Version 3

May 03, 2021

Overview of NCBI's SARS-CoV-2 submission process and the metadata required V.3

Version 1 is forked from Populating the NCBI pathogen metadata template

Ruth E E Timme[1], Emma Griffiths[2], Lee Katz[3]

[1]US Food and Drug Administration; [2]University of British Columbia; [3]CDC

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

GenomeTrakr PHA4GE

Ruth E Timme
US Food and Drug Administration

ABSTRACT

PURPOSE:
This protocol explains the metadata requirements for the following two protocols:

Complete in order (1 then 2):

1. **SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject**
   - Step-by-step instructions for establishing a new NCBI laboratory submission account and for creating and linking a new BioProject to an existing umbrella effort.
   - SARS-CoV-2 raw data submission to SRA (Sequence Read Archive) and metadata to BioSample.

2. **SARS-CoV-2 NCBI consensus submission protocol: GenBank**
   Required: established BioProject and BioSamples
   - Submit SARS-CoV-2 assemblies to NCBI GenBank, linking to existing BioProject, BioSamples, and raw data.

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Griffiths, E. J. et al. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. (2020) doi:10.20944/preprints202008.0220.v1. https://www.preprints.org/manuscript/202008.0220/v1

DOI

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

PROTOCOL CITATION

Ruth E E Timme, Emma Griffiths, Lee Katz 2021. Overview of NCBI's SARS-CoV-2 submission process and the metadata required. **protocols.io** https://dx.doi.org/10.17504/protocols.io.buqtnvwn Version created by Ruth E Timme

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

Griffiths, E. J. et al. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. (2020) doi:10.20944/preprints202008.0220.v1. https://www.preprints.org/manuscript/202008.0220/v1

FORK NOTE

Updated the title and the NCBI BioSample template.

FORK FROM

Forked from Populating the NCBI pathogen metadata template, Ruth E Timme

Citation: Ruth E E Timme, Emma Griffiths, Lee Katz (05/03/2021). Overview of NCBI's SARS-CoV-2 submission process and the metadata required. https://dx.doi.org/10.17504/protocols.io.buqtnvwn

KEYWORDS

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

CREATED

May 03, 2021

LAST MODIFIED

May 03, 2021

PROTOCOL INTEGER ID

49651

Three templates needed for NCBI SARS-CoV-2 submission

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

1.1 Training video:

For the visual learners, here is a 10min video summarizing the entire NCBI submission process:

1.2 Assembling the three NCBI metadata templates for SARS-CoV-2 submission:

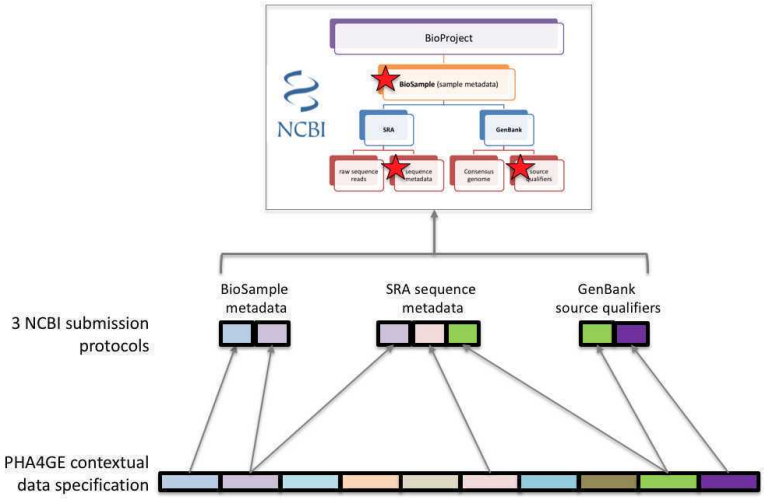
Steps 2-4 provide templates to populate for your submission, however, the primary PHA4GE guidance should be followed first to ensure the correct controlled vocabularies and ontology terms are used to populate these fields.

Guidance included in this protocol:

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

Citation: Ruth E E Timme, Emma Griffiths, Lee Katz (05/03/2021). Overview of NCBI's SARS-CoV-2 submission process and the metadata required. https://dx.doi.org/10.17504/protocols.io.bugtnvwn

PHA4GE contextual data spec. → NCBI templates



BioSample metadata

2 SARS-CoV-2 BioSample submission package:

Download custom version containing the PHA4GE pick-lists and controlled vocabulary:

SARS-CoV-2.cl.1.0_PHA4GEcustom.xlsx

Follow the PHA4GE contextual metadata SOP and guidance posted at NCBI for populating the template.

SRA metadata

3 Populate SRA's batch metadata table:

Download File:

PHA4GE_SRA_template_Feb2021.v2.xlsx

Follow guidance presented in this file for populating the template.

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.

GenBank metadata

4 Populate GenBank source modifier template:

Download file:

PHA4GE_GenBank-source_modifiers_Feb2021.v2.xlsx

Citation: Ruth E E Timme, Emma Griffiths, Lee Katz (05/03/2021). Overview of NCBI's SARS-CoV-2 submission process and the metadata required. https://dx.doi.org/10.17504/protocols.io.bugtnvwn

Follow guidance presented in this file for populating the template.

Citation: Ruth E E Timme, Emma Griffiths, Lee Katz (05/03/2021). Overview of NCBI's SARS-CoV-2 submission process and the metadata required. https://dx.doi.org/10.17504/protocols.io.bugtnvwn