Jul 08, 2020

SOP for populating NCBI submission templates for SARS-CoV-2 (BioSample, SRA, and GenBank)

Forked from Populating the NCBI pathogen metadata template

Ruth Timme[1], Emma Griffiths[2], Duncan MacCannell[3], Stacia Wyman[4]

[1]US Food and Drug Administration; [2]University of British Columbia; [3]US Center for Disease Control and Prevention; [4]Innovative Genomics Institute, UC Berkeley

1 *Works for me* dx.doi.org/10.17504/protocols.io.bf89jrz6

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

Ruth Timme
US Food and Drug Administration

ABSTRACT

PURPOSE: Guidance on how to populate the three NCBI metadata submission templates for SARS-CoV-2 submissions, maximizing interoperability for COVID-19 surveillance.

DOI

dx.doi.org/10.17504/protocols.io.bf89jrz6

PROTOCOL CITATION

Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman 2020. SOP for populating NCBI submission templates for SARS-CoV-2 (BioSample, SRA, and GenBank). **protocols.io** dx.doi.org/10.17504/protocols.io.bf89jrz6

FORK FROM

Forked from Populating the NCBI pathogen metadata template, Ruth Timme

KEYWORDS

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

CREATED

May 11, 2020

LAST MODIFIED

Jul 08, 2020

PROTOCOL INTEGER ID

36865

Three templates needed for NCBI SARS-CoV-2 submission

1 **Guidance for populating the three NCBI metadata templates for SARS-CoV-2 submission.**

START HERE FIRST: Read the **PHA4GE contextual data specification** BEFORE populating your submission templates!

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

Steps 3-4 help to map the NCBI submission template fields to the PHA4GE metadata specification, however, the primary PHA4GE guidance should be followed first to ensure the correct controlled vocabularies and ontology terms are used to populate these fields.

Guidance included in this protocol:

- **Step 2)** BioSample metadata template (modified for PHA4GE)
- **Step 3)** SRA metadata template
- **Step 4)** GenBank source modifier template (modified for PHA4GE)

BioSample metadata

2 PHA4GE pathogen template for BioSample submission

Download template here:

□ **BioSample-Pathogen-PHA4GE_200708.xlsx**

2.1

attributes	guidance	example
sample_name	Sample Name is a unique identifier for the sample (it cannot be left blank). Populate with the same ID as "Isolate", or include just the lab ID:. Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".	Example 1: SARS-CoV-2/human/USA/MI-MDHHS-SC20654/2020 Example 2: MI-MDHHS-SC20654
bioproject_accession	The accession number of the BioProject(s) to which the BioSample belongs (PRJNAxxxxxx). A valid BioProject accession has prefix PRJN, PRJE or PRJD. This cannot be left blank. **Double-check that you are submitting to the correct data BioProject for your laboratory, and not the umbrella BioProject for the entire effort. Populate this field using the values in the PHA4GE specification for "bioproject accession".	PRJNA12345
attribute_package	Specify the pathogen type. "Pathogen.cl" (for clinical or host-associated pathogen) or "Pathogen.env" (for environmental, food or other pathogen). The value provided in this field drives validation of other fields.	Pathogen.cl
organism	Populate this field using the values in the PHA4GE specification for "organism".	Severe acute respiratory syndrome coronavirus 2

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

isolate	Full name of the virus: SARS-CoV-2/source/location/isolateID/year Populate this field using the values in the PHA4GE specification for "isolate".	SARS-CoV-2/human/USA/CA-CDPH-001/2020
collected_by	Full name of laboratory or institute that collected the sample. Populate this field using the values in the PHA4GE specification for "sample collected by".	Utah Public Health Laboratory
collection_date	the date on which the sample was collected; "YYYY-mm-dd", "YYYY-mm", or "YYYY". Including the month or month/day of collection is extremely valuable for accessing seasonality in the database. Populate this field using the values in the PHA4GE specification for "sample collection date". If unknown, put "not collected", or other null value like "missing".	2020-05-15
geo_loc_name	Geographical origin of the sample; use the appropriate name from this list http://www.insdc.org/documents/country-qualifier-vocabulary. Use a colon to separate the country from more detailed information about the location, Populate this field combining the PHA4GE specification fields for "geo_loc name (country)" and "geo_loc name (state/province/region)". Put more country field first, followed by state/province/region, separated by a colon. If unknown, put "not collected", or other null value like "missing".	USA: Utah
isolation_source	Describe the sample as "Clinical", "Animal", or "Environmental". Add the statement "See additional sample source fields for further information".	Clinical, see additional sample source fields for further information
lat_lon	The coordinates of the geographical location of sample or host collection. If known, provide the geographical coordinates of the location where the sample was collected. Specify as degrees latitude and longitude in the format "d[d.dddd] N|S d[dd.dddd] W|E", eg, 38.98 N 77.11 W. Populate this field by combining the PHA4GE specification fields "geo_loc latitude" and "geo_loc_longitute" **DO NOT PROVIDE LAT/LON OF THE INSTITUTION, NOR THE CENTER OF A CITY/REGION WHERE THE SAMPLE WAS COLLECTED. If unknown, put "not collected", or other null value like "missing".	38.98 N 77.11 W not collected

host	If the combined attribute package is being used, this field can be left empty for Pathogen.ev isolates. Populate this field using the values in the PHA4GE specification for "host (scientific name)". This field is only required for host-associated samples (Pathogen.cl specified in attribute_package).	Homo sapiens
host_disease	This field is only required for host-associated samples (Pathogen.cl specified in attribute_package). If the combined attribute package is being used, this field can be left empty for Pathogen.ev isolates. Populate this field using the values in the PHA4GE specification for "host_disease". If the host is healthy or the information is unknown, put "missing"	COVID-19
host_health_state	Information regarding health state of the individual sampled at the time of sampling. Populate this field using the values in the PHA4GE specification for "host health state". If the information is unknown, or can not be shared, leave blank or provide a null value.	asymptomatic
host_disease_outcome	Final outcome of disease, e.g., death, chronic disease, recovery Populate this field using the values in the PHA4GE specification for "host disease outcome". If the information is unknown, or can not be shared, leave blank or provide a null value.	e.g., death, chronic disease, recovery
host_sex	The gender of the host at the time of sample collection. Populate this field using the values in the PHA4GE specification for "host gender". If the information is unknown, or can not be shared, leave blank or provide a null value.	male
host_age	Age of host at the time of sampling Populate this field using the values in the PHA4GE specification for "host age". Provide age in years. Age-binning is also acceptable. If the information is unknown or can not be shared, leave blank or provide a null value.	31
host_subject_id	a unique identifier by which each subject can be referred to, de-identified. Populate this field using the values in the PHA4GE specification for "host subject ID". If the information is unknown or can not be shared, leave blank or provide a null value.	clincal123456

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

anatomical_material	Provide a descriptor if an anatomical part was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, use this look-up service to identify a standardized term: https://www.ebi.ac.uk/ols/ontologies/uberon. Populate this field using the values in the PHA4GE specification for "anatomical material". If not applicable, leave blank.	blood
anatomical_part	An anatomical part of an organism e.g. oropharynx. Populate this field using the values in the PHA4GE specification for "anatomical part". If not applicable, leave blank.	nasopharynx
body_product	A substance excreted/secreted from an organism e.g. feces, urine, sweat. Populate this field using the values in the PHA4GE specification for "body product". If not applicable, leave blank.	feces
passage_history	Number of passages Populate this field using the values in the PHA4GE specification for "passage number". If not applicable, leave blank.	3
lab_host	Name and description of the laboratory host used to propagate the organism or material from which the sample was obtained. This field can be populated by the PHA4GE field "lab_host". If not applicable, leave blank.	Vero E6 cell line

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

passage_method	Passage method. Populate this field using the values in the PHA4GE specification for "passage method". If not applicable, leave blank.	AVL buffer+30%EtOH lysate received from Respiratory Lab. P3 passage in Vero-1 via bioreactor large-scale batch passage. P3 batch derived from the SP-2/reference lab strain.
purpose_of_sampling	The reason that the sample was collected. Populate this field using the values in the PHA4GE specification for "purpose of sampling". If not applicable, leave blank.	Select a value from the pick list in the template. (e.g. diagnostic testing)
environmental_material	A substance obtained from the natural or man-made environment e.g. soil, water, sewage, door handle, bed handrail, face mask. Populate this field using the values in the PHA4GE specification for "environmental material". If not applicable, leave blank.	face mask
environmental_site	An environmental location may describe a site in the natural or built environment e.g. hospital, wet market, bat cave. Populate this field using the values in the PHA4GE specification for "environmental site". If not applicable, leave blank.	hospital room
collection_device	The instrument or container used to collect the sample e.g. swab. Populate this field using the values in the PHA4GE specification for "collection device". If not applicable, leave blank.	swab

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

collection_method	The process used to collect the sample e.g. phlebotomy, necropsy. Populate this field using the values in the PHA4GE specification for "collection method". If not applicable, leave blank.	Bronchoalveolar lavage (BAL)
culture_collection	Name of source institute and unique culture identifier for the organism (SARS-CoV-2). See the description for the proper format and list of allowed institutes, http://www.insdc.org/controlled-vocabulary-culturecollection-qualifier If not applicable, leave blank.	
host_specimen_voucher	Identifier for the physical host specimen. Populate this field using the values in the PHA4GE specification for "host specimen voucher". if not applicable, leave blank.	URI example: http://portal.vertnet.org/o/fmnh/mammals?id=33e55cfe-330b-40d9-aaae-8d042cba7542, INSDC triplet example: UAM:Mamm:52179
description	Optional field for additional description of the sample.	

SRA metadata

3 **Populate SRA's batch metadata table:**

Download template here:
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

Field	**Description**	Example

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

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}	Amplicon-based sequencing of SARS-CoV-2: UT-12345
library_strategy	Overall sequencing strategy or approach. Choose from NCBI pick list	See NCBI SRA pick list. (e.g. WGS, RNA-Seq, Amplicon)
library_source	molecule type used to make the library	See NCBI SRA pick list. (e.g. viral RNA, metagenomic)
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. (single, paired)
platform	Sequencing platform	See NCBI SRA pick list. (e.g., Illumina, Oxford_nanopore, PacBio_SMRT).

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

instrument_model	Name of the sequencing instrument. Populate this field using the values in the PHA4GE specification for "sequencing instrument"	See NCBI SRA pick list. (e.g. Illumina MiSeq, iSeq 100, GridION, MinION, PacBio Sequel II)
Design_description	optional field for free text description of methods	ARTIC PCR-tiling of viral cDNA (V3), sequenced on Illumina MiSeq with DNA Flex library prep-kit. Only reads aligned to SARS-CoV-2 reference (NC_045512.2) retained
Filetype	File format name for the raw sequence data Choose from NCBI pick list	See NCBI SRA pick list. (e.g. Fastq, OxfordNanopore_native, PacBio_HDF5)
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.

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

GenBank metadata

4 **Populate GenBank source modifier template:**

GenBank-source_modifiers-PHA4GE_200708.xlsx

4.1

Source modifier	**Guidance**	**Example**
Sequence_ID	ID to link the fasta sequence to these source modifiers for a batch submission. This ID must match the ID listed as the fasta file header. Suggested: use the sample_name submitted to BioSample, or the short lab ID contained within the full isolate name. Populate this field using the values in the PHA4GE specification for "specimen collector sample ID".	CA-CDPH-001
country	This field can be populated by combining the PHA4GE fields "geo_loc name (country)" and "geo_loc name (state/province/region)". Put more country field first, followed by state/province/region, separated by a colon.	USA: Virginia
host	This field is only required for host-associated samples (Pathogen.cl specified in attribute_package). Leave blank for environmental isolates. Populate this field using the values in the PHA4GE specification for "host (scientific name)".	Homo sapiens
isolate	Full name of the virus: SARS-CoV-2/source/location/isolateID/year Example: SARS-CoV-2/human/USA/CA-CDPH-001/2020 Populate this field using the values in the PHA4GE specification for "isolate".	SARS-CoV-2/human/USA/CA-CDPH-001/2020
collection-date	the date on which the sample was collected; "YYYY-mm-dd", "YYYY-mm", or "YYYY" Populate this field using the values in the PHA4GE specification for "sample collection date".	2020-06-04
isolation-source	"Clinical", "Animal", or "Environmental". Add the statement "See additional sample source fields for further information."	clinical; See additional sample source fields for further information
BioSample	The accession number of the BioSample registered for this sample. Populate this field using the values in the PHA4GE specification for "biosample accession".	SAMN15187145
BioProject	The accession number of the BioProject(s) to which the BioSample belongs. Populate this field using the values in the PHA4GE specification for "bioproject accession".	PRJNA625551

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6

Citation: Ruth Timme, Emma Griffiths, Duncan MacCannell, Stacia Wyman (07/08/2020). SOP for populating NCBI submission templates forÃÂ SARS-CoV-2 (BioSample, SRA, and GenBank). https://dx.doi.org/10.17504/protocols.io.bf89jrz6