NCBI

Jul 08, 2020

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

⅄ Forked from NCBI submission protocol for microbial pathogen surveillance

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

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

| 1 | Works for me | dx.doi.org/10.17504/protocols.io.bf7bjrin |

GenomeTrakr
Tech. support email: **genomeTrakr@fda.hhs.gov**

Ruth Timme
US Food and Drug Administration

ABSTRACT

PURPOSE:
This is a SARS-CoV-2 specific protocol that covers the steps needed to establish a new NCBI submission environment for your laboratory, including the creation of new BioProject(s) and submission groups. Once these are step up, the protocol then walks through the process for submitting raw reads to SRA and sample metadata to BioSample through the Submission portal.

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.

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 (included protocol)
- 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.
- SARS-CoV-2 raw data submission to SRA (Sequence Read Archive) and metadata to BioSample.

2. SARS-CoV-2 NCBI assembly submission protocol: GenBank
Required: established BioProject and BioSamples
- Submit SARS-CoV-2 assemblies to NCBI GenBank, linking to existing BioProject, BioSamples, and raw data.

DOI

dx.doi.org/10.17504/protocols.io.bf7bjrin

PROTOCOL CITATION

Ruth Timme, Emma Griffiths, Heather M Blankenship, Erin L Young, Duncan MacCannell, Stacia Wyman 2020. SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject. **protocols.io**
dx.doi.org/10.17504/protocols.io.bf7bjrin

FORK FROM

Forked from NCBI submission protocol for microbial pathogen surveillance, Ruth Timme

KEYWORDS

BEFORE STARTING

This protocol has three sections:

- **Section 1**: Setting up NCBI accounts (for new users)
- **Section 2**: Single-step data submission to **SRA** for raw reads and associated sequence metadata and to **BioSample** for sample metadata
- **Section 3:** Detailed steps for creating a BioProject (usually done once during the account set-up)

Associated protocols:
- **SOP for populating the NCBI submission templates** for SARS-CoV-2 (e.g. BioSample and SRA metadata)
- GenBank submission protocol for submitting SARS-CoV-2 assemblies or consensus sequences.
- 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 **Set up a new NCBI submission environment for your lab**
    1.1: Create an NCBI user account
    1.2: Set up an NCBI submission user group for your lab
    1.4: Bookmark the link to your submission portal
    1.5. Identify or establish new BioProjects (detailed in **Step 3**)

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

   1.1   **Create an NCBI user account** at NCBI: https://www.ncbi.nlm.nih.gov/account

This will be your own individual user account at NCBI.

> 📄 Some laboratory groups might choose to create a single NCBI user account that everyone in the laboratory uses.

1.2 Establish an NCBI submission user group for your laboratory.

We recommend using this user group for all NCBI submissions related to microbial genome surveillance. This will link your laboratory's NCBI data ownership to the *user group* and not to individuals, allowing anyone in the current group to perform updates or retractions and answer inquiries from the NCBI staff, even if there's been a complete turnover of staff since the original data submission.

User groups also ensure consistent data ownership across BioProjects, BioSamples, and sequence data. If your laboratory has non-overlapping research groups submitting and managing data at NCBI, multiple user groups can be established to track these efforts separately.

Your laboratory might already have a submission group established! Check the "Group" tab in the submission portal, https://submit.ncbi.nlm.nih.gov/groups/. Ask your colleagues to do the same thing, to ensure your laboratory doesn't already have one in place.

Creating a new submission group:

1. Submit an email request to **submit-help@ncbi.nlm.nih.gov** containing the following information:

> 📄 "*Dear NCBI help staff,*

Please establish a new user group for my laboratory.
I'm including the following information to help set up the group:

Short name of the group (abbreviation, e.g. "fda_ny")
Full name of the group (e.g. "NY Wadsworth submission group")
Contact email(s) to start the group
Institution and department or group
Physical address including country
Primary contact person, first and last name plus email.

* if you have existing submissions you want to be converted, please request the ownership change in this email.
i.e., Please assign this new user group to the following BioProjects and linked data.

Thank you,"

2. Look for an email reply entitled "NCBI Submission Portal Group invitation" and click on the enclosed link to accept the invitation.

1.3   Managing your NCBI submission user group.

After a user group has been established it can be edited for membership and permissions by clicking in the "group" tab of the submission portal (https://submit.ncbi.nlm.nih.gov/groups/), then on the Group Id hyperlink, e.g 'fda_ny' in the above example.

Users with admin privileges can update contact information in the "profile" tab and membership in the "Members" tab. New members can be invited by clicking on the "Invite members" link.

This user group should be kept up-to-date as members enter and leave the laboratory.

Permissions levels:
- READ: primarily for collaborators who can see the submissions, but not edit them.
- MODIFY, SUBMIT, DELETE: Permissions to submit, modify, or retract data (members usually have all or none of these permissions)
- ADMIN: Can invite or remove members of the submission group. Ensure that at least one (or more) members of your group have ADMIN privileges.

1.4   Bookmark **"my submissions"** at NCBI: https://submit.ncbi.nlm.nih.gov/subs/. This is the page where you view and track all of your past 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**.

1.5 Identify or establish new BioProjects (Umbrella and/or Data BioProjects)

Umbrella BioProjects. If you are already part of a surveillance network, (e.g. SPHERES, COG-UK, CanCOGeN, etc) you should use one of their established umbrella bioprojects. For reference, here are some of the umbrella projects established for SARS-CoV-2 surveillance:

SPHERES (US): PRJNA615625
CanCOGeN (Canada), PRJNA623807

If you need to establish a new umbrella BioProject, follow instructions in **Step 3** with modifications for creating a new Umbrella BioProject (**Step 3.12**).

Data BioProjects. Does your laboratory have an established data BioProject for this effort? If not please follow instructions in **Step 3** for creating a new one.

Countries with single data projects (not exhaustive):
COG-UK (United Kingdom): PRJEB37886
Turkey: PRJNA636004
Switzerland: PRJEB38472
South Africa: PRJNA624358

📄 **More information:**

Learn more about **data** vs **umbrella** BioProjects in Step 3
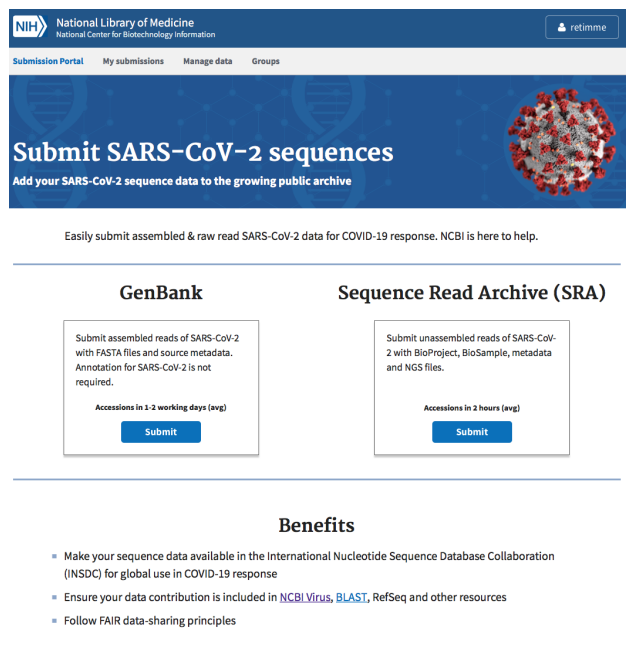

Data submission (BioSample and SRA)

2 Data submission (source metadata and sequence data):

This protocol follows a one-step data submission process where the source metadata is submitted at the same time as

the sequence data.

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

Click "submit" under the Sequence Read Archive (SRA) option

2.1 Download and populate the sample (BioSample) and sequence (SRA) metadata templates:

1. BioSample: PHA4GE pathogen template. (This is an extension of NCBI's combined pathogen package)

    ☐ **BioSample-Pathogen-PHA4GE_200708.xlsx**

2. SRA metadata template for raw sequence data (download directly from NCBI, no modification needed for PHA4GE compliance).

For each SARS-CoV-2 isolate intended for submission, you will create one BioSample and one associated SRA entry. You can submit a single isolate at a time, or as a batch from an entire sequencing run or collection.

> 📄 **Follow PHA4GE guidelines for populating the metadata templates:**
>
> **PHA4GE SARS-CoV-2 Contextual Data Specification for organizing public and non-public SARS-CoV-2 metadata.** **Read through this specification thoroughly before starting your submission.**
>
> **Protocols.io Guidance for translating the PHA4GE spec into NCBI submission templates.** A brief summary of the above specification for NCBI-relevant attributes.
>
> **If your lab is storing metadata in a single spreadsheet (vs a LIMS) you could create one master NCBI metadata submission spreadsheet for *all* isolates, then cut/paste into the

respective submission templates.

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

2.3 **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.

Select the appropriate submission group name (see **Step 1.2** for creating a new submission group), and describe the submitting organization or laboratory name. This will be auto-populated from the contact info you included in your NCBI user account. Click "Continue" to proceed.

2.4 **GENERAL INFO tab:**

1. BioProject: Did you already register a data BioProject for this effort? If not please follow instructions in **Step 3** for creating a new data or umbrella BioProject. Return back to this sub-step with the **data BioProject** accession in hand.

Click "**Yes**" and paste in your data BioProject accession, e.g. PRJNA614995.

Sequence Read Archive (SRA) submission: SUB7513659

New

General Information

BioProject

ℹ 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

ℹ 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?

○ Yes ◉ No

Release date

ℹ **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

ℹ Please allow 24-48 hours for propagation of the data to the NCBI SRA public site.

Continue

2. BioSample: Click "NO" here. You will be registering BioSamples within this current submission

3. Release date: Choose "Release immediately following processing".

4. Click **Continue.**

2.5 **BIOSAMPLE TYPE tab:**

You are choosing the appropriate metadata package here (i.e. what kind of samples are you submitting?).

Select the **"Pathogen affecting public health",** then **"Combined pathogen submission"** for SARS-CoV-2 surveillance data**.**

★ Select the package that best describes your samples:

◉ **Pathogen affecting public health**
Use for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

○ **Clinical or host-associated pathogen**

○ **Environmental, food or other pathogen**

◉ **Combined pathogen submission**
Helper template for batch submissions that include both clinical and environmental samples. In the "attribute_package" field, specify either "Pathogen.cl" (for clinical or host-associated pathogen) or "Pathogen.env" (for environmental, food or other pathogen). Note that mandatory attributes are not completely represented in the template at this time, please refer to the clinical and environmental package definitions for requirements.

If your laboratory only sequences SARS-CoV-2 from human samples you can select the
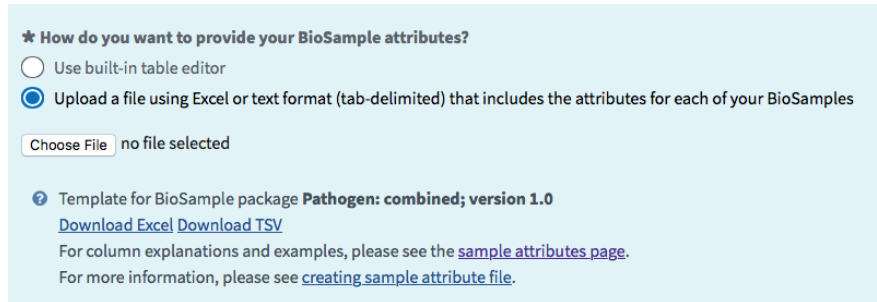
"Clinical or host-associated pathogen" package.

2.6 BIOSAMPLE ATTRIBUTES tab:

Choose "Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples".

Then click "Choose File" and browse to your populated metadata template.

If you have not populated your **PHA4GE BioSample metadata** template yet, download and follow the guidance in **Step 2.1**.

★ **How do you want to provide your BioSample attributes?**

◯ Use built-in table editor

◉ Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples

[Choose File] no file selected

❷ Template for BioSample package **Pathogen: combined; version 1.0**
Download Excel Download TSV
For column explanations and examples, please see the sample attributes page.
For more information, please see creating sample attribute file.

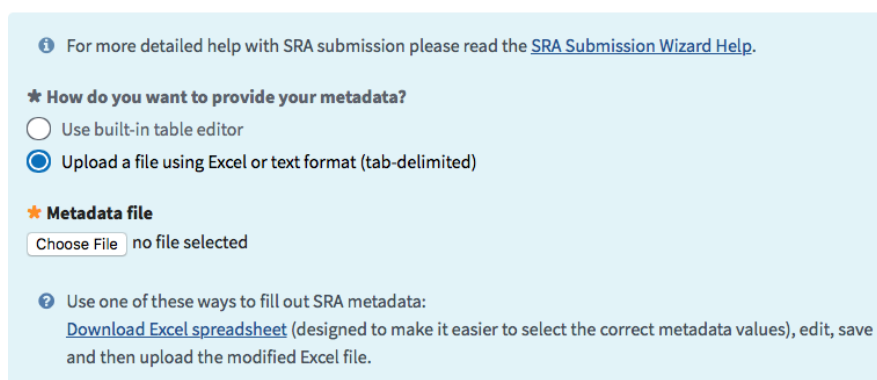**Skip antibiogram sections (not relevant for SARS-CoV-2)

Click "**Continue**".

NCBI will do a validation check on your metadata. Resolve any red "errors" reported back by editing the spreadsheet and replacing the uploaded file. Review any yellow "Warnings" and proceed if everything looks ok.

Click **"Continue".**

2.7 SRA metadata tab:

Choose: "Upload a file using Excel or text format (tab-delimited)"

ℹ 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)

★ **Metadata file**

[Choose File] no file selected

❷ Use one of these ways to fill out SRA metadata:
Download Excel spreadsheet (designed to make it easier to select the correct metadata values), edit, save and then upload the modified Excel file.

Upload the SRA metadata template populated in Step 2.1.

Click **"Continue".**

NCBI will do a validation check on your metadata. Resolve any red "errors" reported back by editing the spreadsheet and replacing the uploaded file. Review any yellow "Warnings" and proceed if everything

looks ok.

Click **"Continue"**.

2.8 Files tab:

Each laboratory will establish it's own path for transferring files.

In general, selecting the web browser option should work for uploading ~48 sequences at a time. For a more stable internet connection, your laboratory can use FTP or Aspera. Directions for doing so pop up after clicking the FTP radio button

> ℹ • **Each file must be listed** in the <u>SRA metadata table you uploaded</u>. If you are uploading a **tar** archive, list each file name, not the archive name.
> • **Unique file names** that **do not contain any sensitive information** should be used for all files. File names as submitted appear publicly when data is retrieved from the cloud.
> • Files can be compressed using **gzip** or **bzip2**, and may be submitted in a **tar** archive, but archiving or compressing your files is not required. **Do not use zip!**
>
> ⭐ **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.
> ○ Amazon S3 bucket

2.9 REVIEW & SUBMIT tab:

Check over your entire submission, then click submit.

If corrections are needed, you can go back and select individual tabs to edit your submission.

> 🗋 If you are having trouble finalizing your submission, contact the relevant NCBI database for assistance and include your submission ID in the email subject (SUB######):
>
> **BioSample** (for source metadata issues): biosamplehelp@ncbi.nlm.nih.gov
> **SRA** (for raw sequence or sequence metadata issues): sra@ncbi.nlm.nih.gov

2.10 **BioSample accessions** will be automatically created upon submission and will be available on the "my submissions" page of the Submission portal by clicking on "## objects" within the submission record. You can also download by clicking the "Download attributes file with BioSample accessions". Accessions will start with SAMNxxxxxxxx. You will also receive an email within 12 hours, but typically much faster, containing these same accessions.

2.11 SRA Accessions:

SRA run accessions will be available on the "my submissions" page of the Submission portal by clicking on "## objects" within the submission record. You can also download by clicking the "Download metadata file with SRA accession". Accessions will start with SRRxxxxxxx." You will also receive an email with these same accessions within 24 hours, but typically much faster, containing these same accessions.

2.12 Important data stewardship and curation notes:

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

- For updates, corrections, or retractions to your BioSample and SRA records, follow the guidance

provided in the NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase. Some edits can be made within the submission portal and others need to be done via email.

> ⚠️ *Caution*: It is possible for a single BioSample to have more than one SRR IDs. Two scenarios include:
> 1. Two runs were submitted for the same isolate/BioSample, which is not generally recommended for surveillance. Follow **Step 3** in the NCBI curation protocol to retract one of them).
> 2. if the initial submission was retracted and new a new run was submitted. It's important to keep track of both IDs, even if one was retracted.

BioProject Creation

3   Create a new BioProject

BioProjects are an organizing tool at NCBI that pulls together different kinds of data submitted across multiple NCBI databases. Each BioProject has a unique URL, providing a home page with a title, description, links to lab websites, publications, funding resources associated with a particular project, along with links to the deposited data. A basic **data BioProject** holds actual sequence data, assemblies, and their associated metadata. An **umbrella BioProject** is a way to group two or more data BioProjects together, which is useful for coordinating disease surveillance and for looking across the grouped BioProjects in a single view.

This protocol describes the steps for creating a new *data* BioProject linked to an existing *umbrella* BioProject (usually established by a coordinating group, e.g. SPHERES or CanCOGen).

*If you need to create a new Umbrella BioProject, modifications are summarized in **Step 3.12.**

3.1
Navigate to the "My Submissions" page, https://submit.ncbi.nlm.nih.gov/subs/, and click "BioProject" in the "Start a new submission" box.

3.2   Click the "New submission" box:

3.3 Submitter tab:

Populate with submitter info. An NCBI "submitter" is the name of the person or submission group who is managing the submissions, not a supervisor or PI.

Select the appropriate submission group name (see **Step 1.2** for creating a new submission group), and describe the submitting organization or laboratory name. This will be auto-populated from the contact info you included in your NCBI user account.

3.4 Project type tab:

Project data type: Genome sequencing and assembly.

Sample scope:

For a **Data** *BioProject*: select multi-isolate. This will allow you to submit multiple isolates or strains of the same species.

3.5 Target tab:

For a **Data** *BioProject*: Populate ONLY the Organism name here: "Severe acute respiratory syndrome coronavirus 2".

Leave the strain info and Description fields blank.

3.6 General info tab:

Click "Release immediately following processing".

Include a brief title describing the effort.
- **Data** *BioProject Title*: e.g., "SARS-CoV-2 WGS: Utah Public Health Laboratory".

Public Description: e.g., "Whole-genome sequencing of XXX pathogens as part of XXXX surveillance effort."

Relevance: medical.

Is your project part of a larger initiative that is already registered at NCBI?
- Data BioProjects. Click "Yes" and include a brief description and umbrella BioProject accession

number (e.g. for SPHERES use PRJNA615625, or refer to **Step 1.5** for full list). This will properly link your data project to the umbrella.

External links: Include a link to your laboratory's website here.

3.7 *BioSample tab:*

Leave blank!! You will create biosamples separately.

3.8 *Publications tab:*

If relevant, include publications from your laboratory.

3.9 *Review and Submit tab:*

Check if everything looks correct and edit if necessary, then click "**submit**."

3.10 **The BioProject accession** will be available within a few minutes on the "my submissions" page of the Submission portal, "PRJNAxxxxxx." You will also receive an email containing the new accession.

3.11 If you are part of a coordinated surveillance effort, like SPHERES, please alert the coordinating body that a new BioProject was created under an existing umbrella.

For SPHERES, contact oamd@cdc.gov.

3.12 **Creating a new Umbrella BioProject:**

Proceed as outlined in the above steps with the following modifications:

PROJECT TYPE tab:

For an Umbrella BioProject: select multi-species. This will allow you to link multiple data BioProjects representing different species under a single umbrella.

TARGET tab:

For an Umbrella BioProject: Leave the Organism name field blank. Include a list or description of species you intend to include in this effort. E.g. "bacterial foodborne pathogens", or "SARS-Cov-2"

GENERAL INFO tab:

Umbrella BioProject Title: e.g. "Microbial pathogen surveillance at NY State Dept. of Health, Wadsworth Center."

Is your project part of a larger initiative that is already registered at NCBI?

- *For an Umbrella BioProject*: click "NO"

The last step is to email bioprojecthelp@ncbi.nlm.nih:

Example email:

> "Dear BioProject and PD help teams,
>
> Please convert the PRJNA##### to an Umbrella BioProject. Our laboratory will be

submitting data under the XXX effort (SARS-CoV-2, GenomeTrakr, Vet-LIRN, NARMS, HAI, or more general pathogen surveillance).

I'd be happy to provide any additional details you might need.

Thank you, "

After the conversion is complete you can use the new Umbrella accession to properly link any new data BioProjects being created.

3.13 Important data stewardship and curation notes:

- Develop an internal method for storing and tracking your BioProject accessions! They are required for every BioSample and sequence data submission to ensure proper linkage.

- Bookmark URLs for each of your BioProjects to monitor the public-facing view of your submissions. e.g. Virgina DCLS's SARS-CoV-2 BioProject: https://www.ncbi.nlm.nih.gov/bioproject/625551

- For updates to your BioProjects, follow the guidance provided in the NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase. Some edits can be made within the submission portal and others need to be done via email.