protocols.io

NCBI submission protocol for microbial pathogen surveillance V.3

Ruth Timme[1], Maria Balkey[2], Robyn Randolph[3], Sai Laxmi Gubbala Venkata[4], William Wolfgang[4], Errol Strain[5]

[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,

[4]Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA, [5]FDA

Upload image

Version 3

May 07, 2020

1 | *Works for me*   dx.doi.org/10.17504/protocols.io.bf3xjqpn

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

Ruth Timme
US Food and Drug Administration

ABSTRACT

PURPOSE: Step-by-step instructions for submitting WGS microbial pathogen sequence data to NCBI and to the NCBI Pathogen Detection portal.

SCOPE: to provide a standardized protocol for NCBI submission for use by any other laboratory submitting WGS data for species under active surveillance within NCBI's Pathogen Detection. (This includes GenomeTrakr, NARMS, Vet-LIRN laboratories and other networks).

This protocol covers the following submission procedures:
- Create new BioProjects
- BioSample/metadata submission
- Sequence submission

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

There are a couple benefits to having the BioSample set up prior to submitting sequences, especially for new submitters to NCBI. Establishing the BioSample separately from uploading sequences allows you to upload available metadata for a large collection of isolates before they have been sequenced, which is common when sequencing historic collections. Also, submitting your BioSamples first, then submitting your sequence data separates the validation and troubleshooting at NCBI – enabling you to focus on errors for one database, rather than two. For these reasons we recommend starting with separating biosample and sequence submissions.
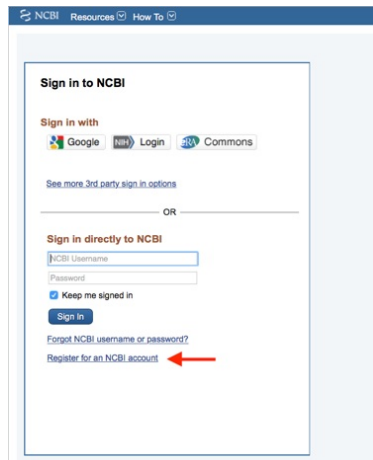
As users gain confidence, they can try the single step submission on smaller batches of samples. In cases where only a limited number of samples are going to be submitted (~a dozen or fewer samples under a single BioProject), it is possible to create the BioSample and submit sequence data in a single submission. This method can save time and effort by initiating a single submission vs two. However, validation issues can often hold up the entire submission, which could be complex to untangle for new users.

"**Ingredients**" to have in place before starting your submissions

1   **Bookmark NCBI's general submission instructions** for the Pathogen Detection portal. These can supplement steps in this protocol:

https://www.ncbi.nlm.nih.gov/pathogens/submit-data/

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

This will be your own individual user account at NCBI.

*Some laboratory groups might create a single user account that everyone in the laboratory will use.

1.2   **Establish an NCBI submission user group for your laboratory.**

We recommend using this user group for all NCBI submissions related to microbial genome surveillance. This will link your laboratory's NCBI data ownership to the *user group* and not to individuals, allowing anyone in the current group to perform updates or retractions and answer inquiries from the NCBI staff, even if there's been a complete turnover of staff since the original data submission.

User groups also ensure consistent data ownership across BioProjects, BioSamples, and sequence data. If your laboratory has non-overlapping research groups submitting and managing data at NCBI, multiple user groups can be established to track these efforts separately.

Your laboratory might already have a submission group established! Check the "Group" tab in the submission portal, https://submit.ncbi.nlm.nih.gov/groups/. Ask your colleagues to do the same thing, to ensure your laboratory doesn't already have one in place.

Creating a new submission group:

1. Submit an email request to **submit-help@ncbi.nlm.nih.gov** containing the following information:

  "*Dear NCBI help staff,*

Please establish a new user group for my laboratory.
I'm including the following information to help set up the group:

Short name of the group (abbreviation, e.g. "fda_ny")
Full name of the group (e.g. "NY Wadsworth microbial pathogen submission group")
Contact email(s) to start the group
Institution and department or group
Physical address including country
Primary contact person, first and last name plus email.

* if you have existing submissions you want to be converted, please request the ownership change in this email.
i.e., Please assign this new user group to the following BioProjects and linked data.

Thank you,"

2. Look for an email reply entitled "NCBI Submission Portal Group invitation" and click on the enclosed link to accept the invitation.

1.3 Managing your NCBI submission user group.

After a user group has been established it can be edited for membership and permissions by clicking in the "group" tab of the submission portal (https://submit.ncbi.nlm.nih.gov/groups/), then on the Group Id hyperlink, e.g 'fda_ny' in the above example.

Users with admin privileges can update contact information in the "profile" tab and membership in the "Members" tab. New members can be invited by clicking on the "Invite members" link.

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

1.4 Bookmark "my submissions" at NCBI: https://submit.ncbi.nlm.nih.gov/subs/. This is your landing page for all new NCBI submissions.

If you see a blank page with a yellow box in the upper right corner saying "please login", click this link and login using the credentials created in **Step 1.1**.

1.5 **Umbrella BioProjects**. If you are already part of a surveillance network, (e.g. GenomeTrakr, NARMS, Vet-LIRN, or PulseNet) you should use one of their established umbrella bioprojects. For reference, here are the GenomeTrakr Umbrella BioProjects for each species under surveillance.

• *Salmonella enterica* PRJNA183844
• *Listeria monocytogenes* PRJNA514048
• *Escherichia coli* and *Shigella* PRJNA230919
• *Vibrio parahaemolyticus* PRJNA245885
• *Campylobacter* sp. PRJNA258021
• *Clostridium botulinum* PRJNA290488

If you need to establish a new umbrella BioProject follow the BioProject creation in **Step 2** with modifications for creating a new Umbrella BioProject, then email the PRJNA accession to **pd-help@ncbi.nlm.nih.gov** and ask to have it flagged for submission to the Pathogen Detection pipeline. Any new data projects linked under this umbrella will inherit the linkage to Pathogen Detection.

1.6 **Download submission templates:**

Link for all submission templates: https://submit.ncbi.nlm.nih.gov/templates

Two-step submission process (two metadata templates):

Metadata: Download the metadata spreadsheet needed for creating new biosample accessions. This document is called the "combined pathogen package template" at NCBI:

■ https://www.ncbi.nlm.nih.gov/biosample/docs/templates/packages/Pathogen.combined.1.0.xlsx

Sequence files: Ensure your raw sequencing files were generated from an Illumina platform instrument (MiSeq, NextSeq, HiSeq, etc.) and download SRA's batch metadata table:

■ ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata_acc.xlsx

One-step submission process (one metadata template):

Combined BioSample/SRA submission (SRA template for submitting BioSample *and* SRA at the

same time):

- ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata.xlsx

BioProject Creation

2   **Create a new BioProject**

Would you like your new BioProject linked to the GenomeTrakr network (relevant for foodborne pathogens)?

- **YES**: Proceed to **Step 2.1**, following instructions for creating new data BioProject(s) and link them to the established GenomeTrakr Umbrella BioProjects, which are already flagged for NCBI Pathogen Detection.

- **NO**: Are you part of another broad surveillance effort that might already have existing Umbrella BioProjects (e.g. Vet-LIRN, or NARMS)?

  YES: Proceed to **Step 2.1**, following instructions for creating new data BioProject(s).
  NO: Proceed to **Step 2.1**, following instructions for creating a new Umbrella **and** data BioProjects.

If you already have data BioProjects established, gather the relevant accessions and proceed to **Step 3**.

2.1
    Navigate to the "My Submissions" page, https://submit.ncbi.nlm.nih.gov/subs/, and click "BioProject" in the "Start a new submission" box.

2.2   **Click the "New submission" box:**

2.3 Submitter tab:

Populate with submitter info. The "submitter" is the name of the person or team who is physically doing the submissions, not a supervisor or PI.

Select the appropriate submission group name (see **Step 1.2** for creating a new submission group), and describe the submitting organization or laboratory name. This will be auto-populated from the contact info you included in your NCBI user account.

2.4 Project type tab:

Project data type: Genome sequencing and assembly.

Sample scope:

- *For a **Data** BioProject*: select multi-isolate. This will allow you to submit multiple isolates or strains of the same species.

- *For an **Umbrella** BioProject*: select multi-species. This will allow you to link multiple data BioProjects representing different species under a single umbrella.

2.5 Target tab:

- *For a **Data** BioProject*: Populate ONLY the Organism name here, usually Genus species, or just Genus if your laboratory does not determine species, e.g., Salmonella enterica. Leave the strain info and Description fields blank.

- *For an **Umbrella** BioProject*: Leave the Organism name field blank. Include a list or description of species you intend to include in this effort. E.g "bacterial foodborne pathogens", or "Salmonella enterica, Listeria monocytogenes, and E.coli"

2.6 General info tab:

Click "Release immediately following processing".

Include a brief title describing the effort.
- ***Data** BioProject Title*: e.g., "GenomeTrakr Project: NY State Dept. of Health, Wadsworth Center".

- ***Umbrella** BioProject Title*: e.g. "Microbial pathogen surveillance at NY State Dept. of Health,

Wadsworth Center."

Public Description: e.g., "Whole genome sequencing of pure-cultured microbial pathogens as part of XXXX surveillance effort."

Relevance: medical.

Is your project part of a larger initiative which is already registered at NCBI?
- Data BioProjects. Click "Yes" and include a brief description and umbrella BioProject accession number (see 5.1.5). This will properly link your data project to the umbrella.

- for an Umbrella BioProject: click "NO" External links: Include a link to your laboratory's website here.

2.7 BioSample tab:

Leave blank!! You will create biosamples separately.

2.8 Publications tab:

If relevant, include publications from your laboratory.

2.9 Review and Submit tab:

Check if everything looks correct and edit if necessary, then click "**submit**."

2.10   **The BioProject accession** will be available on the "my submissions" page of the Submission portal and usually starts with "PRJNAxxxxxx." You will also receive an email containing the new accession.

2.11   **For Umbrella BioProjects creation only:**

The last step for establishment is to email bioprojecthelp@ncbi.nlm.nih and pd-help@ncbi.nlm.nih with two requests.

Example email:

"Dear BioProject and PD help teams, please convert the PRJNA##### to an Umbrella Bioproject and flag it for inclusion in the Pathogen Detection pipeline. Our laboratory will be submitting data under the XXX effort (GenomeTrakr, Vet-LIRN, NARMS, HAI, or more general pathogen surveillance).

I'd be happy to provide any additional details you might need.

Thank you, "

After the conversion is complete you can use the accession created to properly link any new data BioProjects being created, as in Step 2.6.

2.12   If you are part of a coordinated surveillance effort, like GenomeTrakr, please alert the coordinating body that a new BioProject was created under an existing umbrella (e.g, email **GenomeTrakr@fda.hhs.gov**).

2.13   **Important data stewardship and curation notes:**

- Develop an internal method for storing and tracking your BioProject accessions! They are required for every BioSample and sequence data submission to ensure proper linkage.

- Bookmark URLs for each of your BioProjects to monitor the public facing view of your submissions. e.g. FDA's Listeria BioProject: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA514048

- For updates to your BioProjects, follow the accompanying NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase?step=1

BioSample creation

3   **BioSample submission:**

This protocol follows a two-step data submission process (first register BioSamples, then submit raw reads to SRA). This section will focus on submitting only the sample metadata to the BioSample database.

Consult the accompanying "Populating the NCBI pathogen template" protocol for how to fill out this spreadsheet. https://www.protocols.io/view/populating-the-ncbi-pathogen-metadata-template-be9gjh3w

3.1 To register new BioSamples navigate to the "My Submissions" page (https://submit.ncbi.nlm.nih.gov/subs/) and click "BioSample" in the "Start a new submission" box (or navigate directly here: https://submit.ncbi.nlm.nih.gov/subs/biosample/):

3.2 Have you downloaded the metadata template?

If you don't already have the combined (human + food/env) pathogen metadata excel template:

1. Click on "Download batch submission template:

2. Select the "Pathogen affecting public health" and the appropriate package depending on the type of isolates. We recommend using the combined template for simplicity.

3. Populate the metadata spreadsheet for each isolate you intend to submit (you can submit metadata for a single isolate, entire MiSeq run, or for a large collection of isolates you intend to sequence over the next year).

- Follow the guidelines in the accompanying "Populating the NCBI pathogen template" protocol for how to fill out this spreadsheet. https://www.protocols.io/view/populating-the-ncbi-pathogen-metadata-template-be9gjh3w

3.3 **Click the "New submission" box.**

3.4 **Submitter tab:**

Populate with submitter info. The "submitter" is the name of the person, or user group, who is physically doing the submissions, not a supervisor or PI.

Select the appropriate submission group name (see **Step 1.2** for creating a new submission group), and describe the submitting organization or laboratory name. This will be auto-populated from the contact info you included in your NCBI user account. Click "Continue" to proceed.

3.5 General Info tab:

For real-time surveillance efforts, you should click "release immediately following processing". Click the "Batch/Multiple BioSamples" option to upload the metadata spreadsheet.

BioSample submission: SUB6089412
New

1 SUBMITTER **2 GENERAL INFO** 3 SAMPLE TYPE 4 ATTRIBUTES 5 DESCRIPTION 6 REVIEW & SUBMIT

General Information

Release date

ⓘ **Note:** Release of BioProject or BioSample is also triggered by the release of linked data.

★ When should this submission be released to the public?
- ⦿ Release immediately following processing
- ◯ Release on specified date or upon publication, whichever is first

★ Specify if you are submitting a single sample or a file containing multiple samples
- ⦿ Batch/Multiple BioSamples
  You will be asked to upload a tab-delimited text file that describes each of your samples and their attributes.
  Submission template files can be downloaded from the Attributes tab or the templates page.
- ◯ Single BioSample
  You will be asked to manually complete a web form to describe one sample and its attributes.

[Continue]

3.6 **Sample Type tab:**

Click Pathogen affecting public health, combined pathogen submission (or "Environmental, food or other pathogen" if no clinical isolates).

3.7 **Attributes tab:**

Click "upload a file using Excel" then "browse" to upload the file.

Please provide antibiogram data if you have it!

3.8 **Review and Submit tab:**

Check over your submission for errors by clicking on hyperlink, then submit.

3.9 **BioSample accessions** will be available on the "my submissions" page of the Submission portal by clicking on "## objects" within the submission record. You can also download by clicking the "Download attributes file with BioSample accessions". Accessions will start with SAMNxxxxxxxx. You will also receive an email within 12 hours with these same accessions.

3.10 Important data stewardship and curation note:

- Develop an internal method for storing and tracking these BioSample accessions! They are required for making future updates and for submitting sequence data generated from that sample (or isolate in our case).

- For updates to your BioSamples, follow the accompanying NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase?step=2

Sequence data submission

4 Sequence submission

1. Before submission, ensure that your sequences meet the quality control (QC) thresholds for your surveillance network. You can follow your own internal QC process or use FDA's free GalaxyTrakr platform: https://www.protocols.io/view/assessing-sequence-quality-in-galaxytrakr-bdvfi63n

4.1 Navigate to the "My Submissions" page (https://submit.ncbi.nlm.nih.gov/subs/) and click "Sequence Read Archive" in the "Start a new submission" box:

4.2 Populate SRA's batch metadata table, downloaded from Step 1.6

PRO TIPS:

1. If you have sequences to submit that belong to more than one BioProject, create a separate submission + metadata table for each of your BioProjects.
2. *Library_ID tips*: Use a modification of your sample_name to help preserve the correct link between the biosample and the data being submitted. (this would not be necessary for a single-step submission).
3. *Entering fastq filenames in the spreadsheet*: On a Mac, you can directly copy the file names from the folder into a spreadsheet. This is not possible on a PC using copy and paste but can be done with some command-line operation.
4. Finally, it is important to develop a QA/QC step to make sure the files are associated with the correct sample name. For example, use a left function in excel to strip of the appended text in the file name and then use the exact match to make sure the name matches the sample name.

Field	Description
BioSample	BioSample accession (SAMN######)
library_ID	The library name should be a unique ID relevant to your workflow. It can be an autogenerated ID from your LIMS system or a modification of your sample_name (see above pro tip #2).
Title	Short, free text description that identifies the data on public pages. For Example: {methodology} of {organism}: {sample info} e.g. "WGS of a Salmonella isolate"
library_strategy	Overall sequencing strategy or approach. Choose from NCBI pick list, e.g. "WGS"
library_source	molecule type used to make the library Choose from NCBI pick list, e.g. "GENOMIC"
library_selection	Library capture method Choose from NCBI pick list, e.g. "RANDOM"
Library_layout	Choose from NCBI pick list, e.g. "PAIRED"
platform	Sequencing platform Choose from NCBI pick list, e.g. "Illumina"
instrument_model	Name of the sequencing instrument Choose from NCBI pick list, e.g. "Illumina MiSeq"
Design_description	optional field for free text description of methods e.g. "MiSeq deep shotgun sequencing of cultured isolate"
Filetype	File format name for the raw sequence data Choose from NCBI pick list, e.g. "fastq"
Filename	include ALL of the files resulting from this library. **Add additional fields if there are more than two files (e.g. Filename3) genome_r1.fastq (*must be exact)
Filename2	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names for NextSeq data

Save the second sheet (SRA_data) as a TSV (tab-delimited file) for upload in the "SRA metadata" tab within the submission portal.

*NCBI should also accept the original excel formatted file.

Click "New Submission" button at the top of the page:

4.3

4.4 *Submitter tab:*

Select the appropriate submission group name (see **Step 1.2** for creating a new submission group). Most contact info will be auto-populated. Ensure it's complete before proceeding.

4.5 *General tab:*

Click yes for existing BioProject and paste or type in your **data** BioProject accession.

****HINT**: group your submissions by BioProject since you can only specify one BP accession here.

Click **yes** for existing BioSamples (following the two-step submission process).

Click *"***release immediately after processing***"*

4.6 Metadata tab:

Click "Upload a file using Excel or text format (tab-delimited)" and "choose file" saved in **Step 4.2**.

4.7 Files tab:

Check and resolve any validation errors.

4.8 File upload:

You can provide files for submission by:

- Web browser upload via HTTP or Aspera Connect plugin. This is recommended for small batches and the upload is done directly in the web interface.

- FTP or Aspera Command-Line file preload: Use this option if you are uploading files over 10 GB or more than 300 files. Requires beginner experience using a command-line interface. Obtain FTP and Aspera upload instructions in the submission FILES tab. It will include the FTP address, username, and password.

- Amazon S3 bucket.

4.9   *Review and Submit tab*:

Check over entire submission, then click submit.

4.10   **SRA Accessions:**

SRA run accessions will be available on the "my submissions" page of the Submission portal by clicking on "## objects" within the submission record. You can also download by clicking the "Download metadata file with SRA accession". Accessions will start with SRRxxxxxxx." You will also receive an email with these same accessions within 24 hours.

4.11   **Important data stewardship and curation note:**

- Develop an internal method for storing and tracking these SRR accessions! They are required for making updates and identifying the data in any resulting publications.

- Caution: It is possible for a single BioSample to have more than one SRR ID. Two scenarios include 1) two runs were submitted for the same isolate/BioSample (this is **not** recommended for submission to NCBI Pathogen Detection). 2) if the initial submission was retracted and new a new run was submitted. It's important to keep track of both IDs, even if one was retracted.

- For updates to your SRA submissions, follow the accompanying NCBI Curation Protocol: https://www.protocols.io/view/ncbi-data-curation-protocol-bacaiase?step=3