protocols.io

Overview of NCBI's SARS-CoV-2 submission process and the metadata required V.5

Populating the NCBI pathogen metadata template

Ruth Timme[1], Emma Griffiths[2], Lee Katz[3], Michael Weigand[4], Technical Outreach and Assistance for States Team[4]

[1]US Food and Drug Administration; [2]University of British Columbia; [3]CDC; [4]Centers for Disease Control and Prevention

5

Jan 21, 2022

| 1 | ⤳ |

dx.doi.org/10.17504/protocols.io.b35iqq4e

Coronavirus Method Development Community   TOAST_public

Technical Outreach and Assistance for States Team
Centers for Disease Control and Prevention

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

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. Users can modify this protocol to just create a BioSample with no linked raw data.

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.

Version history:
V5: Updated metadata templates (BioSample) for minor edit

DOI

dx.doi.org/10.17504/protocols.io.b35iqq4e

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

Populating the NCBI pathogen metadata template, Ruth Timme

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

protocol ,

Jan 21, 2022

Jan 21, 2022

57226

:

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

Three templates needed for NCBI SARS-CoV-2 submission

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

Direct link to the PHA4GE GitHub repository: https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification/tree/version_3_dev

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
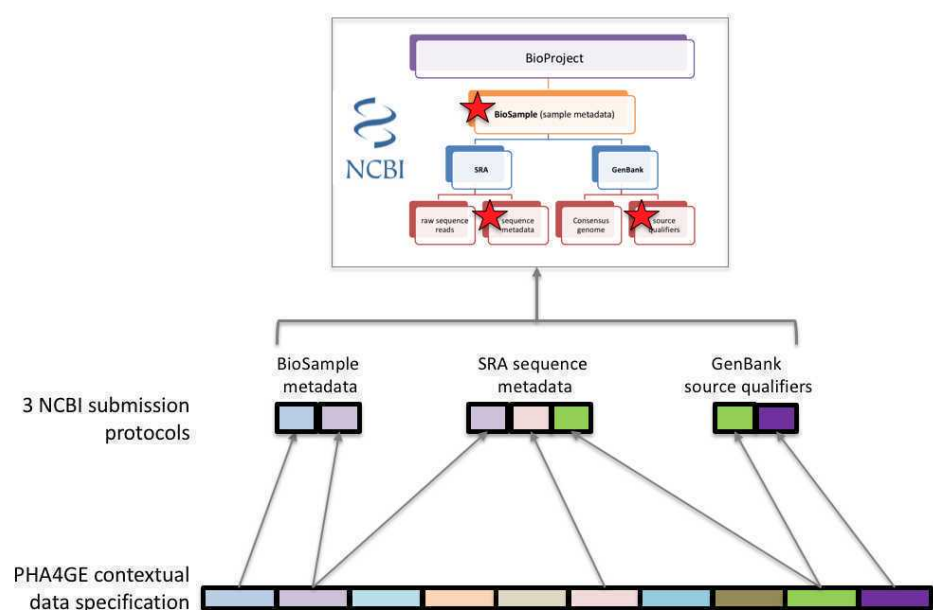
BioSample metadata

2 SARS-CoV-2 BioSample submission package:

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

📄 **SARS-CoV-2.cl.1.0_PHA4GE-V3.1.xlsx**

Follow the PHA4GE contextual metadata SOP and source GitHub repository for guidance in populating the template.

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

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

📄 **SRA_template_PHA4GE-V3.xlsx**

Follow the PHA4GE contextual metadata SOP and source GitHub repository for guidance in 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.

4 **Populate two GenBank templates.**

1. GenBank structured comment (metadata describing the mapping or assembly methods)

📄 **GenBank-structuredComment_PHA4GE-V3.xlsx**

2. GenBank source modifier template. This is a custom version containing PHA4GE guidance and direct linkage to the respective BioSample records. Follow guidance presented in this file for populating the template.

📄 **GenBank-source_modifiers_PHA4GE-V3.xlsx**