Jun 01, 2022 Version 7

NCBI submission protocol for SARS-CoV-2 wastewater data: SRA, BioSample, and BioProject V.7

Version 1 is forked from SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject

DOI
https://dx.doi.org/10.17504/protocols.io.ewov14w27vr2/v7

Ruth E Timme[1], Maria Balkey[1]

[1]US Food and Drug Administration

GenomeTrakr | Coronavirus Method De...

Ruth Timme
US Food and Drug Administration

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.ewov14w27vr2/v7

External link: https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network

Protocol Citation: Ruth E Timme, Maria Balkey 2022. NCBI submission protocol for SARS-CoV-2 wastewater data: SRA, BioSample, and BioProject. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.ewov14w27vr2/v7**Version created by **Ruth Timme**

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
We use this protocol and it's working

Created: May 04, 2022

Last Modified: June 01, 2022

Protocol Integer ID: 61975

Keywords: NCBI submission, pathogen surveillance, genomic epidemiology, SARS-CoV-2, covid-19, SRA, BioSample, BioProject, wastewater, applicable for all wastewater sequence data submission, data submission to ncbi, wastewater sequence data submission, new ncbi submission environment, ncbi submission environment, data public at ncbi, submission protocol for sar, submitting biosample, submitting raw read, sample metadata to biosample, first data submission, wastewater project, raw reads to sra, wastewater collection, bioproject purpose, ncbi, data submission, genome sequencing, sample metadata, submission protocol, pandemic response project, applied nutrition for genometrakr, wastewater, rna extraction, including replicate info, sra metadata template, picklist for pcr concentration unit, laboratory, monitoring sar, submission portal, genometrakr, new attribute to the sra metadata template, minor edits to the biosample, pcr concentration unit, person in the laboratory, biosample

Disclaimer

This method is under development and assessment for suitability of use. It is likely that modifications will be made to improve the method.

Abstract

PURPOSE:

This method was developed at the FDA's Center for Food Safety and Applied Nutrition for GenomeTrakr's pandemic response project, monitoring SARS-CoV-2 variants in wastewater; however, this protocol was written to be broadly applicable for all wastewater sequence data submissions to NCBI. Protocols developed for this project cover wastewater collection, concentration, RNA extraction, RT-qPCR, library prep, genome sequencing, quality control checks, and data submission to NCBI.

This protocol covers the last step of making your data public at NCBI. Specifically, it provides the steps 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 this protocol 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.

Version updates:

V2: minor edits to the BioSample and SRA templates

V3: Adapted the protocol to be more broadly applicable to submitters outside of FDA's wastewater project. Updates were also made to both metadata templates, including a new attribute to the SRA metadata template, called "enrichment_kit".

V4: updates to BioSample and SRA templates: expanded picklists, addition of specimen processing attributes for including replicate info, and the removal of target_extract attribute for reporting level of target found in the sample.

V5: includes guidance for submitting BioSamples with no linked sequence data.

V6: Updated templates. BioSample: added picklist for PCR concentration units. SRA: added new quality control attributes.

V7: SRA and BioSample template updates

Before start

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:

- NCBI Data Curation protocol for making updates, corrections, or retractions to your data.

"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
1.3: Bookmark the link to your Submission Portal
1.4. Identify or establish new BioProjects (detailed in **Step 3**)

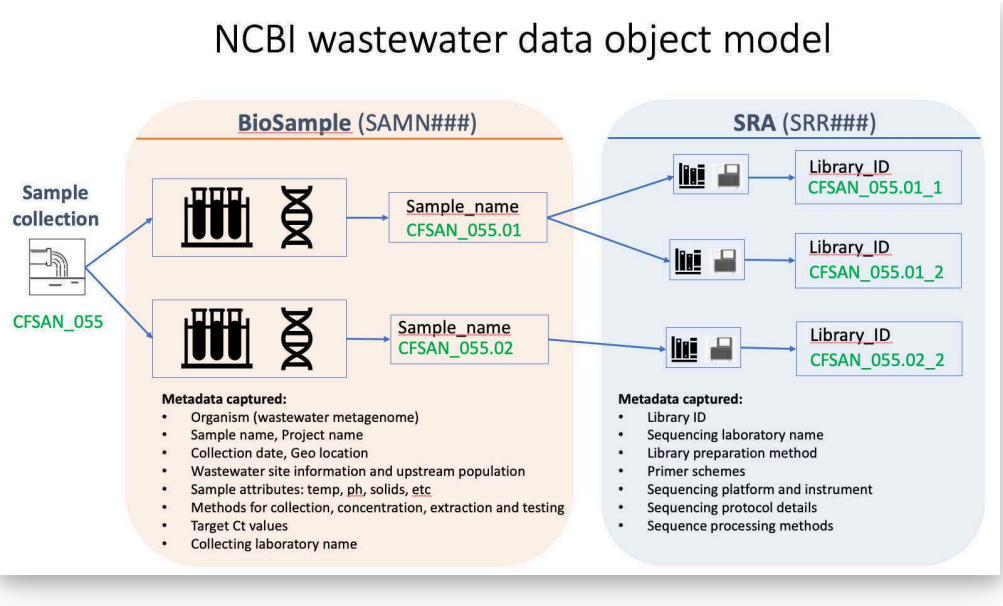
Ready for data submission:
After these steps are complete you can proceed with **BioSample + SRA** data submission in **Step 2.**

1.1 **NCBI data object model established for US government wastewater surveillance:**

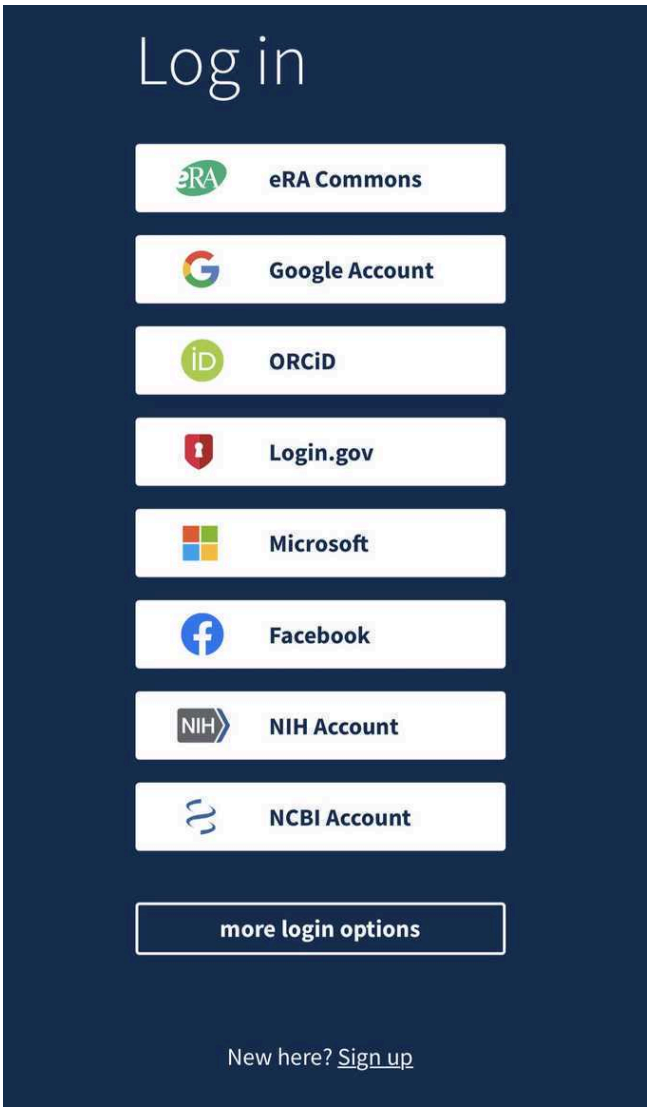
Adhering to this model is **extremely** important, both for standardizing submission protocols AND for helping to standardize data (and metadata location) for downstream analysis.

This structure includes the following NCBI databases:

- BioProject for grouping project-related submissions (e.g. one per laboratory)
- BioSample for storing sample metadata (created at the nucleotide extraction level)
- SRA for raw sequence reads and associated metadata (created at the sequence level)



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



1.3 **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.

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:

Note

"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 owned by this new user group, this is a good*
time to request that ownership change:
i.e., Please assign this new user group to the following BioProjects and linked data (list accessions).

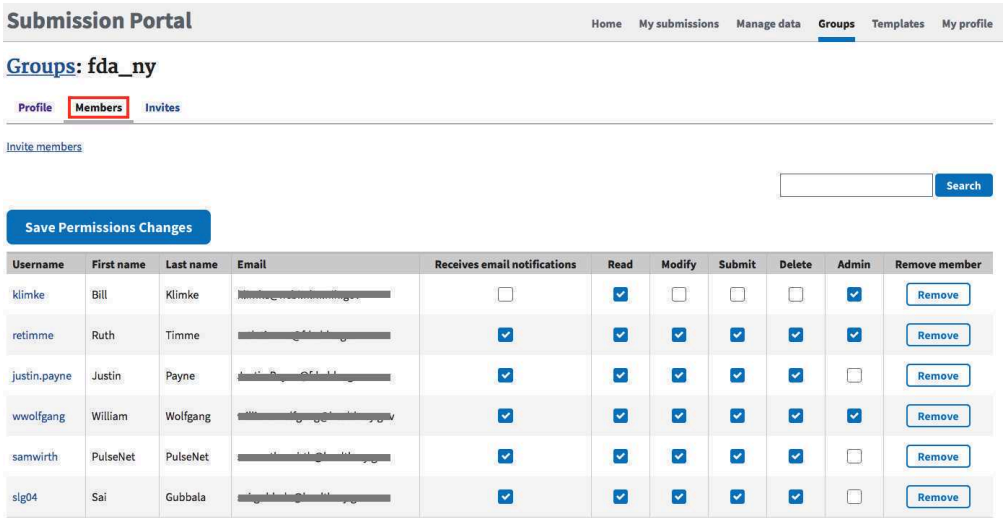
Thank you,"

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

1.4 **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.



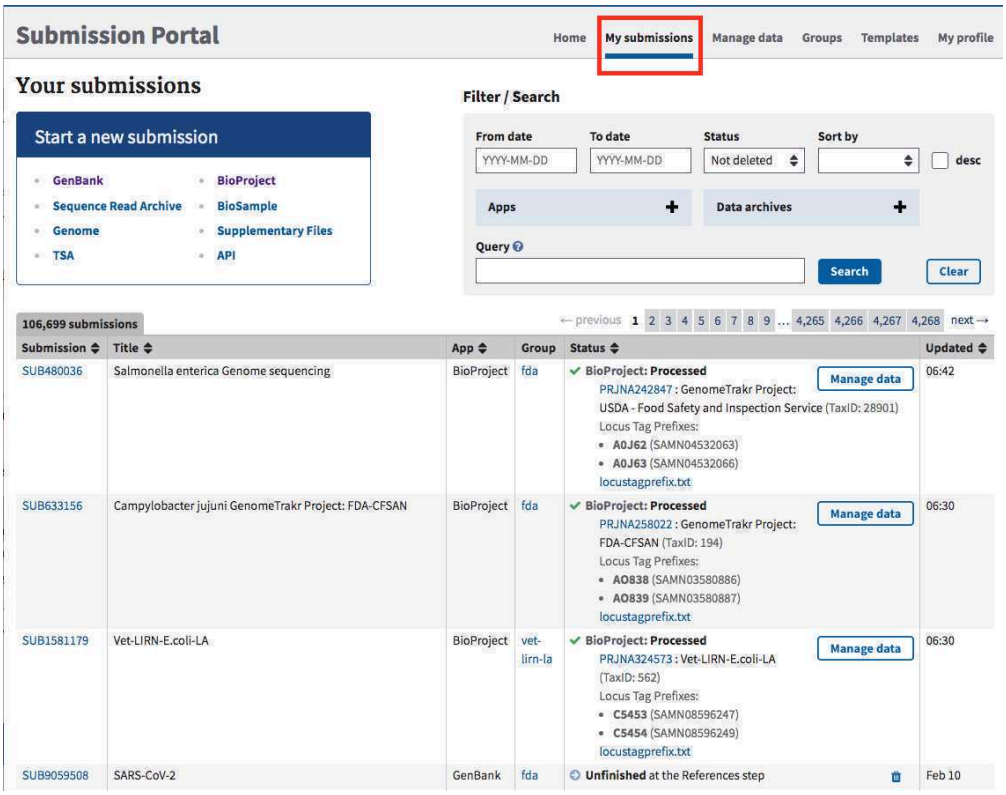
This user group should be kept up-to-date as members enter and leave the laboratory.

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.

1.5 **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**.



1.6 **Identify or establish a new BioProject**

Data BioProjects. Does your laboratory have an established data BioProject for this effort (follow the guidance of your institution or coordinating network (GenomeTrakr, NWSS, etc)? If not please follow instructions in **Step 3** for creating a new one.

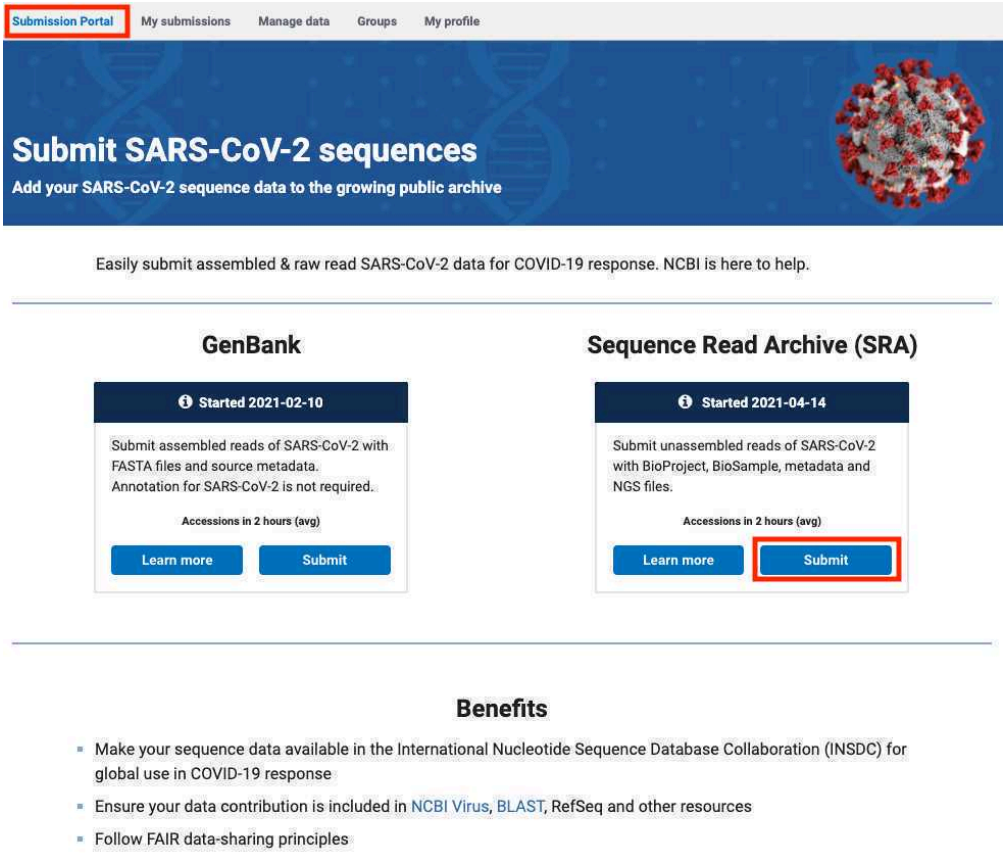
Data submission (BioSample and SRA)

2 **Data submission (source metadata and sequence data):**

This section provides guidance for submitting sequence data + metadata to SRA and BioSample.

****For BioSample-only submissions (samples with low or no detectable target that were not sequenced), follow link provided in the note below.****

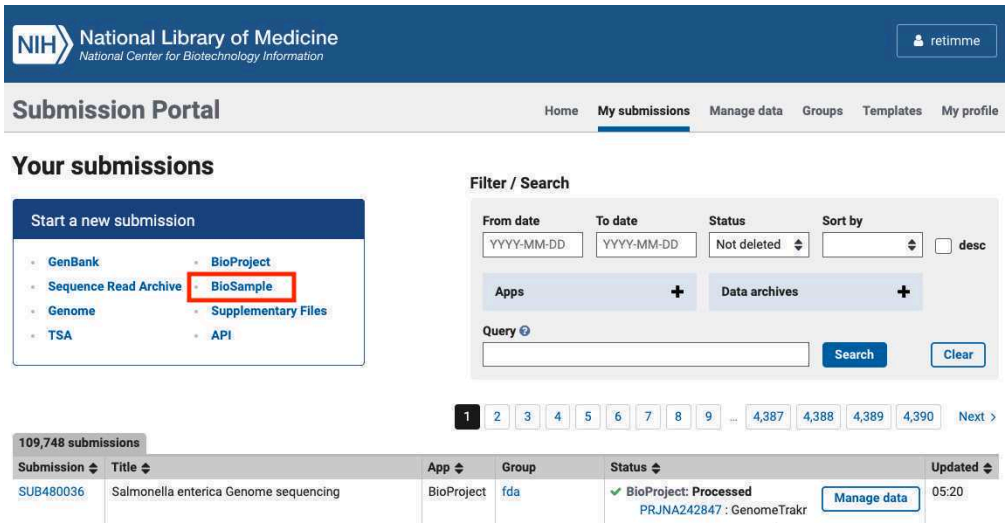
SARS-CoV-2 landing page: **https://submit.ncbi.nlm.nih.gov/sarscov2/**



Click "**Submit**" under the Sequence Read Archive (SRA) option

Note

For BioSample-only submissions: https://submit.ncbi.nlm.nih.gov/subs/



Click "**BioSample**" under the Sequence Read Archive (SRA), then "**New Submission**" and follow prompts for submitting only your BioSample template, available in Step 2.1.

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

1. BioSample custom wastewater template with NWSS/GenomeTrakr guidance and picklists (extension of **NCBI's Generic SARS-CoV-2: wastewater surveillance, v1.0**):

BioSample_ww_template_v1.8.xlsx

2. SRA: custom extension of NCBI's SRA metadata template:

SRA_ww_template_v5.5.xlsx

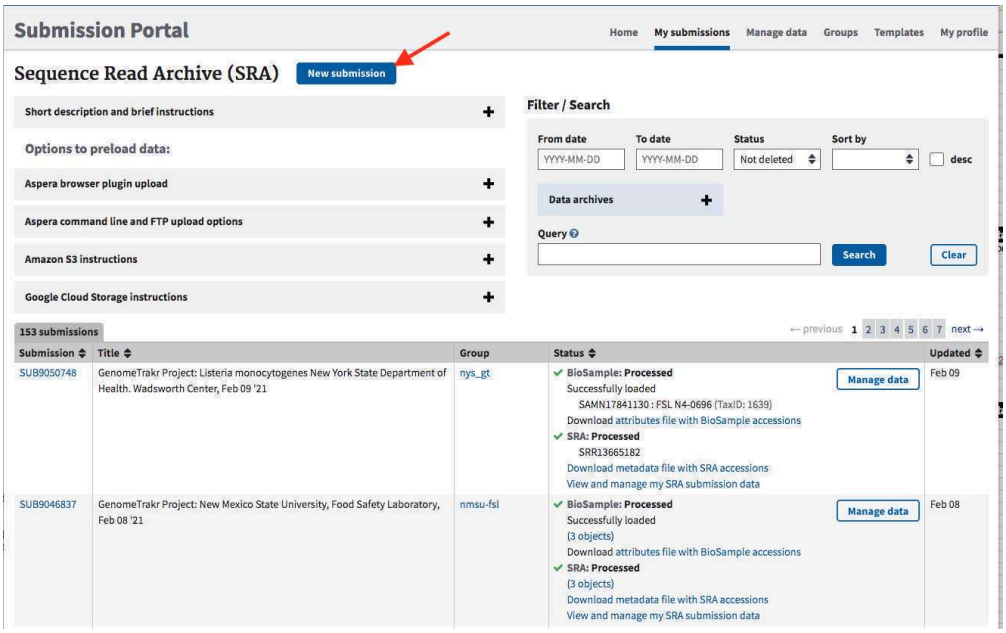
MOST COMMON SCENARIO: For each wastewater sample collected one BioSample and one associated SRA entry will be created. However, BioSamples for this project are actually created at the extraction level with metadata describing the collection → extraction methods. If you wanted to submit data across different collection → extraction methods, you would create separate BioSamples for these different extracts.

TIP: Create a base ID for each sample collection (for example, LABID_001), then add an index to represent each extraction (e.g.LABID_001.01). Every Sample Name from a single Submitter must be unique.

SRA:created at the sequence level, includes metadata for library-prep and sequencing methods. If you wanted to submit data across different sequencing methods from the same extract, you might submit multiple runs to SRA, all linked to the same BioSample.

You can submit a single sample at a time, or as a batch from an entire sequencing run or collection.

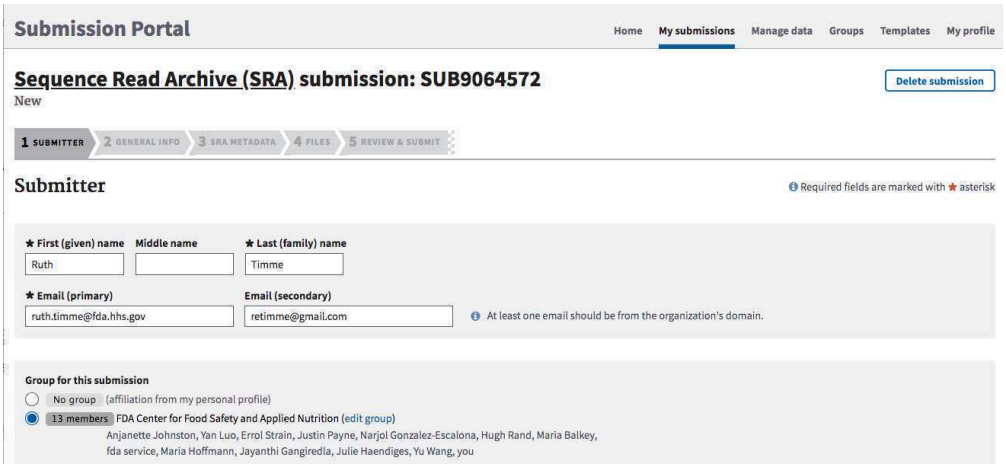
2.2 **Click the “New submission” box.**



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 BioProject for this effort (or has someone else created one for you)? 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 (preformed 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. One 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:

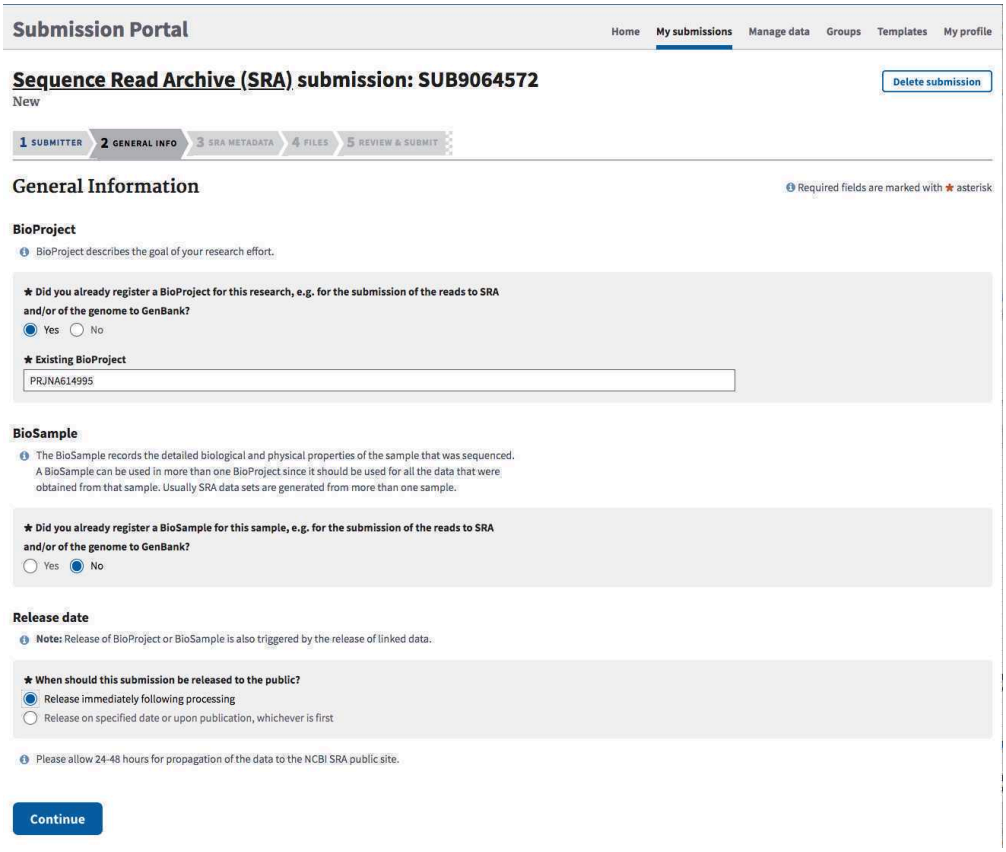
Note

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).

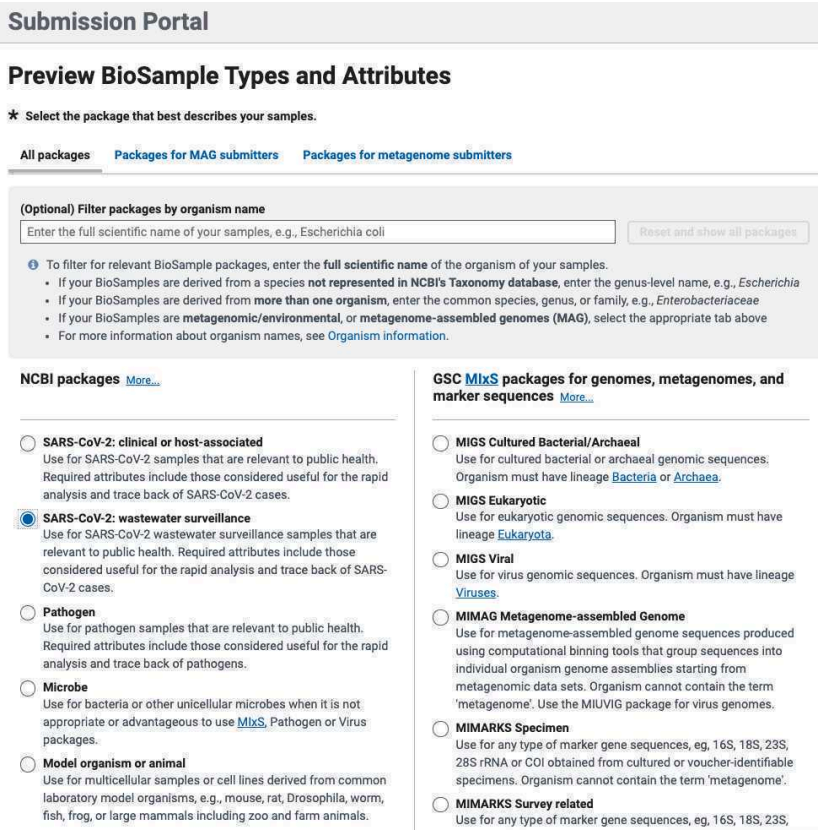


4. Click **Continue.**

2.5 BIOSAMPLE TYPE tab:

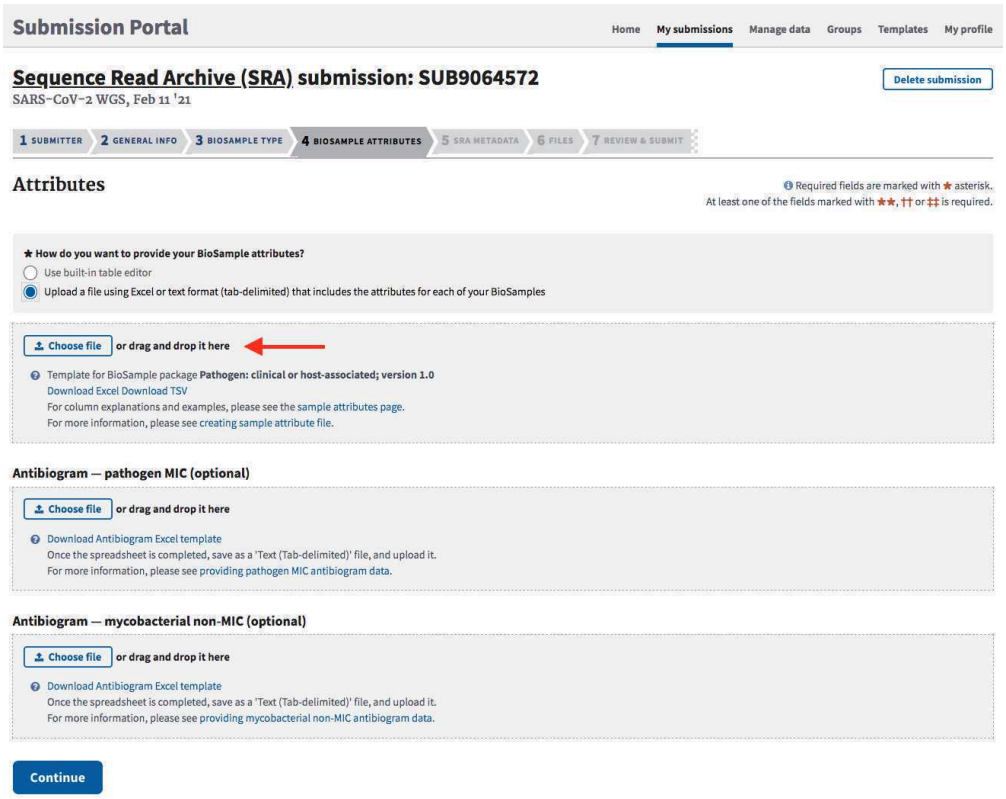
Choose the appropriate metadata package here (i.e. what kind of samples are you submitting?).

Select "**SARS-CoV-2: wastewater surveillance**"



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 PHA4GE BioSample_template Excel file.

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

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

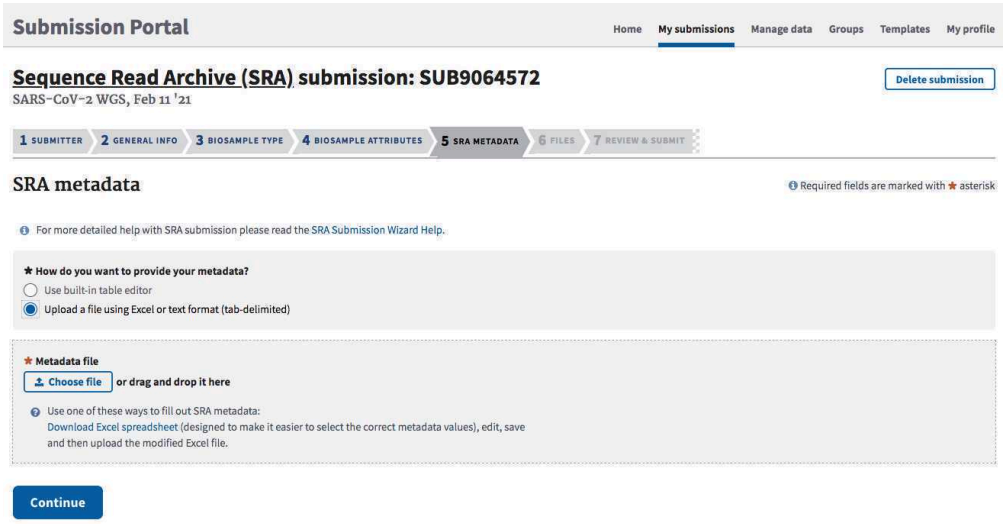
Click "**Continue**".

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)"



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

Click **"Continue".**

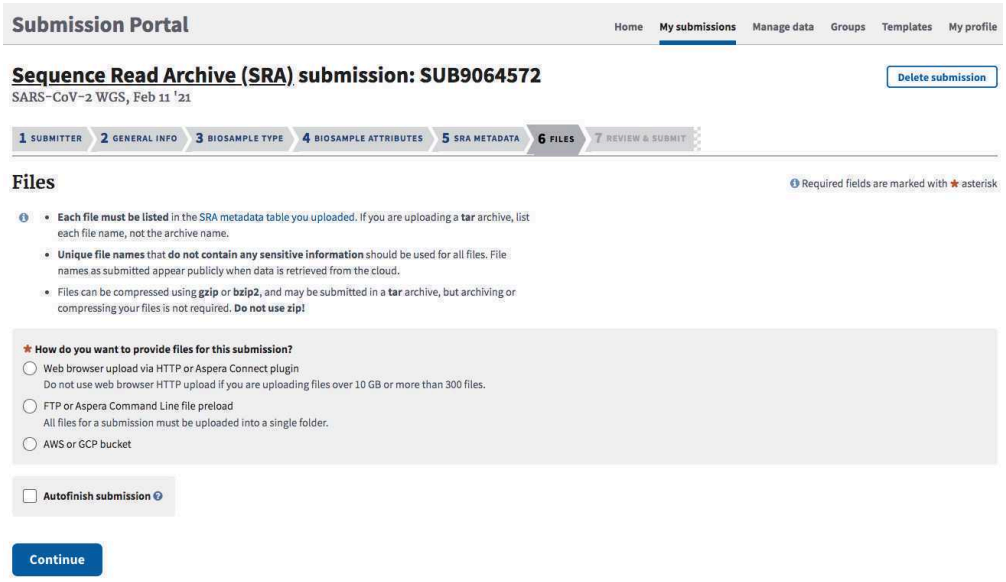
NCBI will do a validation check on your metadata. Resolve any red "errors" reported back by editing the spreadsheet and replacing the uploaded file. Review any yellow "Warnings" and proceed if everything looks ok.

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 a couple dozen samples 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



2.9 **REVIEW & SUBMIT tab:**

Check over your entire submission, then click submit.

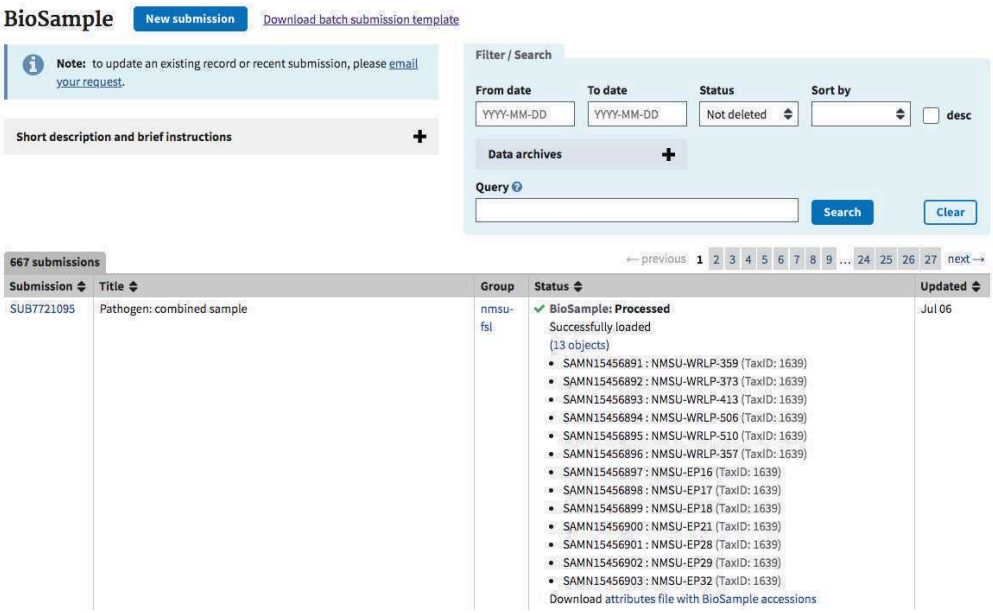
If corrections are needed, you can go back and select individual tabs to edit your submission.

Note

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#######):

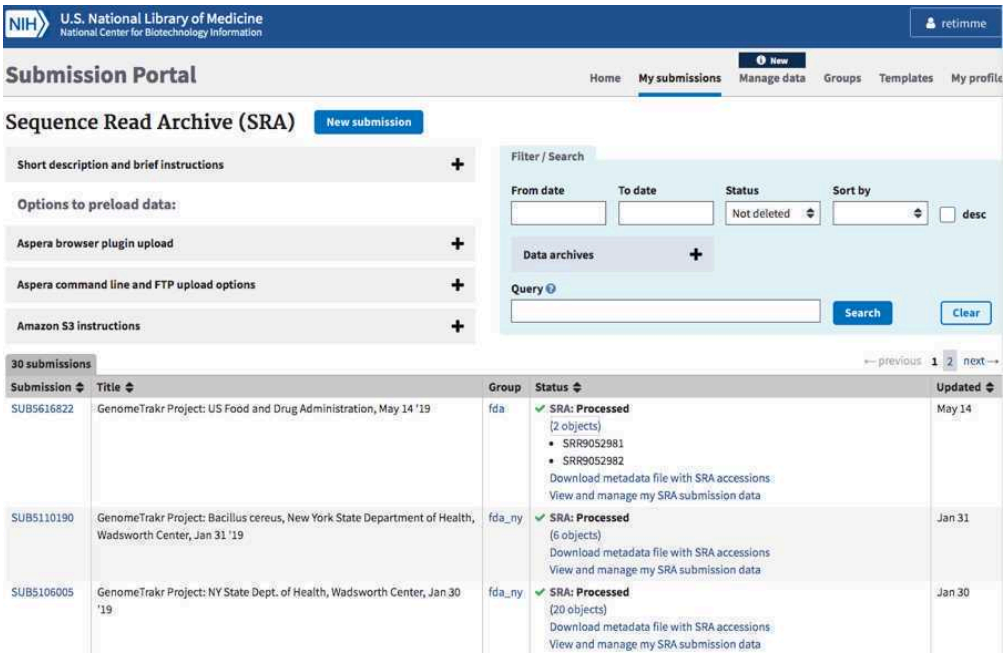
BioSample (for source metadata issues): biosamplehelp@ncbi.nlm.nih.gov
SRA (for raw sequence or sequence metadata issues): sra@ncbi.nlm.nih.gov

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 SAMNxxxxxxxx. You will also receive an email 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 (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.



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.

Safety information

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 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.

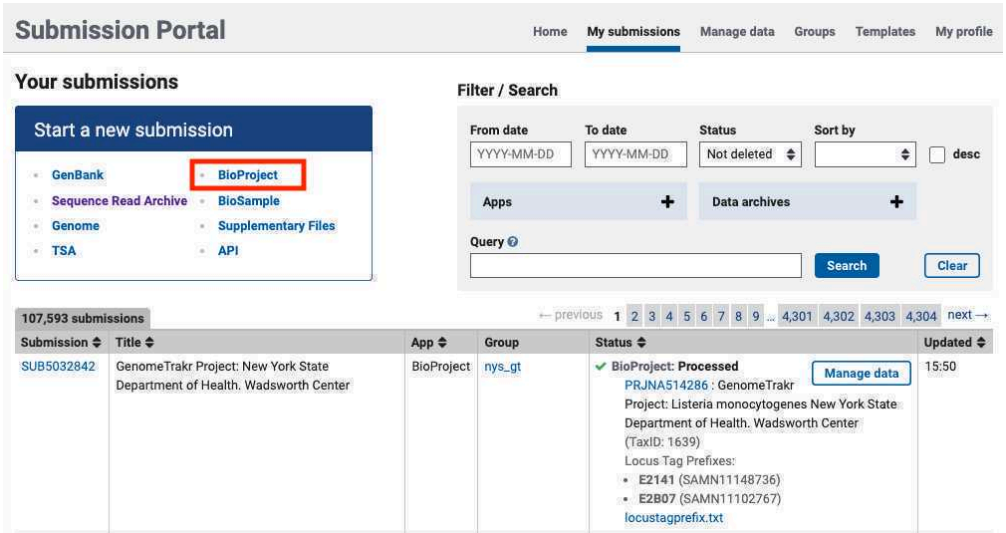
This protocol describes the steps for creating a new *data* BioProject linked to an existing *umbrella* BioProject (usually established by a coordinating group, e.g. CDC NWSS or FDA GenomeTrakr).

Note

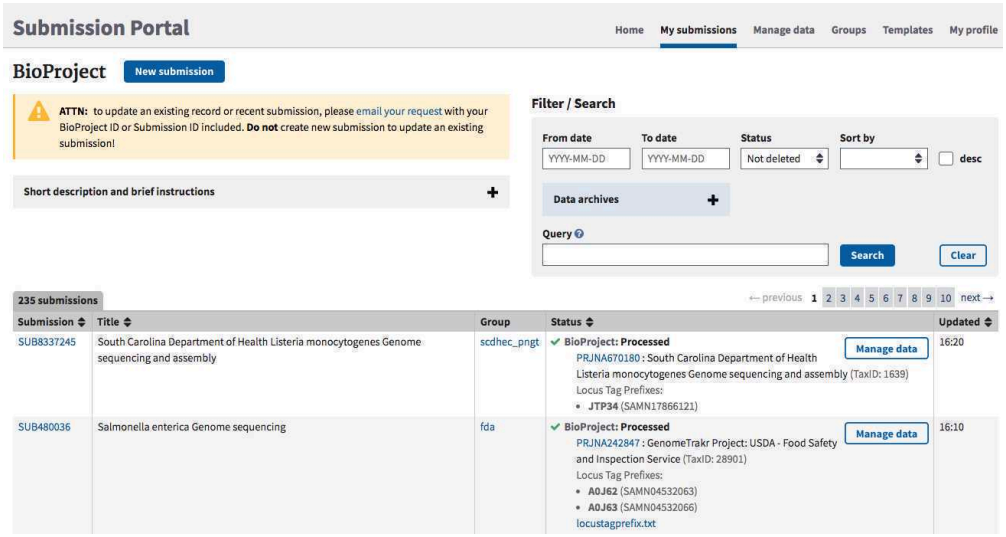
Umbrella BioProjects: If you think need to establish a new umbrella BioProject (for an entirely new project or laboratory network), create a new data BioProject, then send an email to bioprojecthelp@ncbi.hlm.nih.gov with the accession and they will help convert it for you.

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.



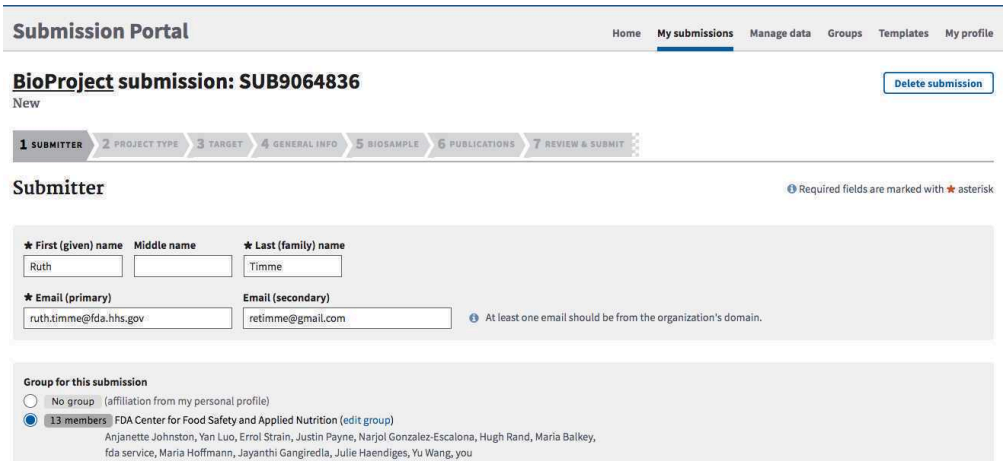
3.2 **Click the “New submission” button:**



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.

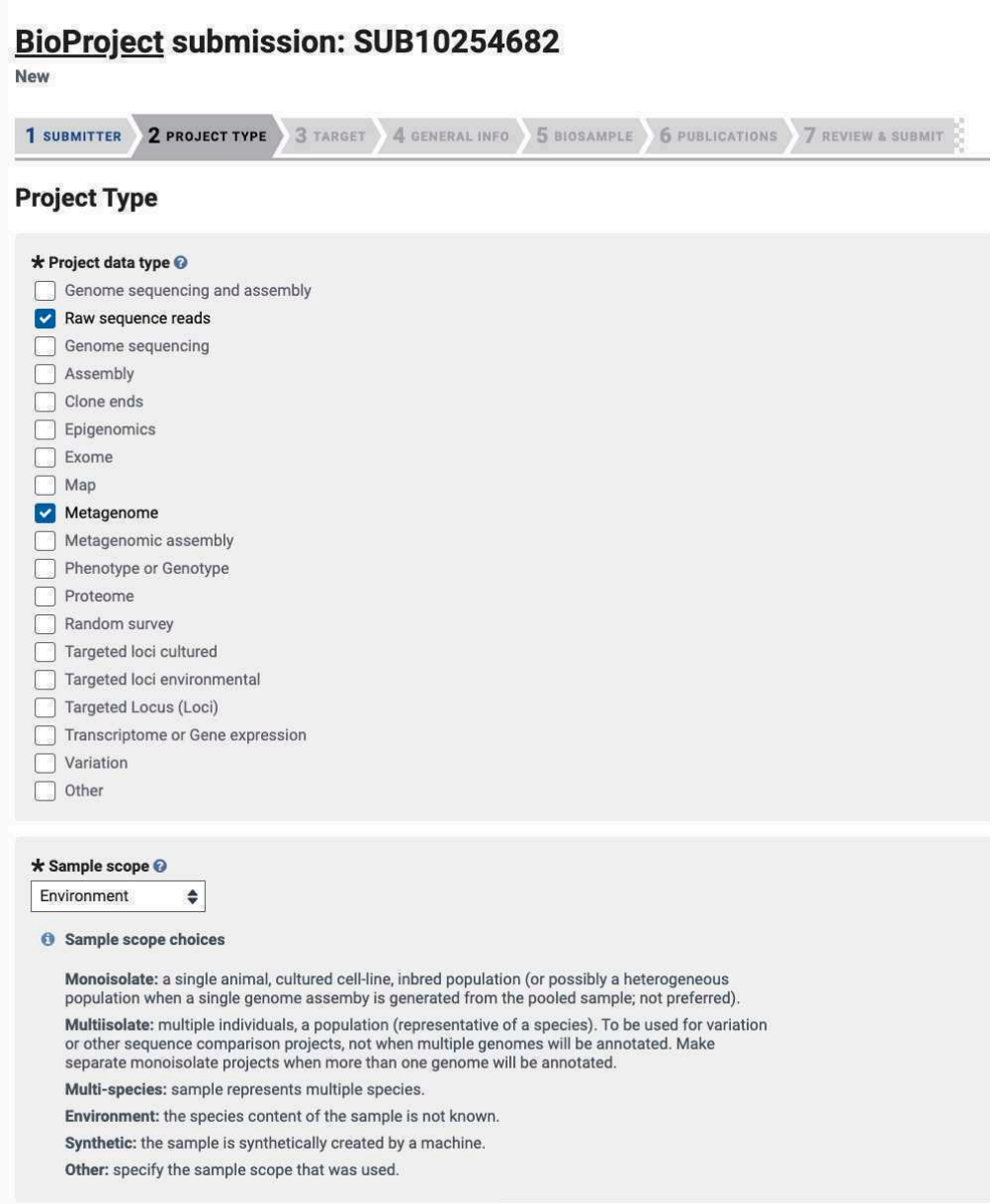
Select the appropriate submission **group name (see **Step 1.2** for creating a new submission group), and describe the submitting organization or laboratory name. This will be auto-populated from the contact info you included in your NCBI user account.



3.4 **PROJECT TYPE tab:**

∗Project data type:
Choose: "**Raw sequence reads**" and "**metagenome**"

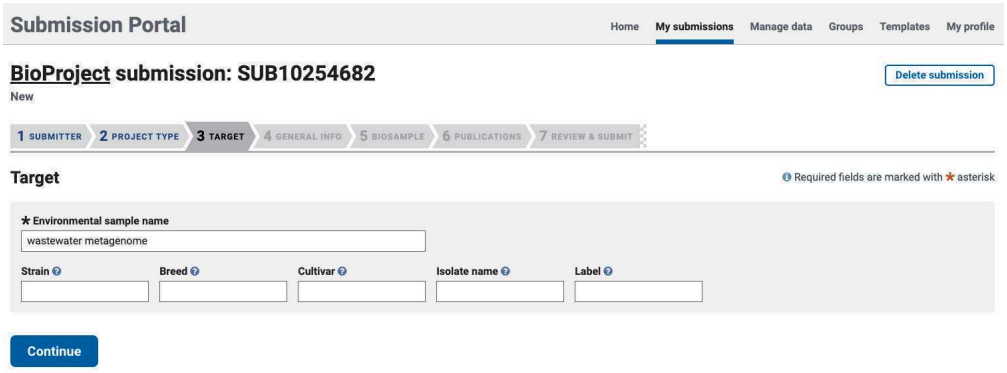
∗Sample scope:
Select **"Environment".**



3.5 **TARGET tab:**

Populate **ONLY** the Environmental sample name here: "wastewater metagenome" for the GenomeTrakr wastewater project.

Leave the strain info fields blank.



3.6 **GENERAL INFO tab:**

Click “Release immediately following processing”.

Project Title: e.g., “GenomeTrakr wastewater project: <YOUR LAB NAME>”.

Public Description: e.g., “Raw sequence data targeting SARS-CoV-2 in wastewater samples. These data were collected as part of the US FDA's pandemic response project for monitoring SARS-CoV-2 variants in wastewater.”

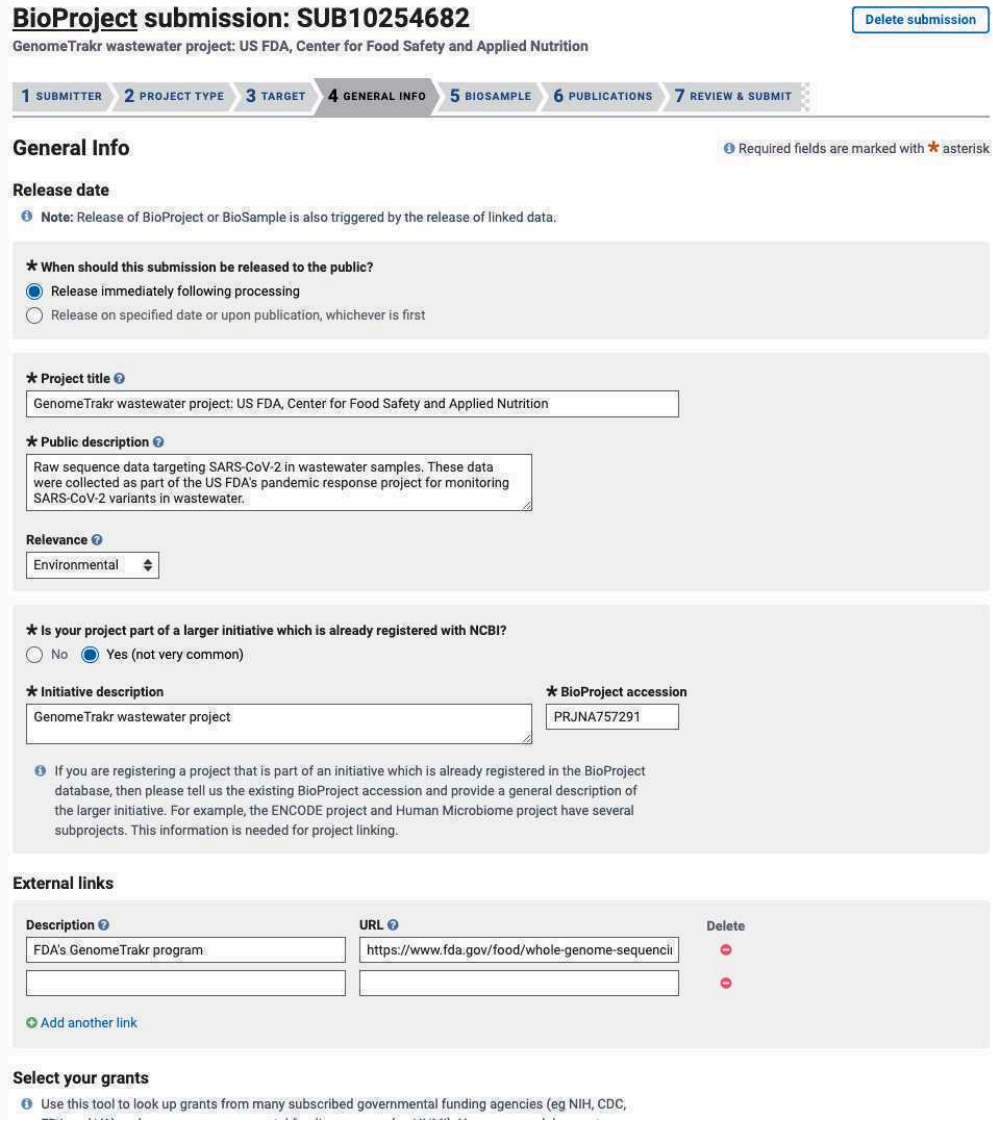
Relevance: Environmental.

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

- Click “Yes”
- *Initiative Description*: "GenomeTrakr wastewater project"
- *BioProject accession:* **PRJNA757291** (or other relevant coordinating umbrella Bioproject)

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

Grants: Add relevant grant information here (i.e. LFFM grant)



3.7 ***BioSample tab:***

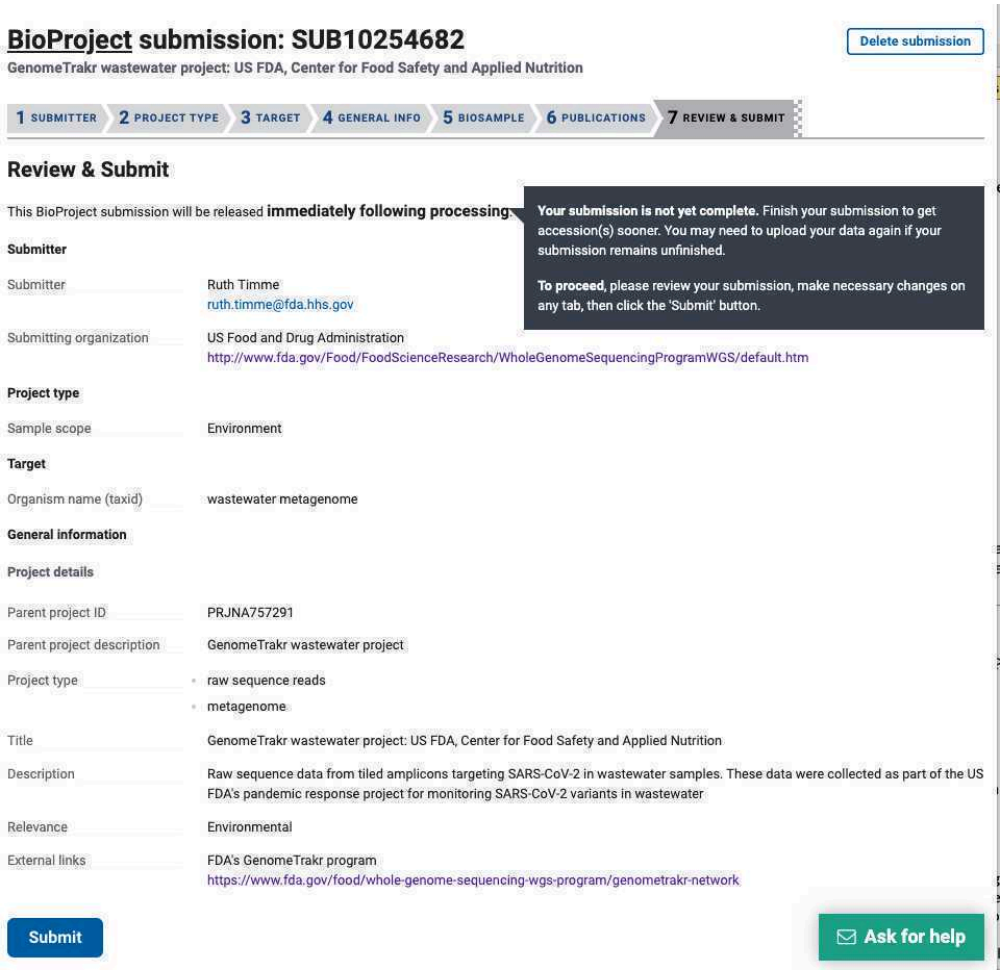
Leave blank!! You will create biosamples separately.

3.8 ***Publications tab:***

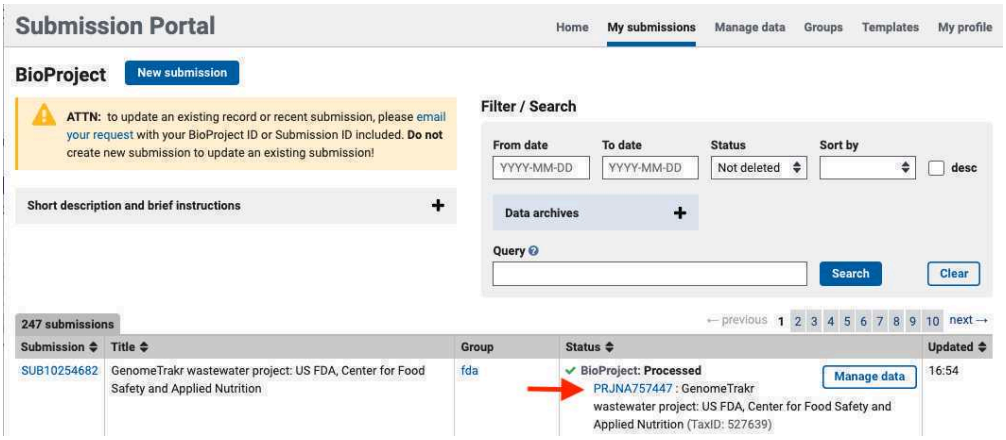
If relevant, include publications from your laboratory.

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 GenomeTrakr, please alert the coordinating body that a new BioProject was created under their existing umbrella.

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. Virgina DCLS's SARS-CoV-2 BioProject:
https://www.ncbi.nlm.nih.gov/bioproject/625551

- Need to make updates to your BioProject? Click the "**Manage Data**" button within the Submission Portal.

- For other updates, ollow the guidance provided in the NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase.