Nov 03, 2022 Version 10

NCBI submission protocol for SARS-CoV-2 wastewater data: SRA, BioSample, and BioProject V.10

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

DOI
dx.doi.org/10.17504/protocols.io.ewov14w27vr2/v10

Ruth Timme[1], Candace Hope Bias[1], Maria Balkey[1]

[1]US Food and Drug Administration

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

Ruth Timme
US Food and Drug Administration

DOI: https://dx.doi.org/10.17504/protocols.io.ewov14w27vr2/v10

External link: **https://www.fda.gov/food/whole-genome-sequencing-wgs-program/wastewater-surveillance-sars-cov-2-variants**

Protocol Citation: Ruth Timme, Candace Hope Bias, 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/v10**Version created by **Ruth Timme**

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

Created: November 03, 2022

Last Modified: November 03, 2022

Protocol Integer ID: 72260

Keywords: NCBI submission, pathogen surveillance, genomic epidemiology, SARS-CoV-2, covid-19, SRA, BioSample, BioProject, wastewater, applicable for any wastewater sequence data submission, wastewater sequence data submission, new ncbi submission environment, ncbi submission environment, data public at ncbi, sample metadata to biosample, submission protocol for sar, submitting biosample, submitting raw read, raw reads to sra, wastewater project, first data submission, bioproject purpose, sample metadata, ncbi, sra metadata template, protocol steps through the process, wastewater, protocol step, new attribute to the sra metadata template, including replicate info, submission protocol, biosample, picklist for pcr concentration unit, applied nutrition, minor edits to the biosample, biosample template, illumina covidseq assay, qiaseq direct sar, direct sar, submission portal, pcr concentration unit, fda, sra template

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 Laboratories; however, this protocol was written to be broadly applicable for any wastewater sequence 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 steps 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
V8: SRA template updates. "Illumina COVIDSeq Assay" added to library_preparation_kit, "QIAseq DIRECT SARS-CoV-2 - Boosted" added to amplicon_PCR_primer_scheme, "low coverage of characteristic mutations" added to quality_control_issues. minor edits to the protocol are also included in this update.
V10: Minor edits to generalize the protocol for broader usage. Picklist updates made to the BioSample and SRA templates.

Troubleshooting

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**)

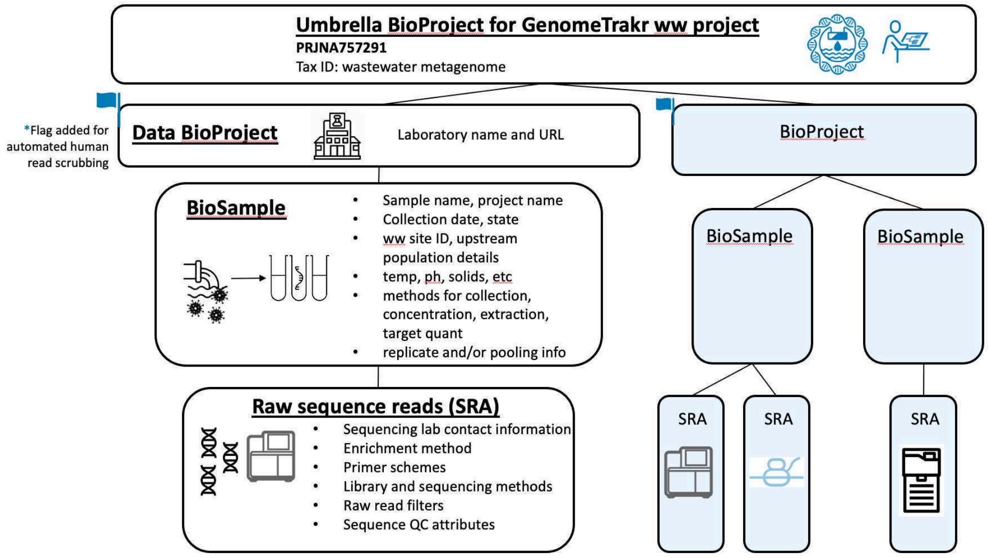
Ready for data submission:
After these steps are complete you can proceed with **BioSample + SRA** data submission in **Step 2.**

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

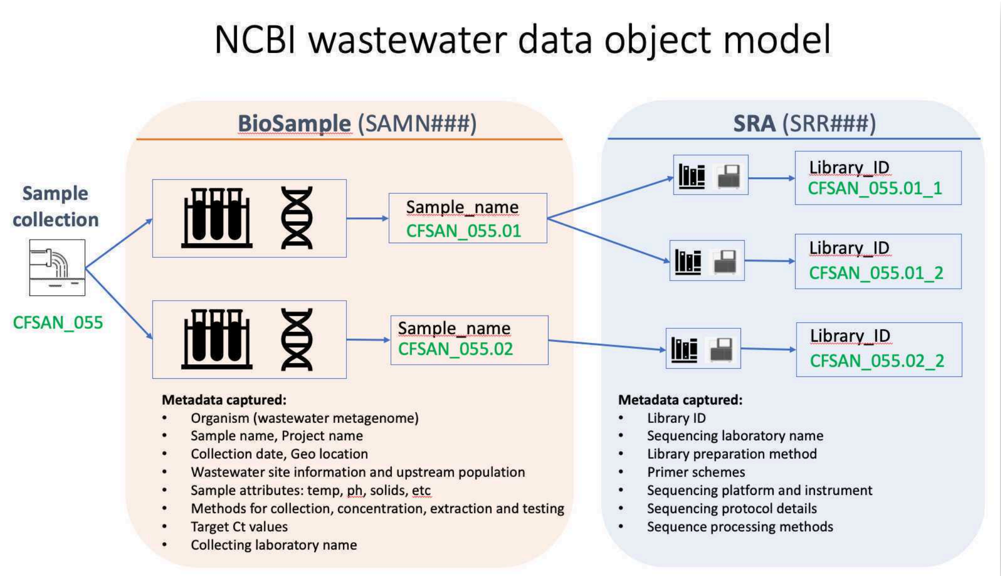
Adhering to this general structure is **extremely** important, both for standardizing submission protocols AND for helping to standardize data (and metadata location) for downstream analysis. *Note: GenomeTrakr Umbrella project listed only as an example. Other efforts may/may not have an umbrella project above the primary data BioProject*

This structure includes the following NCBI databases:

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



Example of structuring wastewater samples, RNA extractions, and sequencing data at NCBI.



NCBI wastewater data object model

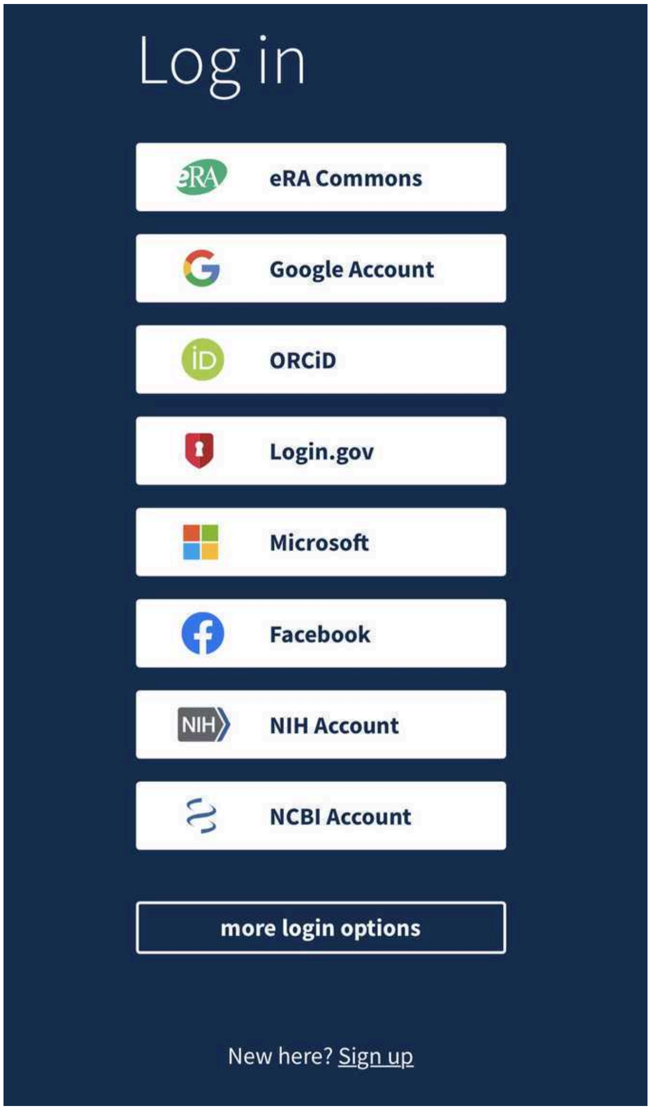
Metadata captured:

- Organism (wastewater metagenome)
- Sample name, Project name
- Collection date, Geo location
- Wastewater site information and upstream population
- Sample attributes: temp, ph, solids, etc
- Methods for collection, concentration, extraction and testing
- Target Ct values
- Collecting laboratory name

Metadata captured:

- Library ID
- Sequencing laboratory name
- Library preparation method
- Primer schemes
- Sequencing platform and instrument
- Sequencing protocol details
- Sequence processing methods

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



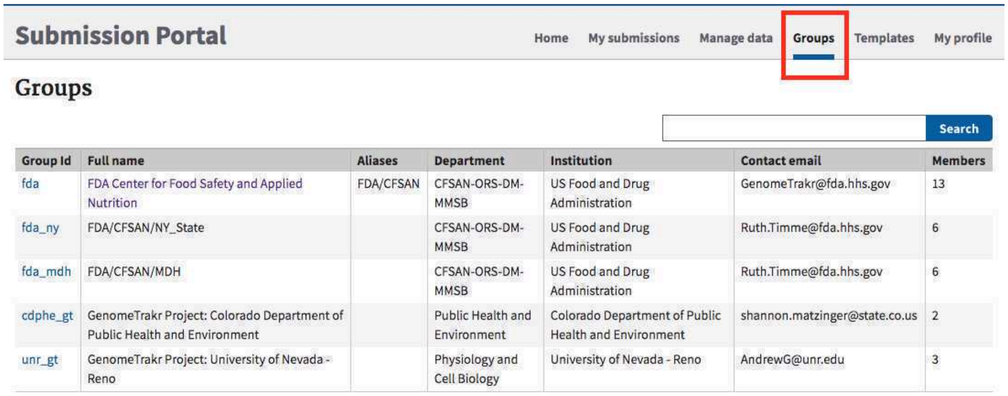
Log in

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).

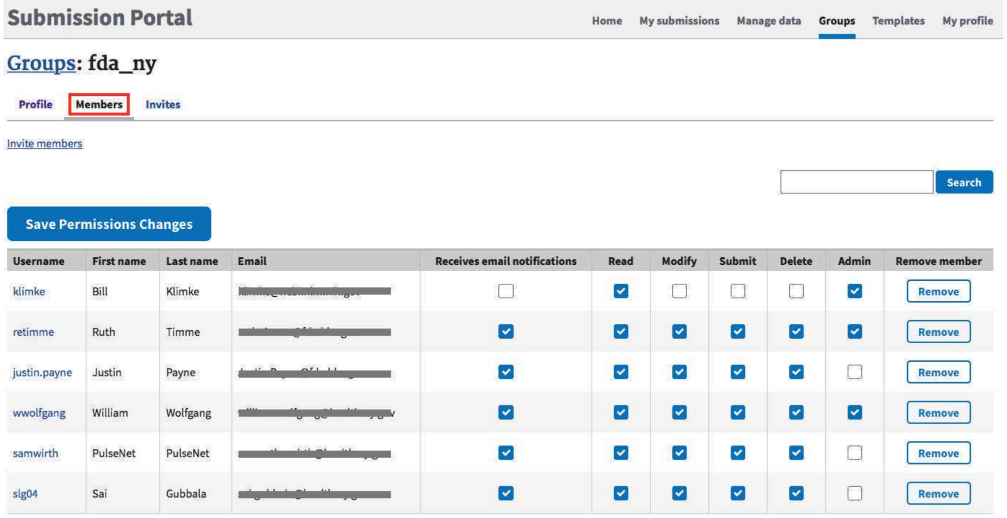
Thank you,"

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.



Groups: fda_ny

Username	First name	Last name	Email	Receives email notifications	Read	Modify	Submit	Delete	Admin	Remove member
klimke	Bill	Klimke	[redacted]	☐	☑	☐	☐	☐	☑	Remove
retimme	Ruth	Timme	[redacted]	☑	☑	☑	☑	☑	☑	Remove
justin.payne	Justin	Payne	[redacted]	☑	☑	☑	☑	☑	☐	Remove
wwolfgang	William	Wolfgang	[redacted]	☑	☑	☑	☑	☑	☑	Remove
samwirth	PulseNet	PulseNet	[redacted]	☑	☑	☑	☑	☑	☐	Remove
slg04	Sai	Gubbala	[redacted]	☑	☑	☑	☑	☑	☐	Remove

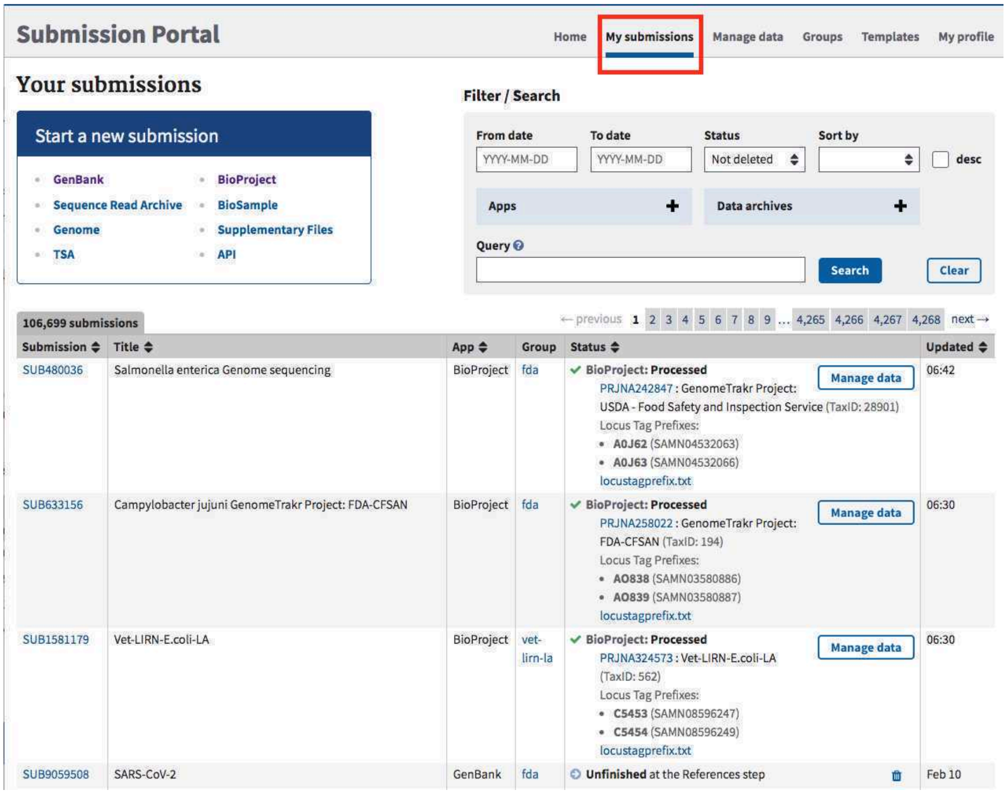
This user group should be kept up-to-date as members enter and leave the laboratory.

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.

1.5 **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**.



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 jujuni GenomeTrakr Project: FDA-CFSAN	BioProject	fda	✓ BioProject: Processed PRJNA258022 : GenomeTrakr Project: FDA-CFSAN (TaxID: 194) Locus Tag Prefixes: • AO838 (SAMN03580886) • AO839 (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

1.6 **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.

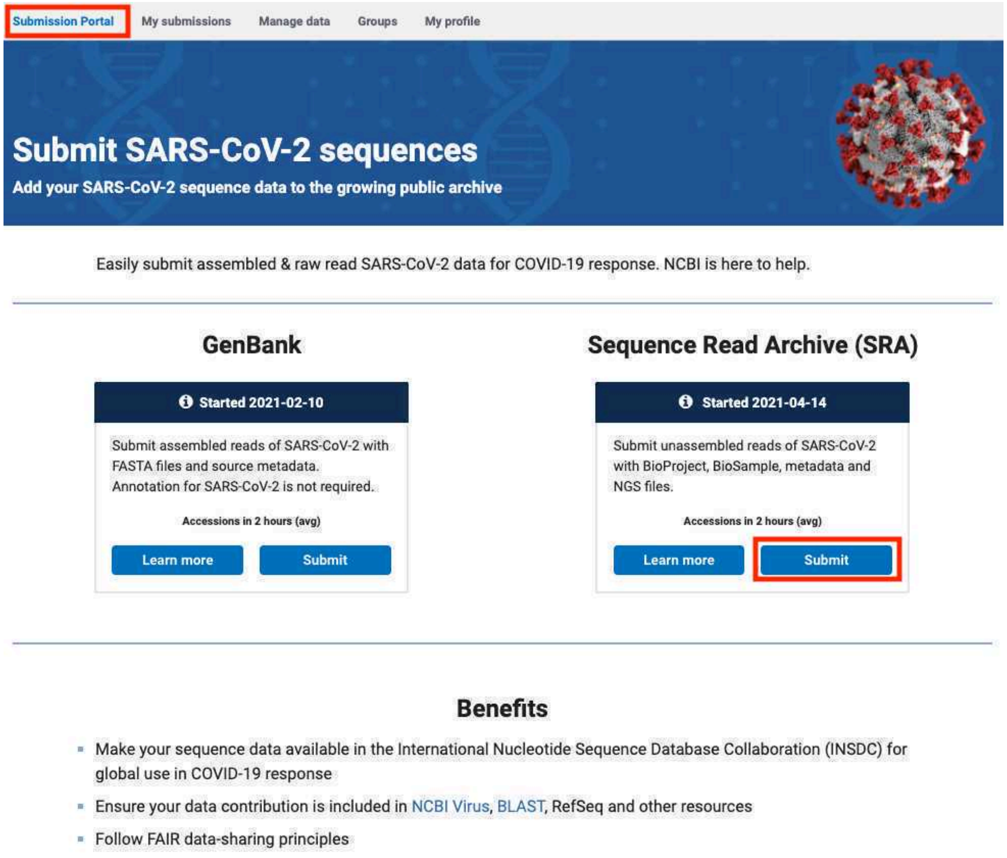
Data submission (BioSample and SRA)

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/



Submit SARS-CoV-2 sequences

Add your SARS-CoV-2 sequence data to the growing public archive

Easily submit assembled & raw read SARS-CoV-2 data for COVID-19 response. NCBI is here to help.

GenBank

Started 2021-02-10

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

Accessions in 2 hours (avg)

Sequence Read Archive (SRA)

Started 2021-04-14

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

Accessions in 2 hours (avg)

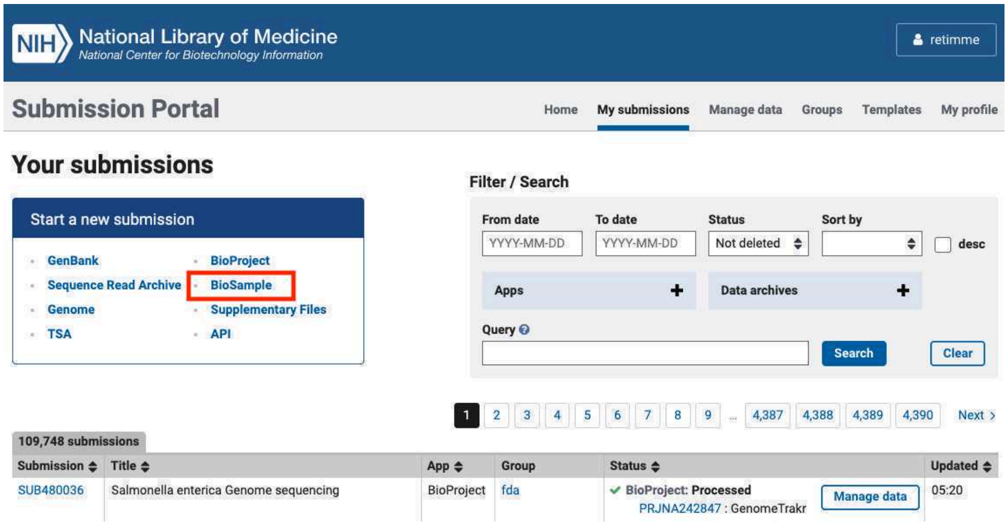
Benefits

- 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

Note

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



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

109,748 submissions

Submission	Title	App	Group	Status	Updated
SUB480036	Salmonella enterica Genome sequencing	BioProject	fda	✓ BioProject: Processed PRJNA242847 : GenomeTrakr	05:20

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_v1.9.xlsx

2. SRA: custom extension of NCBI's SRA metadata template (**see note below for previously registered biosamples):

SRA_ww_template_v5.7.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.

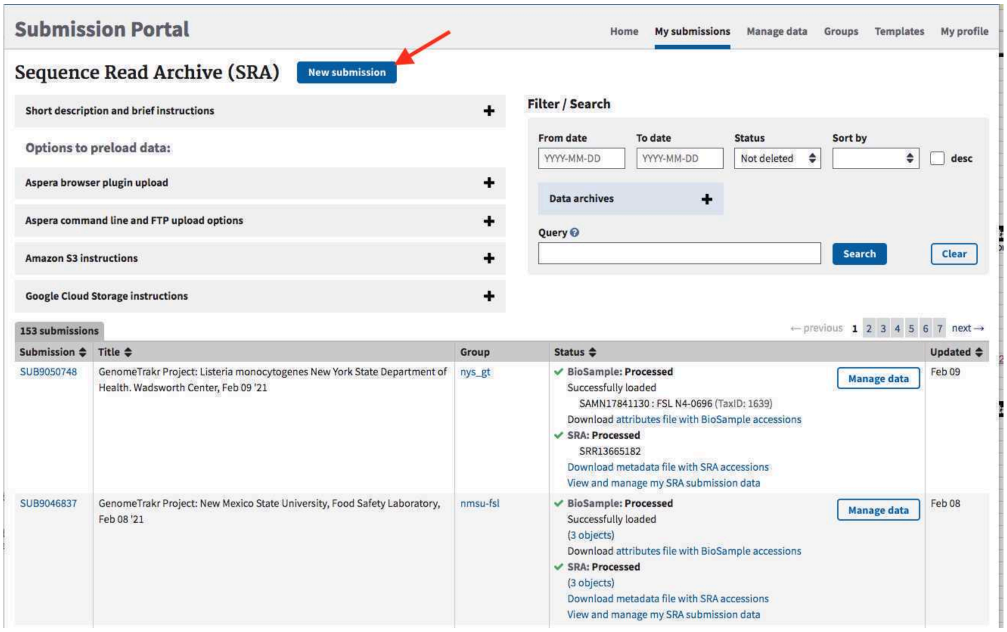
Note

For sequence submissions to previously-registered BioSamples (already have SAMN Ids):

SRA template modification: Change the name of the first column from "sample_name" to "biosample_accession" and populate this column with the respective SAMN#s for the sequences you are uploading..

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.**

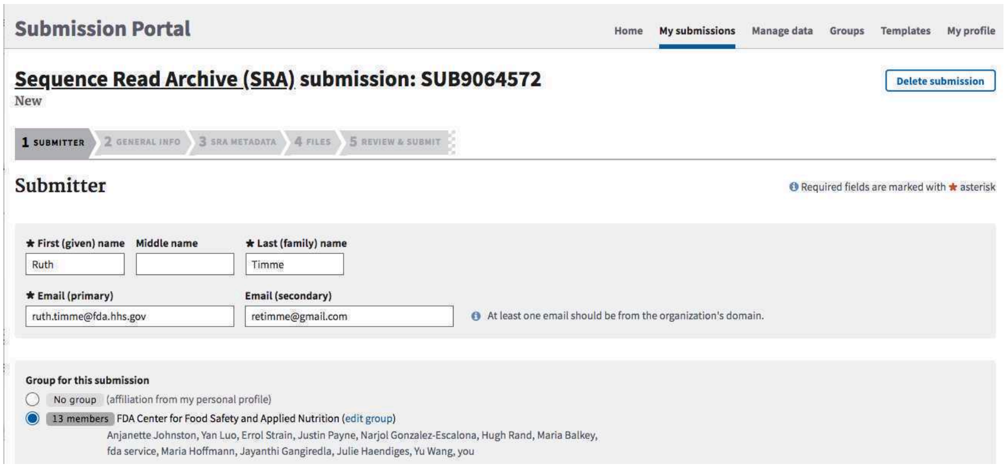


Submission	Title	Group	Status		Updated
SUB9050748	GenomeTrakr Project: Listeria monocytogenes New York State Department of Health. Wadsworth Center, Feb 09 '21	nys_gt	✓ BioSample: Processed Successfully loaded SAMN17841130 : FSL N4-0696 (TaxID: 1639) Download attributes file with BioSample accessions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



Sequence Read Archive (SRA) submission: SUB9064572

New

Submitter

Required fields are marked with ★ asterisk

★ First (given) name: Ruth

Middle name:

★ Last (family) name: Timme

★ Email (primary): ruth.timme@fda.hhs.gov

Email (secondary): retimme@gmail.com

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 Gangiredla, 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 (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, then complete the submission (this will give you some time to establish the human read scrubbing flag prior to data release.)

After you recieve your SRR accessions, send the following email to **sra@ncbi.nlm.nih.gov** asap:

Note

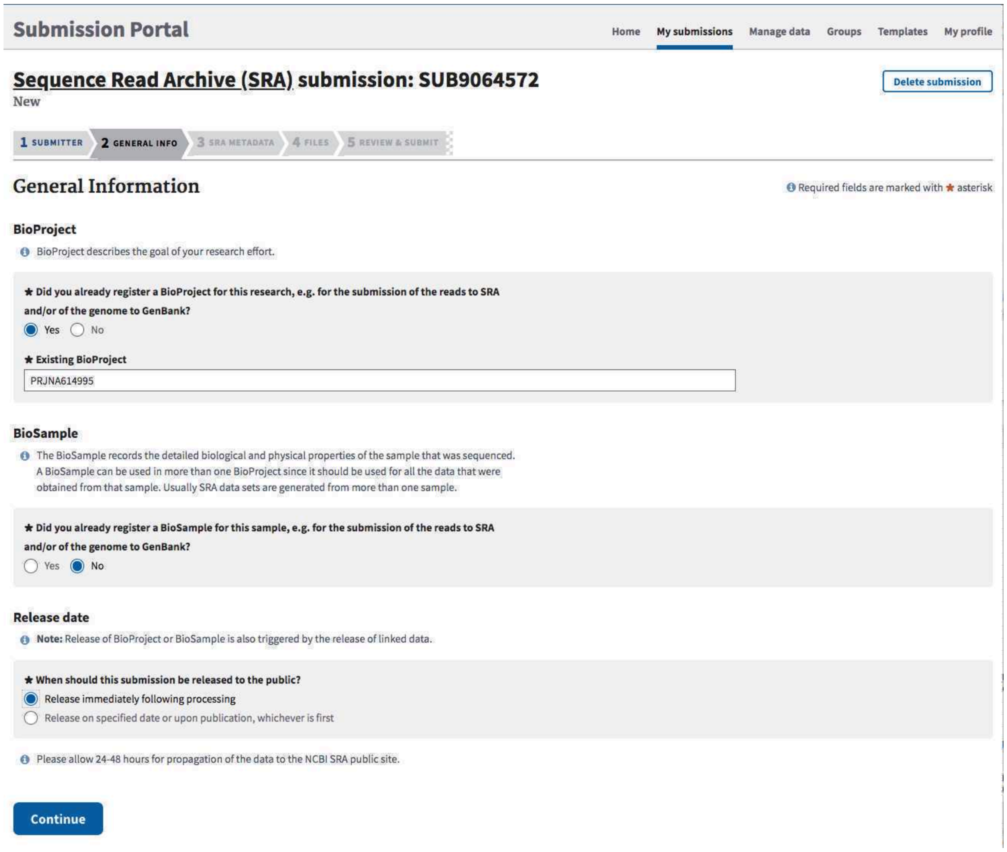
Hi sra,

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

<Include list of SRR accessions>

thanks,

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



Sequence Read Archive (SRA) submission: SUB9064572

New

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

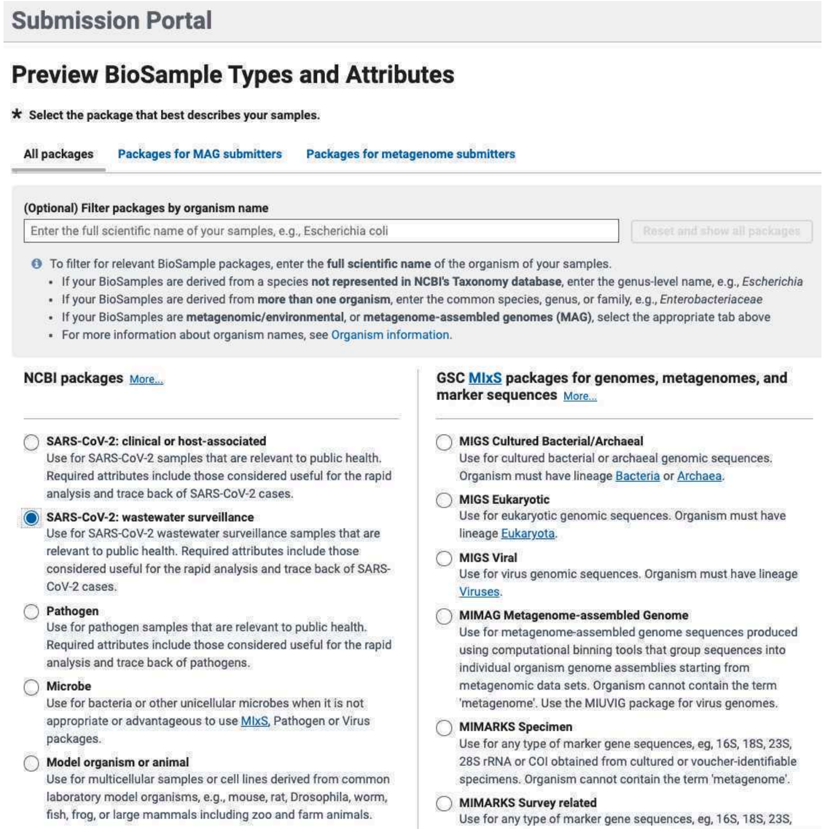
Please allow 24-48 hours for propagation of the data to the NCBI SRA public site.

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**"



Submission Portal

Preview BioSample Types and Attributes

* 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.

● **SARS-CoV-2: wastewater surveillance**
Use for SARS-CoV-2 wastewater surveillance 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 MIxS, 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.

GSC MIxS 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.

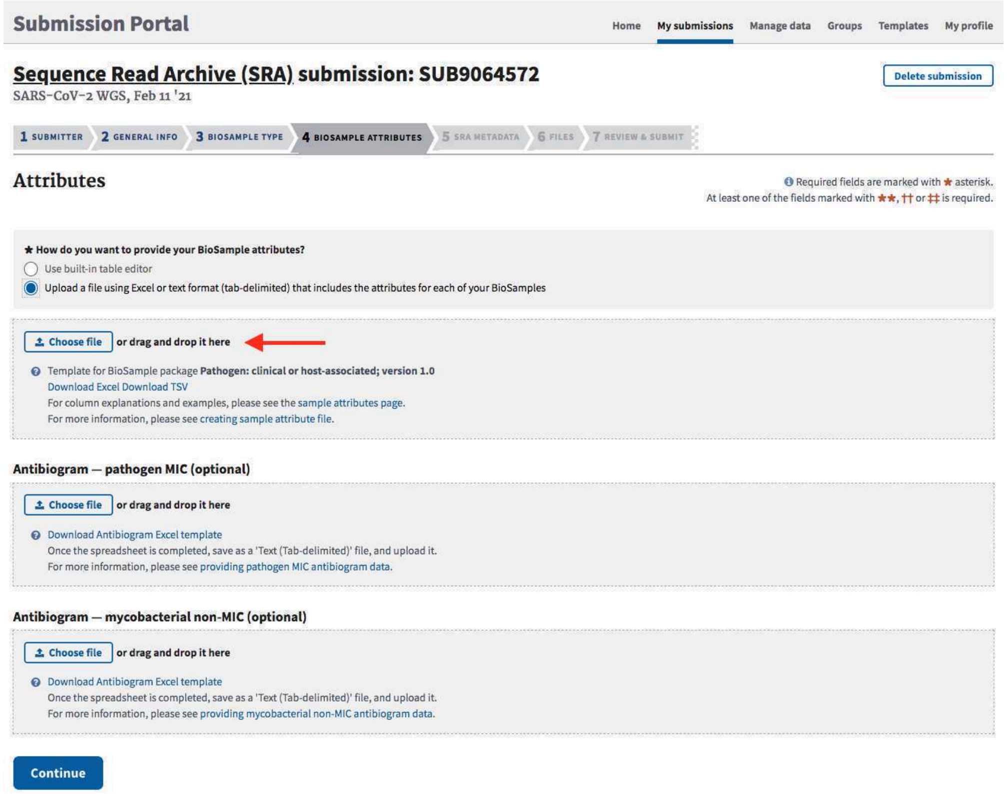
○ **MIMAG Metagenome-assembled Genome**
Use for metagenome-assembled genome sequences produced using computational binning tools that group sequences into individual organism genome assemblies starting from metagenomic data sets. Organism cannot contain the term 'metagenome'. Use the MIUVIG package for 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,

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".



Sequence Read Archive (SRA) submission: SUB9064572
SARS-CoV-2 WGS, Feb 11 '21

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 **Pathogen: 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.

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)

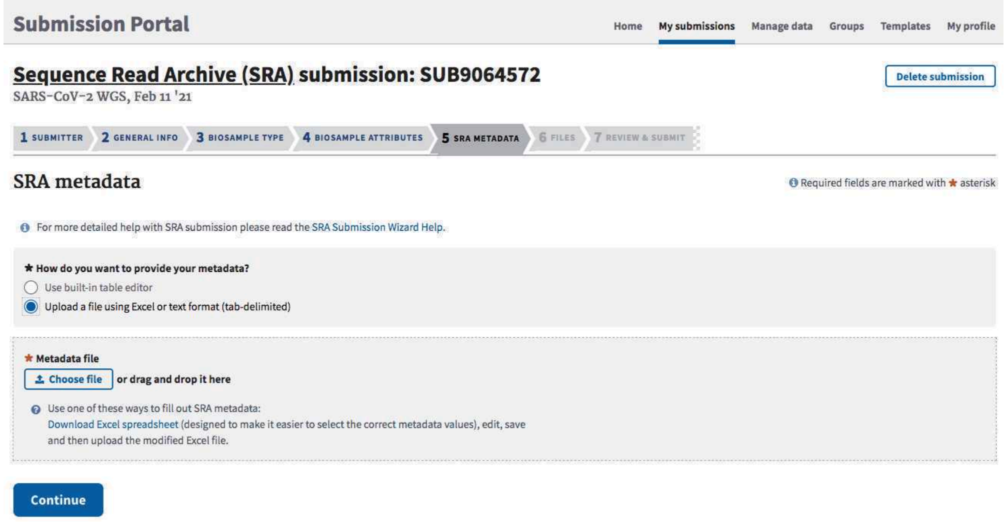
Click "**Continue**".

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)"



Sequence Read Archive (SRA) submission: SUB9064572

SARS-CoV-2 WGS, Feb 11 '21

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

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.

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

Click **"Continue".**

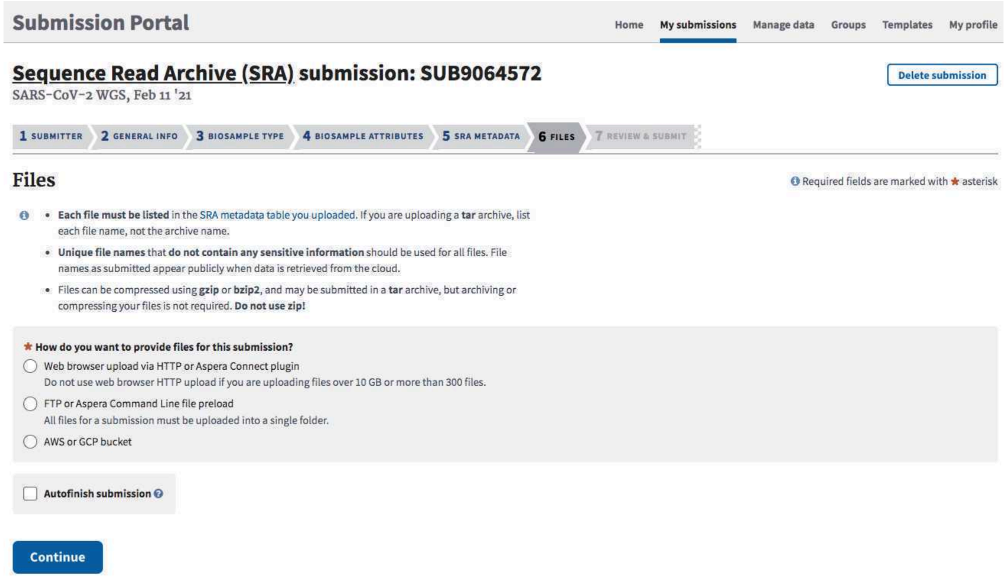
NCBI will do a validation check on your metadata. Resolve any red "errors" reported back by editing the spreadsheet and replacing the uploaded file. Review any yellow "Warnings" and proceed if everything looks ok.

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



Sequence Read Archive (SRA) submission: SUB9064572

SARS-CoV-2 WGS, Feb 11 '21

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**

2.9 REVIEW & SUBMIT tab:

Check over your entire submission, then click submit.

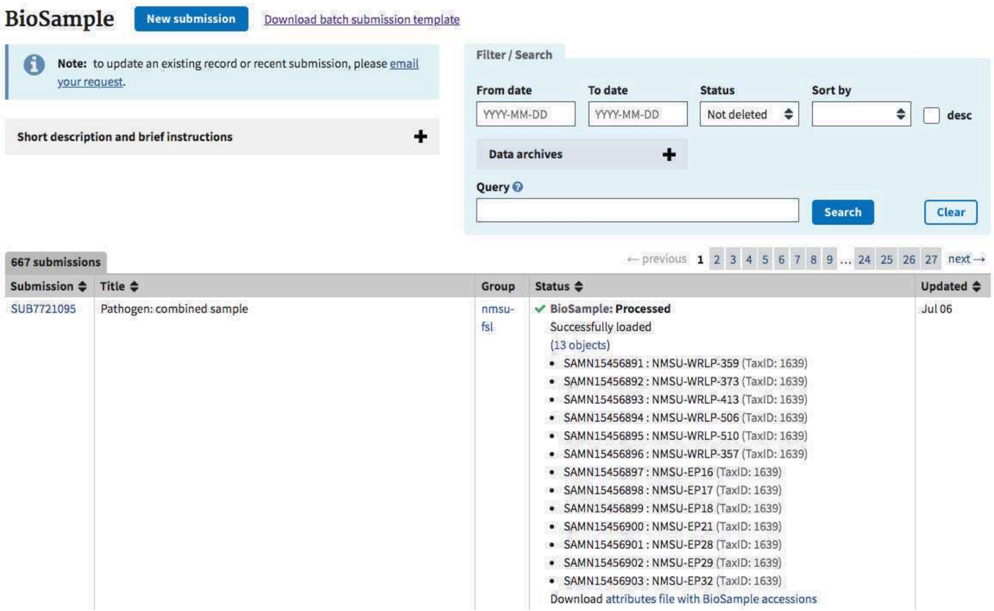
If corrections are needed, you can go back and select individual tabs to edit your submission.

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.



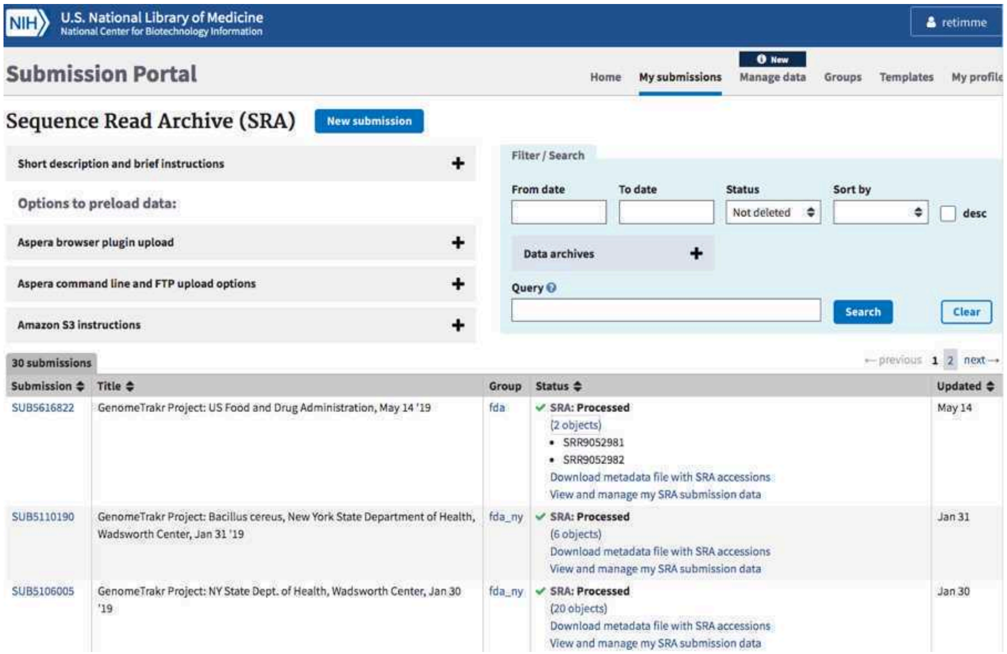
Note: to update an existing record or recent submission, please email your request.

667 submissions

Submission	Title	Group	Status	Updated
SUB7721095	Pathogen: combined sample	nmsu-fsl	✓ BioSample: Processed Successfully loaded (13 objects) • 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.



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.

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.

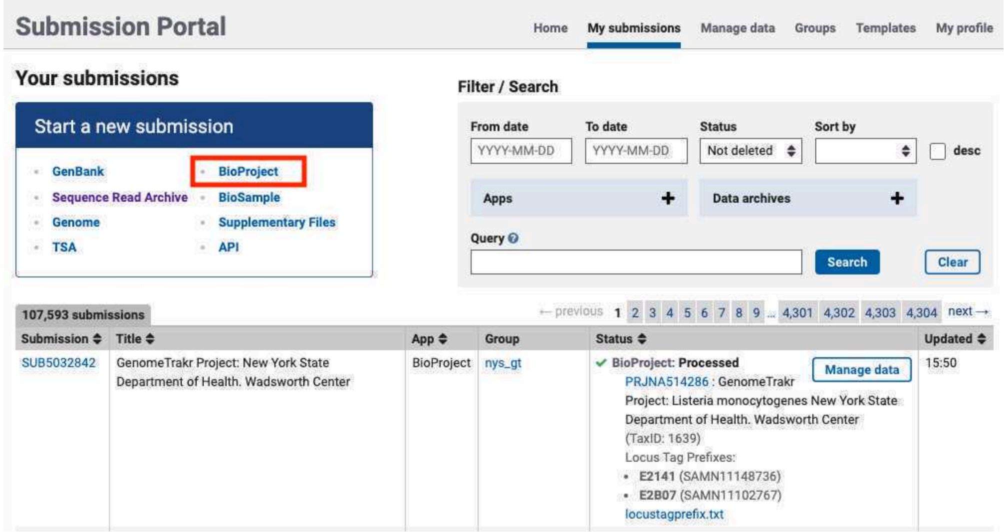
This protocol describes the steps for creating a new *data* BioProject linked to an existing *umbrella* BioProject (usually established by a coordinating group, e.g. CDC NWSS or FDA GenomeTrakr).

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.

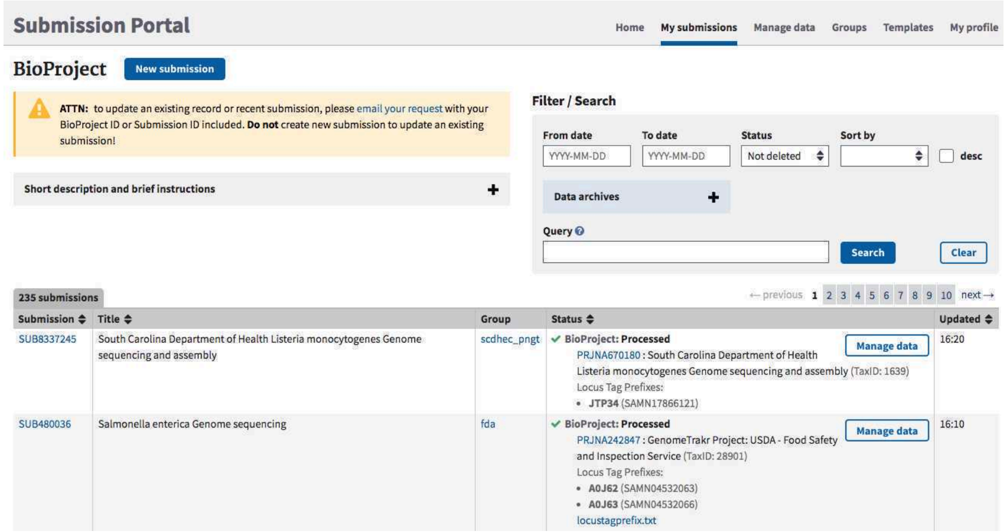
3.1

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



Submission	Title	App	Group	Status	Updated
SUB5032842	GenomeTrakr Project: New York State Department of Health. Wadsworth Center	BioProject	nys_gt	✓ BioProject: Processed Manage data 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:**



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!

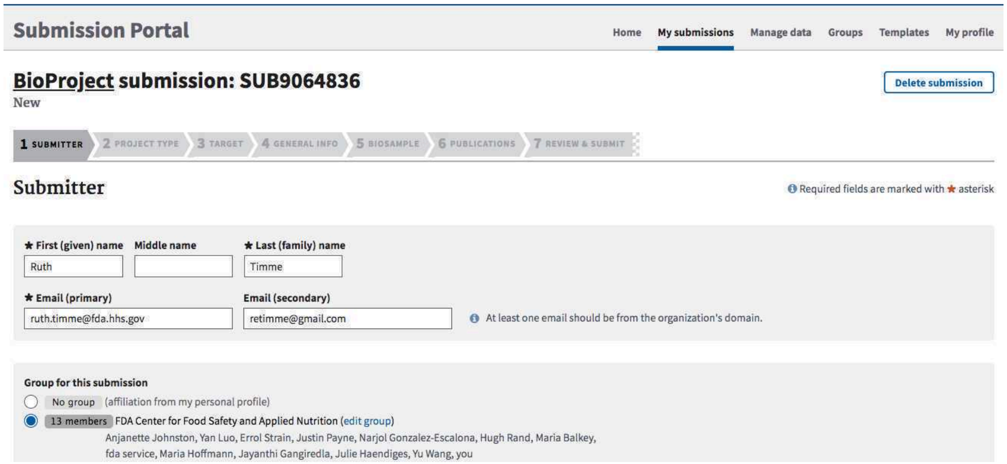
From date: YYYY-MM-DD
To date: YYYY-MM-DD
Status: Not deleted
Sort by:
☐ desc
Query:

Submission	Title	Group	Status	Updated
SUB8337245	South Carolina Department of Health Listeria monocytogenes Genome sequencing and assembly	scdhec_pngt	✓ **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.



BioProject submission: SUB9064836
New

Submitter

Required fields are marked with ★ asterisk

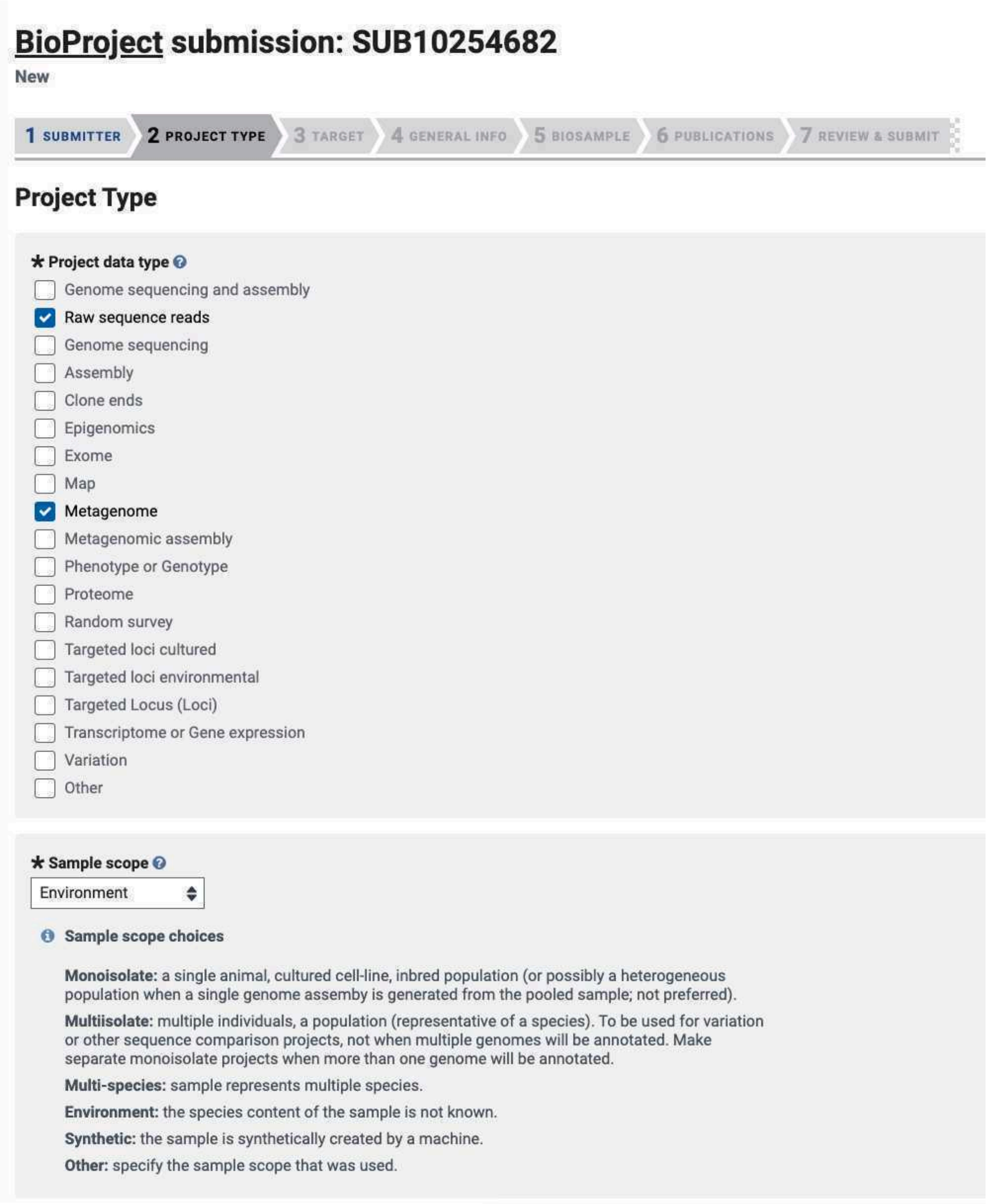
★ First (given) name: Ruth
Middle name:
★ Last (family) name: Timme
★ Email (primary): ruth.timme@fda.hhs.gov
Email (secondary): retimme@gmail.com
At least one email should be from the organization's domain.

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 Gangiredla, Julie Haendiges, Yu Wang, you

3.4 **PROJECT TYPE tab:**

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

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



BioProject submission: SUB10254682
New

Project Type

★ 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
Environment

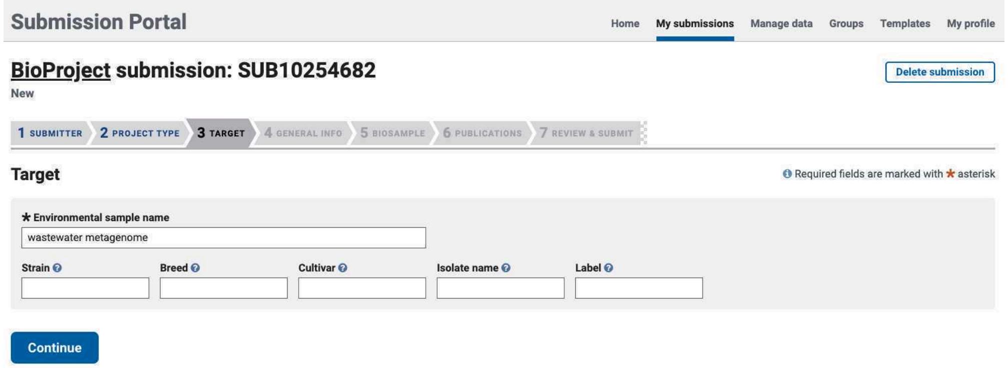
Sample scope choices

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).
Multiisolate: 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.

3.5 **TARGET tab:**

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

Leave the strain info fields blank.



BioProject submission: SUB10254682

New

Target

Required fields are marked with * asterisk

* Environmental sample name: wastewater metagenome

Strain:

Breed:

Cultivar:

Isolate name:

Label:

3.6 **GENERAL INFO tab:**

Click “Release immediately following processing”.

Project Title: e.g., “<Consortium/network name> 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 <Consortium, agency, laboratory name> for monitoring SARS-CoV-2 variants in wastewater.”

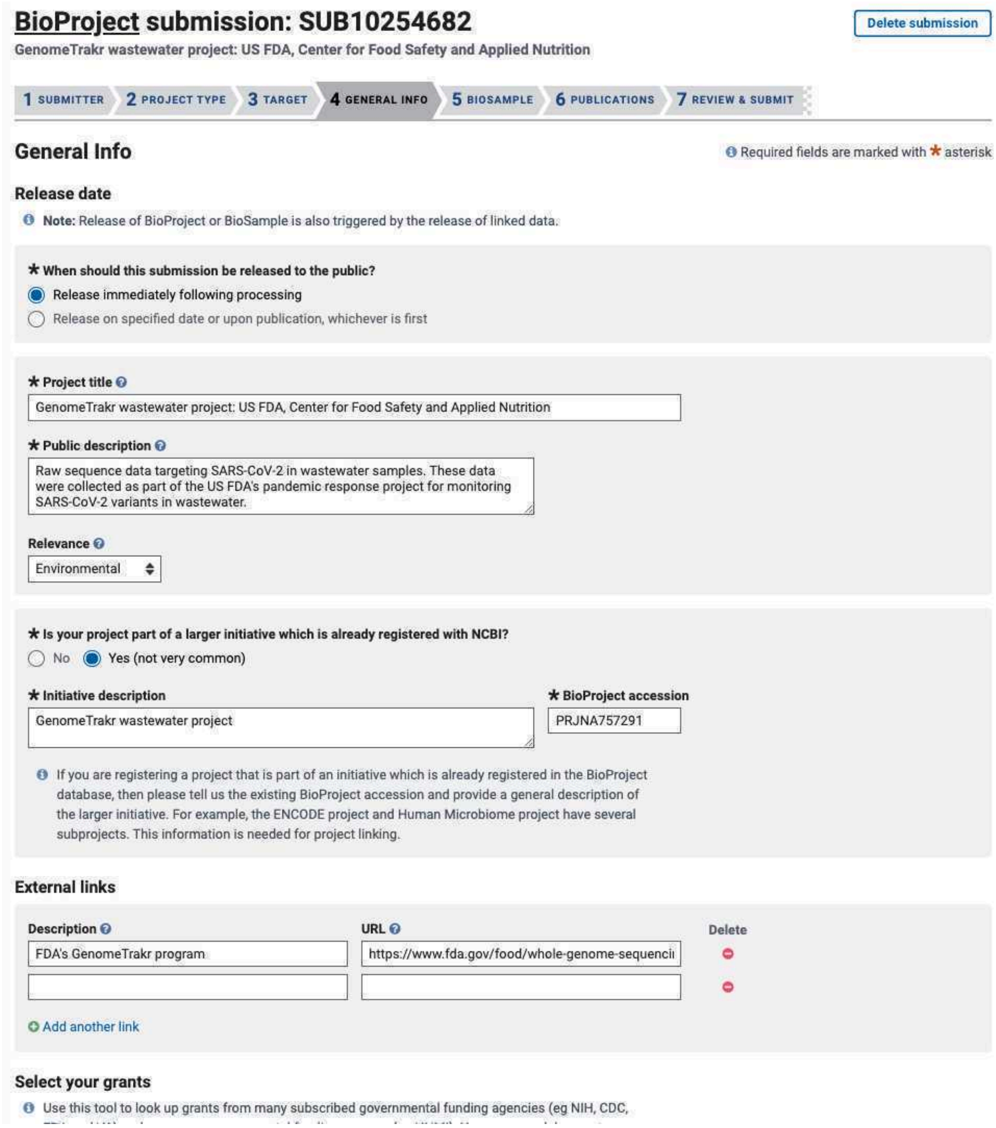
Relevance: Environmental.

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

- Click “Yes”
- *Initiative Description*: "<provide initiative description, e.g. GenomeTrakr, NWSS, etc>"
- *BioProject accession:* **PRJNA#####**, Insert the Umbrella BioProject you want to link to here: (GenomeTrakr: **PRJNA757291,** CDC NWSS**: PRJNA747181** (or other relevant consortium umbrella Bioproject).

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

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



BioProject submission: SUB10254682

GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition

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: GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition

* Public description: 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 which is already registered with NCBI?

○ No ● Yes (not very common)

* Initiative description: GenomeTrakr wastewater project

* BioProject accession: PRJNA757291

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
FDA's GenomeTrakr program	https://www.fda.gov/food/whole-genome-sequencii	⊖
		⊖

Select your grants

Use this tool to look up grants from many subscribed governmental funding agencies (eg NIH, CDC,

3.7 *BioSample tab:*

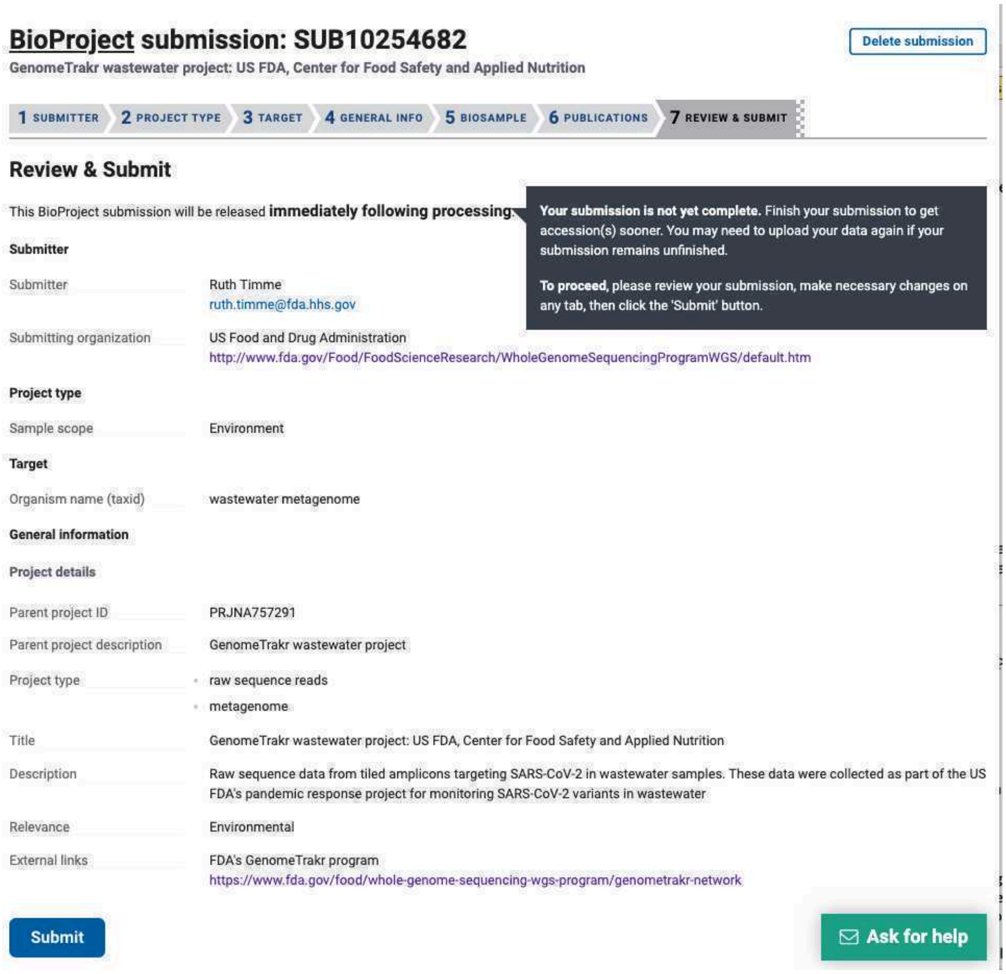
Leave blank!! You will create biosamples separately.

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**.”



BioProject submission: SUB10254682

GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition

Review & Submit

This BioProject submission will be released **immediately following processing**

Your submission is not yet complete. Finish your submission to get accession(s) sooner. You may need to upload your data again if your submission remains unfinished.

To proceed, please review your submission, make necessary changes on any tab, then click the 'Submit' button.

Submitter

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

Sample scope	Environment

Target

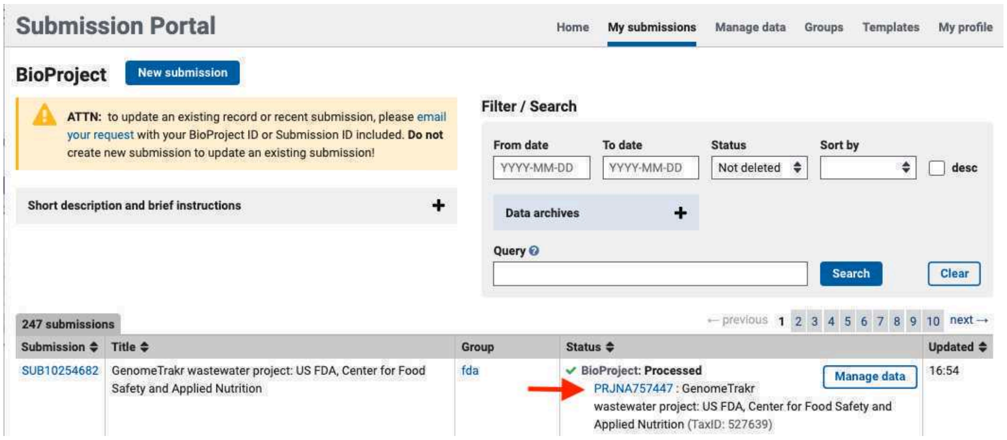
Organism name (taxid)	wastewater metagenome

General information

Project details

Parent project ID	PRJNA757291
Parent project description	GenomeTrakr wastewater project
Project type	raw sequence reads metagenome
Title	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition
Description	Raw sequence data from tiled amplicons 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
External links	FDA's GenomeTrakr program https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network

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.



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!

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

247 submissions

Submission	Title	Group	Status		Updated
SUB10254682	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition	fda	✓ BioProject: Processed PRJNA757447 : GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition (TaxID: 527639)	Manage data	16:54

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.

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	Manage data	16:10