



Jan 13, 2022

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

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


Ruth Timme¹, Emma Griffiths², Duncan MacCannell³, Lee Katz⁴, Michael Weigand³,
Technical Outreach and Assistance for States Team³

¹US Food and Drug Administration; ²University of British Columbia; ³Centers for Disease Control and Prevention; ⁴CDC

1



protocol .

 Technical Outreach and Assistance for States Team
Centers for Disease Control and Prevention

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

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 set 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. [Populate your templates first.](#)

2. [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. Users can modify this protocol to just create a BioSample with no linked raw data.

3. [SARS-CoV-2 NCBI consensus submission protocol: GenBank](#)

Required: established BioProject and BioSamples

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

Version history:

V4: Direct links provided to download metadata templates (instead of hosting duplicate files). Other minor edits throughout the protocol.

Ruth Timme, Emma Griffiths, Duncan MacCannell, Lee Katz, Michael Weigand, Technical Outreach and Assistance for States Team 2022. SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject. **protocols.io**
<https://protocols.io/view/sars-cov-2-ncbi-submission-protocol-sra-biosample-b3tcqniw>

protocol

Griffiths, E. J. et al. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. (2020) doi:10.20944/preprints202008.0220.v1.
<https://www.preprints.org/manuscript/202008.0220/v1>

Added step-by-step instructions to tag submissions at SRA for additional human reads scrubbing

SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject, Technical Outreach and Assistance for States Team

NCBI submission, pathogen surveillance, genomic epidemiology, SARS-CoV-2, covid-19, SRA, BioSample, BioProject

_____ protocol ,

Jan 13, 2022

Jan 13, 2022

56900

:

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

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

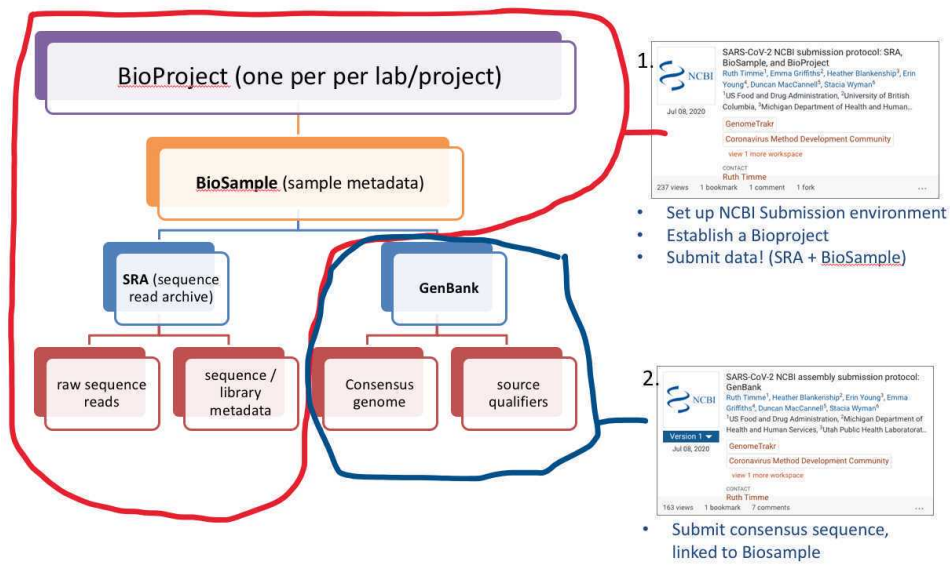
1 Set up a new NCBI submission environment for your lab

- 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:

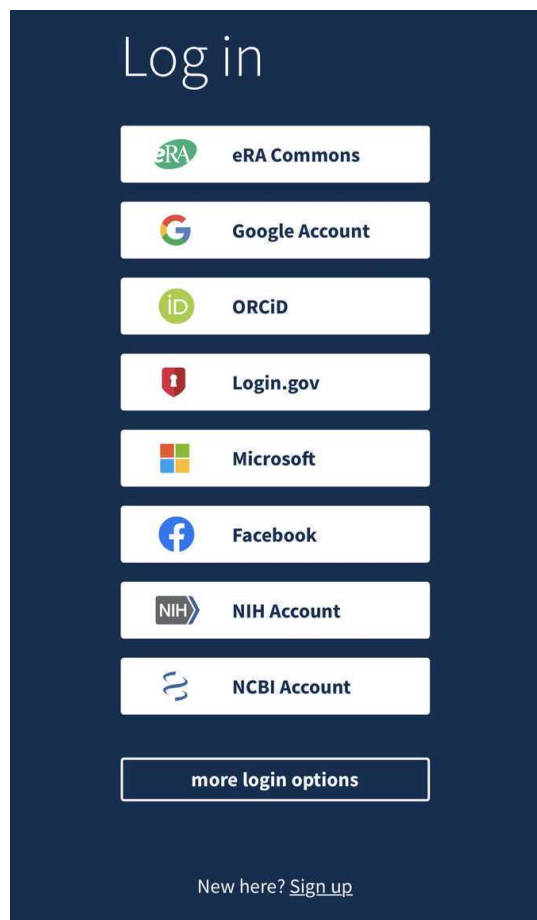
After these steps are complete you can proceed with **BioSample + SRA** data submission in **Step 2**.

NCBI database structure



The NCBI database structure and how this protocol (#1) fits into the submission process.

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



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.

Submission Portal						
			Home	My submissions	Manage data	Groups
Groups						
						Search
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
cdphe_gt	GenomeTrakr Project: Colorado Department of Public Health and Environment		Public Health and Environment	Colorado Department of Public Health and Environment	shannon.matzinger@state.co.us	2
unr_gt	GenomeTrakr Project: University of Nevada - Reno		Physiology and Cell Biology	University of Nevada - Reno	AndrewG@unr.edu	3

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.

Submission Portal

Home My submissions Manage data **Groups** Templates My profile

Groups: fda_ny

Profile **Members** Invites

[Invite members](#)

Username	First name	Last name	Email	Receives email notifications	Read	Modify	Submit	Delete	Admin	Remove member
klimke	Bill	Klimke	[REDACTED]	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="button" value="Remove"/>
retimme	Ruth	Timme	[REDACTED]	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="button" value="Remove"/>
justin.payne	Justin	Payne	[REDACTED]	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="Remove"/>
wwolfgang	William	Wolfgang	[REDACTED]	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="button" value="Remove"/>
samwirth	PulseNet	PulseNet	[REDACTED]	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="Remove"/>
sig04	Sai	Gubbala	[REDACTED]	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="button" value="Remove"/>

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**.

Submission Portal

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: YYYY-MM-DD To date: YYYY-MM-DD Status: Not deleted Sort by: desc

Apps + Data archives +

Query ? Search Clear

106,699 submissions

Submission	Title	App	Group	Status	Updated
SUB480036	Salmonella enterica Genome sequencing	BioProject	fda	✓ BioProject: Processed PRJNA242847 : GenomeTrakr Project: USDA - Food Safety and Inspection Service (TaxID: 28901) Locus Tag Prefixes: • A0J62 (SAMN04532063) • A0J63 (SAMN04532066) locustagprefix.txt	06:42
SUB633156	Campylobacter jejuni GenomeTrakr Project: FDA-CFSAN	BioProject	fda	✓ BioProject: Processed PRJNA258022 : GenomeTrakr Project: FDA-CFSAN (TaxID: 194) Locus Tag Prefixes: • A0838 (SAMN03580886) • A0839 (SAMN03580887) locustagprefix.txt	06:30
SUB1581179	Vet-LIRN-E.coli-LA	BioProject	vet-lirn-la	✓ BioProject: Processed PRJNA324573 : Vet-LIRN-E.coli-LA (TaxID: 562) Locus Tag Prefixes: • C5453 (SAMN08596247) • C5454 (SAMN08596249) locustagprefix.txt	06:30
SUB9059508	SARS-CoV-2	GenBank	fda	⚙ Unfinished at the References step	Feb 10

1.5 Identify or establish a new BioProject

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

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/>

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 the Sequence Read Archive (SRA) option

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

Current excel template versions can be accessed in the [Overview of NCBI's SARS-CoV-2 submission process and the metadata required](#) (Step 2 and Step 3).

Ensure that the BioSample and SRA templates are populated before proceeding to the next step.

2.2 Click the "New submission" box.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) [New submission](#)

Short description and brief instructions +

Options to preload data:

Aspera browser plugin upload +

Aspera command line and FTP upload options +

Amazon S3 instructions +

Google Cloud Storage instructions +

Filter / Search

From date: YYYY-MM-DD To date: YYYY-MM-DD Status: Not deleted Sort by: ☐ desc

Data archives +

Query

153 submissions

Submission	Title	Group	Status	Updated
SUB9050748	GenomeTrakr Project: Listeria monocytogenes New York State Department of Health, Wadsworth Center, Feb 09 '21	nys_gt	✓ BioSample: Processed Successfully loaded SAMN17841130: FSL N4-0696 (TaxID: 1639) Download attributes file with BioSample accessions ✓ SRA: Processed SRR13665182 Download metadata file with SRA accessions View and manage my SRA submission data	Feb 09
SUB9046837	GenomeTrakr Project: New Mexico State University, Food Safety Laboratory, Feb 08 '21	nmsu-fsl	✓ BioSample: Processed Successfully loaded (3 objects) Download attributes file with BioSample accessions ✓ SRA: Processed (3 objects) Download metadata file with SRA accessions View and manage my SRA submission data	Feb 08

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.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) submission: SUB9064572 [Delete submission](#)

New

1 **SUBMITTER** 2 GENERAL INFO 3 SRA METADATA 4 FILES 5 REVIEW & SUBMIT

Submitter Required fields are marked with *

* First (given) name: Middle name: * Last (family) name:

* Email (primary): Email (secondary): At least one email should be from the organization's domain.

Group for this submission
☐ No group (affiliation from my personal profile)
☒ **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 Balkey, fda service, Maria Hoffmann, Jayanthi Gangireddi, Julie Haendiges, Yu Wang, you

Click "Continue" to proceed.

2.4 GENERAL INFO tab:

1. BioProject: Did you already register a BioProject for this effort? If not please follow instructions in **Step 3** for creating a new BioProject and return back to this step with accession in hand.

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

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

submission

3. Release date:

**BioProjects established for wastewater data can be flagged for automated human-read scrubbing (performed prior to public release. tool here: <https://github.com/ncbi/sra-human-scrubber>). This flag needs to be set along side the first data submission for that BioProject. Once the flag is set, subsequent data submissions will get automatically scrubbed.

First submission? Choose "Release on specified date", then enter a date 1-week in the future. This will give you some time to establish the flag prior to data release.

Send the following email to **sra@ncbi.nlm.nih.gov** asap:

Hi sra,

Please add the human read scrubbing analysis flag to my BioProject <paste in your bioproject accession here>, then release my HUPed SRA submissions.

thanks,

Otherwise, choose "Release immediately following processing" (for all subsequent submissions).

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) submission: SUB9064572 Delete submission

New

1 SUBMITTER 2 **GENERAL INFO** 3 SRA METADATA 4 FILES 5 REVIEW & SUBMIT

General Information

Required fields are marked with * asterisk

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**

PRJNA614995

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?**

☐ Yes ☒ No

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

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

Continue

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 **"SARS-CoV-2: clinical or host-associated"**

Sequence Read Archive (SRA) submission: SUB9550561

SARS-CoV-2 WGS, Apr 28 '21

[Delete submission](#)

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

Sample Type

Required fields are marked with * asterisk

* Select the package that best describes your samples.

All packages Packages for MAG submitters Packages for metagenome submitters

(Optional) Filter packages by organism name

Enter the full scientific name of your samples, e.g., *Escherichia coli*[Reset and show all packages](#)

- To filter for relevant BioSample packages, enter the **full scientific name** of the organism of your samples.
 - If your BioSamples are derived from a species **not represented in NCBI's Taxonomy database**, enter the genus-level name, e.g., *Escherichia*
 - If your BioSamples are derived from **more than one organism**, enter the common species, genus, or family, e.g., *Enterobacteriaceae*
 - If your BioSamples are **metagenomic/environmental**, or **metagenome-assembled genomes (MAG)**, select the appropriate tab above
 - For more information about organism names, see [Organism information](#).

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.
- ☐ **Pathogen**
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.
- ☐ **Microbe**
Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use [MAGS](#), Pathogen or Virus packages.
- ☐ **Model organism or animal**
Use for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, *Drosophila*, worm, fish, frog, or large mammals including zoo and farm animals.
- ☐ **Metagenome or environmental**
Use for metagenomic and environmental samples when it is not appropriate or advantageous to use [MAGS](#) packages.
- ☐ **Invertebrate**
Use for any invertebrate sample.
- ☐ **Human**
WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is

GSC [MAGS](#) packages for genomes, metagenomes, and marker sequences [More...](#)

- ☐ **MIGS Cultured Bacterial/Archaeal**
Use for cultured bacterial or archaeal genomic sequences. Organism must have lineage [Bacteria](#) or [Archaea](#).
- ☐ **MIGS Eukaryotic**
Use for eukaryotic genomic sequences. Organism must have lineage [Eukaryota](#).
- ☐ **MIGS Viral**
Use for virus genomic sequences. Organism must have lineage [Viruses](#).
- ☐ **MIMAG Metagenome-assembled Genome**
Use for metagenome-assembled genome sequences produced using computational binning tools that group sequences into individual organism genome assemblies starting from metagenomic data sets. Organism cannot contain the term 'metagenome'. Use the MIMAG package for virus genomes.
- ☐ **MIMARKS Specimen**
Use for any type of marker gene sequences, eg, 16S, 18S, 23S, 28S rRNA or COI obtained from cultured or voucher-identifiable specimens. Organism cannot contain the term 'metagenome'.
- ☐ **MIMARKS Survey related**
Use for any type of marker gene sequences, eg, 16S, 18S, 23S, 28S rRNA or COI obtained directly from the environment, without culturing or identification of the organisms. Organism must be a metagenome, where lineage starts with [unclassified sequences](#)

****Metadata packages are in development for environmental and wastewater samples.****
This step will be updated as soon as these packages are available.

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".

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) submission: SUB9064572 [Delete submission](#)

SARS-CoV-2 WGS, Feb 11 '21

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

Attributes

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

★ 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](#) or drag and drop it here

Template for BioSample package Pathogens: clinical or host-associated; 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](#).

Antibiogram — pathogen MIC (optional)

[Choose file](#) or drag and drop it here

Download Antibiogram Excel template
 Once the spreadsheet is completed, save as a "Text (Tab-delimited)" file, and upload it.
 For more information, please see [providing pathogen MIC antibiogram data](#).

Antibiogram — mycobacterial non-MIC (optional)

[Choose file](#) or drag and drop it here

Download Antibiogram Excel template
 Once the spreadsheet is completed, save as a "Text (Tab-delimited)" file, and upload it.
 For more information, please see [providing mycobacterial non-MIC antibiogram data](#).

[Continue](#)

Then click "Choose File" and browse to your populated PHA4GE BioSample_template Excel file.

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

**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)"

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) submission: SUB9064572 [Delete submission](#)

SARS-CoV-2 WGS, Feb 11 '21

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

SRA metadata

Required fields are marked with * asterisk

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](#) or drag and drop it here

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.

[Continue](#)

Upload the SRA metadata template populated in Step 2.1 (Excel file works here).

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 its 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

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Sequence Read Archive (SRA) submission: SUB9064572 [Delete submission](#)

SARS-CoV-2 WGS, Feb 11 '21

1 SUBMITTER 2 GENERAL INFO 3 BIOSAMPLE TYPE 4 BIOSAMPLE ATTRIBUTES 5 SRA METADATA 6 FILES 7 REVIEW & SUBMIT

Files

Required fields are marked with * asterisk

- Each file must be listed in the SRA metadata table you uploaded. 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.

☐ AWS or GCP bucket

☐ Autofinish submission

[Continue](#)

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 (usually within 2 hours). You can also download by clicking the “Download attributes file with BioSample accessions”. Accessions will start with SAMNxxxxxxx. You will also receive an email containing these same accessions.

BioSample [New submission](#) [Download batch submission template](#)

Note: to update an existing record or recent submission, please [email your request](#).

Short description and brief instructions +

Filter / Search

From date: YYYY-MM-DD To date: YYYY-MM-DD Status: Not deleted Sort by: desc

Data archives +

Query Search Clear

667 submissions

Submission	Title	Group	Status	Updated
SUB7721095	Pathogen: combined sample	nmsu-fsl	BioSample: Processed Successfully loaded (13 objects) <ul style="list-style-type: none">SAMN15456891 : NMSU-WRLP-359 (TaxID: 1639)SAMN15456892 : NMSU-WRLP-373 (TaxID: 1639)SAMN15456893 : NMSU-WRLP-413 (TaxID: 1639)SAMN15456894 : NMSU-WRLP-506 (TaxID: 1639)SAMN15456895 : NMSU-WRLP-510 (TaxID: 1639)SAMN15456896 : NMSU-WRLP-357 (TaxID: 1639)SAMN15456897 : NMSU-EP16 (TaxID: 1639)SAMN15456898 : NMSU-EP17 (TaxID: 1639)SAMN15456899 : NMSU-EP18 (TaxID: 1639)SAMN15456900 : NMSU-EP21 (TaxID: 1639)SAMN15456901 : NMSU-EP28 (TaxID: 1639)SAMN15456902 : NMSU-EP29 (TaxID: 1639)SAMN15456903 : NMSU-EP32 (TaxID: 1639) Download attributes file with BioSample accessions	Jul 06

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 (usually within 2 hours). You can also download by clicking the “Download metadata file with SRA accession”. Accessions will start with SRRxxxxxxx.” You will also receive an email containing these same accessions.

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

Submission Portal

Home My submissions Manage data Groups Templates My profile

Sequence Read Archive (SRA) [New submission](#)

Short description and brief instructions +

Options to preload data:

Aspera browser plugin upload +

Aspera command line and FTP upload options +

Amazon S3 instructions +

Filter / Search

From date To date Status Not deleted Sort by desc

Data archives +


Query Search Clear

30 submissions

Submission	Title	Group	Status	Updated
SUB5616822	GenomeTrakr Project: US Food and Drug Administration, May 14 '19	fda	✓ SRA: Processed (2 objects) • SRR9052981 • SRR9052982 Download metadata file with SRA accessions View and manage my SRA submission data	May 14
SUB5110190	GenomeTrakr Project: Bacillus cereus, New York State Department of Health, Wadsworth Center, Jan 31 '19	fda_ny	✓ SRA: Processed (6 objects) Download metadata file with SRA accessions View and manage my SRA submission data	Jan 31
SUB5106005	GenomeTrakr Project: NY State Dept. of Health, Wadsworth Center, Jan 30 '19	fda_ny	✓ SRA: Processed (20 objects) Download metadata file with SRA accessions View and manage my SRA submission data	Jan 30

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).

Umbrella BioProjects: If you think need to establish a new umbrella BioProject (for an entirely new project or laboratory network), send an email to bioprojecthelp@ncbi.nlm.nih.gov and they will help create one for you.

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.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

Your submissions

Start a new submission

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

Filter / Search

From date: YYYY-MM-DD To date: YYYY-MM-DD Status: Not deleted Sort by: ☐ desc

Apps + Data archives +

Query Search Clear

107,593 submissions

Submission	Title	App	Group	Status	Updated
SUB5032842	GenomeTrakr Project: New York State Department of Health, Wadsworth Center	BioProject	nys_gt	✓ BioProject: Processed PRJNA514286 : GenomeTrakr Project: Listeria monocytogenes New York State Department of Health, Wadsworth Center (TaxID: 1639) Locus Tag Prefixes: • E2141 (SAMN11148736) • E2B07 (SAMN11102767) locustagprefix.txt	15:50

3.2 Click the “New submission” button:

Submission Portal Home **My submissions** Manage data Groups Templates My profile

BioProject **New submission**

ATTN: to update an existing record or recent submission, please email your request with your BioProject ID or Submission ID included. Do not create new submission to update an existing submission!

Short description and brief instructions +

Filter / Search

From date: YYYY-MM-DD To date: YYYY-MM-DD Status: Not deleted Sort by: ☐ desc

Data archives +

Query Search Clear

235 submissions

Submission	Title	Group	Status	Updated
SUB8337245	South Carolina Department of Health Listeria monocytogenes Genome sequencing and assembly	scdhc_prgt	✓ BioProject: Processed PRJNA670180 : South Carolina Department of Health Listeria monocytogenes Genome sequencing and assembly (TaxID: 1639) Locus Tag Prefixes: • JTP34 (SAMN17866121)	16:20
SUB480036	Salmonella enterica Genome sequencing	fda	✓ BioProject: Processed PRJNA242847 : GenomeTrakr Project: USDA - Food Safety and Inspection Service (TaxID: 28901) Locus Tag Prefixes: • A0J62 (SAMN04532063) • A0J63 (SAMN04532066) locustagprefix.txt	16:10

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.

The screenshot shows the 'Submitter' tab in the NCBI Submission Portal for a new BioProject submission (SUB9064836). The interface includes a navigation bar with links: Home, My submissions (active), Manage data, Groups, Templates, and My profile. Below the submission ID, there is a 'Delete submission' button. A progress bar shows seven steps: 1 SUBMITTER (active), 2 PROJECT TYPE, 3 TARGET, 4 GENERAL INFO, 5 BIOSAMPLE, 6 PUBLICATIONS, and 7 REVIEW & SUBMIT. The 'Submitter' section contains fields for: First (given) name (Ruth), Middle name, Last (family) name (Timme), Email (primary) (ruth.timme@fda.hhs.gov), and Email (secondary) (retimme@gmail.com). A note states: 'At least one email should be from the organization's domain.' Below these fields is the 'Group for this submission' section, with two options: 'No group (affiliation from my personal profile)' and '13 members, FDA Center for Food Safety and Applied Nutrition (edit group)'. The selected group lists members: Anjanette Johnston, Yan Luo, Errol Strain, Justin Payne, Nariol Gonzalez-Escalona, Hugh Rand, Maria Balkey, fda service, Maria Hoffmann, Jayanthi Gangiredla, Julie Haendiges, Yu Wang, you.

3.4 PROJECT TYPE tab:

*Project data type:

Choose: "**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.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

BioProject submission: SUB9064836 [Delete submission](#)

New

1 SUBMITTER 2 **PROJECT TYPE** 3 TARGET 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 REVIEW & SUBMIT

Project Type ⓘ Required fields are marked with * asterisk

★ **Project data type** ⓘ

☒ Genome sequencing and assembly

☐ Raw sequence reads

☐ Genome sequencing

☐ Assembly

☐ Clone ends

☐ Epigenomics

☐ Exome

☐ Map

☐ Metagenome

☐ Metagenomic assembly

☐ Phenotype or Genotype

☐ Proteome

☐ Random survey

☐ Targeted loci cultured

☐ Targeted loci environmental

☐ Targeted Locus (Loci)

☐ Transcriptome or Gene expression

☐ Variation

☐ Other

★ **Sample scope** ⓘ

Monoisolate

ⓘ **Sample scope choices**

Monoisolate: a single animal, cultured cell-line, inbred population (or possibly a heterogeneous population when a single genome assembly is generated from the pooled sample; not preferred).

Multisolate: multiple individuals, a population (representative of a species). To be used for variation or other sequence comparison projects, not when multiple genomes will be annotated. Make separate monoisolate projects when more than one genome will be annotated.

Multi-species: sample represents multiple species.

Environment: the species content of the sample is not known.

Synthetic: the sample is synthetically created by a machine.

Other: specify the sample scope that was used.

☒ Autogenerate locus tag prefix ⓘ

3.5 TARGET tab:

For a *Data BioProject*.

Populate **ONLY** the Organism name here: "Severe acute respiratory syndrome coronavirus 2", or higher level taxonomic classification for broader surveillance.

Leave the strain info and Description fields blank.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

BioProject submission: SUB9064836 [Delete submission](#)

New

1 SUBMITTER 2 PROJECT TYPE 3 **TARGET** 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 REVIEW & SUBMIT

Target ⓘ Required fields are marked with * asterisk

★ **Organism name** ⓘ

Severe acute respiratory syndrome coronavirus 2

Strain ⓘ Breed ⓘ Cultivar ⓘ Isolate name ⓘ Label ⓘ

Description ⓘ

[Continue](#)

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.

The screenshot shows the 'Submission Portal' for a BioProject submission, specifically for SUB9064836. The 'General Info' tab is active, showing fields for 'Release date', 'Project title', 'Public description', 'Relevance', 'Initiative description', 'BioProject accession', and 'External links'. The 'Release date' section has a note and two radio buttons: 'Release immediately following processing' (selected) and 'Release on specified date or upon publication, whichever is first'. The 'Project title' field contains 'SARS-CoV-2 genome surveillance: Utah Public Health Laboratory'. The 'Public description' field contains 'Genome sequencing of XXX pathogens as part of XXXX surveillance effort'. The 'Relevance' dropdown is set to 'Medical'. The 'Initiative description' field contains 'SPHERES' and the 'BioProject accession' field contains 'PRJNA615625'. The 'External links' section has a table with columns for 'Description', 'URL', and 'Delete', with one row containing 'Utah Public Health Laboratory' and a 'URL link' field.

Submission Portal Home **My submissions** Manage data Groups Templates My profile

BioProject submission: SUB9064836 [Delete submission](#)

New

1 SUBMITTER 2 PROJECT TYPE 3 TARGET 4 **GENERAL INFO** 5 BIOSAMPLE 6 PUBLICATIONS 7 REVIEW & SUBMIT

General Info Required fields are marked with * asterisk

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

★ Project title

SARS-CoV-2 genome surveillance: Utah Public Health Laboratory

★ Public description

Genome sequencing of XXX pathogens as part of XXXX surveillance effort

Relevance

Medical

★ Is your project part of a larger initiative which is already registered with NCBI?

☐ No ☒ Yes (not very common)

★ Initiative description

SPHERES

★ BioProject accession

PRJNA615625

If you are registering a project that is part of an initiative which is already registered in the BioProject database, then please tell us the existing BioProject accession and provide a general description of the larger initiative. For example, the ENCODE project and Human Microbiome project have several subprojects. This information is needed for project linking.

External links

Description	URL	Delete
Utah Public Health Laboratory	URL link	<input type="button" value="Delete"/>

[Add another link](#)

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.”

The screenshot shows the 'Review & Submit' tab for a BioProject submission. The submission ID is SUB9064836, titled 'SARS-CoV-2 genome surveillance: Utah Public Health Laboratory'. The progress bar indicates steps 1 through 7, with '7 REVIEW & SUBMIT' being the current step. The form contains the following information:

- Submitter:** Ruth Timme (ruth.timme@fda.hhs.gov)
- Submitting organization:** US Food and Drug Administration (http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/default.htm)
- Project type:** Multiisolate
- Target:** Organism name (taxid): Severe acute respiratory syndrome coronavirus 2
- General information:**
 - Project details:**
 - Parent project ID: PRJNA615625
 - Parent project description: SPHERES
 - Project type: genome sequencing and assembly
 - Title: SARS-CoV-2 genome surveillance: Utah Public Health Laboratory
 - Description: Genome sequencing of XXX pathogens as part of XXXX surveillance effort
 - Relevance: Medical
 - External links: Utah Public Health Laboratory (https://uphl.utah.gov/)

A 'Submit' button is located at the bottom left. A callout box on the right says: 'Your submission 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.'

3.10 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.

The screenshot shows the 'Your submissions' page in the Submission Portal. The page header includes the NIH logo and navigation links: Home, My submissions (active), Manage data, Groups, Templates, and My profile. A 'Start a new submission' button is visible. Below it, a list of submission types is shown: GenBank, Sequence Read Archive, Genome, TSA, BioProject, BioSample, Supplementary Files, and API. A filter/search box is on the right. The main table lists submissions, with the following row highlighted:

Submission	Title	App	Group	Status	Updated
SUB9064836	GenomeTrakr Project: Minnesota Department of Health	BioProject	fda	BioProject: Processed	16/22

The status 'BioProject: Processed' is highlighted with a red arrow. The table also shows a 'Query ID' and a 'Search' button.

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 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. Virginia 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>. Most edits can be made within the Submission Portal by clicking the **"Manage Data"** button. Others need to be done via email.

SUB480036	Salmonella enterica Genome sequencing	BioProject	fda	<div><div>✓ BioProject: Processed</div><div>PRJNA242847 : GenomeTrakr Project: USDA - Food Safety and Inspection Service (TaxID: 28901)</div><div>Locus Tag Prefixes:</div><div><ul style="list-style-type: none">• A0J62 (SAMN04532063)• A0J63 (SAMN04532066)</div><div>locustagprefix.txt</div></div> <div>Manage data</div> <td>16:10</td>	16:10
-----------	---------------------------------------	------------	-----	---	-------