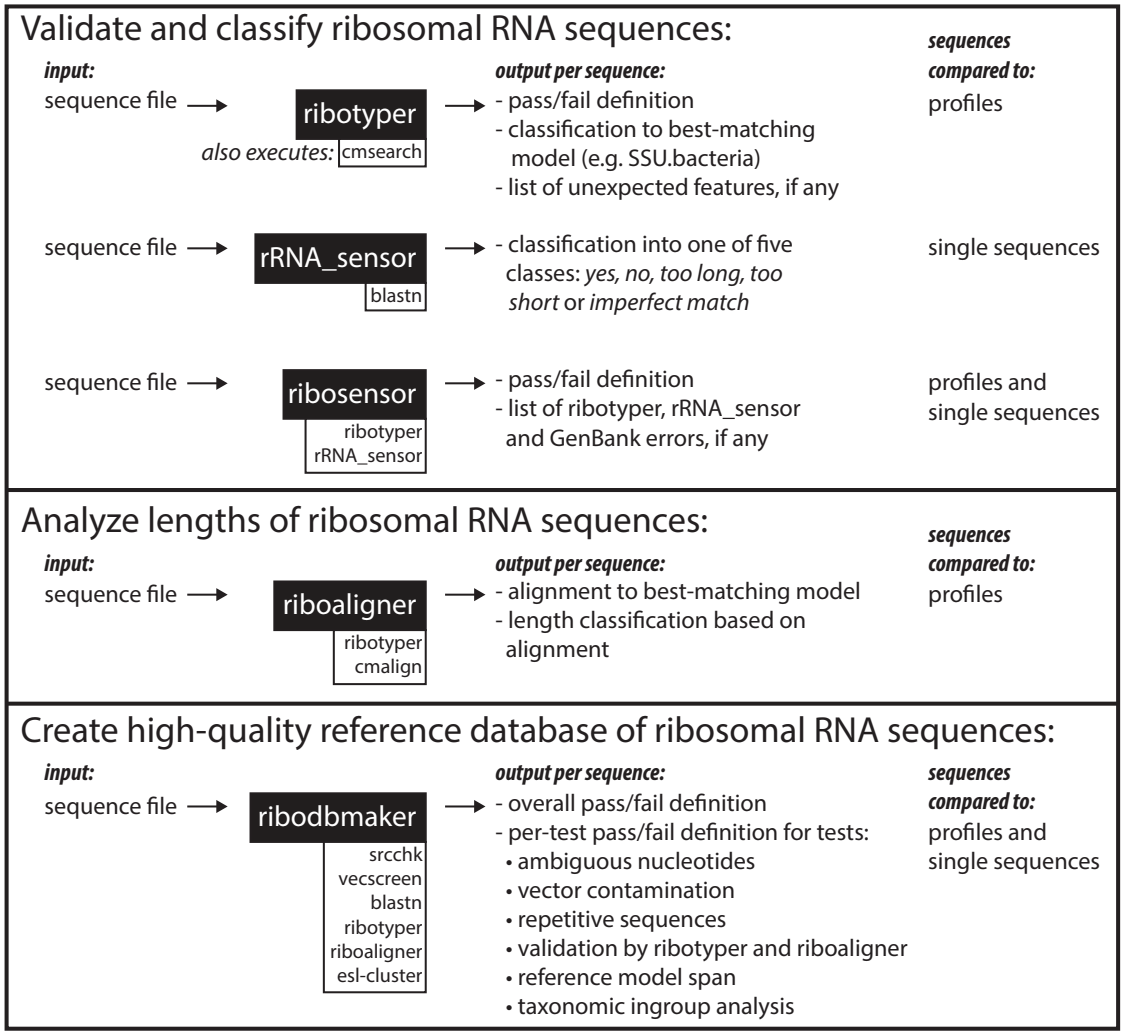
Schäffer *et al. BMC Bioinformatics* (2021) 22:400
https://doi.org/10.1186/s12859-021-04316-z

BMC Bioinformatics

SOFTWARE Open Access

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

Alejandro A. Schäffer[1,2], Richard McVeigh[2], Barbara Robbertse[2], Conrad L. Schoch[2], Anjanette Johnston[2], Beverly A. Underwood[2], Ilene Karsch-Mizrachi[2] and Eric P. Nawrocki[2*]



Ribovore will (hopefully) help drive addition and improvement of Rfam rRNA models

Table 3 Profile models used by Ribovore

Model name	Gene	Taxonomy group	#Seqs	Length	Rfam
SSU_rRNA_archaea	SSU rRNA	Archaea	86	1477	RF01959
SSU_rRNA_bacteria	SSU rRNA	Bacteria	99	1533	RF00177
SSU_rRNA_eukarya	SSU rRNA	Eukarya	91	1851	RF01960
SSU_rRNA_microsporidia	SSU rRNA	Euk-Microsporidia	46	1312	RF02542
LSU_rRNA_archaea	LSU rRNA	Archaea	91	2990	RF02540
LSU_rRNA_bacteria	LSU rRNA	Bacteria	102	2925	RF02541
LSU_rRNA_eukarya	LSU rRNA	Eukarya	88	3401	RF02543
SSU_rRNA_mitochondria_metazoa	SSU rRNA	Mito-Metazoa	83	954	–
SSU_rRNA_mitochondria_amoeba	SSU rRNA	Mito-Amoeba	2	1861	–
SSU_rRNA_mitochondria_chlorophyta	SSU rRNA	Mito-Chlorophyta	2	1200	–
SSU_rRNA_mitochondria_fungi	SSU rRNA	Mito-Fungi	4	1603	–
SSU_rRNA_mitochondria_kinetoplast	SSU rRNA	Mito-Kinetoplast	3	624	–
SSU_rRNA_mitochondria_plant	SSU rRNA	Mito-Plant	4	1951	–
SSU_rRNA_mitochondria_protist	SSU rRNA	Mito-Protist	2	1677	–
SSU_rRNA_chloroplast	SSU rRNA	Chloroplast	94	1488	–
SSU_rRNA_chloroplast_pilostyles	SSU rRNA	Chloroplast	1	1531	–
SSU_rRNA_cyanobacteria	SSU rRNA	Bac-Cyanobacteria	49	1487	–
SSU_rRNA_apicoplast	SSU rRNA	Euk-Apicoplast	3	1463	–

'#seqs' is the number of sequences in the multiple alignment used to build the model. 'length' is the number of reference model positions. Abbreviations in 'taxonomy group' column: 'Bac' is Bacteria, 'Euk' is Eukarya and 'Mito' is Mitochondria

Acknowledgements

NCBI - viral annotation
Alejandro Schäffer (now NCI)

Linda Yankie
Vincent Calhoun
Sergiy Gotvyanskyy
Susan Schafer
Ilene Mizrachi
Colleen Bollin
Beverly Underwood
Prakash Keranahalli
Vasuki Gobu
Alex Kotliarov

Rodney Brister
Eneida Hatcher

Lara Shonkwiler
Sophia Hu

Wratko Hlavina
Ron Patterson

NCBI - leadership
David Landsman
Kim Pruitt
Steve Sherry
Jim Ostell
David Lipman

NLM - leadership
Patti Brennan
Jerry Sheehan
Valerie Florance

Software developers
Sean Eddy (HMMER/Infernal/Easel)
Travis Wheeler (HMMER)
Tom Madden and BLAST team
William Pearson (FASTA/glsearch)
Michael Farrar (HMMER/glsearch)

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.