Version 1

Jul 08, 2020

SARS-CoV-2 NCBI assembly submission protocol: GenBank V.1

Ruth Timme[1], Heather M Blankenship[2], Erin L Young[3], Emma Griffiths[4], Duncan MacCannell[5], Stacia Wyman[6]

[1]US Food and Drug Administration; [2]Michigan Department of Health and Human Services; [3]Utah Public Health Laboratoraty; [4]University of British Columbia; [5]Centers for Disease Control and Prevention; [6]Innovative Genomics Institute, UC Berkeley

1 Works for me dx.doi.org/10.17504/protocols.io.bg2tjyen

GenomeTrakr Coronavirus Method Development Community 1 more workspace

Ruth Timme
US Food and Drug Administration

ABSTRACT

PURPOSE:
This protocol covers the steps for submitting a SARS-CoV-2 assembly to NCBI's GenBank

For new submitters, there's quite a bit of groundwork that needs to be established before a laboratory can start its first data submission. We recommend that one person in the laboratory take a few days to get everything set up in advance of when you expect to do your first data submission.

Two protocols cover the PHA4GE guidance for SARS-CoV-2 submission to NCBI (Raw sequence data, metadata, and assemblies).

If you need a pipeline for frequent or large volume submissions, follow Step 1 in the **SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject** to get your NCBI submission environment established, then contact gb-admin@ncbi.nlm.nih.gov to set up an account for submitting through the API.

These protocols cover submission using NCBI's Submission Portal web-interface.

Complete in order (1 then 2):

1. SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject

- Step-by-step instructions for establishing a new NCBI laboratory submission account and for creating and linking a new BioProject to an existing umbrella effort.
- Submit SARS-CoV-2 raw data to SRA (Sequence Read Archive) and metadata to BioSample.

2. SARS-CoV-2 NCBI assembly submission protocol: GenBank (included protocol)

Required: established BioProject and BioSamples

- Submit SARS-CoV-2 consensus sequences to NCBI GenBank, linking to existing BioProject, BioSamples, and raw data.

DOI

dx.doi.org/10.17504/protocols.io.bg2tjyen

PROTOCOL CITATION

Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman 2020. SARS-CoV-2 NCBI assembly submission protocol: GenBank. **protocols.io** dx.doi.org/10.17504/protocols.io.bg2tjyen

KEYWORDS

NCBI submission, pathogen surveillance, SARS-CoV-2, covid-19, genomic epidemiology, GenBank

LICENSE

CREATED

Jun 01, 2020

LAST MODIFIED

Jul 08, 2020

PROTOCOL INTEGER ID

37683

BEFORE STARTING

This protocol has two sections:

Section 1: ensuring your NCBI submission environment is established
Section 2: SARS-CoV-2 submission of assemblies or consensus sequences to GenBank.

Associated protocols:

- SOP for populating the NCBI submission templates (e.g. source modifiers for GenBank)
- NCBI submission to BioProject, SRA, and BioSample. Also includes NCBI account set-up for new users (Step 1)
- NCBI Data Curation protocol for making updates, corrections, or retractions to your data.

Link to PHA4GE contextual data specification

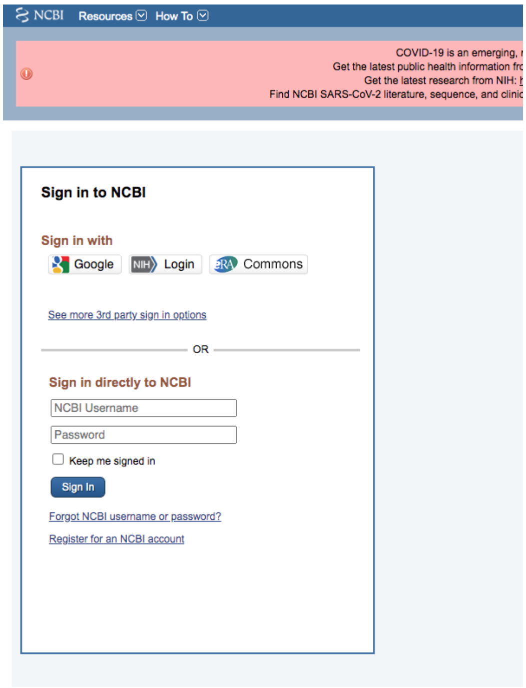
"**Ingredients**" to have in place before starting your submissions

1 **1.1:** Ensure you have a working NCBI user account
1.2: Identify your NCBI submission user group or establish a new one if necessary.
1.3: Bookmark the link to your submission portal
1.4. BioSample + BioProject assessions in-hand

After these steps are complete you can proceed with data submission in **Step 2.**

1.1 Sign in to your **NCBI user account**: https://www.ncbi.nlm.nih.gov/account/

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen



COVID-19 is an emerging,
Get the latest public health information fr
Get the latest research from NIH:
Find NCBI SARS-CoV-2 literature, sequence, and clinic

Sign in to NCBI

Sign in with

See more 3rd party sign in options

OR

Sign in directly to NCBI

NCBI Username

Password

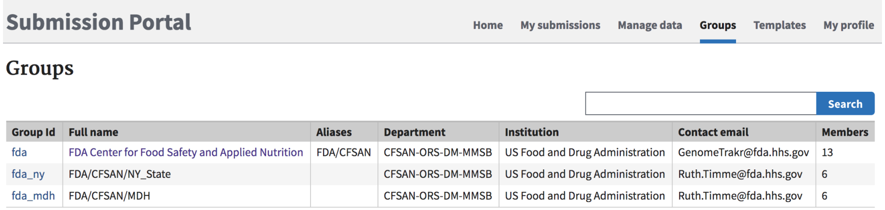
□ Keep me signed in

Forgot NCBI username or password?

Register for an NCBI account

1.2 **Ensure you have an NCBI user group established and correct permissions are assigned for you to submit.**

List of submission groups: https://submit.ncbi.nlm.nih.gov/groups/



Groups

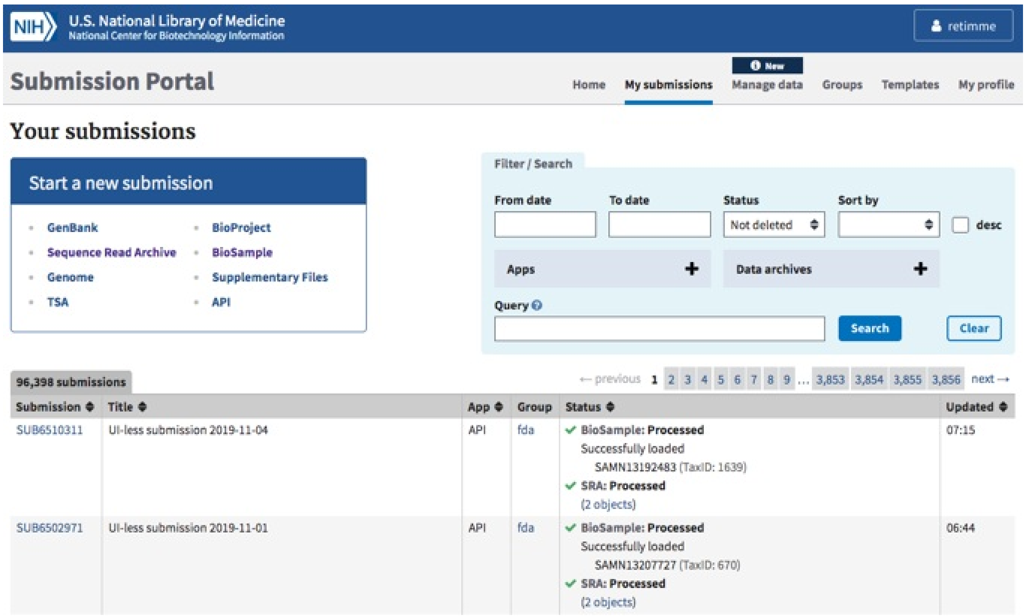
Group Id	Full name	Aliases	Department	Institution	Contact email	Members
fda	FDA Center for Food Safety and Applied Nutrition	FDA/CFSAN	CFSAN-ORS-DM-MMSB	US Food and Drug Administration	GenomeTrakr@fda.hhs.gov	13
fda_ny	FDA/CFSAN/NY_State		CFSAN-ORS-DM-MMSB	US Food and Drug Administration	Ruth.Timme@fda.hhs.gov	6
fda_mdh	FDA/CFSAN/MDH		CFSAN-ORS-DM-MMSB	US Food and Drug Administration	Ruth.Timme@fda.hhs.gov	6

If you don't have a submission group established, please follow this protocol to create one for your laboratory group:

https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7bjrin

1.3 **Bookmark "my submissions"** at NCBI: https://submit.ncbi.nlm.nih.gov/subs/. This is your landing page for all new NCBI submissions.

If you see a blank page with a yellow box in the upper right corner saying "please login", click this link and login using the credentials created in **Step 1.1**.


NIH U.S. National Library of Medicine
National Center for Biotechnology Information
retimme
Submission Portal
New
Home
My submissions
Manage data
Groups
Templates
My profile

Your submissions

Start a new submission

- GenBank
- Sequence Read Archive
- Genome
- TSA
- BioProject
- BioSample
- Supplementary Files
- API

Filter / Search

From date | To date | Status: Not deleted | Sort by | □ desc

Apps ✚ | Data archives ✚

Query

Search | Clear

96,398 submissions

← previous 1 2 3 4 5 6 7 8 9 … 3,853 3,854 3,855 3,856 next →

Submission	Title	App	Group	Status	Updated
SUB6510311	UI-less submission 2019-11-04	API	fda	✓ BioSample: Processed Successfully loaded SAMN13192483 (TaxID: 1639) ✓ SRA: Processed (2 objects)	07:15
SUB6502971	UI-less submission 2019-11-01	API	fda	✓ BioSample: Processed Successfully loaded SAMN13207727 (TaxID: 670) ✓ SRA: Processed (2 objects)	06:44

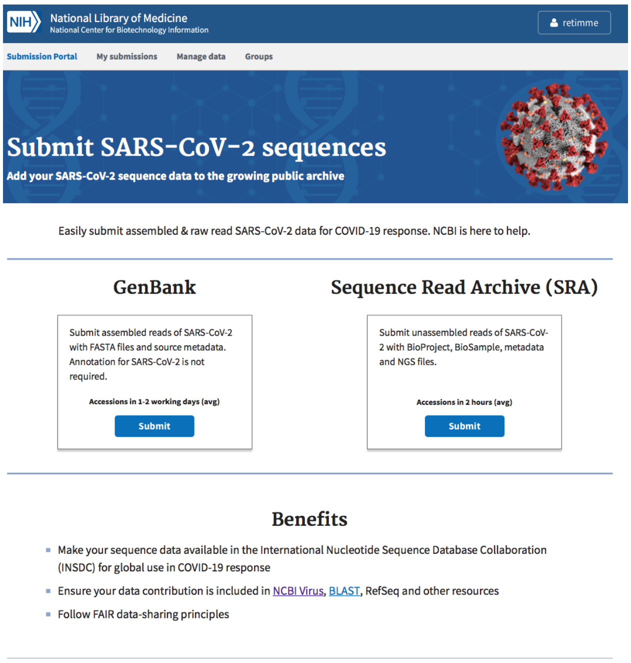
1.4 **1. Identify your lab's BioProject accession.** Does your laboratory have an established BioProject for this effort?

If not please follow instructions in **Step 3** of https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7bjrin for creating a new one.

Data submission (assemblies to GenBank)

2 GenBank assembly submission of SARS-CoV-2:

SARS-CoV-2 landing page: https://submit.ncbi.nlm.nih.gov/sarscov2/


NIH National Library of Medicine
National Center for Biotechnology Information
retimme
Submission Portal
My submissions
Manage data
Groups

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 for COVID-19 response. NCBI is here to help.

GenBank

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

Accessions in 1-2 working days (avg)

Submit

Sequence Read Archive (SRA)

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

Accessions in 2 hours (avg)

Submit

Benefits

- Make your sequence data available in the International Nucleotide Sequence Database Collaboration (INSDC) for global use in COVID-19 response
- Ensure your data contribution is included in NCBI Virus, BLAST, RefSeq and other resources
- Follow FAIR data-sharing principles

Click "submit" under GenBank.

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen

2.1 **For all sequences you intend to submit at this time:**

1. Gather associated BioSample accessions and metadata previously registered in https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7bjrin along with three pieces of information describing the sequencing method and assembly:

1. **Sequencing method.** Populate with the PHA4GE field "sequencing instrument"
2. **Assembly program/pipeline.** Populate with the name from the PHA4GE field "assembly method"
3. **Version** of the assembly program. Populate with the version from the PHA4GE field "assembly method"

BioSample Accession	sample_name	seq. method	assembly program	assembly version or date
SAMN15460792	CA-IGI-0042	MinION	ARTIC-nCoV-bioinformaticsSOP	1.1.0
SAMN15460793	CA-IGI-0031	MinION	ARTIC-nCoV-bioinformaticsSOP	1.1.0

Example of two BioSamples and associated sample_name IDs

2. Concatenate all SARS-CoV-2 consensus sequences into a single fasta file, where the fasta headers contain the "sample_name" submitted to the BioSample.

Example FASTA file for two sequence submissions:

```
>CA-IGI-0042
ATCGATCGGTACCTAAGGATCGATCGGTACCTAAGGATCGATCGGTACCTAAGG....
>CA-IGI-0031
ATCGATCGGTACCTAAGGATCGATCGGTACCTAAGGATCGATCGGTACCTAAGG....
```

2.2 **Download and populate the PHA4GE GenBank source modifiers template:**

□ **GenBank-source_modifiers-PHA4GE_200708.xlsx**

Guidance:

- Follow **Step 4** in SOP for populating the NCBI submission templates for populating the source modifiers.
- Refer to **PHA4GE contextual data specification** where relevant.

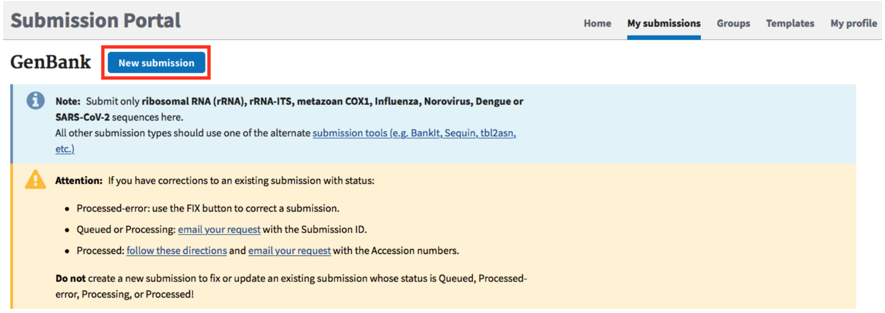
Populate the metadata spreadsheets for each isolate you intend to submit (you can submit metadata for a single isolate, entire MiSeq run, or for a large collection of isolates you intend to submit in batch).

**Ensure that the BioProject and BioSample(s) were registered using the same NCBI user group. If you are not listed as an owner on the BioProject/BioSample(s) you will not be able to properly link the new assembly data to existing records.

Save the excel spreadsheet as a tab-delimited text file (.tsv) and ensure that the date field is formatted correctly (e.g. 2020-04-20) in the text file.

2.3 **Click the "New submission" box.**

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen



Submission Portal

GenBank

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, Sequin, tbl2asn, etc.)

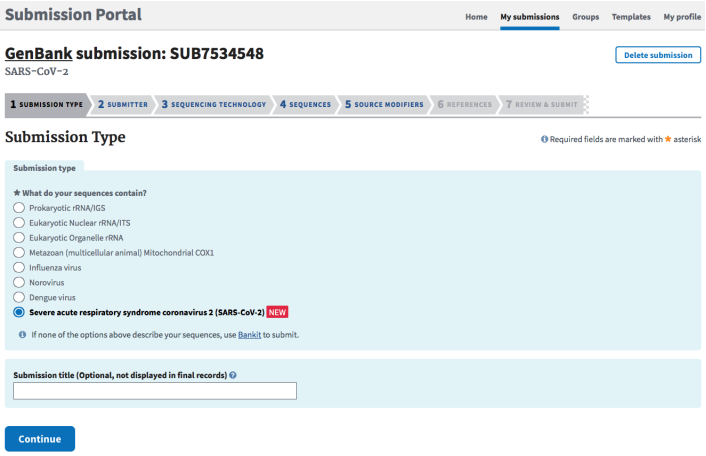
Attention: If you have corrections to an existing submission with status:

- Processed-error: use the FIX button to correct a submission.
- Queued or Processing: email your request with the Submission ID.
- Processed: follow these directions and email your request with the Accession numbers.

Do not create a new submission to fix or update an existing submission whose status is Queued, Processed-error, Processing, or Processed!

2.4 SUBMISSION TYPE tab:

Select "SARS-CoV-2" option.



Submission Portal

GenBank submission: SUB7534548
SARS-CoV-2

Submission Type

Required fields are marked with * asterisk

Submission type

* What do your sequences contain?
○ Prokaryotic rRNA/IGS
○ Eukaryotic Nuclear rRNA/ITS
○ Eukaryotic Organelle rRNA
○ Metazoan (multicellular animal) Mitochondrial COX1
○ Influenza virus
○ Norovirus
○ Dengue virus
● Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) NEW

If none of the options above describe your sequences, use Bankit to submit.

Submission title (Optional, not displayed in final records)

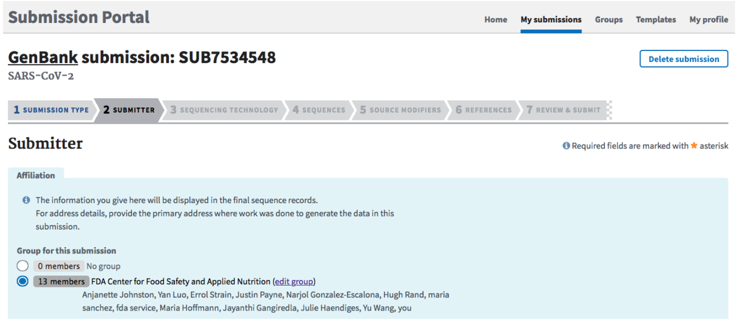
2.5 SUBMITTER tab:

Populate with submitter info. The “submitter” is the name of the person, or user group, who is physically doing the submissions, not a supervisor or PI.

**Must be the same person or group that submitted the associated BioSamples and BioProject.

Select the appropriate submission group name for your laboratory and check the contact information below.

If you do not have a submission group available to click, see **Steps 1.2-1.3 in the SRA submission protocol to establish a new one for your laboratory, or to add your name to a group already established for your lab.



Submission Portal

GenBank submission: SUB7534548
SARS-CoV-2

Submitter

Required fields are marked with * asterisk

Affiliation

The information you give here will be displayed in the final sequence records.
For address details, provide the primary address where work was done to generate the data in this submission.

Group for this submission
○ 0 members No group
● 13 members FDA Center for Food Safety and Applied Nutrition (edit group)
Anjanette Johnston, Yan Luo, Errol Strain, Justin Payne, Narjol Gonzalez-Escalona, Hugh Rand, maria sanchez, fda service, Maria Hoffmann, Jayanthi Gangiredla, Julie Haendiges, Yu Wang, you

Click "Continue" to proceed.

2.6 SEQUENCING TECHNOLOGY tab:

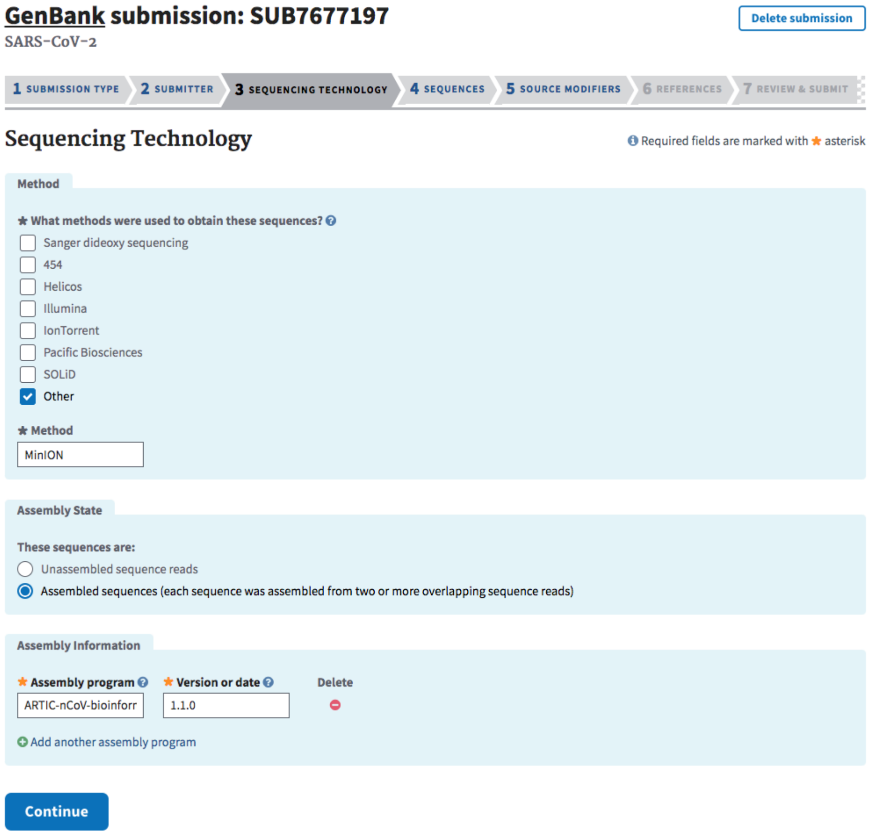
This information will get populated as a structured comment on the GenBank record.

Pull the sequencing method and assembly information gathered in **Step 2.1.**

Method: sequencing technology or platform.

Assembly State: Click "Assembled sequences".

Assembly information: Specify program/pipeline AND version.



GenBank submission: SUB7677197

SARS-CoV-2

Sequencing Technology

Required fields are marked with * asterisk

Method

* What methods were used to obtain these sequences?

□ Sanger dideoxy sequencing
□ 454
□ Helicos
□ Illumina
□ IonTorrent
□ Pacific Biosciences
□ SOLiD
☑ Other

* Method: MinION

Assembly State

These sequences are:

○ Unassembled sequence reads
● Assembled sequences (each sequence was assembled from two or more overlapping sequence reads)

Assembly Information

* Assembly program: ARTIC-nCoV-bioinforr
* Version or date: 1.1.0

2.7 SEQUENCES tab:

Release date: Click "Release immediately following processing" for all routine surveillance isolates.

Sequences:

Sequences can be uploaded one at a time (one per submission), or as a batch upload in a single concatenated FASTA file (https://submit.ncbi.nlm.nih.gov/genbank/help/#fasta) when you are submitting multiple isolates in one submission. See **Step 2.1** for guidance on formating your FASTA

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen

file.

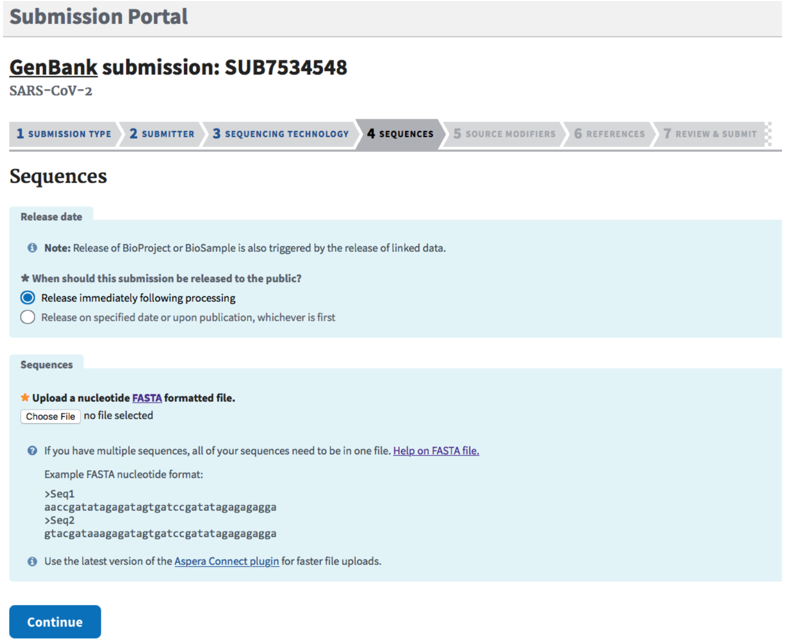
Fasta headers must include a unique ID that links the sequence to the source modifiers

For example:

FASTA header:
>CA-IGI-0042

Source modifier template
ID from Sequence_ID column in metadata workbook: CA-IGI-0042

Click "Choose File" to browse and upload your .fasta file:



Submission Portal

GenBank submission: SUB7534548
SARS-CoV-2

Sequences

Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

★ When should this submission be released to the public?
● Release immediately following processing
○ Release on specified date or upon publication, whichever is first

Sequences

★ Upload a nucleotide FASTA formatted file.
Choose File no file selected

If you have multiple sequences, all of your sequences need to be in one file. Help on FASTA file.

Example FASTA nucleotide format:

```
>Seq1
aaccgatatagagatagtgatccgatatagagagagga
>Seq2
gtacgataaagagatagtgatccgatatagagagagga
```

Use the latest version of the Aspera Connect plugin for faster file uploads.

Click "Continue" and respond to any validation issues.

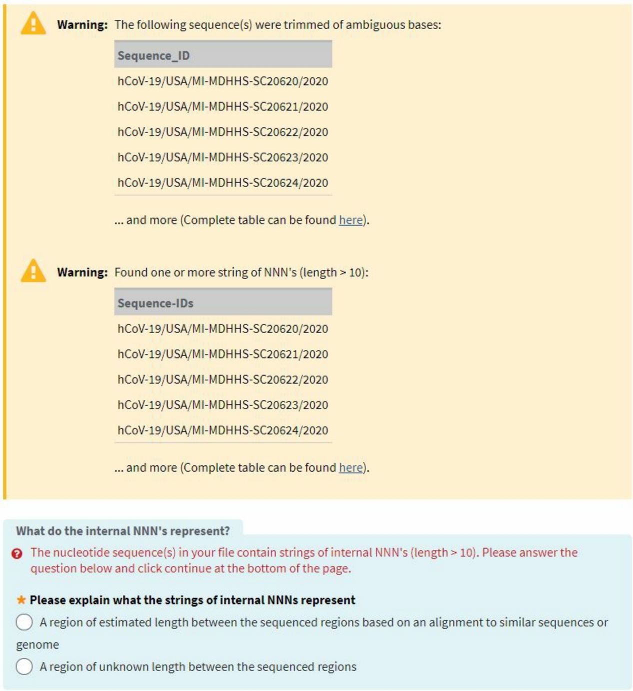
Common validation issues:

Ambiguous bases were trimmed warning. Ambiguous bases are non- A, T, G, or Cs. NCBI trims terminal Ns first at the 5' end, then looks to see if 50% of the first 10 bases are ambiguous and trim to last ambiguous. If more than 30% of the first 50 are ambiguous, we trim to the last ambiguous and then recheck the 5' end. If that is fine, we follow the same steps on the 3' end. This procedure is run again if we trimmed vector from an end. NCBI removes sequences that are greater than 50% ambiguous after the trimming. They also remove sequences with internal vector.

String of NNNs: If your assembly contains strings of internal NNNs (from mapping to a reference genome), you will get a warning asking for you for more information:

Click "A region of estimated length between the sequenced regions based on an alignment to similar sequences or genome" if the NNNs were caused by the reference-based assembly.

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen



⚠ **Warning:** The following sequence(s) were trimmed of ambiguous bases:

Sequence_ID
hCoV-19/USA/MI-MDHHS-SC20620/2020
hCoV-19/USA/MI-MDHHS-SC20621/2020
hCoV-19/USA/MI-MDHHS-SC20622/2020
hCoV-19/USA/MI-MDHHS-SC20623/2020
hCoV-19/USA/MI-MDHHS-SC20624/2020

... and more (Complete table can be found here).

⚠ **Warning:** Found one or more string of NNN's (length > 10):

Sequence-IDs
hCoV-19/USA/MI-MDHHS-SC20620/2020
hCoV-19/USA/MI-MDHHS-SC20621/2020
hCoV-19/USA/MI-MDHHS-SC20622/2020
hCoV-19/USA/MI-MDHHS-SC20623/2020
hCoV-19/USA/MI-MDHHS-SC20624/2020

... and more (Complete table can be found here).

What do the internal NNN's represent?

The nucleotide sequence(s) in your file contain strings of internal NNN's (length > 10). Please answer the question below and click continue at the bottom of the page.

★ **Please explain what the strings of internal NNNs represent**

○ A region of estimated length between the sequenced regions based on an alignment to similar sequences or genome

○ A region of unknown length between the sequenced regions

Click "Continue" again.

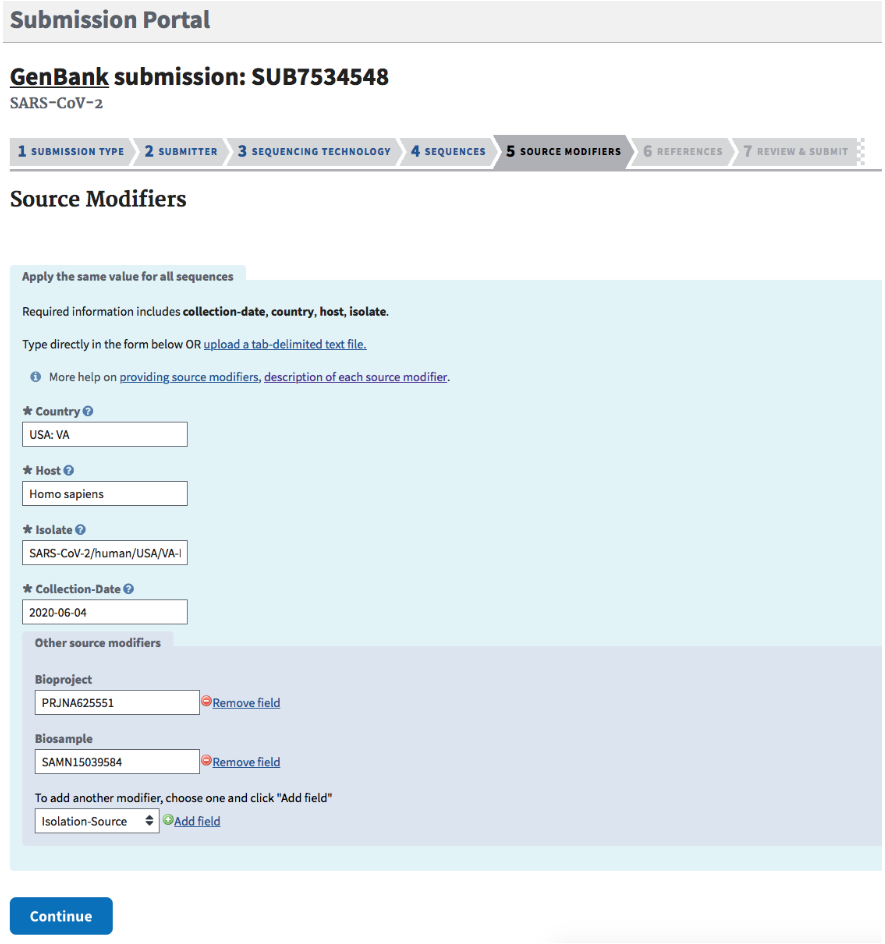
2.8 SOURCE MODIFIERS tab:

Guidance for populating this metadata outlined in **Step 2.2.**

For a single submission: In the "Other source modifiers" Box, click Add field to add "BioProject" and "BioSample". Then populate these six fields following guidance in SOP for populating the NCBI submission templates, **Step 4**.

For a batch submission. Upload the csv file created from populating the **PHA4GE GenBank source modifiers template** in **Step 2.2.** Upload this file by clicking on the "upload a tab-delimited text file" link. Ensure that the first column in this spreadsheet, "Sequence_ID" contains an ID that matches *exactly* the ID used in your FASTA file headers.

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen



Submission Portal

GenBank submission: SUB7534548

SARS-CoV-2

1 SUBMISSION TYPE 2 SUBMITTER 3 SEQUENCING TECHNOLOGY 4 SEQUENCES 5 SOURCE MODIFIERS 6 REFERENCES 7 REVIEW & SUBMIT

Source Modifiers

Apply the same value for all sequences

Required information includes **collection-date, country, host, isolate.**

Type directly in the form below OR upload a tab-delimited text file.

More help on providing source modifiers, description of each source modifier.

★ Country: USA: VA

★ Host: Homo sapiens

★ Isolate: SARS-CoV-2/human/USA/VA-I

★ Collection-Date: 2020-06-04

Other source modifiers

Bioproject: PRJNA625551

Biosample: SAMN15039584

To add another modifier, choose one and click "Add field"

Isolation-Source

Click Continue.

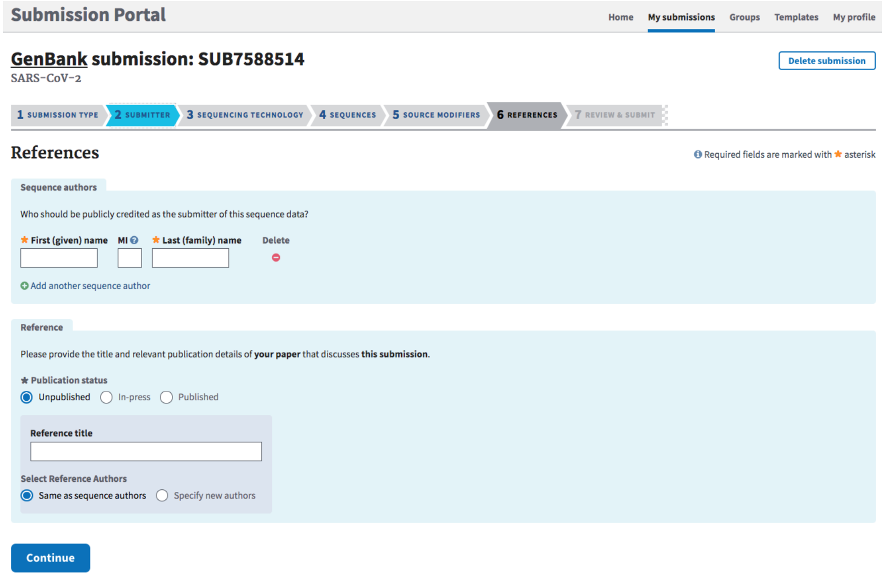
ERRORS: If you are not listed as an owner on the BioProject/BioSample you will see an error here stating that these are "Unknown". If you do not have a submission group available to click, see **Steps 1.2-1.3** in the SRA submission protocol to establish a new one for your laboratory, or to add your name to a group already established for your lab.

2.9 **REFERENCES tab:**

Sequence Authors: Enter names here from your NCBI submission user group (can be a sub-set of the names or the full submission group list).

Reference: For routine surveillance submissions choose "Unpublished", leave "Reference title" blank, and choose "same as sequence authors".

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen



GenBank submission: SUB7588514
SARS-CoV-2

References

Required fields are marked with * asterisk

Sequence authors

Who should be publicly credited as the submitter of this sequence data?

* First (given) name | MI | * Last (family) name | Delete

Add another sequence author

Reference

Please provide the title and relevant publication details of **your paper** that discusses **this submission.**

* Publication status
● Unpublished ○ In-press ○ Published

Reference title

Select Reference Authors
● Same as sequence authors ○ Specify new authors

Click Continue.

2.10 REVIEW & SUBMIT tab:

Check over entire submission, then click submit.

2.11 GenBank accessions:

The status of your submission will be updated once it is processed (track the status of your submissions under the "My Submissions" tab: https://submit.ncbi.nlm.nih.gov/subs/).

Sequences with no annotation issues will be listed as **Processed** and the GenBank accessions will be emailed to you and listed on the submissions page. Submissions with annotation discrepancies will be marked as **Error** and a Fix button will appear. A report is emailed to you and listed on the submissions page with the detailed issues. If the data is incorrect, click the Fix button and you will return to the sequences page of your submission to upload a corrected file.

If you have evidence that the discrepancy is due to a naturally occurring mutation, send an email to **gb-admin@ncbi.nlm.nih.gov** with the SUB number and evidence.

2.12 Important data stewardship and curation notes:

- Develop an internal method for storing and tracking your GenBank accessions! They are required for making future updates to your records.

For updates to your GenBank records follow the NCBI Curation Protocol hosted by GenomeTrakr: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase

Citation: Ruth Timme, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SARS-CoV-2 NCBI assembly submission protocol: GenBank. https://dx.doi.org/10.17504/protocols.io.bg2tjyen