



VERSION 2
DEC 21, 2023

OPEN  ACCESS



DOI:
dx.doi.org/10.17504/protocols.io.36wgq3ejylk5/v2

Protocol Citation: Ruth Timme, Maria Balkey, Julie Haendiges, Brian Sauders, Tina.Pfefer 2023. Populating NCBI template for submissions using BioNumerics . **protocols.io** <https://dx.doi.org/10.17504/protocols.io.36wgq3ejylk5/v2> Version created by [Maria Balkey](#)

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

Populating NCBI template for submissions using BioNumerics V.2

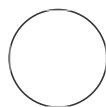
Maria Ruth Timme¹, Balkey¹, Tina.Pfefer² Brian Julie Haendiges¹, Sauders²,

¹US Food and Drug Administration;

²New York State Department of Agriculture & Markets

GenomeTrakr

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



Maria Balkey

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: Dropdown menus from controlled vocabulary added to the ncbi_update submission sheet

Protocol status: Working
We use this protocol and it's working

Created: Dec 20, 2023

Last Modified: Dec 21, 2023


PROTOCOL integer ID:
92592

Keywords: NCBI submission,
BioNumerics, biosample,
SRA, metadata, bioproject

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

 GT_BioNumerics_spreadsheet_v3.xlsx33KB

Create the fields NCBI_bioproject, Attribute_package, Organism_name, NCBI_LabID, SourceCountryState, Latitude_longitude, Reference_material, Culture_collection, ProjectName, SequencedBy or Description if **they are not** in the BioNumerics interface and are needed to process the metadata for your isolates.

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

2 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.

- 2.1 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:

consortium

Role:

owner

Contact first name:

Contact last name:

Contact e-mail:

FTP upload directory:

submit/Production

< Back

Next >

Cancel

Fig 1. NCBI Submission Template: BioProject and Organization

A	B	C
Name of Field in BioNumerics Template	Description	Example

A	B	C
BioProject accession	Identifier for NCBI data collection that contains data associated to GenomeTrakr. Specific for organism and lab submitter	PRJNA514285
Organization name	Surveillance Program (example is the default value for GenomeTrakr submissions)	GenomeTrakr
SPUID namespace	Surveillance Program (example is the default value for GenomeTrakr submissions)	GenomeTrakr
Type	organization type (example is the default value for GenomeTrakr submissions)	consortium
Role	laboratory responsibility (example is the default value for GenomeTrakr submissions)	owner
Contact first name	First name for Lab POC for NCBI submissions. Lab might choose to create alias name for WGS team	First Name
Contact last name	Last name Lab POC for NCBI submissions. Lab might choose to create alias name for WGS team	Last Name
Contact e-mail	email for Lab POC for NCBI submissions. Lab might choose to create alias name for WGS team	first.last@lab.gov
FTP upload directory	Name of directory at NCBI FTP site (example is the default value for GenomeTrakr submissions)	submit/Production

Table 1. Guidelines for Bioproject and Organization metadata

2.2 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.**

2.3 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.

The screenshot shows a window titled "NCBI submission settings" with a question mark icon and a close button (X). Below the title bar, the section "BioSample" is highlighted, with the subtitle "Information fields to be used as BioSample metadata." Below this, there are ten rows, each with a label and a dropdown menu:

- 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)

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

Fig 2. NCBI Submission Template:BioSample

A	B	C	D
Name of Field in BioNumerics Template	Description	Name of Field in BioNumerics DataBase	Example of metadata value
Submitter Provided Unique ID	Local lab strain ID	Entry Key	21B00181-5
BioSample accession (output)	NCBI accession will get populated upon submission to NCBI	NCBI_ACCESSION (field content)	SAMN17385051
Organism name	Genus – species for organism	Organism_name (field content)	Listeria monocytogenes
Title	Organism name	Organism_name (field content)	Listeria monocytogenes
Attribute package	Sample category	Pathogen: environmental/food/other; version 1.0	Pathogen: environmental/food/other; version 1.0
Strain name	PNUSA identifier (automatically populates at the time of registration)	WGS_id (field content)	PNUSAL008933
Serovar (optional)	Serotyping information for Escherichia coli and Salmonella enterica	Serovar (field content)	missing
Isolate (optional)	Field is not required for GenomeTrakr	<Missing>	missing
Isolate namea alias (optional)	Optional identifier for collaboration projects	Isolate_name_alias (field content)	21B00181-5; RS_21290
Project name	Name of the project within which the sequencing was organized	ProjectName (field content)	

Table 2. Guidelines for BioSample metadata

2.4 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. **Host or host disease are provided only for isolates obtained from human, indicate "missing" for isolates from food or environmental sources.**

NCBI submission settings
?
X

BioSample (continued)
Information fields to be used as BioSample metadata.

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:

<Missing> ▼

Host disease:

<Missing> ▼

Sequenced by:

SequencedBy (field content) ▼

Source name / type:

SourceType (field content) ▼

< Back
Next >
Cancel

Fig 3. NCBI Submission Template: BioSample 2

A	B	C	D
---	---	---	---

A	B	C	D
Name of Field in BioNumerics Template	Description	Name of Field in BioNumerics DataBase	Example of metadata value
Collected by	Full name of laboratory that collected the sample or has taken over curation of the isolate.	NCBI_LabID (field content)	NY Department of Agriculture and Markets
Collection date	Date on which the sample was collected.	IsolateDate (field content)	2020
Geographical location	Country and State for sample collection	SourceCountryState (field content)	USA:NY
Geographical coordinates	latitude and longitude for site of collection. Missing if it is not provided	<Missing>	missing
Isolation source	Detailed description for sample product or environmental source	SourceSite (field content)	cheese
Host	Only provided for human isolates	<Missing>	missing
Host disease	Only provided for human isolates	<Missing>	missing
Sequenced by	The name of the agency that generated the sequence, e.g., Centers for Disease Control and Prevention	Sequenced by (field content)	
Source name / type	Controlled vocabulary describing the isolation_source. Choose the best fit term: Human, Animal, Food, Environmental, Other.	SourceType (field content)	

Table 3. Guidelines for BioSample metadata (2)

2.5 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.

3 NCBI submission settings – Submission Template

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

4 Import data

- 4.1 Import the ***GenomeTrakr Metadata form for BioNumerics*** (GT_BioNumerics_spreadsheet_v2.csv) according to PulseNet Instructions.
- 4.2 When importing rules, the field source should match destination fields.
- 4.3 In the importing links section, choose the -key- for linking records to database entries.
- 4.4 Proceed with sequencing data import according to PulseNet Instructions.
- 4.5 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.
- 4.6 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.

