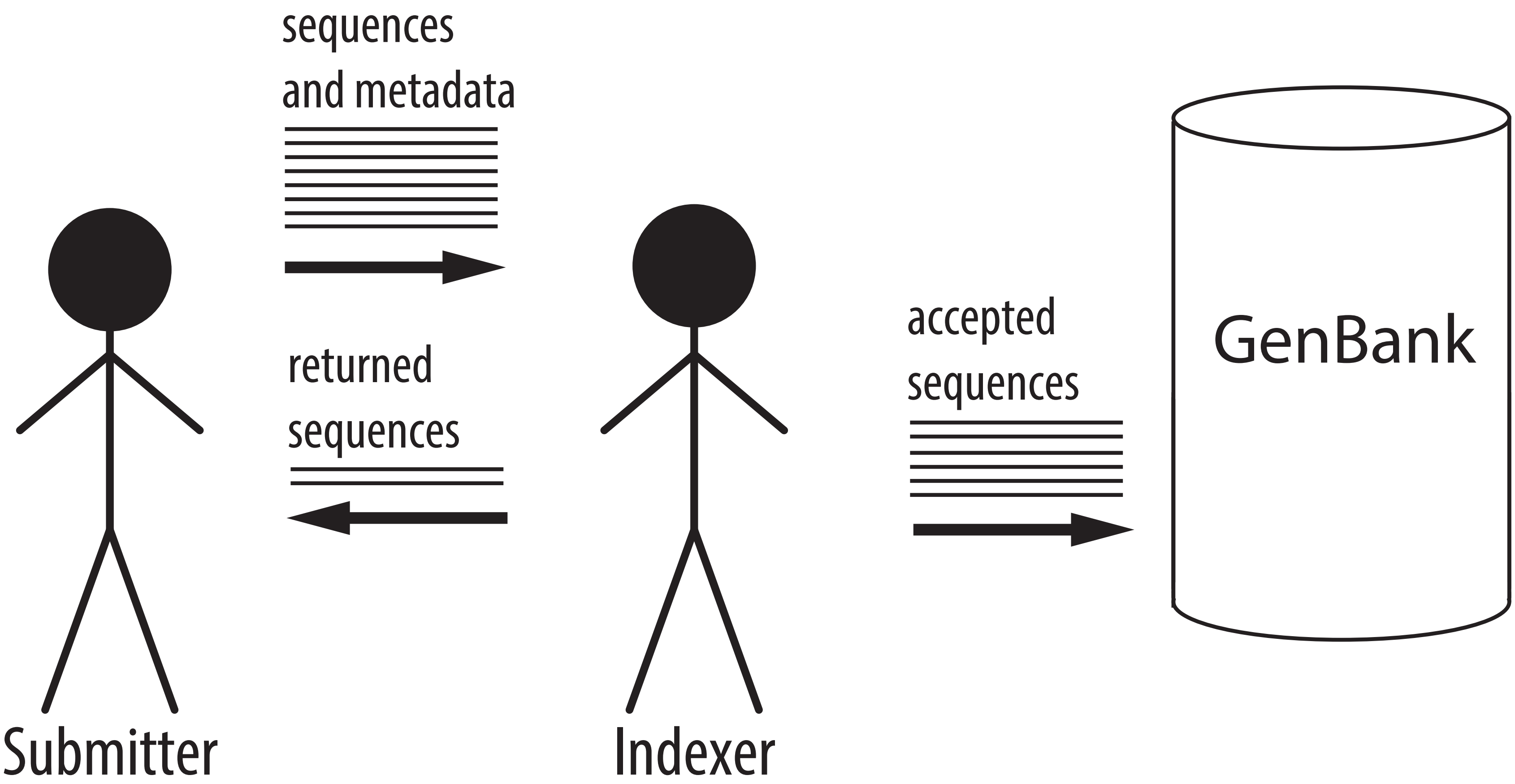
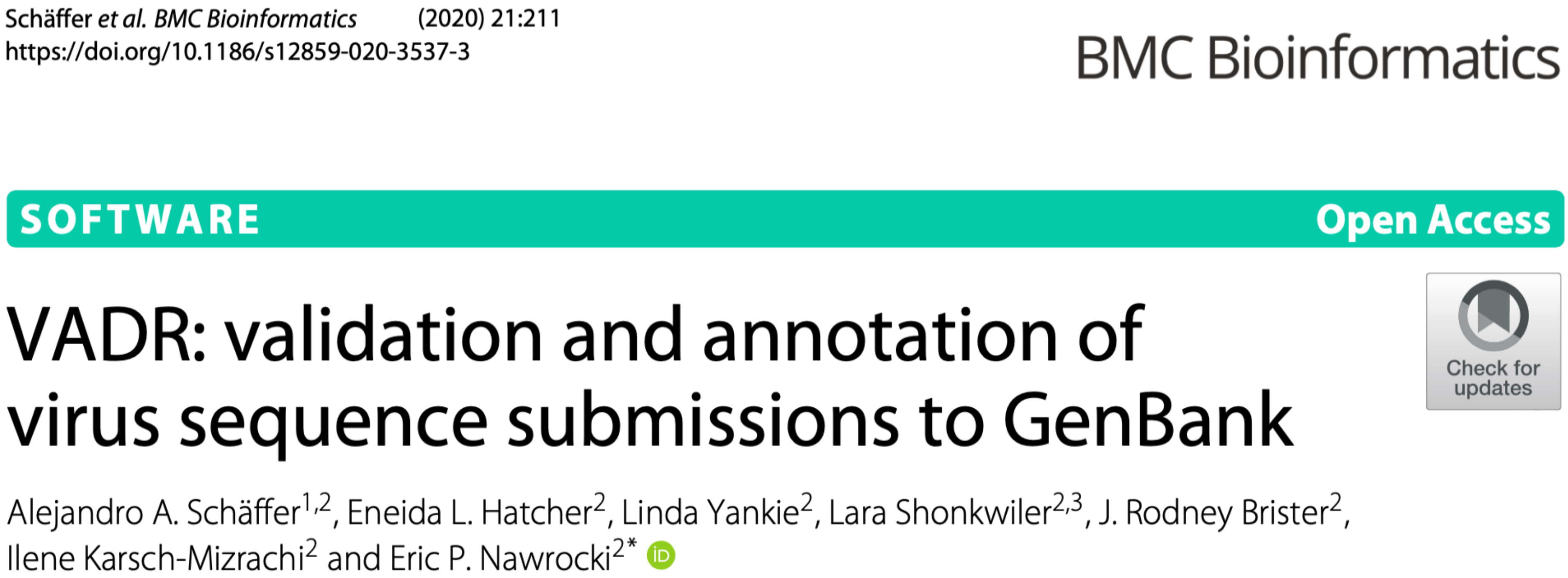
Validation and annotation of SARS-CoV-2 sequences for GenBank using VADR

Eric Nawrocki

National Center for Biotechnology Information
National Library of Medicine

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

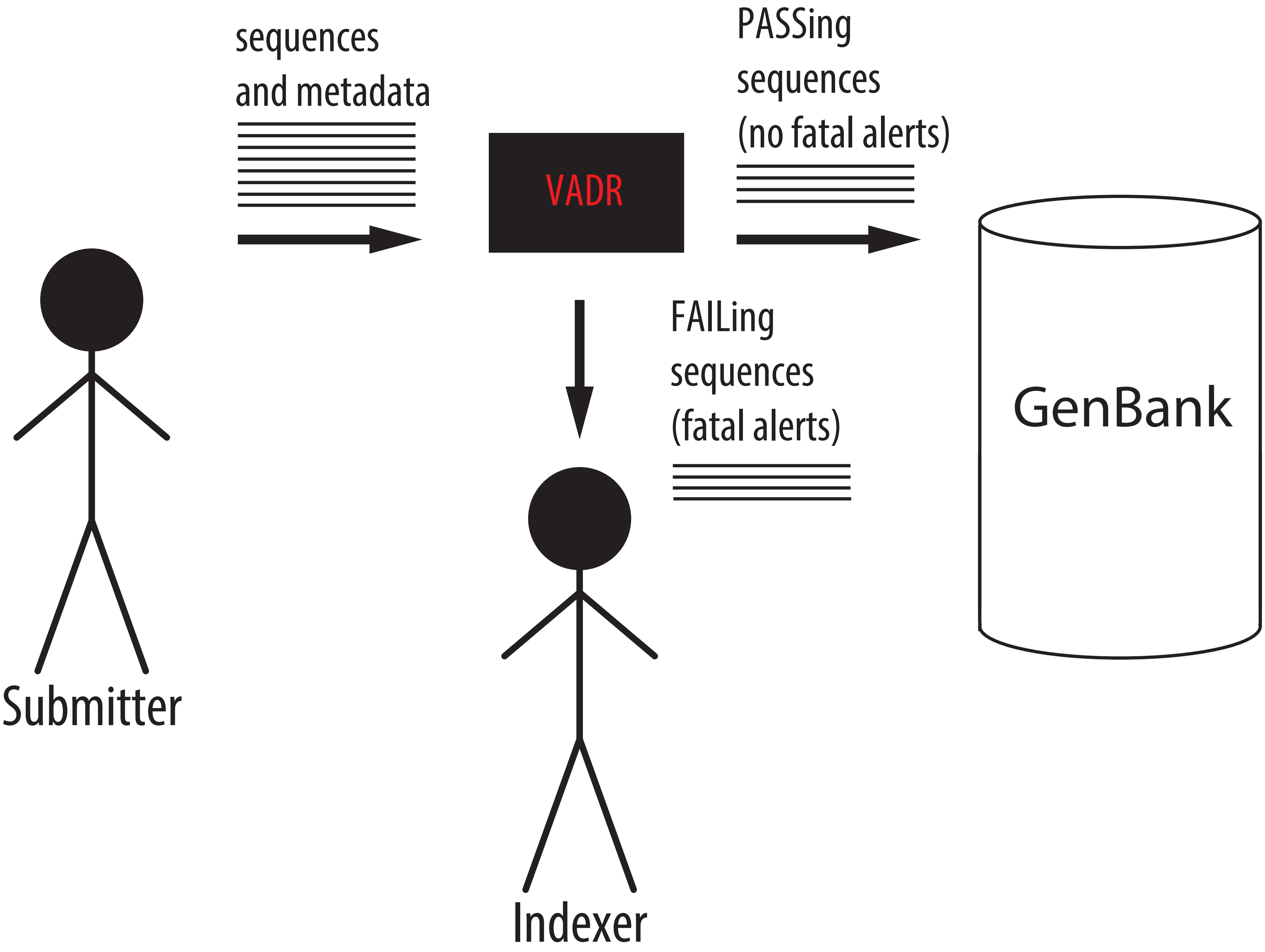
SOFTWARE Open Access

VADR: validation and annotation of virus sequence submissions to GenBank

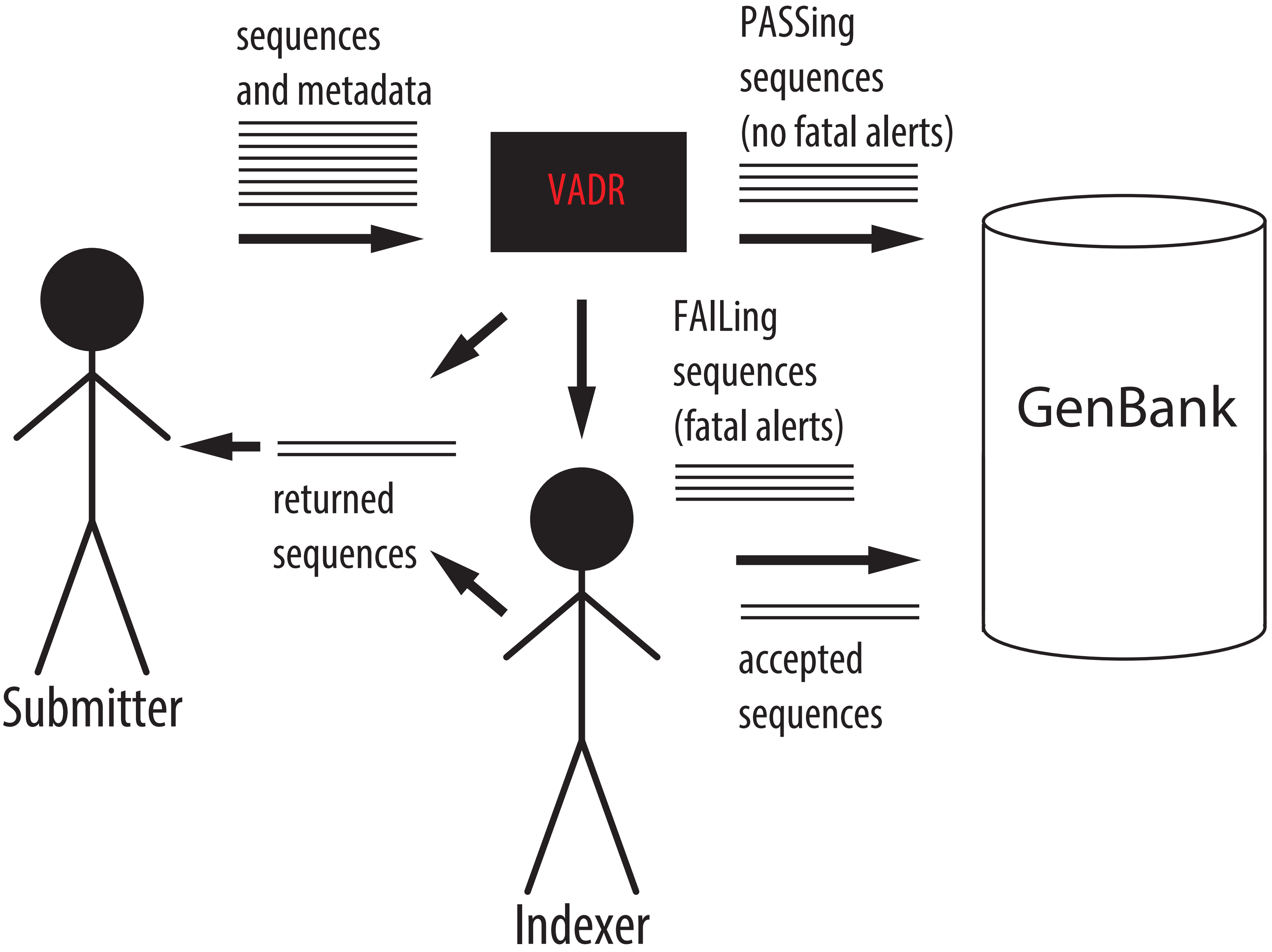
Alejandro A. Schäffer[1,2], Eneida L. Hatcher[2], Linda Yankie[2], Lara Shonkwiler[2,3], J. Rodney Brister[2], Ilene Karsch-Mizrachi[2] and Eric P. Nawrocki[2*]

- general tool for reference-based annotation of viral sequences
- used for dengue virus and norovirus submissions since 2018
- used for SARS-CoV-2 submissions since March 2020

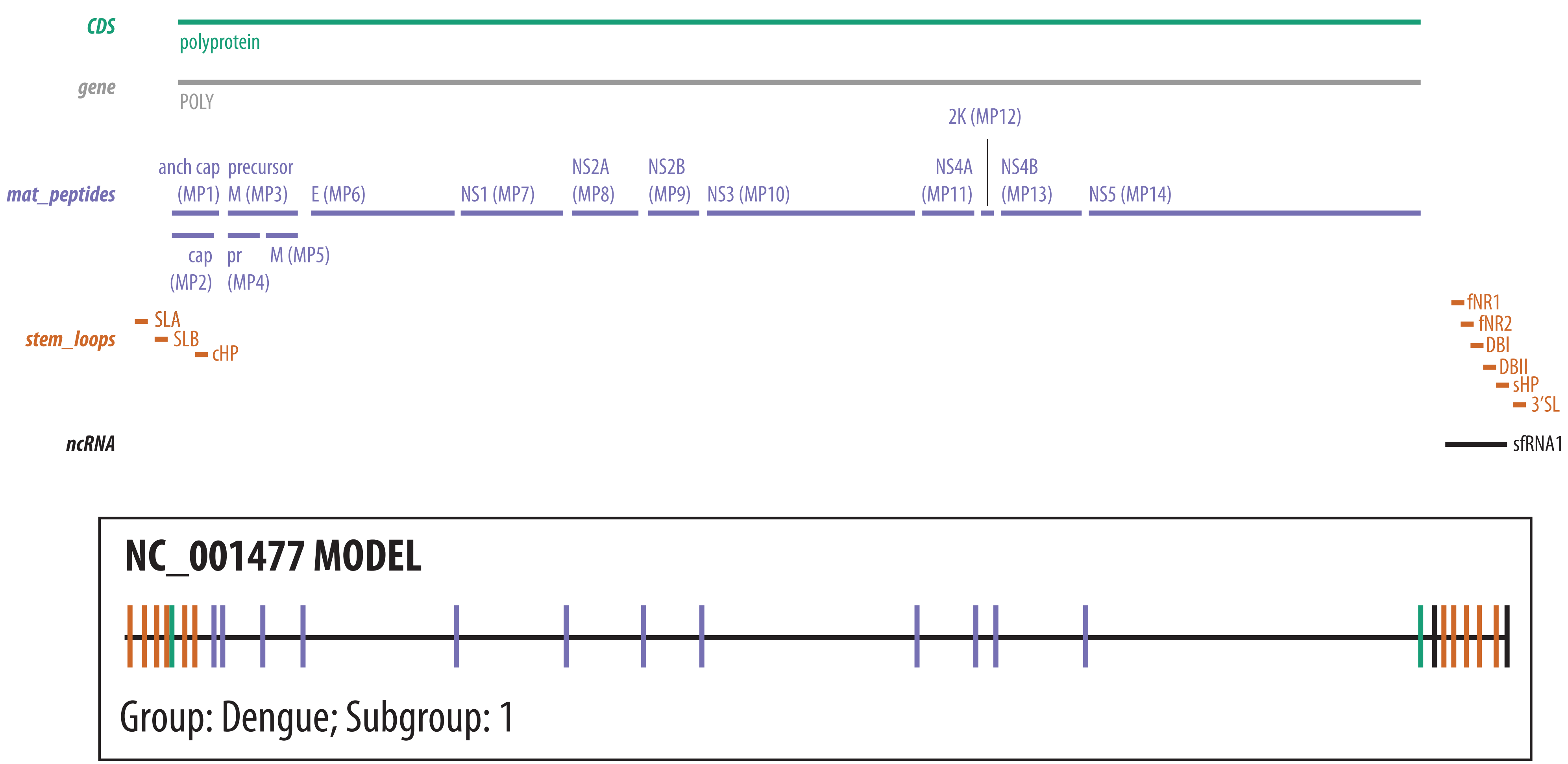
VADR assists GenBank indexers:
Each sequence PASSes or FAILs



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



VADR builds a reference model of a RefSeq and its features



CDS

polyprotein

gene

POLY

mat_peptides

anch cap (MP1)

precursor M (MP3)

E (MP6)

NS1 (MP7)

NS2A (MP8)

NS2B (MP9)

NS3 (MP10)

NS4A (MP11)

2K (MP12)

NS4B (MP13)

NS5 (MP14)

cap (MP2)

pr (MP4)

M (MP5)

stem_loops

SLA

SLB

cHP

fNR1

fNR2

DBI

DBII

sHP

3'SL

ncRNA

sfRNA1

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:

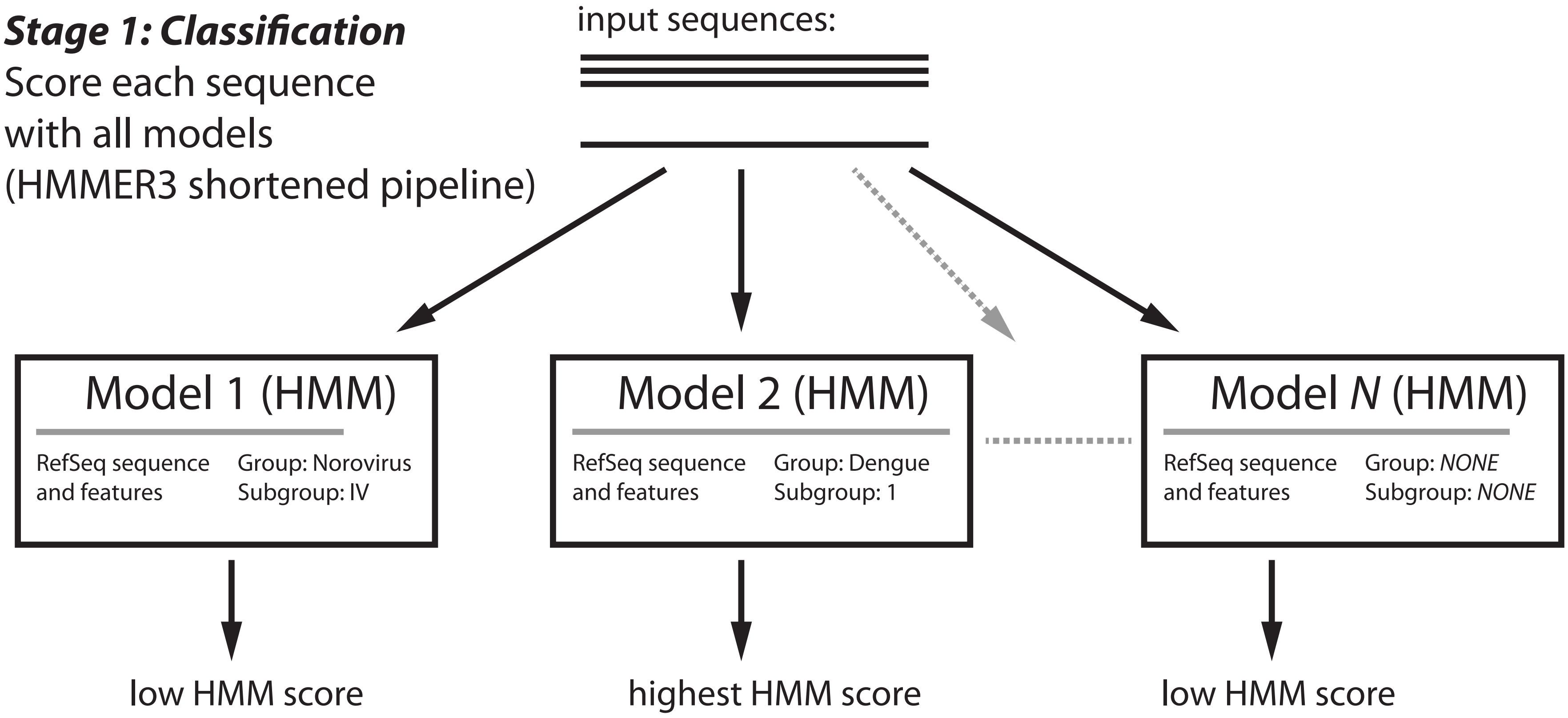
1. **Classification**

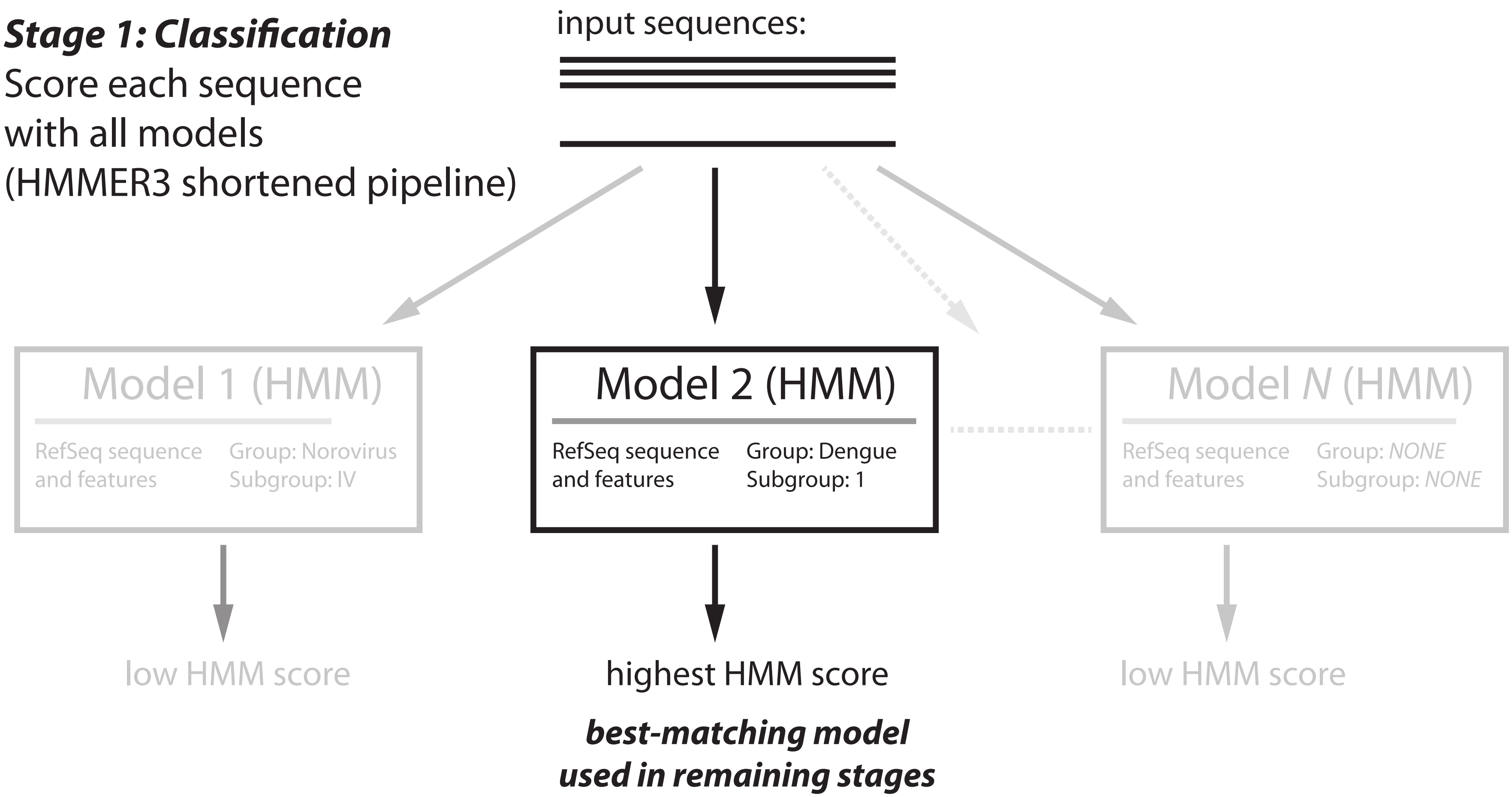
2. **Coverage determination**

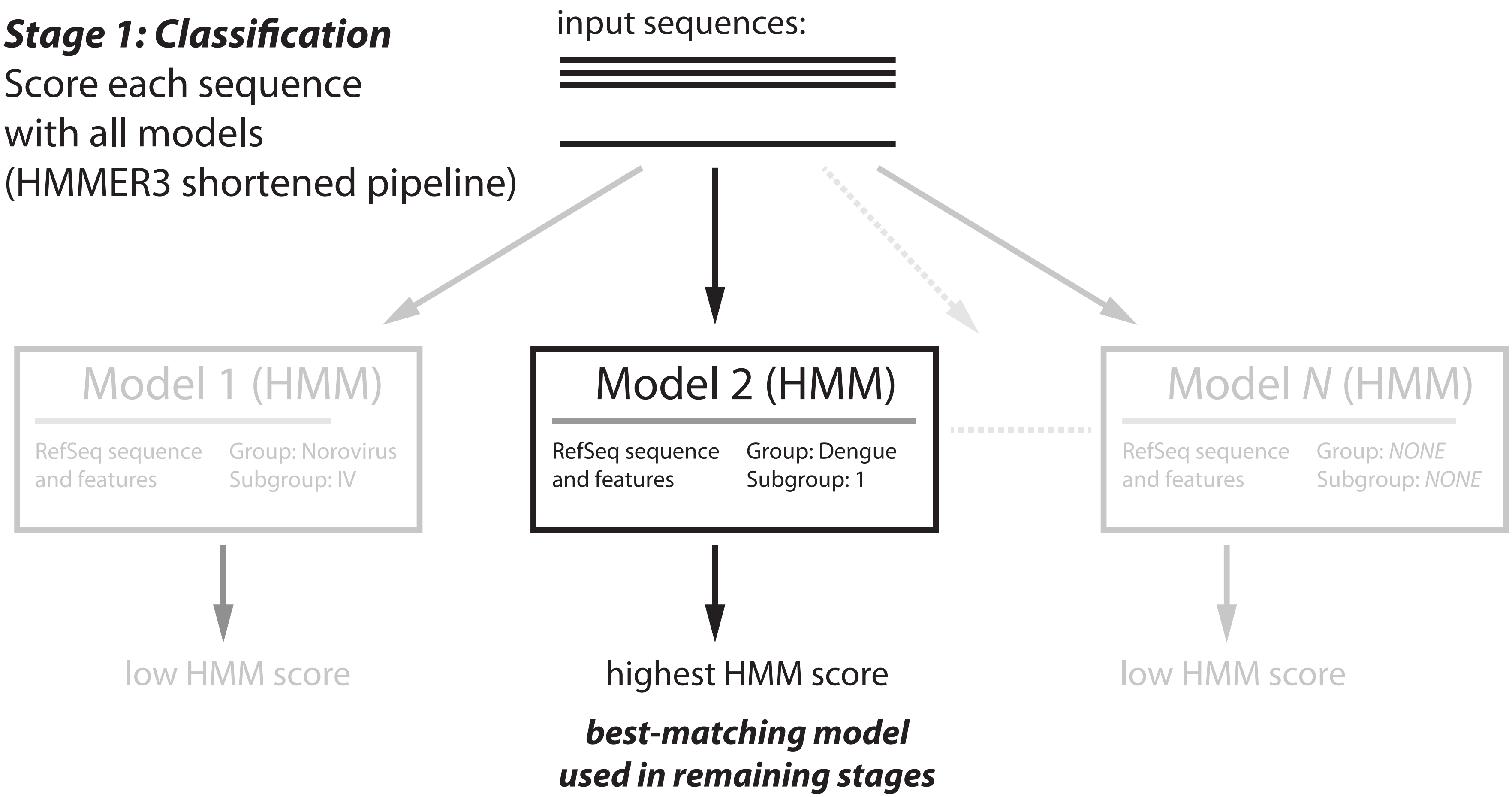
3. **Alignment**

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:
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

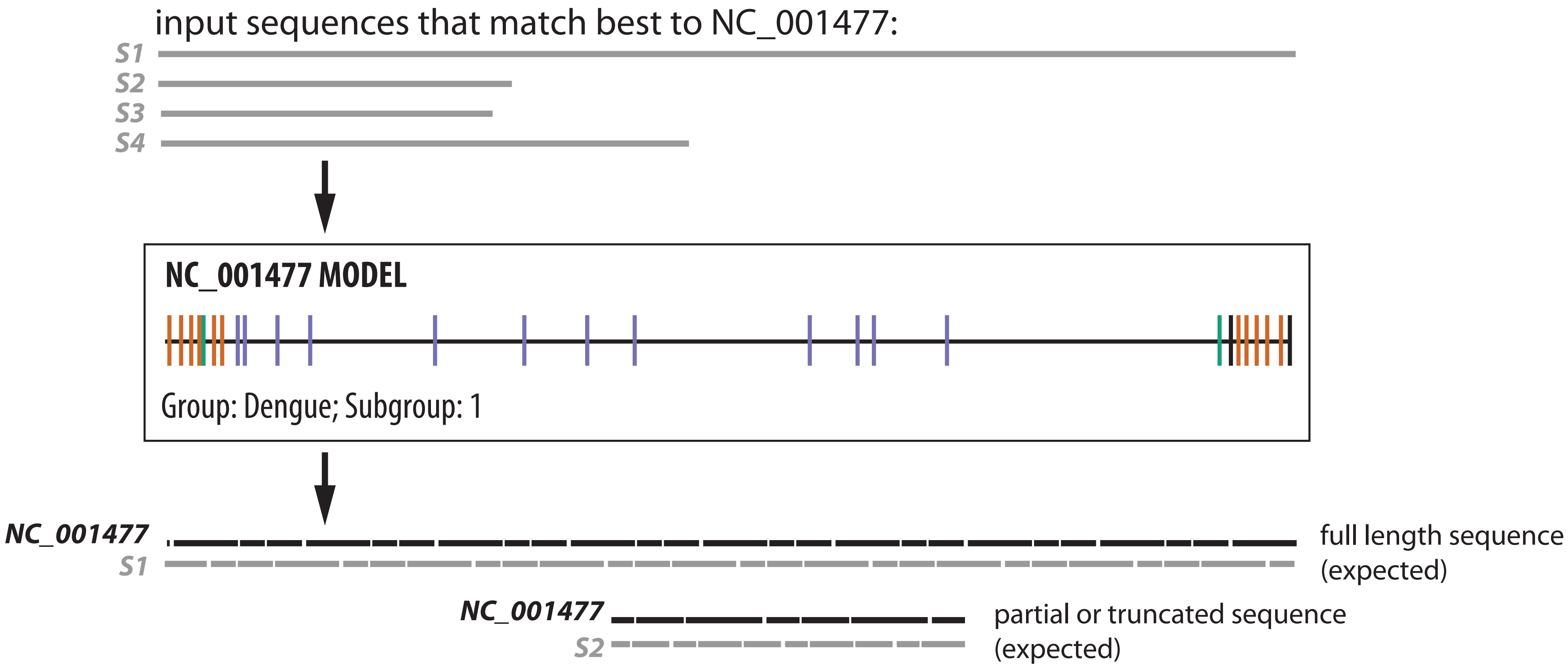

Stage 1: Classification
Score each sequence
with all models
(HMMER3 shortened pipeline)
input sequences:
Model 1 (HMM)
RefSeq sequence and features
Group: Norovirus
Subgroup: IV
Model 2 (HMM)
RefSeq sequence and features
Group: Dengue
Subgroup: 1
Model N (HMM)
RefSeq sequence and features
Group: NONE
Subgroup: NONE
low HMM score
highest HMM score
low HMM score
best-matching model
used in remaining stages



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)



input sequences that match best to NC_001477:

NC_001477 MODEL

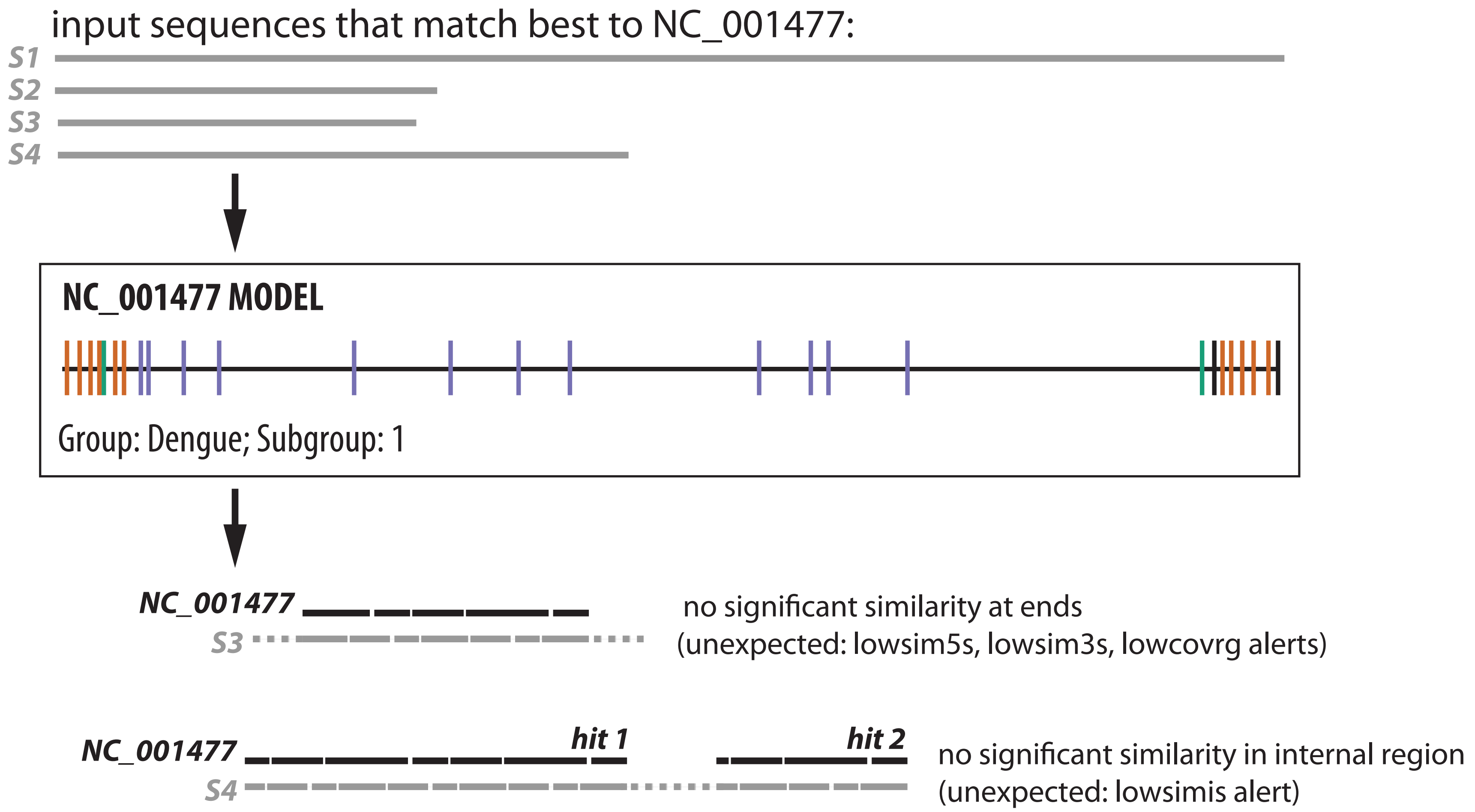
Group: Dengue; Subgroup: 1

full length sequence (expected)

partial or truncated sequence (expected)

Stage 2: Coverage determination

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

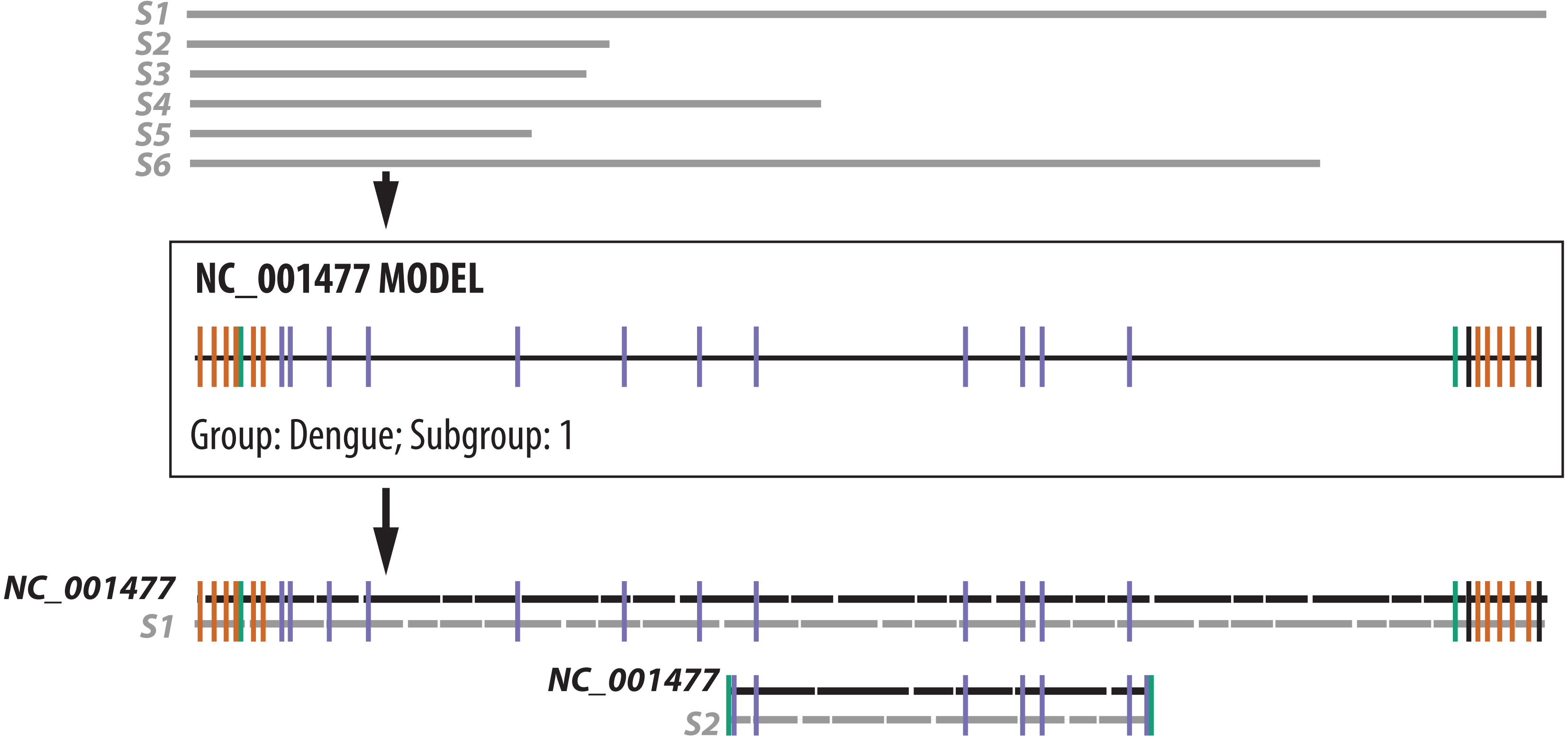

input sequences that match best to NC_001477:
S1
S2
S3
S4
NC_001477 MODEL
Group: Dengue; Subgroup: 1
NC_001477
S3
no significant similarity at ends
(unexpected: lowsim5s, lowsim3s, lowcovrg alerts)
NC_001477
hit 1
hit 2
S4
no significant similarity in internal region
(unexpected: lowsimis alert)

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)

input sequences that match best to NC_001477:

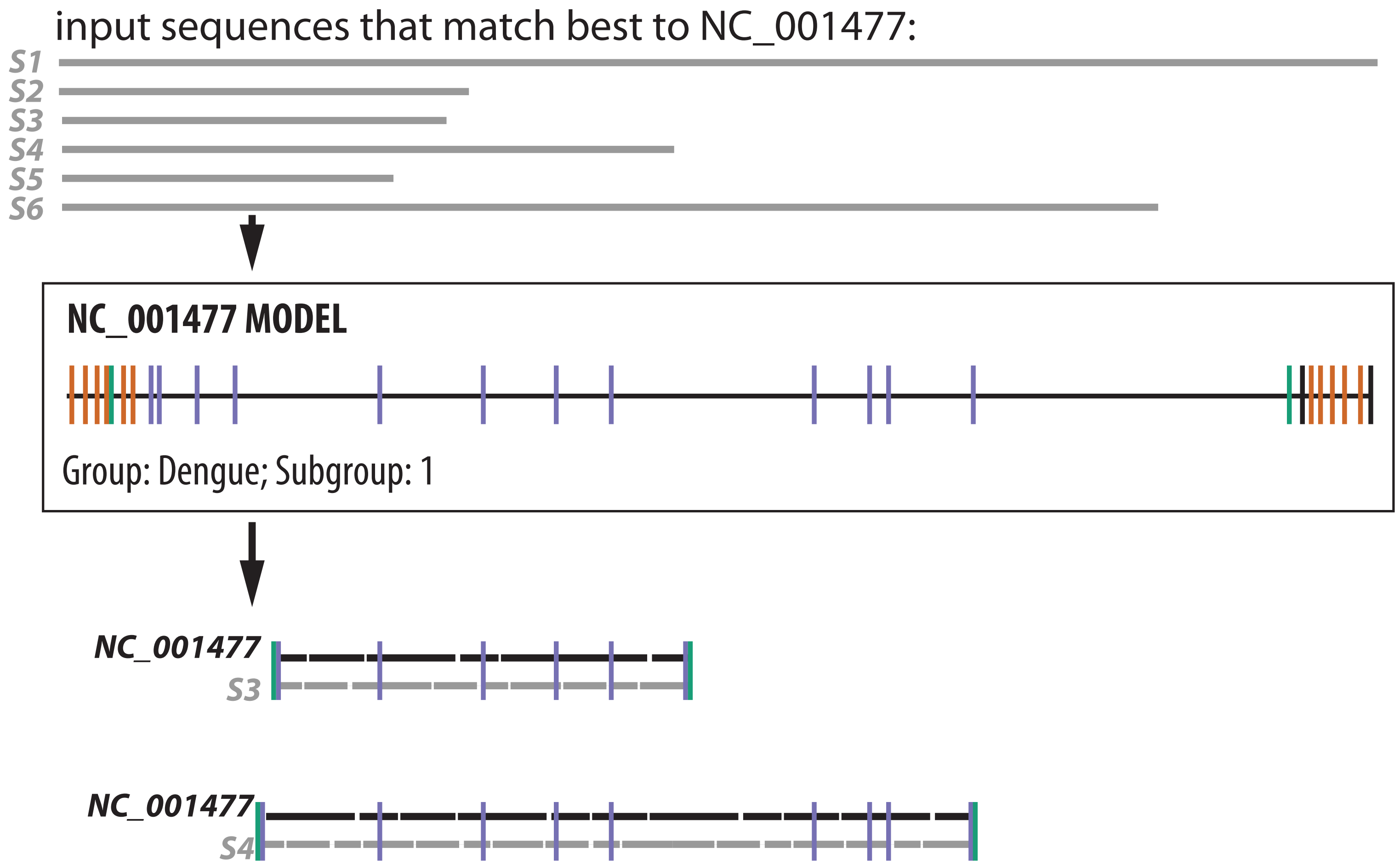


NC_001477 MODEL

Group: Dengue; Subgroup: 1

Stage 3: Alignment and feature mapping

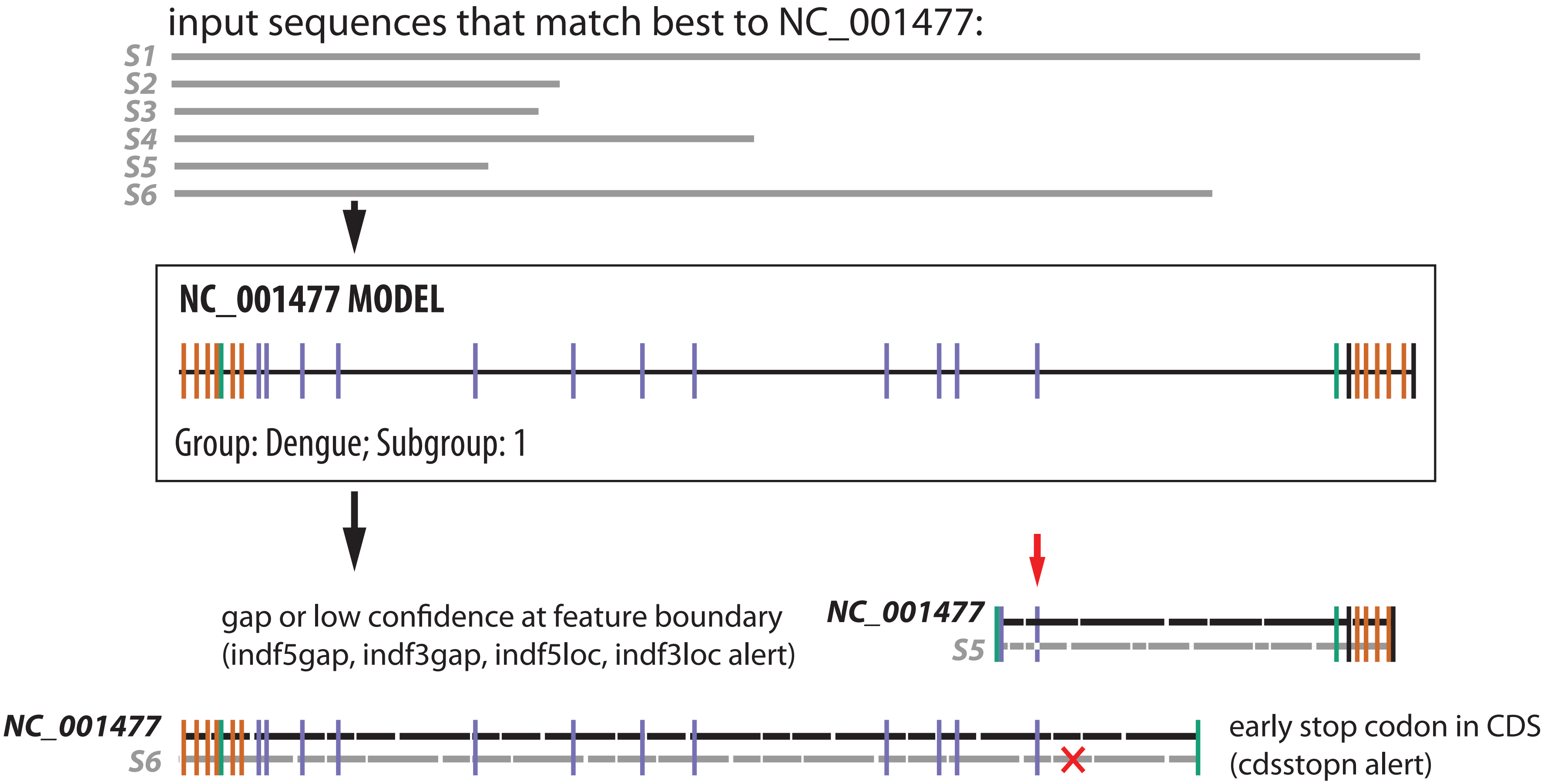
Align each sequence to its best-matching model (Infernal’s cmalign)



input sequences that match best to NC_001477:

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:

Group: Dengue; Subgroup: 1

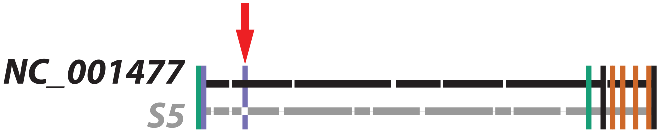
gap or low confidence at feature boundary
(indf5gap, indf3gap, indf5loc, indf3loc alert)

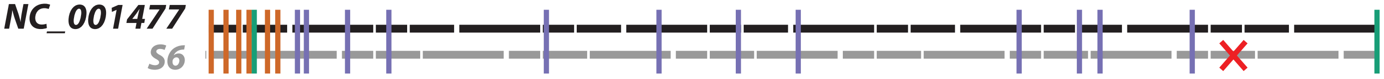
early stop codon in CDS
(cdsstopn alert)

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)



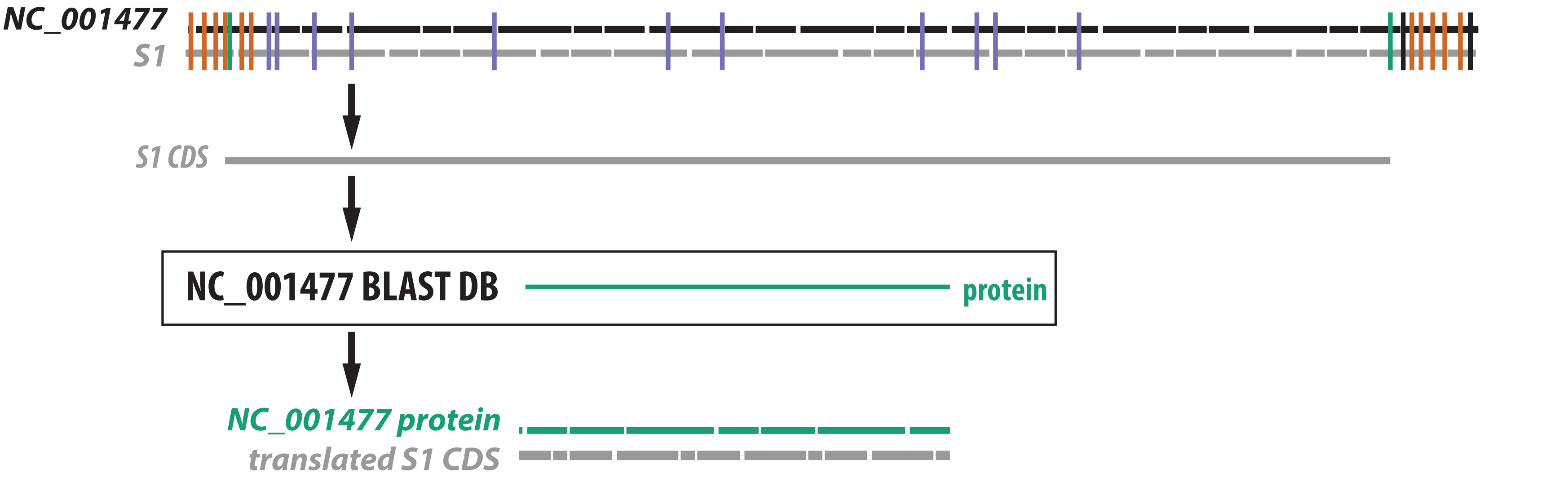


early stop codon in CDS
(cdsstopn alert)

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



NC_001477

S1

S1 CDS

NC_001477 BLAST DB

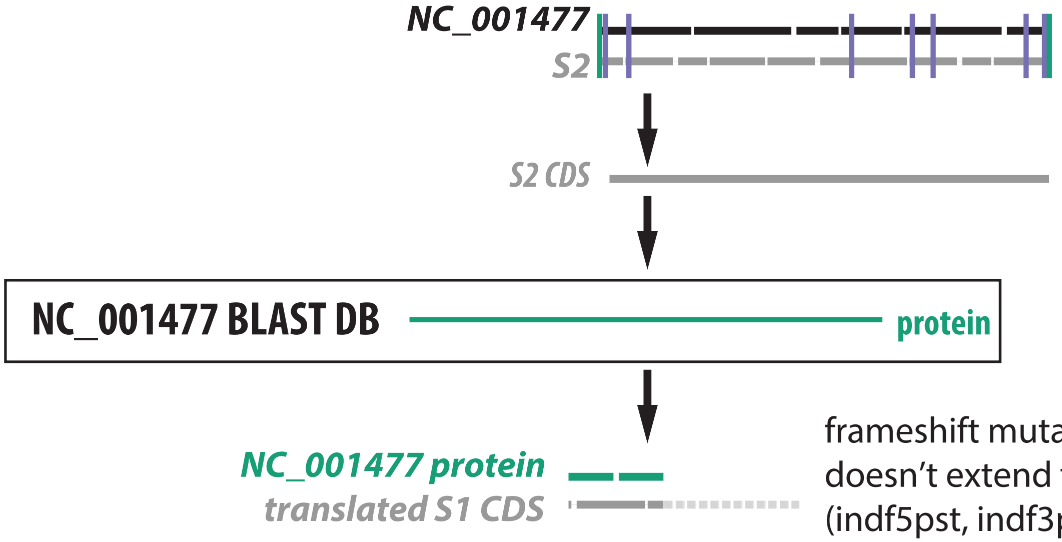
protein

NC_001477 protein

translated S1 CDS

Stage 4: Protein validation (Alejandro Schäffer)

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

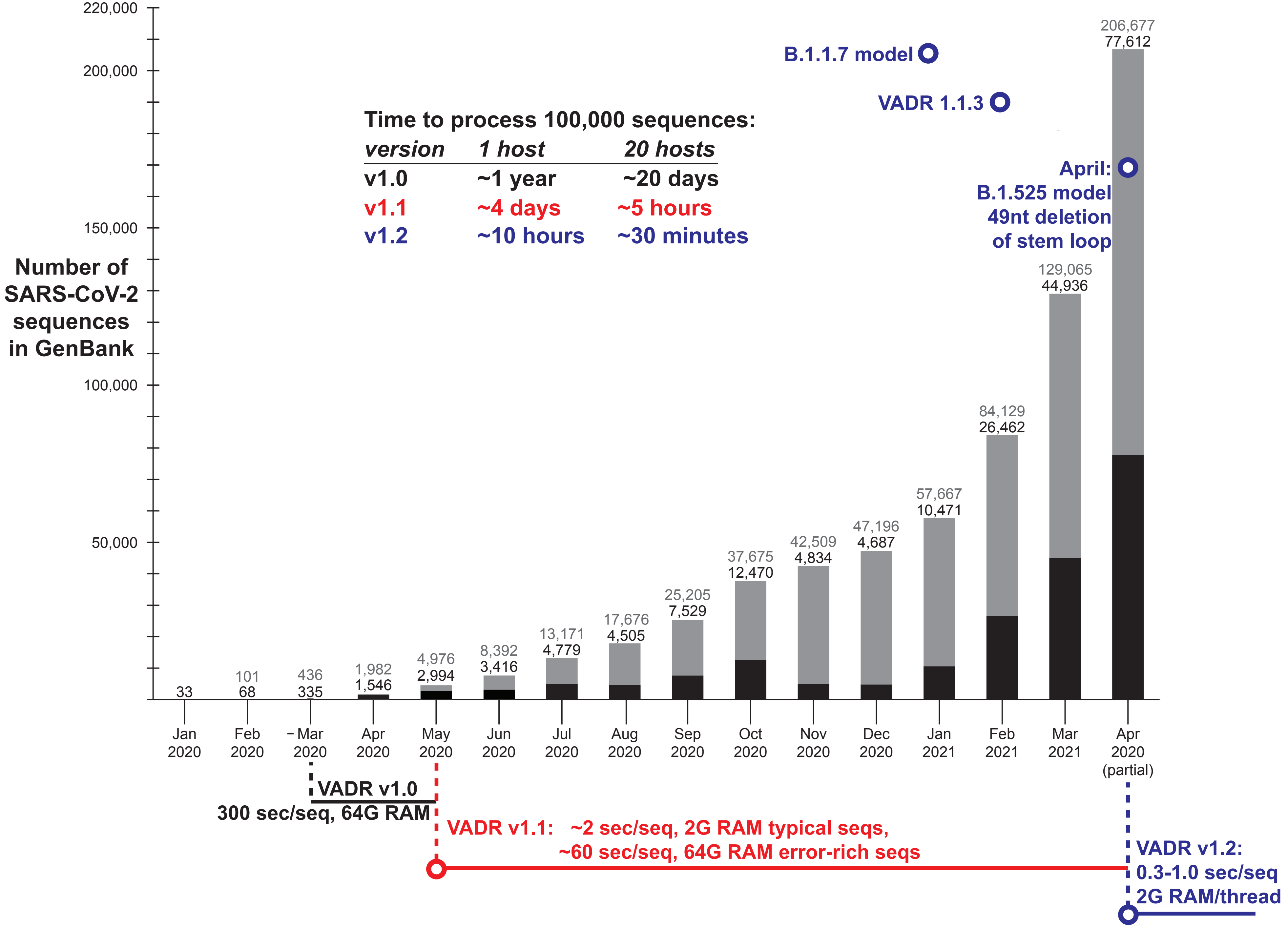


frameshift mutation; BLASTX alignment doesn't extend full length of CDS prediction (indf5pst, indf3pst alerts)

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

FRAMESHIFT EXAMPLE?

SARS-CoV-2 motivated speed-ups and other improvements



Time to process 100,000 sequences:

version	*1 host*	*20 hosts*
v1.0	~1 year	~20 days
v1.1	~4 days	~5 hours
v1.2	~10 hours	~30 minutes

Acknowledgements

NCBI - GenBank
Linda Yankie
Vince Calhoun
Ilene Mizrachi
Colleen Bollin
Beverly Underwood
Susan Schafer
Vasuki Gobu
Sergiy Gotvyanskyy
Alex Kotliaro

Alejandro Schaffer (now NCI)

NCBI - Virus
Rodney Brister
Eneida Hatcher
Ryan Connor
Lydia Fleischmann

NLM - leadership
Patti Brennan
Steve Sherry
Kim Pruitt
David Landsman

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

NIH NLM

NCBI
National
Center for
Biotechnology
Information