Apr 22, 2020

Populating the NCBI pathogen metadata template V.3

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

[1]US Food and Drug Administration, [2]NY Wadsworth Laboratory

1 Works for me dx.doi.org/10.17504/protocols.io.be9gjh3w

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

Ruth Timme
US Food and Drug Administration

ABSTRACT

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

SCOPE: This protocol provides detailed instructions for populating the core metadata fields within the Pathogen metadata package.

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.*

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 surveillaince 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).

Citation: Ruth Timme, Maria Balkey, William Wolfgang, Errol Strain (04/22/2020). Populating the NCBI pathogen metadata template. https://dx.doi.org/10.17504/protocols.io.be9gjh3w

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

Background

1 **Download the pathogen metadata package from NCBI:**

Navigate to BioSample submission: https://submit.ncbi.nlm.nih.gov/subs/biosample

Click on "Download batch submission template", then select the "**Pathogen affecting public health**" and the appropriate package depending on the type of isolates. We recommend using the combined template for simplicity.

Direct link to download the packages: https://submit.ncbi.nlm.nih.gov/biosample/template

Follow the GenomeTrakr guidance below for populating the minimum set of metadata fields. Following these guidelines closely will ensure that your submissions will be fully interoperable within the rest of the database.

Pathogen package attributes

2 **strain**

This is the authoritative ID used within NCBI Pathogen Detection and for thePulseNet/GenomeTrakr networks. Although the strain ID can have any format, we suggest that it be unique, concise, and consistent within your laboratory (e.g. CFSAN123456). There are downstream advantages to the name being entirely alpha-numeric, so avoid special characters if possible.

3 **sample_name**

Sample Name is another unique identifier for the pure culture isolate and required by NCBI for BioSample submission (it cannot be left blank). It can have any format, but we suggest that it be the same as the strain name or contain another identifier important to the isolate or submitting laboratory. NCBI validates this attribute for uniqueness, so you cannot use "missing, or "not collected". This identifier is NOT available in NCBI-PD.

4 **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/wwwtax.cgi. Check spelling carefully! Levels of valid organism names are as follows:

Genus species:
Salmonella enterica
Listeria monocytogenes

Genus species and subspecies:
Salmonella enterica subsp. *enterica*

Determined serotype or serovar (trad or WGS-based):
Escherichia coli O104:H7
Salmonella enterica subsp. *enterica* serovar Agnoa
Salmonella enterica subsp. *diarizonae* serovar 16:z10:e,n,x,z15
Listeria monocytogenes 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.

Citation: Ruth Timme, Maria Balkey, William Wolfgang, Errol Strain (04/22/2020). Populating the NCBI pathogen metadata template.
https://dx.doi.org/10.17504/protocols.io.be9gjh3w

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 Guidenance for other taxonomy attributes in the pathogen package:

- serovar
- serotype

NCBI will autopopulate these fields based on the name included in the organism field, or updates made to the BioSample.

Leave these attributes blank!

5 **collected_by**

Name of laboratory that sequenced the isolate (or institute that collected the sample). Abbreviations are ok if they are well-known in the community (e.g. FDA or CDC)

6 **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.

6.1 **host**

*For Pathogen.cl only: "Homo sapiens" if clinical isolate.

6.2 **host_disease**

*For Pathogen.cl only: Name of relevant disease, e.g., Salmonella gastroenteritis. This field must use controlled vocabulary provided at:http://bioportal.bioontology.org/ontologies/1009orhttp://www.ncbi.nlm.nih.gov/mesh. Label this field "not collected" if unknown for clinical isolates. Leave blank for all Pathogen.env isolates.

7 **collection_date**

Date of sampling in 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 valuble for accessing seasonality in the database.

8 **geo_loc_name**

Geographical origin of the sample using controlled vocabulary: http://www.insdc.org/country. Use a colon to separate the country or ocean from more detailed information about the location, e.g., "Canada: Vancouver". Country and state are required for GenomeTrakr isolates from the US, e.g. "USA: CA".

Packaged food guidance: list the country or state of origin listed on the label. If no originating source is listed then include the location of purchase (country: state).

Citation: Ruth Timme, Maria Balkey, William Wolfgang, Errol Strain (04/22/2020). Populating the NCBI pathogen metadata template. https://dx.doi.org/10.17504/protocols.io.be9gjh3w

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

9 **isolation_source**

Describes the physical, environmental and/or local geographical sample from which the organism was derived. Avoid generic terms such as patient isolate, sample, food, surface, clinical, product, source, environment.

Food samples: provide a precise description of the food without including product brands or firm names. E.g. bagged romaine lettuce, chicken breast, frozen shrimp, cilantro, ground turmeric, etc. Specifiy type of vegtable, milk, cheese, flour, and seafood. Do not use acronyms.

Environmental samples: specify natural geographic features. E.g. agricultural soil, fresh water stream, irrigation pond, river sediment, etc.

Facility or farm inspection samples: "Environmental swab" is the standard term in the database for facility inspection samples, however include more information if possible. E.g. environmental swab from sink at retail food establishment, environmental sponge from floor drain at food processing facility. For farm samples specify the type of sample collected, e.g. irrigation water from farm, or soil from farm.

Animal clinical samples: provide the type of specimen, organ (nasal swab, skin ulcer, fecal, spleen biopsy, etc) and host binomial in parentheses. E.g. nasal swab (*Equus ferus caballus*).

Animal feed: specifiy the type of feed, incluing the intended animal. E.g. bovine animal feed, or poultry feed.

10 **bioproject_accession**

The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxxx). This cannot be left blank.

**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.

11 **lat_lon**

Provide latitude and longitude to support geo_loc_name. This field is required to be populated by NCBI. However, if this level of detail is not availabe, GenomeTrakr recommends including "missing" or "not collected" here.

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

Citation: Ruth Timme, Maria Balkey, William Wolfgang, Errol Strain (04/22/2020). Populating the NCBI pathogen metadata template. https://dx.doi.org/10.17504/protocols.io.be9gjh3w

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