VERSION 5

FEB 26, 2024

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.36wgq5jb5gk5/v5

Protocol Citation: Tina Lusk Pfefer, Ruth Timme, Candace Hope Bias, Errol Strain, Maria Balkey 2024. NCBI Bacterial Pathogen Data Curation Protocol: SOP for Editing GenomeTrakr Submissions. **protocols.io** https://dx.doi.org/10.17504/protocols.io.36wgq5jb5gk5/v5Version created by Ruth Timme

NCBI Bacterial Pathogen Data Curation Protocol: SOP for Editing GenomeTrakr Submissions V.5

In 5 collections

Tina Lusk Pfefer[1], Ruth Timme[2], Candace Hope Bias[1], Errol Strain[3], Maria Balkey[1]

[1]Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA;
[2]US Food and Drug Administration;
[3]U.S. Food and Drug Administration, College Park, Maryland, USA

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

Ruth Timme
US Food and Drug Administration

DISCLAIMER

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

MANUSCRIPT CITATION:
Timme, R.E., Wolfgang, W.J., Balkey, M. et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook 2, 20 (2020). https://doi.org/10.1186/s42522-020-00026-3. Timme R.E., Sanchez Leon M., Allard M.W. (2019) Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. In: Bridier A. (eds) Foodborne Bacterial Pathogens. Methods in Molecular Biology, vol 1918. Humana, New York, NY. https://doi.org/10.1007/978-1-4939-9000-9_17

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

Created: Jan 09, 2024

Last Modified: Feb 26, 2024

PROTOCOL integer ID: 93197

Keywords: NCBI submission, GenomeTrakr, curation, genomic pathogen surveillance

ABSTRACT

PURPOSE: After data are submitted to NCBI submitters often encounter the need to update, retract, or replace these records. This is called data curation. This protocol provides instructions for making data curation requests at NCBI.

SCOPE: This protocol applies specifically to NCBI pathogen genome submissions falling within the scope of Pathogen Detection efforts (see here). Briefly, this includes whole genome sequence data submissions of bacterial pathogens, which is the primary submission type for FDA's GenomeTrakr network.

Version history:
V5: Significant edits to the protocol including new guidance for primary contacts at NCBI. This protocol was also forked, with the current version focused on whole genome sequence data for bacterial pathogens, and the other protocol (in development) focusing on other data types for pathogens (metagenomic, targeted amplicon, other enrichment panels).
V4: Clarifying protocol for SRA retraction.
V3: Update to BioSample section, providing further guidance on updating taxonomic names.
V2. Edit submissions using the NCBI portal (Manage data). Moved "how to find my data" content to a new protocol

BEFORE START INSTRUCTIONS

This protocol applies specifically to NCBI pathogen genome submissions falling within the scope of Pathogen Detection efforts (see here). Briefly, this includes whole genome sequence data submissions of bacterial pathogens.

For curation requests for data that align with these criteria, the NCBI Pathogen Detection team will serve as your primary contact at NCBI: pd-help@ncbi.nlm.nih.gov. They will coordinate with other NCBI databases to manage each curation request, covering BioSample, BioProject, SRA, GenBank, and Pathogen Detection.

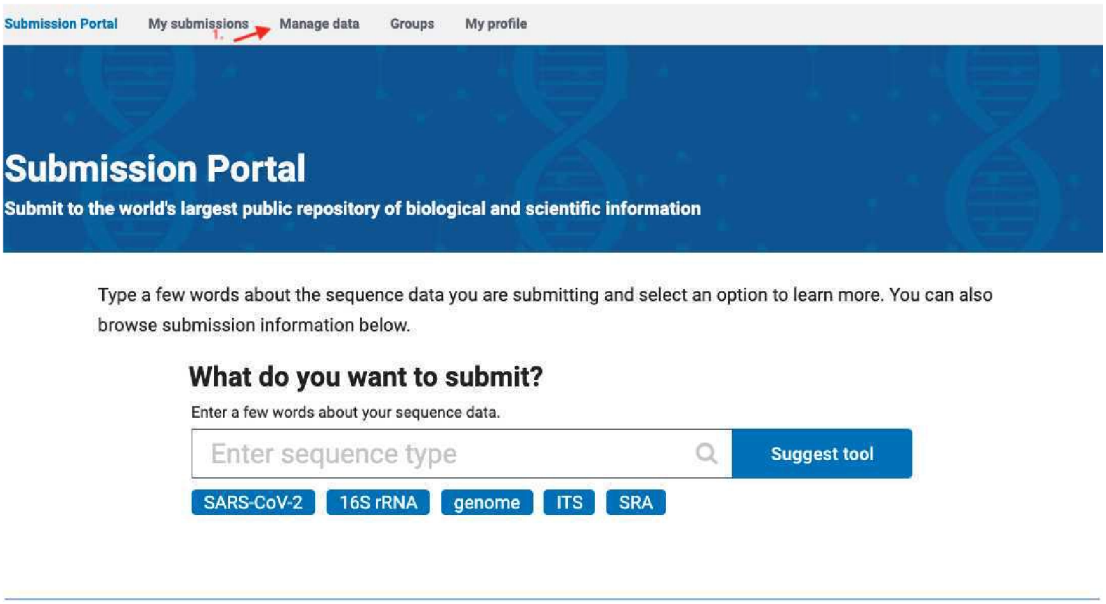
For NCBI submissions that fall outside the purview of the Pathogen Detection pipeline, including viral genomes, targeted amplicon datasets, data derived from NGS pathogen panels, or, specifically, SARS-CoV-2 in wastewater, the curation process will be performed by each respective database team. Protocol in development.

BioProject Curation for BioProjects linked to NCBI Pathogen Detection

1 How to make edits to BioProject records:

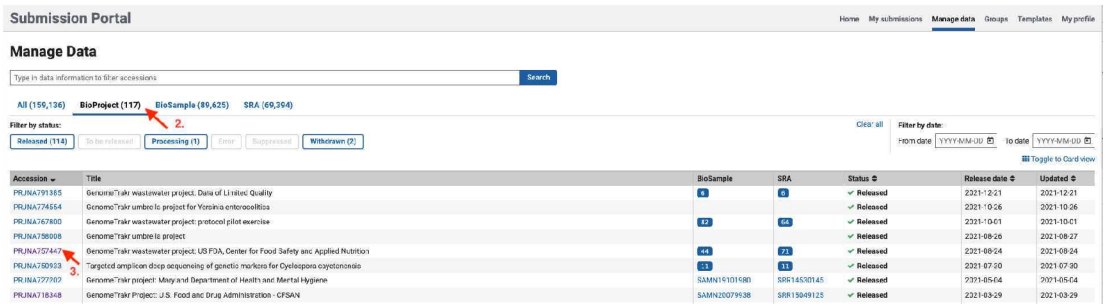
1.1 To edit Title, Organism, Description, URL, or publications for your BioProject, follow steps 1-6 below.

1. Click on the "Manage Data" tab within the submission portal, or navigate directly to "Manage Data": https://dataview.ncbi.nlm.nih.gov

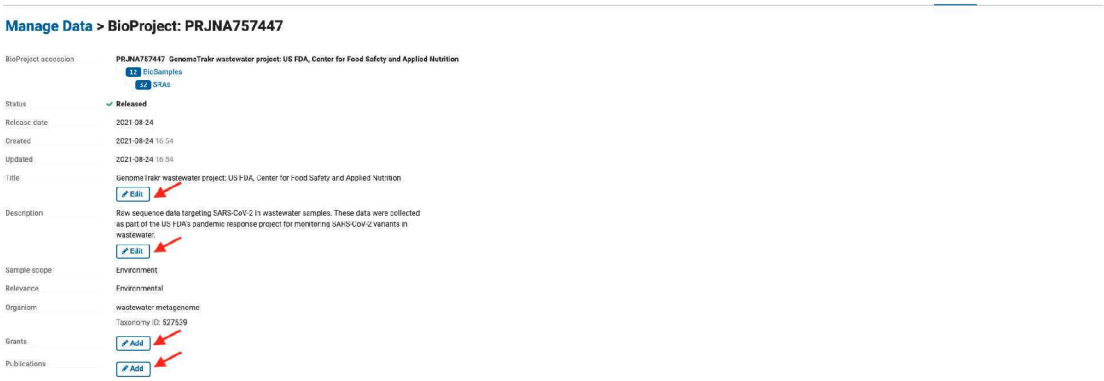


2. In the menu, select the "**BioProject (##)"** tab. A complete list of your NCBI group bioprojects will be displayed.

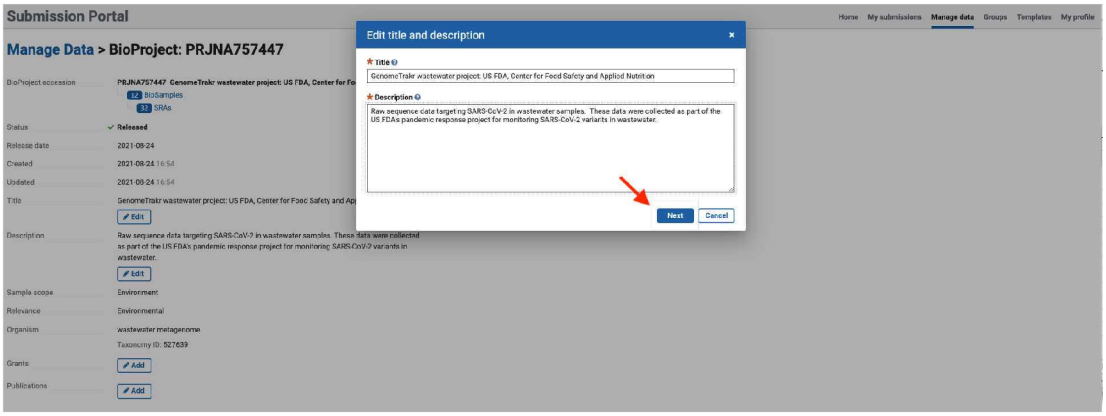
3. Click on the BioProject that you need to edit.

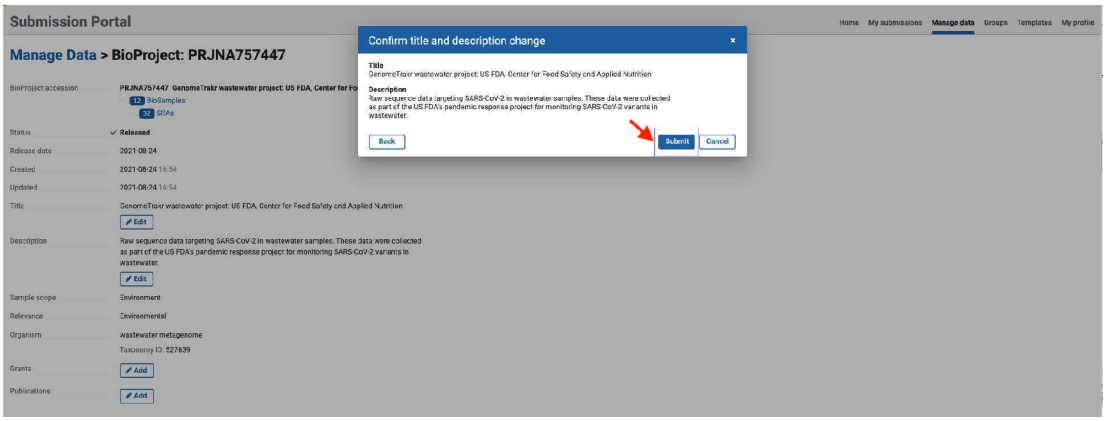


4. Fields available for editing will be displayed after selecting a BioProject.

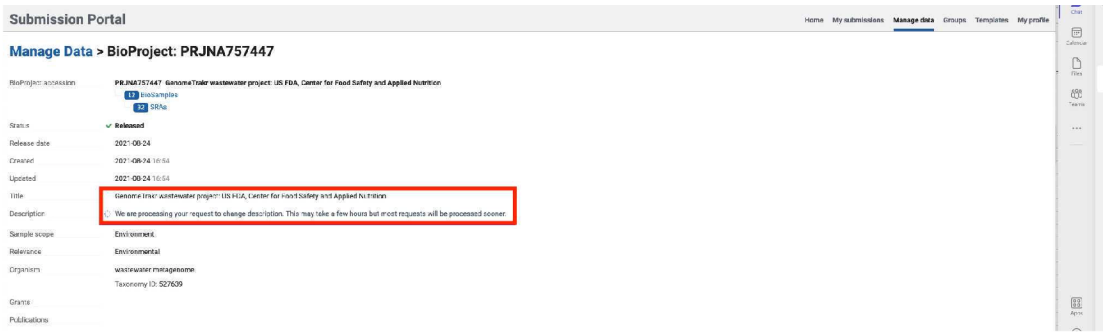


5. Click in any of the edit/add fields and proceed to add the corresponding BioProject information. Once the information is changed or added, click next and submit.





6. A confirmation prompt will indicate that your updates are in progress.



1.2 **To request additional assistance with your BioProject, follow steps 1 and 2 below. This includes, but is not limited to:**

- **Questions about errors or processing of a BioProject submission**
- **Convert a Data BioProject to an Umbrella BioProject**
- **Re-assign a BioProject from one Umbrella BioProject to another**

1. For Pathogen Detection submissions ONLY:
Send an email to PD-help (**pd-help@ncbi.nlm.nih.gov**), so they can ensure all linked records are changed (GenBank, etc.). Include the BioProject accession in the email subject line.

2. For all other submissions (non-Pathogen Detection), send an email to:
bioprojecthelp@ncbi.nlm.nih.gov. Include the BioProject accession in the email subject line.

BioSample Curation for records included NCBI Pathogen Detection

2 **How to edit BioSamples:**

2.1 All edits or updates to PD BioSample records are submitted via email to PD-help:
TO: pd-help@ncbi.nlm.nih.gov
Send all change and retraction requests to PD-help, so they can ensure all linked records are changed (GenBank, etc.).

Use this email for the following tasks. Include your lab and the request date in your subject line for easy tracking, eg “FDA BioSample update, Dec 10, 2019”.

- Questions about validation errors or processing of a BioSample submission.
- Update, correct, or add fields/attributes to a BioSample(s)
- Retraction
- Add a linkage or re-assign linkage to a BioProject
- Add or change a strain or isolate field to an existing BioSample where one has been lacking (necessary for the isolate's assembly to appear in GenBank). NOTE, there is now a list of terms that results in a failure to process the isolate and it will not be processed at all in Pathogen Detection. Do not use these terms in the strain/isolate fields:

1. bacteria
2. sp.
3. strain
4. environmental
5. soil
6. clinical isolate
7. NA
8. whole organism
9. Microbial
10. Any kind of taxonomic information, such as genus name or species name

- Taxonomic updates: send to "pd-help@ncbi.hlm.nih.gov" on these requests to ensure taxonomic changes get propagated fully across NCBI databases. The organism’s name should include the binomial name (Genus species), subspecies where present, plus serovar/serotype information. In cases where the BioSample attributes serovar/serotype were populated (e.g. with traditional serotyping results), ensure they are also updated as needed. Special note about Salmonella enterica isolates: please submit or update serotyping information in the serovar field, not the serotype field.

You will receive a confirmation email that the updates were performed. These types of transactions are common for this database, so do not hesitate to submit requests as needed.

2.2 How to retract one or multiple BioSamples

Note

TO: **pd-help@ncbi.nlm.nih.gov**

Dear PD-Help,

Please retract the following BioSamples due to sample mix-ups (or other reason):

SAMN########
SAMN########
SAMN########
SAMN########

Thank you,
Ruth

2.3 How to update content in metadata fields or add new fields/attributes to a BioSample record(s):

Note

TO: pd-help@ncbi.nlm.nih.gov

Dear PD-Help,

Please update the attached BioSample records.

Thanks,
Ruth

Attach a tab-delimited text file with the BioSample accessions in the first column and fields to update the right. You can attach a table to update one or multiple records at a time.

Examples:

FDA_biosample_update_20220203_fb.txt

(adding "sequenced_by" and "project_name" to a biosample)

- The following table will update the collection date and isolation source on one BioSample record:

	BioSample	collection_date	isolation_source
	SAMN12987335	2019-10-12	cilantro

Tab-delimited table for updating a BioSample record.

2.4 **Re-assign a BioSample from one BioProject to another:**
Submit an update request with the new BioProject accession(s) specified in a column. If the BioSample has associated SRA or GenBank data, then please also request that these objects get reassigned to the new BioProject.

Note

TO: pd-help@ncbi.nlm.nih.gov

Dear PD-Help,

Please process the attached BioSample updates and ***remove all previous BioProject links.***

Thanks,
Ruth

SRA curation for records included in NCBI Pathogen Detection

3 **SRA updates and retractions:**

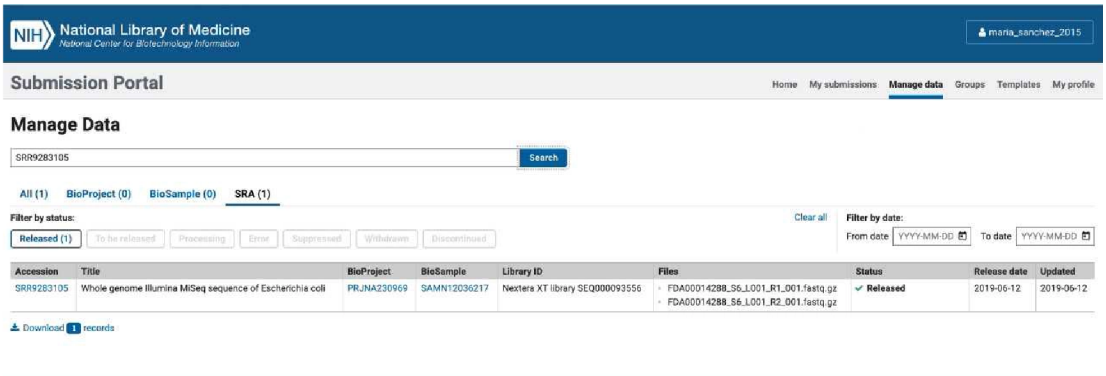
3.1 **Make updates within the submission portal:**

The following types of updates can be made within the submission portal under the “Manage data” tab:

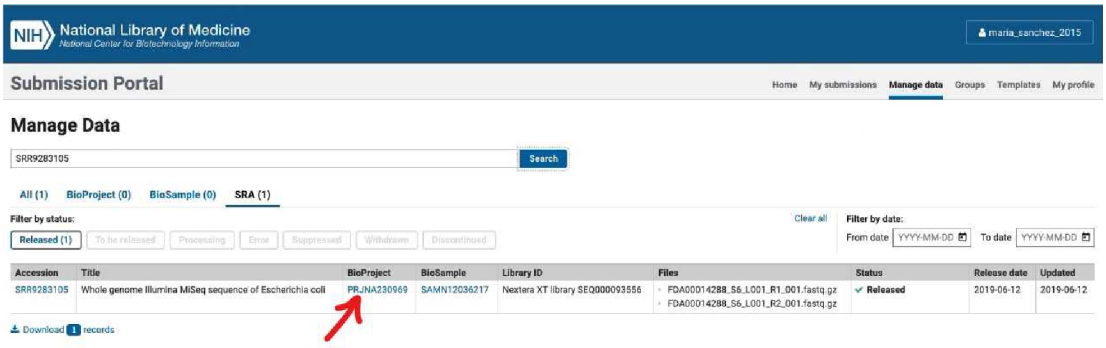
- Sequence metadata, such as library ID, library strategy, sequencing platform or instrument.
- Associated BioSample or BioProject accession numbers
- Release date

1. Click on the "Manage Data" tab within the submission portal, or navigate directly to "Manage Data": https://dataview.ncbi.nlm.nih.gov

2. Query for SRR accession you'd like to update:



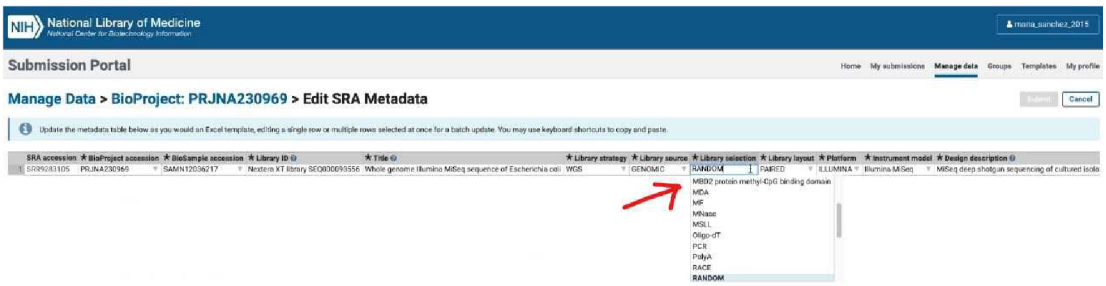
3. Click on the BioProject accession link:



4. All the SRA records submitted to this BioProject can now be edited! Scroll down the BioProject page until the list of SRA records in that BioProject becomes visible and search for the one(s) you want to edit. Select the records you want to edit by clicking the check box beside them.

Once you've made your selection(s), click 'Edit metadata'.

5. You can now edit the metadata directly for this record. For example, if you need to correct a sample-swap you can enter the correct BioSample accession here and the sequence will get re-parented. There are drop-down lists for some attributes.



When you make a change, the field will turn yellow. When you are done making changes, click 'Submit'.

3.2 SRA retraction

An SRA record should *only* be retracted for the following reasons:

1. Discovery of poor quality data. Lab intends to re-generate data (starting at appropriate wet-lab step, re-isolation, DNA extraction, library prep, or sequencing) and re-submit the data.
2. Sample mix-ups that cannot be resolved by re-parenting or correcting the BioSamples. Lab intends to re-generate (starting at appropriate wet-lab step, re-isolation, DNA extraction,

library prep, or sequencing) and re-submit the data.
3. Discovery of multiple runs per isolate. Laboratory would like to have only one run per isolate in the system. No re-sequencing planned.

DO NOT retract an SRA submission, then attempt to re-submit the same files. This will get flagged as a duplicate within NCBI's validation check and will be rejected.

Emails for SRA retraction: pd-help@ncbi.nlm.nih.gov
Send all retraction requests to PD-help, so they can ensure all linked records are retracted (GenBank, etc.).

Emails should include a list of SRR accessions to retract and *reason for retraction* (i.e. sample mix-up, quality of data, etc.).

Email template:

Note

TO: pd-help@ncbi.nlm.nih.gov

SUBJECT: FDA SRA retractions, Dec 10, 2019

Dear PD-Help,

Please retract the following SRR accessions and any linked assemblies or PD analyses due to XXX issue. This request has been submitted using the NCBI submission portal.

We will re-sequence these isolates and re-submit new data.

SRRXXXXXX1
SRRXXXXXX2
SRRXXXXXX3

Thanks,
Ruth

3.3 To move SRA data from one BioProject to another, if not able to do so in the portal:

In the event that submission portal does not allow, and this is not for the specific BioSample attribute in OHE for BioProject Accession, do the following (Note: This a costly change, and labs should ensure this is a rare change):

Send an email to pd-help@ncbi.nlm.nih.gov
Send all move requests to PD-help, so they can ensure all linked records are retracted (GenBank, etc.).