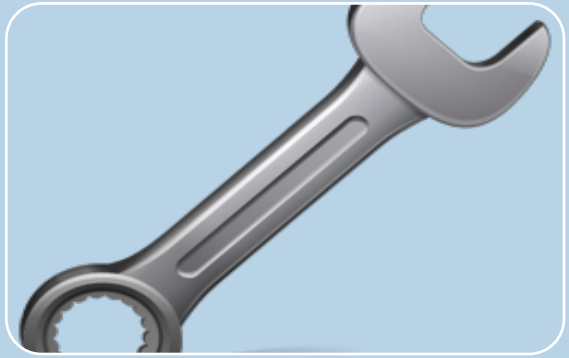
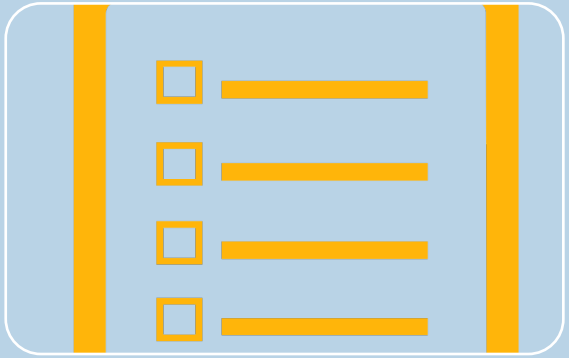
SARS-CoV-2 sequence submission & annotation

The National Center for Biotechnology Information (NCBI)
Eric Nawrocki, Linda Yankie & Yuriy Skripchenko
Oct. 21st 2020

Overview



Introduce helpful tool for SARS-CoV-2 annotation: VADR!



Share submission trends & tips for fast, error-free submission

Continue dialogue on NCBI submission with this community

VADR: Viral Annotation DefineR

- General tool for reference-based annotation of viral sequences
- Aligns sequences globally with respect to a RefSeq or profile
- Originally designed for dengue virus and norovirus but used for SARS-CoV-2 since March 2020

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

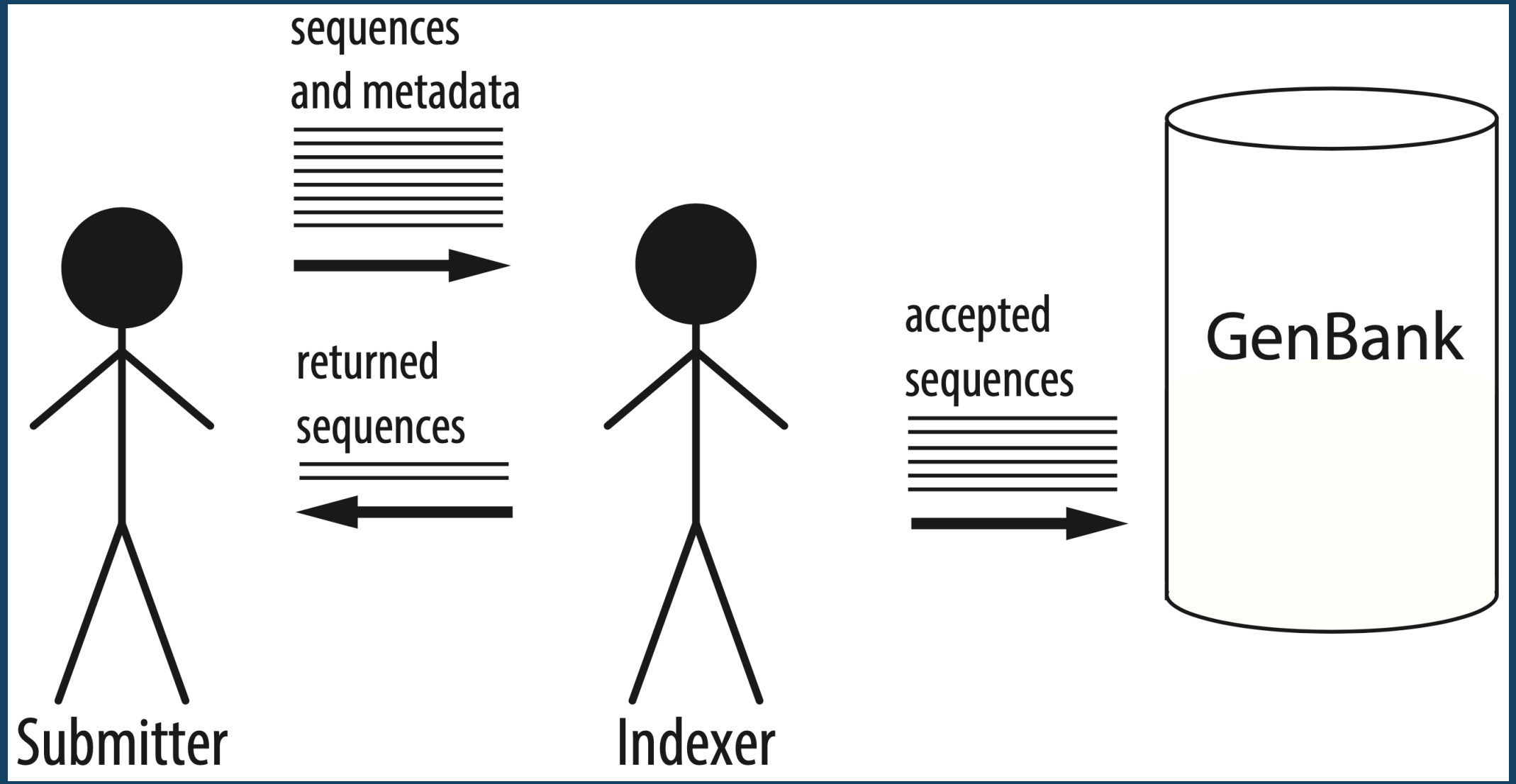
BMC Bioinformatics

SOFTWARE Open Access

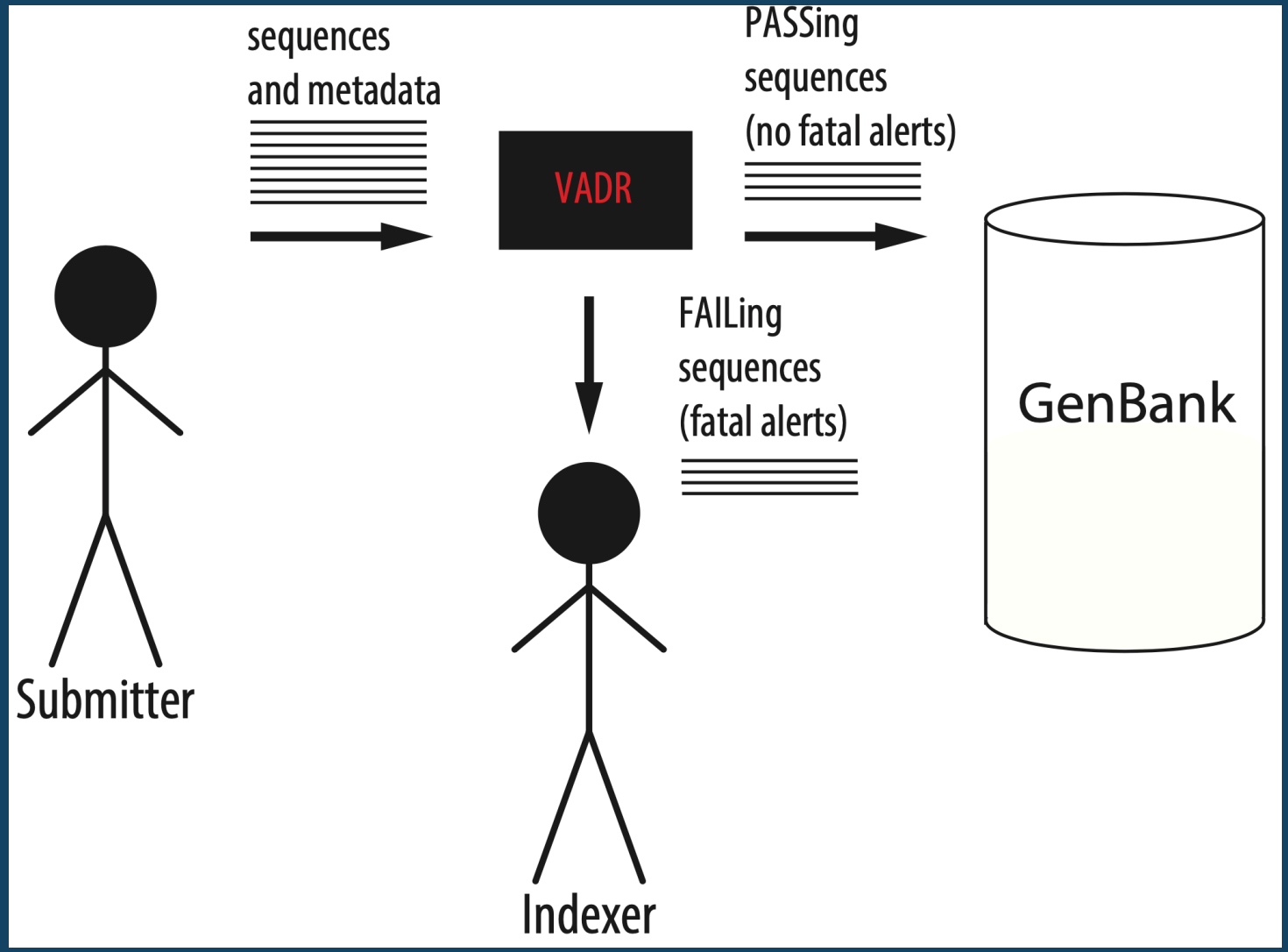
VADR: validation and annotation of virus sequence submissions to GenBank

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*]

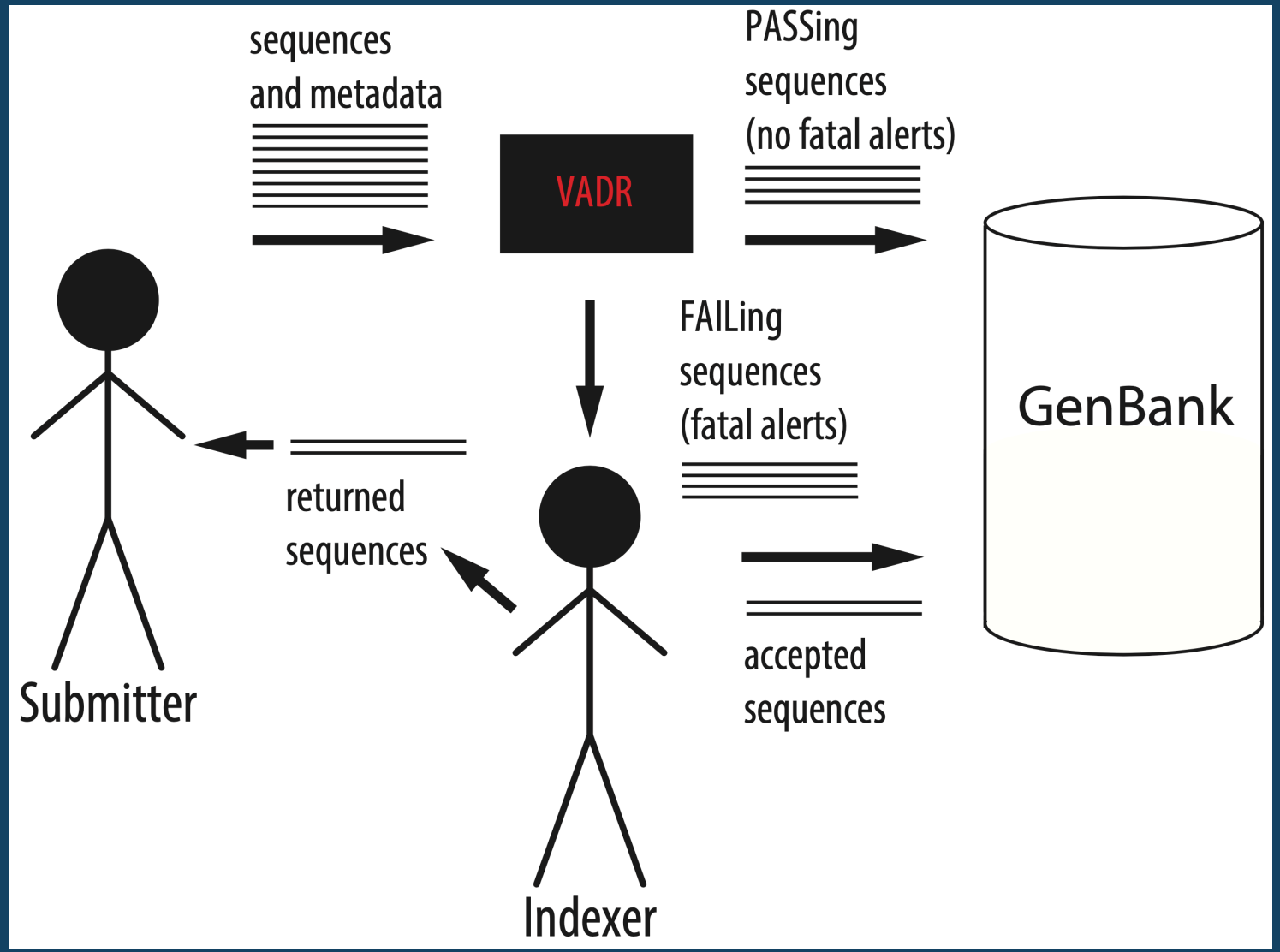
GenBank indexers handle incoming sequence submissions



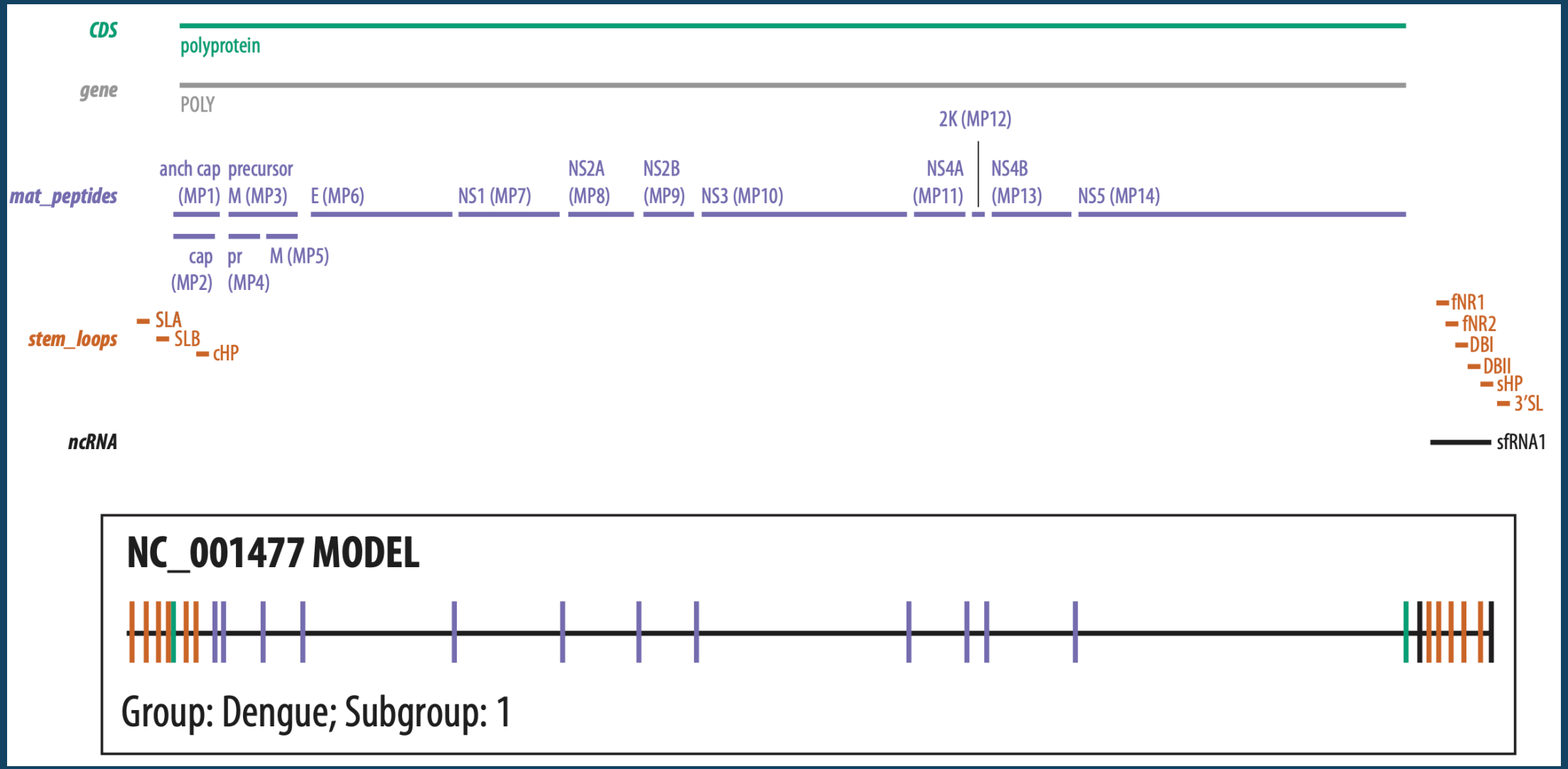
VADR assists GenBank indexers: Each sequence PASSes or FAILs



Indexers decide fate of FAILed sequences



VADR builds homology model (covariance model) of a RefSeq & stores feature info

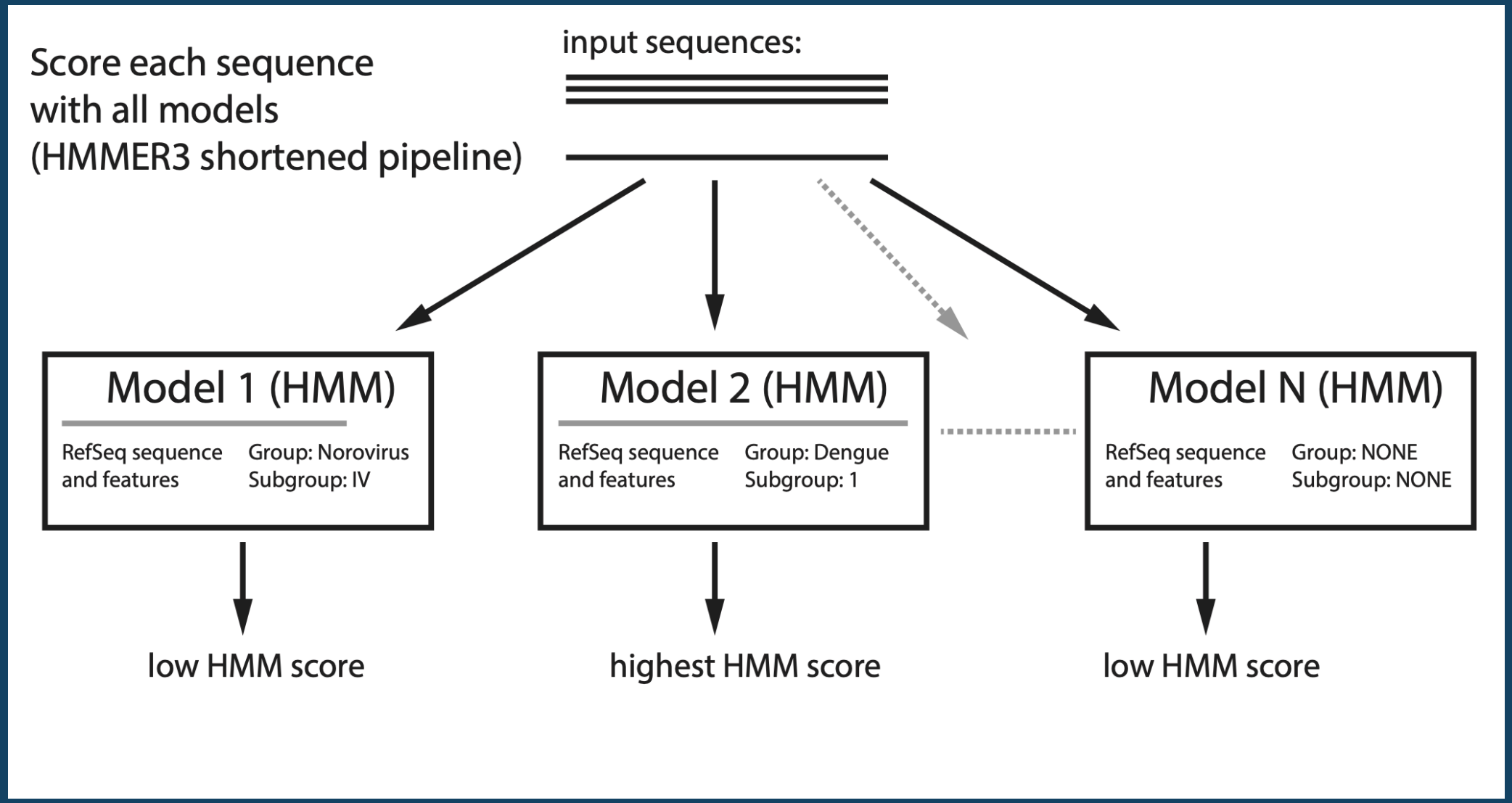


VADR annotates each sequence with features from its best matching model

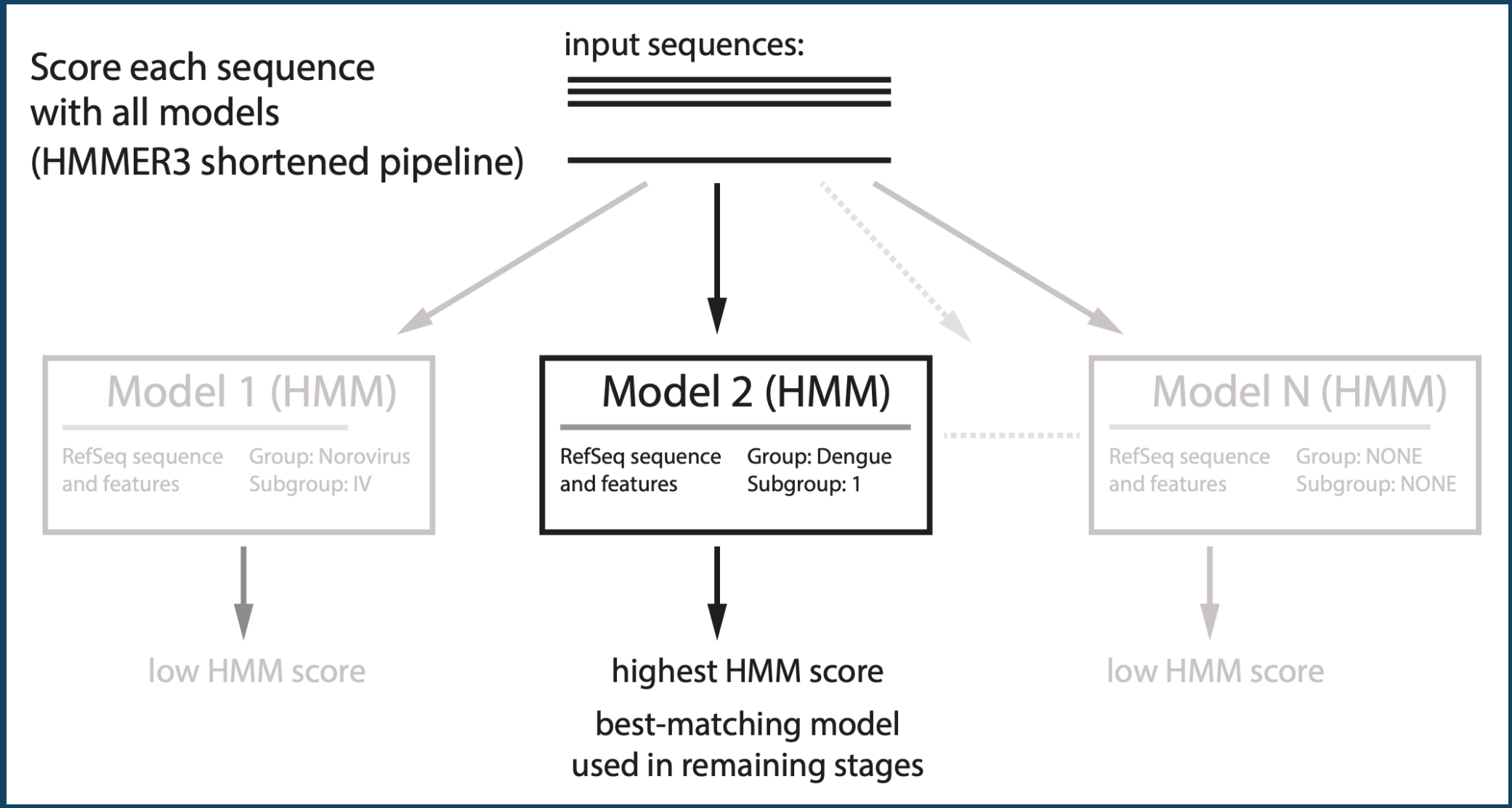
- For each sequence S:
  1. **Classification**: compare S to all models to find best matching model M
  2. **Coverage determination**: search M against S to find 'hits'
  3. **Alignment**: align S to M and map features from M to S
  4. **Protein validation**: compare predicted CDS in S to proteins from M using BLASTX

Different types of alerts are identified and reported at each stage

Stage 1: Classification



Stage 1: Classification

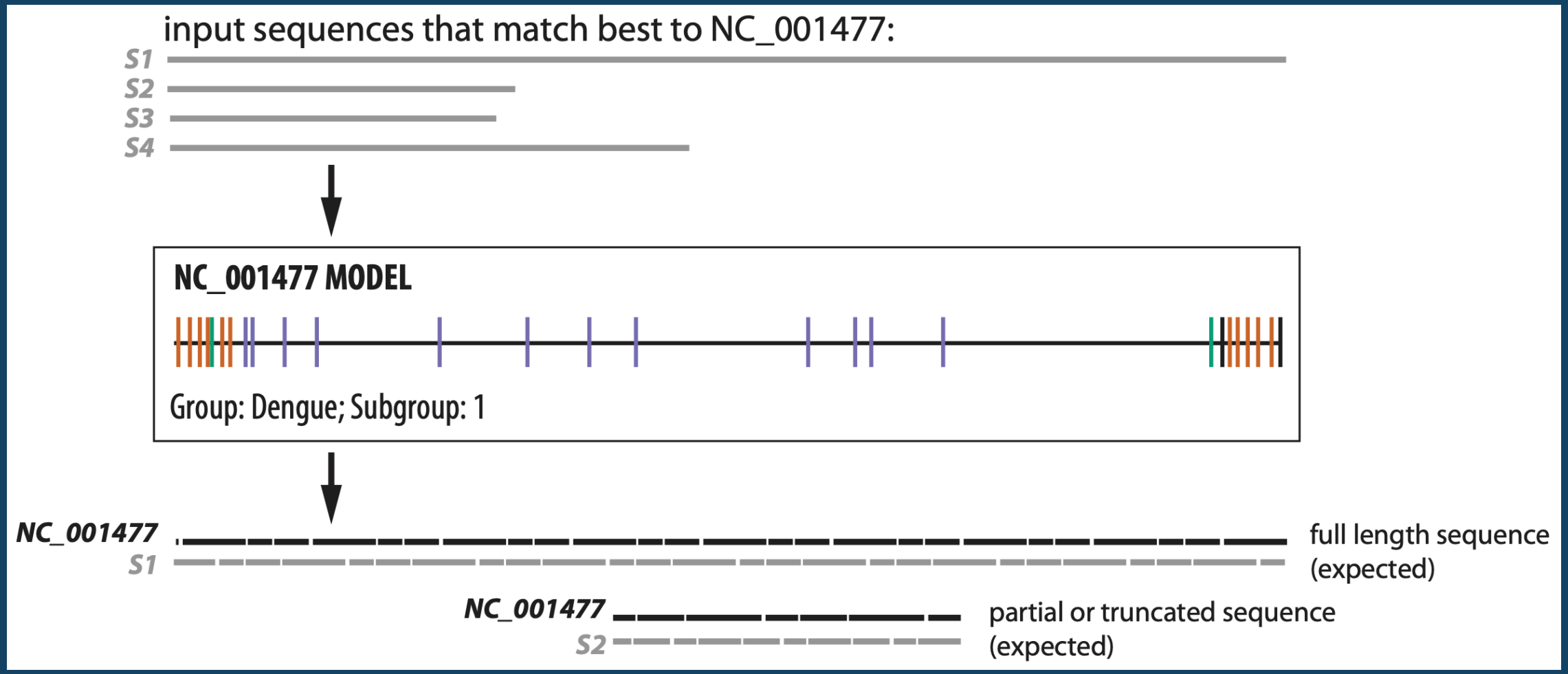


Alerts reported in classification stage

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

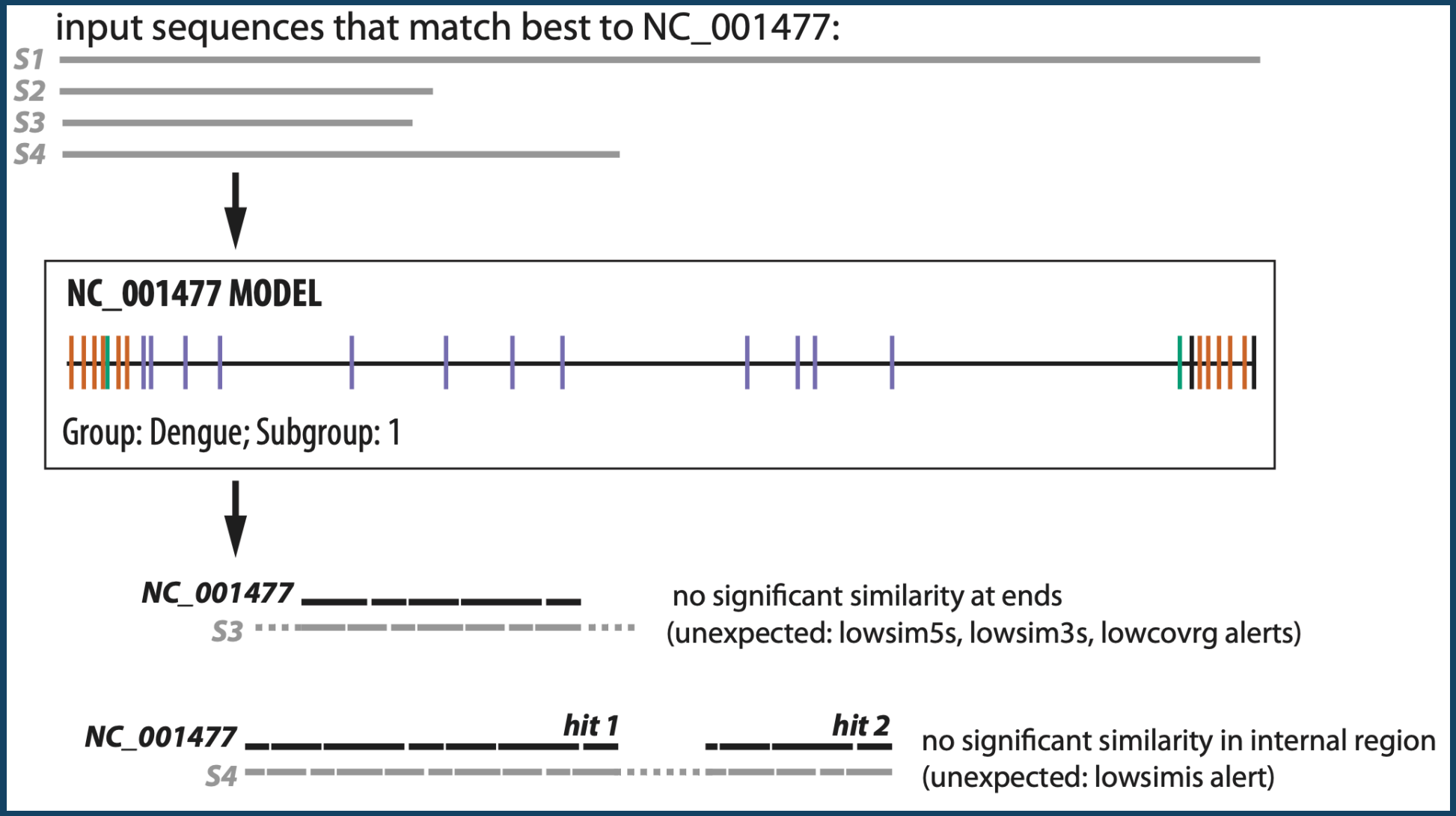
Stage 2: Coverage determination

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



Stage 2: Coverage determination

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

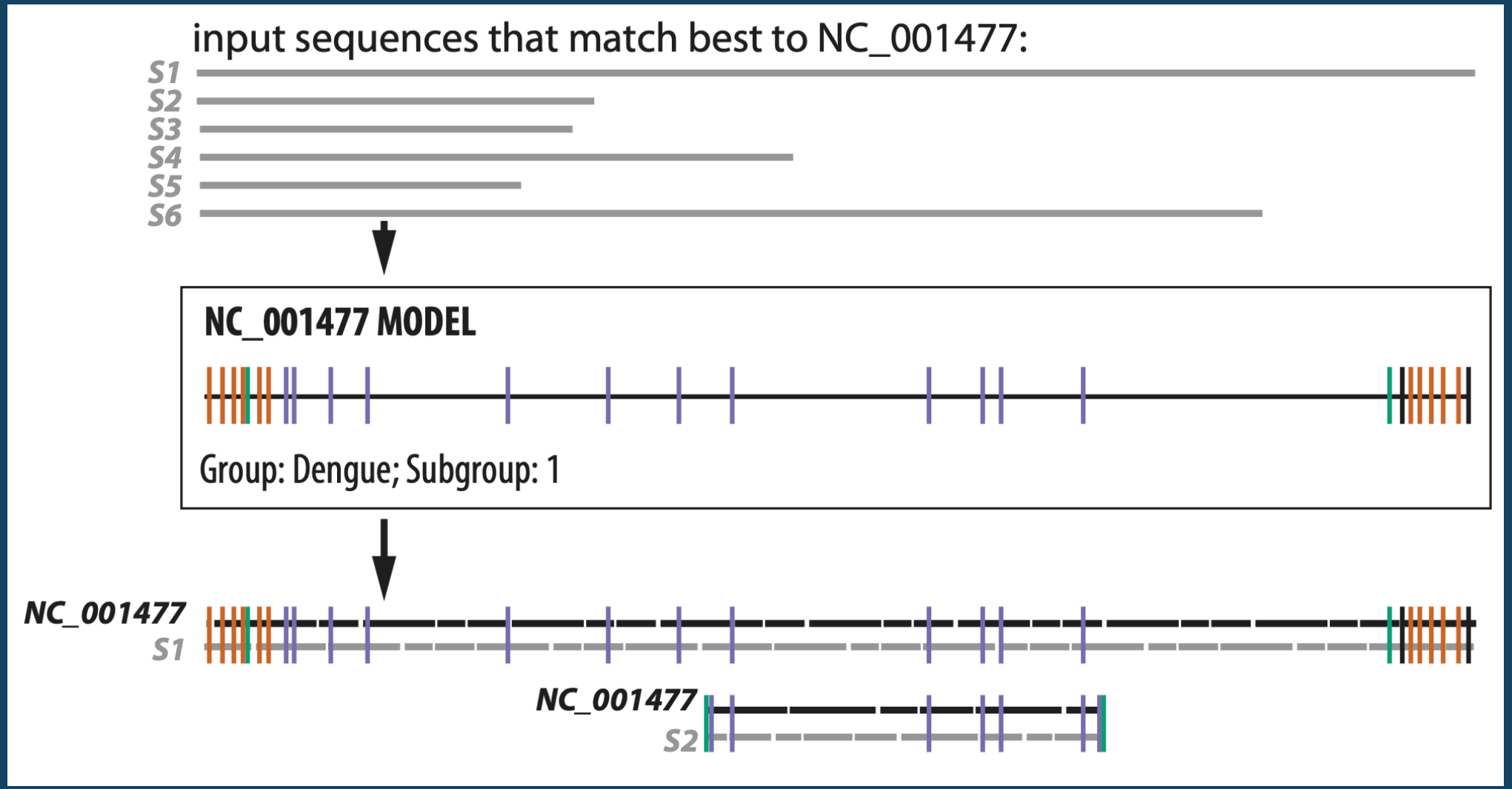


Alerts reported in coverage determination stage

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

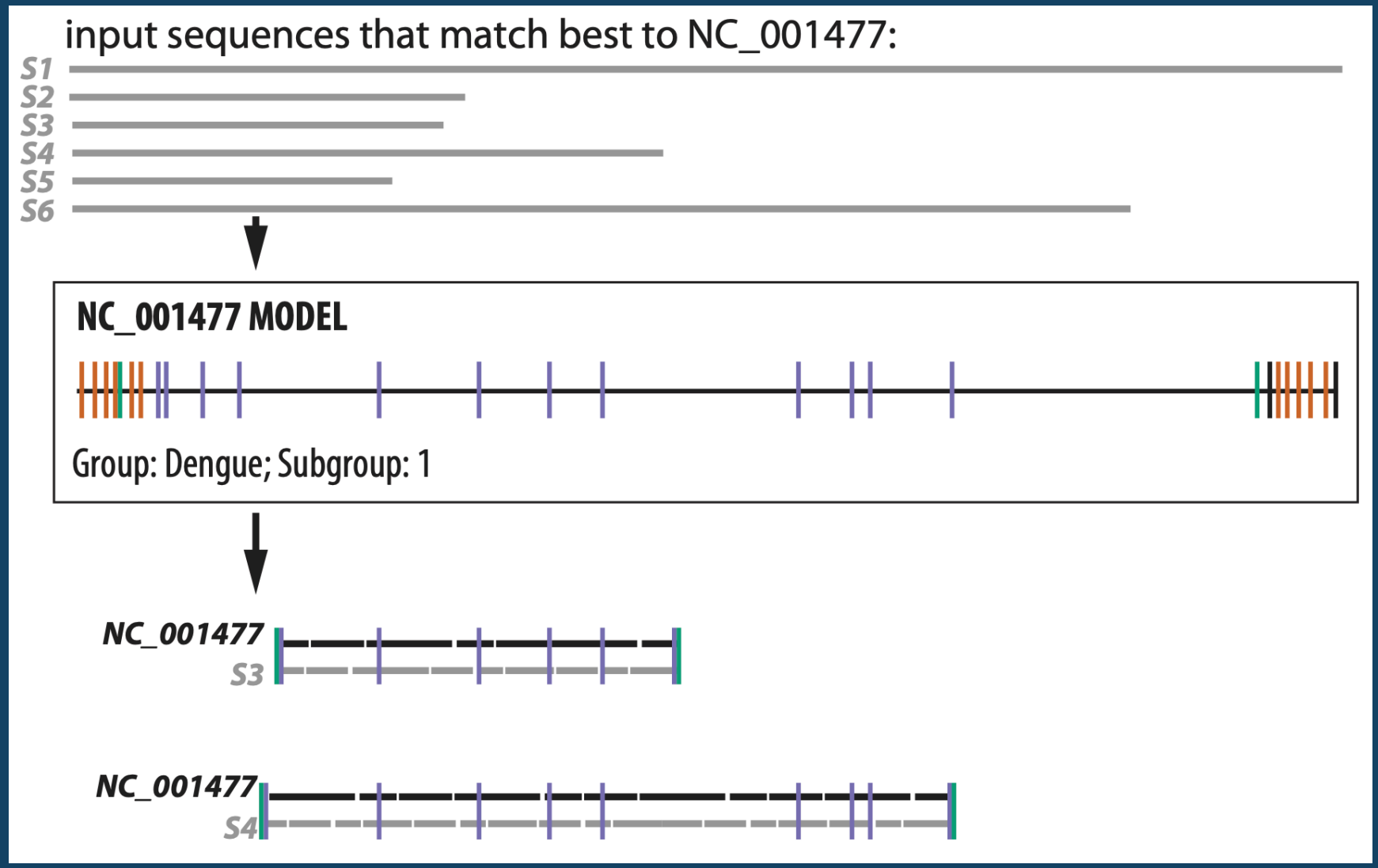
Stage 3: Alignment & feature mapping

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



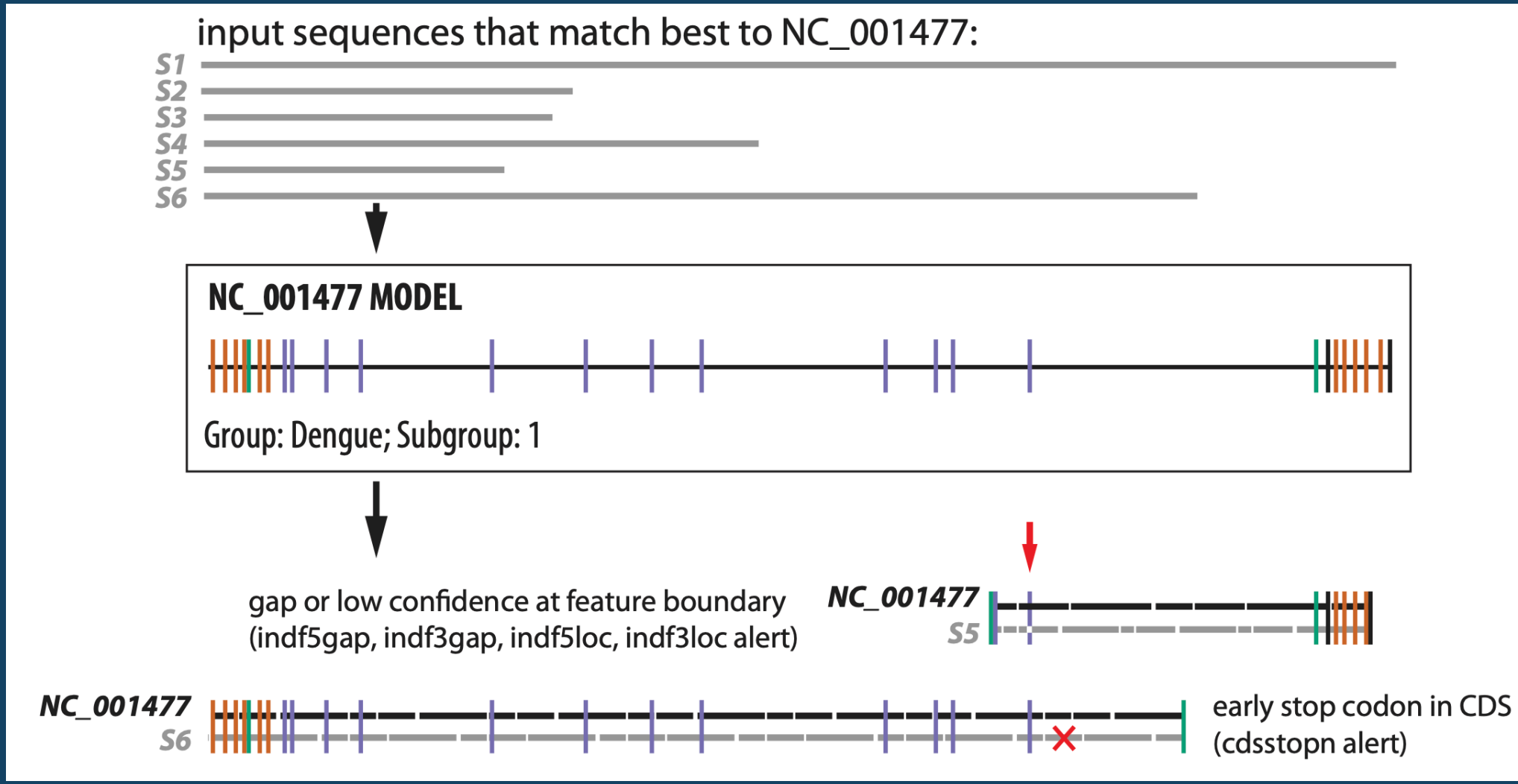
Stage 3: Alignment & feature mapping

- Problems caught in coverage determination stage could be missed in alignment stage



Stage 3: Alignment & feature mapping

- Parsing the alignment allows new types of alerts to be detected

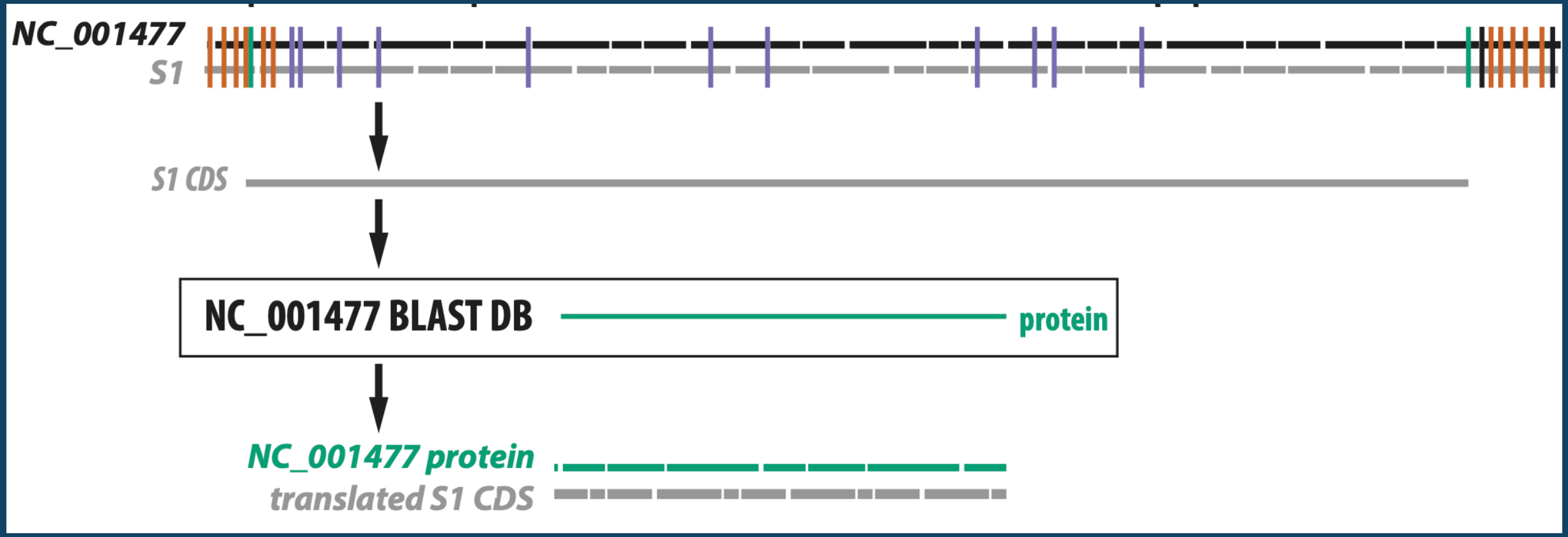


Alerts reported in alignment stage

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

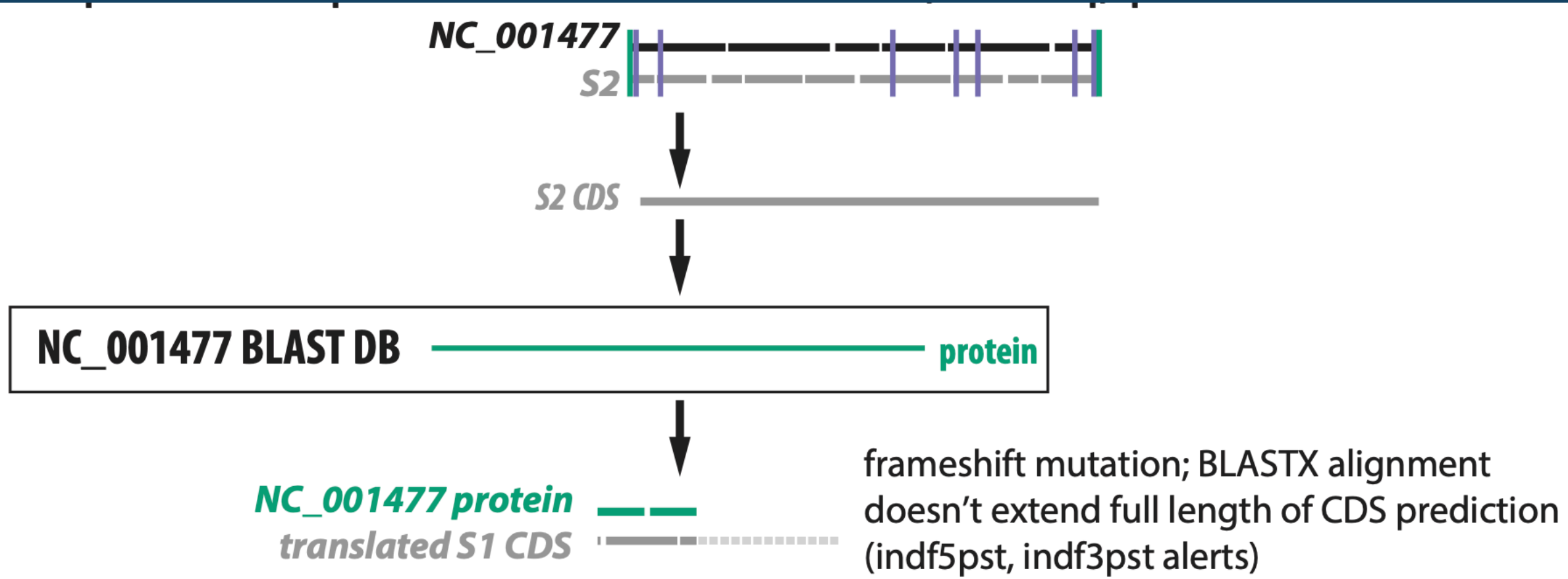
Stage 4: Protein validation (Alejandro Schaffer)

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



Stage 4: Protein validation (Alejandro Schaffer)

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



Alerts reported in protein validation stage

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

Enhancements for SARS-CoV-2 analysis

- Acceleration based on a BLASTN alignment, reduces processing time from several minutes to 2s per typical sequence
- Replacement of stretches of Ns with 'expected' nucleotides from the RefSeq when possible

virus	% of nucleotides that are Ns	% of seqs w/stretch of Ns >= 50 nt	number of sequences
SARS-CoV-2	1.12%	***26.4%***	28,936
Norovirus	0.296%	0.628%	43,937
Dengue virus	0.0037%	0.0070%	101,701

Fatal alert counts - SARS-CoV-2

Errors - Submitter

Category	Name	Count
sarscov2_annotation	ERROR_CDS_HAS_FRAMESHIFT	289
sarscov2_annotation	ERROR_CDS_HAS_STOP_CODON	422
sarscov2_annotation	ERROR_DELETION_OF_NT	41
sarscov2_annotation	ERROR_DISCONTINUOUS_SIMILARITY	12
sarscov2_annotation	ERROR_DUPLICATE_REGIONS	26
sarscov2_annotation	ERROR_INDEFINITE_ANNOTATION	5
sarscov2_annotation	ERROR_INDEFINITE_ANNOTATION_END	254
sarscov2_annotation	ERROR_INDEFINITE_ANNOTATION_START	228
sarscov2_annotation	ERROR_INSERTION_OF_NT	18
sarscov2_annotation	ERROR_LOW_COVERAGE	4
sarscov2_annotation	ERROR_LOW_FEATURE_SIMILARITY	58

Errors - Submitter

Category	Name	Count
sarscov2_annotation	ERROR_LOW_FEATURE_SIMILARITY_END	13
sarscov2_annotation	ERROR_LOW_FEATURE_SIMILARITY_START	10
sarscov2_annotation	ERROR_LOW_SIMILARITY	5
sarscov2_annotation	ERROR_LOW_SIMILARITY_END	137
sarscov2_annotation	ERROR_LOW_SIMILARITY_START	103
sarscov2_annotation	ERROR_MISASSEMBLED	23
sarscov2_annotation	ERROR_MUTATION_AT_END	131
sarscov2_annotation	ERROR_MUTATION_AT_START	133
sarscov2_annotation	ERROR_PEPTIDE_TRANSLATION_PROBLEM	239
sarscov2_annotation	ERROR_REVCOMPLEM	24
sarscov2_annotation	ERROR_UNEXPECTED_LENGTH	290

Example of SARS-CoV-2 failure that was trimmed by indexer

ExampleSeq1 – FAILed due to DUPLICATE_REGIONS, LOW_SIMILARITY_START and LOW_SIMILARITY

alert desc	alert detail
DUPLICATE_REGIONS	similarity to a model region occurs more than once [15307-15350 (len 44 >= 20) hits 1 and 2 (model:2..29876,15307..15350 seq:25..29899,29938..29981)]
LOW_SIMILARITY_START	significant similarity not detected at 5' end of the sequence [low similarity region of length 24 (1..24)]
LOW_SIMILARITY	internal region without significant similarity [low similarity region of length 38 (29900..29937)]

Example of SARS-CoV-2 failure that was trimmed by indexer

ExampleSeq1 – FAILed due to DUPLICATE_REGIONS, LOW_SIMILARITY_START and LOW_SIMILARITY

```
                nearly identical to NC_045512.1 27697..27719 (off by a single U)
5' end:         vvvv vvvvvvvvvvvvvvv
ExampleSeq1 caaga-acuuacucuccaauuuuucUUAAAGGUUUAUACCUUCCCAGGUA
#=GC RF     .....A..................UUAAAGGUUUAUACCUUCCCAGGUA
Identical                           **************************

                                                                                identical to NC_045512.1 15316..15350
3′ end:                                                                         vvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvvv
ExampleSeq1 AGAAUGACAAUAAGGUAGCGAAAUUCCUUGUCAACUAAUUAuugaccugcacgaaaggcgcaugccuaacaugcuuagaauuauggccucacuugu
#=GC RF     AGAAUGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAA....................................................
Identical   ********** **   *   ***          **  **  *
```

Example of SARS-CoV-2 failure that was eventually accepted

ExampleSeq2 – FAILed due to CDS_HAS_STOP_CODON, submitter verified

```
#alert              alert
#desc               detail
#----------------- ------
CDS_HAS_STOP_CODON in-frame stop codon exists 5' of stop position predicted by
                   homology to reference, revised to 27893..27916 (stop shifted 342 nt)

CDS_HAS_STOP_CODON stop codon in protein-based alignment [stop codon(s) end at
                   position(s) 27916]
```

Example of SARS-CoV-2 failure that was eventually accepted

ExampleSeq2 – FAILed due to CDS_HAS_STOP_CODON, submitter verified

```
CDS_HAS_STOP_CODON: ORF8 stop shifts by 342nt

                                        27916
                                          v
ExampleSeq2   [27893...ACATGAAATTTCTTGTTTTCTTATGAATCATCACAACTGTAGCTGCATTTCACC...28258]
NC_045512     [27894...ACATGAAATTTCTTGTTTTCTTAGGAATCATCACAACTGTAGCTGCATTTCACC...28259]
identical              ********************** ******************************
```

Limitations

- Makes inferences in nucleotide space, not protein space
- Model (RefSeq) must be representative
- Divergent sequences, novel stop codons, etc. are problematic
- Requires 64Gb of RAM (!) for some SARS-CoV-2 sequences

Important features of VADR

Try it out! Review your sequences before submission

- PASS/FAIL decision for each sequence with informative alert messages
- User control over fatal/non-fatal alerts and thresholds
- Standalone version for submitters ***and other users***
- General tool applicable to any virus (but performance is dependent on sequence diversity)
- Features can be as short as a single nucleotide, and potentially multiple 'segments' (programmed frameshifts)

Get VADR documentation on GitHub

Coronavirus annotation

Edit | New Page

Eric Nawrocki edited this page on May 13 · 12 revisions

https://github.com/nawrockie/vadr/wiki/Coronavirus-annotation

How to annotate SARS-CoV-2 sequences with VADR

Example annotation of SARS-CoV-2 sequences

Identifying and annotating *Coronaviridae* sequences other than SARS-CoV-2 using a larger VADR model library

Building new *Coronaviridae* models

Pages 4

Find a Page...

Home

Available VADR model files

Coronavirus annotation

Development notes and instructions

+ Add a custom sidebar

Clone this wiki locally

https://github.com/nawr

How to annotate SARS-CoV-2 sequences with VADR:

1. Download and install the latest version of VADR, following the instructions on this page
2. Download the latest coronavirus vadr models (gzipped tarball) from this FTP page, unpack them (e.g. `tar xfz <tarball.gz>`). Note the path to the directory name created (`<coronavirus-models-dir-path>`) for step 3.

GitHub documentation (cont.)

master ▾ vadr / documentation / annotate.md Go to file ⋯

https://github.com/nawrockie/vadr/wiki/Coronavirus-annotation

nawrockie Fixes new frame/codon_start calculation in code that looks for frames... ⋯ Latest commit bd3f940 on Jul 23 History

1 contributor

1044 lines (907 sloc) | 79.6 KB Raw Blame

`v-annotate.pl` example usage, command-line options and alert information

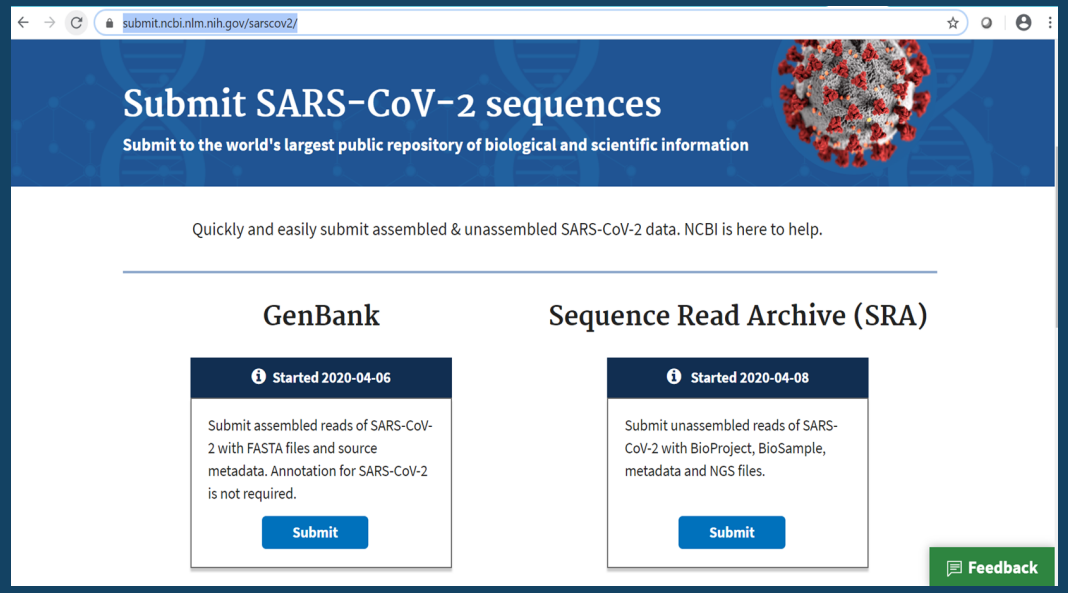
- `v-annotate.pl` example usage
  - example annotation of norovirus sequences
  - example of using `--alt_pass` to change alerts from fatal to non-fatal
  - example of using `-p` to run in parallel mode
- `v-annotate.pl` command-line options
  - basic options
  - options for specifying expected sequence classification
  - options for controlling which alerts are fatal

What VADR means for submissions

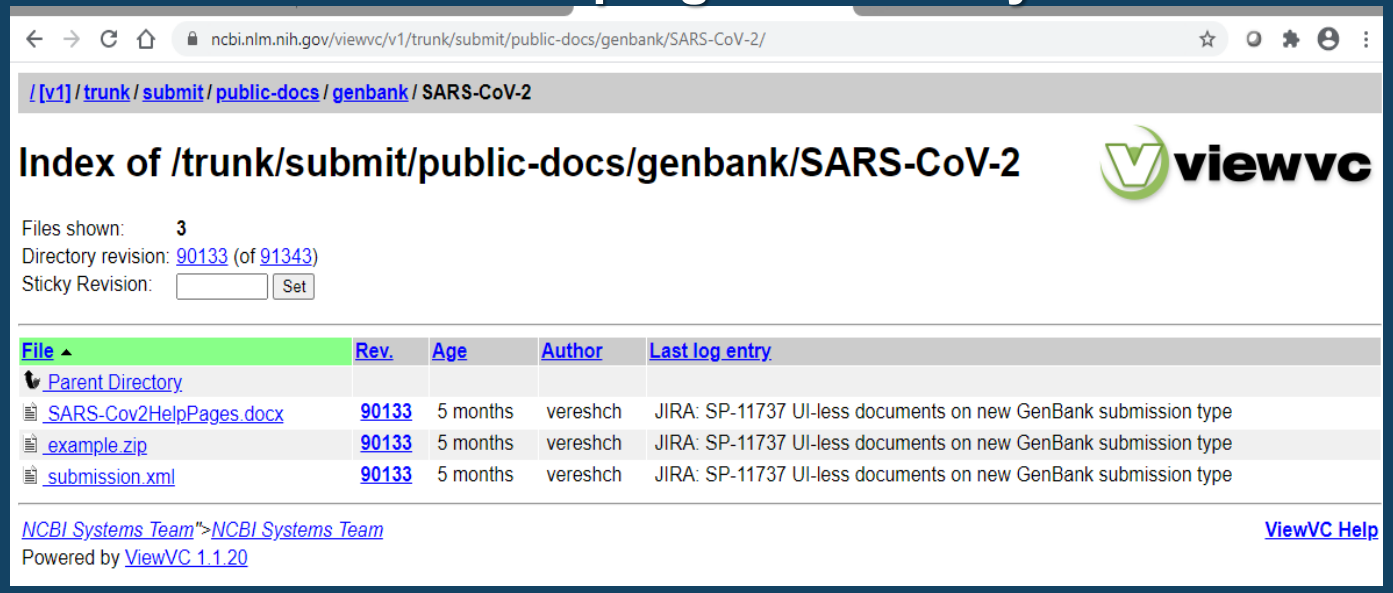
- **Speed -** automating annotation saves time for submitter and results in faster data release
- **Quality control -** VADR checks for sequence variability in coding and non-coding regions
- **Consistency -** standard nomenclature for genes and products allows for easier retrieval and comparisons
- **Value Added -** annotation of mature peptides and RNA structural components

Submitting assembled sequences

Submit on the web



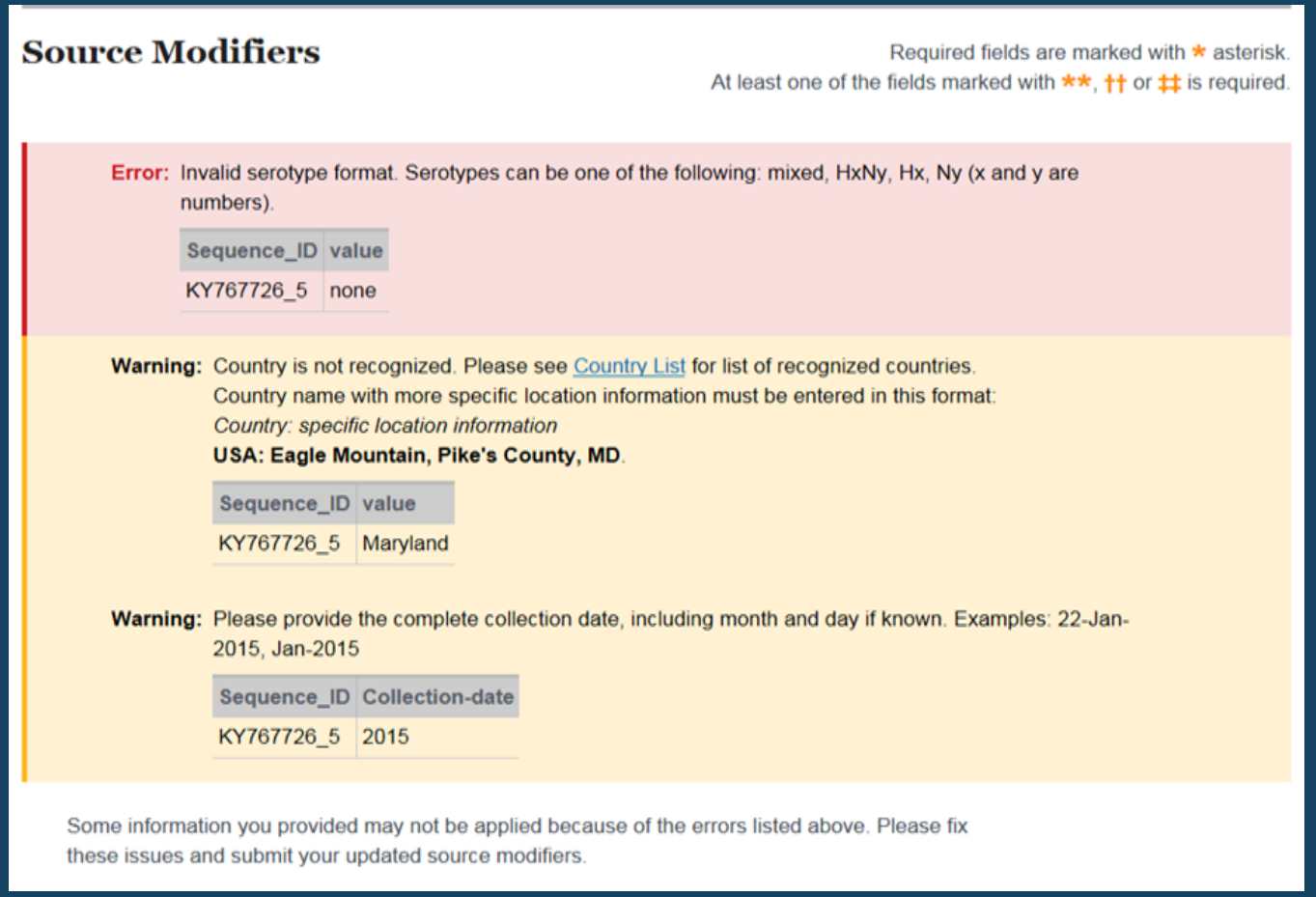
Submit programmatically



- Common requirements: FASTA, source table, submitter information
- **Both accept BioProject, BioSample and SRA run accession in source table**
- Common reporting system: VADR, validator, discrepancy report

Web submission to GenBank

- Forms prompt for required information
- Source information imported as table or use of editable table
- Interactive source and sequence validation
  - Country, date, isolate format
  - Sequence length and vector screening



Source Modifiers

Required fields are marked with * asterisk.
At least one of the fields marked with **, †† or ‡‡ is required.

Error: Invalid serotype format. Serotypes can be one of the following: mixed, HxNy, Hx, Ny (x and y are numbers).

Sequence_ID	value
KY767726_5	none

Warning: Country is not recognized. Please see Country List for list of recognized countries.
Country name with more specific location information must be entered in this format:
Country: specific location information
USA: Eagle Mountain, Pike's County, MD.

Sequence_ID	value
KY767726_5	Maryland

Warning: Please provide the complete collection date, including month and day if known. Examples: 22-Jan-2015, Jan-2015

Sequence_ID	Collection-date
KY767726_5	2015

Some information you provided may not be applied because of the errors listed above. Please fix these issues and submit your updated source modifiers.

Programmatic submissions

- .zip archive file and submission .xml file uploaded via ftp
  - .fsa, .src, .sbt, optional .cmt
- Sequence checks: low quality, vector, length. Source checks: Country, date, isolate format
- If no formatting errors, a temporary SUB number is assigned in report.xml
- VADR validation
- Accessions, files on Submission Portal: https://submit.ncbi.nlm.nih.gov/subs/api/

```
<?xml version="1.0"?>
<Submission>
  <Description>
    <Comment>SARS-CoV-2 test submission</Comment>
    <Organization type="center" role="owner">
      <Name>account name</Name>
    </Organization>
    <Hold release_date="2024-05-25"/>
  </Description>
  <Action>
    <AddFiles target_db="GenBank">
      <File file_path="sarscov2.zip">
        <DataType>genbank-submission-package</DataType>
      </File>
      <Attribute name="wizard">BankIt_SARSCoV2_api</Attribute>
      <Identifier>
        <SPUID spuid_namespace="ncbi-sarscov2-genbank">2020-03-
04.sarscov2</SPUID>
      </Identifier>
    </AddFiles>
  </Action>
</Submission>
```

Common accession reports

2 submissions

Submission	Title	App	Owner	Group	Status	Updated
SUB593912	SARS-CoV-2	GenBank	yankie		✔ **GenBank: Processed** EU865031-EU865032 3 files: • AccessionReport.tsv • flatfile.txt • email.txt	Oct 01
SUB586761	UI-less submission 2020-05-08	API	testsars2-srv	testsars2	✔ **GenBank: Processed** EU864968-EU864969 3 files: • AccessionReport.tsv • flatfile.txt • email.txt	May 08

Common VADR alert reporting

SUB586164

A list of errors in your submission is included below.
Click on the error titles for explanations of the errors and the suggested corrections. Then use the 'Fix' button for SUB586164 to enable editing of your submission and correct the errors.

If you have questions, please write to: gb-admin@ncbi.nlm.nih.gov with "GenBank Submission Portal SUB586164" in the subject line of your email.

If we do not receive a corrected submission by May 20, 2020, SUB586164 will be deleted.

==========

ERRORS

[] CDS_HAS_FRAMESHIFT

396241MT

[] CDS_HAS_STOP_CODON

396241MT

[] INDEFINITE_ANNOTATION_END

396241MT

[] INDEFINITE_ANNOTATION_START

396241MT

[] PEPTIDE_TRANSLATION_PROBLEM

396241MT

[] UNEXPECTED_LENGTH

396241MT

2 submissions

Submission	Title	App	Owner	Group	Status	Updated
SUB618483	SARS-CoV-2	GenBank	yankie		GenBank: Error has errors SUB618483-Report.html Fix	Sep 04
SUB599072	UI-less submission 2020-07-10	API	testsars2-srv	testsars2	GenBank: Error has errors SUB599072-Report.html	Jul 10

Future improvements

- Validator changes for handling Ns will result in less manual review
- Include detailed error report as separate file

```
lcl|396241MT    DELETION_OF_FEATURE     *sequence*      internal deletion of a complete feature [mat_peptide feature number 7: nsp7]
lcl|396241MT    DELETION_OF_FEATURE     *sequence*      internal deletion of a complete feature [mat_peptide feature number 22: nsp7]
lcl|396241MT    CDS_HAS_STOP_CODON      ORF1ab polyprotein      in-frame stop codon exists 5' of stop position predicted by homology to reference [revised to 266..425 (s
lcl|396241MT    POSSIBLE_FRAMESHIFT_HIGH_CONF   ORF1ab polyprotein      high confidence potential frameshift in CDS [nucleotide alignment of positions 11410..20572 (9163
lcl|396241MT    INDEFINITE_ANNOTATION_END       ORF1ab polyprotein      protein-based alignment does not extend close enough to nucleotide-based alignment 3' endpoint [9
lcl|396241MT    UNEXPECTED_LENGTH       ORF1ab polyprotein      length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [20308]
lcl|396241MT    CDS_HAS_STOP_CODON      ORF1a polyprotein       in-frame stop codon exists 5' of stop position predicted by homology to reference [revised to 266..425 (s
lcl|396241MT    POSSIBLE_FRAMESHIFT_HIGH_CONF   ORF1a polyprotein       high confidence potential frameshift in CDS [nucleotide alignment of positions 11410..12500 (1091
lcl|396241MT    INDEFINITE_ANNOTATION_END       ORF1a polyprotein       protein-based alignment does not extend close enough to nucleotide-based alignment 3' endpoint [1
lcl|396241MT    UNEXPECTED_LENGTH       ORF1a polyprotein       length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [12235]
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     leader protein  mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp2    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp3    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp4    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     3C-like proteinase      mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    INDEFINITE_ANNOTATION_END       nsp6    alignment to homology model is a gap at 3' boundary [RF position 11842]
lcl|396241MT    PEPTIDE_ADJACENCY_PROBLEM       nsp6    predictions of two mat_peptides expected to be adjacent are not adjacent [feature stops at seq position 11409 on
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp6    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    UNEXPECTED_LENGTH       nsp6    length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [440]
lcl|396241MT    INDEFINITE_ANNOTATION_START     nsp8    alignment to homology model is a gap at 5' boundary [RF position 12092]
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp8    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    UNEXPECTED_LENGTH       nsp8    length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [293]
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp9    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp10   mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     RNA-dependent RNA polymerase    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     helicase        mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     3'-to-5' exonuclease    mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     endoRNAse       mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     2'-O-ribose methyltransferase   mat_peptide may not be translated because its parent CDS has a problem
lcl|396241MT    PEPTIDE_TRANSLATION_PROBLEM     nsp11   mat_peptide may not be translated because its parent CDS has a problem
```

VADR alerts on confirmed mutations

- Write gb-admin@ncbi.nlm.nih.gov
  - Read coverage over area of mutation
  - SRR number if not already included
- If mutation is common, please let us know!
  - We can add additional sequences to VADR reference set so future submissions with the same mutation will not produce an alert
  - Submitter-confirmed mutations in ORF7a, ORF7b, ORF8, ORF3a, ORF6, ORF10 and S

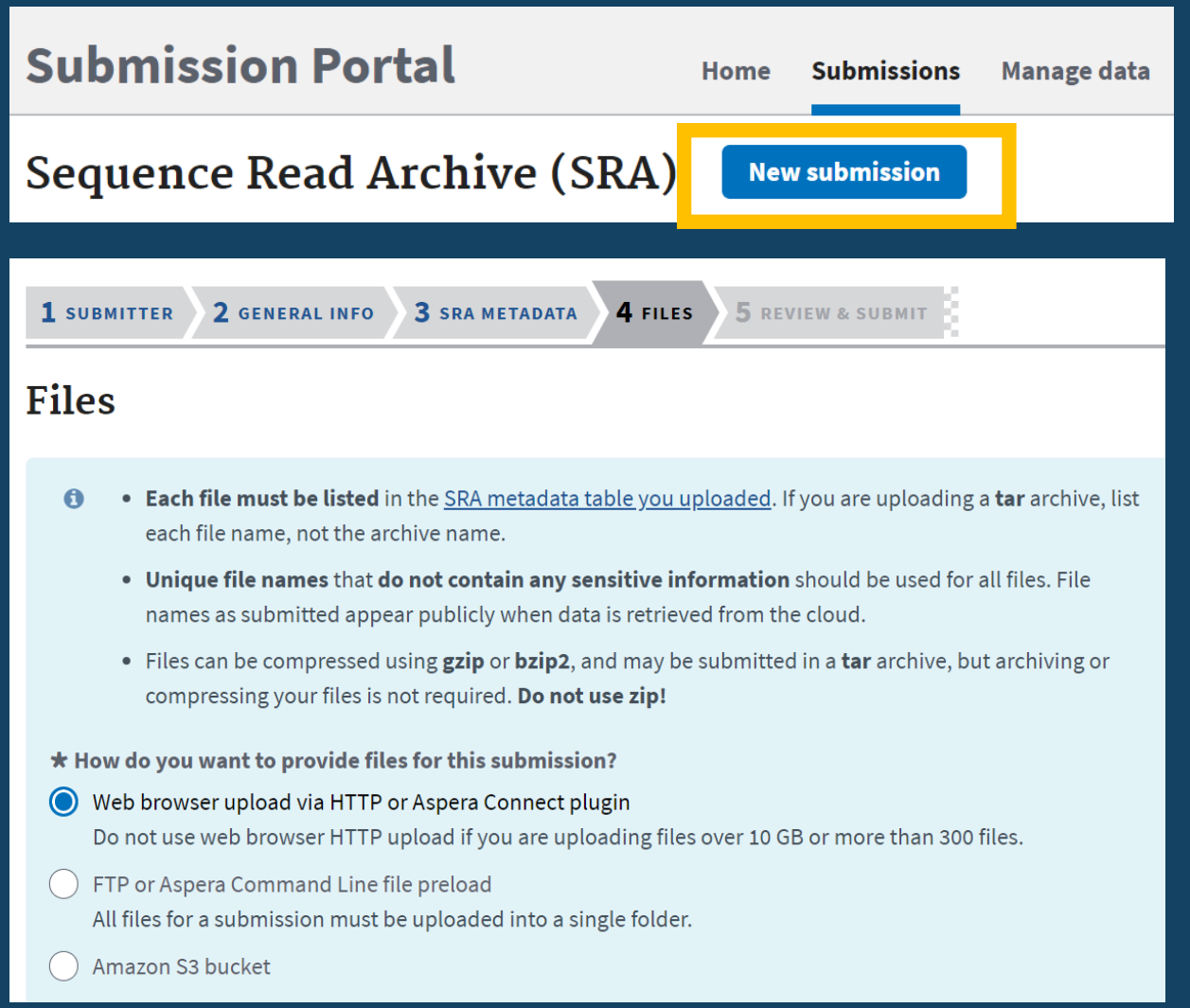
Create data connections ←→

Do you submit to GenBank and SRA?

- BioProject can be used to link your submissions!
  - NCBI-created SPHERES umbrella = **PRJNA615625**
    - New BioProject submissions: Enter this accession # as you submit
    - Already submitted? Email us at submit-help@ncbi.nlm.nih.gov & we'll make the connection
- Use the same BioSample in both your SRA and GenBank submissions

Getting started with SRA submission

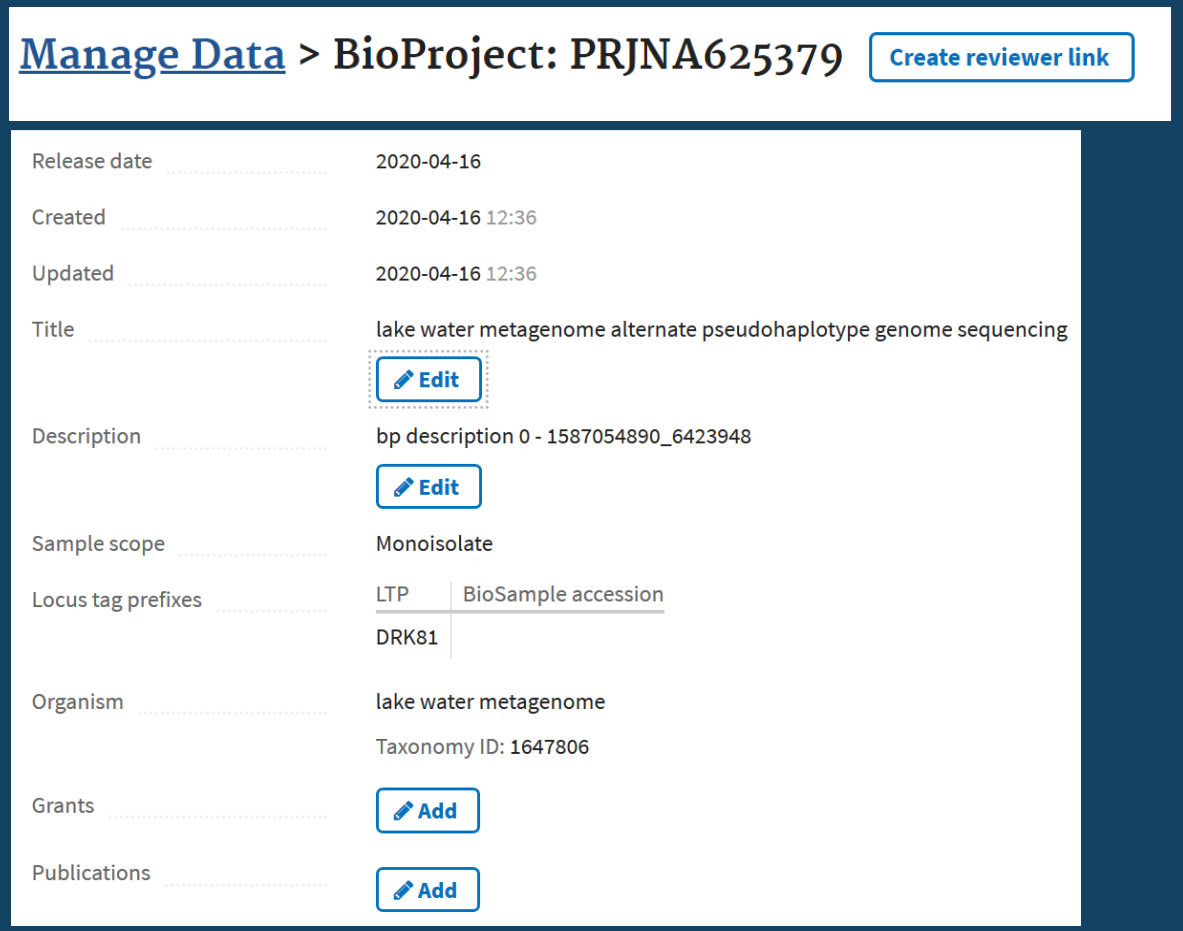
- Submit unassembled reads to SRA
- Register BioProject & BioSample during SRA submission
- Flexible file upload options
  - Web or FTP
  - Cloud-based: AWS now, GCP next month



More on SRA submission & stats

- Modify submission after processing
  - Edit or add BioProject and SRA metadata
  - "Release now" or delay release
- Programmatic submission option also available
- SARS-CoV-2 SRA data: 11K Runs from USA and ~100K global
- Data now on https://registry.opendata.aws/ncbi-covid-19/ !

dataview.ncbi.nlm.nih.gov



We’re here for you!

- Enhancements made per SPHERES input
  - Programmatic submission option(s)
  - File upload drag-and-drop
  - Help documentation edits
- Parking lot – Input needed!
  - BioSample linkage & data flow during GenBank submission
  - Combined SRA & GenBank programmatic submission
  - “Human scrubber” tool for removing human sequences
- Taking volunteers for testing, feedback!

Q&A

Help

- NCBI is here to help with your submission!
  - GenBank gb-admin@ncbi.nlm.nih.gov
  - SRA sra@ncbi.nlm.nih.gov
  - VADR resources on GitHub https://github.com/nawrockie/vadr

- RefSeq
  - Send feedback on RefSeq NC_045512 to Eneida.Hatcher@nih.gov