

Feb 15, 2024

Version 1

Populating NCBI template for submissions using BioNumerics V.1

DOI

<https://dx.doi.org/10.17504/protocols.io.3byl4qn4ovo5/v1>



Ruth Timme¹, Maria Balkey¹, Julie Haendiges¹, Brian Sauders², Tina.Pfefer¹

¹US Food and Drug Administration; ²New York State Department of Agriculture & Markets

GenomeTrakr

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



Maria Balkey

US Food and Drug Administration

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.3byl4qn4ovo5/v1>

Protocol Citation: Ruth Timme, Maria Balkey, Julie Haendiges, Brian Sauders, Tina.Pfefer 2024. Populating NCBI template for submissions using BioNumerics . **protocols.io** <https://dx.doi.org/10.17504/protocols.io.3byl4qn4ovo5/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: February 09, 2024

Last Modified: February 15, 2024

Protocol Integer ID: 94941

Keywords: NCBI submission, BioNumerics, biosample, SRA, metadata, bioproject, biosample update to ncbi, metadata template pulsenet-bionumerics-isolate-metadata, changes in metadata template pulsenet-bionumerics-isolate-metadata, environmental isolates to ncbi, ncbi template for submission, populating ncbi template, submission of genometrakr, bionumerics for submission, genometrakr submission, collecting isolate metadata, using bionumeric, genometrakr, isolate metadata, records to ncbi, bionumeric, familiar with the mandatory metadata field, biosample update, metadata template, genometrakr network management, ncbi, environmental isolate, mandatory metadata field, project name to metadata template, sequencing record, sequencing, sourcecountrystate,

Abstract

PURPOSE: to define the standard operating procedure for collecting isolate metadata using BioNumerics for submission of food/environmental isolates to NCBI.

SCOPE: to provide a standardized procedure to collect isolate metadata using BioNumerics for submission of food/environmental isolates to NCBI.

RESPONSIBILITIES- SOP Responsible Officials: Ruth Timme, Maria Balkey

The GenomeTrakr Network Management will be responsible to monitor GenomeTrakr submissions processed through Bionumerics and ensure that all GT labs are familiar with the mandatory metadata fields required for submission of GenomeTrakr sequencing records to NCBI.

V3: Added dropdown menus from controlled vocabulary to sequenced by and project name to metadata template *PulseNet_Bionumerics_Isolate_Metadata*

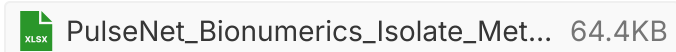
V4: Changes in metadata template *PulseNet_Bionumerics_Isolate_Metadata*.

- Added dropdown menus from controlled vocabulary to collected_by , SourceCountryState
- Added fields: collected by, isolation source
- Added mapping table of attribute names.
- Remove requirement to send biosample update to NCBI to make changes on sequenced by and project name.

1 Metadata SampleSheet preparation

Before uploading your sequencing run or linking NCBI sequencing records at the BioNumerics platform make sure to fill out the metadata spreadsheet form.

Please download the template and guidelines included in the file



Create the fields NCBI_bioproject, Attribute_package, Organism_name, NCBI_LabID, Collected by, SourceCountryState, Latitude_longitude, ProjectName, SequencedBy, Isolation source if **they are not** in the BioNumerics interface.

Once you have filled out the template information, save the **template sheet** as **.csv** and import the metadata to BioNumerics.

- The metadata fields created in Bionumerics will map metadata fields at NCBI. Table 1 describes each of the fields submitted to NCBI along mapping against name of the fields in Bionumerics templates.

A	B	C	D
Field Name at BioNumerics NCBI Submission Prompt	Field Name at NCBI	Field Name in BioNumerics Submission Metadata Template	Description
BioProject accession	BioProject	NCBI bioproject	The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxxx). **Double check that you are submitting to the correct BioProject (the organism name must match the one designated for your BioProject). For species that fall outside of NCBI pathogen detection, we recommend establishing a separate multi-species "research" bioproject for publishing data outside of the structured Pathogen Detection surveillance effort.



A	B	C	D
Attribute package	attribute_package	attribute_package	This field provides the pathogen type (or "isolation type"). Allowed values are "Pathogen.cl" (for human clinical pathogens) or "Pathogen.env" (for environmental, food, or animal clinical isolates). The value provided in this field drives validation of other fields and cannot be left blank.
Strain name	strain	Key	This is the authoritative ID used for foodborne pathogen genomic epidemiology and within NCBI Pathogen Detection. Although the strain ID can have any format, we suggest that it be unique, concise, and consistent within your laboratory (e.g. CFSAN123456).
Serovar	serovar	Serovar	The organism serovar/serotype name should include the most descriptive information you have at time of submission, adhering to proper nomenclature in NCBI taxonomy database: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi . Check spelling carefully!
Isolate name alias	isolate name alias	isolate_name_alias	Other IDs associated with this isolate. Separate with ';' if more than one
Project name	project name	ProjectName	Name of the project within which the sequencing was organized
Collected by	collected by	collected by	Full name of laboratory or agency that collected the sample or has taken over



A	B	C	D
			curation of the physical isolate. The name should be written out in full, (with minor exceptions) and be consistent across multiple submissions. Example: Washington State Department of Health.
Collection / Isolate date	collection date	IsolateDate	Date on which the sample was collected. Populate using ISO 8601 standard: "YYYY-mm-dd", "YYYY-mm" or "YYYY" (e.g., 1990-10-30, 1990-10, or 1990). Including the month or month/day of collection is extremely valuable for accessing seasonality in the database.
Geographical origin	geographic location	SourceCountryState	Populate the geographic origin of the food product. Include the country name if imported, or the "Country: state/territory/province" if domestic. Include multiple locations if necessary, delimited by semi colon.
Geographical coordinates	latitude and longitude	lat_long	The geographical coordinates of the location where the sample was collected. Specify as degrees latitude and longitude in format "d[d.dddd] N S d[dd.dddd] W E", eg, 38.98 N 77.11 W. If information is unavailable for any mandatory field, please enter 'not collected', 'not applicable' or 'missing' as appropriate.
Isolation source	isolation source	isolation source	Free text, short description of sample source. Avoid generic terms such as patient, sample, food, surface, clinical, product, source, or



	A	B	C	D
				environment. Example: bagged romaine lettuce.
	Host	host	host	For human, animal, and plant hosts, include the full taxonomic name of the host when available, "Homo sapiens" or "Bos Taurus". Animal livestock terms are also acceptable entries, e.g. porcine, bovine, equine, etc.
	Host disease	host disease	host_disease	Name of relevant disease, e.g. Salmonella gastroenteritis. Choose an ontological term from https://bioportal.bioontology.org/ontologies/DOID or https://www.ncbi.nlm.nih.gov/mesh . Attribute is mandatory for Pathogen.clin isolates (human clinical isolates) or include "missing" if unknown. Leave blank if not relevant.
	Sequenced by	sequenced by	SequencedBy	The name of the agency that generated the sequence, e.g., Centers for Disease Control and Prevention
	Source name/type	source type	SourceType	Controlled vocabulary describing the isolation_source. Choose the best fit term: Human, Animal, Food, Environmental, Other.
	Organism name	Organism	organism	The organism name should include the most descriptive information you have at time of submission, adhering to proper nomenclature in NCBI taxonomy database: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/ww

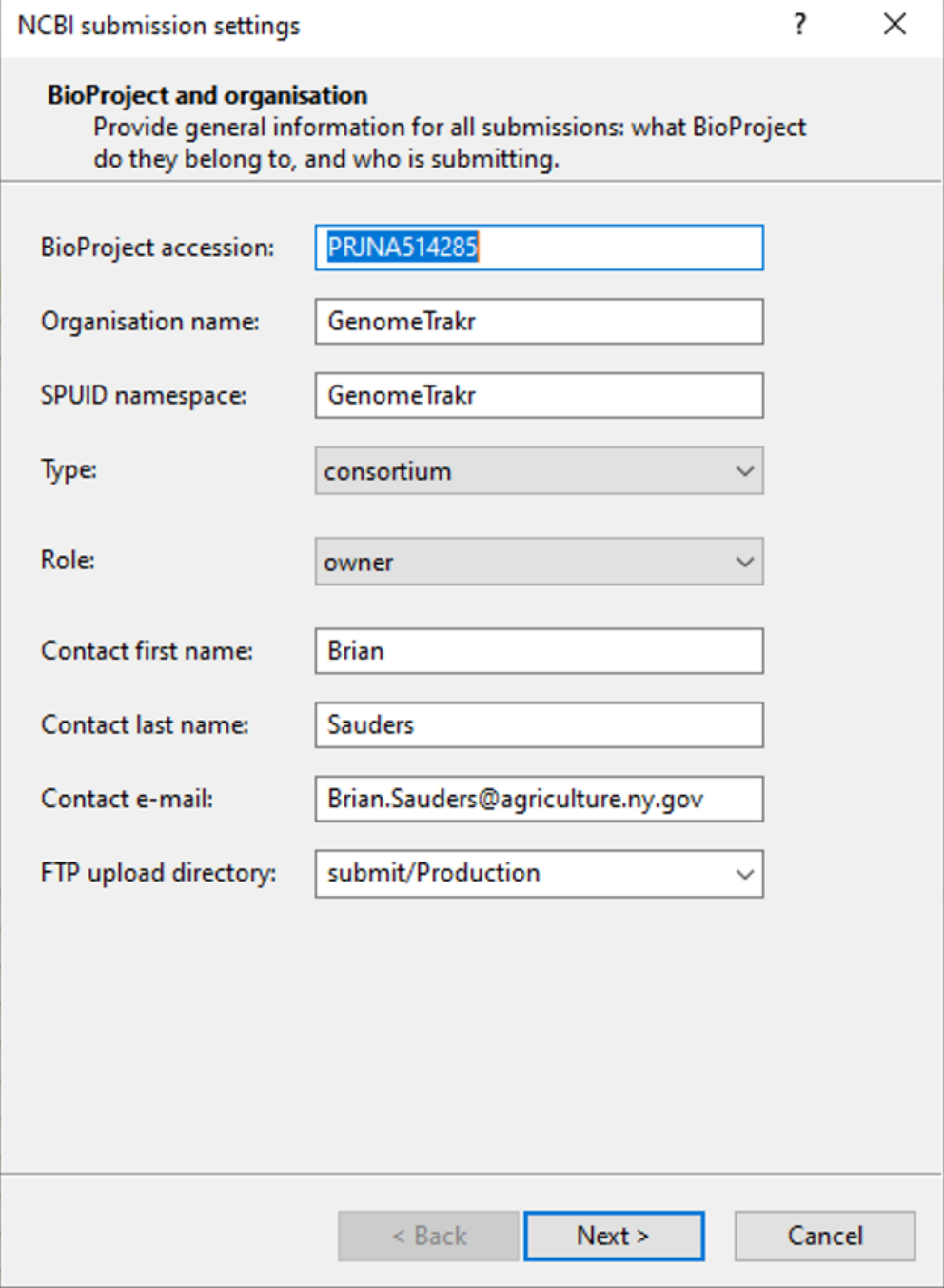
	A	B	C	D
				wtax.cgi. Check spelling carefully! Levels of valid organism names are as follows: Genus species: <i>Salmonella enterica</i> <i>Listeria monocytogenes</i> Genus species and subspecies: <i>Salmonella enterica</i> subsp. <i>enterica</i> Determined serotype or serovar (trad or WGS-based): <i>Escherichia coli</i> O104:H7 <i>Salmonella enterica</i> subsp. <i>enterica</i> serovar Agnoa <i>Salmonella enterica</i> subsp. <i>diarizonae</i> serovar 16:z10:e,n,x,z15 <i>Listeria monocytogenes</i> serotype 1/2a If NCBI doesn't have the desired organism name, enter the name determined by your laboratory. After submission, a "taxonomy consult" will take place to evaluate the new name. Sometimes the organism name is changed to a canonical serovar name and the submission proceeds. It is also possible that the serovar is a novel one not currently in the NCBI database and the Taxonomy team will work with the submitter to get the new name added to the database.

Table 1: Metadata attributes for GenomeTrakr

3 NCBI Submission Settings (Manage submission template)

Create the NCBI metadata template in BioNumerics following PulseNet instructions making sure fields are populated according to GT requirements which are described in the following steps.

- 3.1 **BioProject and Organization:** GenomeTrakr labs by submitting independently become owners of their data and are responsible for managing individual bioprojects for each sequenced organism. The term '**field content**' denotes that the **template value** e.g. BioProject accession is mapping to the **field in BioNumerics** e.g. NCBI_bioproject.



NCBI submission settings ? X

BioProject and organisation
Provide general information for all submissions: what BioProject do they belong to, and who is submitting.

BioProject accession:

Organisation name:

SPUID namespace:

Type:

Role:

Contact first name:

Contact last name:

Contact e-mail:

FTP upload directory:

< Back Next > Cancel

Fig 1. NCBI Submission Template: BioProject and Organization



- 3.2 Laboratories will be submitting to specific bioprojects for lab/organisms. Find the **organism/lab** specific bioproject **under each of the GenomeTrakr umbrella bioprojects** included at <https://www.ncbi.nlm.nih.gov/bioproject/593772>

Make sure to submit to your lab bioproject. **Please don't submit to umbrella bioprojects.**

- 3.3 **BioSample:** Metadata associate to the isolate might require the creation of new fields in BioNumerics. The term '**field content**' denotes that the **template value** e.g. Organism name is mapping to the **field in BioNumerics** e.g. OrganismName. The template values might map to **default values** e.g. Pathogen: environmental/food/other; version 1.0. Make sure to include the metadata associated to the isolates in the **mandatory fields** such as: Submitter Provided Unique ID, BioSample accession (output), Organism name, Title, Attribute package, Strain name, Isolate name alias and Project name. Isolate name alias is a **mandatory field** for **GenomeTrakr submissions**. Provide serovar when available.

NCBI submission settings
?
X

BioSample
Information fields to be used as BioSample metadata.

Submitter Provided Unique ID:
Entry Key

BioSample accession (output):
NCBI_ACCESSION (field content)

Organism name:
Listeria monocytogenes

Title:
Listeria monocytogenes

Attribute package:
Pathogen: environmental/food/oth

Strain name:
Entry Key

Serovar (optional):
<Missing>

Isolate (optional):
WGS_id (field content)

Isolate name alias (optional):
<Missing>

Project name:
ProjectName (field content)

< Back
Next >
Cancel

Fig 2. NCBI Submission Template:BioSample



- 3.4 **BioSample:** Make sure to include the metadata associated to the isolates in the **mandatory fields** such as: Collected by, Collection / Isolate date, Collection / Isolate date format, Title, Geographical origin, Isolate source, Sequenced by and Source name/type. Isolate name alias is a mandatory field for GenomeTrakr submissions. Provide Geographical coordinates when available. For human, animal, and plant hosts, include the full taxonomic name of the host when available, "Homo sapiens" or "Bos Taurus". Animal livestock terms are also acceptable entries, e.g. porcine, bovine, equine, etc.

The screenshot shows a window titled "NCBI submission settings" with a question mark icon and a close button. Below the title bar, the text "BioSample (continued)" is displayed, followed by "Information fields to be used as BioSample metadata." The form contains several fields, each with a label and a dropdown menu:

- Collected by:** CollectedBy (field content)
- Collection / Isolation date:** IsolatDate (field content)
- Collection / Isolation date format:** YYYY-mm-dd
- Geographical origin:** SourceCountry:SourceState (field cc
- Geographical coordinates:** <Missing>
- Isolation source:** TypeDetails (field content)
- Host:** Host (field content)
- Host disease:** Host disease (field content)
- Sequenced by:** SequencedBy (field content)
- Source name / type:** SourceType (field content)

At the bottom of the window, there are three buttons: "< Back", "Next >" (which is highlighted with a blue border), and "Cancel".

Fig 3. NCBI Submission Template:BioSample

3.5 NCBI submission settings – SRA Experiment and Run

Populate fields for SRA Experiment and Run according to PulseNet instructions.

NCBI submission settings
?
X

SRA Experiment and Run

Input SRS experiment type:

wgs

Output SRS experiment type:

wgs

Output SRS accession field:

SRR_id (field content)

Instrument model:

Illumina MiSeq

Library name:

Illumina DNA Prep

Library strategy:

WGS

Library source:

GENOMIC

Library selection:

RANDOM

Library layout:

PAIRED

< Back

Next >

Cancel

Fig 4. NCBI Submission Template forBioNumerics, SRA Experiment and run: Make sure to map collection attributes to the corresponding fields.

4 **NCBI submission settings – Submission Template**

Save submission template according to PulseNet Instructions as **-GenomeTrakr-Template-**.

5 **Import data**

- 5.1 Import the ***GenomeTrakr Metadata form for BioNumerics*** according to PulseNet Instructions.
- 5.2 When importing rules, the field source should match destination fields.
- 5.3 In the importing links section, choose the **-key-** for linking records to database entries.
- 5.4 Proceed with sequencing data import according to PulseNet Instructions.
- 5.5 Submit data to NCBI according to PulseNet Instructions. If NCBI accessions are not available at BioNumerics in 1 business day, please contact NCBI and PulseNet to troubleshoot issues with submissions.
- 5.6 Contact GenomeTrakr by email genometrakr@fda.hhs.gov if issues with submissions are delayed for more than 3 days. GenomeTrakr can support urgent submissions if needed.