

Nov 01, 2021

Version 7

Guidance for populating GenomeTrakr metadata templates (BioSample and SRA) V.7

 In 1 collection

DOI

<https://dx.doi.org/10.17504/protocols.io.by2mpyc6>

Ruth Timme¹, Maria Balkey¹, William Wolfgang², Errol Strain¹

¹US Food and Drug Administration; ²Wadsworth Center NYSDOH

GenomeTrakr

Springer Nature Books



Ruth Timme

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.by2mpyc6>

Protocol Citation: Ruth Timme, Maria Balkey, William Wolfgang, Errol Strain 2021. Guidance for populating GenomeTrakr metadata templates (BioSample and SRA). protocols.io <https://dx.doi.org/10.17504/protocols.io.by2mpyc6> Version created by Ruth Timme



Manuscript citation:

Timme, R.E., Wolfgang, W.J., Balkey, M. et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook 2, 20 (2020). <https://doi.org/10.1186/s42522-020-00026-3>

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: October 13, 2021

Last Modified: November 10, 2021

Protocol Integer ID: 54061

Keywords: GenomeTrakr, metadata, Pathogen package, NCBI Pathogen Detection, INSDC, genometrakr metadata template, populating genometrakr metadata template, extended pathogen package, biosample metadata, generic sequence metadata template for sra submission, pathogen package, draft one health enteric package, generic sequence metadata template, picklists in the genometrakr, genometrakr, health enteric package, interoperability for foodborne pathogen surveillance, sra metadata, metadata package, foodborne pathogen surveillance, biosample, incremental update file for the draft, sra submission, guideline

Disclaimer

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

Abstract

PURPOSE: Guidance on how to populate NCBI's metadata packages, maximizing interoperability for foodborne pathogen surveillance.

SCOPE: This protocol provides detailed instructions for populating the following two templates:

1. **BioSample metadata:** guidelines to populate the GenomeTrakr-extended pathogen package.
2. **SRA metadata:** NCBI's generic sequence metadata template for SRA submissions.

Versions:

v6: Added the One Health Enteric package presented at IAFP 2021 meeting.

v7: Updated the picklists in the GenomeTrakr-extended pathogen package, "**GT-pathogen package-OHE**

v0.2.2.xlsx" and added an incremental update file for the **DRAFT One Health Enteric Package** that includes extensive edits compared to v6.



Materials

Gather the following contextual information for each pure culture isolate:

1. organism name
2. lab name that collected the sample
3. collection date
4. collection source
5. Geographic location of sample collection

Before start

Before collecting sequence data for your isolates, ensure that you can provide the minimum metadata recommended by your coordinating surveillance body. The INSDC, in collaboration with the Global Microbial Identifier (GMI) (<https://www.globalmicrobialidentifier.org>), recommends using the Pathogen metadata template for pathogen surveillance submissions: (NCBI: <https://www.ncbi.nlm.nih.gov/pathogens/submit-data/> and EMBL-EBI: <https://www.ebi.ac.uk/ena/submit/pathogen-data>).

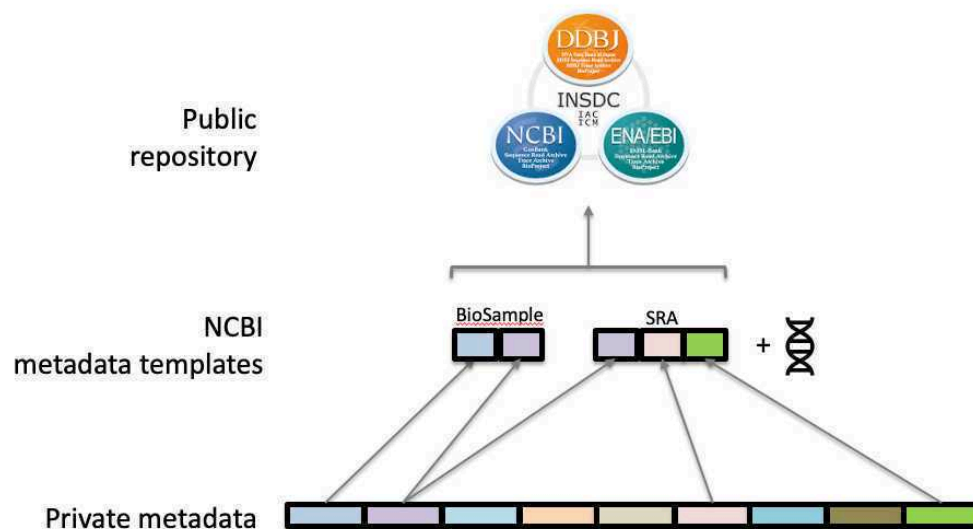
Overview

- 1 **Guidance for organizing and populating the metadata templates required for direct submission to NCBI. This guidance is applicable for most enterics and/or microbial pathogens.**

******If your laboratory uses the BioNumerics platform for submission, please follow this [protocol](#).******

Two metadata templates are required:

1. BioSample metadata (metadata describing the sample source and submitter)
2. SRA metadata (metadata describing the sequence data collection)



BioSample metadata template

- 2 **Template for BioSample submission:**

Download the **GenomeTrakr-extended pathogen package** and follow the guidance included in this template.

 GT-pathogen package-OHE v0.2.2....

- 2.1 **DRAFT One Health Enteric Package, announced at IAFP 2021 (and the fall GenomeTrakr meeting) is ready for review and comment.** Please download the file, review the attributes, and provide feedback here or email directly to ruth.timme@fda.hhs.gov.



One Health Enteric Package-DRAF...

Safety information

NOTE: DO NOT USE THIS VERSION FOR NCBI SUBMISSION. We expect the v1.0 release to be ready for use in Fall 2021.

SRA sequence metadata template

3 Template for SRA metadata submission:

Download the "**Metadata spreadsheet with sample names**" file from the NCBI Submission Templates page:

<https://submit.ncbi.nlm.nih.gov/templates/>

And follow the guidance in the following table:

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. *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.
3. 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.

3.1

A	B	C
Field	Description	Example
sample_name	Include the same ID here as you entered for "sample_name" in the	UT-12345



A	B	C
	BioSample submission template. Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".	
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. Populate this field using the values in the PHA4GE specification for "library_id".	UT-12345.6
Title	Short, free text description that identifies the data on public pages. For Example: {methodology} of {organism}: {sample_name}	WGS of Salmonella enterica: UT-12345
library_strategy	Overall sequencing strategy or approach. Choose from NCBI pick list	See NCBI SRA pick list. (e.g. WGS)
library_source	molecule type used to make the library	See NCBI SRA pick list. (e.g. Genomic)
library_selection	Library capture method	See NCBI SRA pick list. (e.g. random, PCR)
Library_layout	Choose from NCBI pick list	See NCBI SRA pick list, choose "paired"
platform	Sequencing platform	See NCBI SRA pick list. (e.g., Illumina).
instrument_model	Name of the sequencing instrument.	See NCBI SRA pick list. (e.g. Illumina MiSeq, iSeq 100)
Design_description	optional field for free text description of methods	
Filetype	File format name for the raw sequence data Choose from NCBI pick list	See NCBI SRA pick list. (e.g. Fastq)



A	B	C
Filename	include ALL of the files resulting from this library. **Add additional fields if there are more than two files (e.g. Filename3). Populate this field using the values in the PHA4GE specification for "r1 fastq filename".	genome_r1.fastq (*must be exact)
Filename2	genome_r2.fastq (*must be exact) Populate this field using the values in the PHA4GE specification for "r2 fastq filename".	genome_r2.fastq (*must be exact)
Filename3-8	list other fastq file names (e.g. 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.