Mar 17, 2020

NCBI submission protocol for microbial pathogen surveillance

Ruth Timme[1], Maria Balkey[2], Sai Laxmi Gubbala Venkata[3], Robyn Randolf[4], William Wolfgang[3], 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]Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA, [4]Association of Public Health Laboratories, Silver Spring, Maryland, USA, [5]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.9aph2dn

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

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

"**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

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

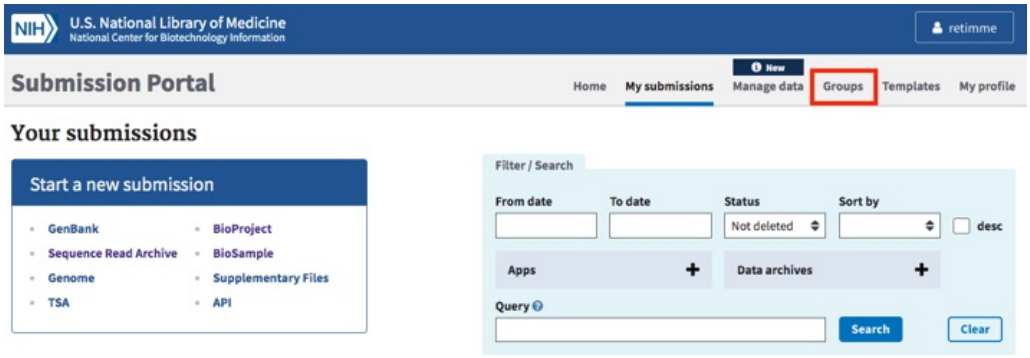
This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

1.2 **Create an NCBI submission user group** for your laboratory. This will allow multiple people to have edit privileges for your NCBI data submissions (multiple people could edit/retract data). This is highly recommended.

Submit a request via email to submit-help@ncbi.nlm.nih.gov containing the following information:

- Short name of the group (abbreviation, e.g. FDA-CFSAN)
- Full name of the group (The FDA-CFSAN GenomeTrakr 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 and email. You can also provide an optional phone number.

Editing a user group. After a user group has been established you can freely edit membership by clicking in the "group" tab of the submission portal, https://submit.ncbi.nlm.nih.gov/subs where you will find the list of groups containing your membership. You can add /delete members from groups for which you are an owner or have edit privileges.

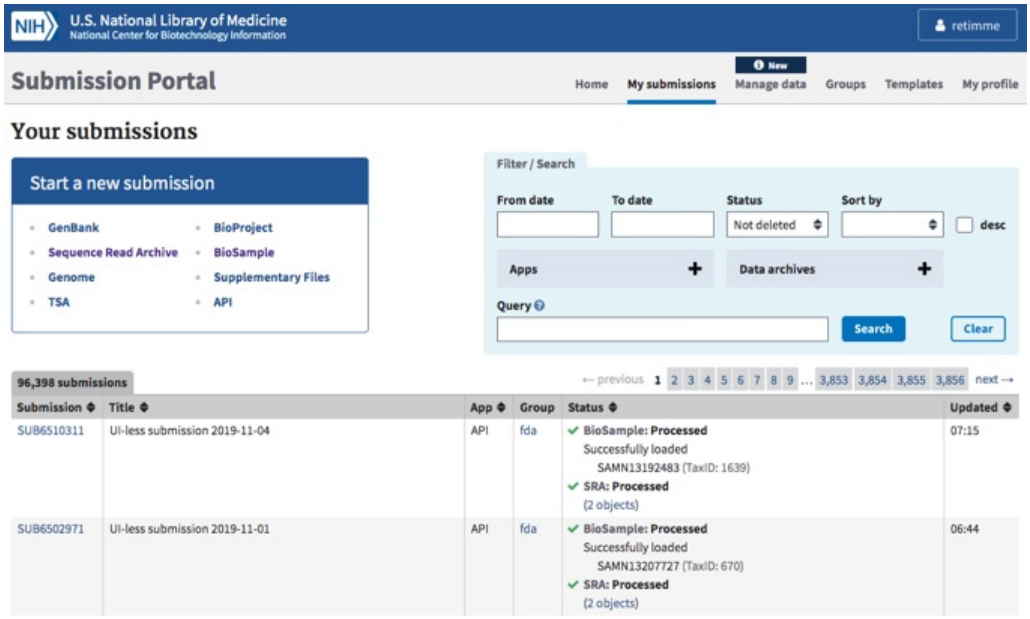


Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

1.3 **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 1.2.



1.4 **Umbrella BioProjects**. If you are already part of a surveillance network, e.g. GenomeTrakr, you should use one of their established umbrella bioprojects. 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.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

1.5 ⚠

Download submission templates:

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

Combined BioSample/SRA submission (SRA template for submitting BioSample AND SRA within the same submission):

- 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)?

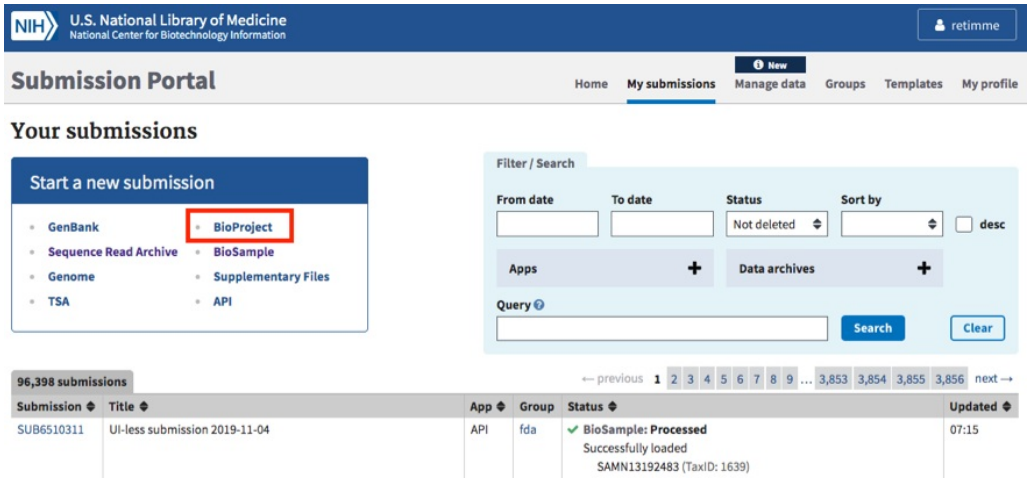
  YES: Proceed to **Step 2.1**, following instructions for creating new data BioProject(s).
  NO: Proceed to **Step 2.1**, following instructions for creating a new Umbrella **and** data BioProjects.

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

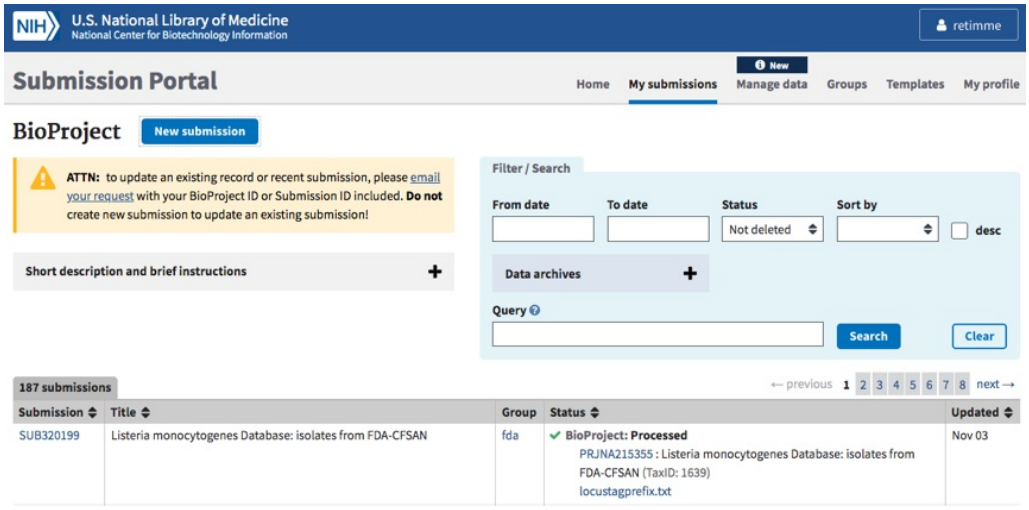
Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

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 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.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

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.”

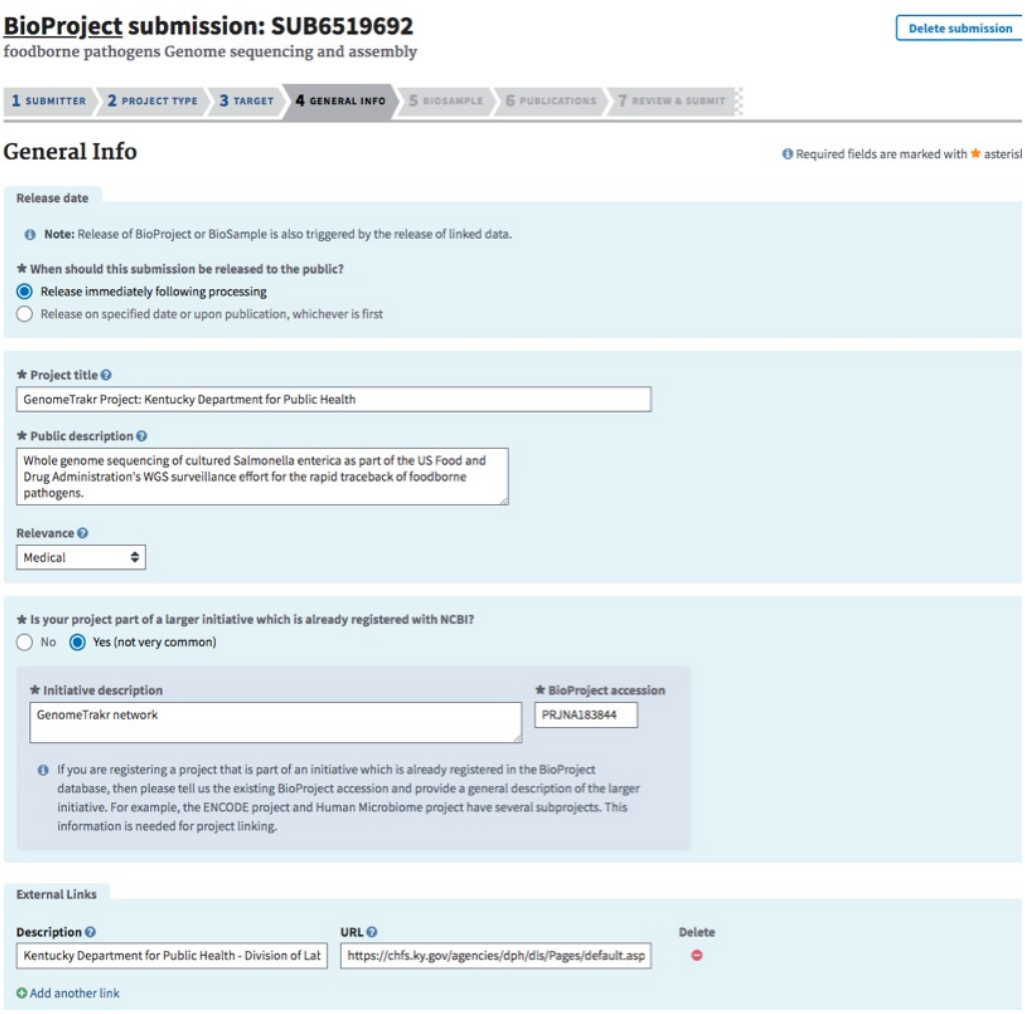
Relevance: medical.

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.

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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



2.7 ***BioSample tab:***

Leave blank!! You will create biosamples separately.

2.8 ***Publications tab:***

If relevant, include publications from your laboratory.

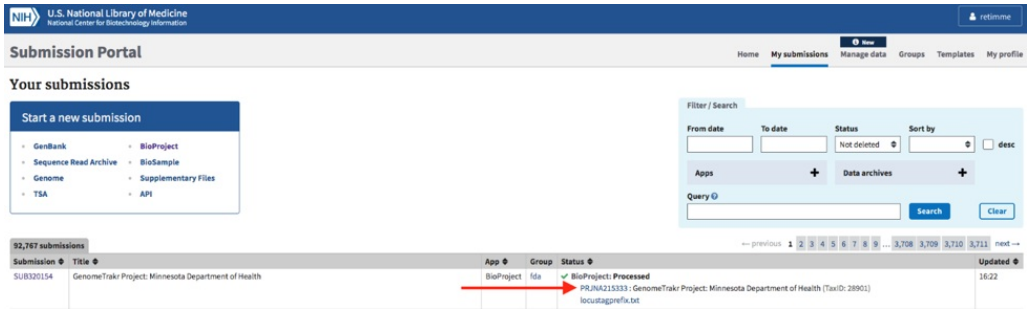
2.9 ***Review and Submit tab:***

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

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

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).

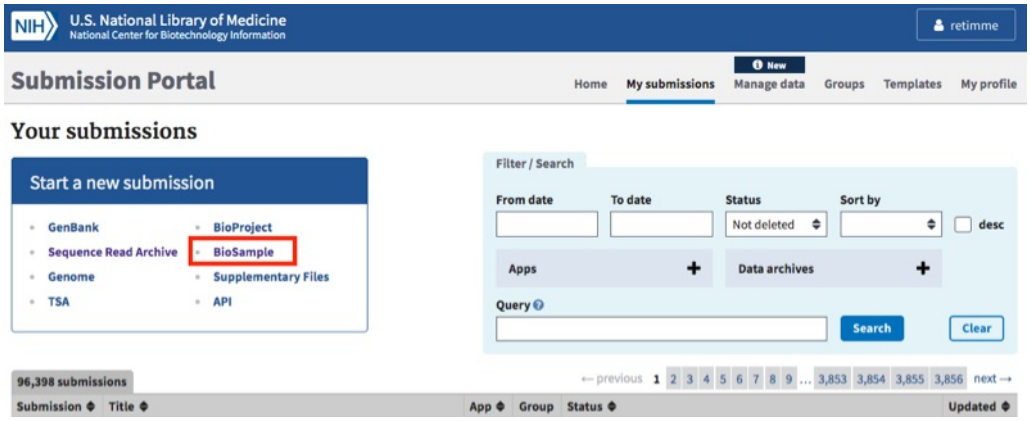
Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

BioSample creation

3 **BioSample/metadata submission:**

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/):



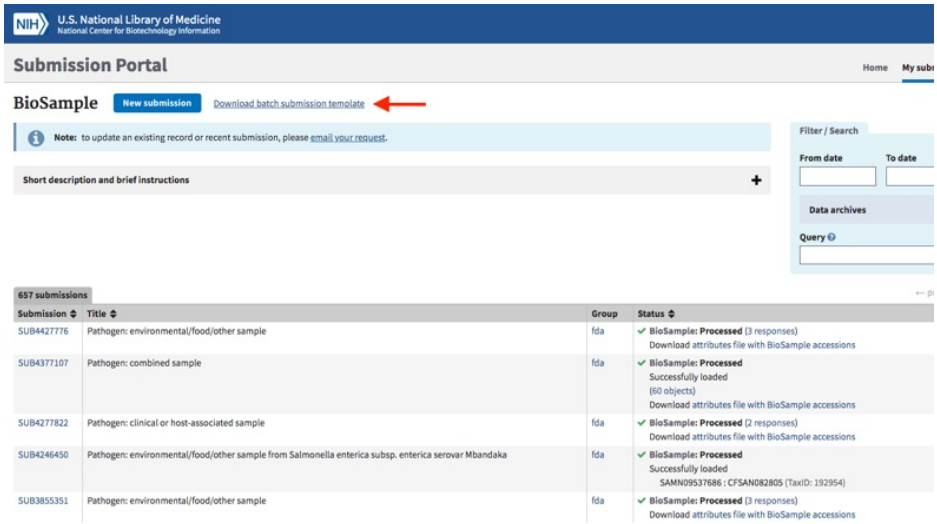
Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

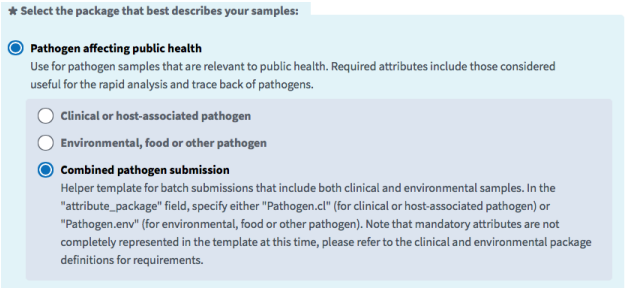
3.1 Have you downloaded the metadata template?

If you don't already have the combined 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.



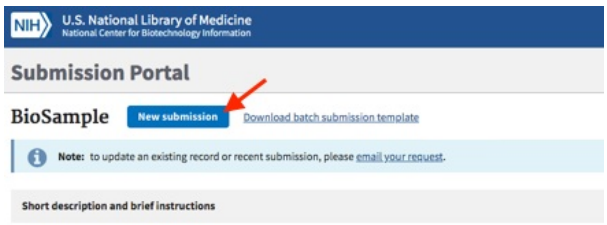
Populate the metadata spreadsheet for each isolate you intend to submit (you can submit metadata for a single MiSeq run, or for a large collection of isolates you intend to sequence).

Ensure that text is included for **ALL** mandatory fields. Include the word "missing, or "not collected" if data are not available for a given mandatory field (see **Table 3** in the main document for the core set of required fields – you are welcome, and encouraged, to include more metadata in the optional fields).

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

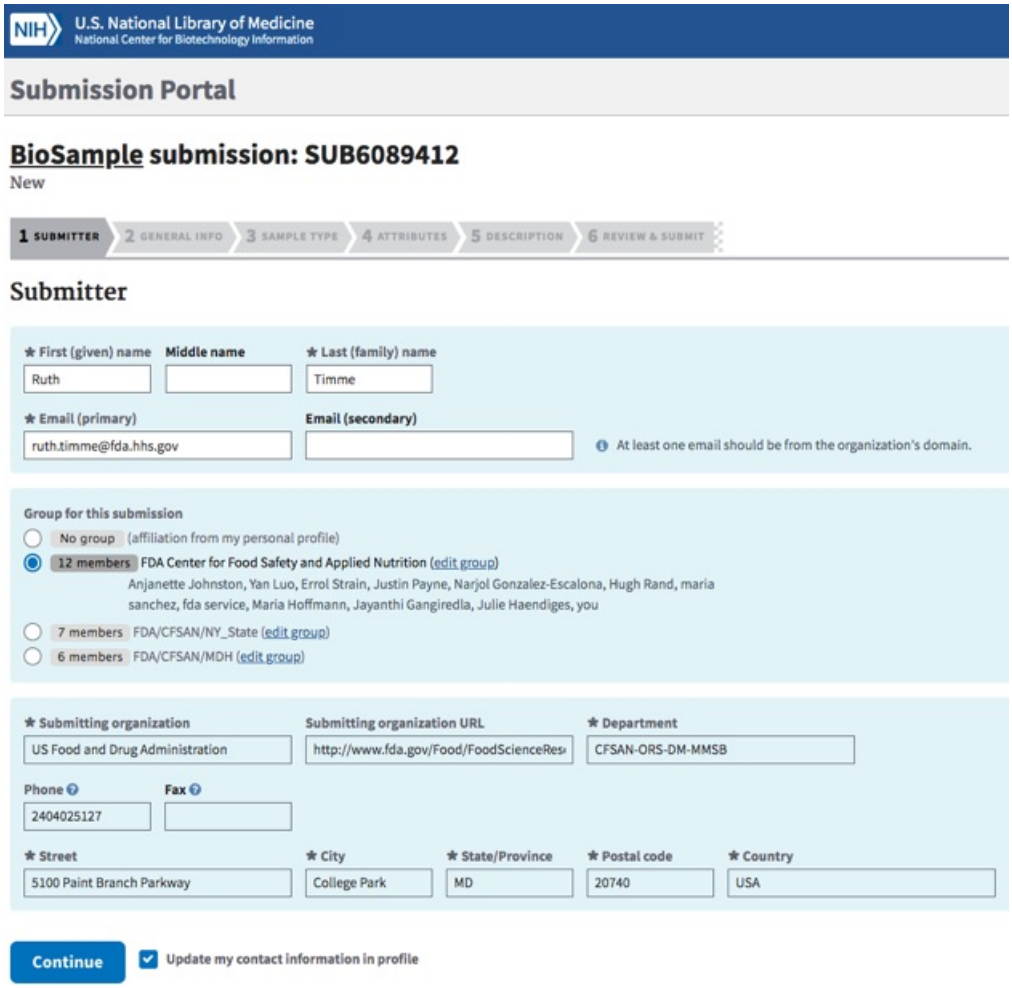
This is an open access protocol distributed under the terms of the **Creative Commons Attribution License** (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

3.2 **Click the “New submission” box.**



3.3 **Submitter tab:**

Populate with submitter info. The “submitter” is the name of the person 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.

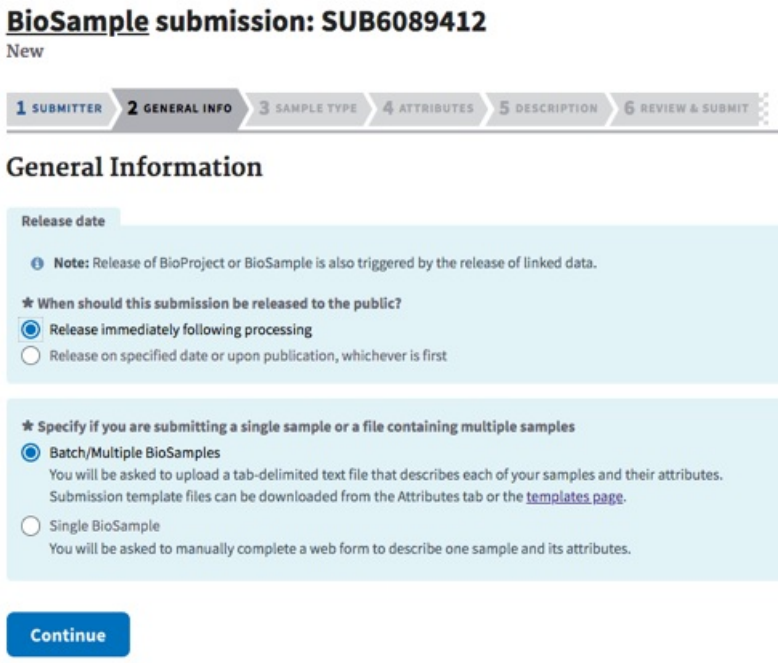


Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

3.4 **General Info tab:**

For real-time surveillance efforts you should click “release immediately following processing” (e.g. GenomeTrakr). Click the “Batch/Multiple BioSamples” option to upload the metadata spreadsheet.



3.5 **Sample Type tab:**

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

3.6 **Attributes tab:**

Click “upload a file using Excel” then “browse” to navigate to upload file.

Please provide antibiogram data if you have it!

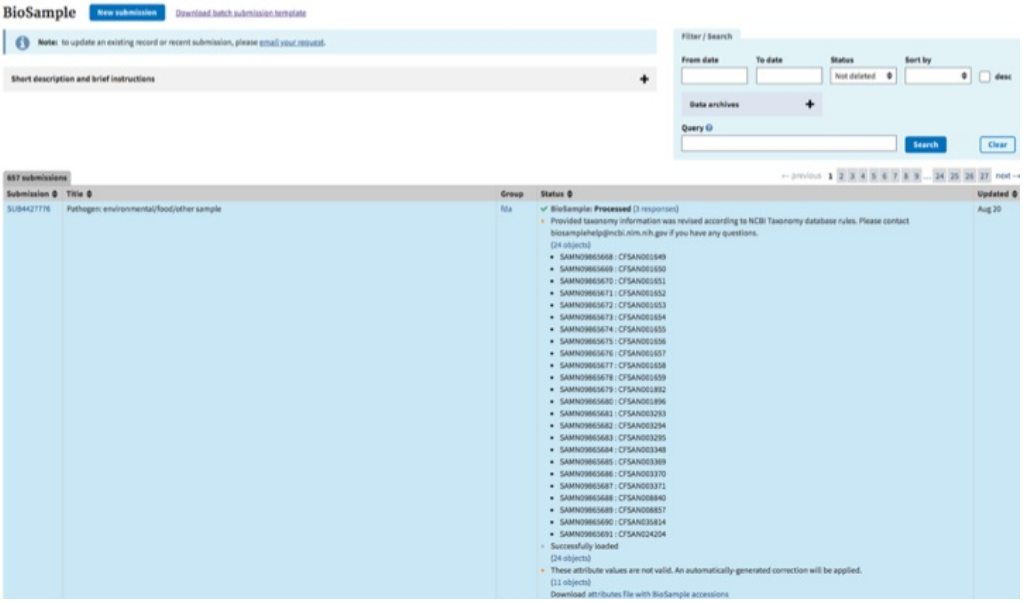
3.7 **Review and Submit tab:**

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

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

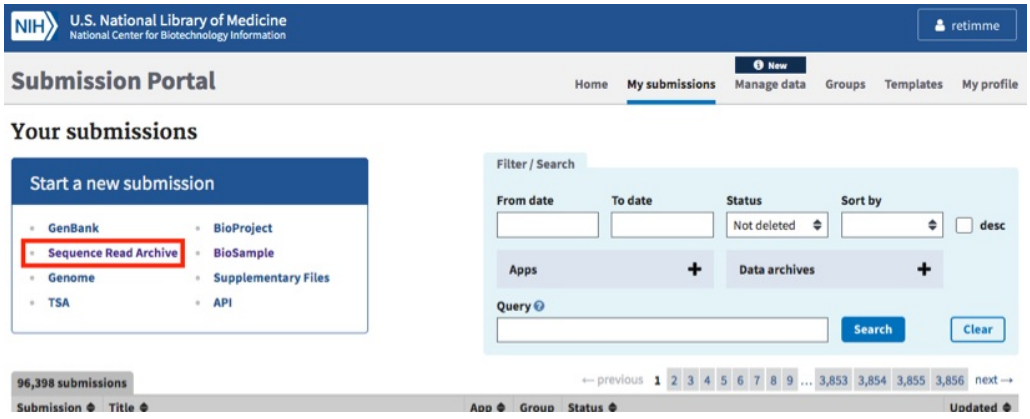
3.8 **BioSample accessions** will be available on the “my submissions” page of the Submission portal by clicking on “## objects” within the submission record. You can also download by clicking the “Download attributes file with BioSample accessions”. Accessions will start with SAMNxxxxxxxx. You will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



Sequence data submission

4 Sequence submission

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:



Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

4.1 **Populate SRA's batch metadata table**, downloaded from **Step 1.5**

If you have sequences to submit that belong to more than one BioProject, create a separate table for each of your BioProjects. We recommend using the sample name for the library ID so that you can copy and paste the biosample and sample name from the csv file returned from NCBI after the Biosample submission. This should help preserve the correct link between a biosample name and the sample as we create this second metadata sheet. (this would not be neccessary in a single step submission).

Entering fastq filenames in the spreadsheet: On a Mac you can directly copy the files 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.

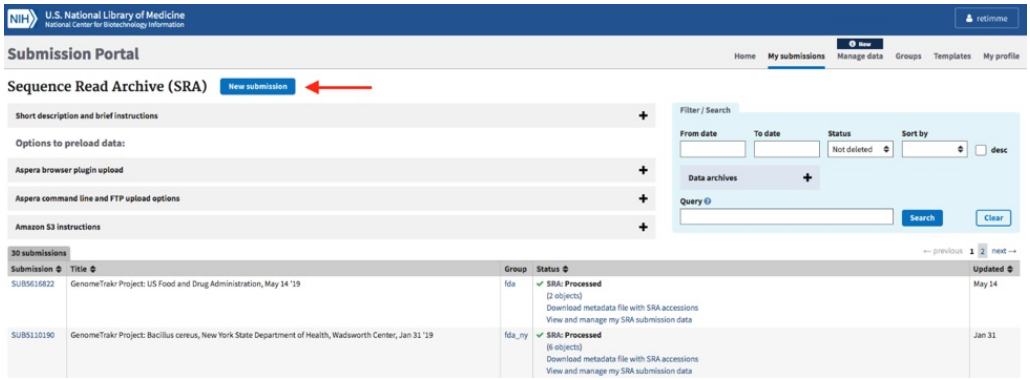
Finally it is important to develop a QA/QC step to make sure the files are associated with the correct sample name. We do this using left function in excel to strip of the appended text in the file name and then use exact match to make sure the name matches the sample name.

Field	Description
BioSample	BioSample accession (SAMN######)
Library ID	unique name for library (can be ID from LIMS)
Title	Free Text - GenomeTrakr looks like this, "Nextera XT library SEQ0000#####"
Instrument	Illumina MiSeq (or other Illumina instrument)
Strategy	WGS
Source	GENOMIC
selection	RANDOM
Layout	PAIRED
Platform:	Choose from the dropdown menu. e.g. MiSeq
Design_description	Free text. e.g. "MiSeq deep shotgun sequencing of cultured isolate"
Filetype	fastq
Filename	genome_r1.fastq (*must be exact)
Filename2	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names for NextSeq data

Save 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.

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



Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

4.3 ***Submitter tab:***

Populate with personal and institutional information. The “submitter” is the name of the person 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.

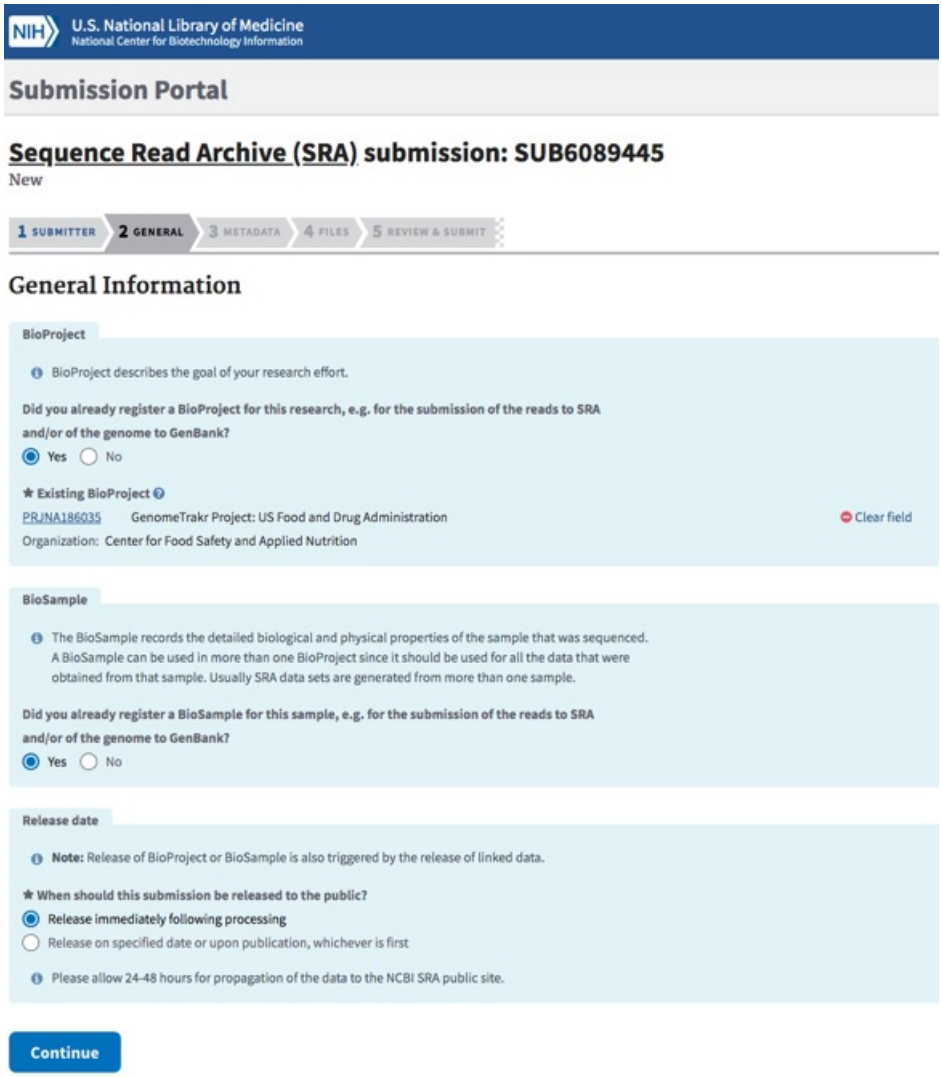
4.4 ***General tab:***

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.

Click “**release immediately after processing**”



Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

4.5 ***Metadata tab:***

Click “Upload a file using Excel or text format (tab-delimited)” and “choose file” saved in **Step 4.1**.

4.6 ***Files tab:***

Check and resolve any validation errors

4.7 ***File upload:***

You can provide files for submission by:

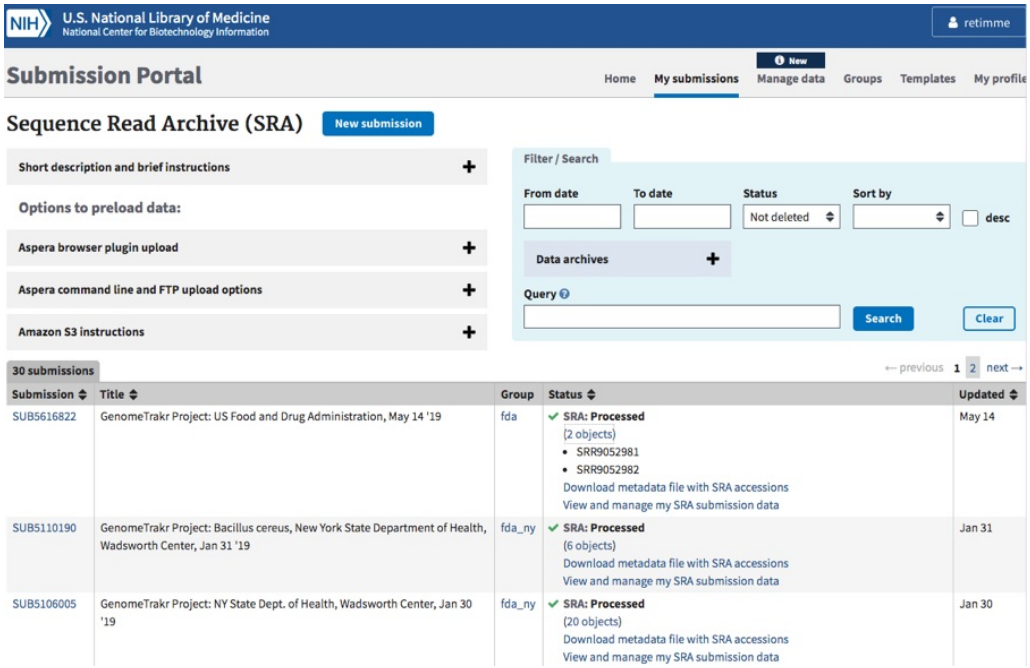
- Web browser upload via HTTP or Aspera Connect plugin: Recommended for small batches. The upload is done directly in the GUI interface.
- FTP or Aspera Command Line file preload: Use this option if you are uploading files over 10 GB or more than 300 files. Requires beginner experience with command line. Obtain FTP and Aspera upload instructions in the submission FILES tab. It will include ftp address, username and password.
- Amazon S3 bucket.

4.8 ***Review and Submit tab:***

Check over entire submission, then click submit.

4.9 **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.



This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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

unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Citation: Ruth Timme, Maria Balkey, Sai Laxmi Gubbala Venkata, Robyn Randolf, William Wolfgang, Errol Strain (03/17/2020). NCBI submission protocol for microbial pathogen surveillance. https://dx.doi.org/10.17504/protocols.io.9aph2dn

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