Jan 18, 2023 Version 3

NCBI data curation protocol - SOP for editing GenomeTrakr submissions V.3

DOI

dx.doi.org/10.17504/protocols.io.36wgq5jb5gk5/v3

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

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

GenomeTrakr Springer Nature Books

Ruth Timme
US Food and Drug Administration

DOI: https://dx.doi.org/10.17504/protocols.io.36wgq5jb5gk5/v3

Protocol Citation: Ruth Timme, Candace Hope Bias, Errol Strain, Tina Pfefer, Maria Balkey 2023. NCBI data curation protocol - SOP for editing GenomeTrakr submissions. **protocols.io https://dx.doi.org/10.17504/protocols.io.36wgq5jb5gk5/v3**Version created by **Ruth Timme**

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: December 13, 2022

Last Modified: January 18, 2023

Protocol Integer ID: 73931

Keywords: NCBI submission, GenomeTrakr, curation, genomic pathogen surveillance, ncbi data curation protocol, editing genometrakr submissions purpose, editing genometrakr submission, bioproject biosample sequence read archive, data curation protocol, genometrakr submissions purpose, updating taxonomic name, genometrakr submission, data curation, called data curation, biosample section, ncbi submitter, taxonomic name, ncbi, curation, database, manage data, edit submission, relevant database, new protocol, data

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: 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 keeping these records up-to-date for each relevant database at NCBI.

SCOPE: This protocol covers curation for the following NCBI databases:

- BioProject
- BioSample
- Sequence Read Archive

V2. Edit submissions using the NCBI portal (Manage data). Moved "how to find my data" content to a new protocol.
V3: Update to BioSample section, providing further guidance on updating taxonomic names.

Troubleshooting

Before start

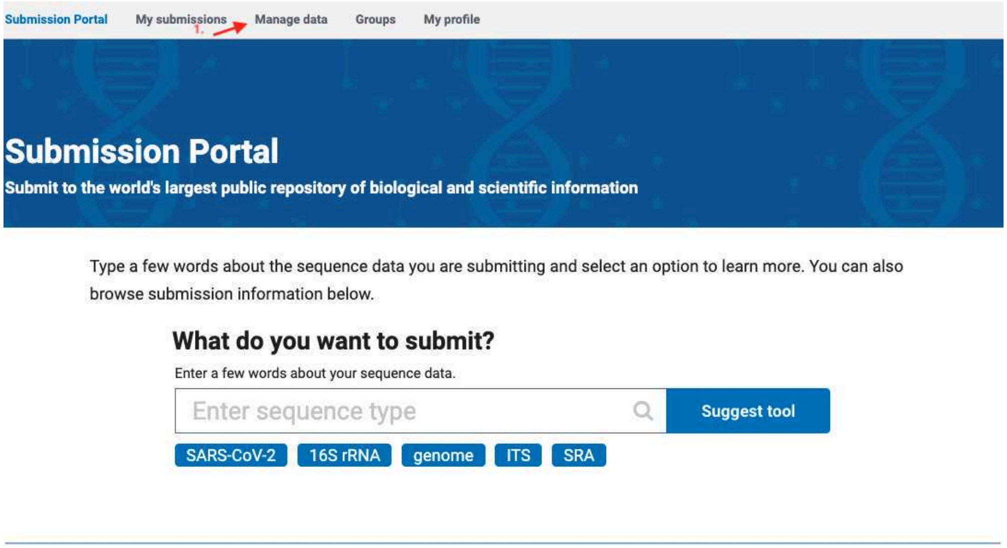
Most updates to existing NCBI submissions are performed through email requests to each respective NCBI database (e.g. BioSample, BioProject, Sequence Read Archive, and Pathogen Detection). NCBI curators within each respective database expect these emails to update and retract data. It is their job to help the data stay current, so do not hesitate to correct errors when they are spotted.

BioProject Curation

1 **How to edit a BioProject**

1.1

1. Click on the "Manage Data" tab within the submission portal, or navigate directly to "Manage Data": https://dataview.ncbi.nlm.nih.gov to edit Title, Organism, Description, URL, or publications for your BioProject.



Submission Portal

Submit to the world's largest public repository of biological and scientific information

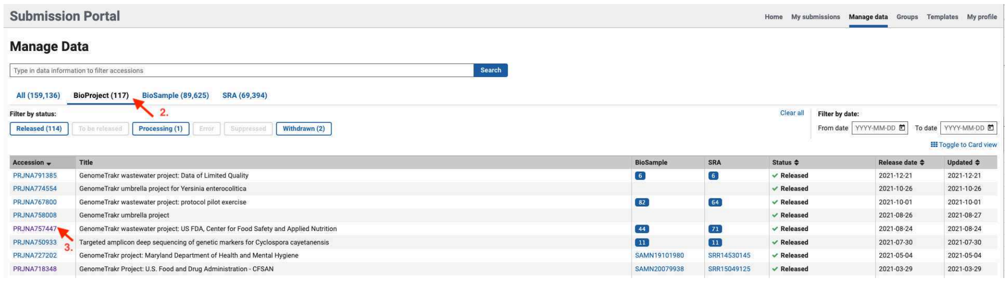
Type a few words about the sequence data you are submitting and select an option to learn more. You can also browse submission information below.

What do you want to submit?

Enter a few words about your sequence data.

2. In the menu, select BioProject, a complete list of your NCBI group bioprojects will be displayed.

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



Manage Data

Type in data information to filter accessions

All (159,136) BioProject (117) BioSample (89,625) SRA (69,394)

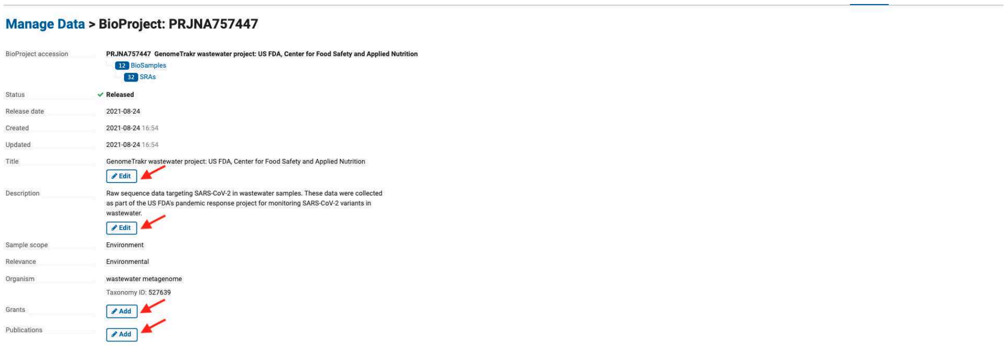
Filter by status: Released (114) To be released Processing (1) Error Suppressed Withdrawn (2)

Clear all

Filter by date: From date YYYY-MM-DD To date YYYY-MM-DD

Accession	Title	BioSample	SRA	Status	Release date	Updated
PRJNA791385	GenomeTrakr wastewater project: Data of Limited Quality	6	6	✓ Released	2021-12-21	2021-12-21
PRJNA774554	GenomeTrakr umbrella project for Yersinia enterocolitica			✓ Released	2021-10-26	2021-10-26
PRJNA767800	GenomeTrakr wastewater project: protocol pilot exercise	82	64	✓ Released	2021-10-01	2021-10-01
PRJNA758008	GenomeTrakr umbrella project			✓ Released	2021-08-26	2021-08-27
PRJNA757447	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition	44	71	✓ Released	2021-08-24	2021-08-24
PRJNA750933	Targeted amplicon deep sequencing of genetic markers for Cyclospora cayetanensis	11	11	✓ Released	2021-07-30	2021-07-30
PRJNA727202	GenomeTrakr project: Maryland Department of Health and Mental Hygiene	SAMN19101980	SRR14530145	✓ Released	2021-05-04	2021-05-04
PRJNA718348	GenomeTrakr Project: U.S. Food and Drug Administration - CFSAN	SAMN20079938	SRR15049125	✓ Released	2021-03-29	2021-03-29

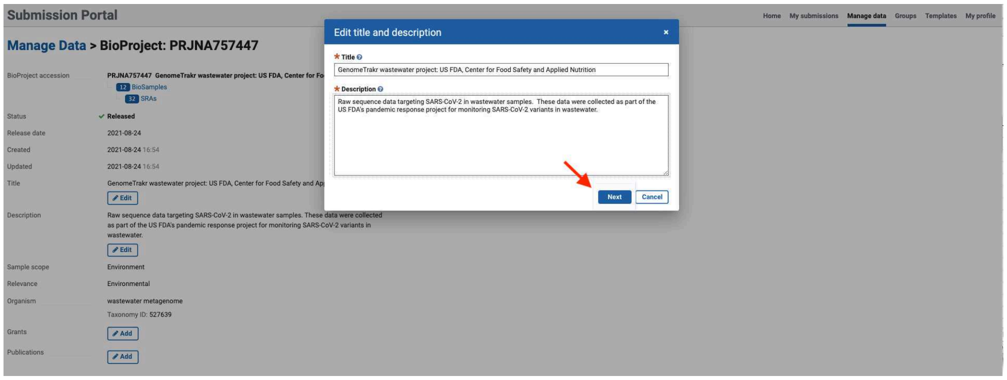
4. Fields available for editing will be displayed after selecting a BioProject.



Manage Data > BioProject: PRJNA757447

BioProject accession	PRJNA757447 GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition 12 BioSamples 32 SRAs
Status	✓ Released
Release date	2021-08-24
Created	2021-08-24 16:54
Updated	2021-08-24 16:54
Title	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition
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.
Sample scope	Environment
Relevance	Environmental
Organism	wastewater metagenome Taxonomy ID: 527639
Grants	
Publications	

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.



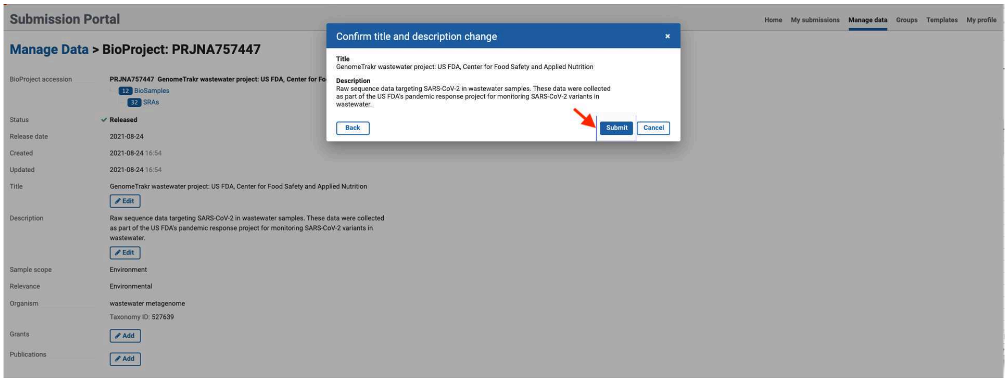
Edit title and description

* Title

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

* 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.



Manage Data > BioProject: PRJNA757447

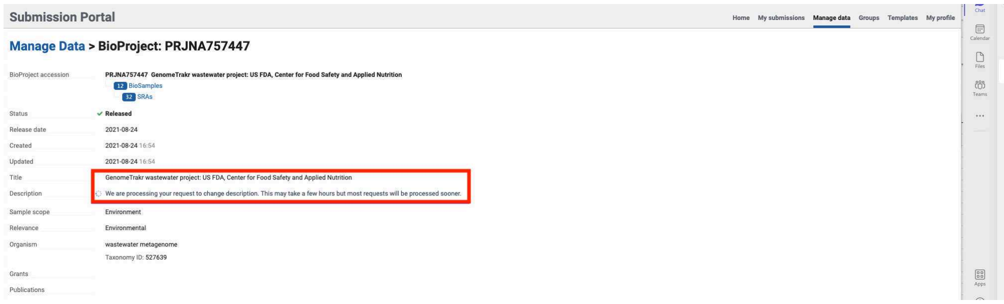
BioProject accession	PRJNA757447 GenomeTrakr wastewater project: US FDA, Center for Fo 12 BioSamples 32 SRAs
Status	✓ Released
Release date	2021-08-24
Created	2021-08-24 16:54
Updated	2021-08-24 16:54
Title	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition
Description	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.
Sample scope	Environment
Relevance	Environmental
Organism	wastewater metagenome Taxonomy ID: 527639
Grants	
Publications	

Confirm title and description change

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

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.

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



Manage Data > BioProject: PRJNA757447

BioProject accession	PRJNA757447 GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition 12 BioSamples 32 SRAs
Status	✓ Released
Release date	2021-08-24
Created	2021-08-24 16:54
Updated	2021-08-24 16:54
Title	GenomeTrakr wastewater project: US FDA, Center for Food Safety and Applied Nutrition
Description	We are processing your request to change description. This may take a few hours but most requests will be processed sooner.
Sample scope	Environment
Relevance	Environmental
Organism	wastewater metagenome Taxonomy ID: 527639
Grants	
Publications	

1.2 **Email for BioProject: bioprojecthelp@ncbi.nlm.nih.gov**

Use this email for the following tasks, include the BioProject accession in the email subject:

- 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

BioSample curation

2 **How to edit BioSamples.**

2.1 All edits or updates to BioSample records are submitted via email to the **BioSample database: biosamplehelp@ncbi.nlm.nih.gov.**

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).
- Taxonomic updates: Include "**pd-help@ncbi.hlm.nih.gov**" on these requests to ensure taxonomic changes get propagated fully across NCBI databases. The organism name should include the 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**

Email: biosamplehelp@ncbi.nlm.nih.gov

Dear BioSampleHelp,

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)**

Email: **biosamplehelp@ncbi.nlm.nih.gov**

Dear BioSampleHelp,

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:

FILE FDA_biosample_update_20220203...

(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.

Dear BioSampleHelp,

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

Thanks,
Ruth

SRA curation

3 **SRA updates and retractions**

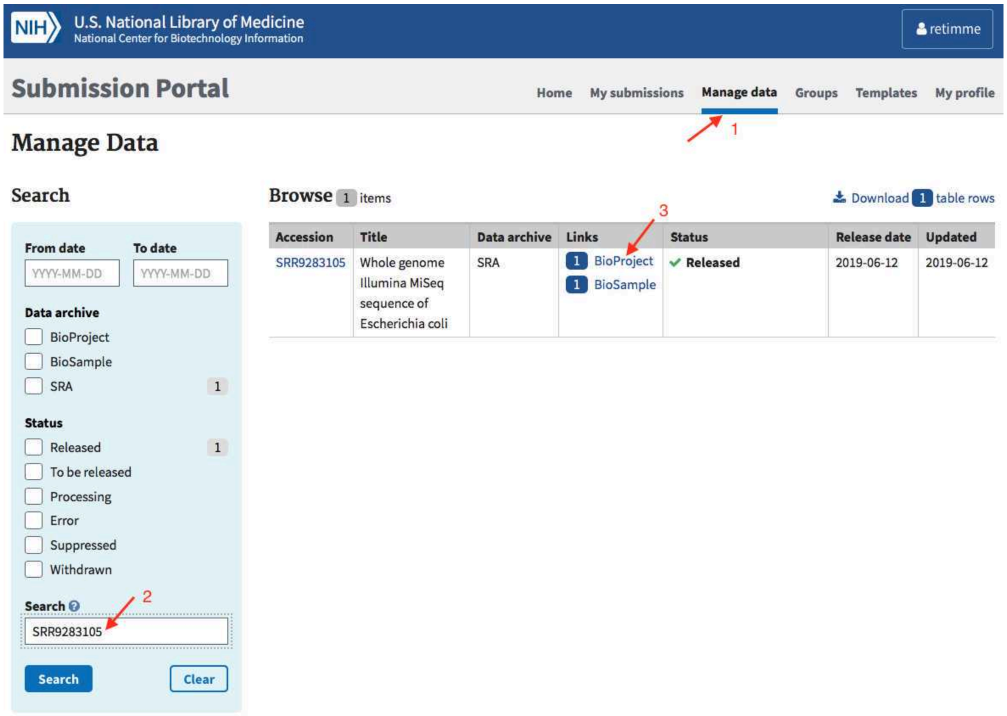
3.1 The following types of updates can be made within the submission portal under the “Manage data” tab:

- 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:

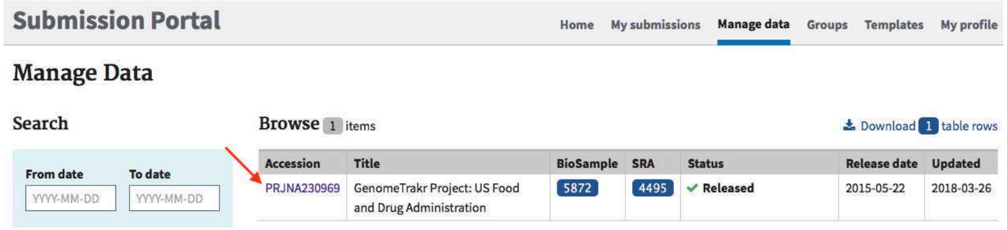
3. Click on the resulting "BioProject" link.



Search: SRR9283105

Accession	Title	Data archive	Links	Status	Release date	Updated
SRR9283105	Whole genome Illumina MiSeq sequence of Escherichia coli	SRA	1 BioProject 1 BioSample	✓ Released	2019-06-12	2019-06-12

4. Click on the BioProject accession link:



Submission Portal

Manage Data

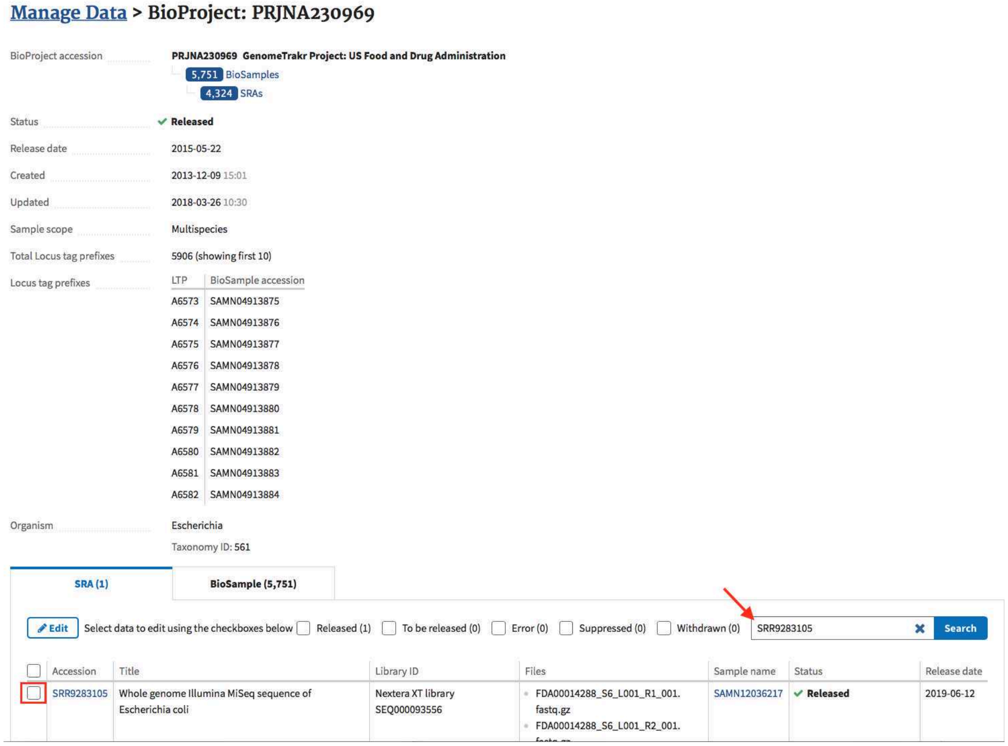
Search

From date YYYY-MM-DD To date YYYY-MM-DD

Browse 1 items — Download 1 table rows

Accession	Title	BioSample	SRA	Status	Release date	Updated
PRJNA230969	GenomeTrakr Project: US Food and Drug Administration	5872	4495	✓ Released	2015-05-22	2018-03-26

5. All the SRA records submitted to this BioProject can now be edited! Search for the one(s) you want and click the box to edit.



Manage Data > BioProject: PRJNA230969

BioProject accession	PRJNA230969 GenomeTrakr Project: US Food and Drug Administration 5,751 BioSamples 4,324 SRAs
Status	✓ Released
Release date	2015-05-22
Created	2013-12-09 15:01
Updated	2018-03-26 10:30
Sample scope	Multispecies
Total Locus tag prefixes	5906 (showing first 10)
Organism	Escherichia Taxonomy ID: 561

Locus tag prefixes

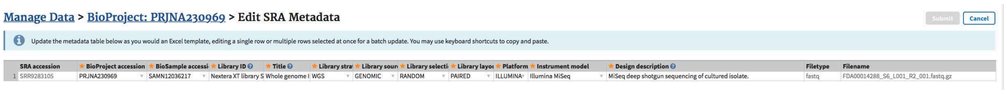
LTP	BioSample accession
A6573	SAMN04913875
A6574	SAMN04913876
A6575	SAMN04913877
A6576	SAMN04913878
A6577	SAMN04913879
A6578	SAMN04913880
A6579	SAMN04913881
A6580	SAMN04913882
A6581	SAMN04913883
A6582	SAMN04913884

SRA (1)
BioSample (5,751)
Edit
Select data to edit using the checkboxes below
□ Released (1) □ To be released (0) □ Error (0) □ Suppressed (0) □ Withdrawn (0)
Search

SRR9283105

□	Accession	Title	Library ID	Files	Sample name	Status	Release date
□	SRR9283105	Whole genome Illumina MiSeq sequence of Escherichia coli	Nextera XT library SEQ000093556	FDA00014288_S6_L001_R1_001.fastq.gz FDA00014288_S6_L001_R2_001.fastq.gz	SAMN12036217	✓ Released	2019-06-12

6. You can now edit the metadata directly for this record. If you need to correct a sample-swap you can enter the correct BioSample accession here and the sequence will get re-parented.



Manage Data > BioProject: PRJNA230969 > Edit SRA Metadata

Update the metadata table below as you would an Excel template, editing a single row or multiple rows selected at once for a batch update. You may use keyboard shortcuts to copy and paste.

	SRA accession	BioProject accession	BioSample accessi	Library ID	Title	Library strat	Library sourc	Library selecti	Library layo	Platform	Instrument model	Design description	Filetype	Filename
1	SRR9283105	PRJNA230969	SAMN12036217	Nextera XT library S	Whole genome I	WGS	GENOMIC	RANDOM	PAIRED	ILLUMINA	Illumina MiSeq	MiSeq deep shotgun sequencing of cultured isolate.	fastq	FDA00014288_S6_L001_R2_001.fastq.gz

3.2 Editing/updating custom SRA metadata attributes

Email: sra@ncbi.nlm.nih.gov

> *Dear SRA,*
>
> *Please update the attached SRA records.*
>
> *Thanks,*
> *Ruth*

Attach a tab-delimited text file with the SRR accessions in the first column and attributes to update included as additional columns (*****only include columns you want to update*****).

Examples:

FDA_SRA_update_20210203_ct.txt (adding custom wastewater attributes)

FDA_SRA_update_20210203_fb.txt (updating core SRA metadata attributes)

The following table will update or add the custom attributes used for the covid wastewater project:

	A	B	C	D	E
	Run	enrichment_kit	amplicon_PCR_primer_scheme	library_preparation_kit	dehosting_method
	SRR17540870	NEBNext ARTIC SARS-CoV-2 RT-PCR Module	NEB VarSkip Short	Illumina DNA prep	SRA human read removal tool

Tab-delimited table for updating an SRA record.

3.3 Email contact for SRA database: sra@ncbi.nlm.nih.gov

Use this email for the following tasks. Include your lab and the request date in your subject line for easy tracking, e.g. “FDA SRA retractions, Dec 10, 2019”.

- Questions about validation errors or processing of an SRA submission.
- Retractions

3.4 **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 and will be rejected.

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

*Although the data submissions appear visibly linked at NCBI (you can navigate between databases with links on each record) the data may not be linked in a way that works with retractions. Therefore, if you need to retract a bad SRA run, you should also request that all other data (such as GenBank assemblies or Pathogen Detection analyses) also be retracted, even if you didn't submit them yourself.

Email template:

Dear SRA,

Please retract the following SRR accessions and any linked assemblies or PD analyses due to XXX issue.
We will re-sequence these isolates and re-submit new data.

SRRXXXXXX1
SRRXXXXXX2
SRRXXXXXX3

Thanks,

Ruth