Mar 23, 2020

NCBI data curation protocol

Ruth Timme[1], Maria Balkey[2], Sai Laxmi Gubbala Venkata[3], Robyn Randolph[4], William Wolfgang[3], Errol Strain[4]

[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]Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA, [4]Center for Veterinary Medicine, U.S. Food and Drug Administration, College Park, Maryland, USA

1 Works for me dx.doi.org/10.17504/protocols.io.bacaiase

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

Ruth Timme
US Food and Drug Administration

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

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

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.

BEFORE STARTING

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.

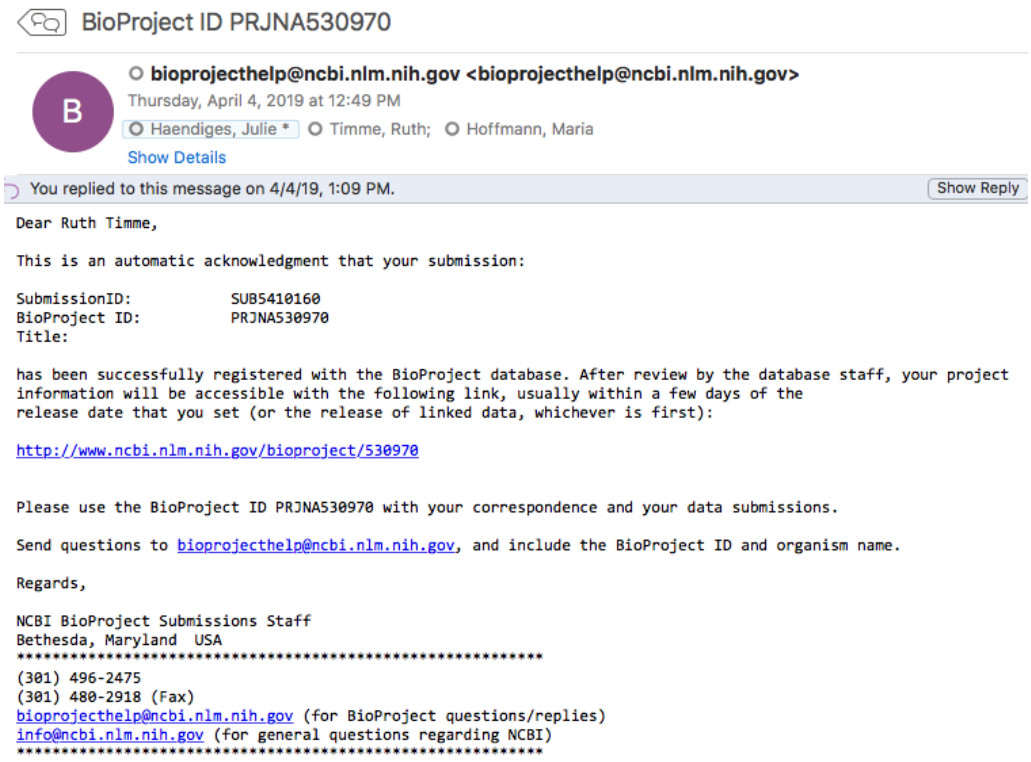
BioProject Curation

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.**

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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

"BioProject ID PRJNA######."



BioProject ID PRJNA530970

B

bioprojecthelp@ncbi.nlm.nih.gov <bioprojecthelp@ncbi.nlm.nih.gov>

Thursday, April 4, 2019 at 12:49 PM

Haendiges, Julie * Timme, Ruth; Hoffmann, Maria

Show Details

You replied to this message on 4/4/19, 1:09 PM. Show Reply

Dear Ruth Timme,

This is an automatic acknowledgment that your submission:

SubmissionID: SUB5410160
BioProject ID: PRJNA530970
Title:

has been successfully registered with the BioProject database. After review by the database staff, your project information will be accessible with the following link, usually within a few days of the release date that you set (or the release of linked data, whichever is first):

http://www.ncbi.nlm.nih.gov/bioproject/530970

Please use the BioProject ID PRJNA530970 with your correspondence and your data submissions.

Send questions to bioprojecthelp@ncbi.nlm.nih.gov, and include the BioProject ID and organism name.

Regards,

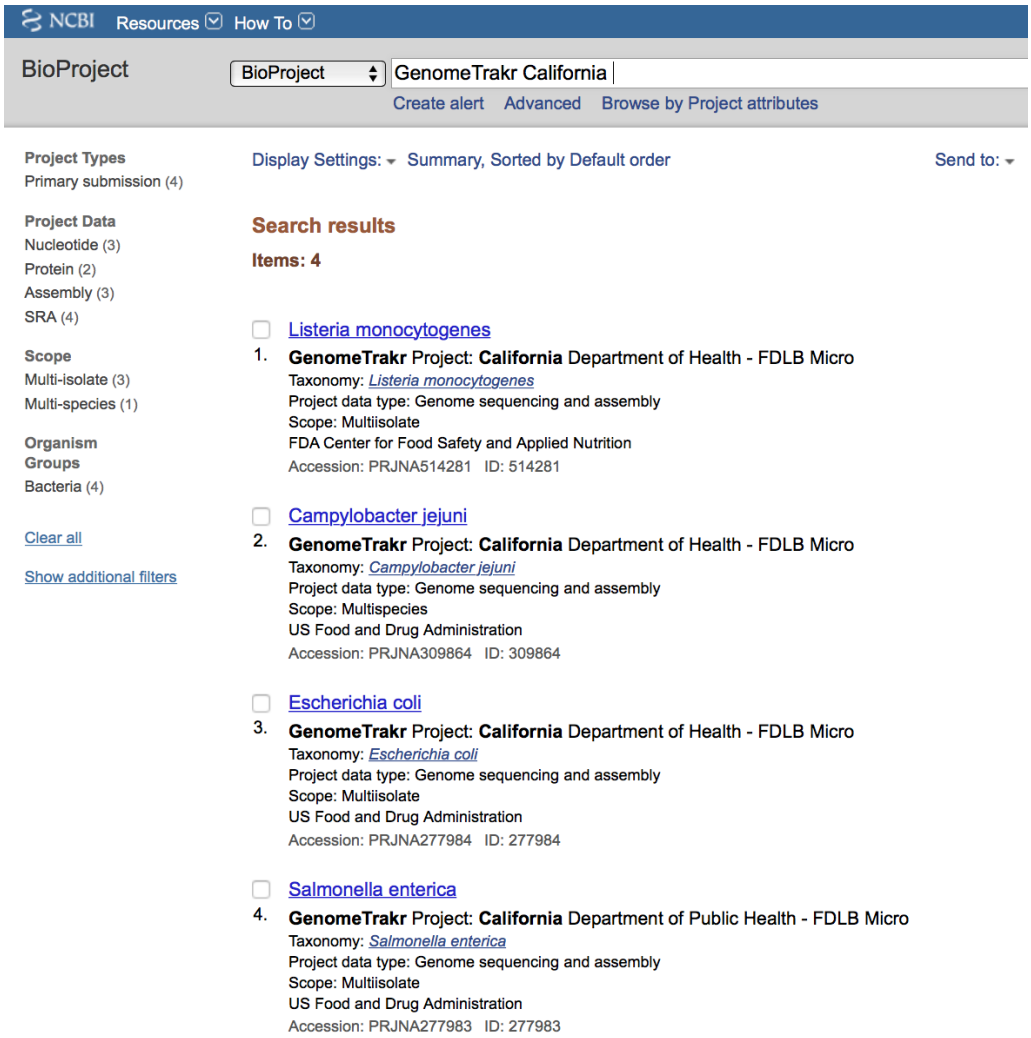
NCBI BioProject Submissions Staff
Bethesda, Maryland USA
****
(301) 496-2475
(301) 480-2918 (Fax)
bioprojecthelp@ncbi.nlm.nih.gov (for BioProject questions/replies)
info@ncbi.nlm.nih.gov (for general questions regarding NCBI)
****

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.

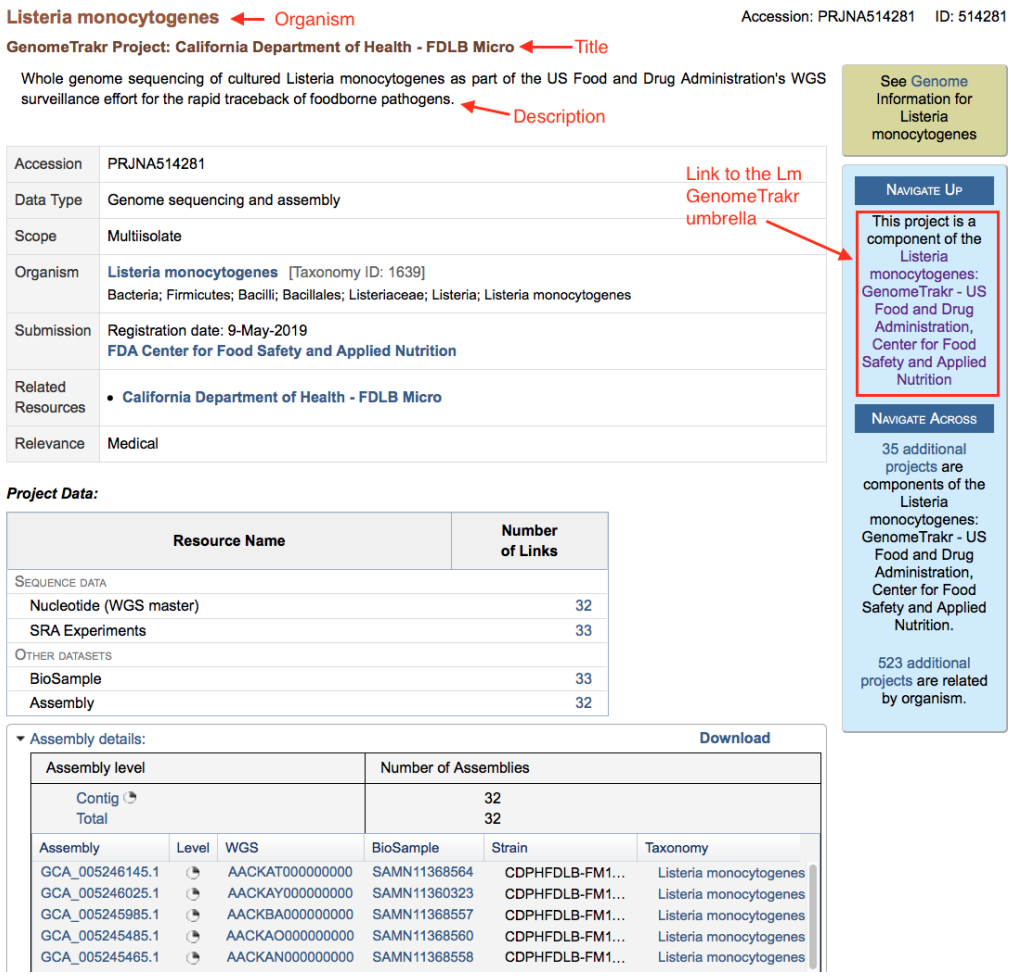
Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase



Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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



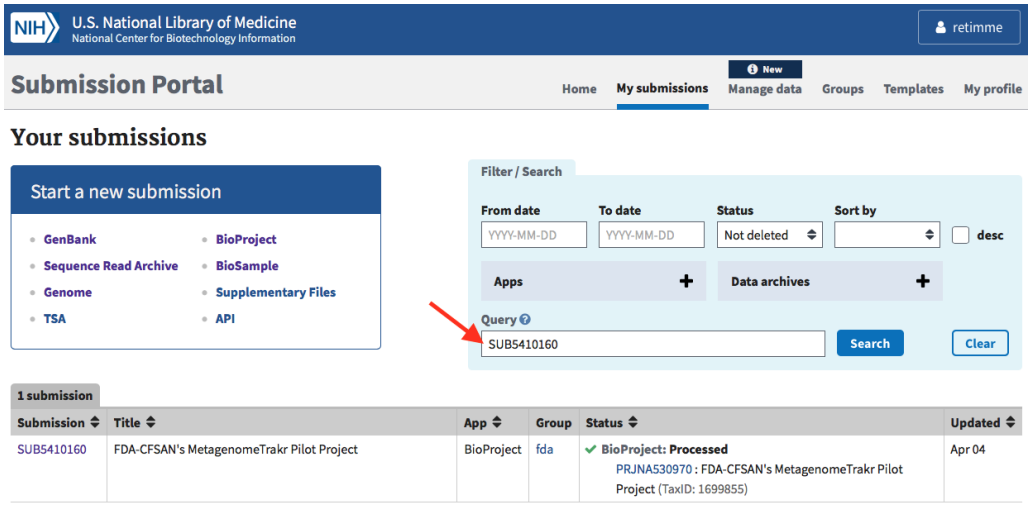
Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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 form this query to check the processing status for this BioProject.



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.**

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

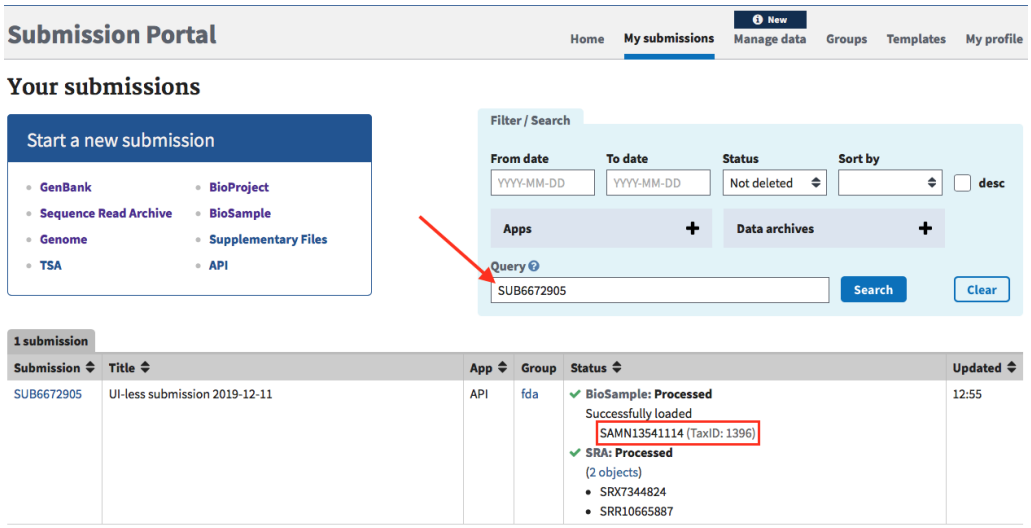
2.1 You can find your BioSample accessions in two places.

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

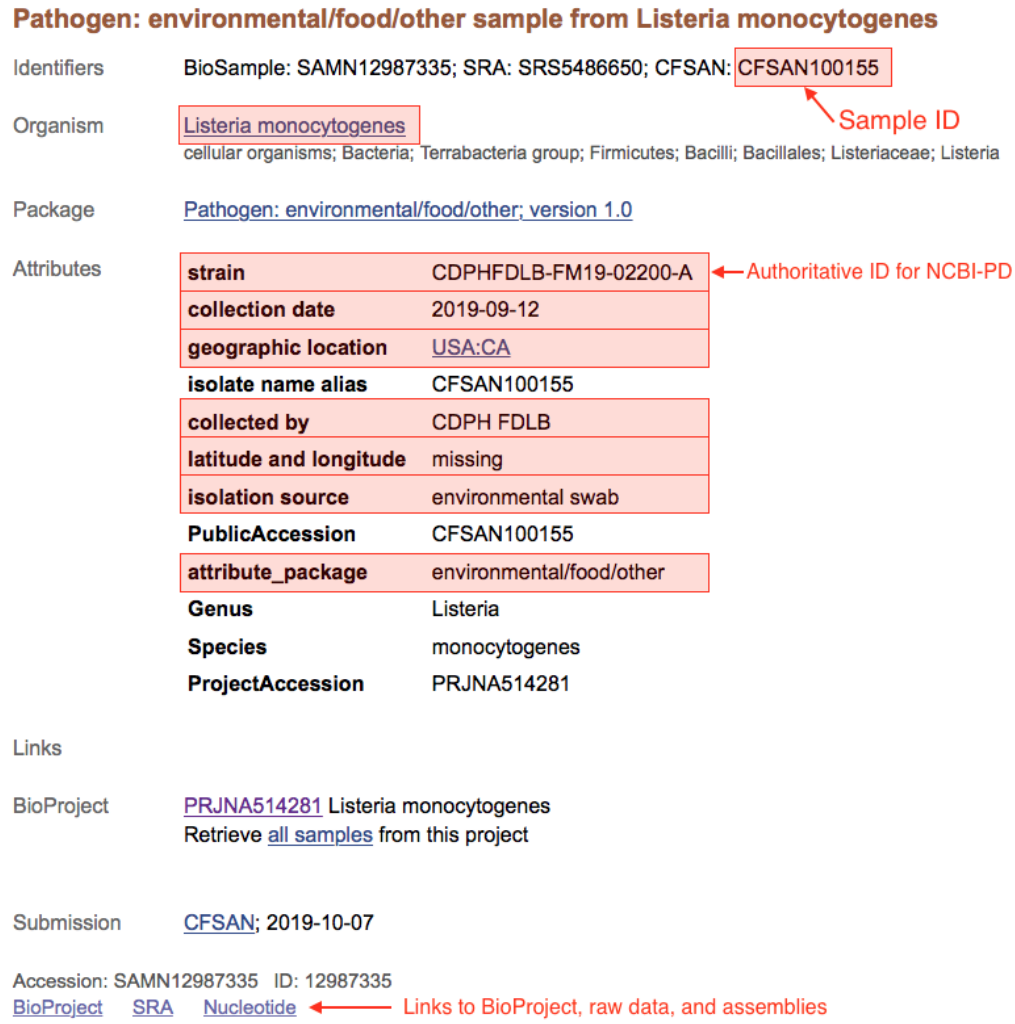


Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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").



Mandatory metadata fields are highlighted in red.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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 udpate 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.**

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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:

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)

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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 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:

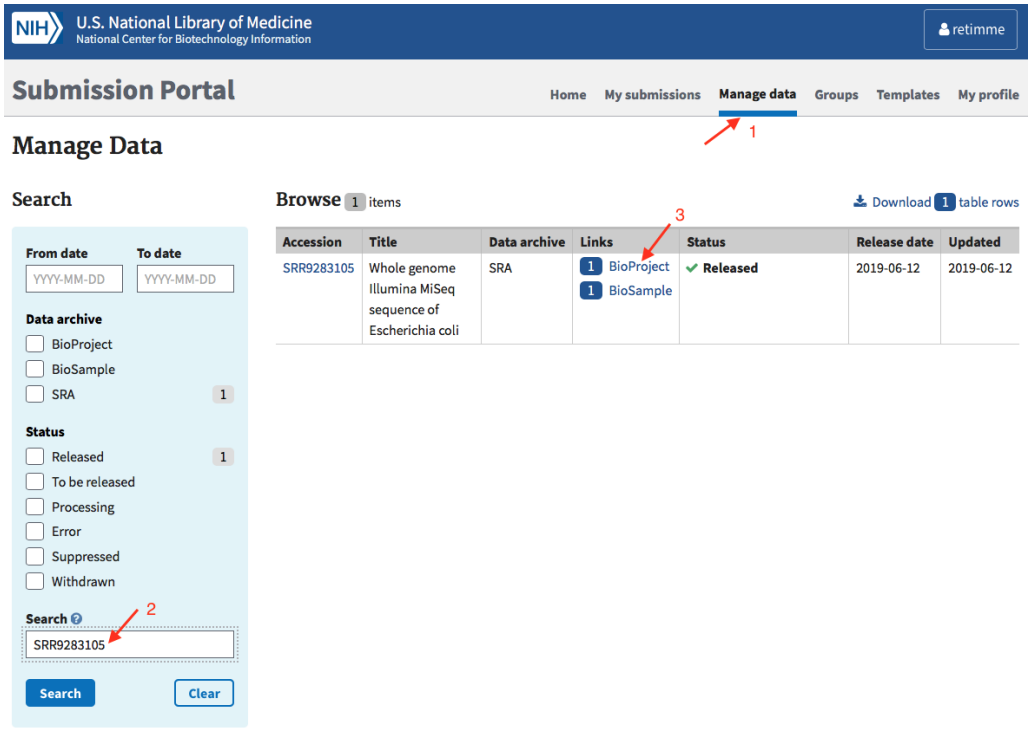
- Sequence metadata, such as library ID, library strategy, sequencing platform or instrument
- Associated BioSample or BioProject accession numbers
- Release date

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

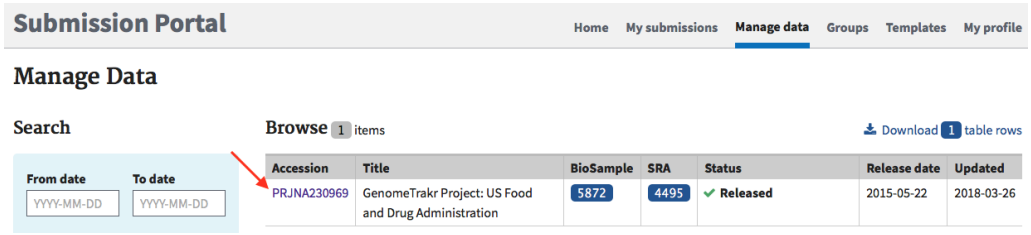
2. Query for SRR accession you'd like to update:

3. Click on the resulting "BioProject" link.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

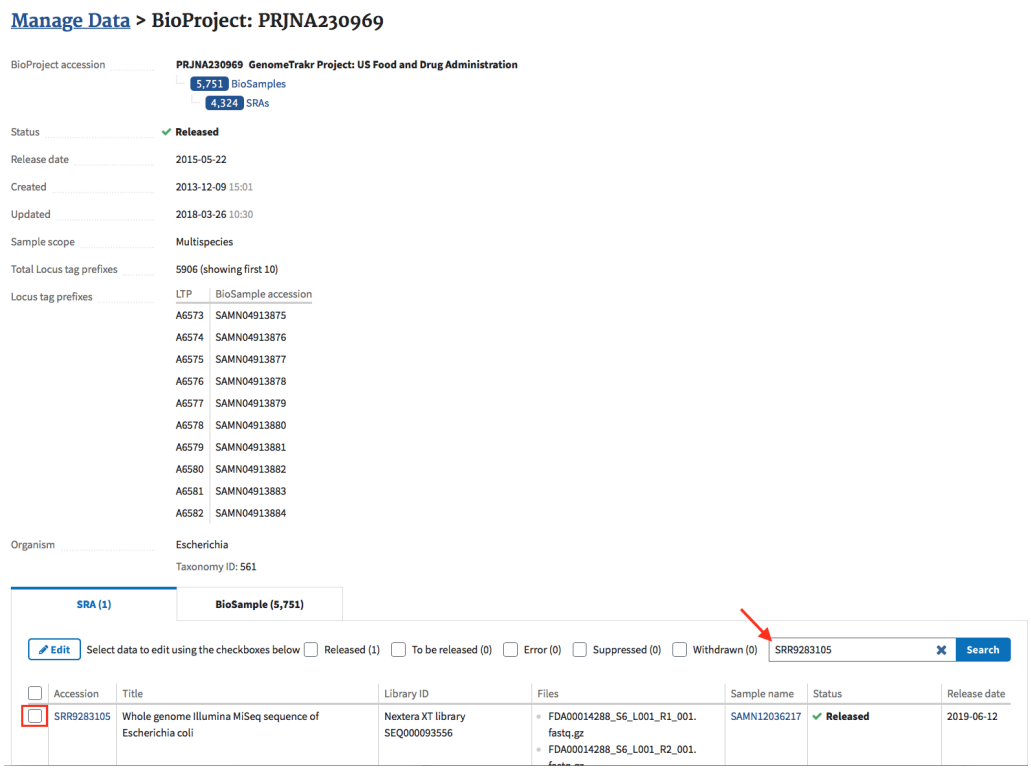


4. Click on the BioProject accession link:

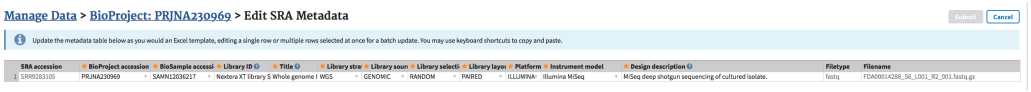


5. All the SRA records submitted to this BioProject can now be edited! Search for the one(s) you want and click the box to edit.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase



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.



Pathogen Detection

4

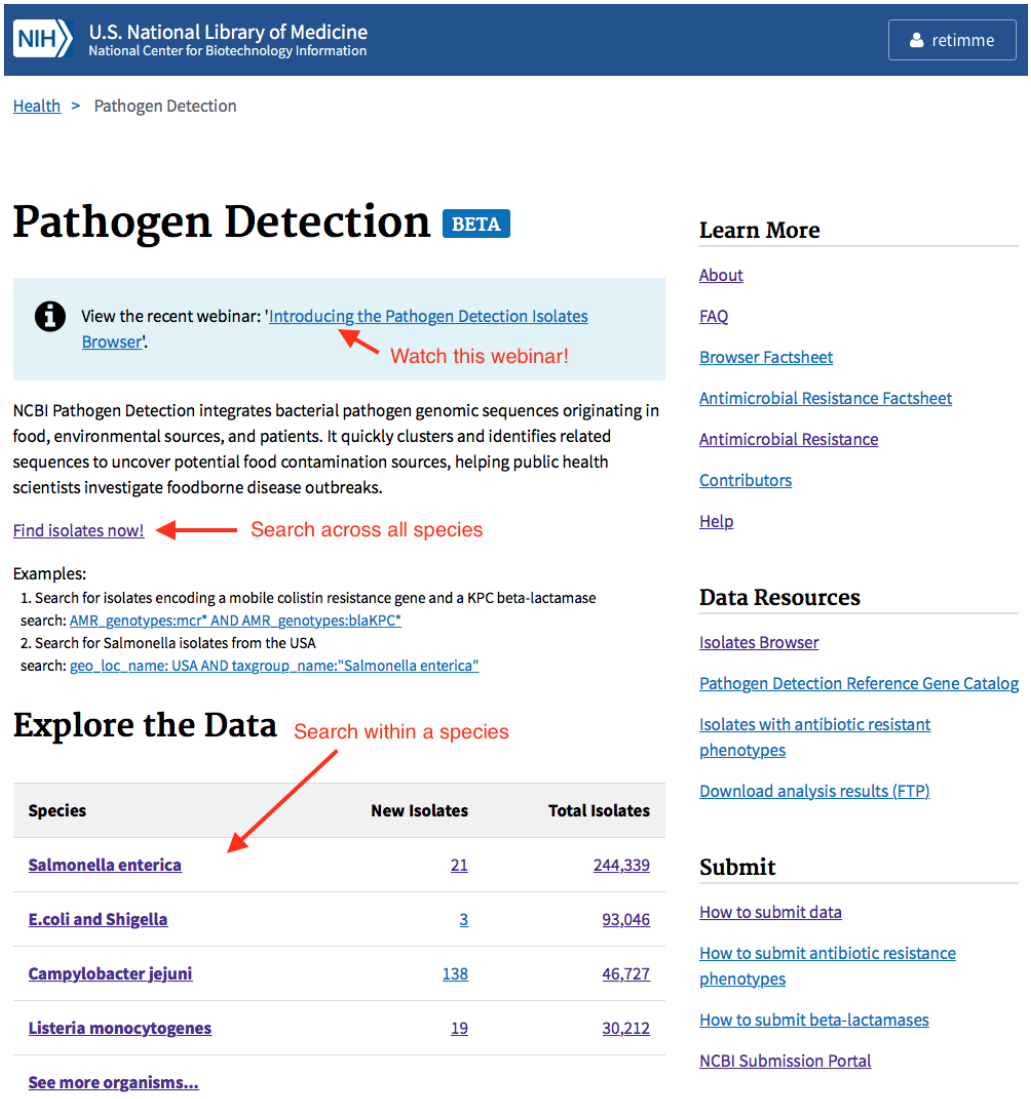
The Pathogen Detection curation protocol includes instructions for finding your data within the surveilliance platform and identifying quality control issues that might have prevented your data from being processed.

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.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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:

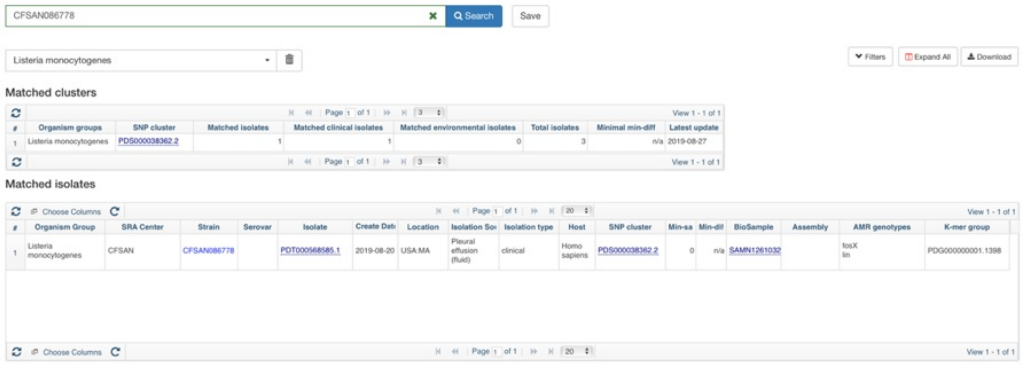


Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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



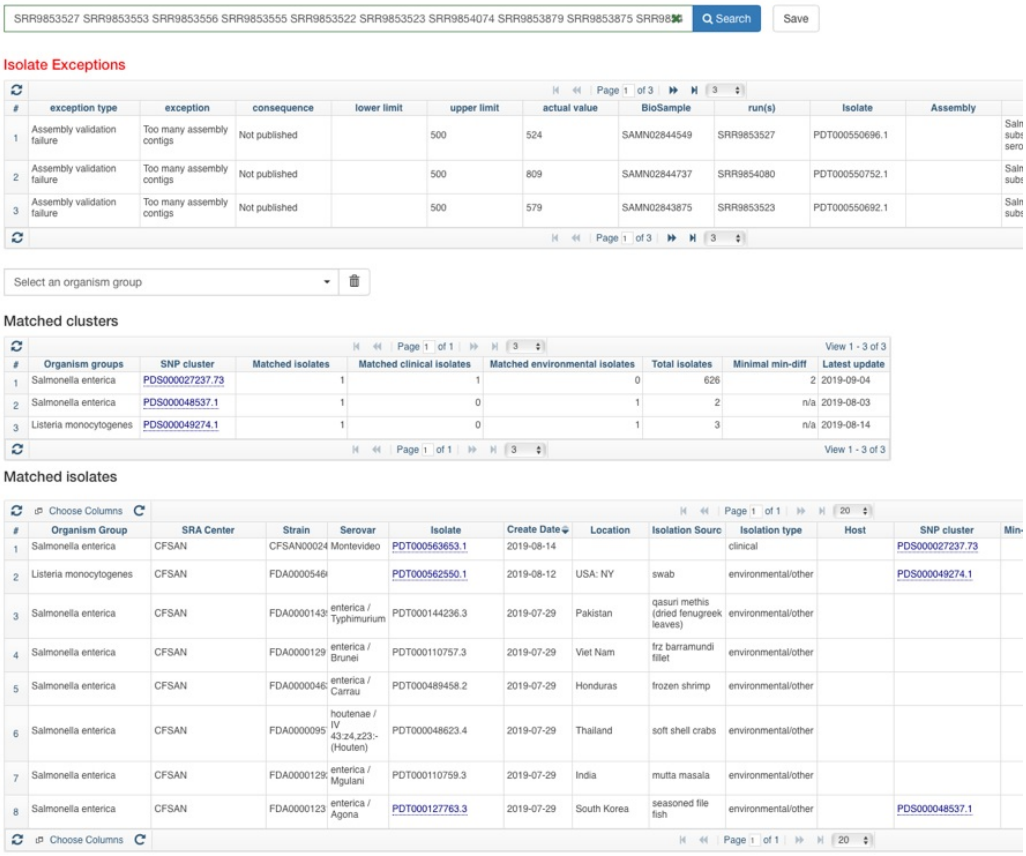
4.3 Exceptions table

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:

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase



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.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase

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>**/**PDG0000000XX.XXXX**/Exceptions/**PDG0000000XX.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.*

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolph, William Wolfgang, Errol Strain (03/23/2020). NCBI data curation protocol. https://dx.doi.org/10.17504/protocols.io.bacaiase