INSTITUTO NACIONAL DE SALUD

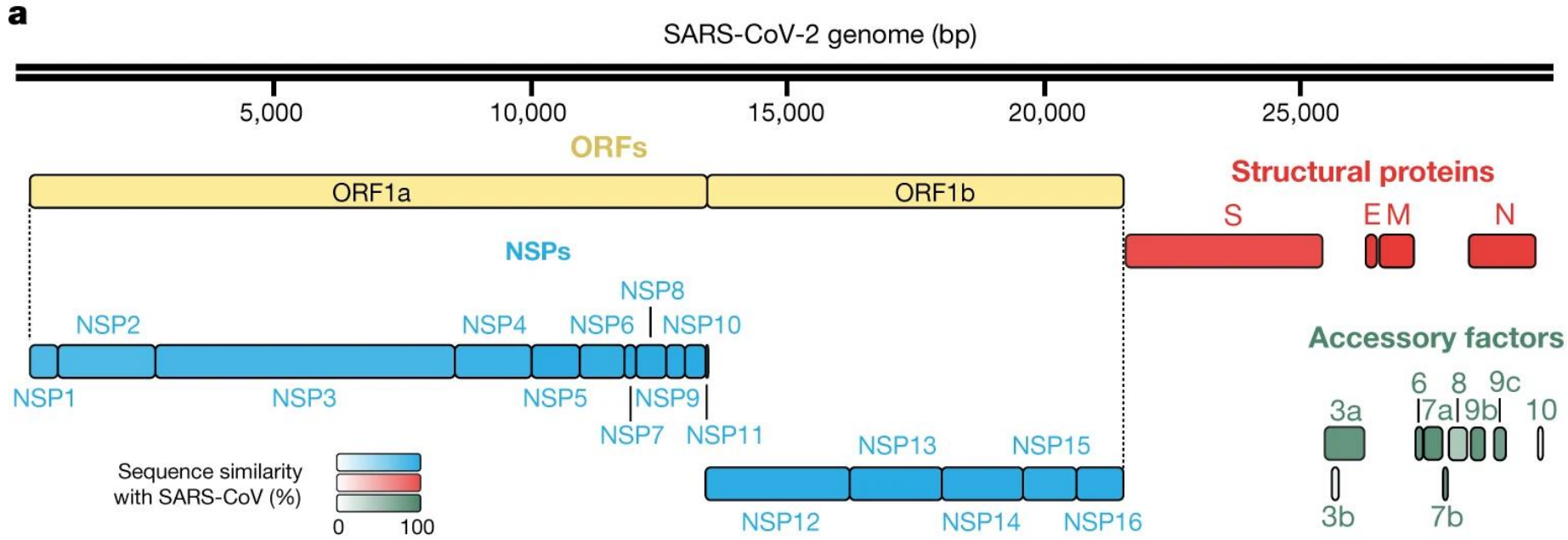
Tutorial anotación de genomas

Adrián C. Rodríguez Ararat.
Asistente de investigación grupo Natura Universidad Icesi
PhDc Ciencias Biomédicas Universidad de Valle
acrodriguez@icesi.edu.co

UNIVERSIDAD ICESI

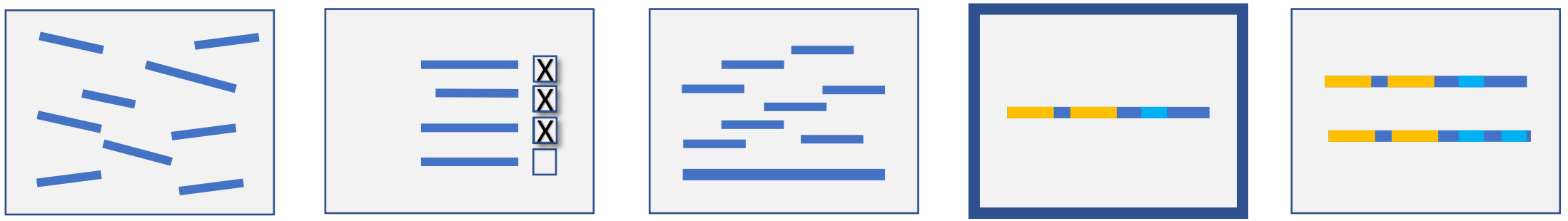
Que es anotación de genomas

- Es el proceso de identificar la ubicación de genes y descubrir su función, así como la anotación de otras características genéticas de una secuencia genómica cruda.



Gordon, D.E., Jang, G.M., Bouhaddou, M. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583,** 459–468 (2020). https://doi.org/10.1038/s41586-020-2286-9

Como es anotado un genoma



Secuenciación | Control calidad | Ensamble | **Anotación** | Comparación

- Selección genoma
- Anotación automática (Maker pipeline, Prokka, …)
- Anotación manual (Apollo, Gbrowse, Galaxy)

Herramientas para anotación de genomas

- Identificación de genes ARN
  - RNAmmer (HHMM – 5S rRNA – European rRNA db Project – CBS server)

    Especies principales de ARN de organismos en diferentes reinos.
  - tRNAScanSE (genes tRNA genomas completos ampliamente aceptado)
- Encontrar genes/ORFs
  - Prodigal (genes procariotas, sitio de inicio de traducción – Genbank - SIB)
  - GeneMark (Procariotas, eucariotas, virus, fagos… HHMM, Heurístico, - > MAKER2, BRAKER1 = RNAseq based y BRAKER2 = protein based)
  - MetageneAnnotator

VADR: Viral Annotation DefineR

- Anotación basada en referencia de secuencias virales
- Realiza alineamiento de secuencias global contra RefSeq o perfil
- Creado para dengue virus y norovirus -> SARS-CoV-2 (marzo 2020)

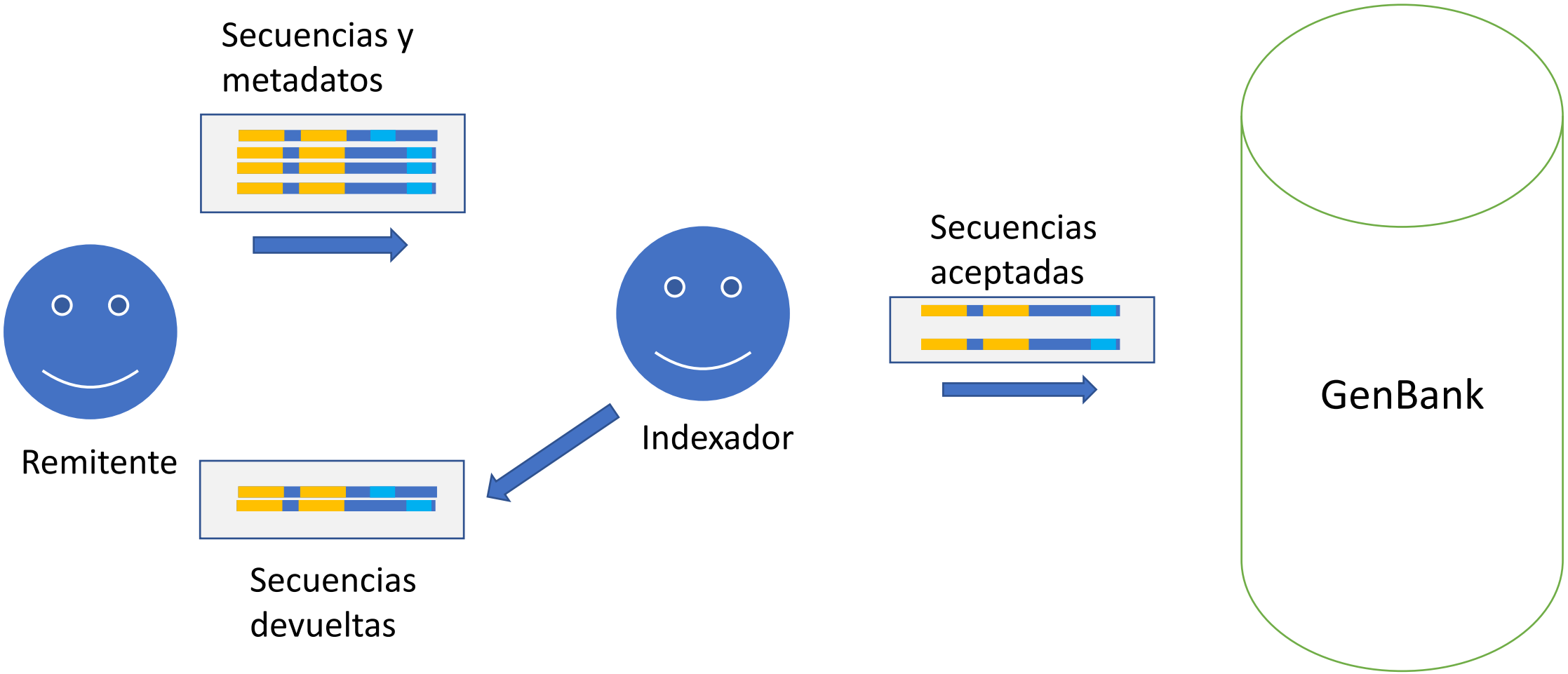
BMC Bioinformatics

BMC Bioinformatics. 2020; 21: 211.
Published online 2020 May 24. doi: 10.1186/s12859-020-3537-3
PMCID: PMC7245624
PMID: 32448124

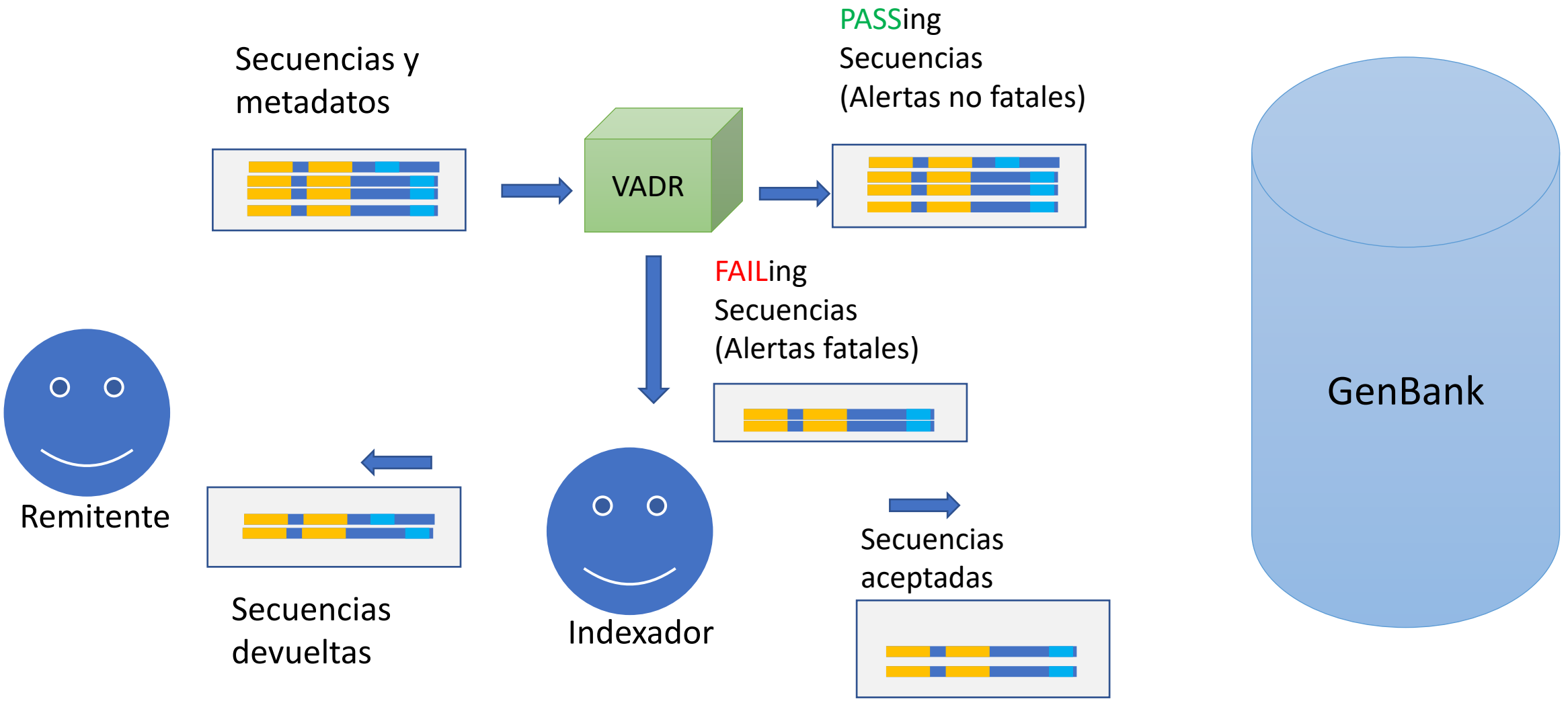
VADR: validation and annotation of virus sequence submissions to GenBank

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]

Manejo de secuencias sometidas a GenBank Indexers



VADR asiste GenBank Indexers:



VARD anota usando el mejor modelo

- Para cada secuencia S:

    1- Clasificación (S comparación con modelos M)

    2- Cobertura (busca S contra M determina ‘hits’)

    3- Alineamiento (Alineamiento S a M y mapeo características de M a S)

    4- validación mediante proteína (compara pCDS en S a proteínas de M usando BLASTX).

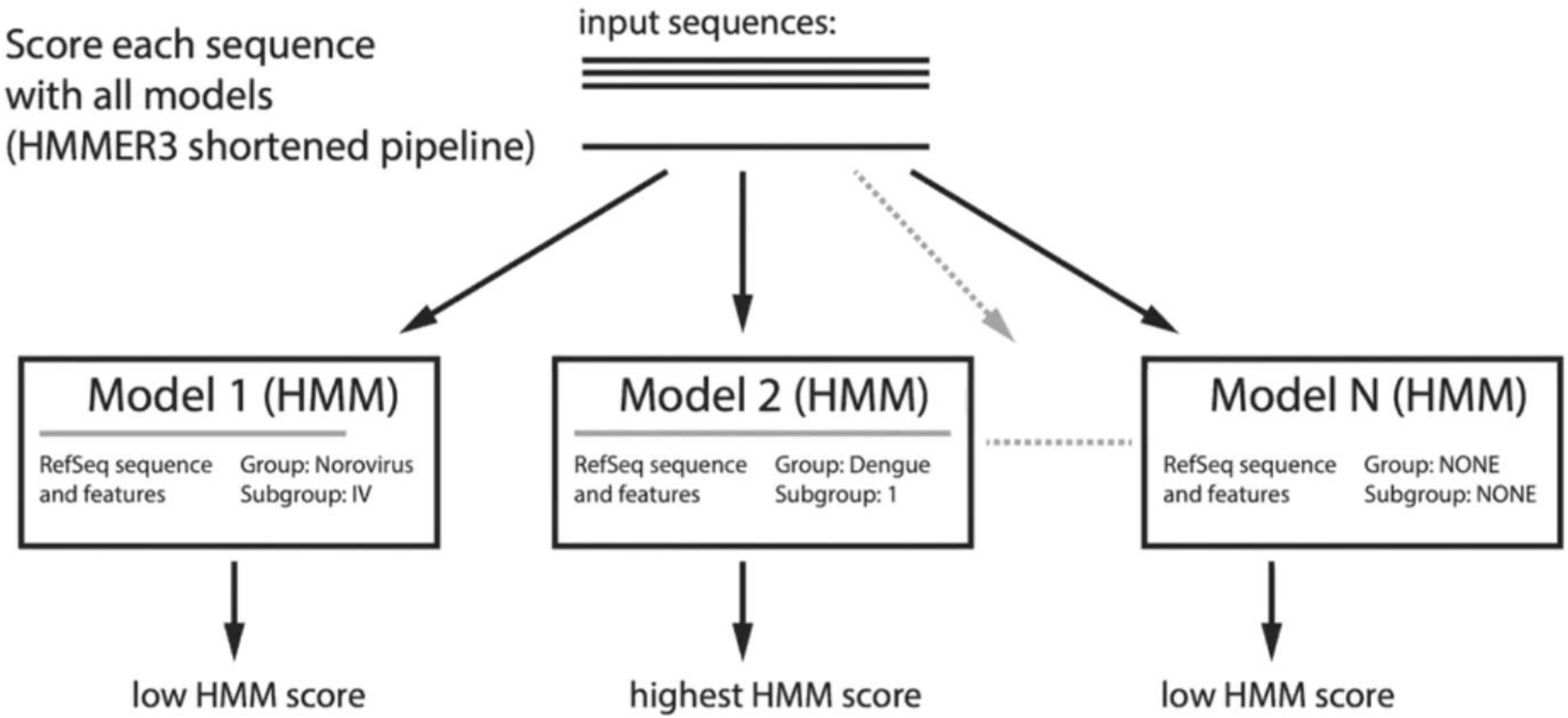
VADR – Brinda alertas en cada proceso.

COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research

Tool	Advancing SARS-CoV-2 research by
Detection and annotation	
PriSeT	computing SARS-CoV-2 specific primers for RT-PCR tests
CoVPipe	reproducible, reliable and fast analysis of NGS data
poreCov	reducing time-consuming bioinformatic bottlenecks in processi
VADR	validation and annotation of SARS-CoV-2 sequences
V-Pipe	reproducible NGS-based, end-to-end analysis of genomic divers
Haploflow	detection and full-length reconstruction of multi-strain infectior
VIRify	identifying viruses in clinical samples
VBRC genome analysis tools	visualizing differences between coronavirus sequences at differe
VIRULIGN	fast, codon-correct multiple sequence alignment and annotatio
Rfam COVID-19	annotating structured RNAs in coronavirus sequences and predi
UniProt COVID-19	providing latest knowledge on proteins relevant to the disease f
Pfam	protein detection and annotation for outbreak tracking and stuc

Franziska Hufsky, *et al.,* Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research, *Briefings in Bioinformatics*, Volume 22, Issue 2, March 2021, Pages 642–663, https://doi.org/10.1093/bib/bbaa232

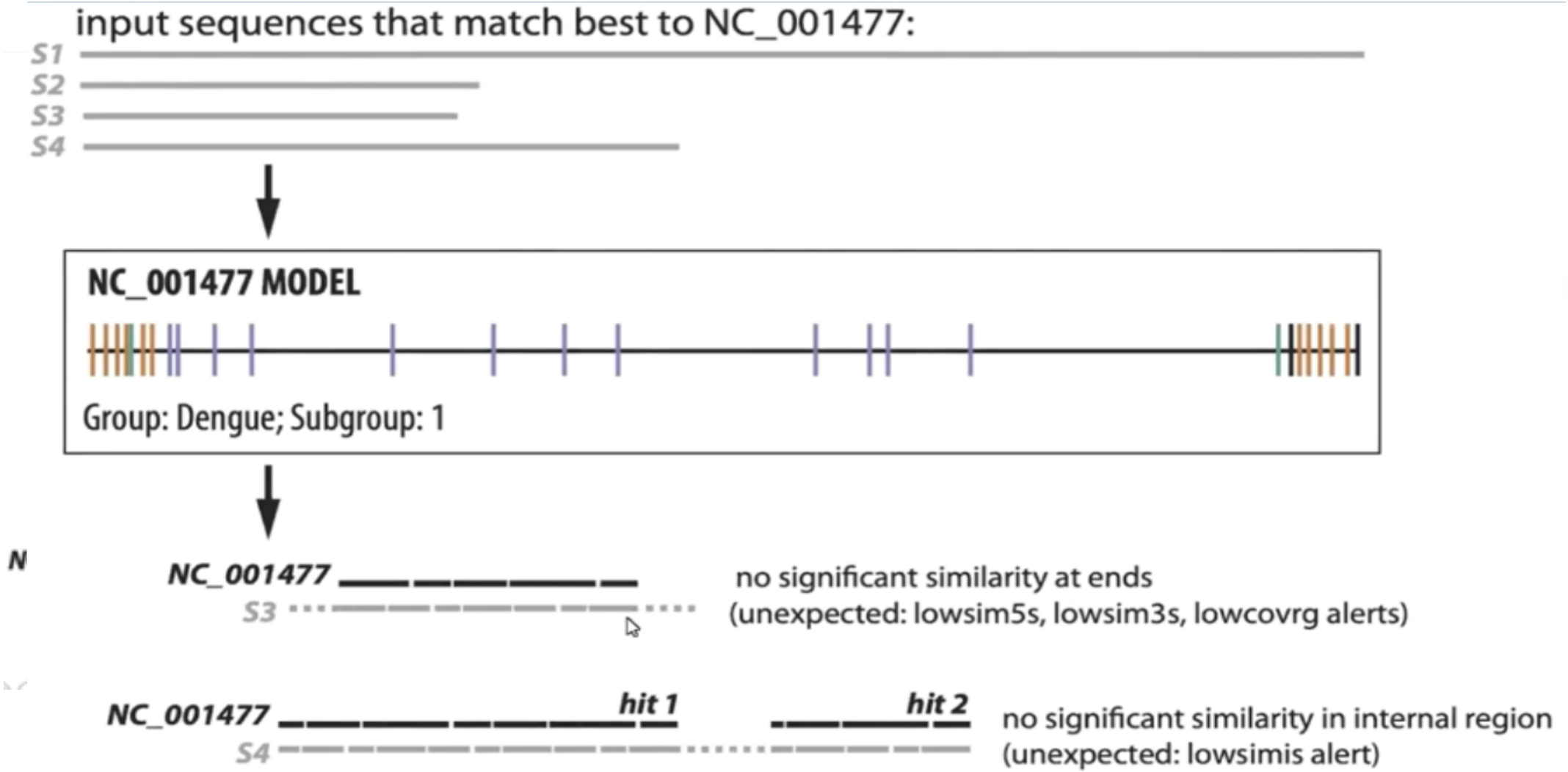
1- Clasificación (Alertas)



1- Clasificación (Alertas)

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

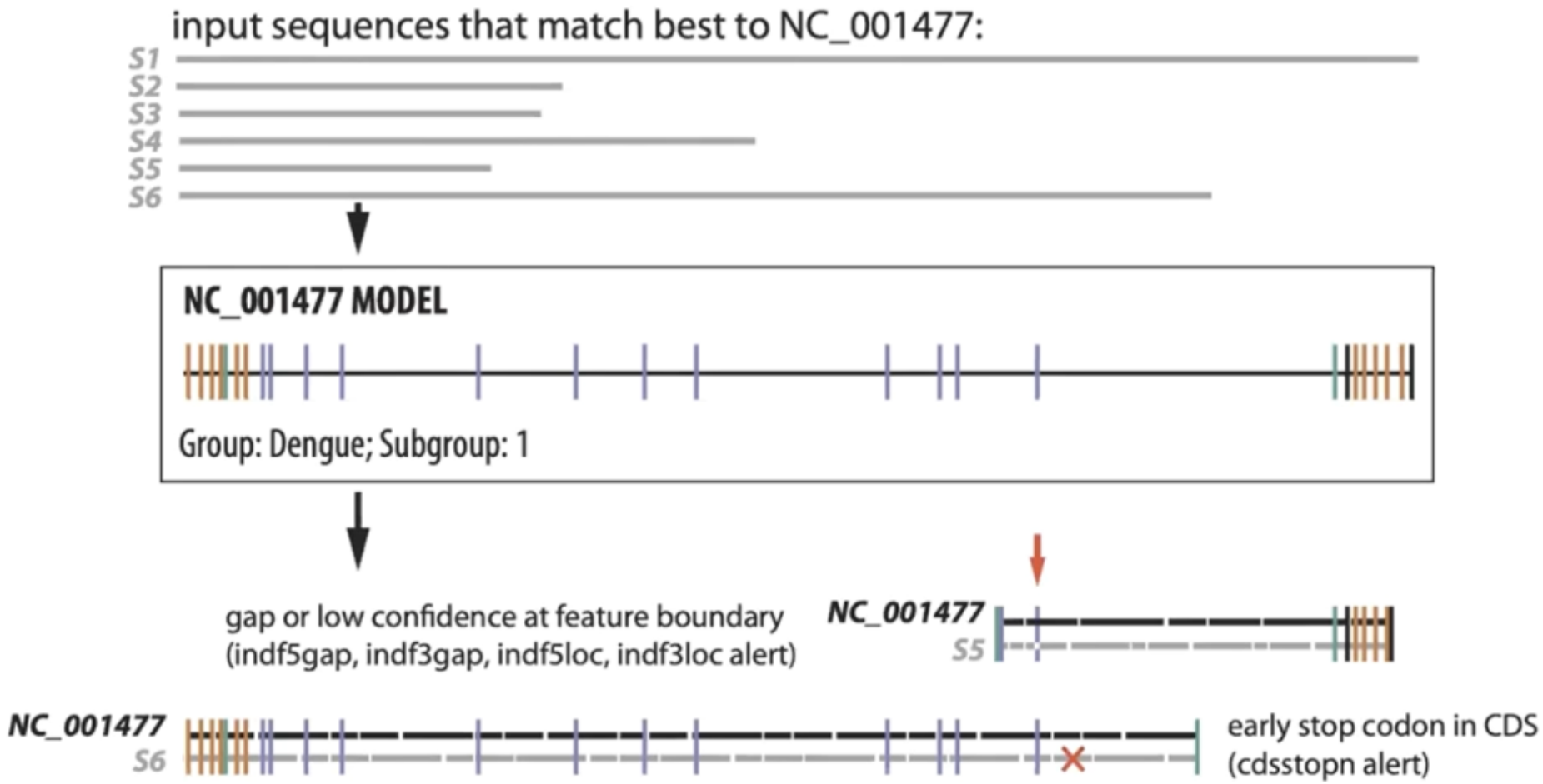
1- Coverturas (Alertas)



1- Coverturas (Alertas)

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

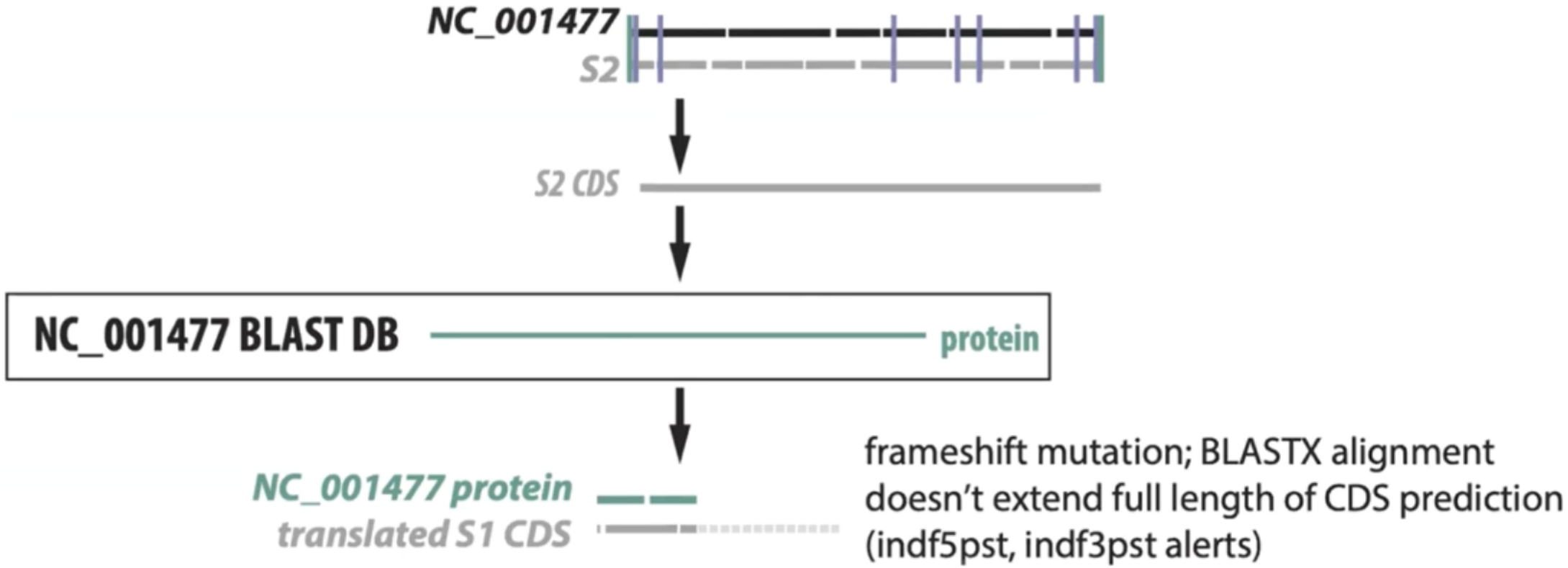
3- Alineamiento (Alertas)



3- Alineamiento (Alertas)

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

1- Validación (Alertas)



1- Validación (Alertas)

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

Limitaciones

- Hace inferencias en el espacio de nucleótidos, no en el espacio de proteínas
- El modelo (RefSeq) tiene que ser representativo
- Secuencias divergentes, codones de parada nuevos, etc. son problemáticos
- Capacidad de computo >64GB para algunas secuencias de SASR-CoV-2

Referencias

- VADR https://github.com/ncbi/vadr/blob/master/documentation/annotate.md#examplebasic
- VADR SARS-CoV2 annotation https://github.com/ncbi/vadr/wiki/Coronavirus-annotation
- VADR - Biowulf https://hpc.nih.gov/apps/VADR.html
- Schäffer, A. A. *et al.* VADR: validation and annotation of virus sequence submissions to GenBank. *BMC Bioinformatics* **21**, (2020).
- SPHERES Webinar, NCBI Submission & Annotation for SARS-CoV-2. https://www.youtube.com/watch?v=Gr20D_NcWPg&t=389s