Version 4

Apr 15, 2021

Guidance for populating GenomeTrakr metadata templates (BioSample and SRA) V.4

In 1 collection

Ruth E Timme[1], Maria Balkey[1], William Wolfgang[2], Errol Strain[1]

[1]US Food and Drug Administration; [2]Wadsworth Center NYSDOH

In Development dx.doi.org/10.17504/protocols.io.bnhemb3e

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

Ruth Timme
US Food and Drug Administration

ABSTRACT

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

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

1. **BioSample metadata**: guidelines to populate the custom One Health Enteric Package, which is a modification of NCBI's combined pathogen package.

2. **SRA metadata:** NCBI's generic sequence metadata template for SRA submissions.

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

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

DOI

dx.doi.org/10.17504/protocols.io.bnhemb3e

PROTOCOL CITATION

Ruth E 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.bnhemb3e
Version created by Ruth Timme

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

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

COLLECTIONS

GenomeTrakr data collection and submission workflow

KEYWORDS

GenomeTrakr, metadata, Pathogen package, NCBI Pathogen Detection, INSDC

CREATED

Oct 15, 2020

LAST MODIFIED

Apr 15, 2021

PROTOCOL INTEGER ID

43270

PARENT PROTOCOLS

Part of collection

GenomeTrakr data collection and submission workflow

MATERIALS TEXT

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 STARTING

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

Overview

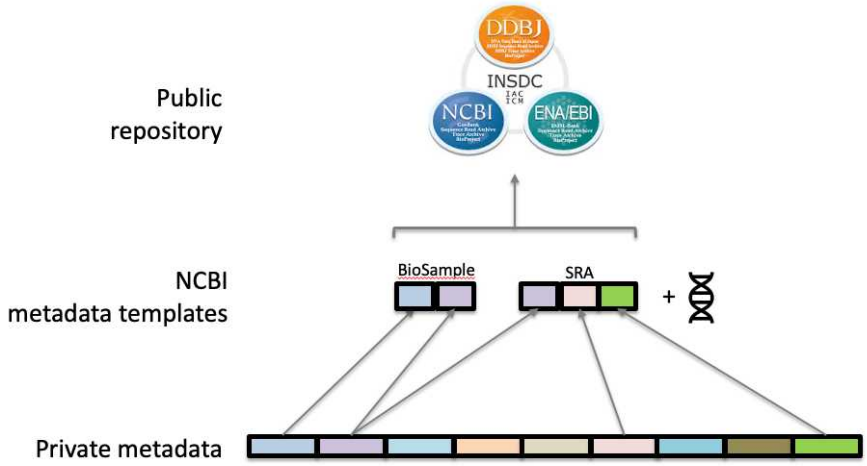
1 **Guidance for organizing and populating the metadata templates required for direct submission to NCBI. This guidance is applicable for most enterics and/or microbial pathogens.**

****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)

Citation: Ruth E Timme, Maria Balkey, William Wolfgang, Errol Strain (04/15/2021). Guidance for populating GenomeTrakr metadata templates (BioSample and SRA). https://dx.doi.org/10.17504/protocols.io.bnhemb3e



BioSample metadata

2 Template for BioSample submission:

Download the One Health Enteric Package and follow the guidance included in this template.

One Health Enteric Package v0.2.xlsx

SRA sequence metadata template

3 Populate SRA's batch metadata table:

Download template here and follow the guidance in the following table:
ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata.xlsx

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

Citation: Ruth E Timme, Maria Balkey, William Wolfgang, Errol Strain (04/15/2021). Guidance for populating GenomeTrakr metadata templates (BioSample and SRA). https://dx.doi.org/10.17504/protocols.io.bnhemb3e

A	B	C
Field	Description	Example
sample_name	Include the same ID here as you entered for "sample_name" in the BioSample submission template. Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".	UT-12345
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)
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)	

Citation: Ruth E Timme, Maria Balkey, William Wolfgang, Errol Strain (04/15/2021). Guidance for populating GenomeTrakr metadata templates (BioSample and SRA). https://dx.doi.org/10.17504/protocols.io.bnhemb3e

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.

Citation: Ruth E Timme, Maria Balkey, William Wolfgang, Errol Strain (04/15/2021). Guidance for populating GenomeTrakr metadata templates (BioSample and SRA). https://dx.doi.org/10.17504/protocols.io.bnhemb3e