SARS CoV-2 submission overview for Africa Pathogen Genomics Institute (PGI)

The National Center for Biotechnology Information (NCBI)
Rick Lapoint, Linda Yankie
Lydia Fleischmann

19 October, 2021

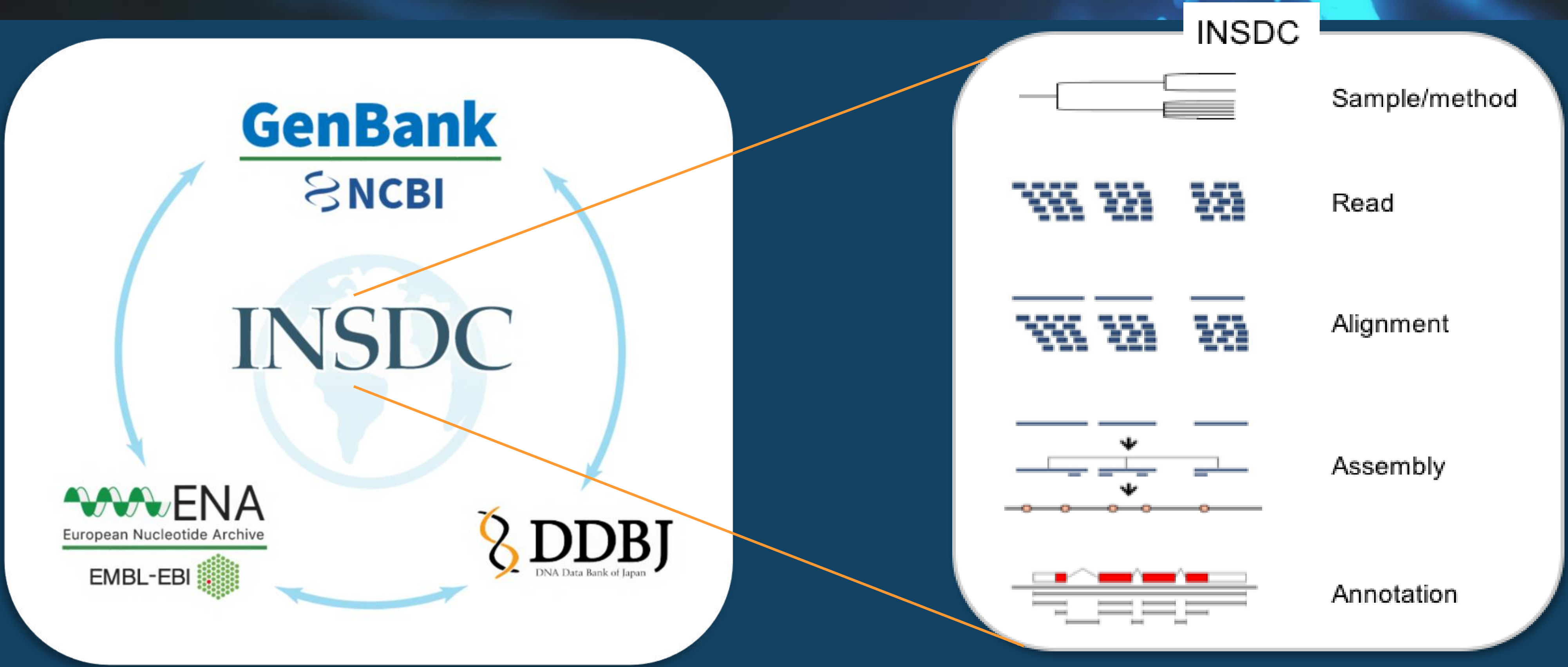
Introduction

The National Center for Biotechnology Information **(NCBI)**, is part of the National Library of Medicine **(NLM)** at the National Institutes of Health **(NIH)** in Bethesda, Maryland, USA.

NCBI is part of the International Nucleotide Sequence Database Collaboration **(INSDC)**.

INSDC Partners:
- The European Bioinformatics Institute
- The DNA Data Bank of Japan

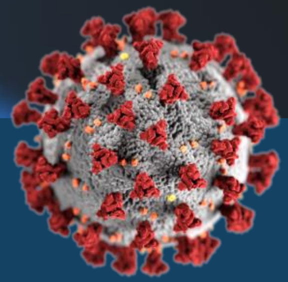
International Nucleotide Sequence Database collaboration



- Regular data exchange
- Embrace data standards
- Open and unrestricted access

- Globally comprehensive coverage
- Scientific database of record
- Public forum for the scientific process

Benefits of PGI submission to NCBI



- Your data is valuable! NCBI working with submitters around the world to increase SARS-CoV-2 submission
  - Emerging variants worldwide tracked more easily with data from around the globe
- Put more open, public data into the hands of global researchers working on pandemic surveillance, response and therapeutics
- Establish submission workflows to make it easier to submit other types of sequence data post-pandemic

NCBI’s archives

- Sequence Read Archive (SRA) The largest publicly-available repository of next generation sequence (NGS) data

- GenBank Archive of assembled nucleotide sequence data and annotations with descriptive metadata including genome and transcriptome assemblies

Metadata Resources

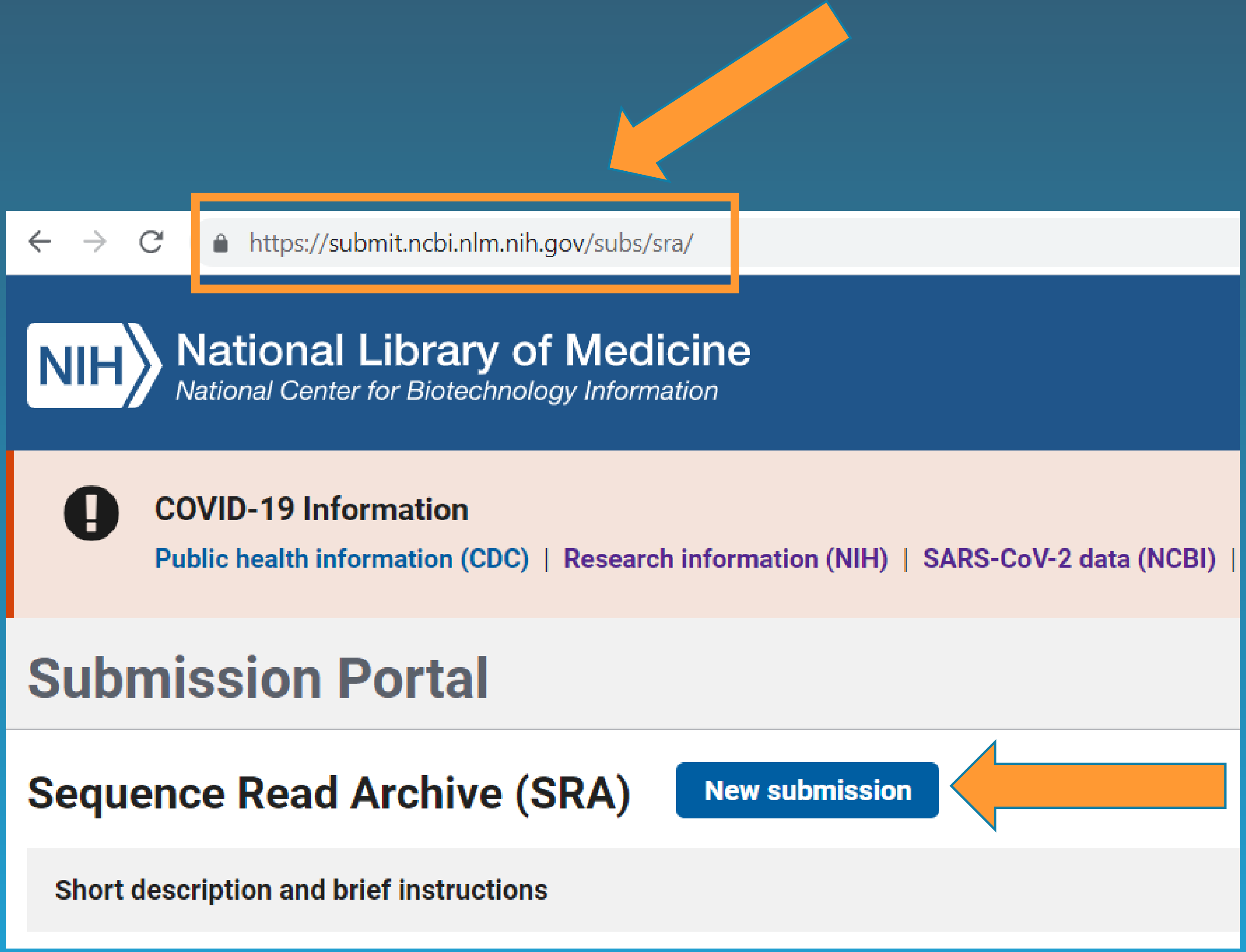
- BioProject Collection of biological data related to an initiative which provides users with links to the diverse data types generated for that project

- BioSample Stores descriptive information about biological materials used in experimental assays

SRA

Submit unassembled reads of SARS-CoV-2 with BioProject, BioSample, metadata and Next-Generation Sequencing (NGS) NGS files.

Accessions in 2 hrs. (avg)



https://submit.ncbi.nlm.nih.gov/subs/sra/

NIH National Library of Medicine
National Center for Biotechnology Information

COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) |

Submission Portal

Sequence Read Archive (SRA)

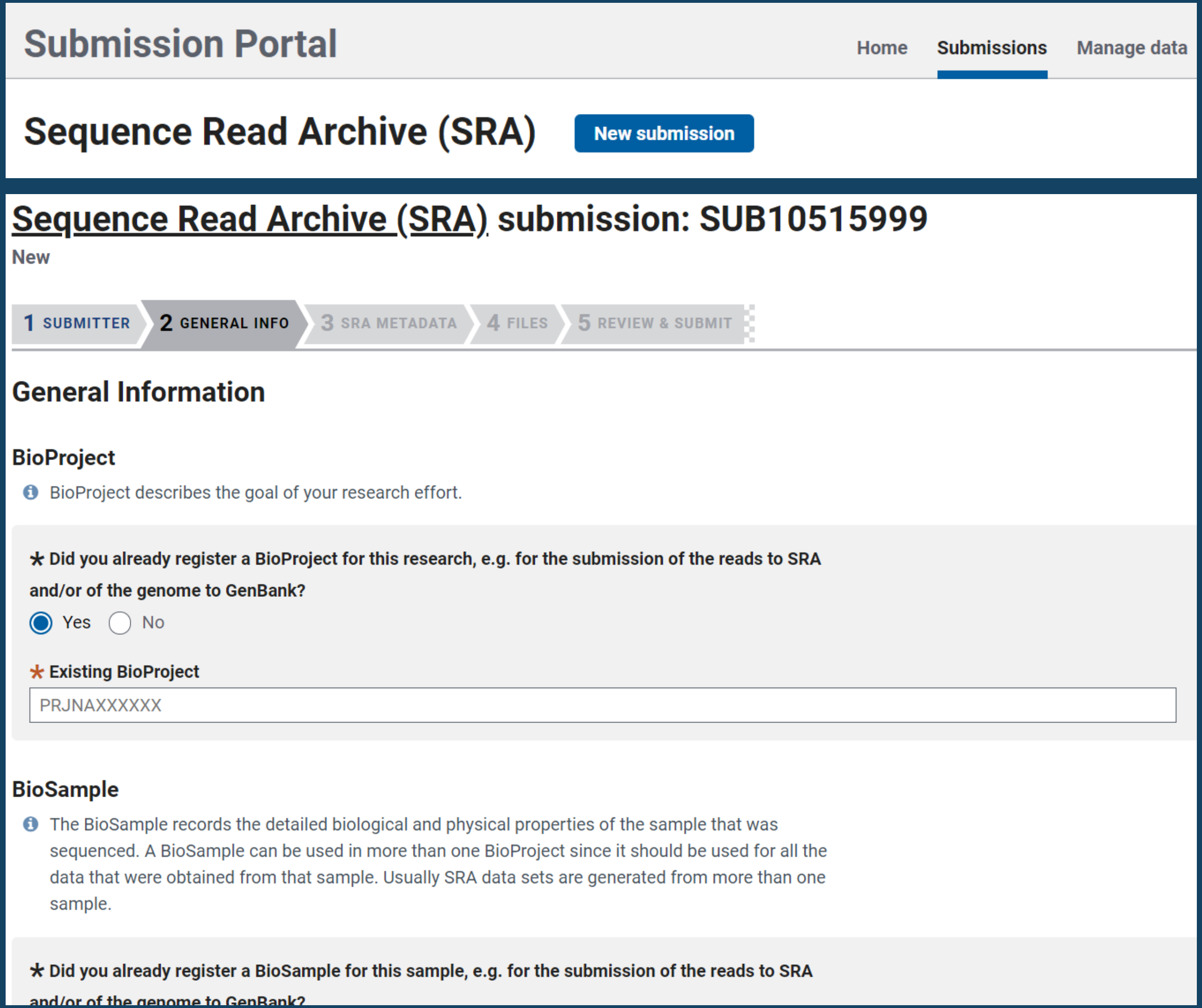
Short description and brief instructions

Getting started with SRA submission

- Guided workflow on web

- Register BioProject & BioSample *during SRA submission*

- Complete a few easy steps:
  - Submitter, general info
  - BioProject / BioSample
  - Metadata
  - File upload



Sequence Read Archive (SRA)

Sequence Read Archive (SRA) submission: SUB10515999

New

General Information

BioProject

BioProject describes the goal of your research effort.

* Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?

● Yes ○ No

* Existing BioProject

PRJNAXXXXXX

BioSample

The BioSample records the detailed biological and physical properties of the sample that was sequenced. A BioSample can be used in more than one BioProject since it should be used for all the data that were obtained from that sample. Usually SRA data sets are generated from more than one sample.

* Did you already register a BioSample for this sample, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?

BioProject & BioSample: Connect your data

We encourage you submit to SRA and GenBank!

- BioProjects are a description of your study and a single place to find links to the archived data for that study.

- BioSamples describe the biologically, or physically, unique specimen that was sequenced.
    - Use the same BioSample in both your SRA and GenBank submissions

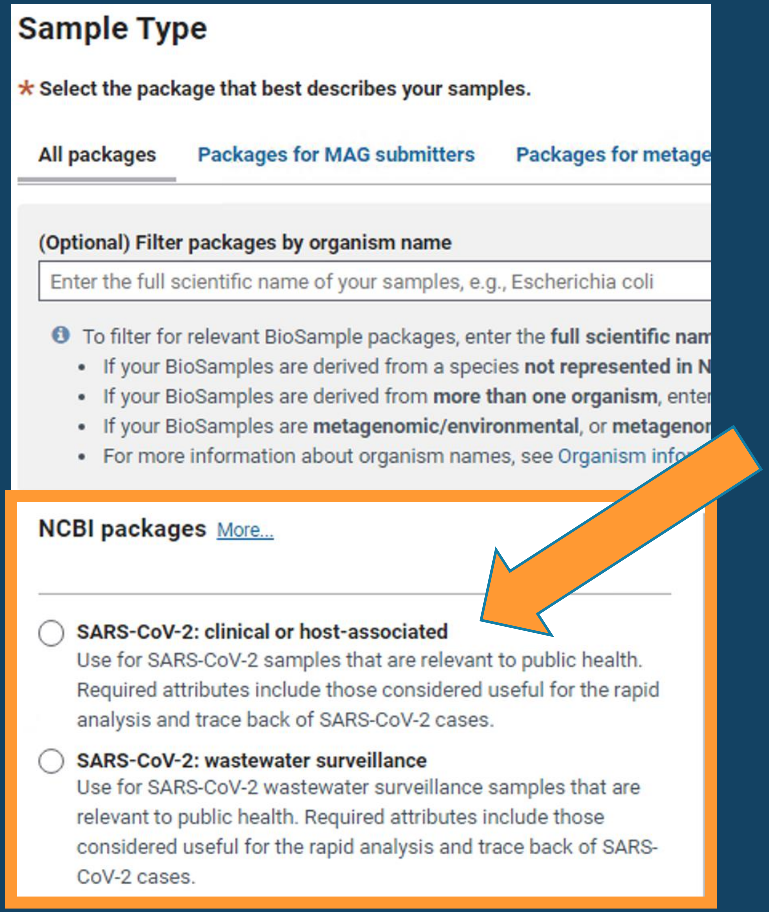
SARS-CoV-2 BioSample packages

- “Packages” are a collection of attributes to help submitters & researchers

- Select the "SARS-CoV-2 clinical or host-associated package" as your BioSample type

- Specific to SARS-CoV-2 BioSamples
  - Source – host, location, isolation, etc.
  - Host travel history (location and dates)
  - Prior infection
  - Antiviral treatment
  - Cycle threshold value – result from diagnostic SARS-CoV-2 RT-PCR test, e.g., 2

Preview BioSample packages at: https://submit.ncbi.nlm.nih.gov/biosample/template/



Sample Type

* Select the package that best describes your samples.

(Optional) Filter packages by organism name

Enter the full scientific name of your samples, e.g., Escherichia coli

To filter for relevant BioSample packages, enter the **full scientific nam**

- If your BioSamples are derived from a species **not represented in N**
- If your BioSamples are derived from **more than one organism**, ente
- If your BioSamples are **metagenomic/environmental**, or **metagenom**
- For more information about organism names, see Organism info

NCBI packages More...

○ **SARS-CoV-2: clinical or host-associated**
Use for SARS-CoV-2 samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

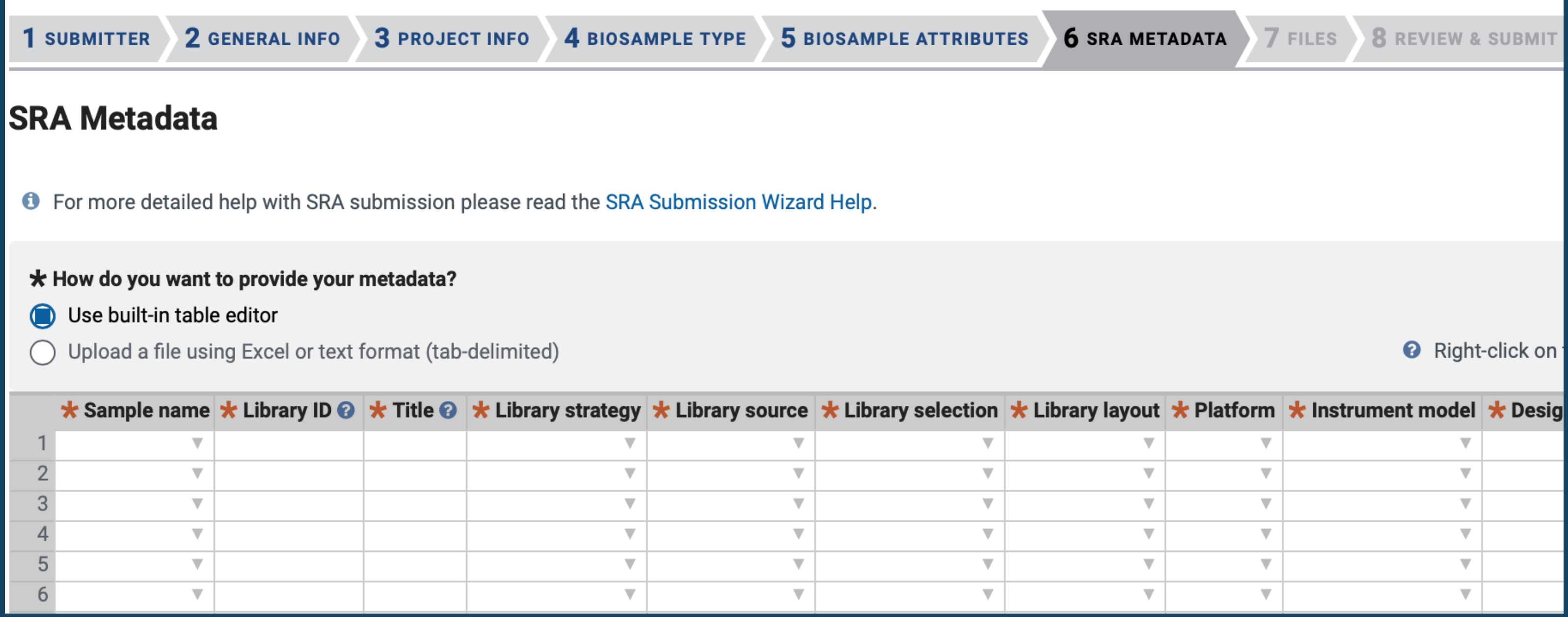
○ **SARS-CoV-2: wastewater surveillance**
Use for SARS-CoV-2 wastewater surveillance samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of SARS-CoV-2 cases.

Adding SRA metadata

- Library ID

- Controlled vocabulary
  - Library strategy
  - Library source
  - Library selection
  - Library layout
  - Instrument platform and model

An editable table in SRA submission works like Microsoft Excel



SRA Metadata

For more detailed help with SRA submission please read the SRA Submission Wizard Help.

* How do you want to provide your metadata?

● Use built-in table editor

○ Upload a file using Excel or text format (tab-delimited)

Right-click on

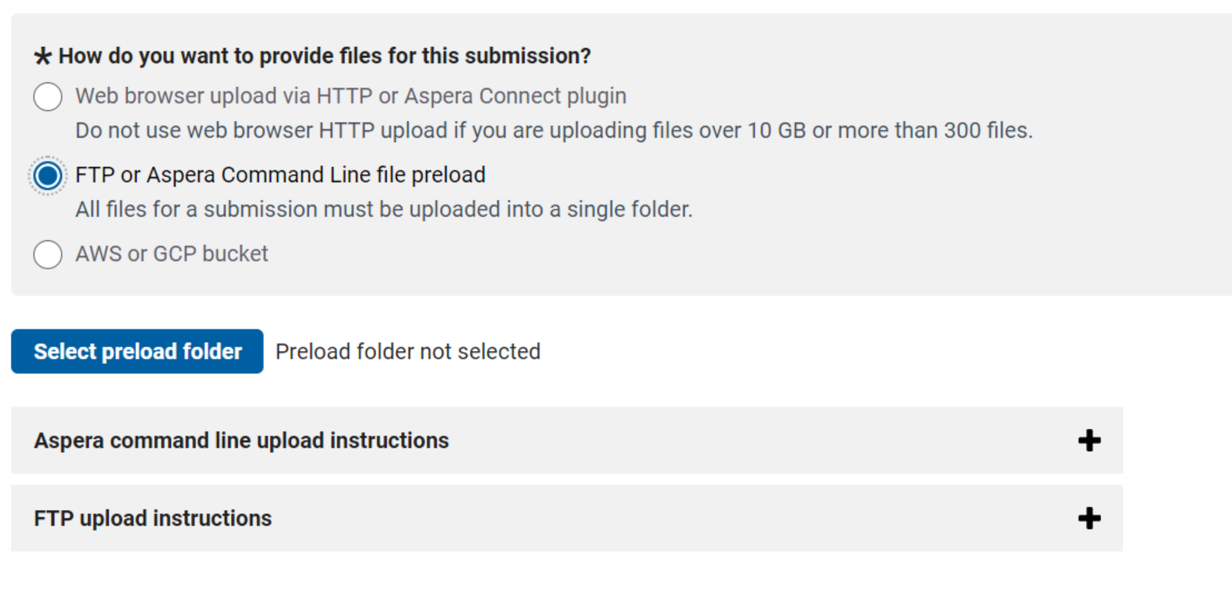
	* Sample name	* Library ID	* Title	* Library strategy	* Library source	* Library selection	* Library layout	* Platform	* Instrument model	* Desig
1										
2										
3										
4										
5										
6										

Uploading files to SRA

- Flexible file upload options
  - Command line Aspera
  - Web browser or FTP
  - Cloud-based: Transfer from Amazon Web Services (AWS), Google Cloud Platform (GCP) buckets

- Create a preload folder in advance of, or during, submission

When you select an option, the screen will refresh to provide prompts that help you move forward



*** How do you want to provide files for this submission?**

○ Web browser upload via HTTP or Aspera Connect plugin
Do not use web browser HTTP upload if you are uploading files over 10 GB or more than 300 files.

● FTP or Aspera Command Line file preload
All files for a submission must be uploaded into a single folder.

○ AWS or GCP bucket

Preload folder not selected

Aspera command line upload instructions +
FTP upload instructions +

Human read removal available to remove host reads from SARS-CoV-2 sequences

- The human read removal tool (HRRT) outputs a sequence file in which all reads that are identified as potentially of human origin are removed
  https://github.com/ncbi/sra-human-scrubber

- Based on the SRA Taxonomy Analysis Tool
  https://doi.org/10.1101/2021.02.16.431451

- Notify the SRA when your submission completes to have your reads screened

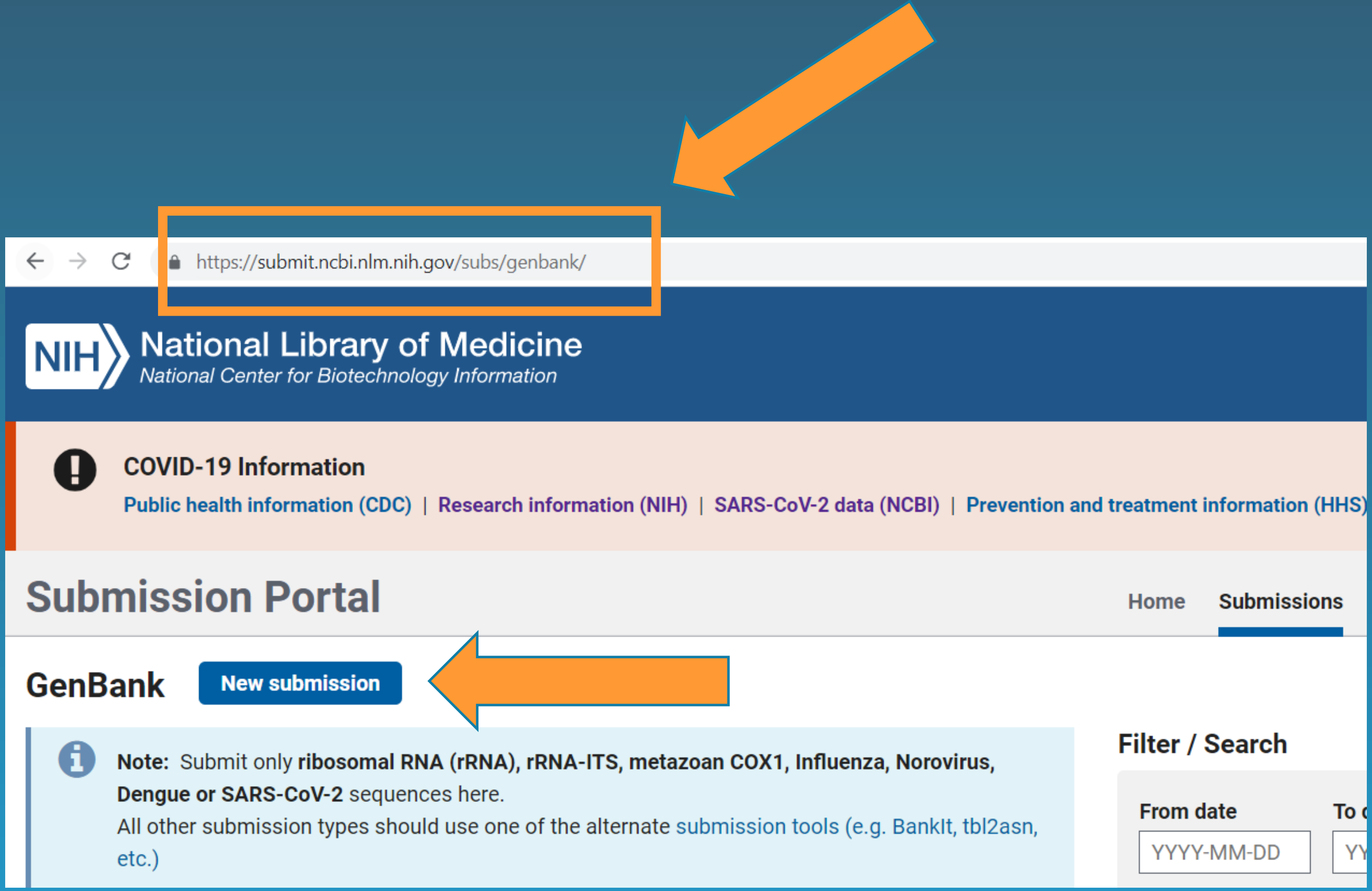
DEMO

GenBank

Submit assembled reads of SARS-CoV-2 with FASTA files and source metadata.

Gene annotation for SARS-CoV-2 **is not required**.

Accessions in 2 hrs. (avg)



https://submit.ncbi.nlm.nih.gov/subs/genbank/

NIH National Library of Medicine
National Center for Biotechnology Information

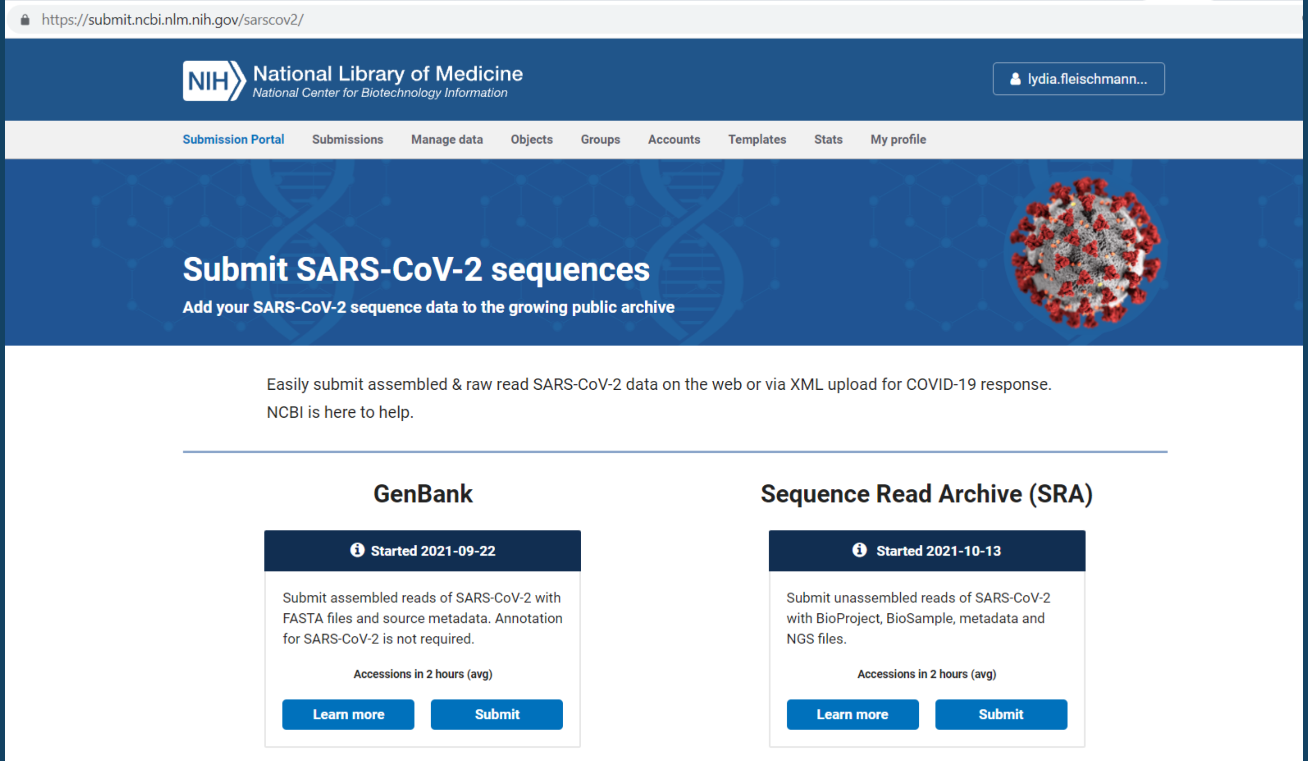
COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

Submission Portal

Note: Submit only **ribosomal RNA (rRNA), rRNA-ITS, metazoan COX1, Influenza, Norovirus, Dengue or SARS-CoV-2** sequences here.
All other submission types should use one of the alternate submission tools (e.g. BankIt, tbl2asn, etc.)

Submitting assembled sequences

This onboarding site helps you prepare to submit



Submit SARS-CoV-2 sequences

Add your SARS-CoV-2 sequence data to the growing public archive

Easily submit assembled & raw read SARS-CoV-2 data on the web or via XML upload for COVID-19 response. NCBI is here to help.

GenBank

Started 2021-09-22

Submit assembled reads of SARS-CoV-2 with FASTA files and source metadata. Annotation for SARS-CoV-2 is not required.

Accessions in 2 hours (avg)

Sequence Read Archive (SRA)

Started 2021-10-13

Submit unassembled reads of SARS-CoV-2 with BioProject, BioSample, metadata and NGS files.

Accessions in 2 hours (avg)

- Requirements: FASTA, source table, submitter information
- BioProject, BioSample and SRA run accession listed in source table
- Reporting & quality checks to assist you in submission

Viral Annotation DefineR (VADR) for annotation & sequence quality checks

Publicly-available tool https://github.com/ncbi/vadr

- General tool applicable to a wide variety of viruses, supporting submission automation & providing informative alert messages

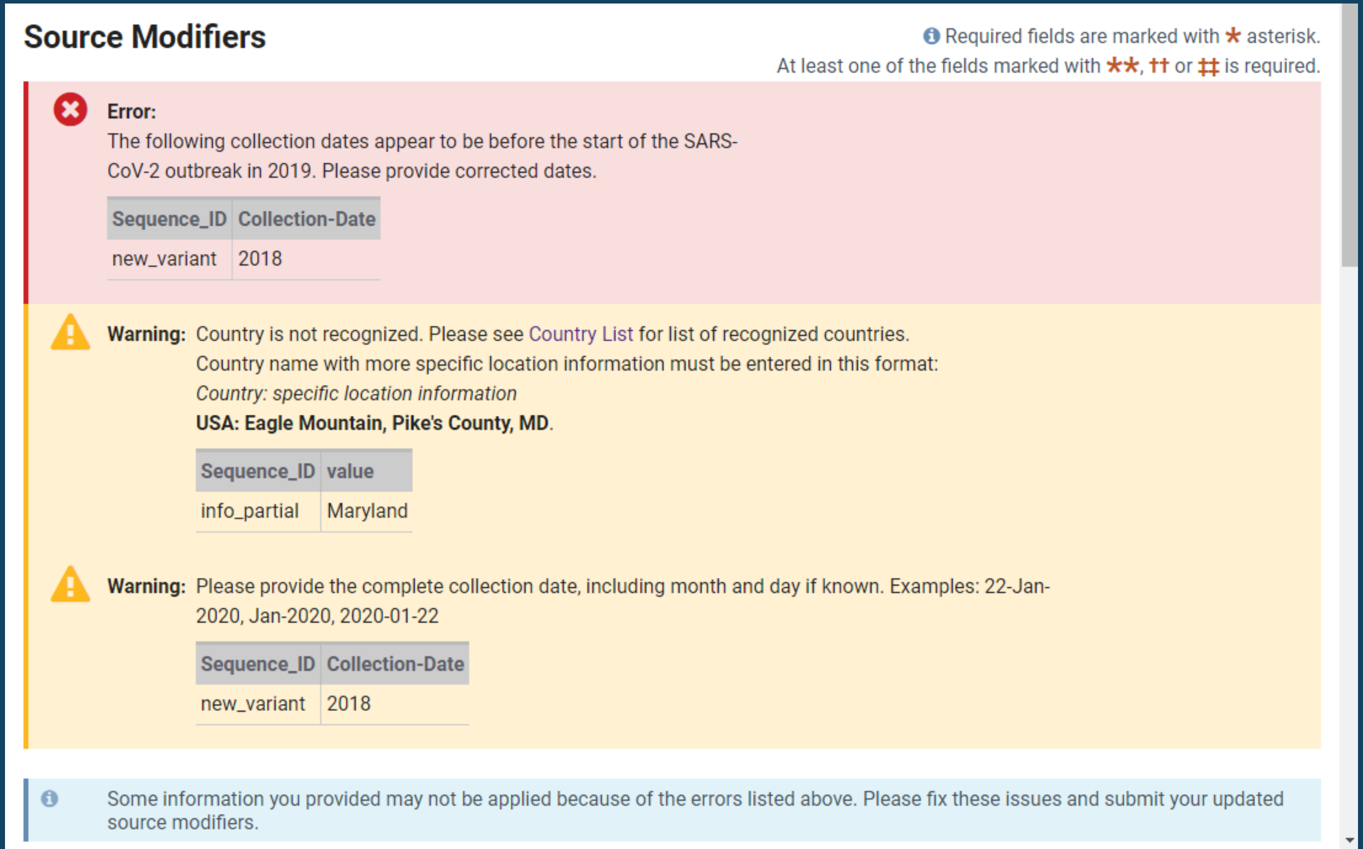
- Provides consistent annotation, including mature peptides and RNA features

- Designed so parameters can be adjusted over time based on viral evolution

Web submission to GenBank

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

Real-time validations guide you during submission



Source Modifiers

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

Error:
The following collection dates appear to be before the start of the SARS-CoV-2 outbreak in 2019. Please provide corrected dates.

Sequence_ID	Collection-Date
new_variant	2018

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.

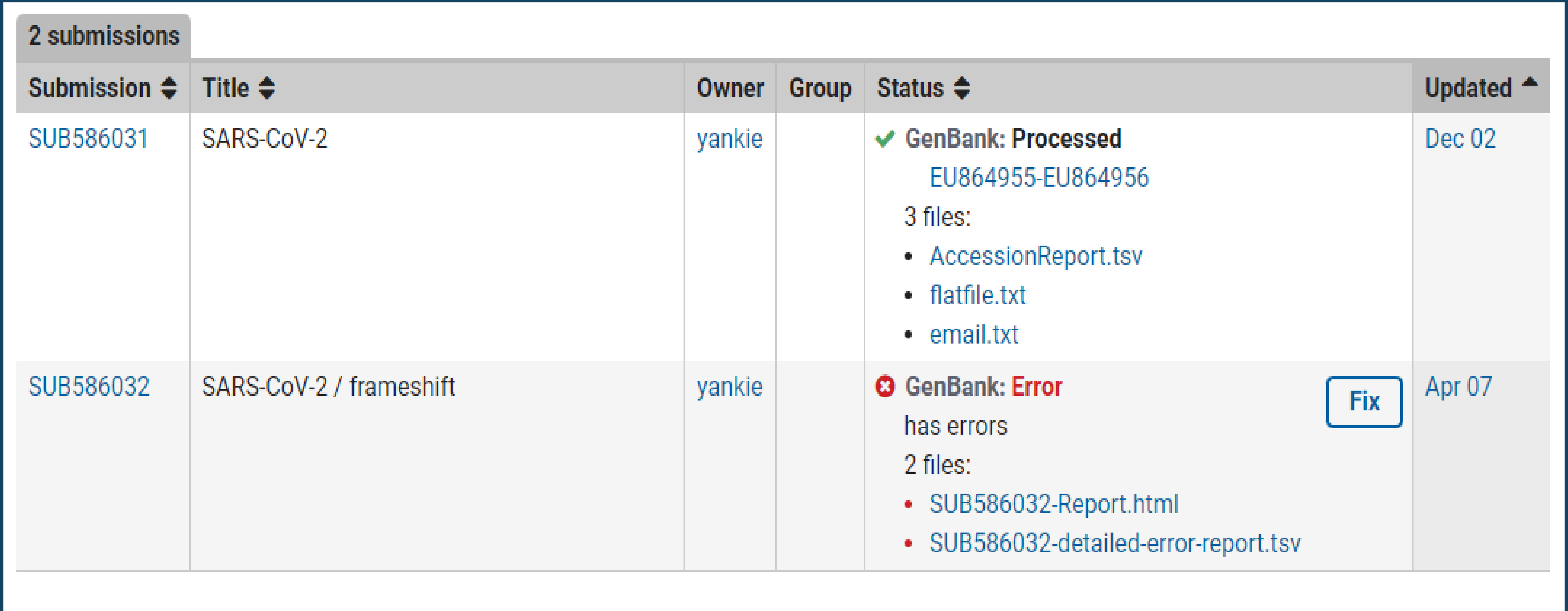
Sequence_ID	value
info_partial	Maryland

Warning: Please provide the complete collection date, including month and day if known. Examples: 22-Jan-2020, Jan-2020, 2020-01-22

Sequence_ID	Collection-Date
new_variant	2018

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

Submission Portal view

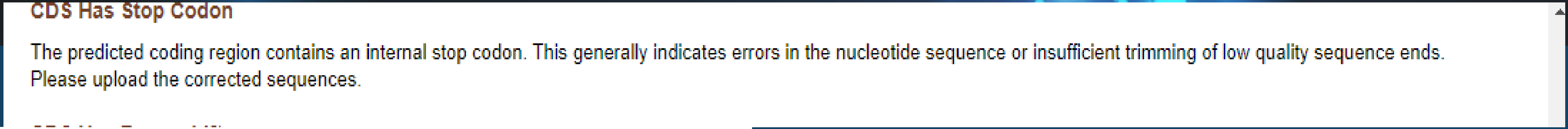


2 submissions

Submission ⬍	Title ⬍	Owner	Group	Status ⬍	Updated ▲
SUB586031	SARS-CoV-2	yankie		✓ GenBank: Processed EU864955-EU864956 3 files: • AccessionReport.tsv • flatfile.txt • email.txt	Dec 02
SUB586032	SARS-CoV-2 / frameshift	yankie		⊗ GenBank: Error Fix has errors 2 files: • SUB586032-Report.html • SUB586032-detailed-error-report.tsv	Apr 07

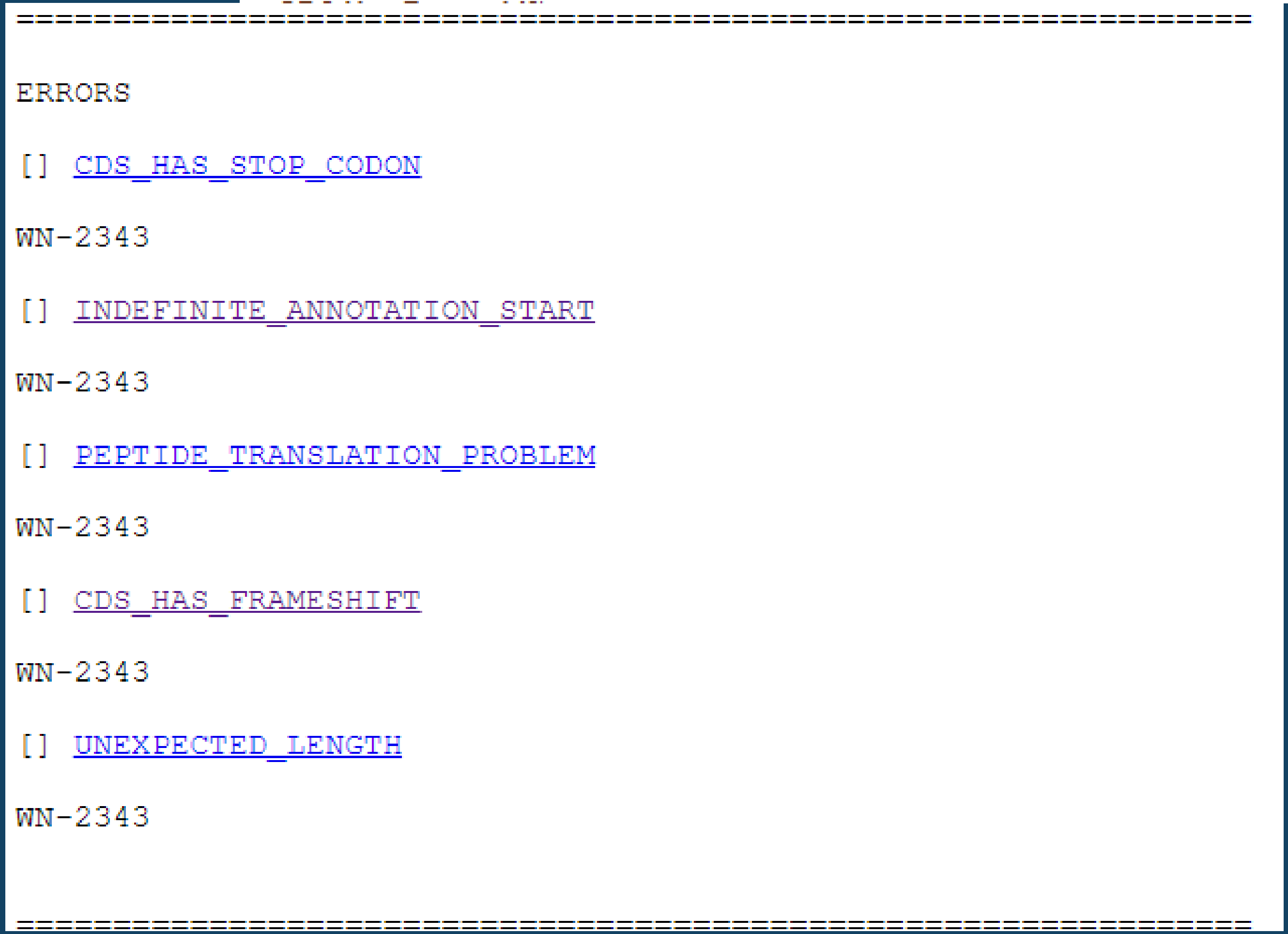
Example view of GenBank SARS-CoV-2 submissions in Submission Portal. Fix option for errors

Alert report

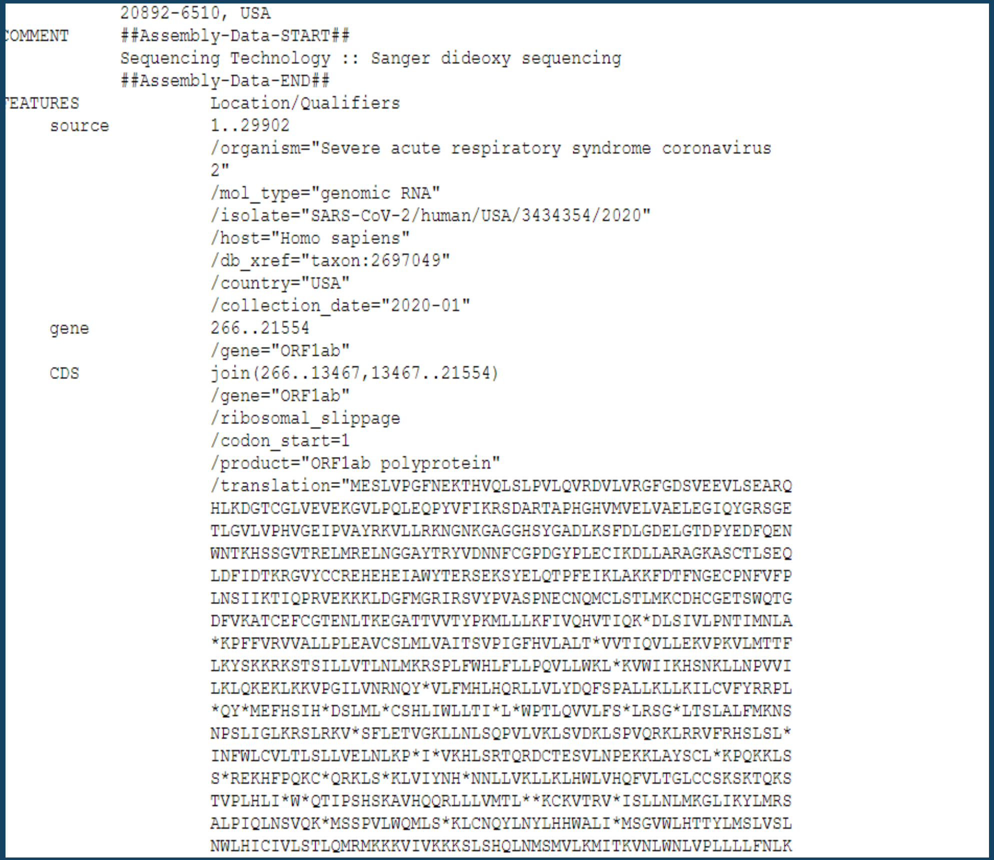


CDS Has Stop Codon

The predicted coding region contains an internal stop codon. This generally indicates errors in the nucleotide sequence or insufficient trimming of low quality sequence ends. Please upload the corrected sequences.

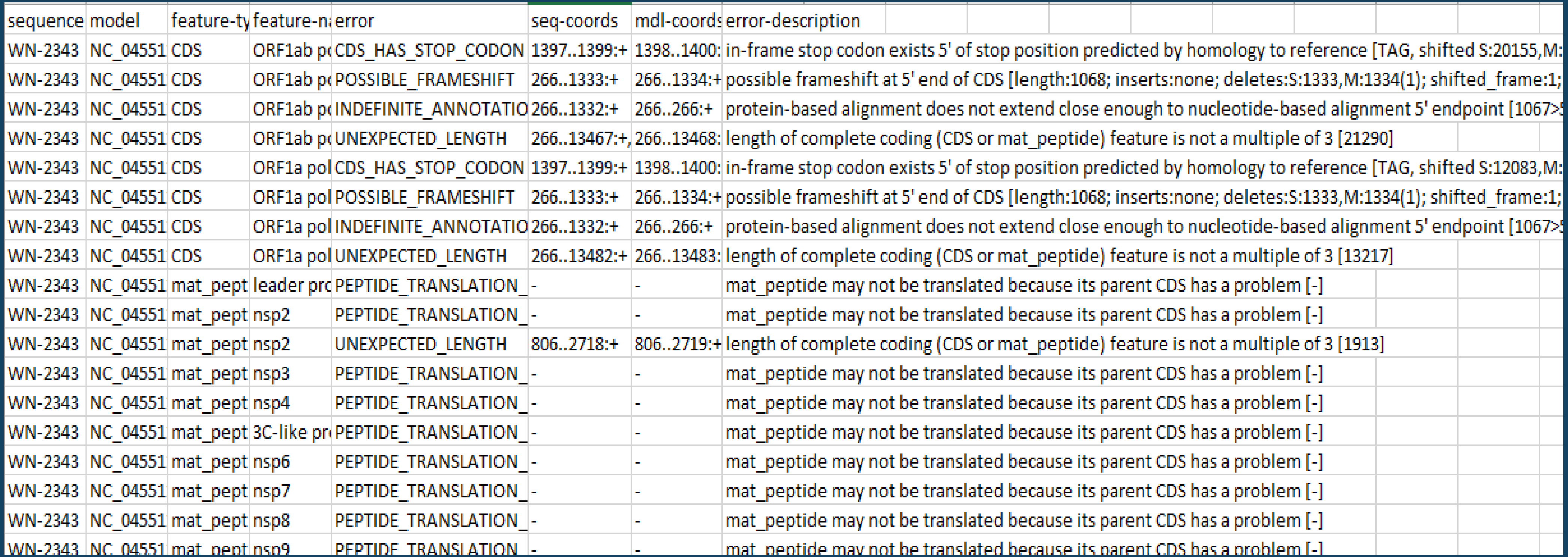


```
==============================================================

ERRORS

[] CDS_HAS_STOP_CODON

WN-2343

[] INDEFINITE_ANNOTATION_START

WN-2343

[] PEPTIDE_TRANSLATION_PROBLEM

WN-2343

[] CDS_HAS_FRAMESHIFT

WN-2343

[] UNEXPECTED_LENGTH

WN-2343


==============================================================
```



```
            20892-6510, USA
OMMENT      ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
EATURES              Location/Qualifiers
     source          1..29902
                     /organism="Severe acute respiratory syndrome coronavirus
                     2"
                     /mol_type="genomic RNA"
                     /isolate="SARS-CoV-2/human/USA/3434354/2020"
                     /host="Homo sapiens"
                     /db_xref="taxon:2697049"
                     /country="USA"
                     /collection_date="2020-01"
     gene            266..21554
                     /gene="ORF1ab"
     CDS             join(266..13467,13467..21554)
                     /gene="ORF1ab"
                     /ribosomal_slippage
                     /codon_start=1
                     /product="ORF1ab polyprotein"
                     /translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
                     HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
                     TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN
                     WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQ
                     LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
                     LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG
                     DFVKATCEFCGTENLTKEGATTVVTYPKMLLLKFIVQHVTIQK*DLSIVLPNTIMNLA
                     *KPFFVRVVALLPLEAVCSLMLVAITSVPIGFHVLALT*VVTIQVLLEKVPKVLMTTF
                     LKYSKKRKSTSILLVTLNLMKRSPLFWHLFLLPQVLLWKL*KVWIIKHSNKLLNPVVI
                     LKLQKEKLKKVPGILVNRNQY*VLFMHLHQRLLVLYDQFSPALLKLLKILCVFYRRPL
                     *QY*MEFHSIH*DSLML*CSHLIWLLTI*L*WPTLQVVLFS*LRSG*LTSLALFMKNS
                     NPSLIGLKRSLRKV*SFLETVGKLLNLSQPVLVKLSVDKLSPVQRKLRRVFRHSLSL*
                     INFWLCVLTLSLLVELNLKP*I*VKHLSRTQRDCTESVLNPEKKLAYSCL*KPQKKLS
                     S*REKHFPQKC*QRKLS*KLVIYNH*NNLLVKLLKLHWLVHQFVLTGLCCSKSKTQKS
                     TVPLHLI*W*QTIPSHSKAVHQQRLLLVMTL**KCKVTRV*ISLLNLMKGLIKYLMRS
                     ALPIQLNSVQK*MSSPVLWQMLS*KLCNQYLNYLHHWALI*MSGVWLHTTYLMSLVSL
                     NWLHICIVLSTLQMRMKKKVIVKKKSLSHQLNMSMVLKMITKVNLWNLVPLLLLFNLK
```

Detailed report from VADR



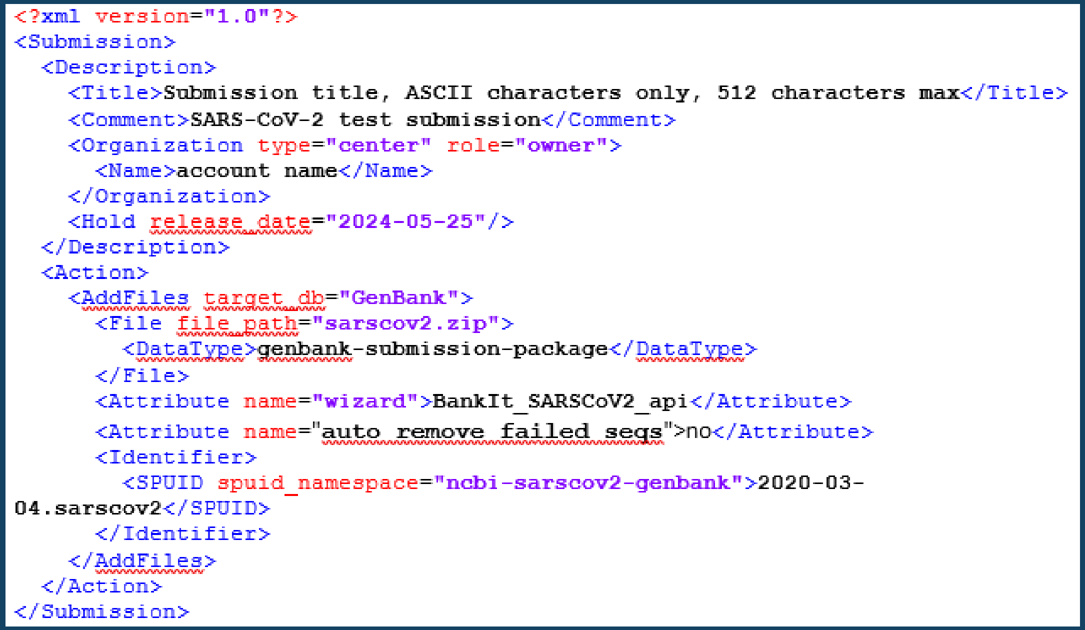
sequence	model	feature-ty	feature-na	error	seq-coords	mdl-coords	error-description
WN-2343	NC_04551	CDS	ORF1ab p	CDS_HAS_STOP_CODON	1397..1399:+	1398..1400:	in-frame stop codon exists 5' of stop position predicted by homology to reference [TAG, shifted S:20155,M:
WN-2343	NC_04551	CDS	ORF1ab p	POSSIBLE_FRAMESHIFT	266..1333:+	266..1334:+	possible frameshift at 5' end of CDS [length:1068; inserts:none; deletes:S:1333,M:1334(1); shifted_frame:1;
WN-2343	NC_04551	CDS	ORF1ab p	INDEFINITE_ANNOTATIO	266..1332:+	266..266:+	protein-based alignment does not extend close enough to nucleotide-based alignment 5' endpoint [1067>
WN-2343	NC_04551	CDS	ORF1ab p	UNEXPECTED_LENGTH	266..13467:+,	266..13468:	length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [21290]
WN-2343	NC_04551	CDS	ORF1a pol	CDS_HAS_STOP_CODON	1397..1399:+	1398..1400:	in-frame stop codon exists 5' of stop position predicted by homology to reference [TAG, shifted S:12083,M:
WN-2343	NC_04551	CDS	ORF1a pol	POSSIBLE_FRAMESHIFT	266..1333:+	266..1334:+	possible frameshift at 5' end of CDS [length:1068; inserts:none; deletes:S:1333,M:1334(1); shifted_frame:1;
WN-2343	NC_04551	CDS	ORF1a pol	INDEFINITE_ANNOTATIO	266..1332:+	266..266:+	protein-based alignment does not extend close enough to nucleotide-based alignment 5' endpoint [1067>
WN-2343	NC_04551	CDS	ORF1a pol	UNEXPECTED_LENGTH	266..13482:+	266..13483:	length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [13217]
WN-2343	NC_04551	mat_pept	leader prc	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp2	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp2	UNEXPECTED_LENGTH	806..2718:+	806..2719:+	length of complete coding (CDS or mat_peptide) feature is not a multiple of 3 [1913]
WN-2343	NC_04551	mat_pept	nsp3	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp4	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	3C-like pr	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp6	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp7	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp8	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]
WN-2343	NC_04551	mat_pept	nsp9	PEPTIDE_TRANSLATION_	-	-	mat_peptide may not be translated because its parent CDS has a problem [-]

Note on mutations in non-essential genes

Programmatic submission

- Available for GenBank & SRA submission
- Recommended if you submit a large volume of data, regularly
- Submission .xml file uploaded via FTP
- Accessions, files available on Submission Portal: https://submit.ncbi.nlm.nih.gov/subs/api/
- Contact NCBI at the emails on the ‘Help’ slide to explore this option

Example GenBank XML for SARS-CoV-2



```
<?xml version="1.0"?>
<Submission>
  <Description>
    <Title>Submission title, ASCII characters only, 512 characters max</Title>
    <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>
      <Attribute name="auto_remove_failed_seqs">no</Attribute>
      <Identifier>
        <SPUID spuid_namespace="ncbi-sarscov2-genbank">2020-03-
04.sarscov2</SPUID>
      </Identifier>
    </AddFiles>
  </Action>
</Submission>
```

DEMO

We’re here for you!

- Submitter feedback has factored into numerous improvements during the pandemic
  - “Auto-remove” error sequences GenBank feature
  - VADR enhancements per virus evolution
  - FASTAedit command line tool
  - Programmatic submission option(s)
  - File upload drag-and-drop
  - Help documentation edits

- Always accepting volunteers for testing & feedback!

Help

- Get started
  - https://submit.ncbi.nlm.nih.gov/sarscov2/

- 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/ncbi/vadr

Thank you!