

Mar 23, 2020 Version 1

NCBI data curation protocol V.1

✓ Book Chapter

 In 1 collection

DOI

dx.doi.org/10.17504/protocols.io.bacaiase

Ruth Timme¹, Maria Balkey², Sai Laxmi Gubbala Venkata³, Robyn Randolph⁴, William Wolfgang⁵, Errol Strain⁴

¹US Food and Drug Administration;

²Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA;

³Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA;

⁴Center for Veterinary Medicine, U.S. Food and Drug Administration, College Park, Maryland, USA;

⁵Wadsworth Center NYSDOH

GenomeTrakr

Springer Nature Books



Ruth Timme

US Food and Drug Administration

OPEN  ACCESS



DOI: dx.doi.org/10.17504/protocols.io.bacaiase

Protocol Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain 2020. NCBI data curation protocol. [protocols.io https://dx.doi.org/10.17504/protocols.io.bacaiase](https://dx.doi.org/10.17504/protocols.io.bacaiase)

Manuscript citation:

Timme, RE, Wolfgang, WJ, Balkey, M, Venkata, SLG, Randolph, R, Allard, M, Strain, E. Optimizing open data to support OneHealth: Best practices to ensure interoperability of genomic data from microbial pathogens. In prep.

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

This protocol has is one of four that are currently being tested with the GenomeTrakr direct submission pilot. Please comment if you find errors, steps that need clarification, or curation areas we might have missed.



Created: December 10, 2019

Last Modified: November 10, 2021

Protocol Integer ID: 30818

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

Disclaimer

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

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.

The submission staff at each respective NCBI database handle incoming submissions and curation updates. These are the people whom submitters interface with for routine submissions, data retractions, and updates to records.

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

- BioProject
- BioSample
- Sequence Read Archive
- Pathogen Detection

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.

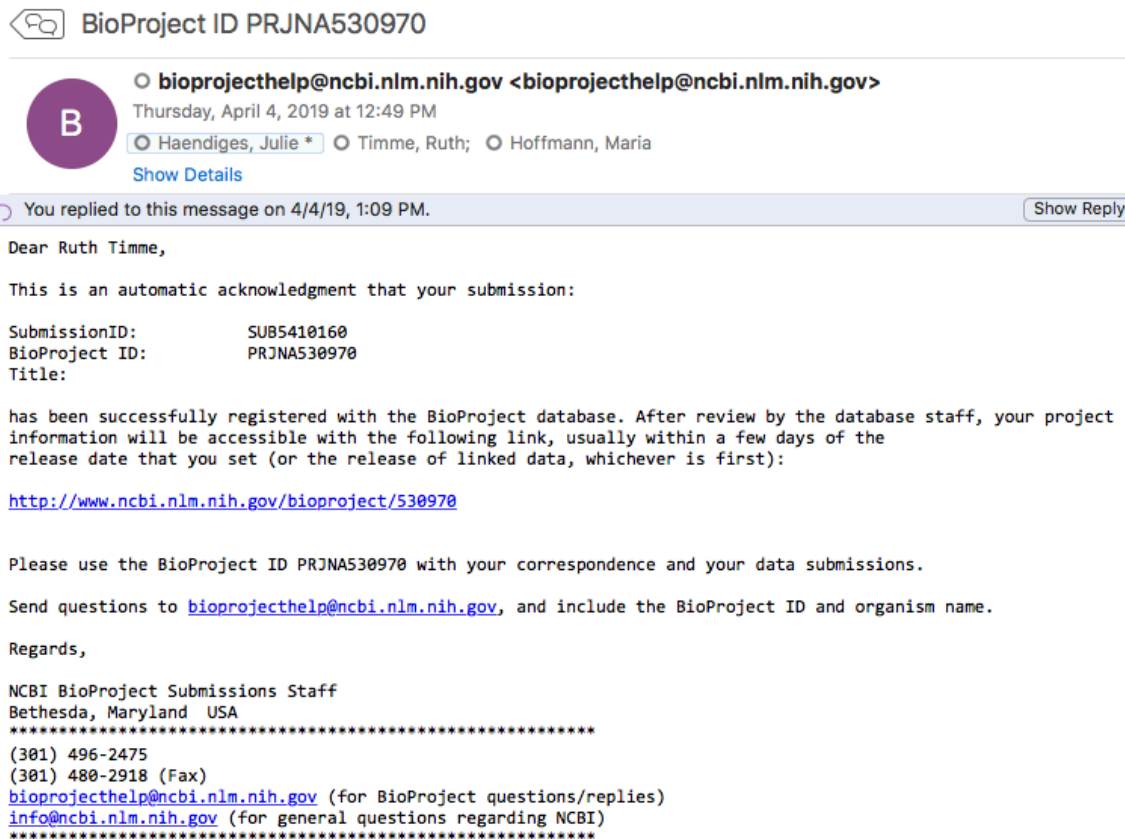


BioProject Curation

1 The BioProject protocol details how to check if your BioProjects were submitted correctly and how to track and update them once they are live.

1.1 Look for an email subject in the following format to retrieve your accession number:

"BioProject ID PRJNA#####."



1.2 Query the BioProject database to ensure your BioProjects are live and linked properly with their umbrella projects (if relevant): <https://www.ncbi.nlm.nih.gov/bioproject>.

Search using **free text** that you know appears in the description section of your BioProject, or using the **accession** returned to you via email or submission portal (e.g. PRJNA530970).



Here's an example of all GenomeTrakr bioprojects created for the California Department of Public Health. Each of these are data BioProjects linked to their respective species-specific GenomeTrakr Umbrellas.

The screenshot shows the NCBI BioProject search interface. The search term 'GenomeTrakr California' is entered in the search bar. The results are displayed in a list format, showing 4 items. The first item is 'Listeria monocytogenes', which is a GenomeTrakr Project: California Department of Health - FDLB Micro. The second item is 'Campylobacter jejuni', which is a GenomeTrakr Project: California Department of Health - FDLB Micro. The third item is 'Escherichia coli', which is a GenomeTrakr Project: California Department of Health - FDLB Micro. The fourth item is 'Salmonella enterica', which is a GenomeTrakr Project: California Department of Public Health - FDLB Micro. The search results are sorted by Default order.

NCBI Resources How To

BioProject BioProject GenomeTrakr California

Create alert Advanced Browse by Project attributes

Project Types
Primary submission (4)

Project Data
Nucleotide (3)
Protein (2)
Assembly (3)
SRA (4)

Scope
Multi-isolate (3)
Multi-species (1)

Organism Groups
Bacteria (4)

Clear all
Show additional filters

Display Settings: Summary, Sorted by Default order Send to:

Search results
Items: 4

☐ [Listeria monocytogenes](#)

1. **GenomeTrakr Project: California Department of Health - FDLB Micro**
Taxonomy: [Listeria monocytogenes](#)
Project data type: Genome sequencing and assembly
Scope: Multiisolate
FDA Center for Food Safety and Applied Nutrition
Accession: PRJNA514281 ID: 514281

☐ [Campylobacter jejuni](#)

2. **GenomeTrakr Project: California Department of Health - FDLB Micro**
Taxonomy: [Campylobacter jejuni](#)
Project data type: Genome sequencing and assembly
Scope: Multispecies
US Food and Drug Administration
Accession: PRJNA309864 ID: 309864

☐ [Escherichia coli](#)

3. **GenomeTrakr Project: California Department of Health - FDLB Micro**
Taxonomy: [Escherichia coli](#)
Project data type: Genome sequencing and assembly
Scope: Multiisolate
US Food and Drug Administration
Accession: PRJNA277984 ID: 277984

☐ [Salmonella enterica](#)

4. **GenomeTrakr Project: California Department of Public Health - FDLB Micro**
Taxonomy: [Salmonella enterica](#)
Project data type: Genome sequencing and assembly
Scope: Multiisolate
US Food and Drug Administration
Accession: PRJNA277983 ID: 277983

- 1.3 Each of the California data BioProjects listed above are linked to their respective species-specific GenomeTrakr Umbrellas.

For example, here is the *Listeria monocytogenes* data BioProject listed above, showing the linkage to the GenomeTrakr umbrella bioproject.

<https://www.ncbi.nlm.nih.gov/bioproject/514281>

Listeria monocytogenes ← Organism

Accession: PRJNA514281 ID: 514281

GenomeTrakr Project: California Department of Health - FDLB Micro ← Title

Whole genome sequencing of cultured *Listeria monocytogenes* as part of the US Food and Drug Administration's WGS surveillance effort for the rapid traceback of foodborne pathogens. ← Description

Accession	PRJNA514281
Data Type	Genome sequencing and assembly
Scope	Multiisolate
Organism	Listeria monocytogenes [Taxonomy ID: 1639] Bacteria; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria; Listeria monocytogenes
Submission	Registration date: 9-May-2019 FDA Center for Food Safety and Applied Nutrition
Related Resources	<ul style="list-style-type: none"> California Department of Health - FDLB Micro
Relevance	Medical

Link to the Lm
GenomeTrakr
umbrella

See [Genome Information for Listeria monocytogenes](#)

NAVIGATE UP

This project is a component of the *Listeria monocytogenes*: GenomeTrakr - US Food and Drug Administration, Center for Food Safety and Applied Nutrition

NAVIGATE ACROSS

35 additional projects are components of the *Listeria monocytogenes*: GenomeTrakr - US Food and Drug Administration, Center for Food Safety and Applied Nutrition.

523 additional projects are related by organism.

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (WGS master)	32
SRA Experiments	33
OTHER DATASETS	
BioSample	33
Assembly	32

Assembly details: Download					
Assembly level			Number of Assemblies		
Contig			32		
Total			32		
Assembly	Level	WGS	BioSample	Strain	Taxonomy
GCA_005246145.1		AACKAT000000000	SAMN11368564	CDPHFDLB-FM1...	Listeria monocytogenes
GCA_005246025.1		AACKAY000000000	SAMN11360323	CDPHFDLB-FM1...	Listeria monocytogenes
GCA_005245985.1		AACKBA000000000	SAMN11368557	CDPHFDLB-FM1...	Listeria monocytogenes
GCA_005245485.1		AACKAO000000000	SAMN11368560	CDPHFDLB-FM1...	Listeria monocytogenes
GCA_005245465.1		AACKAN000000000	SAMN11368558	CDPHFDLB-FM1...	Listeria monocytogenes

- 1.4 If you can't find your BioProjects they might not be live yet or they might have been submitted with a "hold until published (HUP)" date.

Check your "My Submissions" tab for potential processing errors using the submission ID (e.g. SUB5410160) returned in the email correspondence from NCBI (see step 1.1)

<https://submit.ncbi.nlm.nih.gov/subs/>

Click on the correct submission returned from this query to check the processing status for this BioProject.



Start a new submission

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

Filter / Search

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

Apps + Data archives +

Query: SUB5410160 Search Clear

1 submission

Submission	Title	App	Group	Status	Updated
SUB5410160	FDA-CFSAN's MetagenomeTrakr Pilot Project	BioProject	fda	✓ BioProject: Processed PRJNA530970 : FDA-CFSAN's MetagenomeTrakr Pilot Project (TaxID: 1699855)	Apr 04

1.5 Email contact for BioProject: bioprojecthelp@ncbi.nlm.nih.gov

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

- Questions about errors or processing of a BioProject submission.
- Update the Title, Organism, Description, URL, or publications on this BioProject
- Convert to an Umbrella BioProject
- Add a linkage or re-assign linkage to an existing Umbrella BioProject

BioSample curation

2 The BioSample protocol details how to check if your metadata was submitted correctly and how to track, update, or retract them once your submissions are live.

2.1 You can find your BioSample accessions in two places.

1. Email with following subject line: "BioSample accession SAMN#####". There will also be a text file attached with a tab-delimited table listing the Accessions generated during the submission, along with strain ID and organism info. This table can be easily imported into your local database.



Dear fda service,

This is an automatic acknowledgment that your recent submission to the BioSample database has been successfully processed and will be released on the date specified.

BioSample accession: SAMN13541114
Temporary SubmissionID: SUB6672905
Release date: 2019-12-11-05:00, or with the release of linked data, whichever is first

A submission summary and the links by which your BioSample records will be accessible are appended and attached.

Please reference BioSample accession SAMN13541114 when making corresponding sequence data submissions.

Send questions and update requests to biosamplehelp@ncbi.nlm.nih.gov; include the BioSample accession SAMN13541114 in any correspondence.

Regards,

NCBI BioSample Submissions Staff
Bethesda, Maryland USA

(301) 496-2475
(301) 480-2918 (Fax)
biosamplehelp@ncbi.nlm.nih.gov (for BioSample questions/replies)
info@ncbi.nlm.nih.gov (for general questions regarding NCBI)

2. Query your submissionID in "My Submissions":

<https://submit.ncbi.nlm.nih.gov/subs>

The screenshot shows the NCBI Submission Portal interface. At the top, there's a navigation bar with 'Home', 'My submissions', 'Manage data', 'Groups', 'Templates', and 'My profile'. The 'My submissions' tab is active. Below the navigation bar, there's a 'Your submissions' section. On the left, there's a 'Start a new submission' button and a list of links: GenBank, BioProject, Sequence Read Archive, BioSample, Genome, Supplementary Files, TSA, and API. On the right, there's a 'Filter / Search' section. It includes fields for 'From date', 'To date', 'Status', and 'Sort by'. Below these, there are buttons for 'Apps' and 'Data archives'. A red arrow points to the 'Query' input field, which contains the submission ID 'SUB6672905'. To the right of the input field are 'Search' and 'Clear' buttons. Below the search section, there's a table with the following data:

Submission	Title	App	Group	Status	Updated
SUB6672905	UI-less submission 2019-12-11	API	fda	✓ BioSample: Processed Successfully loaded SAMN13541114 (TaxID: 1396) ✓ SRA: Processed (2 objects) • SRX7344824 • SRR10665887	12:55

2.2 Query the BioSample database to ensure your BioSamples are live and linked properly under their respective BioProjects, e.g. SAMN12987335.

<https://www.ncbi.nlm.nih.gov/biosample>

The BioProject ID is hyperlinked at the bottom of the record. If data has been submitted to SRA under this BioSample, a hyperlinked "SRA" will also appear here, as will assemblies submitted to GenBank (listed as "nucleotide").

**Pathogen: environmental/food/other sample from *Listeria monocytogenes***

Identifiers	BioSample: SAMN12987335; SRA: SRS5486650; CFSAN: CFSAN100155		← Sample ID																																				
Organism	<u>Listeria monocytogenes</u> cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Listeriaceae; Listeria																																						
Package	<u>Pathogen: environmental/food/other; version 1.0</u>																																						
Attributes	<table><tr><td>strain</td><td>CDPHFDLB-FM19-02200-A</td><td>← Authoritative ID for NCBI-PD</td></tr><tr><td>collection date</td><td>2019-09-12</td><td></td></tr><tr><td>geographic location</td><td><u>USA:CA</u></td><td></td></tr><tr><td>isolate name alias</td><td>CFSAN100155</td><td></td></tr><tr><td>collected by</td><td>CDPH FDLB</td><td></td></tr><tr><td>latitude and longitude</td><td>missing</td><td></td></tr><tr><td>isolation source</td><td>environmental swab</td><td></td></tr><tr><td>PublicAccession</td><td>CFSAN100155</td><td></td></tr><tr><td>attribute_package</td><td>environmental/food/other</td><td></td></tr><tr><td>Genus</td><td>Listeria</td><td></td></tr><tr><td>Species</td><td>monocytogenes</td><td></td></tr><tr><td>ProjectAccession</td><td>PRJNA514281</td><td></td></tr></table>			strain	CDPHFDLB-FM19-02200-A	← Authoritative ID for NCBI-PD	collection date	2019-09-12		geographic location	<u>USA:CA</u>		isolate name alias	CFSAN100155		collected by	CDPH FDLB		latitude and longitude	missing		isolation source	environmental swab		PublicAccession	CFSAN100155		attribute_package	environmental/food/other		Genus	Listeria		Species	monocytogenes		ProjectAccession	PRJNA514281	
strain	CDPHFDLB-FM19-02200-A	← Authoritative ID for NCBI-PD																																					
collection date	2019-09-12																																						
geographic location	<u>USA:CA</u>																																						
isolate name alias	CFSAN100155																																						
collected by	CDPH FDLB																																						
latitude and longitude	missing																																						
isolation source	environmental swab																																						
PublicAccession	CFSAN100155																																						
attribute_package	environmental/food/other																																						
Genus	Listeria																																						
Species	monocytogenes																																						
ProjectAccession	PRJNA514281																																						
Links																																							
BioProject	<u>PRJNA514281</u> Listeria monocytogenes Retrieve <u>all samples</u> from this project																																						
Submission	<u>CFSAN</u> ; 2019-10-07																																						
Accession: SAMN12987335 ID: 12987335																																							
<u>BioProject</u> <u>SRA</u> <u>Nucleotide</u> ← Links to BioProject, raw data, and assemblies																																							

Mandatory metadata fields are highlighted in red.

2.3 Email contact for 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 to a BioSample(s)
- Retraction



- Add a linkage or re-assign linkage to an existing Umbrella BioProject

Corrections, updates, and retractions are all performed through email. The content, or body of the email, should contain the specific request.

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 multiple requests in one day.

2.4 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

2.5 How to update content in metadata fields or add new fields 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. Ensure the exact same header names are used here as were included in the original BioSample submission, e.g. strain, organism, collected_by, isolation_source, collection_date, geo_loc_name, etc.



- The following table will correct 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.

2.6 Re-assign a BioSample from one BioProject to another

Submit an update request (see 2.5) 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 The SRA protocols details how to check if your raw reads were submitted correctly and how to update or retract them once they are live.

- 3.1 Search the SRA database for the strain ID, BioSample accession, or SRR accession to pull up the submission record (*see NCBI Submission Protocol, Step 4.9 for obtaining SRA accessions*):

Navigate to the SRA homepage: <https://www.ncbi.nlm.nih.gov/sra>

Query using a run accession (e.g. SRR9283105), strain name, or BioSample accession:

**SRX6052810: Whole genome Illumina MiSeq sequence of Escherichia coli**

1 ILLUMINA (Illumina MiSeq) run: 705,622 spots, 328.5M bases, 171Mb downloads

External Id: EXT00360584**Design:** MiSeq deep shotgun sequencing of cultured isolate.**Submitted by:** FDA Center for Food Safety and Applied Nutrition (CFSAN)**Study:** GenomeTrakr Project: US Food and Drug Administration[PRJNA230969](#) • [SRP058582](#) • [All experiments](#) • [All runs](#)[show Abstract](#)**Sample:**[SAMN12036217](#) • [SRS4953076](#) • [All experiments](#) • [All runs](#)*Organism:* [Escherichia coli](#)**Library:***Name:* Nextera XT library SEQ000093556*Instrument:* Illumina MiSeq*Strategy:* WGS*Source:* GENOMIC*Selection:* RANDOM*Layout:* PAIRED**Spot descriptor:****Runs:** 1 run, 705,622 spots, 328.5M bases, [171Mb](#)

Run	# of Spots	# of Bases	Size	Published
SRR9283105	705,622	328.5M	171Mb	2019-06-12

ID: 8095985

"run" accession

Metadata from the sequence run, including the sequencing platform and library prep kit, are included on an SRA record, along with summary stats of the sequencing data. In addition, the linked BioSample and BioProject are also listed under Sample and Study, respectively.

3.2 Email contact for BioSample 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

Updates to SRA records can be performed within the "Manage Data" web portal (see 3.4)

3.3 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 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

3.4 SRA record update



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.

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

Submission Portal

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

Manage Data

Search

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

Data archive

☐ BioProject
☐ BioSample
☐ SRA 1

Status

☐ Released 1
☐ To be released
☐ Processing
☐ Error
☐ Suppressed
☐ Withdrawn

Search ?
SRR9283105 2
Search Clear

Browse 1 items

Download 1 table rows

Accession	Title	Data archive	Links	Status	Release date	Updated
SRR9283105	Whole genome Illumina MiSeq sequence of Escherichia coli	SRA	1 BioProject 1 BioSample 3	✓ Released	2019-06-12	2019-06-12

4. Click on the BioProject accession link:

**Submission Portal**[Home](#) [My submissions](#) **[Manage data](#)** [Groups](#) [Templates](#) [My profile](#)**Manage Data**

Search

Browse **1** items[Download](#) **1** table rows

From date	To date	Accession	Title	BioSample	SRA	Status	Release date	Updated
YYYY-MM-DD	YYYY-MM-DD	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: PRJNA230969BioProject 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)

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

Organism Escherichia

Taxonomy ID: 561

SRA (1)		BioSample (5,751)					
Edit		Select data to edit using the checkboxes below <input type="checkbox"/> Released (1) <input type="checkbox"/> To be released (0) <input type="checkbox"/> Error (0) <input type="checkbox"/> Suppressed (0) <input type="checkbox"/> Withdrawn (0)					
<input type="checkbox"/>	Accession	Title	Library ID	Files	Sample name	Status	Release date
<input type="checkbox"/>	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 access	Library ID	Title	Library str	Library ssun	Library select	Library layer	Platform	Instrument model	Design description	Filetype	Filename
SRR9283105	PRJNA230969	SAMN12036217	Neutera AT library 5 Whole genome	WGS	GENOMIC	RANDOM	PAIRED	ILLUMINA	Illumina MiSeq	MISeq deep shotgun sequencing of cultured isolate.		fastq	FDA00014288_S6_L001_R2_001.fastq.gz

Pathogen Detection

4

The Pathogen Detection curation protocol includes instructions for finding your data within the surveillance platform and identifying quality control issues that might have prevented your data from being processed.

Important!!!: The NCBI-PD staff only need to be contacted once in the beginning to flag the BioProject accession for inclusion to the Pathogen Detection system. They can also field questions about the Pathogen Detection browser, interface, or analyses. However, The NCBI-PD staff *cannot* help resolve data updates, retractions, or submission problems for the other NCBI databases.

4.1 Navigate to the NCBI Pathogen Detection browser:

<https://www.ncbi.nlm.nih.gov/pathogens>

Search for your data by clicking on the “find isolates now” link, using the strain name, BioSample, SRR accession, or any other term present in the metadata. For example, to locate all isolates included in a recent run, paste the list of IDs from a spreadsheet or Word document into the general search field:



Pathogen Detection BETA



View the recent webinar: '[Introducing the Pathogen Detection Isolates Browser](#)'.

Watch this webinar!

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

[Find isolates now!](#) Search across all species

Examples:

1. Search for isolates encoding a mobile colistin resistance gene and a KPC beta-lactamase search: [AMR_genotypes:mcr* AND AMR_genotypes:blaKPC*](#)
2. Search for Salmonella isolates from the USA search: [geo_loc_name:USA AND taxgroup_name:"Salmonella enterica"](#)

Explore the Data Search within a species

Species	New Isolates	Total Isolates
Salmonella enterica	21	244,339
E.coli and Shigella	3	93,046
Campylobacter jejuni	138	46,727
Listeria monocytogenes	19	30,212
See more organisms...		

Learn More

[About](#)

[FAQ](#)

[Browser Factsheet](#)

[Antimicrobial Resistance Factsheet](#)

[Antimicrobial Resistance](#)

[Contributors](#)

[Help](#)

Data Resources

[Isolates Browser](#)

[Pathogen Detection Reference Gene Catalog](#)

[Isolates with antibiotic resistant phenotypes](#)

[Download analysis results \(FTP\)](#)

Submit

[How to submit data](#)

[How to submit antibiotic resistance phenotypes](#)

[How to submit beta-lactamases](#)

[NCBI Submission Portal](#)

4.2 Search results

Results will usually include two tables.

- 1) A "matched cluster" table if the matched isolates appear in an existing cluster, e.g. SNP cluster PDS000038362
- 2) A "matched isolate" table listing all the isolates that contain the search term in their metadata, e.g. strain name CFSAN086778

CFSAN086778

Listeria monocytogenes

Matched clusters

#	Organism groups	SNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	Minimal min-diff	Latest update
1	Listeria monocytogenes	PDS000038362.2	1	1	0	3	n/a	2019-08-27

Matched isolates

#	Organism Group	SRA Center	Strain	Serovar	Isolate	Create Date	Location	Isolation Source	Isolation type	Host	SNP cluster	Min-sa	Min-diff	BioSample	Assembly	AMR genotypes	K-mer group
1	Listeria monocytogenes	CFSAN	CFSAN086778		PDS000038362.2	2019-08-20	USA:MA	Pleural effusion (fluid)	clinical	Homo sapiens	PDS000038362.2	0	n/a	SAMN1261032	tsuX lin		PDS000000001.1398

4.3 Exceptions table

Isolates that do not pass the NCBI-PD quality control check will *not* be added to the NCBI-PD database.

Instead, these isolates will be listed in a third table listing isolates which fail NCBI's validation check, along with the reason(s) for the failure. *Note that the data will still be in the SRA.

For example, a query on the following 15 SRR IDs (SRR9853527 SRR9853553 SRR9853556 SRR9853555 SRR9853522 SRR9853523 SRR9854074 SRR9853879 SRR9853875 SRR9854096 SRR9854066 SRR9854069 SRR9854080 SRR9951128 SRR9951847) reveals that eight passed and 7 got flagged for QC issues, listed in the "Isolate Exceptions" table:

SRR9853527 SRR9853553 SRR9853556 SRR9853555 SRR9853522 SRR9853523 SRR9854074 SRR9853879 SRR9853875 SRR9853876

Isolate Exceptions

#	exception type	exception	consequence	lower limit	upper limit	actual value	BioSample	run(s)	Isolate	Assembly
1	Assembly validation failure	Too many assembly contigs	Not published		500	524	SAMN02844549	SRR9853527	PDT000550696.1	Salm subtrero
2	Assembly validation failure	Too many assembly contigs	Not published		500	809	SAMN02844737	SRR9854080	PDT000550752.1	Salm subtr
3	Assembly validation failure	Too many assembly contigs	Not published		500	579	SAMN02843875	SRR9853523	PDT000550692.1	Salm subtr

Select an organism group

Matched clusters

#	Organism groups	SNP cluster	Matched isolates	Matched clinical isolates	Matched environmental isolates	Total isolates	Minimal min-diff	Latest update
1	Salmonella enterica	PDS000027237.73	1	1	0	626	2	2019-09-04
2	Salmonella enterica	PDS000048537.1	1	0	1	2	n/a	2019-08-03
3	Listeria monocytogenes	PDS000049274.1	1	0	1	3	n/a	2019-08-14

Matched isolates

#	Organism Group	SRA Center	Strain	Serovar	Isolate	Create Date	Location	Isolation Source	Isolation type	Host	SNP cluster	Min
1	Salmonella enterica	CFSAN	CFSAN00024	Montevideo	PDT000563653.1	2019-08-14			clinical		PDS000027237.73	
2	Listeria monocytogenes	CFSAN	FDA0000546		PDT000562550.1	2019-08-12	USA: NY	swab	environmental/other		PDS000049274.1	
3	Salmonella enterica	CFSAN	FDA0000143	enterica / Typhimurium	PDT000144236.3	2019-07-29	Pakistan	gasuri methis (dried fenugreek leaves)	environmental/other			
4	Salmonella enterica	CFSAN	FDA0000129	enterica / Brunel	PDT000110757.3	2019-07-29	Viet Nam	frz barramundi fillet	environmental/other			
5	Salmonella enterica	CFSAN	FDA0000046	enterica / Carrau	PDT000489458.2	2019-07-29	Honduras	frozen shrimp	environmental/other			
6	Salmonella enterica	CFSAN	FDA0000095	houstenae / 43:z4,z23:- (Houten)	PDT000048623.4	2019-07-29	Thailand	soft shell crabs	environmental/other			
7	Salmonella enterica	CFSAN	FDA0000129	enterica / Mgulani	PDT000110759.3	2019-07-29	India	mutta masala	environmental/other			
8	Salmonella enterica	CFSAN	FDA0000123	enterica / Agona	PDT000127763.3	2019-07-29	South Korea	seasoned file fish	environmental/other		PDS000048537.1	

Depending on what QC issue is flagged, re-isolation or re-sequencing might be required. If the sequencing data is determined to be poor quality, then follow the SRA retraction guidelines and re-submit following the SRA submission instructions listed previously.

The columns in the exception table are described here:

Column headers	Description of field
Exception type	Readset validation failure – The SRA run was not valid and could not be used. Assembly validation failure – The pathogen assembly was not valid and could not be used. wgMLST validation failure – The assembly (pathogen or GenBank) could not be used for wgMLST analysis.
Exception	Short message indicating the reason for failing

	validation.
Consequence	Not published – The isolate will not appear in any published organism group (PDG). Not clustered – The isolate will appear in a published organism group (PDG) but will be presented as a singleton (ie no clustering attempted).
Lower limit	Lower limit of the valid range (as relevant).
Upper limit	Upper limit of the valid range (as relevant).
Actual value	Actual value recorded by the system.
Biosample_acc	INSDC accession of the isolate's biosample record.
Run(s)	INSDC accession(s) of the isolate's SRA run record(s).
pathogen target	Pathogen target accession (PDT) for this isolate.
Organism	NCBI taxonomy (scientific_name) of the isolate.
Run center	Submitting organization name (e.g. FDA-CFSAN)

Description of NCBI's exception file. This information was pulled from the README.txt file on August 14th, 2019 located under the following path
[:ftp.ncbi.nlm.nih.gov/pathogen/README.txt](ftp://ftp.ncbi.nlm.nih.gov/pathogen/README.txt).

4.4 Exceptions File:

All QC failures are also aggregated in an exceptions file posted at NCBI's FTP site under the following generic path:

ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/<pathogenName>/PDG00000000XX.XXXX/Exceptions/PDG00000000XX.XXXX.reference_target.exceptions.tsv

For example:

<ftp://ftp.ncbi.nlm.nih.gov/pathogen/Results/Salmonella/PDG000000002.1216/Exceptions/PDG000000002.1216.exceptions.tsv>.

Depending on what flagged QC issue is, re-isolation or re-sequencing *may* be required. If the sequencing data is determined to be poor quality, then follow the SRA retraction guidelines and re-submit following the SRA submission instructions listed previously. The

exceptions file can be sorted by sra center (name of submitting group) enabling a lab to easily identify all of their flagged isolates within each species database.

Note: QC failure within the NCBI-PD may not mean failure for other purposes (i.e BioNumerics analysis and submission at CDC). Look at each failure/exception carefully to determine the appropriate next step.

Note:For organism groups still using legacy kmer clustering, the Exceptions file is far more limited in scope and will found in the ./Clusters directory.

4.5 **Email contact for Pathogen Detection database: pd-help@ncbi.nlm.nih.gov**

Use this email for the following tasks.

- Link a new data or umbrella BioProject to NCBI Pathogen Detection
- General questions or feature requests

The NCBI-PD staff only need to be contacted once in the beginning to flag the BioProject accession for inclusion to the Pathogen Detection system. They can also field questions about the Pathogen Detection browser, interface, or analyses. ***However, The NCBI-PD staff cannot help resolve data updates, retractions, or submission problems. Please follow database-specific instructions for these curation tasks.***