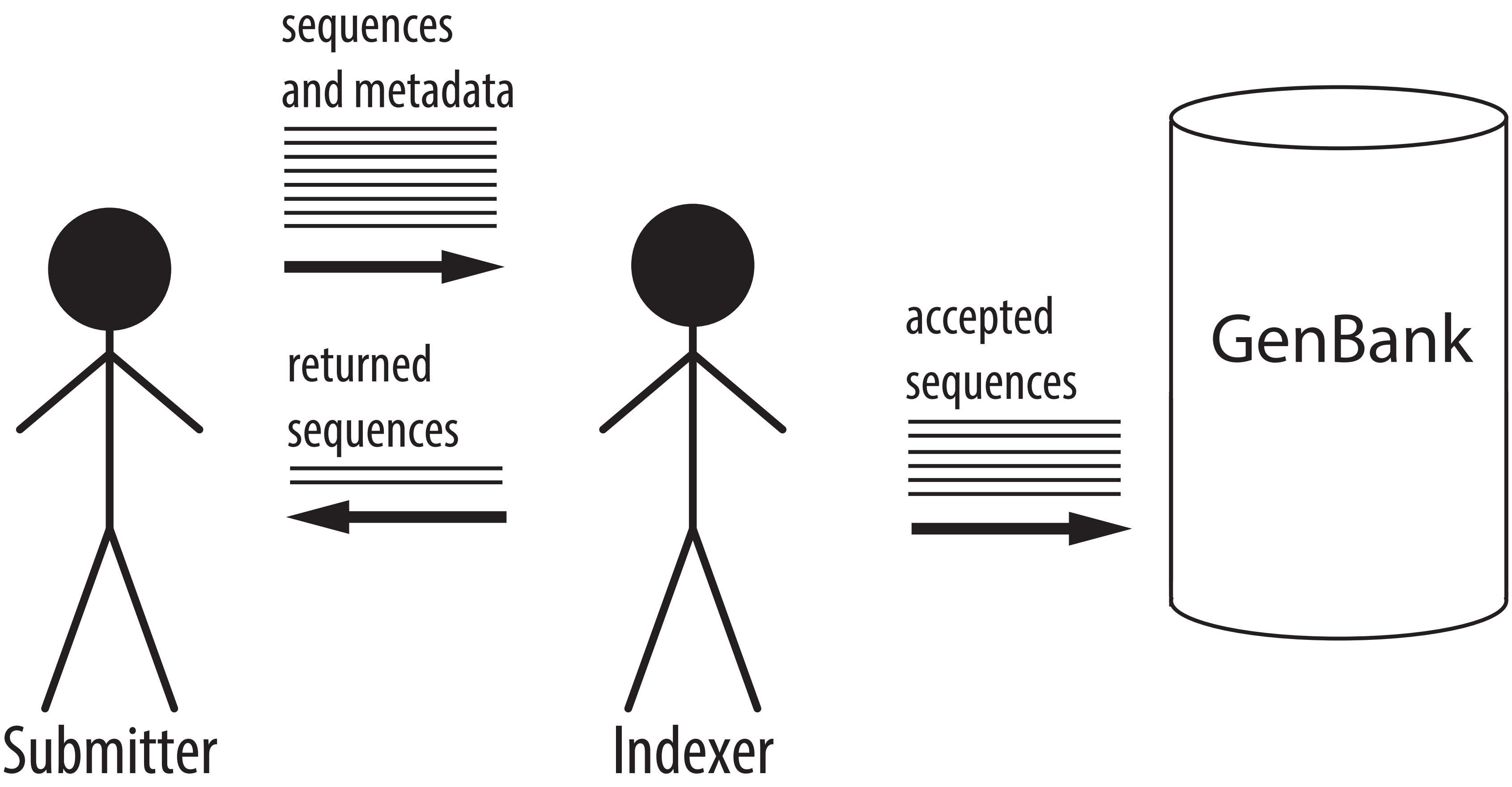
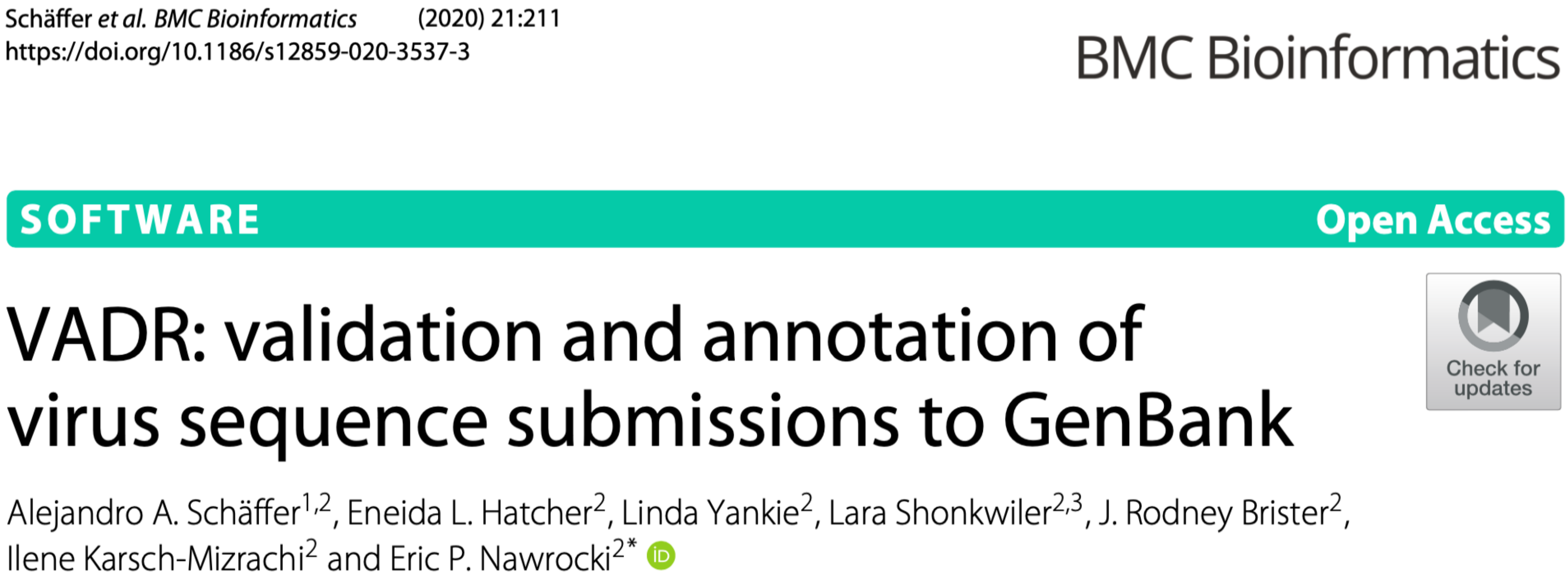
Automated validation and annotation of SARS-CoV-2 sequences for GenBank using VADR

Eric Nawrocki
Staff Scientist

Computational Biology Branch
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

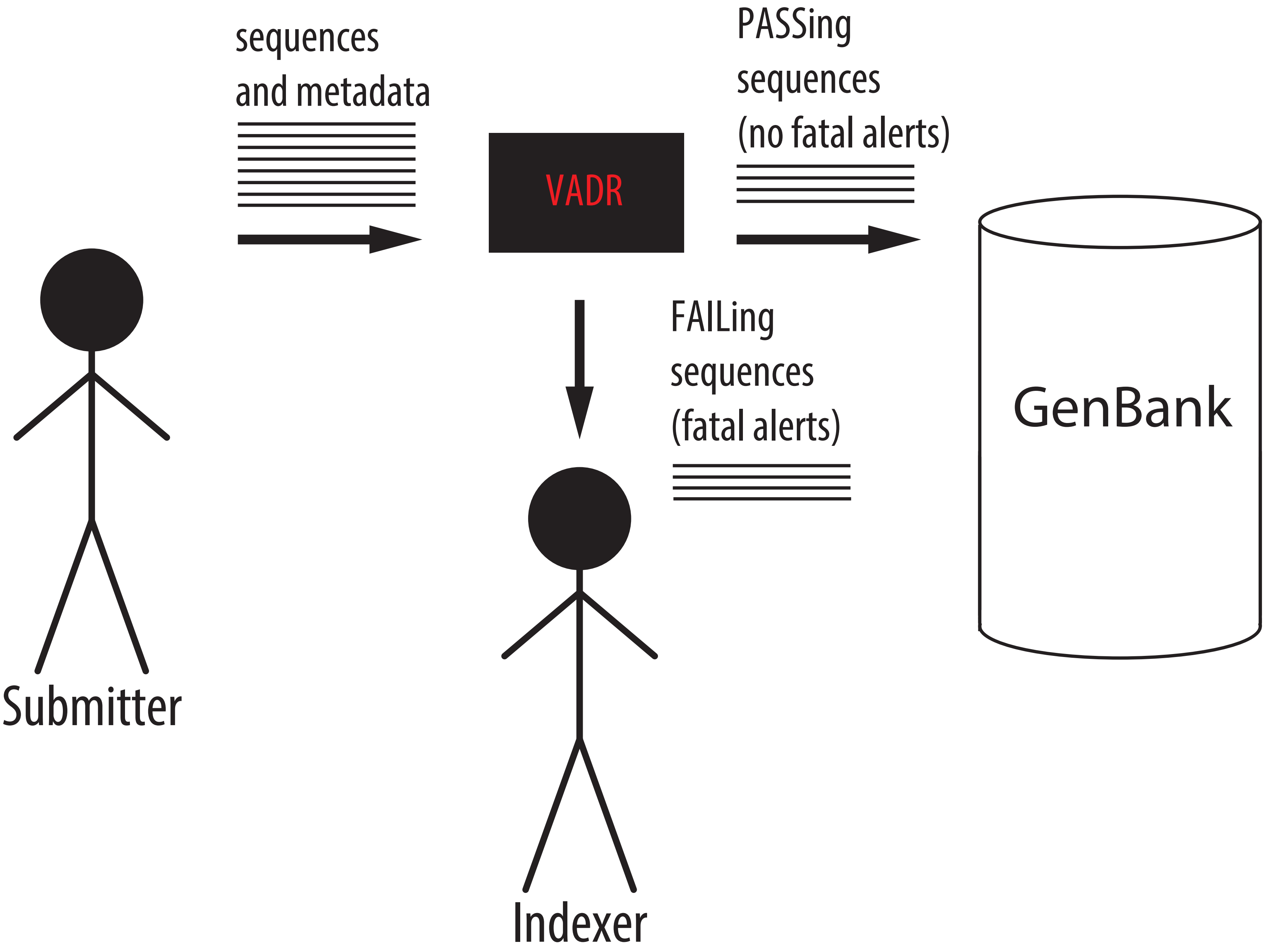
Alejandro A. Schäffer[1,2], Eneida L. Hatcher[2], Linda Yankie[2], Lara Shonkwiler[2,3], J. Rodney Brister[2], Ilene Karsch-Mizrachi[2] and Eric P. Nawrocki[2*]

- general tool for reference-based annotation of viral sequences

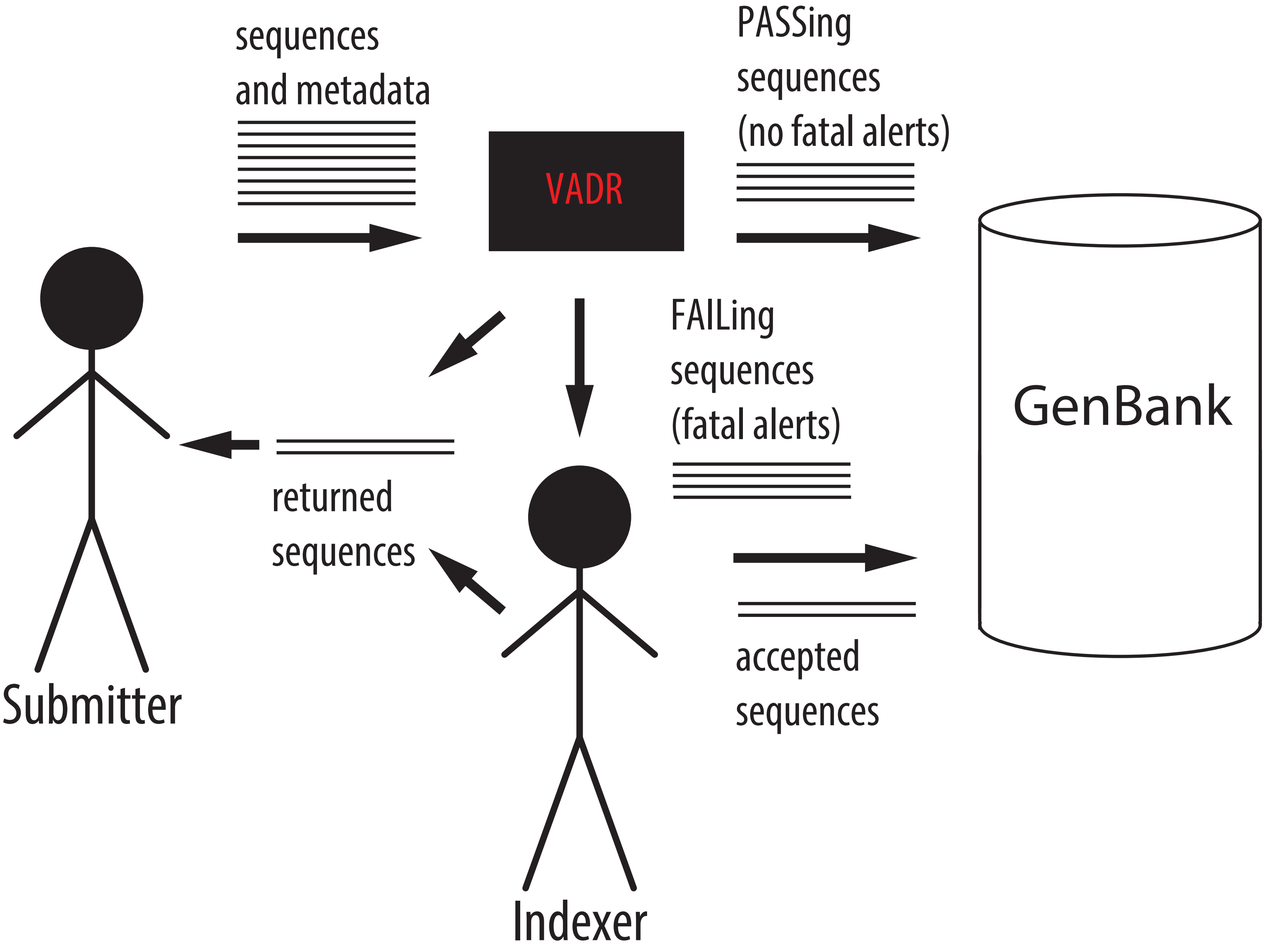
- used for Norovirus and Dengue virus submissions since 2018

- used for SARS-CoV-2 submissions since March 2020

VADR assists GenBank indexers:
Each sequence PASSes or FAILs



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



VADR proceeds over four stages to validate and annotate sequences

- For each sequence S:

1. **Classification**: compare S to all models to find best matching model M

2. **Coverage determination**: search M against S to find 'hits'

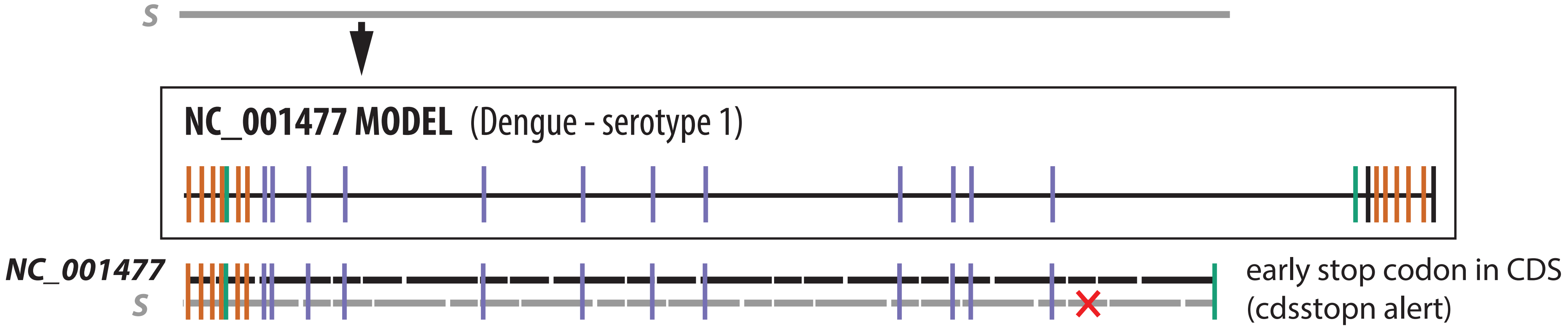
3. **Alignment**: align S to M and map features from M to S

4. **Protein validation**: compare predicted CDS in S to proteins from M using BLASTX

Different types of alerts are identified and reported at each stage

Stage 3: Alignment and feature mapping

Align each sequence to its best-matching model



S

NC_001477 MODEL (Dengue - serotype 1)

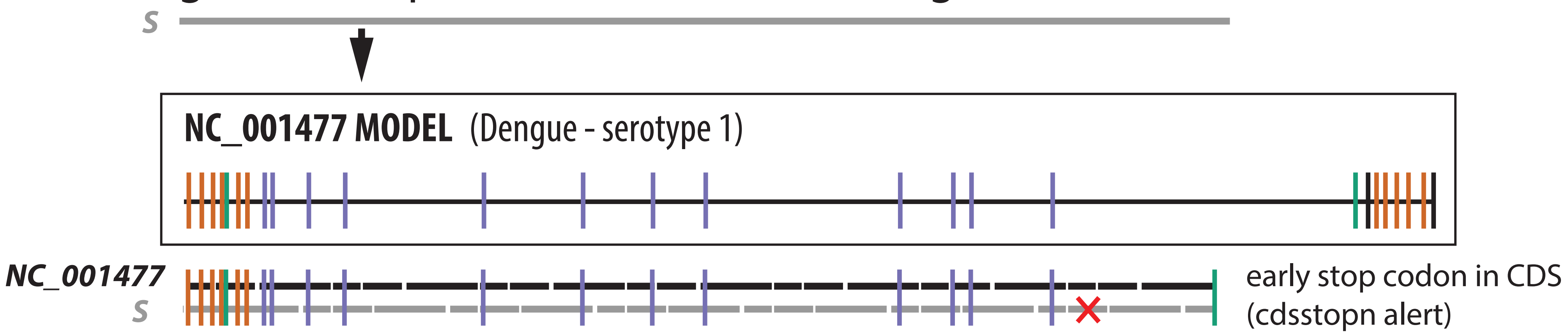
NC_001477

S

early stop codon in CDS
(cdsstopn alert)

Stage 3: Alignment and feature mapping

Align each sequence to its best-matching model



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

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

		#new	#cumulative
month	year	seqs	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

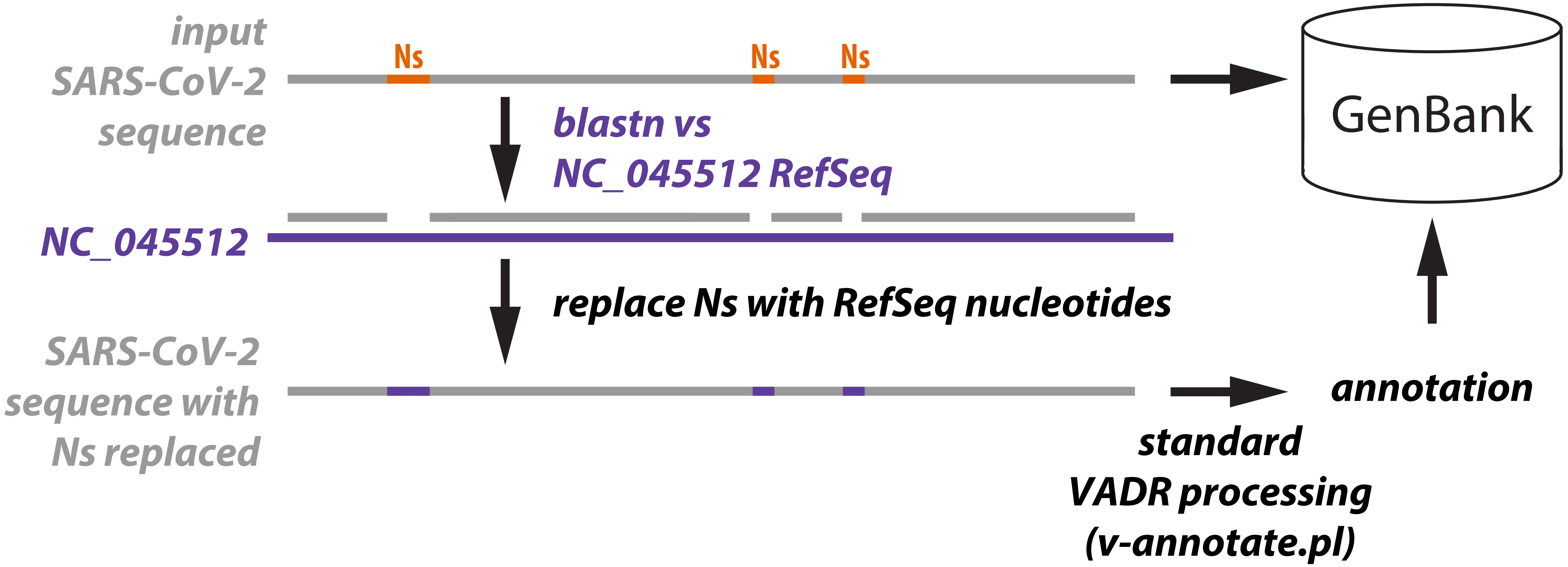
SARS-CoV-2 sequences differ from Norovirus and Dengue virus in several ways that impact VADR processing

	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%

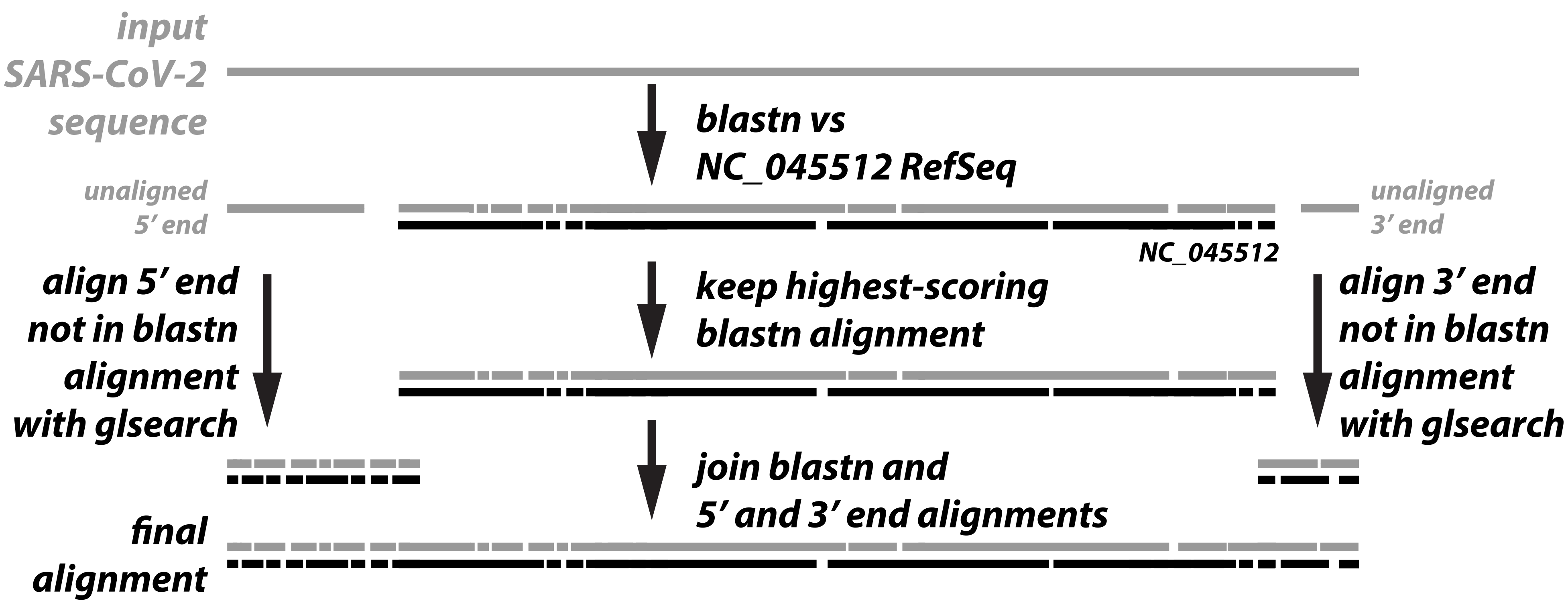
VADR v1.0 performance

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

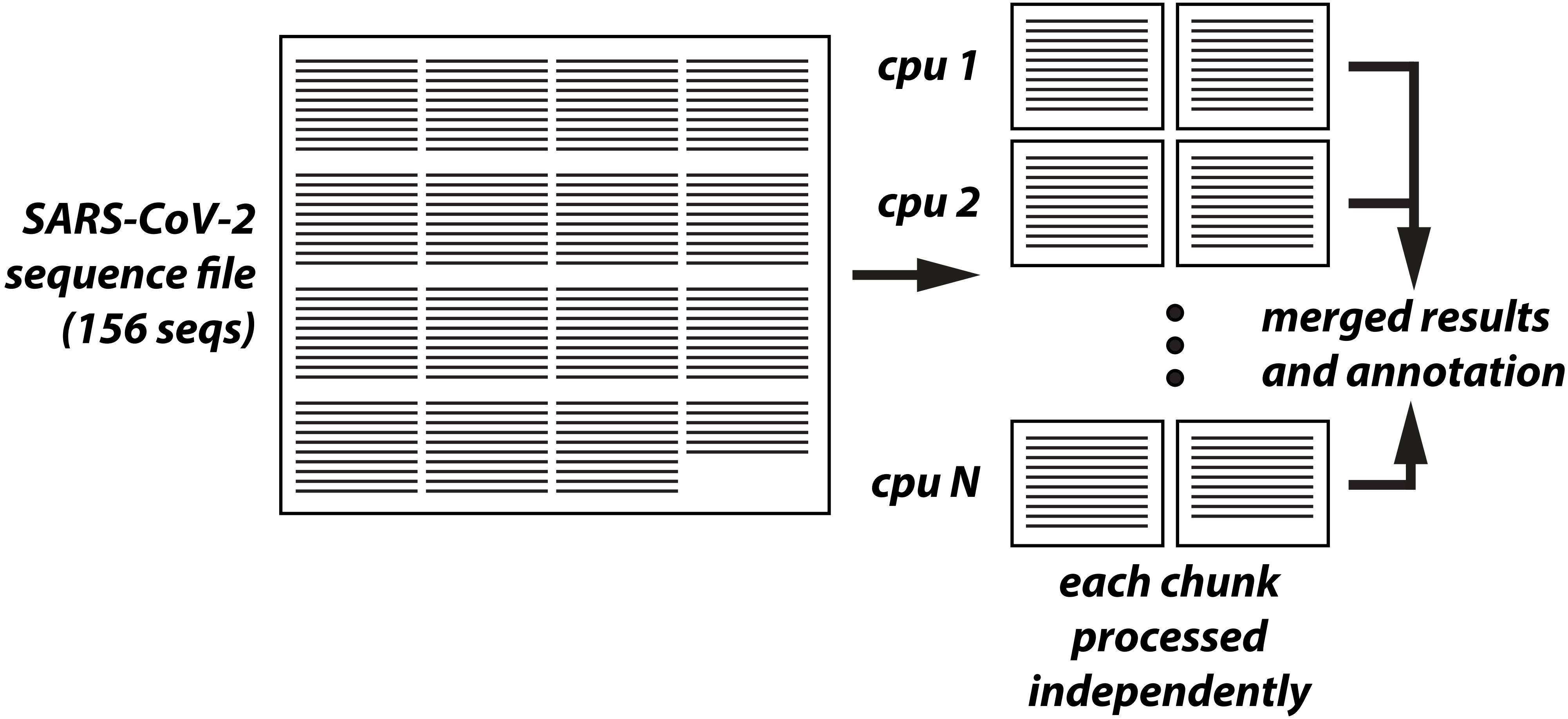

input SARS-CoV-2 sequence
Ns
Ns
Ns
GenBank
blastn vs NC_045512 RefSeq
NC_045512
replace Ns with RefSeq nucleotides
SARS-CoV-2 sequence with Ns replaced
annotation
standard VADR processing (v-annotate.pl)

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

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

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

	seeded	N				secs	hours	speedup
VADR	align-	replace-		#	required	per	per 100K	vs
version	ment?	ment?	glsearch?	cpus	RAM	seq	seqs	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

	seeded	N				secs	hours	speedup
VADR	align-	replace-		#	required	per	per 100K	vs
version	ment?	ment?	glsearch?	cpus	RAM	seq	seqs	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

		#new	#cumulative
month	year	seqs	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)

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.

NIH NLM

NCBI National Center for Biotechnology Information