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

**Ruth E Timme**, Heather M Blankenship, Erin L Young, Emma Griffiths, Duncan MacCannell, Stacia Wyman

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

**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)**
   - 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**


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

**KEYWORDS**

NCBI submission, pathogen surveillance, SARS-CoV-2, covid-19, genomic epidemiology, GenBank
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. Ensure you have a working NCBI user account
2. Identify your NCBI submission user group or establish a new one if necessary.
3. Bookmark the link to your submission portal
4. BioSample + BioProject assessments 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/
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/

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


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


---

**Data submission (assemblies to GenBank)**

2. **GenBank assembly submission of SARS-CoV-2:**


---

**Citation:** Ruth E 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](https://dx.doi.org/10.17504/protocols.io.bg2tjyen)

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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-bf7birin](https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7birin) 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"

<table>
<thead>
<tr>
<th>BioSample Accession</th>
<th>sample_name</th>
<th>seq. method</th>
<th>assembly program</th>
<th>assembly version or date</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMN15460792</td>
<td>CA-IGI-0042</td>
<td>MinION</td>
<td>ARTIC-nCoV-bioinformaticsSOP</td>
<td>1.1.0</td>
</tr>
<tr>
<td>SAMN15460793</td>
<td>CA-IGI-0031</td>
<td>MinION</td>
<td>ARTIC-nCoV-bioinformaticsSOP</td>
<td>1.1.0</td>
</tr>
</tbody>
</table>

   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](image)

   **Guidance:**
   - Follow Step 4 in [SOP for populating the NCBI submission templates](https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7birin) for populating the source modifiers.
   - Refer to [PHA4GE contextual data specification](https://www.protocols.io/edit/sars-cov-2-ncbi-submission-protocol-sra-biosample-bf7birin) 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 E 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.bg2tyen](https://dx.doi.org/10.17504/protocols.io.bg2tyen)

This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
2.4 **SUBMISSION TYPE tab:**

Select "SARS-CoV-2" option.

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.**
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.

---

Citation: Ruth E 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](https://dx.doi.org/10.17504/protocols.io.bg2tjyen)

This is an open access protocol distributed under the terms of the Creative Commons Attribution License ([https://creativecommons.org/licenses/by/4.0/](https://creativecommons.org/licenses/by/4.0/)), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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 formatting your FASTA 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:

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.

![Warning](image)

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.

![Warning](image)
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](https://www.ncbi.nlm.nih.gov/sra/submission/) 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'.
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-bacaisa